



CAROLINA MAGALHÃES LOPES

Bachelor in Science and Computer Engineering

NEWSMQA: A MULTIMODAL QUESTION ANSWERING BENCHMARK OVER NEWS PIECES

MASTER IN COMPUTER SCIENCE

NOVA University Lisbon
March, 2023



DEPARTMENT OF
COMPUTER SCIENCE

NEWSMQA: A MULTIMODAL QUESTION ANSWERING BENCHMARK OVER NEWS PIECES

CAROLINA MAGALHÃES LOPES

Bachelor in Science and Computer Engineering

Adviser: David Semedo
Assistant Professor, NOVA School of Science and Technology

MASTER IN COMPUTER SCIENCE

NOVA University Lisbon

March, 2023

NewsMQA: A Multimodal Question Answering Benchmark over News Pieces

Copyright © Carolina Magalhães Lopes, NOVA School of Science and Technology, NOVA University Lisbon.

The NOVA School of Science and Technology and the NOVA University Lisbon have the right, perpetual and without geographical boundaries, to file and publish this dissertation through printed copies reproduced on paper or on digital form, or by any other means known or that may be invented, and to disseminate through scientific repositories and admit its copying and distribution for non-commercial, educational or research purposes, as long as credit is given to the author and editor.

Gostaria de dedicar esta tese ao meu avô.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude to my advisor Dr. David Semedo for all of the support, and encouragement along with his availability during the course of this thesis. I am grateful that my work has been partially funded by the FCT project NOVA LINCS Ref. UIDB/04516/2020), which granted me a research scholarship of 9 months.

Secondly, I would like to thank all my colleagues and friends that helped with the assembly of my dataset: Diogo Mateus, Diogo Tavares, Diogo Silva, Pedro Valente, Rafael Ferreira, Tiago Soares, Xavier Pacheco, João Antão, Pedro Camponês, Rafael Cabaço, Sofia Canas, Joana Crespo, José Dias, Gonçalo Dias e Gustavo Gonçalves.

Furthermore, I would like to thank my lab besties: Diogos, Pedro and Rafael for the motivation, feedback regarding my "mancha gráfica" and padel games. A special thank you to my second family: Didi, Titi and Xavier for being my best friends and for all your encouragement and patience throughout our college adventure. <3

Finally, I would like to thank my family, mãe, pai e maninho for all their support, and help during all my life.

*“Life is tough, and things don’t always work out well, but we
should be brave and go on with our lives.” (Min Yoongi)*

*“That is fake news, it’s totally fake news, made up, fake.”
(Donald Trump)*

ABSTRACT

News articles are one of the main sources of information available to people all around the world. These are highly diverse documents that cover a wide range of topics and refer to various entities, relationships and concepts spanning through a multitude of domains. The illustrations accompanying these articles play a crucial role in capturing the readers' attention and adding to the storytelling.

In complex domains such as news, question-answering can be an effective method for delivering acute information-seeking. Due to the open-domain and multimodal nature of news, where images complement the textual medium, QA systems are required to dynamically multiplex both visual and linguistic sources to ground answers. To create such a system, we need to combine both natural language and visual information to find an answer. Moreover, the model needs to answer open-ended questions and learn multimodal representations, correlating both images and text elements.

In this thesis, we propose NewsMQA, a novel dataset and benchmark for the task of Multimodal Question-answering for News. NewsMQA differs from existing datasets by fully enclosing the multimodal facets of news and improving on the quality vs. scale trade-off. We adopt a two-part approach that combines human annotation with synthetic question-answer generation through answer roundtrip consistency.

We comprehensively study the created dataset, highlighting its unique characteristics, features, quality, and the research challenges of the task that it supports. To benchmark the dataset, we leverage pre-trained Transformers and propose different strategies to extend these with visual information extracted from corresponding images. We conduct an extensive evaluation of the intricacies and challenges of the dataset and provide insights regarding the impact of enriching the input of these models with image-related information. Finally, we provide a critical discussion regarding the best performing approaches and discuss the task open challenges.

Keywords: News Media, Dataset Creation, Transformers, Multimodal Question Answering, Natural Language Processing

RESUMO

Artigos de notícias são uma das principais fontes de informação disponíveis para pessoas em todo o mundo. São documentos que cobrem uma ampla gama de tópicos e que referem várias entidades, relacionamentos e conceitos que abrangem uma diversidade de domínios. As ilustrações que acompanham estes artigos desempenham um papel crucial em capturar a atenção dos leitores e contribuem para o enriquecimento da narrativa.

Em domínios complexos como as notícias, métodos de resposta a perguntas podem ser bastante eficazes para fornecer informações ao utilizador. Devido à natureza multimodal das notícias, estes sistemas deveriam contemplar tanto estas fontes visuais como as linguísticas para conseguir responder corretamente. Para criar um sistema assim, é necessário abordar tarefas como processamento de linguagem natural e de visão. Além disso, é preciso responder a perguntas abertas e aprender representações multimodais, correlacionando as imagens e o texto.

Os conjuntos de dados existentes para notícias são bastante limitados e ocultam a complexidade do problema. Propomos NewsMQA, um conjunto de dados para a tarefa de Resposta a Perguntas Multimodais para Notícias. Este difere dos conjuntos de dados existentes, contemplando as facetas multimodais das notícias e procurando uma boa relação qualidade-tamanho. Em relação à última, sugerimos uma abordagem que combina anotação humana com geração sintética de perguntas e respostas.

Fornecemos uma análise abrangente do conjunto de dados introduzido, destacando as suas características e desafios que suporta. Para avaliar o nosso conjunto de dados, utilizamos Transformers pré-treinados e propomos estender estes modelos para suportar multimodalidade, incorporando informações extraídas das imagens na sequência de entrada fornecida a estes. Realizamos um conjunto de análises e estudos com os quais avaliamos e discutimos a complexidade e desafios do conjunto de dados e fornecemos a nossa percepção sobre as melhores informações que podemos usar para enriquecer os modelos e melhorar sua performance na resolução da tarefa formulada.

Palavras-chave: Artigos de Notícias, Resposta a perguntas multimodais, Transformers, Construção de datasets, Processamento de Linguagem Natural

CONTENTS

List of Figures	xii
List of Tables	xiv
1 Introduction	1
1.1 Context and Motivation	1
1.2 Problem Formulation and Objective	2
1.3 Challenges and Research Hypothesis	3
1.4 Contributions and Achievements	4
1.5 Document Organization	4
2 Background and Related Work	6
2.1 News Stories Understanding	6
2.2 Language and Vision Representations	8
2.2.1 Contextualized Text Representations	8
2.2.2 Vision Representations	12
2.2.3 Multimodal Representations	14
2.3 News-related Multimodal Tasks	17
2.3.1 News Image Captioning	17
2.3.2 Visual Storylines	18
2.4 Question Answering	18
2.4.1 Question-Answering Settings	19
2.4.2 Retrieving News Context Passages	22
2.4.3 Text-based Question Answering Models	23
2.4.4 Visual Question Answering Models	25
2.5 On the Creation of Question Answering Datasets	26
2.5.1 SQuAD	27
2.5.2 VQA	27
2.5.3 OK-VQA	28

3	News Multimodal QA Dataset	30
3.1	News Multimodal QA Task Formulation	30
3.2	Dataset Creation Pipeline	31
3.2.1	Image-passage Extraction	31
3.2.2	Question-Answer Candidates Generation	33
3.2.3	Human Validated Set	35
3.2.4	Synthetic Set	38
3.3	Crowdsourcing tasks for Question-Answer pairs Generation	41
3.3.1	Using a Single Task for Dataset Samples Generation	41
3.4	Dataset Analysis and Statistics	43
3.4.1	Dataset Overview	43
3.4.2	Human Validated vs. Synthetic sets	43
3.4.3	Dataset challenges	44
3.4.4	Questions' diversity analysis	45
3.4.5	Questions multi-modality analysis	45
3.4.6	Critical Summary	46
4	Transformer-based models for News Multimodal QA	47
4.1	Transformers for Span-based QA	47
4.2	Visual Domain Context	49
4.2.1	Object Detection	49
4.2.2	Image Captions	50
4.3	Multimodal Input Sequence Formatting	51
4.3.1	Span-based Answer Extraction Models	51
4.3.2	Span-based Decoding Models	52
5	Evaluation	54
5.1	Training Procedure	54
5.1.1	Dataset Separation	54
5.1.2	Metrics Analysis for Span-based QA	54
5.1.3	Training Settings and Protocol	55
5.2	Sequence Context-Encoding Analysis	56
5.3	Studying Context Settings	59
5.4	Benchmarking Visual Context	60
5.5	Assessing Dataset Quality	62
6	Conclusions	63
6.1	Limitations and Future Work	64
	Bibliography	65
	Annexes	

I Annex 1	75
------------------	-----------

LIST OF FIGURES

1.1	Multimodal vs. Unimodal questions on the news domain.	2
2.1	Example of a The New York Times article about alternative energy in Africa.	7
2.2	Schematic representation of a recurrent neural network.	8
2.3	Transformer architecture overview.	9
2.4	BERT input representation.	11
2.5	Diagram of T5 text-to-text framework.	12
2.6	Schematic illustration of convolutional and pooling layers.	13
2.7	ViT architecture overview.	14
2.8	The LXMERT model architecture.	15
2.9	Illustration of VL-T5 and VL-BART architectures for visual grounding task.	16
2.10	Pre-training and fine-tuning procedures for BERT.	24
2.11	SpanBERT model architecture.	25
2.12	VisualBERT model architecture.	26
2.13	LXMERT pre-training tasks.	27
2.14	VQA ATM interface for collecting questions-answer pairs.	28
3.1	News Multimodal Question-Answering hybrid dataset creation pipeline. .	31
3.2	Example of an article gatherer query.	32
3.3	Our Mechanical Turk interface for annotating the generated candidates. . .	35
3.4	Interface the annotators interacted with to answer the candidate questions.	38
3.5	Selection of Human Validated subset samples.	39
3.6	Selection of Synthetic subset samples.	40
3.7	Our Mechanical Turk interface for the composition of question-answer pairs.	42
3.8	Two examples from the NewsMQA dataset.	45
3.9	Distribution of the top 5 question types.	45
4.1	Example of the visual cues used to enrich the input sequence of the models.	49
4.2	A comparative analysis of visual cues.	50
4.3	Span-based sequence encoding illustration.	52

4.4	Illustration of sequence encoding.	53
5.1	Training and evaluation reports.	57
5.2	Example of the T5 model's answers employing different context settings. .	61

LIST OF TABLES

2.1	A comparison of QA settings and their supporting datasets.	20
3.1	List of all keywords used to collect the articles for each topic.	32
3.2	Number of articles collected for each topic.	33
3.3	Progress of the number of articles and samples collected.	34
3.4	Number of responses for each classification in the crowdsourcing task. . .	37
3.5	Re-evaluation of the samples that have a relevant question but an invalid answer.	37
3.6	Distribution of news pieces, images and questions per topic.	44
3.7	Comparison between the subsets.	44
3.8	Number of samples whose answer is within one of the components used to generate the candidate pairs.	46
5.1	Number of unique articles and questions per topic in each subset.	55
5.2	Comparison between Exact match and F1 score results for a set of examples.	56
5.3	Impact of two sample pre-processing settings on BERT Model Performance.	58
5.4	Performance on zero-shot on the Synthetic test set, over different context set- tings.	59
5.5	Performance in the NewsMQA full dataset adding visual context.	60
5.6	Performance on the Human Validated, with different fine-tuning and input settings	62

INTRODUCTION

1.1 Context and Motivation

Over the past decades, the internet has become the primary and largest repository of information for people all around the world, transforming the way we access and consume information. This is most evident in the transition of newspapers and news outlets to the digital world, which has vastly increased the amount of articles we have access to. As a result, new search paradigms, such as question answering, have emerged towards helping users quickly find the information they need.

News articles cover a wide range of topics, addressing current events of general interest such as world championships or specific subjects such as technology releases. However, these articles are complex documents that refer to various entities, relationships, and concepts. Moreover, they are considered multimodal sources of information as they are composed of textual and visual elements. These visual elements play a crucial role as they complement news articles with an extra dimension that brings not only authenticity but also additional information [1, 2].

Search engines such as Google News bring all the articles and updates that are spread all over different platforms into one place, helping users traverse these large amounts of data by filtering the articles relevant to a topic and making it easier for users to find relevant content quickly and efficiently. Although these can be very useful for general searches, there is an upper bound to their specificity since they only allow for document retrieval. While reading an article about a certain event, the user may desire to quickly find an answer to a particular question about that given event or topic, rather than reading the whole document, and possibly searching other articles for an answer.

Systems should support conducting open-ended questioning within that same subject or document, providing users with relevant answers to their questions. As a collection of multimodal data, curious readers should also be able to question the articles regarding both their textual and visual elements. Utilizing both text and visual components to answer questions regarding news stories is a task that this thesis aims to address, providing a more comprehensive and interactive experience for users.

Headline: "Study Gauges Plastic Levels in Oceans"

Passage: "When the survey teams looked for plastics floating in the water that were the size of grains of sand, however, they were surprised to find far fewer samples than expected - one-hundredth as many particles [...] suggests that the smaller bits may be swept deeper into the sea or consumed by marine organisms."



Unimodal Question

What is consuming plastic particles?

Answer

marine organisms

Multimodal Question

Which particles are being swept deep into the sea?

Answer

plastic particles

Figure 1.1: Illustration of an instance of our problem. Multimodal vs. Unimodal questions on the news domain. In a multimodal question, the answer is grounded across both visual and textual news elements. Source [3].

1.2 Problem Formulation and Objective

The goal of this thesis is the formulation and benchmark of the novel task of Multimodal Question Answering for News Stories in which, given a question about a news article passage and image, the intent is to extract an excerpt from a passage that answers the question.

Our task emerges as a combination of a) open-ended text-based QA [4, 5], in which answers to a question are extracted from a context passage, and b) Visual question-answering [6–8] (VQA), which seeks to answer questions about an image in a close-ended setting (*i.e.* usually formulated in a categorical setting).

Multimodal question-answering is a wide field of research where a multitude of problems can be defined, therefore, narrowing down the settings of our problem is essential for this project. We assume the system receives a passage from an article that provides the answer to the question. The questions can be **unimodal**, meaning that one can find the answer just by interpreting a single modality (image or text passage), or **multimodal**, requiring a joint interpretation of the two mediums. The output is a span (part of the context), for example, a word, expression or sentence. Figure 1.1 illustrates our problem's framing and the available elements. In this example, we can see the two sorts of questions: a unimodal question that can be answered by skipping the image and a multimodal one that requires a joint interpretation of both the image and the context passage.

1.3 Challenges and Research Hypothesis

In this thesis, we assemble and benchmark NewsMQA (News Multimodal Question Answering), a dataset that reflects the difficulty of the proposed task.

For the assembly of the dataset, we draw inspiration from influential works in the area, including SQuAD [4], VQA [6], and OK-VQA [9]. These works have demonstrated the effectiveness of using crowdsourcing platforms to assemble large datasets of visual questions and corresponding answers. We also take inspiration from the work of [10–12], which employs a different approach to synthesizing visual question-answering datasets using the images’ captions. We aim for scale and quality, with the latter being assured through a significant portion of human-curated samples. To achieve this, we adapt the techniques mentioned above and approaches for hybrid dataset construction to our task, by applying a two-part approach: 1) multimodal synthetic question and answer generation component and 2) two-step crowdsourced question-answer annotation.

Regarding the benchmark of this dataset, a model capable of answering multimodal questions in the context of news articles brings the following challenges:

- The textual elements (question and passage) are crucial in the process of reasoning and finding the correct answer. Therefore, the model must have natural language processing capabilities.
- The model must also be able to identify and interpret the various elements of the image and create a detailed representation.
- The system needs to be able to select part of the given passage. Moreover, it is also necessary to define the criteria for which we consider a span valid. For example, if we have a picture of a banana and the question "what fruit is this?", we could answer "banana" or "this fruit is a banana"; as these are both valid answers, we should consider both answers correct.

We propose to address these challenges by leveraging deep learning methods that analyze textual and visual contents [13–16]. Recently, Transformer-based architectures [17] have demonstrated to be highly effective at modelling image and text semantics, both individually [13, 16] and combined [15, 18]. These have also proved to be able to solve various types of problems such as classification, regression and span prediction. [14]

Our work builds upon the rich representations provided by pre-trained Transformer models. We aim to improve them to respond to complex questions that require multimodal interpretation of information. Specifically, we propose to extend existing Span-based Transformer models to support multimodality by incorporating visual information extracted from images. To achieve this, we will need to reformulate the information provided to the models: instead of relying solely on the passage’s textual input, our models will need to take into account both the passage and visual information. Our approach

is based on the hypothesis that by incorporating visual information into existing Transformer models, we can enable them to more accurately and efficiently answer complex questions that require a multimodal interpretation of information.

1.4 Contributions and Achievements

As a result of the work performed in this thesis, we highlight the following contributions:

- **Extensive study of Question Answering Datasets and proposed models:** We review the existing tasks and respective datasets within the question-answering research area as well as explore the current models employed to solve these problems. (Chapter 2)
- **Propose a novel task and dataset:** We propose the novel dataset of News Multimodal Question Answering (NewsMQA) and we exploit its statistics and challenges. (Chapter 3)
- **Novel formatting of Transformer input for Multimodal Question Answering:** We propose to reformulate the information provided to the models by incorporating visual information into existing Transformer models (Chapter 4)
- **Benchmark and comprehensive study over our dataset:** We benchmark the NewsMQA dataset and conduct a set of studies and analyses over the dataset to evaluate the challenges it brings to the research area. (Chapter 5)
- **Submission under review for ACL 2023:** *"NewsMQA - An Open-ended Multimodal Question-Answering Benchmark for the News Domain"*, with authors: Carolina Lopes and David Semedo (Available in Annex I).

1.5 Document Organization

This document is organized into the following chapters:

Chapter 2 - Background and Related Work: Provides an overview and discussion about relevant existing models for language and vision representation and an analysis of techniques used for dataset construction.

Chapter 3 - News Multimodal QA Dataset: Describes the process behind our dataset assembly and an analysis of its key features and statistics.

Chapter 4 - Transformer-based models for News Multimodal QA: Introduces the backbone models used for benchmarking and study of our dataset. Moreover, it explores the incorporation of visual context into the backbone models.

Chapter 5 - Evaluation: Presents our training procedure, and a set of analyses and studies to evaluate and discuss the intricacies and challenges of our dataset.

Chapter 6 - Conclusions: Summarizes the conclusions of this thesis, reflects the limitations of our dataset and presents approaches to solve the task, and raises possible future works.

BACKGROUND AND RELATED WORK

In this chapter, we discuss previous research works to understand which already established methodologies can be helpful when approaching problems like the one raised in the context of our work. Firstly, we introduce the concept of news stories and their components. In the following section, we present state-of-the-art techniques and models for representing text, images and how to combine both. Next, we present relevant works within the news stories domain that approach multimodal tasks. In the fourth section, we introduce the problem of question answering, review documented problem variants involving a single (text-only) or multiple modalities (text and image) covering distinct domains (e.g. news, Wikipedia, and others), and discuss state-of-the-art approaches designed to solve them. Finally, we examine methods and protocols used in constructing question-answering datasets, towards deriving a sound methodology to assemble a novel dataset, which will be used to tackle the variant proposed in this thesis.

2.1 News Stories Understanding

A news article is a collection of informative text and illustrations about a current topic or specific event, usually conveyed by multiple media platforms: newspapers, magazines, television, radio, internet, among others. These are deemed one of the main sources of information, covering all aspects of our society and the world and resulting in highly diverse documents spanning a multitude of domains.

In Figure 2.1 is an example of a news article from a The New York Times newspaper. Regarding the structure and written information, an article generally consists of a headline, a lead paragraph and finally, the body of the news itself. Usually, the lead paragraph reveals the fundamental elements, always trying to answer basic questions, such as: what, when, where, how, why and who. The body follows and comprises the speech itself, delving deeper into what was introduced by the lead and providing new details about the story. In this section, the journalist usually arranges the storyline and presents the various topics in descending order of importance. The information presented by the lead is analysed more in-depth, and new details are provided, often referencing past articles

The New York Times

The World Needs to Quit Oil and Gas. Africa Has an Idea: Rich Countries First.

As negotiators at the Glasgow climate talks try to agree on greenhouse gas cuts, African leaders say poorer countries can't be expected to remake their systems as quickly as wealthy ones.

 Give this article    336



A new solar panel in Lagos, Nigeria, last year. "It's the bigger emitters that should have the responsibility to cut," one analyst said of fossil fuels. "We should be sensitive to history." Akintunde Akinleye/EPA, via Shutterstock

Figure 2.1: Example of a The New York Times news article about alternative energy in Africa ¹. It starts with a title, an entry image and the lead paragraph.

that address the same topic, similar topics, or that stand relevant to the current one.

Illustrative figures play a crucial role, both visually and in the storytelling [1, 2]. Images capture the readers' attention and invite them to notice the article. News with visual elements like a photo or graphic are much more likely to be read than one with just blocks of text. If there is a newsgroup on the same topic, the story that most people will gravitate to will be the one with an image [19]. An ideal illustration of a news story is an image that accompanies the story told in words and adds information in ways that words cannot. Images are capable of showing the relationships and emotions of the story constituents. Sometimes, with the help of a good caption, the images themselves can even tell the story by themselves.

We aspire to create a system capable of answering questions that force the understanding and correlation of both text and image of a news article and, as far as we know, such a system has not been developed and made available.

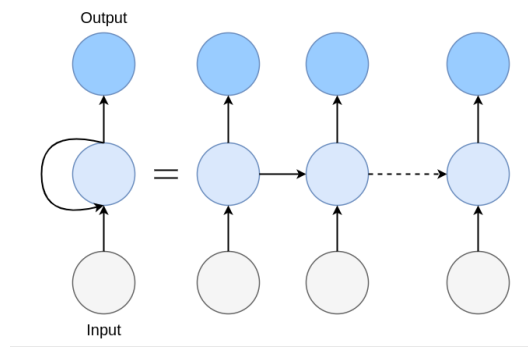


Figure 2.2: Schematic representation of a recurrent neural network (adapted from [20]).

2.2 Language and Vision Representations

In the context of our work, we tackle a task that requires processing both natural language and vision in order to find an answer. For this reason, in this section, we discuss state-of-the-art pre-training models for text, image and multimodal processing. In its empirical form, embedding is a mapping of a discrete variable to a low dimensional numeric vector. These representations provide the ability to correlate multiple inputs by calculating their distance in a vector space. A simple practical example would be calculating the similarity between two text documents or images.

2.2.1 Contextualized Text Representations

Natural Language Processing is a relevant research area as it has numerous uses in the computational representation and analysis of human language, for example, translation, retrieval, summarizing and question answering.

Depending on where and how a word is used in a sentence, it can have distinct meanings. Contextual embeddings take the context used into account and assign each word a representation based on it, capturing the uses of words in varying situations.

2.2.1.1 Recurrent Neural Network-based approaches

Recurrent neural networks (RNNs) are a type of deep neural network specially designed to deal with sequential data such as text sequences and have drawn attention due to their flexibility in capturing nonlinear relationships [20–22].

An RNN is very similar to a typical feed-forward neural network, except that the current layer uses past inputs information. At each step, the neurons receive an input vector and the output vector of the previous step. However, when a sequence is too long, RNNs have a short-term memory caused by the vanishing gradient problem, which results in losing information and difficulties capturing long-term dependencies. Architectures like Long Short-Term Memory Units (LSTM) and Gated Recurrent Units (GRU) are used to overcome this problem, as they are capable of learning which data is most relevant.

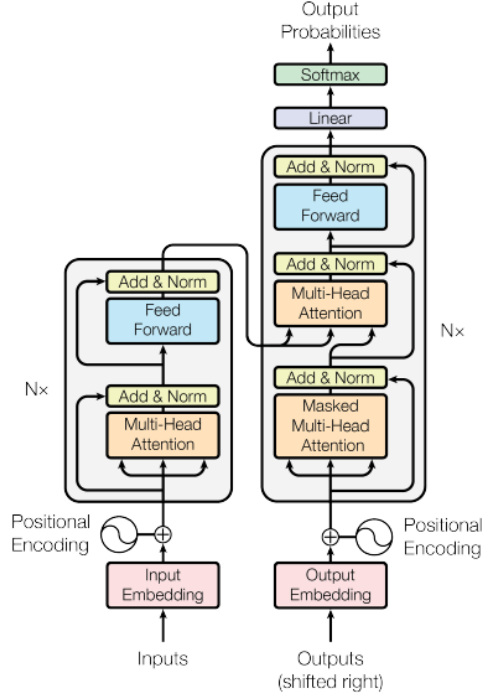


Figure 2.3: Transformer architecture overview. [17]

Peters et al. [23] presented ELMo, a strategy for high-quality learning of deep context-dependent representations based on RNNs. The ELMo word representations are functions that take into account the entire input sentence, so it is possible to model the complex characteristics of word usage and understand how these usages vary across linguistic contexts. Their results showed that ELMo representations can be easily added to existing models and used to solve six diverse and challenging NLP tasks: question answering, textual entailment, semantic role labelling, coreference resolution, named entity extraction and sentiment analysis.

2.2.1.2 Transformers-based approaches

Recurrent models have an inherently sequential nature that causes computing longer sequences challenging, as they cannot retain information about the first elements. In order to deal with this limitation, the concept of attention mechanism emerged.

An attention function extracts information from the whole sequence, returning a weighted sum of all the values. This allows assigning greater weight or importance to certain input elements, by learning which sequence elements the model should focus to predict the next output element. Unlike RNNs that struggle to retain long-term dependencies due to their implicit modelling of input sequences, Transformers explicitly reason over all the states, thus being able to access information about all the elements of the sequence. This leads to better-contextualized representations for each individual input element.

Vaswani et al. [17] proposed a new architecture, the Transformer, based exclusively on attention mechanisms to draw the dependencies within the input and output sequences, dispensing the recurrence usage. The Transformer architecture, illustrated in Figure 2.3, has an encoder-decoder structure where the encoder (on the left) maps an input sequence of symbol representations to a sequence of continuous representations, and the decoder (on the right) gathers these representations and generates an output sequence one element at a time.

At the bottom of each component stack, it is performed an initial word embedding layer, where each word is mapped to a continuous vector, and a positional encoding so the model is aware of the relative and absolute position of the tokens in the sequence. Both encoder and decoder contain a core block subdivided into a multi-head self-attention mechanism followed by a fully connected feed-forward network repeated N times (in their case, $N=6$). Around each sub-layer, there is a residual connection followed by a normalization layer, which help the training process by stabilizing the network. The decoder incorporates the encoder stack and inserts a third sub-layer that performs multi-head attention over its output.

The number of encoders and decoders used in sequence can lead to an effectiveness improvement. Namely, stacking the encoders and decoders allows learning of different and more high-level representations and can improve how the model can learn to extract and focus on different combinations of attention from its attention heads.

2.2.1.3 BERT - Deep Bidirectional Transformers

Until now, the discussed models are unidirectional, meaning they can only read input sequentially, either from left-to-right or right-to-left. This constraint specifically affects token-level tasks, such as question answering, where it is crucial to incorporate and comprehend the context from both directions. Devlin et al. [13] introduce BERT (Bidirectional Encoder Representations from Transformers), a model based on the Vaswani et al. [17] implementation and designed to pre-train deep bidirectional representations from unlabeled text, jointly considering both left and right directions across all layers. It is one of the most well-known models for language representation and presents impressive results on various natural language processing tasks.

The input representation can unambiguously represent a single sentence or a pair of sentences (e.g. a question and an answer). Figure 2.4 illustrates the construction of the input representation. At the beginning of every input sequence is always a special classification token ([CLS]) and at the end of each sentence is a token [SEP]. To differentiate the two sentences, it was added segment and position embeddings to each token, indicating which sentence it belongs to. The final input representation is formed by adding the corresponding token, segment, and position embeddings.

BERT training process has a pre-training and a fine-tuning phase. The model is pre-trained on two NLP tasks: a) Masked Language Modeling and b) Next Sentence

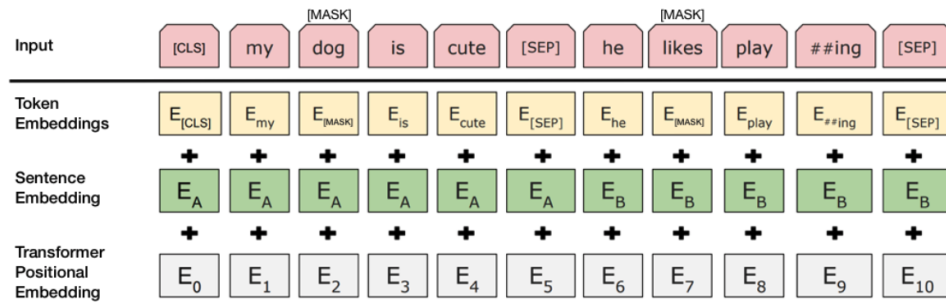


Figure 2.4: BERT input representation. [13]

Prediction. For Masked Language Modeling a), 15% of the input tokens are masked at random, and the model predicts those masked tokens using all the other tokens of the sequence. To mask these tokens, they used a [MASK] token 80% of the time, chose another random token 10% of the time and kept the original token 10% of the time. For the next sentence prediction task b), the model predicts whether sentence B is supposed to come after sentence A. When choosing the A and B sentences, 50% of the time B is the actual sentence that follows A, and 50% of the time it is a random sentence. This task is crucial for tasks such as question answering since it allows the model to understand the relationships between the two sentences.

Thanks to the pre-training process, there is no need for heavily-engineered task-specific architectures, and the model can be fine-tuned by appending task-specific layers, in a classification or regression setting.

2.2.1.4 BERT Variants

Several methods have been proposed to improve BERT. Among these, RoBERTa [24] and XLNet [25] are two models that improved BERT and achieved better performance. They were able to do this by changing the training methodology, and utilizing more training data and computer power. RoBERTa [24] (Robustly optimized BERT approach), to improve the training procedure, eliminate the Next Sentence Prediction (NSP) pre-training task and introduce dynamic masking so the masked token changes during training epochs. XLNet [25] introduces a pre-training method that, in contrast to BERT's masked language model where only the masked (15%) tokens are predicted, enables learning bidirectional contexts by predicting all tokens but in random order.

2.2.1.5 T5 - Transfer Learning with a Unified Text-to-Text Transformers

Raffel et al. [16] introduce T5 (Text-to-Text Transfer Transformer), an encoder-decoder model pre-trained on a mixture of tasks and for which each task is converted into a text-to-text format. Contrary to BERT, T5 is an encoder-decoder text-to-text model which means the model is fed some text for context or conditioning and is then asked to produce some output text. As 2.5 exemplifies, we can employ the model for a variety of tasks by

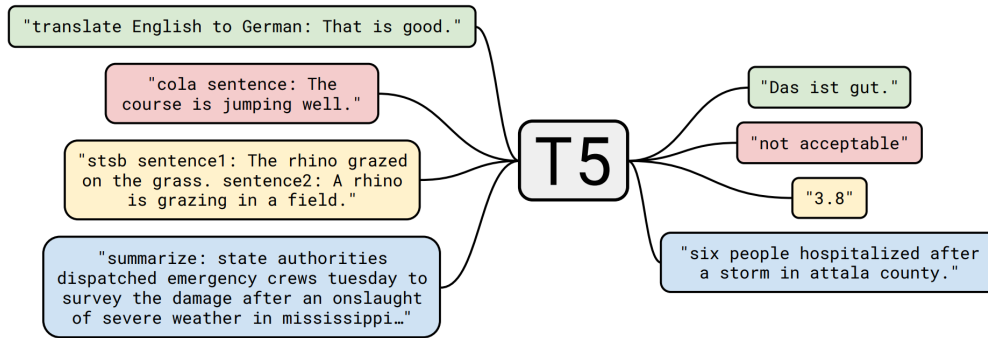


Figure 2.5: Diagram of T5 text-to-text framework. It exemplifies that for any task a text prompt is given as input and the model is trained to generate some target text. [16]

prepending a different prefix to the input corresponding to each task *e.g.*, for translation: *translateEnglishToGerman : text* or for summarization: *summarize : text*. This type of framework allows the usage of the same model, loss function, hyperparameters, etc. across our diverse set of tasks. T5 reached state-of-the-art results in more than 20 natural language processing downstream tasks, including tasks such as machine translation, question answering, summarization, and text classification.

In their study to achieve such a model, Raffel et al. [16] compare pre-training objectives, architectures, unlabeled data sets, transfer approaches, and other factors on dozens of language understanding tasks. Moreover, they also introduce the e “Colossal Clean Crawled Corpus” (C4), the massive dataset T5 was pre-trained on, and which consists of hundreds of gigabytes of clean English text scraped from the web.

2.2.2 Vision Representations

Computer Vision is another relevant research topic that enables systems to interpret meaningful information from visual data that, in recent years, has shown growth in tasks such as object detection, classification and segmentation.

2.2.2.1 Convolutional Neural Networks-based approaches

Convolutional neural networks (CNNs) have shown excellent results in object identification, face recognition and when dealing with time series [21, 22, 26, 27]. Its distinguishing feature is the ability to detect patterns by applying convolutional and pooling layers (exemplified in Figure 2.6).

A convolutional layer aims at recognizing patterns that exist in the data. It applies a finite number of filters (kernels) to the input, returning a feature map with the patterns detected by each filter as output. The values of each filter are adjusted during the training process so that the network can learn which patterns it needs to identify. Pooling layers

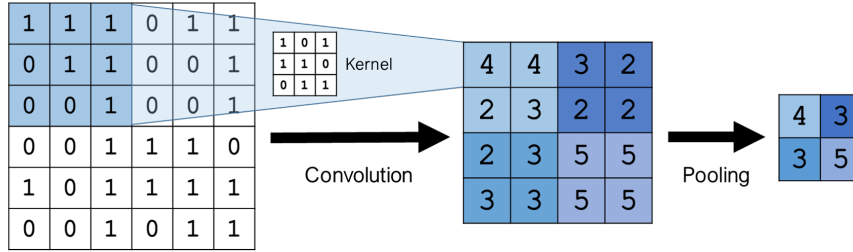


Figure 2.6: Schematic illustration of convolutional and pooling layers. (adapted from [26])

allow reducing the dimensions of the data by joining a portion of the neurons of the previous layer into a single neuron of the next layer, using, for example, max pooling where the highest value is selected. This way, the data is compressed, simplifying the result without losing the most relevant information. Combining these two kinds of layers leads to a gradual reduction in the amount of information to be processed across the network and requires fewer parameters and computations than a fully connected network.

Among the most known CNNs are the Fully Convolutional Networks (FCN), Residual Networks (ResNet) [27] and Visual Geometry Group (VGG) [28], which are usually used for object recognition and image classification.

2.2.2.2 Vision Transformer-based approach

Despite the remarkable results, CNNs present some limitations caused by the pooling layers. Due to these layers, too much potentially valuable information is lost, and relationships between the parts of the image and the notion of the image as a whole are ignored. Inspired by the Vaswani et al. [17] Transformer successes in NLP, Dosovitskiy et al. [29] proposed ViT, a model that applies the standard Transformer to Computer Vision, applying stacks of self-attention layers instead of typical CNNs or RNNs.

In a naive application of self-attention to images, each pixel would consider every other pixel. However, such implementation would require an exponential calculation of combinations as the image size increases. Moreover, not every possible pixel pair has a meaningful relationship to interpret or reason about (for example, two opposing corner pixels). To overcome this problem, the image is split into fixed-size patches, which are then handled by the model in the same way as tokens (words) in an NLP application. As we can see in Figure 2.7, each patch is linearly embedded, its position embedding is added, and the resulting sequence is fed to a standard Transformer encoder. The model was trained on image classification, so an extra learnable classification token ([class]) was added to the sequence.

During the experiments, it was noticed that ViT integrates information throughout the image, even in the lowest layers of Transformers and is more efficient with smaller patches. In addition, the model revealed impressive results in many image classification datasets, matching or exceeding the state-of-the-art models, while being relatively inexpensive to pre-train.

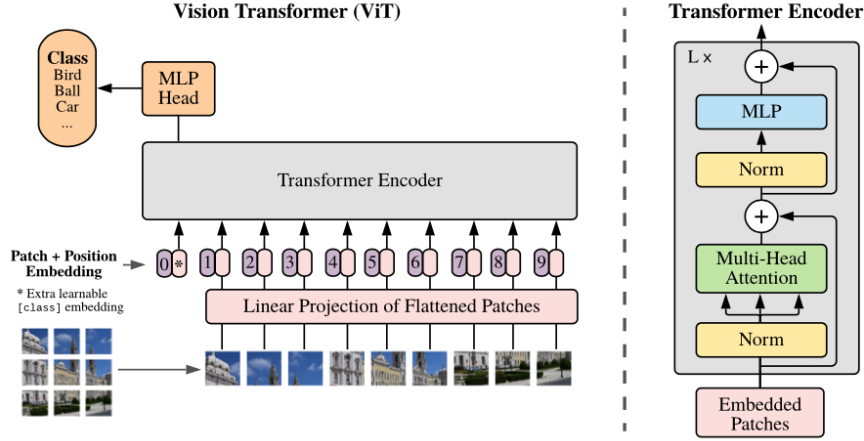


Figure 2.7: ViT architecture overview. [29]

2.2.3 Multimodal Representations

Many models have been designed to retrieve and classify a specific type of data, for example, images or text. However, in a real-world scenario, data comes in different modalities, carrying diverse information. We propose a system capable of identifying and interpreting this distinct but complementary information. Particularly, we aim to link visual and linguistic content and correlate them both.

Transformers have already proven to be able to interpret and reason about text and images separately. Lately, these have been extended to work in a multimodal setting [15, 18, 30], where the core idea is to make self-attention multimodal, enabling the model to attend over different modalities.

One of the first transformer-based pre-trained models designed to tackle multimodal problems with vision and language was LXMERT [18]. This model, represented in Figure 2.8, takes two inputs (an image and its related sentence as a question or caption) and consists of three encoders: a language, an object-relationship and a cross-modality encoder. In a first phase the two inputs are converted into sequences of features. The input text is split into words by the WordPiece tokenizer [31], and its final embedding is given by the word's token and its position in the sentence. Regarding the image, an object detector is used to extract the most relevant features. The final embedding is given by the sum of the feature itself, its position in the image and its RoI (Region of Interest).

The model itself starts with two transformer encoders: a language and an object-relationship encoder. These are based on the Transformer architecture Vaswani et al. [17] and consist of self-attention and feed-forward sub-layers. Then, in order to learn correlations within text and images, the outputs of these encoders are directed to a cross-modality encoder which, in addition to the individual self-attention sub-layers to both modalities, has a cross-attention sub-layer. This last layer helps the model to exchange information between the two modalities so that it learns cross-modality representations.

The model was pre-trained with large amounts of image-and-sentence pairs via five

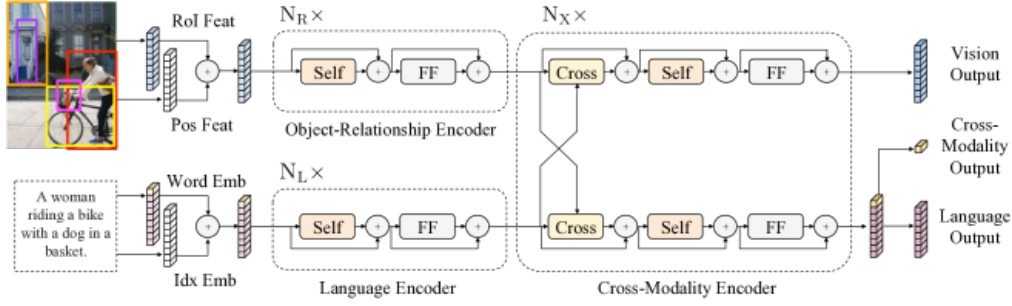


Figure 2.8: The LXMERT model architecture. Adapted from [18].

diverse representative pre-training tasks: one language task, two visions tasks and two cross-modality tasks. When it comes to language, they use a Masked Cross-modality Language Model task, where the model is asked to predict the masked words. These words were selected randomly with a probability of 0.15. For the visual tasks, they masked random objects, masking their RoI feature, with a probability of 0.15. Then they asked the model to predict two properties of these masked objects: their RoI feature value (RoI-feature Regression) and to identify their labels (Detected-label Classification). These two tasks were crucial for the model to learn object relationships and infer decisions using the language information. Finally, they pre-trained the model with two cross-modality tasks: a) Cross-Modality Matching and b) Image Question Answering. In a), they ask the model to predict whether an image and sentence match each other. For this, each sentence was replaced with a probability of 0.5. In b), the model is asked to predict the answer to image-related questions. These tasks are crucial for the model to learn both intra-modality and cross-modality relationships.

There are other alternatives to represent multimodal embeddings using Transformers. ViLBERT [30] and VisualBERT [15] are other two models that are similar to LXMERT. All these models take as input a multimodal sequence, in which image regions are provided as tokens. Moreover, they are all pre-trained in masked cross-modality LM and cross-modality matching.

The difference between these models is in the architecture. While VisualBERT has a single-stream architecture, ViLBERT and LXMERT have a dual-stream architecture. [32] In VisualBERT, a standard BERT architecture receives the concatenation of the visual and linguistic features direct from the image-text input pair. This design allows for an early and unconstrained convergence of cross-modal information.

On the other hand, in ViLBERT and LXMERT, the linguistic input features is first processed by an independent stack of Transformer layers. In the first individual processing of the input features, LXMERT also applies an object relationship encoder. Then, the resulting representations are fed into a cross-modality Transformer where intra and inter-modal interactions are alternated. Furthermore, LXMERT employs additional tasks during pre-training, such as RoI-feature regression and image question answering, which makes a considerable difference in the final performance of the models.

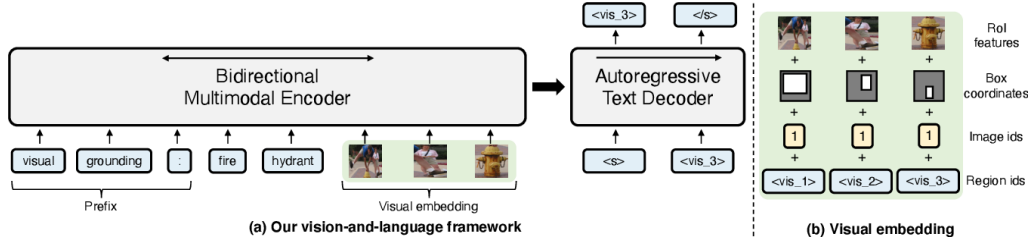


Figure 2.9: Illustration of VL-T5 and VL-BART architectures for visual grounding task. (a) refers to visual embeddings. (b) shows the components of visual embedding. [33]

The methods we have analyzed so far for learning representations using Transformers typically require designing task-specific architectures for each downstream task. However, most of the time, the reasoning skills required by these tasks overlap significantly. Hence, to alleviate these hassles of designing task-specific architectures, Cho et al. [33] propose a unified framework for vision-and-language learning that learns different tasks in a single architecture via text labels generation.

They introduced the VL-T5 and VL-BART, two models based on two off-the-shelf pre-trained transformer language models: T5 [16] and BART [34]. The original models were modified so that their text encoders could process multimodal data. Such was possible by incorporating image region embeddings as additional input. Figure 2.9 shows the overall framework’s architecture.

For the visual embedding, as presented in Figure 2.9 (b), each image region is encoded as a sum of four types of features: RoI object features, RoI bounding box coordinates, image ids and region ids. The image ids are used to distinct regions from different images, and the regions ids are used to distinct the regions within the same image. Regarding text embedding, both models follow the positional embedding configurations of their text backbone models: BART learns positional embeddings for absolute token positions and adds them to the token embeddings, and T5 adds relative position bias to each self-attention layer.

In contrast to existing methods that train different architectures for each downstream task, these models tackle all tasks within the same language modelling head. For learning a new task, instead of adding extra parameters or designing a new architecture, it is only necessary to add different prefix to the original input text. This augmented input text is then tokenized and encoded as learned embedding.

The models then follow the Transformer encoder-and-decoder architecture [17] to encode the visual and text inputs, and generate label text. As illustrated in Figure 2.9 (a), the encoder receives, as input, the text and visual embeddings and outputs the contextualized joint representations. It consists of a stack of transformer blocks composed of a self-attention layer and a fully-connected layer with residual connections. The auto-regressive decoder iteratively attends to the previously generated tokens (via self-attention) and to the encoder outputs (via cross-attention) and predicts the probability of future text tokens. It is also composed of a stack of transformer blocks similar to the encoder but each

block has an additional cross-attention layer.

The models were pre-trained in a diverse set of tasks: visual question answering, multimodal language modelling, image-text matching, visual grounding and grounded captioning. During the visual question answer task, models generate answers instead of ranking predefined candidates. In the Multimodal language modelling task, part of the input is masked, and the models need to predict the masked text. For VL-T5, 15% of input text tokens were masked and replaced with sentinel tokens (e.g., <text_1>). For VL-BART, 30% of input text tokens were masked with <mask> tokens. The grounding tasks were crucial to teach the model object-text matching and object-level information. Given a region description in visual grounding, the model predicts the id of the related object region. In grounded captioning, given a visual sentinel token (which indicates an image region) as text input, the model needs to generate a corresponding textual description of the image region.

This generative approach revealed remarkable generalization ability on questions that have answers that rarely occur during training.

2.3 News-related Multimodal Tasks

Over the years, more and more research work has been developed within the news domain, and some of these address similar challenges to the ones we encounter in our proposal. Having covered vision and language representation learning methods, we now discuss relevant news tasks that require dealing with the news articles multimodality.

2.3.1 News Image Captioning

Image captioning is a challenging multimodal task. Given an image, the objective is to generate a cohesive narrative of the event that it conveys. Although it is not a question answering problem, this work indicates to be very pertinent since building such a system requires multimodal semantic understanding and linguistic expressiveness.

Tran, Mathews, and Xie [35] proposed a model that combines neural networks, vision subsystems and language representation learning to generate captions for articles' images. Their model consists of a set of pre-trained encoders and a decoder. The encoders generate high-level vector representations of the images, faces, objects, and article text. By doing this, they capture the existing relationships between the objects present in the image. The general image representation is obtained from a ResNet-152 that analyses the different regions of the picture. For the faces and object detection, they use two models specifically designed for these tasks, the MTCNN [36] and YOLOv3 [37], respectively. Finally, to encode the news text, they use RoBERTa [24], a model design for pre-training natural language processing. Then, the decoder correlates all these representations and generates a natural language caption.

Most subtitling systems rely on a fixed vocabulary and can not generate new or rare words. Thanks to RoBERTa, which uses a byte-pair-encoding, the system is capable of generating unseen words, which addresses the linguistic gap of a fixed vocabulary. In the end, the model generates a caption with named entities directly, without needing to use an intermediate model.

Besides the model, they introduced the NYTimes800k dataset, an image captioning dataset that contains 445k articles and 793k figures with captions from the New York Times archive. Within the news image captioning, it is also relevant to mention the GoodNews [38] and Visual News [39] datasets, the other two benchmarks in the area.

The Visual News dataset is the largest one up to date and includes more than one million news images paired with both captions and news article text collected from a diverse set of sources (The Guardian, BBC, USA Today, and The Washington Post). These datasets will be a reference point in building our dataset as they feature both news text and images.

2.3.2 Visual Storylines

So far, we have explored systems that focus predominantly on the written content of the news, both to look for potential answers and to generate subtitles. However, the visual elements are an integral part of the storytelling process in journalism. Considering the amount of related information that continually arrives at the newsroom, the task of creating high-quality and cohesive media plots is becoming a more prominent challenge.

To assist media editors in this arduous task, in [40] is proposed a framework to organize content into cohesive, high-quality, relevant visual storylines. They analyzed various factors to select the best images, such as the photo's relevance to the news, when the photo was taken, its quality, and how the final visual aesthetic would look. They proposed and evaluated four graph-based approaches to storyline creation, pointing to their high performance at creating these.

The results showed that targeting maximum cohesion among all story illustrations was more likely to achieve a better final quality storyline and lead to a better sense of the pertinence of the news story illustration. Most relevant for our work, they demonstrate that users prefer transitions where two images share visual concepts and recognise the power of a good set of illustrations in understanding a news story.

2.4 Question Answering

Question Answering (QA) is a large area of computer science that aims to create a system capable of answering, in a cohesive way, to any question written in natural language. Such a system requires natural language processing (NLP), artificial intelligence and information retrieval capabilities, which are difficult to achieve.

The first and most crucial steps in tackling such a complex problem are understanding what kind of question is being asked and how to answer it, as it influences the entire system’s reasoning. However, there are numerous configurations to consider. The question itself can refer to a phrase, image, video, audio or an entire document. In addition, different systems have different purposes, for example, classification, retrieval or information extraction. For this reason, within the QA area, several strands address diverse types of problems with distinct settings.

This section begins with a categorization of the most pertinent existing question answering systems, analyzing the various formats of their configurations (question, answer, modality and domain). Subsequently, considering the context of our problem, we further analyze already-developed models of two of these approaches: text-based models and visual question-answer models.

2.4.1 Question-Answering Settings

We took into account the separation made by Rogers *et al.* [41], and in Table 2.1, we gather and compare various QA settings and their respective support datasets. In their work, they provide an overview of the various formats and domains of the resources: questions, answers, and input evidence (in terms of their modality).

2.4.1.1 Questions format

Four question formats were outlined: natural language questions, queries, cloze format, and story completion. Most QA datasets have the questions formulated as natural language questions, which means they are written using natural language as if a human would ask them. On the other hand, cloze-format resources consist of a sentence(s) with a masked span that the model needs to predict (fill-in-the-blank style). Some examples of settings with this format are the CBT [42], Booktest [43], and DMQA [44] datasets. Given this format, these datasets can be programmatically generated, resulting in a quick and cost-effective data collection procedure. Despite this advantage, cloze questions are not technically "questions" and, therefore, do not directly target the QA task. Another limitation is that it only targets relationships within a given context, and it is challenging to control the information needed to populate the Cloze.

2.4.1.2 Answer format

Regarding the answer formats, Rogers, Gardner, and Augenstein [41] organise them as extractive, multiple-choice, categorical or free-form. Multiple-choice questions are questions for which a small number of answer options are provided (*e.g.* MCtest [45], QuAIL [46], Visual Madlibs [47] and MovieQA [48] datasets). When the format is categorical, the answers come from a strictly pre-defined set of options, being the most relevant, the DAQUAR [7] and COCO-QA [8] datasets. Another style of categorical output format

Table 2.1: A comparison of QA settings and their supporting datasets. T, text; I: image, Tb: table, V: video; Q: Question, Img: Image.

Inputs	Answer Format	Modality	Domain	Available Datasets
Question + Context	Multiple-choice	T	Fiction	MCTest [45]
Question + Context	Multiple-choice	T	News	QuAIL [46], NewsQuizQA [49]
Question + Context	Extractive	T	Encyclopedia	SQuAD [4]
Question + Context	Extractive	T	News	NewsQA [50], NYT archive [51], NLQuAD [52], ArchivalQA [53]
Cloze + Context	Extractive	T	Fiction	CBT [42], BookTest [43]
Cloze + Context	Extractive	T	News	DMQA [44]
Question + Image	Multiple-choice	I+T	Images	Visual Madlibs [47]
Question + Image	Categorical	I+T	Images	DAQUAR [7], COCO-QA [8]
Question + Image	Extractive / Categorical	I+T	Images	VQA [6]
Q + Image + Knowledge	Extractive / Categorical	I+T	Multi-domain	Visual Genome [54], CLEVR [55], FVQA [56], KVQA [57], GQA [58], HybridQA [59]
Q + Context + Table	Extractive	T+Tb	Encyclopedia	MMQA [60]
Q + Img + Context + Table	Extractive	I+T+Tb	Encyclopedia	ManyModalQA [61]
Q + Img + Context + Table	Extractive	I/T/Tb	Encyclopedia	MovieQA [48]
Q + Plot + Video + Script	Multiple-choice	V+T	Movies	OK-VQA [9]
Q + Image + Context	Extractive	I+T	Multi-domain	[62]
Q + Image + Context	Extractive	I+T	Expert materials	MuMuQA [63]
Q + Img + Caption + Context	Extractive	I+T	News	
(NewsMQA) Question + Image + Context	Extractive	I+T	News	—

is when the set of responses allowed at the test time is the set of answers seen during training. In these circumstances, the final prediction is a classification problem, which simplifies the implementation but is quite limiting since no test question can have an unseen answer (*e.g.* VQA, GQA and CLEVR datasets). The extractive format is the one that interests us the most since it consists in predicting the part of the input that contains a valid answer to the question (*e.g.* a word, sentence, span, or document), that is, a span.

As we can corroborate from Table 2.1, several datasets encompass this type of configuration, however, let's look at its domains, modalities and compare these with the proposed problem settings. We can separate the datasets into two big groups: the single and the multi-modality datasets.

2.4.1.3 Single Modality

Within the single modality group, the most relevant are those listed and highlighted in the first section of the Table 2.1. These datasets helped the research community build

algorithms capable of answering questions requiring human-level comprehension and reasoning skills. They are focused on QA in the news domain and questions that demand reasoning abilities, however, they only reason over text.

2.4.1.4 Multiple Modality

Concerning problems that associate multiple modalities, there is a wide variety of datasets. Since the beginning of 2014, one of the most prevalent topics within the Question Answering area has been Visual Question Answering (VQA), a computer vision task in which a system receives a text-based question about an image and must infer the correct answer. Including this image significantly increases the complexity of the problem, as it is necessary to interpret not only the question but also the image and its elements.

The most relevant works to explore question answering on images are [7, 8, 47], where were introduced the first datasets for the novel problem. Agrawal et al. [6] introduced the VQA dataset and was the first work to propose Visual Question Answering as a problem on its own. It includes about 760k questions and 10 million answers for 255k images from both MS COCO dataset [64], and Amazon Mechanical Turk. These datasets' questions require an understanding of vision, language and commonsense knowledge to answer, however many of them can be accurately answered without using the image due to language biases [65].

Most VQA datasets are focused mainly on recognition, and most questions are about simple counting, colours and other visual detection tasks. However, our proposed problem incorporate, in addition to just visual recognition, also logical reasoning and association with knowledge of the external world.

Datasets as CLEVR [55], FVQA [56], KVQA [57] (Knowledge-Aware Visual Question Answering) and GQA [58] explore the task of world knowledge-aware VQA. Their questions require multi-entity and multi-relation reasoning over large Knowledge Graphs to arrive at an answer. Furthermore, and still, within the multiple modality group, several datasets associate text with images and also with video or tables (*e.g.* HybriQA, MMQA and MovieQA datasets).

2.4.1.5 Towards News Multimodal QA

Recapping, this thesis proposes a dataset on the news domain, that takes a question, an image and a context as input. The system must then provide an answer, in an extractive format *i.e.* the answer in a span of the given context.

One of the most similar VQA dataset setup to our proposal is the OK-VQA dataset [9] (released in 2019), which assembles questions that look beyond what is directly in the image. Some state-of-the-art VQA models were evaluated on this new dataset and showed their performance degrading drastically. These results revealed the need for new methods capable of retrieving relevant facts from a source of knowledge.

Another relevant dataset within this matter is the 2016 Multimodal Question Answering in the cultural heritage domain dataset proposed in [62]. This dataset holds multimodal content about art from the old-Egyptian Amarna period that includes 16 images of the artworks, documents regarding them and over 800 questions combining visual and textual details. The most prominent problem with this dataset is that it is a relatively small set containing a low amount of images and questions.

Although these datasets are multimodal, incorporating images and documents, none targets the news domain, which brings distinct challenges. As discussed in section 2.1, news articles are composed of textual and visual elements. Concerning the textual element, when working with global domains like news, there are a lot of entities and relationships to identify and correlate. Regarding the images, these are specifically chosen by the journalist to illustrate the news. Therefore, we know that they are inherently intertwined with the story, illustrating it and adding information that did not exist before.

Reddy et al. [63] take a step further in this direction with the MuMuQA dataset, by comprising question and answer pairs from the news domain. Similar to our NewsMQA, the dataset is constructed with a combination of synthetic and human-annotated samples. However, their training set is fully synthetic and the human-annotated set is very limited (1384 samples). Moreover, 1) their questions follow an entity-centred setting, which does not capture the full connection heterogeneity between news text and images, 2) the task requires the entirety of the text body, the image but also the image caption as input, which hides a portion of the task’s complexity and 3) their synthetic entity-focused training set generation scheme might implicitly be biased as this generation scheme and actual solution scheme use the same reason over the same components (image, caption and body text) for the visual entity grounding, visual attribute extraction and question decomposition, something we avoid when creating our synthetic set of questions.

This strong relationship between text and images constitutes a significant challenge because to understand and interpret news stories in its entirety, it is necessary to understand this strong relationship between the two elements. We aspire to create a system capable of understanding this strong correlation between the text and image elements and use it to answer questions regarding the articles.

2.4.2 Retrieving News Context Passages

One of the challenges underlying news question answering is finding the document or paragraph in which the answer can be found. Furthermore, the system needs to interact with language-based data and locate the correct answer.

Wang et al. [51] takes into thought the numerous amount of articles available to answer questions about specific topics and events. This work approaches the sub-task of QA that uses news article archives and presents a large scale question answering system called QANA (Question Answering in News Archives) designed for answering event-related questions on these. The proposed retriever-reader system is focused on choosing

the best documents where the correct answer may be and solve this task by exploring the temporal characteristics of both questions and documents.

The system itself consists of a Document Retriever Module that performs keyword extraction and retrieves candidate documents from the archive, a Time-Aware Reranking Module that exploits the timestamp for re-ranking candidate documents and a Document Reader Module that extracts the answers. The first module starts by performing a keyword extraction and then retrieves candidate documents from the underlying document file. In the Time-Aware Reranking Module, temporal information is explored considering various perspectives to rerank the retrieved candidate documents. In the last module, they use the BiDAF model [66] to extract answers of the top N reranked documents and select the most common answer as the final answer. Thanks to these modules, the system is capable of outperforming the existing QA systems. They used a subset of the New York Times archive as the underlying knowledge source and a carefully constructed test set of questions associated with past events.

Within the retrieval systems, Qu *et al.* [67] also presented a dense retrieval system, but this is concentrated on outside-knowledge visual question answering. Given a question-picture pair, the task would be to retrieve, within an extensive collection, the k passages that most appear to have the correct answer. This means that in addition to text, this model also analyzes and considers visual content.

The system holds a dual-encoder architecture and is formed of a query encoder and a passage encoder. Since the query encoder needs to consider both query and image, they use the LXMERT model [18], a model designed to learn cross-modality connections. For the passage encoder, they apply the BERT model [13] as it is one of the best-known models for learning text representation.

With this work, they verified that visual elements are helpful to answer the OK-VQA question and that the image captions can even be more informative than object detection.

In addition to these works, several others address the task of QA for news [49, 50, 52, 53], however, none of these considers questions related to the illustrative figures of the articles and only search for information in the available text to come up with an answer. We will analyze this problem and the existing datasets in more depth in the Section 2.4.

2.4.3 Text-based Question Answering Models

Typically, in a question-answering system, the question and the context from which information is supposed to be extracted to answer are both textual elements. A system with these characteristics needs to have the capacity to process natural language and real-world knowledge/commonsense.

As discussed in section 2.2.1.3, BERT [13] is one of the most well-known models for natural language processing and presents remarkable results in specific tasks such as Question Answering. Thanks to its pre-training process, the model can be easily adjusted and trained on a new task simply by connecting task-specific inputs and outputs.

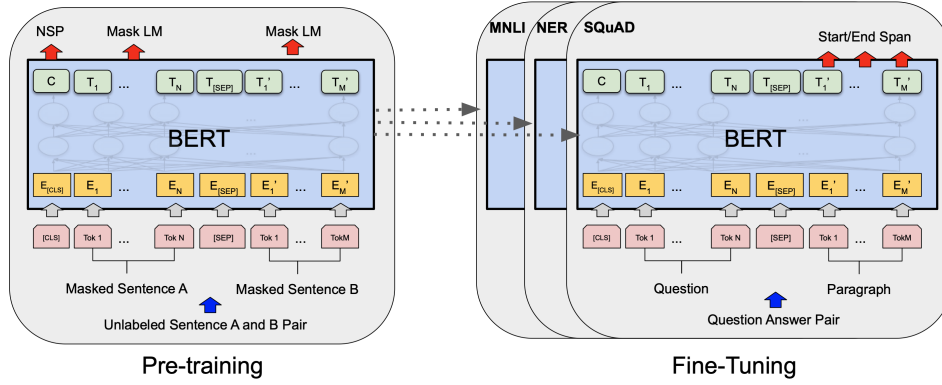


Figure 2.10: Pre-training and fine-tuning procedures for BERT [13].

As represented in Figure 2.10, each downstream task has a separate fine-tuned model. In this fine-tuning process, the model is first initialized with the pre-trained parameters, and all of the parameters are then fine-tuned end-to-end using labelled data from the downstream task. For example, at the input, sentences A and B from pre-training might be sentence pairs in paraphrasing or degenerate text- $\emptyset$ pairs in text classification or sequence tagging.

In the QA scenario, the input sentences are a question and a passage and the output will be a span of the passage, specifically the model will identify the start and end tokens of that span. During the fine-tuning, the model returns, for each token, the probability of it being the answer's start and/or end token. We treat questions that do not have an answer as having an answer span with start and end at the [CLS] token.

2.4.3.1 Spans Prediction from text

As discussed in section 2.4.1.5, interpreting news documents is not a simple problem. The domain is so vast that it is not feasible to train a classifier in a fixed set of answers.

An approach to exploit the text is using a span-based model since these are able to predict entire spans exclusively by interpreting the given context. Moreover, such a system provides a greater variety of answers as well as adequate and well-formed answers in terms of linguistics and vocabulary, as it is not necessary to generate any words or phrases.

Joshi et al. [14] proposed SpanBERT, a pre-training method that extends the usage of BERT to represent and predict text spans. As we can observe from Figure 2.11, SpanBERT's architecture is identical to Bert's; the discrepancy between these two is in the way the models are pre-trained. The first prominent difference is in the training samples, where SpanBERT samples a single contiguous text segment for each example instead of two as in BERT. Consequently, the model is not pre-trained with the Next Sentence Prediction task. This decision brings some benefits since the model is trained with longer full-length contexts. Secondly, while in BERT they mask individual tokens for the masked language modelling task, in SpanBERT they mask continuous spans of text, replacing all those

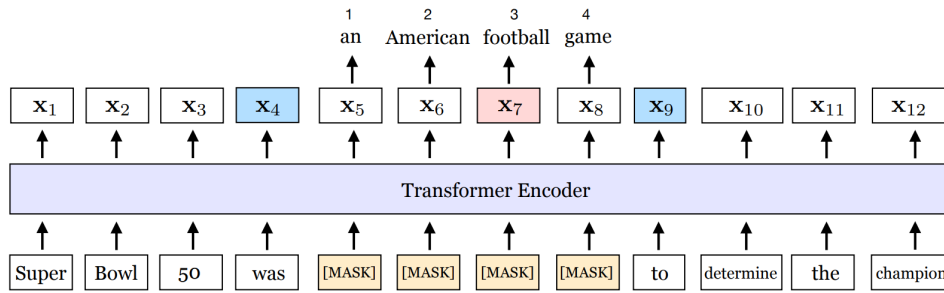


Figure 2.11: SpanBERT model architecture (adapted from [14]).

tokens with a [MASK] or sampled tokens. Finally, they also introduce a novel span-boundary objective (SBO), which helps the model learn how to predict each token in the masked span exclusively using the representations of the observed unmasked tokens.

After the pre-training process, SpanBERT was evaluated in several tasks, including extractive question answering, where similarly to BERT, given a question and a short passage, the model is asked to select a contiguous span of text as an answer. To select this span the model gives each token of the passage a probability of being the answer's start and/or end token. The model showed intriguing results, outperforming all BERT baselines and reaching a better performance on span selection tasks in particular.

2.4.4 Visual Question Answering Models

Visual Question Answering (VQA) is a problem of QA where a system takes an image and a language-based question about it and is asked to find an accurate natural language answer based on that image information. For a system to be capable of solving such a complex and specific task, it must have a series of capabilities, for example, natural language process and detection of objects and entities. Hence, these models are often first pre-trained on tasks such as masked language modelling and image-text matching, and then fine-tuned on particular ones like Visual Question Answering.

Among the systems with the best performance in solving the VQA problem are transformer-based models, particularly, the VisualBERT [15] and LXMERT [18]. Figure 2.12 illustrates the architecture of VisualBERT, which combines the image regions and language with a Transformer. This model was pre-trained with a masked language modelling (identified as Objective 1 in the Figure 2.12) and sentence-image prediction task (identified as Objective 2). Regarding the fine-tuning process, VisualBERT considers the VQA task a classification problem, where the model needs to choose an answer from a limited answer collection. To better fit the pre-training process (where random words from the textual sequence were masked), a [MASK] token is appended after the question, and the representation of the [MASK] token is then fed into an output layer for classification.

It is interesting to note that the SpanBERT and VisualBERT architectures are quite equivalent, differentiating the type of input that the Transformer receives. This reflects

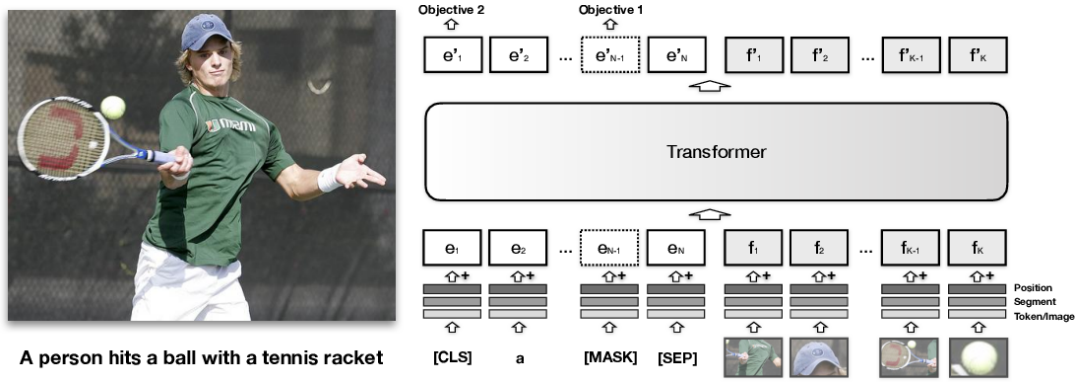


Figure 2.12: VisualBERT model architecture. [15]

the Transformers versatility in interpreting data from different modalities. In contrast to other systems, LXMERT applies VQA as a task right in the pre-training process, not being necessary to make changes to the architecture for fine-tuning in this task. This model was fine-tuned on the VQA and GQA datasets.

As seen in Figure 2.13, the model has three types of output: one for visual features, one for language and one for cross-modality. The language and vision outputs are the feature sequences generated by the cross-modality encoder. For the cross-modality output, they append a special token [CLS] at the beginning of the sentence. Then, the corresponding feature vector of this token in language feature sequences is used as the cross-modality output. Specifically for the VQA task, the model has a multi-class classifier that takes this special token embedding as input. Like with VisualBERT, LXMERT uses a classification answer head, meaning the number of classes corresponds to the possible answers and therefore, the range of answers that the model can give is limited.

Several other works [68–72] present question answering models that require common sense and reason over knowledge from external sources. However, these models seek this external information through Knowledge Bases and not in a specific context as we require.

2.5 On the Creation of Question Answering Datasets

Our proposal introduces a new task within the QA research area, and as we demonstrated in the previous section, there is still no dataset with our required settings. For this reason, a crucial step of our work will be on the contribution of a such novel dataset. However, collecting interesting, diverse, and well-posed questions and respective answers, especially in a multimodal setting, is a significant challenge requiring a well-thought methodology.

Cambazoglu et al. [73] presented an extensive review of publicly available datasets commonly used in QA research and provided background information about how each dataset was created. This study shows various manners to assemble a new dataset, from using frameworks for automatic question generation [49, 53] to even utilizing public

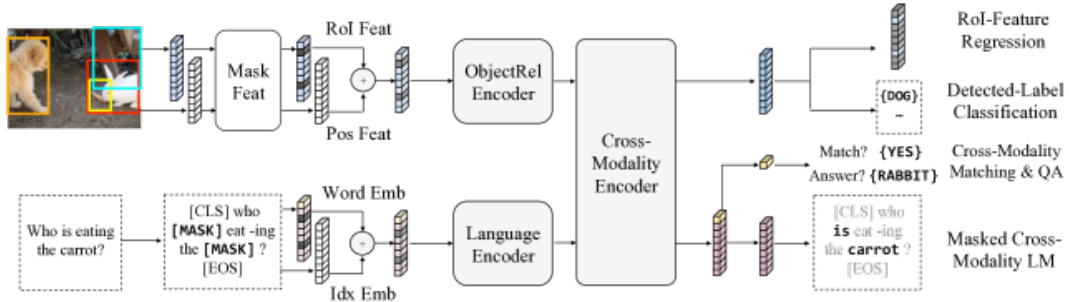


Figure 2.13: LXMERT pre-training tasks. [18]

sources like online forums and games [74, 75].

Among all the presented methods, the most popular one is crowdsourcing [4, 6, 9]. It consists of collecting information, opinions or results from a large distributed group of workers, usually using an online platform. In the dataset creation process, this method allows us to collect human annotations at a considerable scale, from people all around the work with different knowledge and thoughts, mimicking the end-users of our envisioned system.

2.5.1 SQuAD

SQuAD (Stanford Question Answering Dataset) [4] is one of the most known extractive datasets and tests the systems capability to determine whether the answer is within a given passage and refrain from responding if it does not. It consists of 100,000 answerable and 50,000 unanswerable questions based on passages extracted from Wikipedia articles.

To create this dataset, Rajpurkar *et al.* started by gathering 10 000 English Wikipedia articles with the highest PageRank values and selected 536 randomly. Then, they separated these articles into passages with at least 500 characters. The questions and answers were generated by crowdworkers.

For the answerable questions, each crowdworker received a passage and had the task of asking and answering (by highlighting the passage segment) up to five questions based on the passage content. Moreover, for each question in the training and test sets, at least two additional answers were obtained from crowdworkers. In a second phase, for the unanswerable questions, the crowdworkers were tasked to generate up to five questions that cannot be answered by the presented Wikipedia passage.

2.5.2 VQA

The VQA dataset [6] is one of the most widely used datasets to evaluate new visual models, as it has a wide range of questions. It contains over 204K real images from the COCO dataset [64] and 50K abstract scenes from Amazon Mechanical Turk (AMT)². Both real images and scenes include about 20 human models spanning genders, races, and ages

²<https://www.mturk.com/>

Stump a smart robot! Ask a question about this scene that a human can answer, but a smart robot probably can't!

Updated instructions: Please read carefully

Hide Show

We have built a smart robot. It understands a lot about scenes. It can recognize and name all the objects, it knows where the objects are, it can recognize the scene type (e.g., kitchen, beach), people's expressions and poses, and properties of objects (e.g., the color of objects, their texture). Your task is to stump this smart robot! In particular, it already knows answers to some questions about this scene. We will tell you what these questions are.

Ask a question about this scene that this SMART robot probably can not answer, but any human can easily answer while looking at the scene in the image. IMPORTANT: The question should be about this scene. That is, the human should need the image to be able to answer the question – the human should not be able to answer the question without looking at the image.



Your work **will get rejected** if you do not follow the instructions below:

- Do not ask questions that are similar to the ones listed below each image. As mentioned, the robot already knows the answers to those questions for the scene in this image. Please ask about something different.
- Do not repeat questions. Do not ask the same questions or the same questions with minor variations over and over again across images. Think of a new question each time specific to the scene in each image.
- Each question should be a single question. Do not ask questions that have multiple parts or multiple sub-questions in them.
- Do not ask generic questions that can be asked of many other scenes. Ask questions specific to the scene in each image.

Below is a list of questions the smart robot can already answer. Please ask a different question about this scene that a human can answer "if" looking at the scene in the image (and not otherwise), but would stump this smart robot:

Q1: What is unusual about this mustache? (The robot already knows the answer to this question.)

Q2: What is her facial expression? (The robot already knows the answer to this question.)

Q3: Write your question, different from the questions above, here to stump this smart robot.

prev next Page 2/3

Figure 2.14: ATM interface for collecting the questions for the VQA dataset assembly. [6]

with different expressions, over 100 objects and 31 animals in various poses. There are also representations of both indoor and outdoor scenes. The questions' answers were collected using the AMT crowdsourcing marketplace.

Given an image or scene, workers were asked to write a question that was related to the image. In Figure 2.14 we can see an example of the user interface for the questions crowdsourcing. To increase question diversity, the interface shows two previous questions for that same image. In total, three questions from unique workers were collected for each image/scene.

Regarding the answers, they gathered ten answers for each question from unique workers, ensuring that the worker answering a question did not ask it. They requested the workers to provide brief answers and to avoid using conversational language or opinions. They also asked the workers: "Do you think you were able to answer the question correctly?" and gave the choices of "no", "maybe", and "yes". In the appendix is available an example of the answers user interface.

The final dataset includes 614,163 questions and 7,984,119 answers for 204,721 images from the MS COCO dataset and 150,000 questions with 1,950,000 answers for 50,000 abstract scenes.

2.5.3 OK-VQA

As we were able to verify in the previous section, the dataset most similar to ours is the OK-VQA [9]. This dataset enhance others such as VQA [6] and combine visual recognition with information extraction from a source outside the image. To create the OK-VQA

dataset, they used COCO [64] images since these are visually complex and ideal for knowledge-based questions and assemble the questions through the AMT platform.

Firstly, given an image, each worker was tasked with writing a question that could fool a "smart robot" and related to the image content. Most importantly, they specified to question about something that would require some outside knowledge to respond and not about what is in or how many of something there is in the picture. In a second phase, they tasked the workers to label each question-image pair with an answer, and, in the end, they had five answers from different workers for each pair.

After gathering all the questions and answers, they end up filtering out the low-quality ones, removing questions if there were more than five instances of that question answer as the most common one (so that the answer distribution was uniform) and if it had no inter-annotator agreement on the answer. This process ensured that the dataset asked the challenging external knowledge questions and had as minimum bias as possible.

Although they consider the support of external knowledge necessary to answer the questions, they did not include passages in the crowdsourcing tasks. Furthermore, these passages are not available either, forcing the systems to do a prior search for them before answering the questions.

In 2021, Qu *et al.* [67] present a system for passage retrieval for outside-knowledge visual question answering. They considered a collection of 11 million passages, each containing, at most, 384 "word pieces" with intact sentence boundaries. In the end, they provided passages for the OK-VQA dataset.

NEWS MULTIMODAL QA DATASET

In this chapter, we describe the process of assembling our dataset and its key features in detail. We begin by defining our task, including its composition and objectives. Then, we delve into the complete process of assembling the dataset samples. Additionally, we examine alternative methods that were considered for obtaining question-answer pairs. Lastly, we provide a comprehensive overview of the final dataset, highlighting its unique characteristics, features, and specificities.

3.1 News Multimodal QA Task Formulation

News pieces seek to inform users in a concise yet comprehensive manner. As such, they are linguistically and visually rich open-domain documents, dense in knowledge. From a multimodal perspective, the connection between news pieces’ text and illustrations is heterogeneous [76, 77], meaning that there are multiple news text-to-image grounding patterns, that journalists use when choosing a specific image.

Having a dataset that captures these traits is critical to address the challenging task of Multimodal Question Answering in the News domain. Existing datasets have a set of limitations w.r.t. capturing the news multimodality aspects, by either solely focusing on text [50, 53] or not fully grasping the inherent heterogeneity underlying news pieces and images connection [63].

We formulate the task of Multimodal News Question Answering and create a dataset named NewsMQA that embeds the aforementioned traits. Namely, given a question q_i about a news passage p_i and an image v_i , the goal is to extract the excerpt from the passage p_i that answers the question. To support this task, we now need to create the dataset. An instance of this dataset must be composed of four elements: an image and passage from a news piece, a question, and the answer. The question may be unimodal, meaning that one can find the answer just by interpreting a single modality (image or text passage), or multimodal, requiring a joint interpretation of the two modalities.

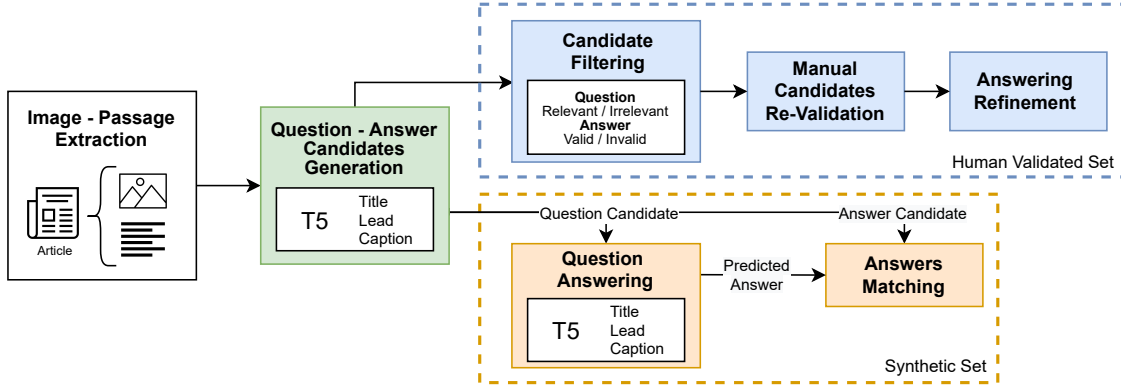


Figure 3.1: News Multimodal Question-Answering hybrid dataset creation pipeline.

3.2 Dataset Creation Pipeline

We seek to create a dataset whose number of samples can support the training of models tailored for the task while ensuring its quality. To strike this trade-off, our dataset creation pipeline (schematized in Figure 3.1) consists of three distinct phases: 1) image-passage extraction; 2) question-answer candidates generation and 3) two-step question-answer annotation. We adopt a two-step approach to filter the generated candidate pairs through both human validation and synthetic question-answer annotation, leading to the formation of a Human Validated and a Synthetic subsets.

In this section, we present a thorough examination of the procedures executed in each phase and the assembly of each subset.

3.2.1 Image-passage Extraction

We construct the dataset with news pieces from NYtimes¹. Namely, we sampled news piece elements (image and passage) from the NYTimes800k dataset [35], which has 444 914 articles written in English and has an average of 963 words/article. Most articles have an image at the top that correlates to the main topic and 39% have at least one more image throughout the text.

3.2.1.1 Topics and news articles sampling

Foremost, we defined a set of topics that we considered relevant and diverse that cover most of the news. Those topics are *art*, *books*, *food*, *movies*, *politics*, *science*, *sports* and *technology*. Then, for each topic, we compiled a list of related keywords (displayed in table 3.1), and retrieve news pieces by searching over their lead paragraphs using those keywords. To this end, we employed an Elasticsearch² Query string, which fetches the articles based on that query, using full-text search. Figure 3.2, we have a representative

¹ <https://www.nytimes.com>

²Elasticsearch, <https://www.elastic.co/>

Topic	keywords
arts	brush, painting, art, expressionism, sketchbook, oil painting, sculpture, painter, landscape, sketch, museum, acrylic-painting, artist, drawing, exhibition, craft, calligraphy, art gallery, portrait, illustrations, design, canvas, mural, ceramics, artwork, graphic design, illustrator, designer, watercolor, artistic, paint
books	bookmark, book-store, books, library, author, novel, best-seller, bookstore, reader, poetry, hardback, publishers, paperback, book, ISBN, poem, best-seller, writer, reading, e-book, booklet
food	meat, catering, vegan, restaurant, recipe, protein, nutrition, bakery, healthy food, diabetes, food, calories, lunch, grocery store, fast-food, gourmet, cooking, breakfast, junk food, organic, vegetarian, obesity, diet, food processor, delicious
movies	cinematography, cinema, theatre, film, actor, Netflix, television, filmmaker, airing, movies, documentary, Oscars, tv, plot, tv-show, Hollywood, cast, visual effects, audition, movie, movie-theatre, tv-drama, guest-star, commercial, script, Disney-movie, popcorn, sitcom
politics	politic, dictator, diplomacy, president, politician, elections, senate, campaign, conservative, voters, prime minister, democracy, corruption, politics, socialism, republican, constitution, minister, congress, council, independence, election
science	observatory, telescope, particle, scientist, international space station, climate change, fossil, atom, biology, astronomy, chemical, research, molecule, organism, physics, global warming, cell, chemistry, science, astrophysics, geology
sports	sports, rugby, gym, Olympics games, tennis, rowing, karate, jogging, player, taekwondo, athlete, athletics, team, workout, soccer, football, sports_competition, hockey, volleyball, coach, Olympic, basketball, golf, skate, racing, boxing
technology	cloud, neural network, hardware, computer, oracle, ai, web, machine learning, Microsoft, mobile, crowdsourcing, website, innovation, android, network, tech, robot, framework, ios, online, internet, Linux, automation, digital, software, azure, java, technology, edge-computing

Table 3.1: The chosen topics and set of keywords used to collect the articles for each of the chosen topics.

example of a query we executed to gather articles related to *sports*. With this query, we retrieve all the articles whose *lead paragraph* include the word *athlete*.

```
{
  "query": {
    "query_string": {
      "query": "(athlete)",
      "default_field": "lead_paragraph"
    }
  }
}
```

Figure 3.2: Example of a query we executed to gather articles related to sports where all the articles whose *lead paragraph* include the word *athlete* are returned.

We gathered a total of 92,828 distinct News pieces across all topics. Table 3.2 reveals the number of articles per topic and reveals that an article can be associated with more than one topic.

topic	arts	books	food	movies	politics	science	sports	technology	total
# articles	19 236	18 516	9 423	27 923	28 375	14 936	21 170	14 361	92 828

Table 3.2: Number of articles collected for each topic.

3.2.1.2 Image-passage pairs selection

Having these articles chosen, the next step is to select suitable pairs of images and passages that contain the necessary information to answer the questions correctly.

We considered all the images in each article and looked forward to choosing the best passages for each one, seeing how much the two are correlated. How to choose this passage? Is the text surrounding the image relevant? In that case, should we select the text above or below the image?

Throughout a news piece, journalists illustrate certain details with images that further help the reader contextualize and visualize the event. These images are thus grounded on their surrounding text and are usually placed before the corresponding text [2]. Therefore, for each image, our ground truth passage is assumed to be the two paragraphs following the image in our dataset.

Beyond the images and passages, we gathered other features, such as the article’s title, lead paragraph and image caption. In the end, each sample is composed of: an image URL, passage text, title, lead paragraph, caption and the topics assigned to the article.

After gathering all the samples, we only kept those that met the following criteria:

- the passage is at least 50 words, ensuring that the samples have enough information about the image and the News story;
- no values and/or invalid characters. We only keep samples whose characters are supported by UTF-8 encoding;

As shown in the second column of table 3.3, at the end of the first phase we created a total of 227,685 samples, from which we can now construct our question-answer pairs.

3.2.2 Question-Answer Candidates Generation

Given the image-passage pairs, we now need to generate questions and corresponding answers that inherit the intrinsic multimodality of the task we propose. To achieve this objective, we adopted an automatic, open-source approach presented by Patil [78]. They present three pipelines for question generation, including (1) question generation, (2) multi-task question answering and generation models, and (3) end-to-end question generation. Moreover, they provide various models for each pipeline, which are readily available in the Hugging Face library. All these models are based on the T5 model Raffel et al. [16], specifically, T5 models fine-tuned in an end-to-end multi-task setting, for question and answer generation and answer extraction [79] on SQuAD [4].

topic	# articles	# samples	# generated questions	answers within the passage
arts	19 236	39 288	36 732	6 810
books	18 516	30 372	31 012	5 409
food	9 423	17 328	17 166	3 956
movies	27 923	48 470	50 945	9 602
politics	28 375	44 219	41 490	7 225
science	14 936	25 496	23 424	4 596
sports	21 170	36 030	39 165	7 183
technology	14 361	23 885	23 155	4 582
total	92 828	227 685	221 517	44 404

Table 3.3: Progress of the number of articles and image-passage samples collected for each topic.

Since T5 only receives text as input, to generate questions that embed multimodal knowledge, the input context must be enriched with grounding information that acts as the bridge between the text and the image itself. For that purpose, we use the news image captions provided by journalists. These news-like captions describe the image by bridging the visual material with the news story [35, 76].

As such, given the i -th sample image-passage pair $\langle v_{di}, p_{di} \rangle$, from a news piece d , we create an input sequence s_{di} comprising the news piece title t_d , lead paragraph h_d and the image caption c_{di} . The first two elements provide general information about the news piece, while the caption c_{di} brings information that connects the two modalities. The passage p_i is kept hidden to ensure that the generated questions focus on the main news story and image caption.

The pipeline we used is divided into two steps: (1) **answer extraction** from the given context and (2) **question generation**, having into consideration the given context and extracted answer. Regarding the answer extraction, a T5-Small model fine-tuned in question answering and generation³ is employed. This model splits the input sequence s_{di} into sentences and extracts an answer a_{sj} for each sentence. In the second step, a T5-Base model fine-tuned in question generation⁴ is utilized to generate a question q_{sj} given the extracted answer a_{sj} and the input sequence s_{di} . Then, on these generated question-answer pairs, we performed a post-processing step to filter out invalid pairs. Namely, we keep only the pairs in which:

- the answer is included in the corresponding reference passage p_{di} ;
- the answer is not within the question itself;
- the question ends with the “?” character.

With this process, we gathered a total of 221 517 question-answer pairs. Of these, we only kept the samples whose responses were **within the passage** p_{di} we defined earlier,

³available on huggingface.co/valhalla/t5-small-qa-qg-hl model

⁴available on huggingface.co/valhalla/t5-base-qg-hl

Help Us Validate the Multimodal Questions About Articles

We are building a system capable of answering multimodal questions regarding news.
Given this image and passage extracted from a news article, tell us if this question and corresponding answer verify our criteria.

----- Consider these **CRITERIA** when validating the Question-Answer pair -----

- Question must be relevant regarding **both visual and textual elements**;
 - Question must have one singular answer;
- Answer must be a **span** (section of the given passage);
 - Answer must respond **correctly** the question;
(check the instructions to better understand our hit criteria)

You must select how relevant the [Question] is **AND** if the [Answer] is valid or not.

Your code:
QA-15353-CL

Instructions Shortcuts

Image



Passage
Most cities around the world denounce, or grudgingly tolerate, painting on public property. But on the Mediterranean island of Malta, the process is encouraged. In the shade of a pedestrian bridge, where old men and women sit on the concrete benches, staring out to sea, a wall has been splashed with color and the spray-painted words NO WAR.

Question
What is the name of the mural on Malta?

Answer
Mediterranean island

Insert your code here:

Select an option

[Question] Relevant + [Answer] Valid	1
[Question] Relevant + [Answer] Invalid	2
[Question] Irrelevant + [Answer] Valid	3
[Question] Irrelevant + [Answer] Invalid	4

Figure 3.3: Our Mechanical Turk interface for annotating the generated candidates.

leaving a final number of 44 404 dataset sample candidates. Table 3.3 presents the progress from the initial number of articles to the final set of 44 404 candidates assembled with this process.

3.2.3 Human Validated Set

In the previous step, we gathered a set of candidates for our dataset. Now, we resort to human annotation to filter and refine these question-answer candidates and assure their respective quality.

3.2.3.1 Candidate Filtering - AMT task

Firstly, we used the Amazon Mechanical Turk platform (AMT) and created a task to filter the generated samples. We chose this platform because it is versatile and has been successfully utilized in the creation of various datasets including for QA tasks involving images [6, 54].

Figure 3.3 shows an example of what a worker would see when performing the task. We start by explaining what we intend to accomplish and describing what the task consists

of: "Given this image and passage extracted from a news article, tell us if this question and corresponding answer verify our criteria."

Then, we list the criteria to consider when classifying the given sample candidate:

- Question must be relevant regarding both visual and textual elements
- Question must have one singular answer
- Answer must be a span (section of the given passage)
- Answer must respond correctly to the question

Following the instructions and before showing the actual sample, we reiterate that the worker "must select how relevant the [Question] is and if the [Answer] is valid or not", and provide the worker with a code created particularly for that sample ("Your code:"). The worker must copy this code and paste it into the right place at the end of the form before submitting it. This way, we filter out workers who actually read the task introduction and instructions, from workers that do not execute the task correctly.

Lastly, we display an image-passage pair, a question-answer pair about them and four options from which the worker has to choose one:

- [Question] Relevant + [Answer] Valid
- [Question] Irrelevant + [Answer] Valid
- [Question] Relevant + [Answer] Invalid
- [Question] Irrelevant + [Answer] Invalid

From the candidates' pool, and considering our budget, we validated 10 000 samples through this task. When sampling the candidates, we did it in a per-topic stratified manner to ensure topic balance. 1,400 related to the topics of Arts, Books, Movies, Politics and Sports; 1100 to Science and Technology; and 800 related to Food.

When launching the task for the public, we counted two qualification requirements regarding the workers that could respond. The HIT Approval Rate (%) for all Requesters' HITs must be greater than 97, and the number of HITs Approved greater than 1000 (where a HIT is a response from the worker considered accepted).

Upon receiving responses from workers, we automatically rejected any worker who did not use the code provided for this sample and those who always responded with the same option. Each time a response was rejected, the respective sample became available again for new workers to respond correctly. In the end, 10 707 assignments were done, 707 were rejected and 10 000 were approved. Table 3.4 presents the number of responses for each rating in the approved assignments. 257 distinct workers participated in the task and provided an average of 42 responses each. The average response time for each response was 56.46 seconds. Each worker received \$0.03 for each accepted answer. We spent a total of 400\$ (300\$ on rewards and 100\$ on Fees to Mechanical Turk). Besides the time spent and the answer they gave to the task, we did not require or save any other personal information about the workers.

[Question] Relevant + [Answer] Valid	[Question] Relevant + [Answer] Invalid	[Question] Irrelevant + [Answer] Valid	[Question] Irrelevant + [Answer] Invalid
4776	3044	1529	651

Table 3.4: Analysis of the number of responses for each classification in the AMT crowd-sourcing task.

At the end of this step, we kept only the samples that workers classified as having a relevant question, filtering a total of 2180 low-quality candidates.

3.2.3.2 Manual Candidates Re-Validation

Through the task, we intended for the workers to evaluate whether the generated question and answer met the defined criteria. However, this is a complex and cognitively demanding task, as the worker had to meticulously analyse all elements, including verifying if the question was multimodal, linking the passage and image, and determining the accuracy of the proposed answer.

As such, we conducted a second human validation step over the candidates, with a restricted group of annotators. We asked 4 annotators to re-validate all candidates whose question was annotated as "relevant", which corresponds to a total of 7820 candidates, with 4776 "valid" and 3044 "invalid" answer samples. The validation task was framed in the exact same manner as for the AMT workers, having the same criteria used to assess the relevance of a given question and the correctness of an answer.

Table 3.5 displays the results. 3293 samples were annotated as having relevant questions. From these, 1547 had valid answers and the remaining 1746 invalid ones.

Re-Validated Samples	[Question] Relevant + [Answer] Valid	[Question] Relevant + [Answer] Invalid	[Question] Irrelevant
7820	1547	1746	4527

Table 3.5: Analysis of the re-evaluation of the samples AMT workers considered to have a relevant question but an invalid answer.

3.2.3.3 Answering Refinement

As a result of previous steps, 6707 invalid candidates were eliminated and 1547 qualified candidates were selected for our dataset. Additionally, there were 1746 candidates whose answers were classified as incorrect. To avoid losing these samples, we perform one final task. We asked 16 annotators to answer the 1746 samples whose question was deemed as "relevant", but with an "invalid" answer. The annotators interacted with the interface shown in Figure 3.4. Given an image, passage and question, the task consisted in providing a correct answer considering the following instructions:

We are building a system capable of answering multimodal questions regarding news.
Given an image and passage extracted from a news article, write the answer for these questions that are related to both image and text.

- Assume that the entities identified in the question exist in the article story

About the answer

- Must be a span (select a continuous part of the given passage and copy it to the answer column)
- The span must respond correctly to the question

Place a " --- " if:

- The question is not relevant / does not make sense
- There is no correct answer for the question within the given context
- Some lines are in a gray color, you don't need to answer these questions, we already marked them as invalid

image	passage	question	answer
	Trace the history of the finest pieces of postwar art, and many of them lead back to Ileana Sonnabend, the gallery owner and collector, who died in 2007. Andy Warhols Liz #1 (Early Colored Liz), which sold at Sothebys for \$18 million last month? Sonnabend. Robert Rauschenbergs 20th-century mixed-media collage Canyon, which the Museum of Modern Art won over the Metropolitan Museum last year? Sonnabend. In honor of that gift from the Sonnabend family, the Modern is presenting an exhibition devoted to Sonnabends career, which spanned half a century. She is widely credited with introducing postwar American art to Europe in the 1960s and Italian Arte Povera to the United States. She also supported the Conceptual, new-media and performance-based art of the 1970s. The show, Ileana Sonnabend: Ambassador for the New, explores her eye through works that she presented in her galleries in Paris and New York. Among the 40 artists included are Jasper Johns, Vito Acconci, John Baldessari and Jeff Koons. (Through April 21, 212-708-9400, moma.org.)	What was the name of the gallery owner and collector who died in 2007?	Ileana Sonnabend

Figure 3.4: Interface the annotators interacted with to answer the candidate questions.

- The answer must be a span (select a continuous part of the given passage and copy it to the answer column);
- The span must respond correctly to the question.

As a result of the filtering, re-validation, and answer refinement processes, we gathered a total of 3293 high-quality $\langle v_{di}, p_{di}, q_i, a_i \rangle$ samples for what we named the Human Validated subset, each including an image v_{di} , a passage p_{di} that refers to the image, a question q_i and its answer a_i . Figure 3.5 showcases a selection of samples from this subset.

3.2.4 Synthetic Set

Although we managed to get pairs with high-quality assurance, the past method ended up being quite expensive and time-consuming. Moreover, we earned only a few samples comparing the number of initial question-answer pairs generated and the time we spent on the re-evaluation. Hence, we adopted an automatic approach to efficiently and cost-effectively complement and scale up the dataset with additional samples.

The synthetic QA generation procedure employed is based on ensuring roundtrip answer consistency [10]. This strategy has shown to be effective in closely related tasks, such as visual question-answering [11]. The rationale is to have two different models, one model generating a question and an answer, and a second model predicting the answer to the generated question. If the two answers match, consistency is achieved.

Regarding the question generation, we used the remaining 34404 candidates generated with the pipeline described in section 3.2.2. Recapitulating, this model received

3.2. DATASET CREATION PIPELINE

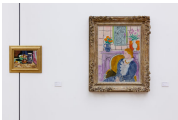
 Passage: In Five Feet Apart, Stella (Haley Lu Richardson) is a teenager with cystic fibrosis. Will (Cole Sprouse), a newcomer to the hospital where she resides, also has the disease and is undergoing a drug trial for an antibiotic-resistant infection. They don't mesh, at first. She is obsessive-compulsive and effusive; she dutifully takes her meds and keeps a YouTube diary about her life. He is a brooding artist and rule-breaker convinced that they are breathing borrowed air. Because of the risk of cross-infection, they must stay six feet apart. Five feet, as measured by the pool cue they carry between them, will be their eventual concession to love and living dangerously. Question: What separates them? Groundtruth: cystic fibrosis	 Passage: I didn't install controls for light switches, blinds or fireplace because it can get pricey (Control4 switches cost \$160 each), but they are also on a totally different level of sophistication. But I did set up a Control4 entertainment system, which proved to be quite useful, though not cheap. The grand total for this setup, including a Control4 hub (\$750), the Control4 Door Station (\$800) and a touch-screen controller (\$900), was about \$3,000 including installation. Was it worth it? Well, a few weeks ago my mother-in-law was babysitting and couldn't figure out how to find her television show, so I logged in with my phone, turned on Homeland and set the volume so it wasn't too loud. And guess where I was? Sitting on an airplane. Question: What is the cost of this device? Groundtruth: \$900
 Passage: He grins impishly out of lithographs or woodcuts or gets translated rather literally into wood, as in White Gloves, 4 Wheels (2007), where the figure perches on a pallet outfitted with wheels and spattered with paint, his arms thrust outward. In The Crying Sand (2006), Pinocchio's neutral coloring links him bluntly with his fictional wooden origins, while in another wooden piece, Pinocchio at Night (2004), the artist attempts to turn him into a darker character, literally, by coloring the sculpture with dark enamel paint. It should be noted that the prints made to accompany a 2006 translation of Pinocchio, which Mr. Dine illustrated, are more interesting. There are 36 hand-painted lithographs on view, along with layouts of the text; the book itself is for sale in the museums. Here, Mr. Dine's images serve as expressive accompaniments to a translated version of the original text, in which Pinocchio fights off knife-wielding assassins, does prison time and gets eaten by a fish before making it to the promised land of boyhood. Question: What is the name of the character in which Mr. Dine's inspired? Groundtruth: Pinocchio	 Passage: Beyond this North-South bifurcation, the earliest paintings here also reflect the extraordinary variety of artistic heritages Bearden was exposed to as a young man. Paintings like Untitled (Harvesting Tobacco) (circa 1940) and Untitled (Husband and Wife) (circa 1941) feature large, full-bodied figures like those of George Grosz, the German-born migr painter who was Bearden's mentor at the Art Students League in the early 1930s and who encouraged Bearden to look at Dutch and Flemish painting. The Family (circa 1941), with its horizontal orientation and multiple figures, is reminiscent of the Mexican muralists, who were influential in New York in the 1930s. Presage (1944) displays the fragmentation of forms that was the hallmark of Cubism. Bearden started working as a political cartoonist; it was Grosz who persuaded him to switch to fine art and the paintings here were created around the same time that he wrote essays like The Negro Artist and Modern Art (1934) and The Negro Artists Dilemma (1946). Question: What is the title of this piece at the Newark Museum? Groundtruth: Untitled (Harvesting Tobacco)
 Passage: LONDON Leonardo da Vinci's Bust of a Warrior, a pyrotechnic graphic performance by the young artist, is now one of his most famous drawings. It was inspired by a lost cast metal relief by Leonardos master, Verrocchio, and the lines of the drawing are inscribed on the paper in metal. For this is one of the artists works in silverpoint. The drawing is one of the treasures of the British Museum, which has the most important collection in the world of drawings in metalpoint in silver, gold, lead and other metals and which is now staging, in association with the National Gallery of Art in Washington, its first exhibition of nearly 100 works, stretching from the early Renaissance to the present day. Question: What is this treasure of the British Museum? Groundtruth: Leonardo da Vincis Bust of a Warrior	 Passage: MONTREAL When Germany, the worlds top team, meets the United States, the worlds second-ranked squad, in a Womens World Cup semifinal game Tuesday, there will be a matchup of goalkeepers befitting the stakes. On one end will be Hope Solo, whom many, particularly those on her team, consider the best keeper in the world. On the other side will be Nadine Angerer, who two years ago became the first goalkeeper to win FIFAs World Player of the Year award. Question: Who is she? Groundtruth: the first goalkeeper to win FIFAs World Player of the Year award
 Passage: WHAT about an audiobook? Pharrell Williams asked, sitting at the head of a conference table at the Park Avenue South offices of Rizzoli as he looked at the nearly finished galleys for an October release called Pharrell: Places and Spaces Ive Been. Here was a lavish coffee-table book filled with images of the many products he has designed in collaboration with other artists and fashion designers, and interviews between Mr. Williams and the likes of Jay-Z, Anna Wintour and Zaha Hadid, which do not exactly lend themselves to the narrative treatment. But why not? Question: Who sat on the chair designed by Domeau & Pérès? Groundtruth: Pharrell Williams	 Passage: Whats on TV EUPHORIA (2019) 10 p.m. on HBO. Drugs, sex, love and social media fuel this modern, anxiety-ridden teenage drama centered around Rue Bennet (Zendaya), a 17-year-old who is searching for connection after leaving rehab. Rue and her orbit of young friends and classmates struggle to find their identities and sense of purpose within a haze of substance-fueled parties, preparation for college and a hefty amount of family and personal trauma. Sam Levinson (The Wizard of Lies) wrote the eight episodes of the American adaptation of the Israeli series, which includes not only Zendaya, a former Disney Channel star, as the cast lead, but also Drake as an executive producer. Question: What is the name of the model and actress who stars in the American version of Euphoria? Groundtruth: Zendaya
 Passage: OSLO When the Henie Onstad art museum near here received an unexpected call in June 2012 on behalf of the heirs of a French art dealer, the director, Tone Hansen, had no idea that it was the beginning of an odyssey that would end with the museums giving away one of its prize paintings and opening a major exhibition built around that works absence. Visitors to the exhibition, In Search of Matisse, which runs until Dec. 13, will not be seeing Blue Dress in a Yellow Armchair, which the artist painted in 1937 and had been in the museums possession. In March 2014, the painting was returned to the heirs of the French art dealer, Paul Rosenberg. Question: What was the name of this painting that was returned to Paul Rosenberg's heirs in 2014? Groundtruth: Blue Dress in a Yellow Armchair	 Passage: Donald J. Sobol (his parents gave him the middle initial, but it did not stand for anything, his son said) was born on Oct. 4, 1924, in the Bronx. His father owned gas stations, which he later sold to Standard Oil. Mr. Sobol graduated from the Ethical Culture School in Manhattan in 1942 and enrolled at Oberlin College. In the middle of his freshman year he enlisted in the Army and served during World War II as a sergeant in a combat engineer battalion in the Pacific. He returned to Oberlin in 1946 and later gave much of the credit for his career to an English professor there, John Singleton, who gave him a personal course in advanced creative writing. Question: What was the name of the creator? Groundtruth: Donald J. Sobol

Figure 3.5: Selection of Human Validated subset samples.

CHAPTER 3. NEWS MULTIMODAL QA DATASET

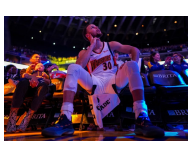
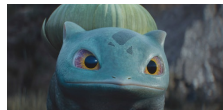
 Passage: "The Terminator" (directed by James Cameron, 1984)"This was my favorite movie for ages. It let the characters and emotion lead the story rather than the tech, which was so rare in sci-fi. The film is also a great primer on Cameron's cutting, action and mood." "Cyborg" (directed by Albert Pyun, 1989)"I saw this as a kid and it was just nuts. It had a really snappy percussive tempo to the fight sequences — that was the Asian influence coming in. They also were creative with how to make the action cool without a huge effects budget." Question: What movie did Vincent Klyn star in? Groundtruth: Cyborg	 Passage: The reality-show villain has evolved. Ladies and gentlemen, meet the YouTube villain. Look no further than Jake Paul, a YouTube-generation successor to platinum-blond miscreants like Spencer Pratt of "The Hills" and Jonny Fairplay of "Survivor," mixed with a little bit of the "Jackass" crew. Question: What is the name of the YouTuber who has taken a heel turn in recent months? Groundtruth: Jake Paul
 Passage: In some cases, more radical action might be necessary. One top auto analyst recommended on Thursday that General Motors, which has lost money in Europe for 12 straight years, sell or even close down its European operations before they undermine the company's overall comeback from its government bailout and bankruptcy. "One of the worst things in the auto industry is owning a cash burning, resource consuming business," the analyst, Adam Jonas of Morgan Stanley, wrote in a lengthy research report about G.M.'s Opel unit. "One of the best things is getting rid of that business." Question: What brand did G.M. want to leave Europe? Groundtruth: Opel	 Passage: ANCHORAGE — An independent candidate, Bill Walker, has won the Alaska governor's race after an aggressive challenge to Gov. Sean Parnell, a Republican. Mr. Walker, a lawyer and former mayor of Valdez, was ahead on election night, but the race was too close to call until thousands of absentee and questioned ballots were counted this week. He is the first candidate unaffiliated with a party since statehood to win the office. Question: What is the name of the candidate who won the governor's race in Alaska? Groundtruth: Bill Walker
 Passage: Mr. Goodman, 61, opened Trash and Vaudeville in 1975, in what was formerly Limbo, a hippie clothing shop. He was a recent graduate of the Fashion Institute of Technology and borrowed \$1,000 from his parents for the down payment. He named it Trash and Vaudeville because people often asked, "What's that trash you're wearing?" For years, Mr. Goodman spent his days "rag picking," searching through warehouses of clothes from previous decades for music-inspired looks: Hawaiian shirts, gabardine shirts, go-go boots. Question: What was the name of the clothing store that left its East Village home? Groundtruth: Trash and Vaudeville	 Passage: The Golden State Warriors followed up their second consecutive N.B.A. title — and third in four years — by adding DeMarcus Cousins, the four-time All-Star. It became commonplace to declare that they had ruined the league by removing any suspense to the season. A starting lineup of five All-Stars, two of which are former recipients of the Most Valuable Player Award, was just too much. There was palace intrigue, however. Kevin Durant and Draymond Green feuded in public, Cousins was routinely exposed on defense, Klay Thompson had a rough start and Stephen Curry's shot seemed to fail him before he began wearing contacts on the court. Question: What is the name of the greatest 3-point shooter in N.B.A. history? Groundtruth: Stephen Curry
 Passage: "Fast Color" is many things — a dreamlike dystopian drama; a warm celebration of family and female power; a teasing hint of superhero-origin story — none of which fully explains its gentle grip on our attention. What it is not, is speedy, with a narrative that moves carefully, even languidly, and visuals so beautifully patient that we have time to memorize their texture and import. Special effects are all the more so for being used sparingly, and the movie's few characters feel no need to be constantly explaining themselves. For fairly long stretches, there are no major incidents — which is not the same as saying that nothing happens. Set in a near-future America parched and wilting from eight years of drought, the movie drops us into a car beside an anxiously fleeing Ruth Question: What is the name of the character in Fast Color? Groundtruth: Gugu Mbatha-Raw	 Passage: These supporters argue machine intelligence will produce so much economic surplus that we could collectively afford to liberate much of humanity from both labor and suffering. The most idealistic thinkers see the plan as a way to foster the sort of quasi-utopian future we've only encountered in science fiction universes like that of "Star Trek." As computers perform more of our work, we'd all be free to become artists, scholars, entrepreneurs or otherwise engage our passions in a society no longer centered on the drudgery of daily labor. Question: What is the Atlas Robot? Groundtruth: machine intelligence
 Passage: A gaggle of these friendly, bulbous creatures come to Pikachu's rescue, falling all over one another. "We took a series of clips of bulldog puppies jumping on top of each other and copied their movements with a rotoscope," Nordby said. "The playfulness and top-heaviness you have with bulldogs, where they don't quite seem like they're ever quite comfortable on their feet, worked really well with the body shape of the bulbasaur." Snubbull Question: What was the inspiration for Sonic the Hedgehog? Groundtruth: bulldog puppies	 Passage: He is also, at least implicitly, a severe critic of the laziness and mediocrity that characterize so much popular culture. He criticizes partly by example, by avoiding the usual kid-movie clichés and demonstrating that a clear, accessible story can also be thoughtful and unpredictable. "Ratatouille" features no annoying sidekick and no obtrusive celebrity voice-work, and while Remy is cute, he can also be prickly, demanding and insecure. Moreover, his basic moral conflict — between family obligation and individual ambition — is handled with unusual subtlety and complexity, so that the reassurances and resolutions of the movie's end feel earned rather than predetermined. Question: What is the rat with the palate in "Ratatouille"? Groundtruth: Remy

Figure 3.6: Selection of Synthetic subset samples.

as input the title t_d , lead paragraph h_d and image caption v_{di} . Now, apart from these elements, the second answer support model takes the generated question as input and predicts the answer. The model used for this is a T5-Base⁵ model fine-tuned for question answering in the SquAD v1 dataset. Through this procedure, we created a Synthetic set of 7883 samples. Figure 3.6 presents a variety of samples from this subset.

3.3 Crowdsourcing tasks for Question-Answer pairs Generation

The process of collecting questions and answers about the generated images and passages was a time-consuming and challenging endeavour that involved multiple iterations and attempts. In this section, we introduce another approach we have experimented with for generating the question-answer pairs. We will also outline the reasons for testing this method and discuss why it was ultimately unsuccessful.

3.3.1 Using a Single Task for Dataset Samples Generation

The initial method that was explored in our experiments was the utilization of the ATM platform to launch a single task to collect the question-answer pairs. In the task, we instructed the workers to directly generate two questions and respective answers based on a given image-passage pair.

As shown in the example of the task in Figure 3.7, we start by presenting a brief introduction of our project, explaining what we aim to accomplish with the task's result and a general idea of what we intend the user to do. We also give the worker a code created particularly for that pair ("Your code:"). The worker must copy this code and paste it into the right place at the end of the form before submitting it.

After this introductory section, we direct the worker to review the instructions checklist, which outlines the guidelines for the acceptable questions and answers. The presented standards are:

- Questions must require an analysis of both visual and textual elements
- Answers must be a span (section of the given passage)
- Do not repeat questions
- Questions must have one singular answer (i.e. do not ask questions that have multiple parts)

Finally, we indicate the image and passage pair the worker must consider and leave space to write their questions and answers. At the bottom, we can verify the place where it is supposed to leave the initially given code ("Insert your code here:").

⁵available on <https://huggingface.co/valhalla/t5-base-squad>

Instructions

Shortcuts

Instructions

Follow the instructions below:

- Questions must require an analysis and knowledge of both visual and textual elements;
- Answers must be a span (section of the given passage);
- Do not write questions similar to the one listed;
- Do not repeat questions;
- Questions must have one singular answer (i.e. do not ask questions that have multiple parts);

Help Us Answer Multimodal Questions About News

We are building a system capable of answering multimodal questions regarding news. Given this image and passage extracted from a news article, write up to two questions that are not answerable by analysing the image or the text individually and the corresponding answers.

(check the instructions to better understand our hit criteria)

Your code:
QA-10741-CL

Image:



Passage:

The Whitney curatorial team Carter Foster, Barbara Haskell, Henriette Huldish and Dana Miller, led by Donna De Salvo has given the third floor an umbrella title: "The Pure Products of America Go Crazy." And from the opening images of Old Glory and cosmetic surgery (Jasper Johns and Andy Warhol, respectively) the ensemble does indeed feel more than a little unhinged. Some of the dementia is just zaniness, crazy-lite. This would describe Man Ray's surrealist "La Fortune," with its billiard table zooming like a flight deck off into space; and Laurie Simmons's photomontage of a walking camera; and a bunch of quirky short films by Jack Goldstein, who died in 2003. (He's one of several recently deceased artists in the show: Mark Lombardi, Agnes Martin and Nam June Paik are others.)

Question 1

Answer 1

Question 2

Answer 2

Insert your code here:

Submit

Figure 3.7: Our Mechanical Turk interface for the composition of question-answer pairs.

The estimated time required to read the introduction/instructions and a passage is approximately 1 minute and 10 seconds and 1 minute, respectively. Moreover, the time required to formulate a question that meets the specified criteria and confirm that the answer is a contiguous span within the passage has not been accounted for.

We did a small test with 20 samples to check the workers' behaviour and confirm if we could use this task to generate all the question-answer pairs. Unfortunately, the results of this approach indicate it may not be a good method to implement. The generated questions were either overly simplistic or contained spelling errors, and the answers were sometimes not found within the passage. Upon evaluating the average time taken by workers to complete the task, we observed that it took an average of 2 minutes. Why did this happen?

The reason for this outcome can be attributed to the difficulty of the task, which incorporates multiple complex steps, resulting in a time-intensive process, presenting challenges for the workers. It requires the worker to understand the instructions thoroughly, carefully read the entire passage, and generate one or two questions and answers that are related to both the image and the passage. Additionally, the requirement of accessing the task instructions through a separate tab may have also hindered its effective execution.

With this first attempt, we can conclude that workers look for tasks that are simple to comprehend and respond to, as well as have convenient access to the necessary information for their successful completion.

3.4 Dataset Analysis and Statistics

In this section, we explore the statistical properties of our dataset, starting with an overview of the full NewsMQA dataset. Then, we compare the Human Validated and Synthetic subsets and analyse the samples' distribution throughout the selected topics. Later we explore the type of questions on the full dataset, the challenges it raises and the influence each component had on the question generation.

3.4.1 Dataset Overview

Table 3.6 summarizes the final NewsMQA dataset distribution, detailing the number of samples for both the Human Validated and Synthetic subsets, across the different topics. NewsMQA comprises a total of 11 176 questions (3293 manually and 7883 automatically annotated) about 9986 image-passage pairs from 9531 different The New York Times articles ranging from March 2005 to August 2019.

3.4.2 Human Validated vs. Synthetic sets

Despite conveying the exact same elements, the two sets inherently have distinct characteristics. Table 3.7 compares the average question and answer lengths, as well as their

Table 3.6: Distribution of news pieces (N), images (I) and questions (Q) per topic.

	Human Validated			Synthetic		
	# N	# I	# Q	# N	# I	# Q
arts	590	609	710	1068	1141	1199
books	571	593	737	1090	1154	1249
food	353	364	439	556	596	637
movies	659	675	801	2064	2141	2253
politics	498	500	562	1329	1372	1446
science	345	349	414	698	723	761
sports	417	425	488	1258	1306	1399
tech	400	414	498	689	719	769
Total	2832	2889	3293	7159	7461	7883

Table 3.7: Comparison between the subsets regarding the question and answer key stats.

	Avg. Length / Vocab Size		
	Questions	Answers	Passage
Human Validated	10.83 / 7455	5.9 / 7538	113.1 / 46256
Synthetic	10.81 / 13926	1.9 / 7005	103.9 / 86555
NewsMQA	10.82 / 17054	3.1 / 12168	106.7 / 102722

vocabulary diversity.

As can be observed, the Human Validated set comprises questions with longer answers and a greater variety of unique words. Therefore we expect it to have a consequently increased difficulty compared to the Synthetic subset. Having more diverse and creative answers on the Human Validated set is expected, especially since paraphrasing is trivial for humans, as well as the capacity to digest and synthesize information. However, even though the employed synthetic approach answers are shorter and less diverse, under a curriculum learning framing, they are still useful for training.

3.4.3 Dataset challenges

Figure 3.8 presents two samples from the NewsMQA dataset that highlight the different types of questions found within it, each with their own unique challenges for answering correctly. The question on the left is unimodal and can be answered solely by having access to the passage. The image does not add any kind of information that could generate a different answer.

On the other hand, the question on the right is multimodal and requires both the passage and image information to determine the correct answer, as the question’s reference is not evident without the accompanying visual context. This is a good example of how images can greatly impact the context of a question and the answer that is given. In this case, the art piece shown in the image is ‘The Garden’, and the correct answer is also ‘The Garden’. However, if the image had depicted a different art piece, the answer would have to be adjusted accordingly. For instance, if the art piece in the picture was a different one altogether, the answer would have to reflect that. Conversely, if the picture showed an

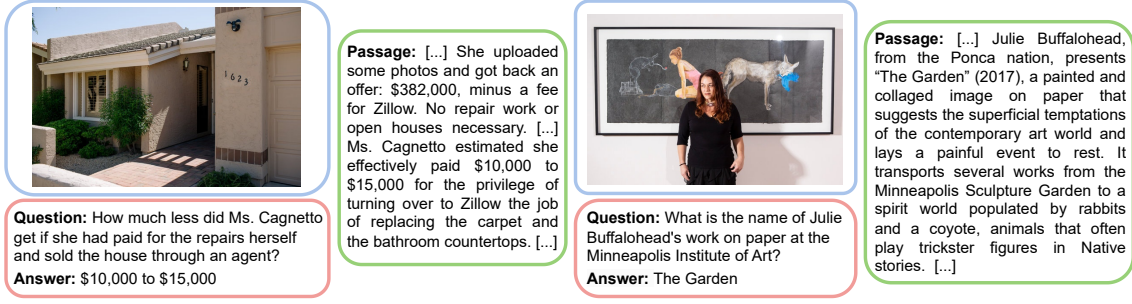


Figure 3.8: Two examples from the NewsMQA dataset with the corresponding image, passage, question and answer.

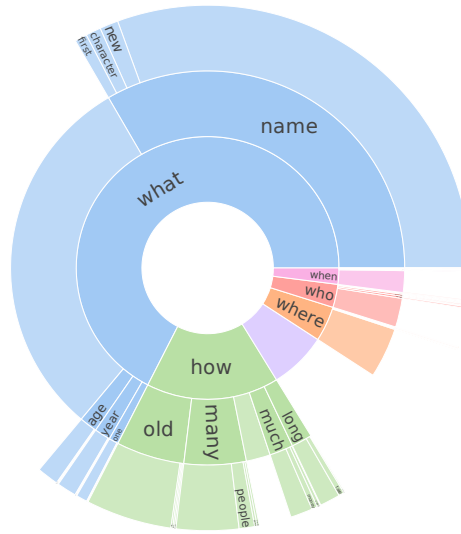


Figure 3.9: Distribution of the top 5 question types from their first words without counting the stopwords. The empty areas cluster the remaining terms.

art piece that was not mentioned in the passage or not in the "Minneapolis Institute of Art", the image would be irrelevant to the answer and the model should focus solely on the information in the passage to answer the question.

3.4.4 Questions' diversity analysis

Figure 3.9 depicts the distribution of question types covered in the NewsMQA dataset. The top 5 most represented question types are "what"(which represents 67,47% of the totality of question types), followed by "how", "where", "who"and "when". These show the broad coverage in NewsMQA of the different question types on the spectrum of the news domain.

3.4.5 Questions multi-modality analysis

NewsMQA is constructed such that the heterogeneous nature of the connection between images and text is inherited. As such, the context sequence used to generate candidate

Table 3.8: Analysis of the number of samples whose answer to the question is solely within one of the components used to generate the candidate pairs.

	Human Validated	Synthetic
Passage	3293 (100%)	7883 (100%)
Article’s title	167 (5.07%)	822 (10.43%)
Lead paragraph	1129 (34.28%)	2975 (37.74%)
Image caption	633 (19.22%)	2560 (32.47%)
Various components	1364 (41.43%)	1526 (19.36%)

question-answer pairs was built with a combination of unimodal (title, lead paragraph) and multimodal (image caption) elements.

Table 3.8 shows the impact of each of these components on the generation of the question-answer pairs. It can be seen that $\approx 19\%$ and $\approx 32\%$ of the answers, on the Human Validated and Synthetic subset, respectively, were extracted exclusively from the caption. Then, on average 44% of the answers were extracted from either the title or lead paragraph. Finally, an average of 30% of the answers appear in multiple context components.

3.4.6 Critical Summary

In this chapter, we explained in detail the process to assemble our dataset. News Multimodal Question Answering (NewsMQA) is a multimodal dataset that comprises a total of 11,176 questions about 9,531 articles from The New York Times. We have divided the dataset into two subsets: Human Validated, which contains 3,293 questions, and Synthetic, which contains 7,883 questions. The dataset includes both unimodal and multimodal question types, which aim to address the diverse and multimodal aspects of news stories. We will make NewsMQA available for public use.

TRANSFORMER-BASED MODELS FOR NEWS MULTIMODAL QA

This chapter focuses on the setup for benchmarking our dataset using transformer-based models. Firstly, we introduce the models that we employ for the benchmarking process. Then, we explore different visual context features that we use to enrich the models' input sequences. Lastly, we explain in detail our approach to incorporate these visual contexts into the models' input sequence.

4.1 Transformers for Span-based QA

In this thesis, we formulate the Multimodal QA problem in the news domain. This is particularly challenging due to the open-domain and multimodal nature of news articles. With images often complementing the textual medium, QA models must dynamically integrate visual and linguistic sources to ground their answers. These models require natural language and vision processing to find answers and must be able to learn multimodal representations that correlate both images and text elements. Additionally, such models need to be capable of answering open-ended questions and selecting a passage's span as the answer, which further adds to the complexity of the task.

Several vision-linguistic models inspired by the Transformers architecture have been proposed to tackle multimodal tasks such as image-text matching and VQA [15, 18, 30, 80–82] introduced in Section 2.2.3. When applied to a downstream QA task, these often turn the problem into a classification one, *i.e.* each candidate answer is treated as a class, requiring the model to select the answer from a set of pre-defined options. This is not suitable for open-ended QA like our proposed task, especially in the news domain, which is characterized by an extensive answer vocabulary and diverse answer length.

Marino et al. [9] has made significant strides in regard to the Multimodal QA problem. However, their approach entails searching through multiple documents to provide answers to a given question and the questions are not related to news articles. On the

other hand, Reddy et al. [63] has proposed an extension to the Oscar multimodal transformer [82] for span-based answer extraction. Yet, their approach requires the image caption as part of the input, which is not a requirement in our approach. Additionally, their method does not seem yet to be publicly available. Thus, to benchmark our NewsMQA dataset, we adopt three models as backbone models and enrich them with multimodal information. These were already introduced in Chapter 2, but let’s recapitulate them.

BERT [13] is one of the most widely used baselines for NLP tasks. It was pre-trained with Masked Language Modeling (MLM), where the model is trained to predict the missing tokens within the sequence, and Next Sentence Prediction (NSP), which involves predicting whether a sentence should come after another. We explain the training process and architecture in more detail in Section 2.2.1.3. Thanks to its architecture and pre-training process, the model may be fine-tuned by appending task-specific heads, in a classification or regression setting.

Extending BERT usage to represent and predict text spans, **SpanBERT** [14] presents a different pre-training process. Firstly, while in MLM individual tokens are masked, in SpanBERT they employ span masking and mask continuous spans of text for the model to predict. Moreover, they introduce a Span Boundary Objective (SBO), which involves predicting each token of a masked span using only the representations of the observed unmasked tokens at the boundaries. Finally, instead of using BERT’s NSP, which uses 2 segments, they pre-trained with longer full-length contexts of up to 512 tokens. We explore the model in more detail in Section 2.4.3.

Regarding the downstream QA task, both BERT and SpanBERT take a question and a context as input. During fine-tuning, the models are trained to locate the start and end tokens of the answer within the given context. They do this by assigning each token in the input sequence two probabilities: the probability of that token being the response starting token and the probability of being the response ending token.

Contrary to these first two encoder models based on span extraction, **T5** [16] is an encoder-decoder generative model. This was pre-trained with Deshuffling, where the sequence words are shuffled and the model is trained to predict the original text and Corrupting Span objective, similar to BERT’s MLM where the model is trained to predict the missing words within the input sequence. Downstreaming on QA, instead of locating an answer within a given context, T5 generates an answer based on the context information.

Despite the fact that these models are not inherently capable of accepting multimodalities, we have utilized them as backbones and put forward customized context-encoding schemes to incorporate the multimodal information necessary for solving the NewsMQA task. In specific, in all our experiments, we used as backbones the Large variants of T5 ¹, SpanBERT ², and BERT ³, which have 770, 340 and 340 million parameters,

¹To fine-tune the T5-Large, we used this repository:
github.com/huggingface/transformers/tree/main/examples/pytorch/question-answering

²available on huggingface.co/mrm8488/spanbert-large-finetuned-squadv1

³available on huggingface.co/bert-large-uncased-whole-word-masking-finetuned-squad

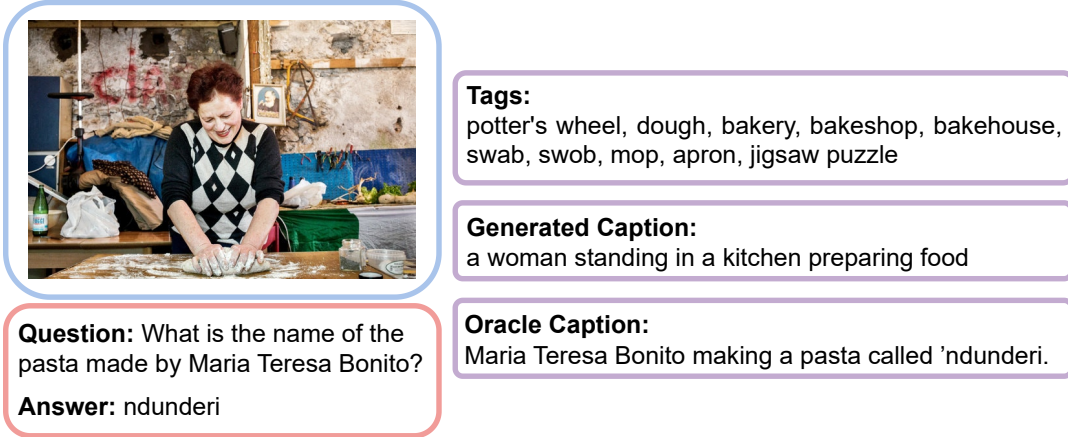


Figure 4.1: Example of the different visual cues we use to enrich the input sequence of the models. The visual cues are object detection tags and two variations of captions. This is a sample from the Synthetic subset.

respectively, all fine-tuned on the SQuAD v1 dataset [4].

4.2 Visual Domain Context

While backbone models receive a text sequence as input, NewsMQA comprises questions that require a multimodal answer grounding, *i.e.* interpreting the image is crucial. In these cases, the samples' context needs to be enriched such that visual domain information is embedded. Following Gui et al. [83], we propose to enrich each sample context with visual cues from the image. In particular, we decided to explore the use of object detection and captions as alternatives. Figure 4.1 displays a sample of the dataset (in this case just the question, answer and image) and an example of these various visual cues about the sample's image. In this section, we present and discuss the different visual cues we integrated into the input sequence of the models, with the aim of improving their performance.

4.2.1 Object Detection

Our initial strategy to augment the models' input with image-related information consists of including a list of objects detected within the picture. To achieve this, we utilized the Vision Transformer (ViT) model⁴ [29], which was pre-trained on ImageNet-21k, containing 14 million images and 21,843 classes, and subsequently fine-tuned on ImageNet 2012, with 1 million images and 1,000 classes. The ViT model takes an image, v_i , as input, and outputs a list of tags, each accompanied by a corresponding score. From this output, we extracted the top 10 labels with the highest scores.

Figures 4.1 and 4.2 illustrate examples of the object detection tags extracted. These examples also highlight the limitations of object detection tags and demonstrate how

⁴available on huggingface.co/google/vit-base-patch16-224

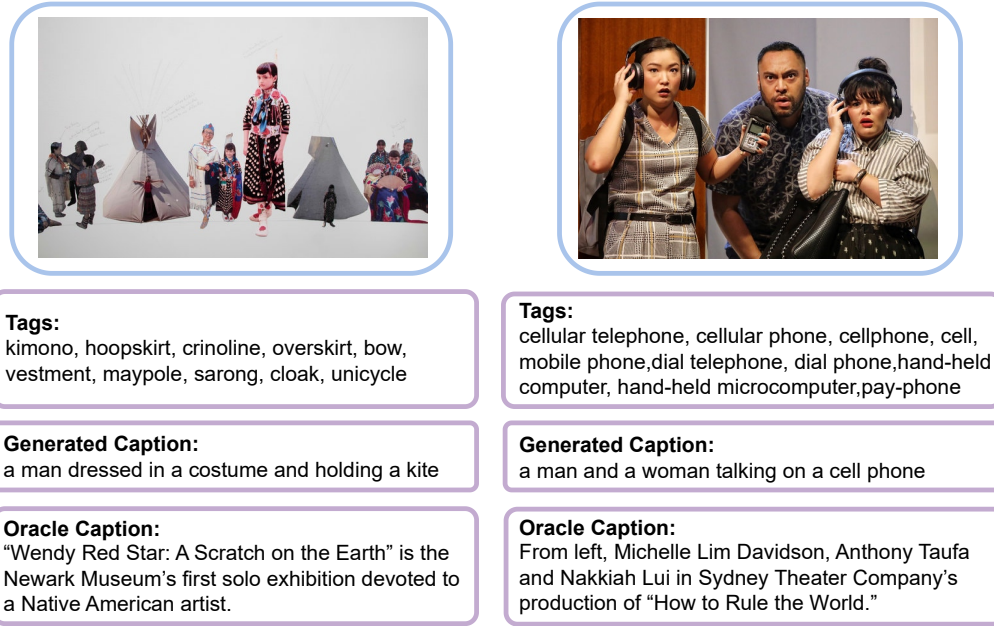


Figure 4.2: A comparative analysis of visual cues we present and the information each provides about the images.

these tags, despite conveying a wide range of terms, may not be the most useful for answering open-ended multimodal questions about news stories. While these tags can be useful for more typical visual question-answering tasks that focus solely on the image, they do not provide the specific knowledge needed for multimodal questions that require the identification of entities and specific concepts.

4.2.2 Image Captions

Alongside object detection tags, we also experimented with the integration of two different types of captions. One is automatically generated by an image captioning model, which provides a general description of the image content. The other is human-written, which provides additional context and information about the news story.

Generated Captions: We embed automatically generated image captions to each backbone input sequence. These captions were extracted employing a GPT-2 model⁵ [84] fine-tuned on the MS COCO dataset [64], with a ViT image backbone [29], which given an image v_i , returns a COCO-like caption c_i .

Original (Oracle) Captions: The most suitable captions for this task will be the article's original ones available on the NYTimes800k dataset. Entities play an important role in these captions, with 97% containing at least one named entity. Names of people are the most frequent entity type, comprising a third of all entities. Furthermore, 71% contain at least one face and 68% mention at least one person's name. Although these captions

⁵available on huggingface.co/nlpconnect/vit-gpt2-image-captioning

provide relevant insights, we can not forget they present privileged information as they were used to generate the dataset samples.

In Figure 4.2 we present two examples of both types of captions for the same image and compare the two. While the journalists themselves formed the original captions to provide a contextualized and specific understanding of the image linking it to the article story, the generated captions tend to describe the scene in a general manner, thus not covering specific entities or news story details as concepts or locations, but capturing explicit visual notions (e.g. objects, buildings, etc.).

One advantage of using original captions is that they capture key concepts and notions of the news story, providing additional information that may be relevant to answering multimodal questions. However, original captions may also introduce bias or subjectivity in interpreting the image and may not provide a complete description of the visual content. On the other hand, generated captions offer a more objective and standardized description of the visual content, which can be useful for certain types of visual questions. However, these may not capture the relevant entities or concepts needed to fully answer more complex questions.

4.3 Multimodal Input Sequence Formatting

Given a news passage and an image as context, our task requires being able to multiplex both textual and visual information to answer a question. One way to accomplish this is to enrich the input sequence with extra contextual information, in particular, elements that capture the knowledge depicted in the image [76, 83]. This requires carefully designed sequence encoding schemes to effectively model multimodal information and deal with the model’s input sequence size limitations. As the answer a_i needs to be extracted from the passage p_i , at training and inference time, we treat the additional knowledge about the image in the same manner as the questions and restrict the models such that answers are exclusively extracted from the passage. In this section, we further explain the approach employed to organize the input data that is fed to the models.

4.3.1 Span-based Answer Extraction Models

For the span-based answer extraction backbones (e.g. BERT and SpanBERT), we followed the input sequence configuration employed during the fine-tuning of these models: $[CLS] \text{ question } [SEP] \text{ passage } [SEP]$. It starts with a special token CLS , then the question and the passage, both followed by SEP tokens. The SEP marker between the question and passage sections serves to delimit the space within which the model can select the response span. In the case of these models, that space is the passage section. As illustrated in Figure 4.3 when integrating information related to images, we append this information to the question section, thereby ensuring that the model can utilize it to acquire knowledge without directly extracting the answer from it.



Figure 4.3: Span-based sequence encoding illustration.

There may be cases when the passages are too long, exceeding the model’s maximum token length. This is further aggravated when the input sequence is augmented with visual information. When this occurs, we create sub-sequences splitting the full passage into sub-passages, that are each appended to the sample’s question. The process is exemplified in Figure 4.4, each sub-sequence keeps the sample’s question and a passage division. The last sequence for the sample contains the question, the rest of the passage and padding tokens until the maximum length. To maintain the contexts as cohesive as possible, we adopt a sliding-window approach with a stride of 64 tokens between consecutive sub-sequences.

Now that we have the input sequence itself, we now need to provide the token indexes that represent the beginning and the end of the answer. The indexes are exclusively within the passage section since the models are supposed to learn how to indicate the correct answer within a given context.

When evaluating each sample, depending on the divisions, we may have several candidate answers. The model returns what it found to be the best answer for each division and we define the final prediction as the candidate answer span with greater confidence. Due to the breaking of the original samples’ sequences in sub-sequences, the correct answer span may not be on the sub-sequence passage. For this special case, as in [5], the model is trained to predict the index (0,0) as the *start* and *end* tokens, which represent the [CLS] token.

4.3.2 Span-based Decoding Models

Discriminative models, as described in the previous section, are used to predict the probability of a particular class label (in our case, to determine the probability of a token being the start or end of an answer) based on a given input. Generative models, such as T5, on the other hand, are used to generate new data that follows the distribution of the training samples (in our case, to generate a response to the given question).

T5 model employs a text-to-text approach, meaning the input sequence follows a prompt: *question : question context : passage </s>*. The sequence is composed of two central sections: the question and the passage. To incorporate image information, as shown in Figure 4.3, we adopt a similar approach as for answer extraction models and append the additional knowledge after the question. In addition to the input sequence, we also provide the models with the actual answer tokenized.

Given that this is a generative model, it does not provide a measure of confidence for the generated answer. As a result, it is not feasible to use the same approach of dividing

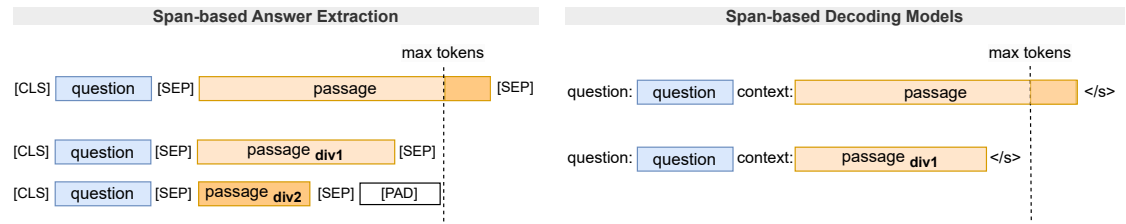


Figure 4.4: Illustration of sequence encoding: explanation of passage division for the extraction and decoding models when samples exceed the maximum number of tokens allowed by the models.

longer passages, as is done with answer extraction models, due to this inability to select an answer with a higher degree of certainty. During the evaluation process, we established the number of beams at four and adopted the generated answer as the final answer for the given sample.

EVALUATION

In this Chapter, we present a set of analyses and studies with which we aim to evaluate and discuss the intricacies and challenges of our dataset and provide insights into the best approaches for training models to solve our formulated task. Firstly, we elaborate on our training procedure, specifically the dataset split, metrics and protocol employed to fine-tune the models. Then, regarding our studies, we conducted a sequence context-encoding analysis, to gain deeper insights into the most practical combination of context-encoding settings for training the models, and also explore the impact of different context settings on the models' performance. Furthermore, we benchmark our dataset with various textual and visual context combinations. This analysis aims to provide a comprehensive understanding of the effectiveness of the existing models on our formulated task, comparing the performance employing the different textual and visual context approaches. Ultimately, we assess the quality of the dataset by exploring the difficulty levels of the Human Validated and Synthetic subsets.

5.1 Training Procedure

In this section, we explain how we train and evaluate the aforementioned backbone models. We explain how we split our dataset for training, introduce common metrics employed to evaluate span-based QA models and describe our training protocol.

5.1.1 Dataset Separation

NewsMQA is composed of two distinct subsets: Human Validated and Synthetic. Each subset was split into 70% for training, 15% for validation and 15% for testing. When evaluating the full NewsMQA, we combine the corresponding training, validation and test sets. Table 5.1 reveals the number of articles and questions each subset comprehends.

5.1.2 Metrics Analysis for Span-based QA

Measuring the answer correctness in QA systems is a challenging task, as a question can have multiple valid forms for the correct answer. Although in span-based QA the answer

Table 5.1: Number of unique articles and questions per topic in each subset. (N: News articles, Q: Questions)

	Human Validated subset						Synthetic subset					
	Train		Validation		Test		Train		Validation		Test	
	#N	#Q	#N	#Q	#N	#Q	#N	#Q	#N	#Q	#N	#Q
arts	423	484	110	112	112	114	779	854	179	179	166	166
books	419	511	120	125	98	101	794	879	186	187	182	183
food	254	295	68	70	71	74	391	436	101	101	98	100
movies	478	548	139	142	110	111	1471	1569	355	356	326	328
politics	359	392	80	83	86	87	962	1018	203	203	224	225
science	255	287	64	69	55	58	493	519	111	112	127	130
sports	312	350	65	67	68	71	909	987	218	222	188	190
technology	286	346	74	76	73	76	505	541	101	103	123	125
Total	2066	2309	481	492	484	492	5146	5519	1175	1182	1172	1182

must be confined within the given passage, there still may not be a single correct answer to a question. Following the literature on span-based QA [4, 5], we report two evaluation metrics: Exact Match (EM) and F1 score. Let’s analyze them in more detail.

As the name suggests, Exact Match is a binary metric that measures the proportion of predicted answers that exactly match the ground truth answer in a given dataset. Given a dataset sample, the EM score is 1 if the predicted answer from the model exactly matches the ground truth and 0 otherwise. This metric is especially strict, as Table 5.2 exemplifies, a deviation by even one character leads to a score of 0.

The F1 score is based on measuring the word overlap between the labelled and predicted answer. It ranges from 0 to 1, with higher values indicating better performance (examples on Table 5.2) and it is calculated as follows:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

F1 combines both precision and recall, which refers to the proportion of correct predictions and correct answers predicted by the model, respectively. Specifically, precision is calculated as the ratio between the number of shared words and the total number of words in the prediction, while recall is calculated as the ratio between the number of shared words and the total number of words in the ground truth. As we can verify through the examples in Table 5.2, this approach is less stringent than the Exact Match score and aligns better with a human judgment regarding the similarity of two answers. It is also important to state that for both metrics the text is normalized, that is, articles and punctuation are removed, and all text is changed to lowercase.

5.1.3 Training Settings and Protocol

In all our experiments, the models were fine-tuned in the NewsMQA dataset (full or a particular subset), for 20 epochs. The three models were fine-tuned with a batch size of 8 for both training and validation and a maximum input sequence of 512 tokens. We set a

Table 5.2: Comparison between Exact match and F1 score results for a set of examples.

Ground Thruth	Prediction	EM	F1
Myrlie Evers-Williams	Myrlie Evers-Williams	100.0	100.0
BLACK MARKET	Black Market	100.0	100.0
Golden State Warriors	Golden State	0.0	80.0
Blue Is the Warmest Color	Blue	0.0	40.0
1939	1940	0.0	0.0

learning rate of $1e-6$, $1e-4$ and a weight decay of 0.01 and $1e-2$ for the span-based models and T5, respectively. We determined the values for the learning rate and weight decay through an initial analysis where we tested several values. We chose to use a combination of values that allowed the models to effectively learn while also stagnating in terms of loss and evaluation metrics. Specifically, we tested different combinations of learning rates and weight decay values and selected the ones that produced the best results in terms of both model performance and convergence during training. In this way, we arrived at those values to train the models on our dataset.

During the training process, at the end of every epoch on the validation set, we report the EM and F1 metrics and save a checkpoint. In the end, we select the checkpoint model with the highest EM score. Fine-tuning on the full NewsMQA dataset took approximately 2 hours for the span-based models (BERT and SpanBERT) and approximately 5 hours for T5, on an NVIDIA A100 GPU with 40GB RAM.

The graphics in Figure 5.1 show the training loss and the evaluation metrics throughout the epochs. Overall, these metrics reports suggest that T5 models may initially perform well in training, but may struggle to generalize to new data. In contrast, extractive models may be more reliable in terms of generalization but may require more iterations to converge during training. While the T5 model exhibits fast learning and rapid stagnation during training, it can be more volatile in the validation set. In contrast, extractive models tend to reach stagnation quickly in the validation set.

5.2 Sequence Context-Encoding Analysis

To better understand which input settings are more suitable for training the span-based models, we carried out an extensive analysis of the impact two sample pre-processing settings have on the model’s performance. We explored:

- *Input Divisions* - The number of after-sample truncation divisions used to train the model. This first setting correlates with the approach to adopt when the passages are very long explained in Section 4.3. Within this setting, we study training the models only with the first division (*First*) or with all the created divisions (*All*)

5.2. SEQUENCE CONTEXT-ENCODING ANALYSIS

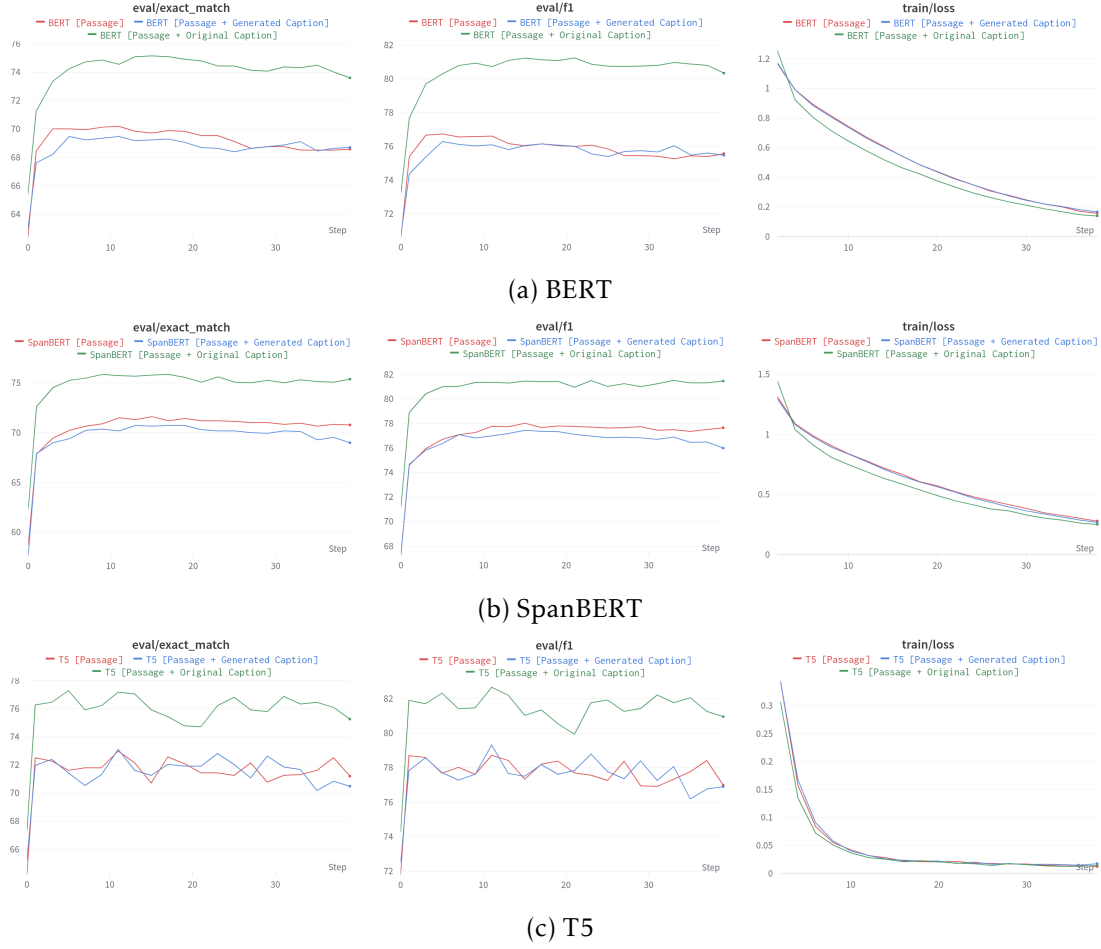


Figure 5.1: The image shows graphics reporting the evaluation metrics (exact match and F1 score) and training loss for each model, as well as the results for three different context settings: only the passage (shown in red), passage and generated captions (shown in blue), and passage and oracle captions (shown in green).

- *All Divisions contain Answer* - Whether or not all those divisions contain the answer. This second setting filters the obtained divisions, selecting only the ones that contain the answer (*Yes*) or all of them even if they do not contain the answer (*No*).

As explained in the previous Section 4.3, the sample division approach only makes sense to be studied on encoder models. Since the BERT and Span-BERT models have identical architectures and training procedures, we chose to carry out this study solely with BERT. Specifically, we use the BERT-Large¹ model fine-tuned on SQuAD available on Hugging Face.

Table 5.3 displays the results of this analysis, starting by identifying the dataset used to fine-tune the model: the Human Validated subset, Synthetic subset or the NewsMQA full dataset. For each of these sets and each combination of settings, it reveals the results after fine-tuning. The (*Training Set Size*) Column indicates the number of samples in the

¹available on huggingface.co/bert-large-uncased-whole-word-masking-finetuned-squad

Table 5.3: Impact of two sample pre-processing settings on **BERT** Model Performance. *Input Divisions*: number of after-samples truncation divisions, *All Divisions contain Answer*: whether or not all sample’s divisions contain the answer, *Training Set Size*: number of initial training samples, *#Divisions*: number of 1st and 2nd divisions generated for those settings combinations

Fine-tuning Subset	Input Divisions	All Divisions Contain Answer	Training Set Size	#Divisions 1st 2nd	EM	F1
Human Validated	First	Yes	2309	2309 0	55.69	69.96
	First	No		2309 0	55.69	69.96
	All	Yes		2309 0	55.69	69.94
	All	No		2309 14	56.10	70.01
Synthetic	First	Yes	5519	5519 0	75.21	78.91
	First	No		5519 0	75.21	78.91
	All	Yes		5519 1	75.97	79.64
	All	No		5519 19	76.48	79.80
NewsMQA	First	Yes	7828	7828 0	70.01	76.61
	First	No		7828 0	70.07	76.61
	All	Yes		7828 1	70.07	76.62
	All	No		7828 33	70.07	76.64

training set of that dataset. We also show the number of *1st* and *2nd Divisions* used to train the model assembled through each combination of settings. As we can see, the total final number of samples sometimes is greater than the number of samples in the training set. Such a case occurs when a sample is long enough to be divided, and more than one of those divisions contains the answer. Note that the correct answer may appear more than once within the context. Let’s use the fourth line of the table as an example: we employ all divisions even if they do not retain the answer. From the original 2309 samples of the training set of the Human Validated subset, there were a total of 2313 samples that contemplate those settings (2309 *1st divisions* +14 *2nd divisions*).

As we can notice, the answer is always present in the first division of the dataset samples. It is also possible to verify there are solely a few cases in which the context is so long that it has to be separated into two input sequences (33 in the full version of the dataset), and even rarer cases in which the answer is in the second division of the context, only once. Based on the results of our experiments, we can conclude that the differences between the various settings combinations are minimal, but there seems to be a slight preference for the fourth combination in each subset. It is worth noting that these settings have an impact on the final number of samples used to train the model. After careful consideration of the outcomes, we have decided to employ the combination of settings that provides the largest number of examples to train the models and that present greater results. This means we will be using all divisions, even those that do not contain the answer. In these cases, the model should predict that the answer is not within that passage and return the special token [CLS] as the answer (further explanation regarding this on Section 4.3).

Table 5.4: Exact Mach and F1 score performance (%) on **zero-shot** on the **Synthetic** test set, over different context settings. T: title, L: lead paragraph, C: image caption, P: passage, GAP: the difference in performance between T+L+C and P. * Oracle is the T5-Base model used to generate the synthetic dataset samples.

<i>Zero-shot on Synthetic subset</i>						
	T	L	T+L	T+L+C	P	GAP ($\approx$)
Oracle	14.13 / 16.69	49.15 / 51.25	59.31 / 61.49	98.95 / 99.21	68.78 / 73.95	31.2 / 25.7
SpanBERT	14.13 / 16.32	41.79 / 46.65	51.61 / 56.42	77.41 / 84.13	60.15 / 67.62	17.3 / 16.5
BERT	14.81 / 17.18	44.92 / 48.77	55.41 / 59.39	83.76 / 88.14	67.34 / 72.59	16.4 / 15.6
T5	15.31 / 18.49	46.36 / 49.37	56.35 / 59.27	84.52 / 87.65	67.01 / 72.23	17.5 / 15.1

5.3 Studying Context Settings

With this analysis, we aim to explore the impact of different context settings combinations on the models’ performance. We compare the models’ behaviour when having access to the most suitable context to predict the answers, *i.e.* having access to the context used to generate the question and answer (*i.e.*, privilege/oracle information), and solely having access to the task formulated context, which is the passage. The privileged information is the contextual elements utilised to generate the samples such as the articles’ title, lead paragraphs and the original images caption. Note that the answer is always available within the passage.

For this study, we only operated with the Synthetic subset as this is the only one whose all samples have never changed since the moment they were generated and so the combination *title + lead paragraph + image caption* is the most promising. Regarding the models, we first define the upper bound with the model used to generate the questions and answers for the dataset, a T5-Base model that we define as the *Oracle* model. The other models we study are the models presented in the last Chapter in Section 4.1. All these evaluations were performed in a zero-shot setting, *i.e.* with no fine-tuning. For each model we evaluate 5 different context inputs: (T) only the article’s title, (L) only the lead paragraph, (T+L) the article’s title and lead paragraph, (T+L+C) the article’s title, lead paragraph and image caption, and finally (P) only the passage.

Table 5.4 reveals the results obtained. In the first slot of the table is the T5-Base model’s results, which as expected have the highest zero-shot scores among all models in the majority of the settings. Moreover, the results of this model also reflect the influence each component had during the question-and-answer generation. However, it is also important to notice that the oracle model T5 may also have introduced some bias. The two settings that seem to have the major impact are the lead paragraph, where it presents enough information to respond to 49.15% of the samples, and the image caption, as from using the title and lead paragraph without the caption and with the caption there is a difference of almost 40%. These results are directly correlated to the analysis carried out in Table 3.8, where we indicated the number of samples whose answer to the question is

Table 5.5: Exact match / F1 score performance (%) results in the **NewsMQA full dataset** adding visual context. * The original caption is privileged information and is analysed to establish an upper-bound.

<i>NewsMQA full dataset</i>						
	Only Passage	with Tags		with Caption		Captions Gap
		Top 5	Top 10	Generated	Original*	
<i>Zero-shot</i>						
SpanBERT	57.17 / 66.88	57.29 / 66.68	57.23 / 66.77	56.15 / 65.89	59.31 / 69.62	3.16 / 3.73
BERT	63.38 / 71.37	63.32 / 70.96	62.90 / 70.67	61.83 / 69.56	65.17 / 73.03	3.34 / 3.47
T5	63.44 / 71.60	62.54 / 70.78	63.26 / 71.09	63.74 / 71.11	67.44 / 74.56	3.70 / 3.45
<i>Fine-tuned</i>						
SpanBERT	69.77 / 76.29	70.55 / 76.52	70.43 / 76.72	68.93 / 75.39	74.37 / 80.35	5.44 / 4.96
BERT	70.07 / 76.64	68.82 / 75.56	69.24 / 75.89	68.40 / 74.84	75.98 / 81.61	7.58 / 6.77
T5	73.95 / 79.85	72.88 / 78.73	72.10 / 78.34	74.19 / 79.89	77.71 / 83.14	3.52 / 3.25

solely within one of the components used to generate the candidate pairs.

Across all models, a consistent 25% drop in performance is observed when the image caption is not available, evidencing that to solve the task, models are required to use the information provided by the image to reach accurate answers. Lastly, we analyse the results obtained when evaluating the models with the passage as the only context. We can observe that the performance is greater than in the title + lead combination and that T5-Large performs better than BERT/Span-BERT, in all combinations and metrics. Moreover, we can verify the performance gap between using the best combination possible (title + lead + caption) and the one required by the NewsMQA task (just the passage). Notwithstanding, we can see that there is a significant margin for improvement.

5.4 Benchmarking Visual Context

In this Section, we benchmark the outcome of enriching each sample context with visual information, beyond the baseline of just having access to the passage. Specifically, we discuss the impact of incorporating the image information discussed in Section 4.2 and analyze its effect on the overall behaviour of the models.

We experimented with incorporating the top 5 and top 10 tags achieved using the object detection model, the COCO-like generated captions and the original caption of the news articles. Recalling, all this additional information is joined alongside the question, which means that span-based models can gain knowledge from these captions but not use it to define the start and end tokens of the answer (this is regarding the BERT and SpanBERT models). It's crucial to keep in mind that the original news captions are considered privileged information since they were taken into account during the question generation process and bridge the visual and textual components. Nevertheless, this information remains highly relevant to this analysis as it contains highly valuable insights that aid in identifying the answer.

	Question: What is the name of Anna Maria's Italian Restaurant? Groundtruth: Anna Maria's Italian Restaurant		Question: What is the final installment of a trilogy of Bowie documentaries from the English filmmaker Francis Whately? Groundtruth: DAVID BOWIE: FINDING FAME
Passage: [...] THESE days, when waiters bring out a cake with candles and sing Happy Birthday to a customer, most other diners barely lift their heads. Its hard to get anyone to sing along, even those at the birthday persons own table. Not so at Anna Maria's Italian Restaurant in Larchmont. [...] Answer: Larchmont	Passage + Generated Caption: people sitting at tables in a restaurant Answer: Larchmont	Passage: [...] DAVID BOWIE: FINDING FAME 9 p.m. on Showtime. I think theres a lot of searching to find the individual within oneself, [...] Finding Fame, the final installment of a trilogy of Bowie documentaries from the English filmmaker Francis Whately, explores a period of Bowies career [...] Answer: Finding Fame	Passage + Generated Caption: a woman in a green shirt with a flower in her hair Answer: Finding Fame
 Passage: [...] The author of "Gone Girl" says it's hard to get over something first read at age 12: "I will probably be clutching 'Flowers in the Attic' in my gnarled hands on my deathbed." What books are currently on your night [...] Answer: Flowers in the Attic	Passage + Oracle Caption: OUTSIDE LOOKING IN Anna Maria's Italian Restaurant, in Larchmont, which seats about 40. Answer: Anna Maria's Italian Restaurant	 Passage: [...] It is not too early to ask whether President Obama's robust supply of political capital has begun to dissipate. Predecessors with bigger victory margins have lost it this quickly. Few remember the early travails of Franklin Roosevelt after he swept 57 percent of the vote and all but six states against Herbert Hoover in 1932. [...] Answer: Obama	Passage + Oracle Caption: David Bowie in "David Bowie: Finding Fame." Answer: Finding Fame
	Question: What is the name of the book that Gillian Flynn wrote? Groundtruth: Gone Girl		Question: What is the name of the president who arrived in Washington with lots of political capital? Groundtruth: Obama
	Passage + Generated Caption: a woman in a red dress with a smile on her face Answer: Flowers in the Attic		Passage + Generated Caption: a man in a uniform is holding a gun Answer: Obama
	Passage + Oracle Caption: Gillian Flynn Answer: Gone Girl		Passage + Oracle Caption: Candidate Mojo vs. Presidential Mojo Franklin Roosevelt, pictured a few days before Election Day in 1932, was in political trouble by February, 1933. Answer: Franklin Roosevelt

Figure 5.2: Example of the T5 model’s (fine-tuned on NewsMQA full dataset) answers employing different context settings: solely the passage, passage and generated caption, and passage and the oracle caption. With these examples, we can notice the influence captions can have on the response the model offers.

Table 5.5 reports the results of the proposed multimodal context encoding scheme in both a zero-shot and fine-tune on the full NewsMQA dataset. In general, the addition of both tags and generated captions does not appear to have a positive impact. Only with the fine-tuned T5 backbone, does the performance increase slightly by employing the generated captions. As discussed earlier, in Section 4.2, when we look more closely at these captions and tags, we realize they mostly describe the scenario in a very general way (e.g. there is a person, beach, certain vehicles etc.), not identifying the entities present or making a connection to the story to be told in the news.

Given that when models have access to the original news image captions, performance improves significantly, it suggests that the tags and general captions fail to provide enough context information, and sustains that entities play a crucial role in the question answering. Figure 5.2 displays several examples that demonstrate the influence of captions on the model’s answers. The left two examples highlight the oracle captions as they are the only information that helps the model arrive at the correct answer. The top right example demonstrates a failure case where the model does not arrive at the answer using any of the components. Lastly, the bottom right example shows the case where the original caption does not direct the model in the right direction, inducing it into error

Table 5.6: EM / F1 Performance (%) on the Human Validated test set, with different fine-tuning schemes and two different input settings.

Fine-tuning set		Test on Human Validated Subset		
		Passage	Passage + Generated Caption	Gap
SpanBERT	Human Validated	56.10 / 67.70	57.70 / 69.20	1.60 / 1.50
	Synthetic	55.48 / 66.12	55.69 / 66.17	0.21 / 0.05
	NewsMQA	58.74 / 69.33	57.93 / 68.01	0.81 / 1.32
BERT	Human Validated	56.91 / 70.51	56.30 / 70.61	0.61 / 0.10
	Synthetic	54.47 / 65.97	56.30 / 65.91	1.83 / 0.06
	NewsMQA	58.54 / 71.03	56.09 / 68.15	2.45 / 2.88
T5	Human Validated	59.35 / 71.84	55.49 / 69.84	3.86 / 2.00
	Synthetic	59.15 / 66.91	56.50 / 65.12	2.65 / 1.79
	NewsMQA	59.96 / 71.97	60.37 / 71.80	0.41 / 0.17

and leading it to the wrong answer.

Overall, this experiment supports our observation that the connection between text and images in news is fundamentally different, bringing novel challenges. As observed, these are embedded in our proposed NewsMQA task and dataset.

5.5 Assessing Dataset Quality

The aim of this Section is to conduct an analysis of the quality of our dataset by examining and emphasizing the impact and difficulty of the different subsets. More specifically, our focus is on evaluating whether the method used to generate the Synthetic subset is useful in the training of our models.

To accomplish this, we fine-tuned the backbone models individually on each subset (Human Validated, Synthetic, and the full NewsMQA) and evaluated them solely on the Human Validated test set. In addition, we analyze these models with those training and evaluation setups employing two input settings that we studied previously: one that provides only the passage and one that provides both the passage and the generated caption to the model. We used the result of training and evaluating each model on the Human Validated subset as the baseline point for comparing the impact of training with each subset.

Table 5.6 shows the outcomes. Among the different backbones, and regardless of the employed input setting, T5 is consistently the best-performing one. Additionally, we once again confirm the generated captions do not help to enhance the models' outcome. As we can verify, when any of the backbone models are fine-tuned with both subsets, *i.e.* with the full version of the NewsMQA dataset, the performance is consistently higher in terms of exact match, and either maintained or exceeded in terms of f1 score compared to when fine-tuned individually on the Human Validated or Synthetic subsets. This confirms the synthetic sample generation strategy we adopted is effective and leads to performance improvement.

CONCLUSIONS

In this thesis, we introduced the task of open-ended News Multimodal Question Answering, in which given a question about a news passage and an image, the goal is to extract the answer excerpt. To support the task, we introduced NewsMQA, a novel multimodal QA dataset that encompasses the multimodal facets of news and improves the quality vs. scale trade-off with a mix of Human Validated and Synthetic samples.

Through a comprehensive study of the existing question-answering research area, we reviewed the tasks and datasets and explored the current models employed to solve these problems. NewsMQA departs from previous datasets, by truly capturing the full multimodality spectrum inherent to the news domain. We conducted an extensive analysis in which we identified both the unique characteristics and intrinsic research challenges of the task. To solve our task, we proposed and explored providing different input formatting that incorporate visual information coming from the image. In particular, we explored the use of object detection and captions as alternatives to augment the models' input with image-related information. Our approach involved including a list of objects detected within the picture, as well as including automatically extracted image captions into each backbone input sequence. We also experimented with using the original article captions available in the NYTimes800k dataset provided by journalists.

We implemented our approach in existing Transformers models that had already demonstrated encouraging results in question-answering tasks, BERT [13], SpanBERT [14] and T5 [16]. To evaluate the effectiveness of this approach, we benchmarked the NewsMQA dataset and conducted a set of studies and analyses to assess the challenges it presents. The results of our experiments confirmed both the quality of the NewsMQA dataset as well as the utility of the synthetically generated samples. They also provided insight into the strengths and limitations of the proposed models and highlight potential areas for future improvement.

Finally, we have submitted a paper that is currently under review for ACL 2023. We believe that our work will contribute to the development of more effective and comprehensive question-answering models in natural language processing and help to address the challenges posed by multimodal information sources.

6.1 Limitations and Future Work

Despite the good overview of our dataset and the results of the models in our task, we believe that there is still room for improvement. Annotating samples for the multimodal QA task in the news domain is not only a labour-intensive but also cognitively demanding task. Questions and answers require deep knowledge of each news story, which is challenging for such a vast domain. This naturally imposed a set of limitations on our work. Specifically, to reach a significant scale, we resorted to synthetic sample generation. As evidenced by our experiments, both the quality and difficulty of synthetic samples are lower when compared to human-annotated samples.

In terms of benchmarking, while we followed the ideas of state-of-the-art approaches in close-ended multimodal QA that leverage large closed-source pre-trained models [83] (GPT-3), it would be interesting to evaluate state-of-the-art multimodal foundational backbones such as FLAVA [85]. Since this requires adapting these models for span-based extraction, we leave this for future work. Based on the methods currently available and the results obtained in our benchmark, future work should also focus on studying visual context-enriching mechanisms, towards developing multimodal question-answering models that can effectively utilize both image and passage modalities to solve tasks such as ours.

BIBLIOGRAPHY

- [1] N. S. Dahmen, K. C. Miller, and B. Walth. “The Power of Images? Visual Journalists’ Assessment of the Impact of Imagery”. In: *Visual Communication Quarterly* 28.1 (2021), pp. 34–44. DOI: [10.1080/15551393.2020.1862662](https://doi.org/10.1080/15551393.2020.1862662). eprint: <https://doi.org/10.1080/15551393.2020.1862662>. URL: <https://doi.org/10.1080/15551393.2020.1862662> (cit. on pp. 1, 7).
- [2] P. Tolmie et al. “Supporting the Use of User Generated Content in Journalistic Practice”. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI ’17. Denver, Colorado, USA: Association for Computing Machinery, 2017, pp. 3632–3644. ISBN: 9781450346559. DOI: [10.1145/3025453.3025892](https://doi.org/10.1145/3025453.3025892). URL: <https://doi.org/10.1145/3025453.3025892> (cit. on pp. 1, 7, 33).
- [3] J. Schwartz. *Study Gauges Plastic Levels in Oceans*. Dec. 2014. URL: <https://www.nytimes.com/2014/12/11/science/new-research-quantifies-the-oceans-plastic-problem.html> (cit. on p. 2).
- [4] P. Rajpurkar et al. “SQuAD: 100,000+ Questions for Machine Comprehension of Text”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2383–2392. DOI: [10.18653/v1/D16-1264](https://doi.org/10.18653/v1/D16-1264). URL: <https://aclanthology.org/D16-1264> (cit. on pp. 2, 3, 20, 27, 33, 49, 55).
- [5] P. Rajpurkar, R. Jia, and P. Liang. “Know What You Don’t Know: Unanswerable Questions for SQuAD”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 784–789. DOI: [10.18653/v1/P18-2124](https://doi.org/10.18653/v1/P18-2124). URL: <https://aclanthology.org/P18-2124> (cit. on pp. 2, 52, 55).
- [6] A. Agrawal et al. “VQA: Visual Question Answering”. In: *Int. J. Comput. Vision* 123.1 (May 2017), pp. 4–31. ISSN: 0920-5691. DOI: [10.1007/s11263-016-0966-6](https://doi.org/10.1007/s11263-016-0966-6). URL: <https://doi.org/10.1007/s11263-016-0966-6> (cit. on pp. 2, 3, 20, 21, 27, 28, 35).

- [7] M. Malinowski and M. Fritz. “A Multi-World Approach to Question Answering about Real-World Scenes Based on Uncertain Input”. In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*. NIPS’14. Montreal, Canada: MIT Press, 2014, pp. 1682–1690 (cit. on pp. 2, 19–21).
- [8] T. Du et al. “CocoQa: Question Answering for Coding Conventions Over Knowledge Graphs”. In: *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. 2019, pp. 1086–1089. DOI: [10.1109/ASE.2019.00108](https://doi.org/10.1109/ASE.2019.00108) (cit. on pp. 2, 19–21).
- [9] K. Marino et al. “OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 3190–3199. DOI: [10.1109/CVPR.2019.00331](https://doi.org/10.1109/CVPR.2019.00331) (cit. on pp. 3, 20, 21, 27, 28, 47).
- [10] C. Alberti et al. “Synthetic QA Corpora Generation with Roundtrip Consistency”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 6168–6173. DOI: [10.18653/v1/P19-1620](https://doi.org/10.18653/v1/P19-1620). URL: <https://aclanthology.org/P19-1620> (cit. on pp. 3, 38).
- [11] S. Changpinyo et al. “All You May Need for VQA are Image Captions”. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 1947–1963. DOI: [10.18653/v1/2022.naacl-main.142](https://doi.org/10.18653/v1/2022.naacl-main.142). URL: <https://aclanthology.org/2022.naacl-main.142> (cit. on pp. 3, 38).
- [12] S. Changpinyo et al. “All You May Need for VQA are Image Captions”. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 1947–1963. DOI: [10.18653/v1/2022.naacl-main.142](https://doi.org/10.18653/v1/2022.naacl-main.142). URL: <https://aclanthology.org/2022.naacl-main.142> (cit. on p. 3).
- [13] J. Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423). URL: <https://aclanthology.org/N19-1423> (cit. on pp. 3, 10, 11, 23, 24, 48, 63).
- [14] M. Joshi et al. “SpanBERT: Improving Pre-training by Representing and Predicting Spans”. In: *Transactions of the Association for Computational Linguistics* 8 (2020),

- pp. 64–77. DOI: [10.1162/tac1_a_00300](https://doi.org/10.1162/tac1_a_00300). URL: <https://aclanthology.org/2020.tac1-1.5> (cit. on pp. 3, 24, 25, 48, 63).
- [15] L. H. Li et al. “VisualBERT: A Simple and Performant Baseline for Vision and Language”. In: *CoRR* abs/1908.03557 (2019). arXiv: [1908.03557](https://arxiv.org/abs/1908.03557). URL: <http://arxiv.org/abs/1908.03557> (cit. on pp. 3, 14, 15, 25, 26, 47).
- [16] C. Raffel et al. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”. In: *CoRR* abs/1910.10683 (2019). arXiv: [1910.10683](https://arxiv.org/abs/1910.10683). URL: <http://arxiv.org/abs/1910.10683> (cit. on pp. 3, 11, 12, 16, 33, 48, 63).
- [17] A. Vaswani et al. “Attention is All You Need”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS’17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 6000–6010. ISBN: 9781510860964 (cit. on pp. 3, 9, 10, 13, 14, 16).
- [18] H. Tan and M. Bansal. “LXMERT: Learning Cross-Modality Encoder Representations from Transformers”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 5100–5111. DOI: [10.18653/v1/D19-1514](https://doi.org/10.18653/v1/D19-1514). URL: <https://aclanthology.org/D19-1514> (cit. on pp. 3, 14, 15, 23, 25, 27, 47).
- [19] T. Olesen. “More Than Words?: A Test of the Effect of Emotionally Charged Photographs”. In: *Nordicom Review* 39.2 (2018), pp. 35–48. DOI: [doi:10.2478/nor-2018-0010](https://doi.org/10.2478/nor-2018-0010). URL: <https://doi.org/10.2478/nor-2018-0010> (cit. on p. 7).
- [20] Z. C. Lipton. “A Critical Review of Recurrent Neural Networks for Sequence Learning”. In: *CoRR* abs/1506.00019 (2015). arXiv: [1506.00019](https://arxiv.org/abs/1506.00019). URL: <http://arxiv.org/abs/1506.00019> (cit. on p. 8).
- [21] J. Ferreira et al. “Faster than LASER - Towards Stream Reasoning with Deep Neural Networks”. In: *CoRR* abs/2106.08457 (2021). arXiv: [2106.08457](https://arxiv.org/abs/2106.08457). URL: <https://arxiv.org/abs/2106.08457> (cit. on pp. 8, 12).
- [22] R. Ferreira et al. “Deep Neural Networks for Approximating Stream Reasoning with C-SPARQL”. In: *EPIA*. 2021 (cit. on pp. 8, 12).
- [23] M. E. Peters et al. “Deep Contextualized Word Representations”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 2227–2237. DOI: [10.18653/v1/N18-1202](https://doi.org/10.18653/v1/N18-1202). URL: <https://aclanthology.org/N18-1202> (cit. on p. 9).

- [24] L. Zhuang et al. “A Robustly Optimized BERT Pre-training Approach with Post-training”. English. In: *Proceedings of the 20th Chinese National Conference on Computational Linguistics*. Huhhot, China: Chinese Information Processing Society of China, Aug. 2021, pp. 1218–1227. URL: <https://aclanthology.org/2021.ccl-1.108> (cit. on pp. 11, 17).
- [25] Z. Yang et al. “XLNet: Generalized Autoregressive Pretraining for Language Understanding”. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019 (cit. on p. 11).
- [26] S. Albawi, T. A. Mohammed, and S. Al-Zawi. “Understanding of a convolutional neural network”. In: *2017 International Conference on Engineering and Technology (ICET)*. 2017, pp. 1–6. DOI: [10.1109/ICEngTechno1.2017.8308186](https://doi.org/10.1109/ICEngTechno1.2017.8308186) (cit. on pp. 12, 13).
- [27] Z. Wang, W. Yan, and T. Oates. “Time series classification from scratch with deep neural networks: A strong baseline”. In: *2017 International Joint Conference on Neural Networks (IJCNN)*. 2017, pp. 1578–1585. DOI: [10.1109/IJCNN.2017.7966039](https://doi.org/10.1109/IJCNN.2017.7966039) (cit. on pp. 12, 13).
- [28] K. Simonyan and A. Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *CoRR* abs/1409.1556 (2015) (cit. on p. 13).
- [29] A. Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *CoRR* abs/2010.11929 (2020). arXiv: [2010.11929](https://arxiv.org/abs/2010.11929). URL: <https://arxiv.org/abs/2010.11929> (cit. on pp. 13, 14, 49, 50).
- [30] J. Lu et al. “ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks”. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019 (cit. on pp. 14, 15, 47).
- [31] Y. Wu et al. “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation”. In: *CoRR* abs/1609.08144 (2016). arXiv: [1609.08144](https://arxiv.org/abs/1609.08144). URL: <http://arxiv.org/abs/1609.08144> (cit. on p. 14).
- [32] E. Bugliarello et al. “Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs”. In: *Transactions of the Association for Computational Linguistics* 9 (Sept. 2021), pp. 978–994. ISSN: 2307-387X. DOI: [10.1162/tac1_a_00408](https://doi.org/10.1162/tac1_a_00408). eprint: https://direct.mit.edu/tac1/article-pdf/doi/10.1162/tac1_a_00408/1963734/tac1_a_00408.pdf. URL: https://doi.org/10.1162/tac1%5C_a%5C_00408 (cit. on p. 15).
- [33] J. Cho et al. “Unifying Vision-and-Language Tasks via Text Generation”. In: *CoRR* abs/2102.02779 (2021). arXiv: [2102.02779](https://arxiv.org/abs/2102.02779). URL: <https://arxiv.org/abs/2102.02779> (cit. on p. 16).

- [34] M. Lewis et al. “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 7871–7880. DOI: [10.18653/v1/2020.acl-main.703](https://doi.org/10.18653/v1/2020.acl-main.703). URL: <https://aclanthology.org/2020.acl-main.703> (cit. on p. 16).
- [35] A. Tran, A. Mathews, and L. Xie. “Transform and Tell: Entity-Aware News Image Captioning”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020, pp. 13032–13042. DOI: [10.1109/CVPR42600.2020.01305](https://doi.org/10.1109/CVPR42600.2020.01305) (cit. on pp. 17, 31, 34).
- [36] K. Zhang et al. “Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks”. In: *IEEE Signal Processing Letters* 23.10 (2016), pp. 1499–1503. DOI: [10.1109/LSP.2016.2603342](https://doi.org/10.1109/LSP.2016.2603342) (cit. on p. 17).
- [37] J. Redmon and A. Farhadi. “YOLOv3: An Incremental Improvement”. In: *CoRR* abs/1804.02767 (2018). arXiv: [1804.02767](https://arxiv.org/abs/1804.02767). URL: <http://arxiv.org/abs/1804.02767> (cit. on p. 17).
- [38] A. F. Biten et al. “Good News, Everyone! Context Driven Entity-Aware Captioning for News Images”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 12458–12467. DOI: [10.1109/CVPR.2019.01275](https://doi.org/10.1109/CVPR.2019.01275) (cit. on p. 18).
- [39] F. Liu et al. “Visual News: Benchmark and Challenges in News Image Captioning”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 6761–6771. DOI: [10.18653/v1/2021.emnlp-main.542](https://doi.org/10.18653/v1/2021.emnlp-main.542). URL: <https://aclanthology.org/2021.emnlp-main.542> (cit. on p. 18).
- [40] G. Marcelino et al. “Assisting News Media Editors with Cohesive Visual Storylines”. In: *Proceedings of the 29th ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 3257–3265. ISBN: 9781450386517. URL: <https://doi.org/10.1145/3474085.3475476> (cit. on p. 18).
- [41] A. Rogers, M. Gardner, and I. Augenstein. “Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension”. In: *arXiv preprint arXiv:2107.12708* (2021) (cit. on p. 19).
- [42] F. Hill et al. “The Goldilocks Principle: Reading Children’s Books with Explicit Memory Representations”. In: *CoRR* abs/1511.02301 (2016) (cit. on pp. 19, 20).
- [43] O. Bajgar, R. Kadlec, and J. Kleindienst. “Embracing data abundance: BookTest Dataset for Reading Comprehension”. In: *CoRR* abs/1610.00956 (2016). arXiv: [1610.00956](https://arxiv.org/abs/1610.00956). URL: <http://arxiv.org/abs/1610.00956> (cit. on pp. 19, 20).

- [44] K. M. Hermann et al. “Teaching Machines to Read and Comprehend”. In: *CoRR* abs/1506.03340 (2015). arXiv: [1506.03340](https://arxiv.org/abs/1506.03340). URL: <http://arxiv.org/abs/1506.03340> (cit. on pp. 19, 20).
- [45] M. Richardson, C. J. Burges, and E. Renshaw. “MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text”. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, Oct. 2013, pp. 193–203. URL: <https://aclanthology.org/D13-1020> (cit. on pp. 19, 20).
- [46] A. Rogers et al. “Getting Closer to AI Complete Question Answering: A Set of Prerequisite Real Tasks”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.05 (2020), pp. 8722–8731. DOI: [10.1609/aaai.v34i05.6398](https://doi.org/10.1609/aaai.v34i05.6398). URL: <https://ojs.aaai.org/index.php/AAAI/article/view/6398> (cit. on pp. 19, 20).
- [47] L. Yu et al. “Visual Madlibs: Fill in the blank Image Generation and Question Answering”. In: *CoRR* abs/1506.00278 (2015). arXiv: [1506.00278](https://arxiv.org/abs/1506.00278). URL: <http://arxiv.org/abs/1506.00278> (cit. on pp. 19–21).
- [48] M. Tapaswi et al. “MovieQA: Understanding Stories in Movies through Question-Answering”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 4631–4640. DOI: [10.1109/CVPR.2016.501](https://doi.org/10.1109/CVPR.2016.501) (cit. on pp. 19, 20).
- [49] A. D. Lelkes, V. Q. Tran, and C. Yu. “Quiz-Style Question Generation for News Stories”. In: *Proceedings of the Web Conference 2021*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 2501–2511. ISBN: 9781450383127. URL: <https://doi.org/10.1145/3442381.3449892> (cit. on pp. 20, 23, 26).
- [50] A. Trischler et al. “NewsQA: A Machine Comprehension Dataset”. In: *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 191–200. DOI: [10.18653/v1/W17-2623](https://doi.org/10.18653/v1/W17-2623). URL: <https://aclanthology.org/W17-2623> (cit. on pp. 20, 23, 30).
- [51] J. Wang et al. “Answering Event-Related Questions over Long-Term News Article Archives”. In: *Advances in Information Retrieval*. Ed. by J. M. Jose et al. Cham: Springer International Publishing, 2020, pp. 774–789. ISBN: 978-3-030-45439-5 (cit. on pp. 20, 22).
- [52] A. Soleimani, C. Monz, and M. Worring. “NLQuAD: A Non-Factoid Long Question Answering Data Set”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, Apr. 2021, pp. 1245–1255. DOI: [10.18653/v1/2021.eacl-main.106](https://doi.org/10.18653/v1/2021.eacl-main.106). URL: <https://aclanthology.org/2021.eacl-main.106> (cit. on pp. 20, 23).

-
- [53] J. Wang, A. Jatowt, and M. Yoshikawa. “ArchivalQA: A Large-scale Benchmark Dataset for Open Domain Question Answering over Archival News Collections”. In: *CoRR* abs/2109.03438 (2021). arXiv: [2109.03438](https://arxiv.org/abs/2109.03438). URL: <https://arxiv.org/abs/2109.03438> (cit. on pp. 20, 23, 26, 30).
- [54] R. Krishna et al. “Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations”. In: *Int. J. Comput. Vision* 123.1 (May 2017), pp. 32–73. ISSN: 0920-5691. DOI: [10.1007/s11263-016-0981-7](https://doi.org/10.1007/s11263-016-0981-7). URL: <https://doi.org/10.1007/s11263-016-0981-7> (cit. on pp. 20, 35).
- [55] J. Johnson et al. “CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 1988–1997. DOI: [10.1109/CVPR.2017.215](https://doi.org/10.1109/CVPR.2017.215) (cit. on pp. 20, 21).
- [56] P. Wang et al. “FVQA: Fact-Based Visual Question Answering”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.10 (2018), pp. 2413–2427. DOI: [10.1109/TPAMI.2017.2754246](https://doi.org/10.1109/TPAMI.2017.2754246) (cit. on pp. 20, 21).
- [57] S. Shah et al. “KVQA: Knowledge-Aware Visual Question Answering”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (July 2019), pp. 8876–8884. DOI: [10.1609/aaai.v33i01.33018876](https://doi.org/10.1609/aaai.v33i01.33018876). URL: <https://ojs.aaai.org/index.php/AAAI/article/view/4915> (cit. on pp. 20, 21).
- [58] D. A. Hudson and C. D. Manning. “GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 6693–6702. DOI: [10.1109/CVPR.2019.00686](https://doi.org/10.1109/CVPR.2019.00686) (cit. on pp. 20, 21).
- [59] W. Chen et al. “HybridQA: A Dataset of Multi-Hop Question Answering over Tabular and Textual Data”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1026–1036. DOI: [10.18653/v1/2020.findings-emnlp.91](https://doi.org/10.18653/v1/2020.findings-emnlp.91). URL: <https://aclanthology.org/2020.findings-emnlp.91> (cit. on p. 20).
- [60] A. Talmor et al. “MultiModalQA: Complex Question Answering over Text, Tables and Images”. In: *arXiv preprint arXiv:2104.06039* (2021) (cit. on p. 20).
- [61] D. Hannan, A. Jain, and M. Bansal. “ManyModalQA: Modality Disambiguation and QA over Diverse Inputs”. In: *CoRR* abs/2001.08034 (2020). arXiv: [2001.08034](https://arxiv.org/abs/2001.08034). URL: <https://arxiv.org/abs/2001.08034> (cit. on p. 20).
- [62] S. Sheng, L. Van Gool, and M.-F. Moens. “A Dataset for Multimodal Question Answering in the Cultural Heritage Domain”. In: *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 10–17. URL: <https://aclanthology.org/W16-4003> (cit. on pp. 20, 22).

- [63] R. G. Reddy et al. “MuMuQA: Multimedia Multi-Hop News Question Answering via Cross-Media Knowledge Extraction and Grounding”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.10 (June 2022), pp. 11200–11208. DOI: [10.1609/aaai.v36i10.21370](https://ojs.aaai.org/index.php/AAAI/article/view/21370). URL: <https://ojs.aaai.org/index.php/AAAI/article/view/21370> (cit. on pp. 20, 22, 30, 48).
- [64] T.-Y. Lin et al. “Microsoft COCO: Common Objects in Context”. In: *ECCV. European Conference on Computer Vision*, Sept. 2014. URL: <https://www.microsoft.com/en-us/research/publication/microsoft-coco-common-objects-in-context/> (cit. on pp. 21, 27, 29, 50).
- [65] K. Kafle and C. Kanan. “Visual question answering: Datasets, algorithms, and future challenges”. In: *Computer Vision and Image Understanding* 163 (2017). Language in Vision, pp. 3–20. ISSN: 1077-3142. DOI: <https://doi.org/10.1016/j.cviu.2017.06.005>. URL: <https://www.sciencedirect.com/science/article/pii/S1077314217301170> (cit. on p. 21).
- [66] M. J. Seo et al. “Bidirectional Attention Flow for Machine Comprehension”. In: *CoRR* abs/1611.01603 (2016). arXiv: [1611.01603](https://arxiv.org/abs/1611.01603). URL: <http://arxiv.org/abs/1611.01603> (cit. on p. 23).
- [67] C. Qu et al. “Passage Retrieval for Outside-Knowledge Visual Question Answering”. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 1753–1757. ISBN: 9781450380379. URL: <https://doi.org/10.1145/3404835.3462987> (cit. on pp. 23, 29).
- [68] P. Xiong et al. “Visual Query Answering by Entity-Attribute Graph Matching and Reasoning”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 8349–8358. DOI: [10.1109/CVPR.2019.00855](https://doi.org/10.1109/CVPR.2019.00855) (cit. on p. 26).
- [69] F. Gardères et al. “ConceptBert: Concept-Aware Representation for Visual Question Answering”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 489–498. DOI: [10.18653/v1/2020.findings-emnlp.44](https://doi.org/10.18653/v1/2020.findings-emnlp.44). URL: <https://aclanthology.org/2020.findings-emnlp.44> (cit. on p. 26).
- [70] M. Ziaeeefard and F. Lecue. “Towards Knowledge-Augmented Visual Question Answering”. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 1863–1873. DOI: [10.18653/v1/2020.coling-main.169](https://doi.org/10.18653/v1/2020.coling-main.169). URL: <https://aclanthology.org/2020.coling-main.169> (cit. on p. 26).

- [71] K. Marino et al. “KRISP: Integrating Implicit and Symbolic Knowledge for Open-Domain Knowledge-Based VQA”. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 14106–14116. DOI: [10.1109/CVPR46437.2021.01389](https://doi.org/10.1109/CVPR46437.2021.01389) (cit. on p. 26).
- [72] V. Shevchenko et al. “Reasoning over Vision and Language: Exploring the Benefits of Supplemental Knowledge”. In: *Proceedings of the Third Workshop on Beyond Vision and LANGUAGE: inTEgrating Real-world kNowledge (LANTERN)*. Kyiv, Ukraine: Association for Computational Linguistics, Apr. 2021, pp. 1–18. URL: <https://aclanthology.org/2021.lantern-1.1> (cit. on p. 26).
- [73] B. B. Cambazoglu et al. “A Review of Public Datasets in Question Answering Research”. In: *SIGIR Forum* 54.2 (Aug. 2021). ISSN: 0163-5840. DOI: [10.1145/3483382.3483389](https://doi.org/10.1145/3483382.3483389). URL: <https://doi.org/10.1145/3483382.3483389> (cit. on p. 26).
- [74] M. Joshi et al. “TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 1601–1611. DOI: [10.18653/v1/P17-1147](https://doi.org/10.18653/v1/P17-1147). URL: <https://aclanthology.org/P17-1147> (cit. on p. 27).
- [75] A. Fan et al. “ELI5: Long Form Question Answering”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 3558–3567. DOI: [10.18653/v1/P19-1346](https://doi.org/10.18653/v1/P19-1346). URL: <https://aclanthology.org/P19-1346> (cit. on p. 27).
- [76] C. Bartolomeu, R. Nóbrega, and D. Semedo. “Understanding News Text and Images Connection with Context-Enriched Multimodal Transformers”. In: *Proceedings of the 30th ACM International Conference on Multimedia*. MM ’22. Lisboa, Portugal: Association for Computing Machinery, 2022, pp. 5821–5832. ISBN: 9781450392037. DOI: [10.1145/3503161.3548430](https://doi.org/10.1145/3503161.3548430). URL: <https://doi.org/10.1145/3503161.3548430> (cit. on pp. 30, 34, 51).
- [77] A. Lommatzsch et al. “NewsImages: Addressing the Depiction Gap with an On-line News Dataset for Text-Image Rematching”. In: *Proceedings of the 13th ACM Multimedia Systems Conference*. MMSys ’22. Athlone, Ireland: Association for Computing Machinery, 2022, pp. 227–233. ISBN: 9781450392839. DOI: [10.1145/3524273.3532891](https://doi.org/10.1145/3524273.3532891). URL: <https://doi.org/10.1145/3524273.3532891> (cit. on p. 30).
- [78] S. Patil. *Question Generation using transformers*. Version 1.0.0. URL: https://github.com/patil-suraj/question_generation (cit. on p. 33).

- [79] I. Akermi, J. Heinecke, and F. Herledan. “Transformer based Natural Language Generation for Question-Answering”. In: *Proceedings of the 13th International Conference on Natural Language Generation*. Dublin, Ireland: Association for Computational Linguistics, Dec. 2020, pp. 349–359. URL: <https://aclanthology.org/2020.inlg-1.41> (cit. on p. 33).
- [80] Y.-C. Chen et al. *UNITER: UNiversal Image-TExt Representation Learning*. 2019. DOI: [10.48550/ARXIV.1909.11740](https://arxiv.org/abs/1909.11740). URL: <https://arxiv.org/abs/1909.11740> (cit. on p. 47).
- [81] W. Su et al. *VL-BERT: Pre-training of Generic Visual-Linguistic Representations*. 2019. DOI: [10.48550/ARXIV.1908.08530](https://arxiv.org/abs/1908.08530). URL: <https://arxiv.org/abs/1908.08530> (cit. on p. 47).
- [82] X. Li et al. “Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks”. In: *ECCV*. Aug. 2020. URL: <https://www.microsoft.com/en-us/research/publication/oscar-object-semantics-aligned-pre-training-for-vision-language-tasks/> (cit. on pp. 47, 48).
- [83] L. Gui et al. “KAT: A Knowledge Augmented Transformer for Vision-and-Language”. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 956–968. DOI: [10.18653/v1/2022.naacl-main.70](https://aclanthology.org/2022.naacl-main.70). URL: <https://aclanthology.org/2022.naacl-main.70> (cit. on pp. 49, 51, 64).
- [84] NLP Connect. *vit-gpt2-image-captioning (Revision 0e334c7)*. 2022. DOI: [10.57967/hf/0222](https://huggingface.co/nlpconnect/vit-gpt2-image-captioning). URL: <https://huggingface.co/nlpconnect/vit-gpt2-image-captioning> (cit. on p. 50).
- [85] A. Singh et al. “FLAVA: A Foundational Language and Vision Alignment Model”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 15638–15650 (cit. on p. 64).



ANNEX 1

This Annex reveals our submission to the ACL 2023 conference.

NewsMQA - An Open-ended Multimodal Question-Answering Benchmark for the News Domain

Anonymous ACL submission

Abstract

In complex domains such as news, acute information-seeking can be delivered through question-answering (QA). However, due to its open-domain and multimodal nature, where images complement the textual medium, QA systems are required to have the flexibility to dynamically multiplex both open-domain visual and linguistic sources to ground answers. Existing news QA datasets are limited to a particular type of answer grounding, which hides the complexity of the problem. We propose NewsMQA, a novel dataset for the task of Multimodal Question-answering for News, comprising 11 176 question-answer pairs. NewsMQA departs from existing datasets by 1) fully encompassing the multimodal facets of news, and 2) improving on the quality vs. scale trade-off. For the latter, we propose a two-part approach that combines human annotation with synthetic question-answer generation through answer roundtrip consistency. We conduct an extensive analysis and benchmarking of the proposed dataset and identify both its unique characteristics and the intrinsic research challenges of the task that it supports.¹

1 Introduction

Question-answering (QA) models provide an intuitive and natural information-seeking medium, where a user can satisfy acute information needs by posing questions. In particular domains, it is often the case that information needs have *heterogeneous groundings*, *i.e.* spanning over different modalities. This is the case of news pieces, which are open-domain rich multimodal documents, often composed of a title, lead paragraph, text body, and illustrating images, aiming to best inform users in a concise and accurate manner. Images act as an essential visual stimulus, conveying information that goes beyond language, giving readers the

Headline: "Study Gauges Plastic Levels in Oceans"

Passage: "When the survey teams looked for plastics floating in the water that were the size of grains of sand, however, they were surprised to find far fewer samples than expected - one-hundredth as many particles [...] suggests that the smaller bits may be swept deeper into the sea or consumed by marine organisms."

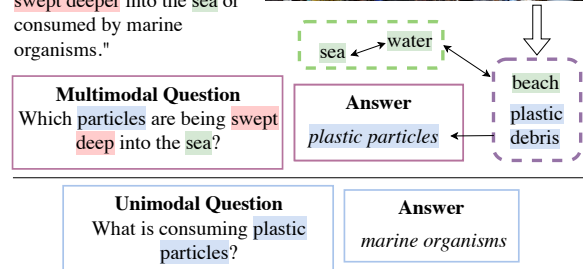


Figure 1: Multimodal vs. Unimodal questions on the news domain. In a multimodal question, the answer grounded across both visual and textual news elements. Source (Schwartz, 2014).

opportunity to "witness" the events (Messaris and Abraham, 2001; Liu et al., 2020).

This leads to a formulation of the novel task of Multimodal Question-answering for News, which emerges as a combination of a) open-ended text-based QA (Chen et al., 2017; Rajpurkar et al., 2016, 2018), in which answers to a question are extracted from a context passage, and b) Visual question-answering (Antol et al., 2015) (VQA), which seeks to answer questions about an image in a closed-ended setting (*i.e.* usually formulated in a categorical setting). Figure 1, illustrates this, where an unimodal question can be answered by skipping the image, while the multimodal one requires a joint interpretation of both the image and the context passage. This setting poses a challenge for QA approaches, which not only need to cope with the open-domain nature of the task and capture the

¹Code and dataset will be made available after publication.

broadness of textual constructs and visual material but also have the flexibility to dynamically multiplex both vision and language sources to ground answers.

The vast availability of adequate datasets has enabled significant advances in both VQA (Antol et al., 2015; Marino et al., 2019; Wang et al., 2018) and open-ended span-based QA (Rajpurkar et al., 2016). This is not the case for multimodal question-answering datasets, where the available datasets are a) not open-domain and thus too focused on highly specific domains (Sheng et al., 2016), b) restricted to the existence of human entities in images, thus not covering the full spectrum of news content (Reddy et al., 2022), or c) are very limited in terms of human-curated training and evaluation samples (Sheng et al., 2016; Reddy et al., 2022).

We formalise the task of News Multimodal QA and propose NewsMQA, a novel dataset for open-domain news multimodal question-answering, covering a wide range of news topics while evidencing the inherent multimodality of news documents. Figure 1 provides a sample from the dataset. In the construction of NewsMQA, we aim for scale and quality, with the latter being assured through a significant portion of human-curated samples. To achieve this, we adapt previous hybrid dataset construction efforts (Alberti et al., 2019; Changpinyo et al., 2022) to our task, by applying a two part approach: 1) multimodal synthetic question and answer generation component and 2) two-step crowd-sourced question-answer annotation.

We conduct an extensive analysis of the proposed dataset, identifying the unique characteristics of NewsMQA as well as the intrinsic research challenges of the task that it supports.

2 News Multimodal QA Formulation

News pieces seek to inform users in a concise yet comprehensive manner. As such, they are linguistically and visually rich open-domain documents, dense in knowledge. From a multimodal perspective, the connection between news pieces’ text and illustrations is heterogeneous (Bartolomeu et al., 2022; Lommatzsch et al., 2022), meaning that there are multiple news text-to-image grounding patterns, that journalists use when choosing a specific image.

Having a dataset that captures these traits is critical to address the challenging task of Multimodal Question Answering in the news domain. Existing datasets have a set of limitations w.r.t. capturing

the news multimodality aspects, by either solely focusing on text (Wang et al., 2021; Trischler et al., 2017) or failing short to fully capture the inherent heterogeneity underlying news pieces and images connection (Reddy et al., 2022).

We formulate the task of Multimodal News QA and create a dataset that embeds the aforementioned traits. Namely, given a question q_i about a news passage and an image v_i , the goal is to extract the excerpt from the passage p_i that answers the question. To support this task, we create a dataset in which each sample is composed of four elements: an image and passage from a news piece, a question, and the answer to that question. The question can be **unimodal**, meaning that one can find the answer just by interpreting a single modality (image or text passage), or **multimodal**, requiring a joint interpretation of the two mediums.

3 NewsMQA: Dataset Creation Pipeline

Our dataset creation pipeline consists of three distinct phases: 1) images and passage selection; 2) question-answer candidates generation and 3) two-step QA annotation.

3.1 Images and Passage Selection

We construct NewsMQA with news pieces from NYTimes². Namely, we sampled news piece elements (image and passage) from the NYTimes800k dataset (Tran et al., 2020), which has 444 914 English articles, with an average of 963 words/article. Most articles have an image at the top that correlates to the main topic and 39% have at least one more image throughout the text.

Sampling Topics and News pieces. First and foremost, we defined a set of diverse topics representative of the news domain. Those topics are: *art, books, food, movies, politics, science, sports and technology*. Then, for each topic, we collected a list of related keywords, and retrieve news pieces by searching over their lead paragraph with those keywords with an ad-hoc retrieval method.

Images and passage Selection. Given the per-topic retrieved news pieces, we proceed to the creation of $\langle image, passage \rangle$ pairs. Throughout a news piece, journalists illustrate certain details with images that further help the reader contextualize and visualize the event. These images are thus grounded on their surrounding text and are usu-

² <https://www.nytimes.com>

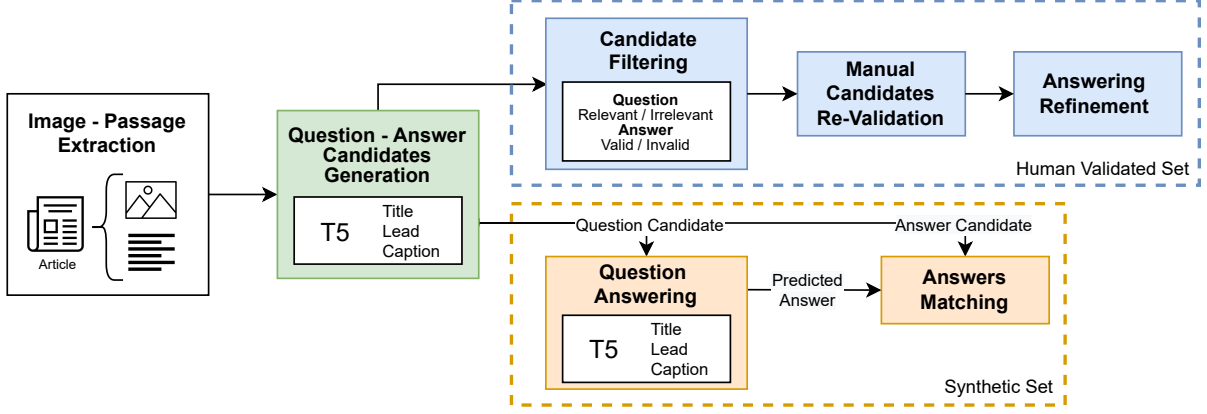


Figure 2: News Multimodal Question-Answering hybrid dataset creation pipeline.

ally placed before the corresponding text (Tolmie et al., 2017). Therefore, for each image, our ground truth passage corresponds to the two paragraphs following the image.

After gathering all the pairs, we only kept those that met the following criteria: (1) no invalid characters, (2) the passage has at least 50 words, ensuring that the samples have enough information about the image and news. With this process, we gathered a total of 227 685 $\langle image, passage \rangle$ samples pairs.

3.2 Question-answer Candidates Generation

Given the extracted pairs, we propose a strategy to automatically generate questions that inherit the intrinsic multimodality of the task. Specifically, we leverage a T5 model Raffel et al. (2022) to autoregressively generate questions, given a controlled context sequence.

To generate questions that embed multimodal knowledge, the context sequence must be enriched with grounding information that acts as the bridge between text and image. For that purpose, we leverage the news image captions provided by journalists. These news-like captions describe the image by bridging the visual material with the news story (Tran et al., 2020; Bartolomeu et al., 2022).

As such, given the i -th sample image-passage pair $\langle v_{di}, p_{di} \rangle$, from a news piece d , we create an input sequence s_{di} comprising the news piece title t_d , lead paragraph h_d and the image caption c_{di} . The first two elements provide general information about the news piece, while the caption c_{di} brings information that connects the two modalities. The passage p_i is kept hidden to ensure that the generated questions focus on the main news story and image caption. Each sequence s_{di} is formatted

such that the answer is prepended, and given as input to a T5 model fine-tuned in an end-to-end multi-task setting, for question+answer generation and answer extraction (Akermi et al., 2020), on SQuAD (Rajpurkar et al., 2016). For each generated image-passage pair $\langle v_{di}, p_{di} \rangle$, we compose and provide as input the sequence s_i , and generate four question-answer pairs per sample.

A post-processing step is performed to filter invalid pairs. Namely, we keep only pairs in which (1) the answer is included in the corresponding reference passage p_{di} , (2) the answer is not within the question itself, (3) generated questions end with the “?” character. With this process, we gathered a total of 44 404 candidates.

3.3 Two-step QA Annotation

We seek to create a dataset whose number of samples can support the training of models tailored for the task while ensuring its quality. To strike this trade-off, we adopt a two-step approach to filter and refine the generated candidate $\langle question, answer \rangle$ pairs through the creation of 1) human refined/validated and 2) synthetically annotated subsets.

3.3.1 Human Validation

We resort to human annotation to filter and refine the generated question-answer candidates and assure their respective quality.

#1 - Candidate Filtering. We created an Amazon MTurk task where we asked workers to filter invalid candidate pairs. Given a candidate tuple (image v_{di} , passage p_i , question q_i and answer a_i), the task consisted in classifying the relevance of the question and the validity of the answer, according to the following criteria: (1) the **question** must be

relevant regarding both visual and textual elements and (2) must have one singular answer; (3) the answer must be a span (section of the given passage) and (4) respond correctly to the question. From the candidates' pool, we sampled and annotated 10 000 question-answer pairs, in a per-topic stratified manner to ensure topic balance. Further details of the crowdsourcing experiment and workers' interface are depicted in Appendix A.1. After this step, we filtered a total of 2180 low-quality candidates.

#2 - Manual Candidates Re-Validation. We observed that the task of deeming whether a generated question meets the defined criteria is cognitively demanding. As such, we conducted a second human validation step over a subset of the candidates, with a restricted group of annotators. Specifically, we asked 4 annotators to re-validate all candidates whose question was annotated as "relevant", which corresponds to a total of 7820 candidates, with 4776 "valid" and 3044 "invalid" answer samples. The validation task was framed in the exact same manner, with the criteria used to assess the relevance of a given question and the correctness of an answer being the same.

As a result of this question re-validation step, a total of 3293 samples were annotated as having relevant questions. From these, 1547 had valid answers and the remaining 1746 invalid ones.

#3 - Answering Refinement. A final manual annotation step is conducted, to guarantee samples with high-quality questions for the proposed task are not thrown away. Therefore, we asked 16 annotators to answer the 1746 candidate samples whose question was deemed as "valid", but had "invalid" answers. As a result, we gathered a total of 3293 high-quality human validated $\langle v_{di}, p_{di}, q_i, a_i \rangle$ samples, each comprising an image v_{di} , a passage p_{di} that refers to that image, a question q_i and its answer a_i .

3.3.2 Synthetic Sample Annotation

We adopt an automatic approach to efficiently and cost-effectively complement and scale up the dataset with additional samples. A synthetic QA generation procedure is employed, based on ensuring roundtrip answer consistency (Alberti et al., 2019). This strategy has shown to be effective in closely related tasks, such as visual question-answering setting (Changpinyo et al., 2022).

Specifically, we leverage two different models: **a)** question-answer generation model and **b)** answer extraction, both being T5 models fine-tuned

Table 1: Distribution of news pieces (N), images (I) and questions (Q) per topic.

	Human Validated			Synthetic		
	# N	# I	# Q	# N	# I	# Q
arts	590	609	710	1068	1141	1199
books	571	593	737	1090	1154	1249
food	353	364	439	556	596	637
movies	659	675	801	2064	2141	2253
politics	498	500	562	1329	1372	1446
science	345	349	414	698	723	761
sports	417	425	488	1258	1306	1399
tech	400	414	498	689	719	769
Total	2832	2889	3293	7159	7461	7883

Table 2: Comparison between the sets regarding the question and answer key stats. HV: Human Validated

	Avg. Length / Vocab Size		
	Questions	Answers	Passage
HV	10.83 / 7455	5.9 / 7538	113.1 / 46256
Synthetic	10.81 / 13926	1.9 / 7005	103.9 / 86555
NewsMQA	10.82 / 17054	3.1 / 12168	106.7 / 102722

in the corresponding tasks. The rationale is to have one model generating a question and an answer, and a second model predicting the answer to the generated question. If the two answers match, consistency is achieved. We provide detail on each model in Appendix A.2. For the question generation, we used the generated question-answer pairs from section 3.2, which received as input the title t_d , lead paragraph h_d and image caption v_{di} . Apart from these elements, the answer extraction model takes the generated question as input and predicts the answer. We applied the aforementioned procedure to the remaining 34404 generated candidates and obtained a total of 7883 synthetic samples.

4 Dataset Overview and Analysis

Table 1 summarizes the final NewsMQA dataset distribution, detailing the number of samples for both the Human Validated and Synthetic subsets, across the different topics. NewsMQA comprises a total of 11 176 questions (3293 manually and 7883 automatically annotated) for 9986 image-passage pairs from 9531 different New York Times articles ranging from March 2005 to August 2019.

Human Validated vs. Synthetic sets. Despite conveying the exact same elements, the two sets inherently have distinct characteristics. Table 2 compares the average question and answer lengths, as well as their vocabulary diversity.

As can be observed, the Human Validated set

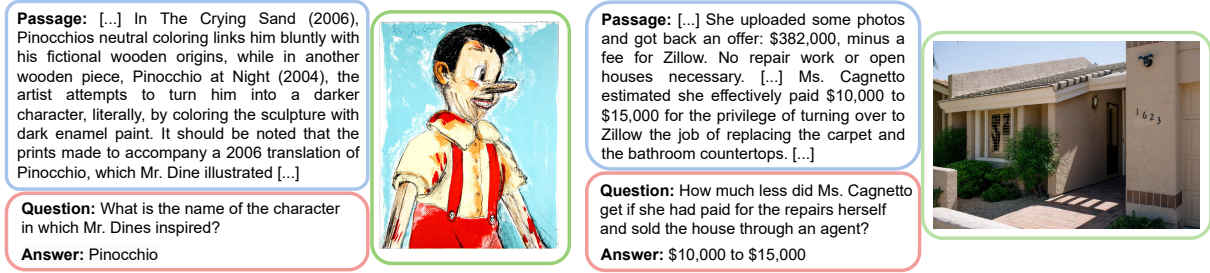


Figure 3: Two examples from the NewsMQA dataset with the corresponding image, passage, question and answer.

Table 3: Analysis of the number of samples whose answer to the question is solely within one of the components used to generate the candidate pairs.

	Human Validated	Synthetic
Passage	3293 (100%)	7883 (100%)
Article's title	167 (5.07%)	822 (10.43%)
Lead paragraph	1129 (34.28%)	2975 (37.74%)
Image caption	633 (19.22%)	2560 (32.47%)
Various components	1364 (41.43%)	1526 (19.36%)

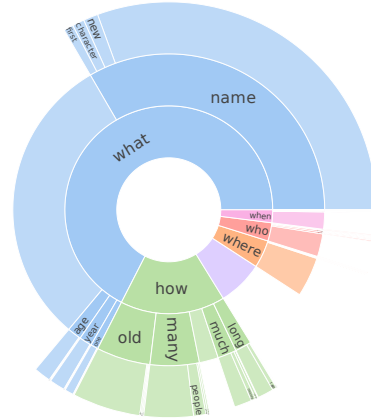


Figure 4: Distribution of the top 5 question types. Empty areas cluster the remaining terms.

comprises questions with longer answers and a greater variety of unique words. Therefore we expect it to have a consequently increased difficulty compared to the Synthetic subset. Having more diverse and creative answers on the Human Validated set is expected, especially since paraphrasing is trivial for humans, as well as the capacity to digest and synthesize information. However, even though the employed synthetic approach answers are shorter and less diverse, under a curriculum learning framing, they are still useful for training.

Questions' diversity analysis. Figure 4 depicts the distribution of question types covered in the NewsMQA dataset. The top 5 most represented question types are "what" (which represents 67.47% of the totality of question types), followed by "how", "where", "who" and "when". These show the broad coverage in NewsMQA of the different question types on the spectrum of the news domain.

Dataset challenges. Figure 3 shows two samples of the NewsMQA dataset. These show two kinds of questions found in our dataset that present different challenges to reaching the goal of answering the given question. While the question on the right is answerable solely by having access to the passage, with the question on the left it is not evident what the question refers to without the image information, needing both to reach to answer correctly. More samples are provided in Appendix A.3.

Questions multimodality analysis. NewsMQA is constructed such that the heterogeneous nature of the connection between images and text is inherited. As such, the context sequence used to generate candidate question-answer pairs was built with a combination of unimodal (title, lead paragraph) and multimodal (image caption) elements.

Table 3 shows the impact of each of these components on the generation of the question-answer pairs. It can be seen that $\approx 19\%$ and $\approx 32\%$ of the answers, on the Human Validated and Synthetic subset, respectively, were extracted exclusively from the caption. Then, on average 44% of the answers were extracted from either the title or lead paragraph. Finally, an average of 30% of the answers appear in multiple context components.

5 Multimodal QA Baselines

5.1 Pre-trained QA Backbones

Several vision-linguistic models inspired by the Transformers architecture have been proposed to tackle multimodal tasks such as image-text matching and VQA (Tan and Bansal, 2019; Lu et al., 2019; Chen et al., 2019; Li et al., 2019; Su et al., 2019; Li et al., 2020). When applied to a down-

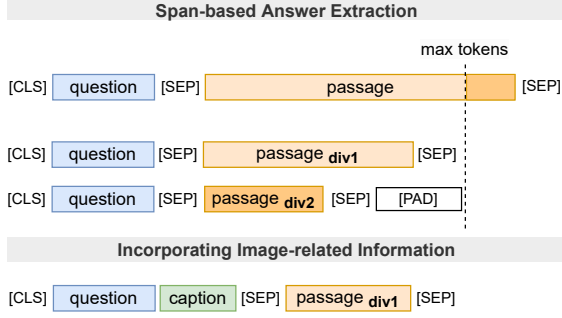


Figure 5: Span-based sequence encoding illustration.

stream QA task, these often turn the problem into a classification one, *i.e.* each candidate answer is treated as a class. This is not suitable for open-ended QA like in NewsMQA, especially in the news domain, which is characterized by a highly wide answer vocabulary and diverse answer length. In (Reddy et al., 2022), an Oscar multimodal transformer (Li et al., 2020) is extended for span-based answer extraction but requires the image caption as input, which is not available. Thus, to benchmark NewsMQA in a principled and realistic manner, we adopt span-based answer extraction backbone models and enrich them with multimodal information.

BERT (Devlin et al., 2019) is one of the most widely used baselines for NLP tasks. It was pre-trained with Masked Language Modeling and Next Sentence Prediction tasks. Thanks to the pre-training process, there is no need for heavily-engineered task-specific architectures. The model can be fine-tuned by appending task-specific heads, in a classification or regression setting.

SpanBERT (Joshi et al., 2020) extends BERT to include span prediction in pre-training. While in BERT individual tokens are masked, in SpanBERT they mask continuous spans of text. Moreover, the model is pre-trained with both longer full-length contexts and a span-boundary objective. These force the model to learn how to predict each token in the masked span exclusively using the representations of the observed unmasked tokens.

T5 (Raffel et al., 2022) is a encoder-decoder generative model pre-trained on an assortment of text-to-text tasks as text classification, machine translation and summarization.

We propose tailored context-encoding schemes for these backbone models, in order to extend their QA contexts with the multimodal information required to solve the NewsMQA task.

5.2 Multimodal Context Sequence Encoding

Given a news passage and an image as context, NewsMQA requires being able to multiplex both textual and visual information to answer a question. One way to accomplish this is to enrich the input sequence with extra contextual information, in particular, elements that capture the information depicted in the image (Bartolomeu et al., 2022; Gui et al., 2022). This requires carefully designed sequence encoding schemes to effectively model multimodal information and deal with the model’s input sequence size limitations.

Visual domain context. While backbone models receive a text sequence as input, NewsMQA comprises questions that require a multimodal answer grounding. In these cases, the samples’ context needs to be enriched such that visual domain information is embedded. Following (Gui et al., 2022), we propose to enrich each sample context with visual cues from the image. In particular, we embed automatically extracted image captions to each backbone input sequence. These are extracted from a fine-tuned GPT-2 model (NLP Connect, 2022) on the MS COCO dataset (Lin et al., 2014), with a ViT image backbone (Dosovitskiy et al., 2021). Given that the answer a_i needs to be extracted from the passage p_i , at training and inference time, we treat the captions in the same manner as the questions and restrict the models such that answers are exclusively extracted from the passage.

Span-based Answer Extraction. In span-based answer extraction backbones (*e.g.* BERT and SpanBERT), the classification head is trained to predict the start and end tokens³, taking contextualized token embeddings as input. Passages may be too long, exceeding the model’s maximum token length. This is further aggravated when the input sequence is augmented with visual information. When this occurs, we create sub-sequences splitting the full passage into sub-passages, that are each appended to the sample’s question. The process is exemplified in Figure 5. To maintain the contexts as cohesive as possible, we adopt a sliding-window approach with a stride of 64 tokens between consecutive sub-sequences.

When evaluating each sample, depending on the divisions, we may have several candidate answers. The final prediction is the candidate answer span with greater confidence. Due to the breaking of

³With two classes: *start* and *end*.

the original samples’ sequences in sub-sequences, the correct answer span may not be on the sub-sequence passage. For this special case, as in (Rajpurkar et al., 2018), the model is trained to predict the index (0,0) as the *start* and *end* tokens.

Span-based Decoding Models. Generative models are fine-tuned to decode the answer span from the context sequence. The T5 backbone takes as input a sequence with a *question* prompt and the question itself followed by the context passage. To incorporate image captions, we adopt a similar approach as for answer extraction models and append the additional knowledge after the question. We use beam-search to decode the answer span, and empirically set the number of beams to 4.

6 Experimental Evaluation

We now conduct a comprehensive benchmarking and analysis of the NewsMQA dataset, assessing and discussing its intricacies and challenges.

6.1 Protocol, Baselines and Metrics

Each NewsMQA subset was divided into 70% for training, 15% for validation and 15% for testing. When evaluating the full NewsMQA, we combine the corresponding training, validation and test sets. Following the literature on span-based QA (Rajpurkar et al., 2016, 2018), we report two metrics: Exact Match (EM) and F1.

For the backbones, we use the large variants of T5, SpanBERT, and BERT, which have 770, 340 and 340 million parameters, respectively, fine-tuned on the SQuAD v1 dataset (Rajpurkar et al., 2016). In all experiments, backbones are fine-tuned in the NewsMQA dataset (full or a particular subset), for 20 epochs. We then select the model with the highest Exact Match score. Further details regarding models and training are in Annex A.5.

6.2 Studying Context Settings

We start by exploring the impact of different context settings, by comparing the setting corresponding to the proposed task (only passage is available), from settings where privileged information is used (original image caption and title). We use the test set of the Synthetic subset and all the baseline models in a zero-shot setting, i.e. w/o fine-tuning. The Oracle (T5-base) provides the upper bound as it was the one used to re-answer the generated question-answer candidates.

Table 4: Exact Match zero-shot performance (%) on the Synthetic test set, over different context settings. T: title, L: lead paragraph, C: image caption, P: passage.

	T	L	T+L	T+L+C	P
Oracle	14.13	49.15	59.31	98.95	68.78
SpanBERT	14.13	41.79	51.61	77.41	60.15
BERT	14.81	44.92	55.41	83.76	67.34
T5	14.38	44.42	53.64	82.57	65.39

Table 4 shows the results, where the contrast between using the title (T) and the lead paragraph (L) with and without access to the image caption (C) can be observed. Across all models, a consistent 25% drop in performance is observed when the image caption is not available, evidencing that to solve the task, models are required to use image information to provide accurate answers. When we have only the passage as context (P), we can observe that the performance is greater than in the title + lead (T+L) setting. Notwithstanding, we can see that there is a significant margin of improvement⁴.

Table 5: Exact match / F1 performance (%) results in the NewsMQA full dataset adding visual context. * The original caption is privileged information.

	Only Passage	with Caption Generated	Original*
<i>Zero-shot</i>			
SpanBERT	57.2 / 66.9	56.2 / 65.9	59.3 / 69.6
BERT	63.4 / 71.4	61.8 / 69.6	65.2 / 73.0
T5	61.3 / 69.5	61.7 / 69.9	65.2 / 73.7
<i>Fine-tuned</i>			
SpanBERT	69.8 / 76.3	68.9 / 75.4	74.4 / 80.4
BERT	70.1 / 76.7	68.4 / 74.8	75.9 / 81.6
T5	72.2 / 78.3	72.5 / 78.7	74.7 / 80.3

6.3 Benchmarking Visual Context

In this section, we benchmark the outcome of enriching each sample context with visual information, in particular, with COCO-like generated captions. These tend to describe the scene in a general manner, thus not covering specific entities or news story details, but capturing explicit visual concepts (e.g. a person, beach, etc.).

Table 5 reports the results of the proposed multi-modal context encoding scheme in both a zero-shot and fine-tune (on the NewsMQA training set). In general, the addition of generated captions does not appear to have a positive impact. Only with

⁴Note that the answer is always available in the passage.

Table 6: Performance on the Human Validated test set (HV), with different fine-tuning schemes. EM / F1 (%)

Fine-tuning set		Test on HV Subset	
		No Caption	Generated
SpanBERT	HV	56.1 / 67.7	57.7 / 69.2
	Synthetic	55.5 / 66.1	55.7 / 66.17
	NewsMQA	58.7 / 69.3	57.9 / 68.01
BERT	HV	56.9 / 70.5	56.3 / 70.6
	Synthetic	54.5 / 65.9	56.3 / 65.9
	NewsMQA	58.5 / 71	56.1 / 68.1
T5	HV	56.7 / 69.7	55.9 / 68.1
	Synthetic	57.5 / 64.8	57.3 / 65.4
	NewsMQA	59.1 / 71.6	58.3 / 70.1

the fine-tuned T5 backbone performance increases slightly. Given that when models have access to the original news image captions, which are privileged information, performance improves significantly, which suggests that general captions fail to provide useful context information. In overall, this experiment supports our observation that the connection between text and images in news is fundamentally different, bringing novel challenges. As observed, these are embedded in the proposed NewsMQA.

6.4 Assessing Dataset Quality

To assess our dataset quality, we explore the impact and difficulty of the Synthetic and Human Validated subsets, and both (full NewsMQA). To accomplish this, we individually fine-tuned each backbone on each set and evaluate them on the Human Validated test set. Table 6 results show that when finetuned on both subsets, performance is consistently higher in terms of EM, than when fine-tuned individually in either the HV or Synthetic subsets. This confirms the synthetic sample generation strategy we adopted leads to performance improvement. Among the different backbones, T5 is consistently the best performing one.

7 Related Work

While open-ended question-answering has been extensively researched in the text domain (Rajpurkar et al., 2016; Chen et al., 2017), in the multimodal domain, literature and data resources are still very limited. Interestingly, in a close-ended setting (*i.e.* QA is treated as a classification problem), multimodal QA tasks like visual question-answering (Antol et al., 2015) and their extensions (*e.g.* include commonsense knowledge (Shah et al., 2019), or external knowledge (Marino et al., 2019; Wang et al., 2018; Shah et al., 2019)), have been

extensively researched, with numerous datasets being now available. This vein of work focuses on a particular setting in which the goal is to answer image-centric general questions and whose answers are grounded on both the visible objects and factual knowledge around these. In this sense, both OK-VQA (Marino et al., 2019) and FVQA (Wang et al., 2018) resemble our proposed NewsMQA. However, they formulate the QA task in a close-ended setting and are restricted to both general and factual knowledge. In (Sheng et al., 2016) an art-focused multimodal QA dataset is proposed. Despite covering a challenging topic, its domain is very narrow (old-Egyptian Amarna period), as well as its size (800 questions). In the news domain, linguistic and visual elements are not only highly diverse but are also inherently intertwined with the story. The MuMuQA (Reddy et al., 2022) dataset takes a step further in this direction, by comprising question and answer pairs from the news domain. However, 1) questions follow an entity-centred setting, which does not capture the full connection heterogeneity between news text and images and 2) the task requires an image-caption as input, what hides the complexity of the task. The training set is fully synthetic and the human annotated set is very limited (1384 samples). NewsMQA exposes the full multimodality of news and comprises 3293 manually annotated + 7883 synthetically generated samples.

8 Conclusions and Future Work

In this paper we introduced the task of open-ended News Multimodal Question Answering, in which given a question about a news passage and an image, the goal is to extract the answer excerpt. To support the task, we introduced NewsMQA, a novel multimodal QA dataset that fully encompasses the multimodal facets of news and improves the quality vs. scale trade-off with a mix of Human Validated and Synthetic samples. NewsMQA departs from previous datasets, by truly capturing the full multimodality spectrum inherent to the news domain. An extensive analysis is conducted, in which we identified both the unique characteristics and intrinsic research challenges of the task. We confirmed both the quality of the NewsMQA dataset as well as the utility of the synthetically generated samples.

In future work we will study visual context-enriching mechanisms, towards effectively bridging both the news image and passage, and use it to ground answers.

9 Limitations

Annotating samples for the multimodal QA task in the news domain is not only a labour intensive but also cognitively demanding task. Questions and answers require deep knowledge of each news story, which is challenging for such a wide domain. This naturally imposed a set of limitations on our work. Specifically, to reach a significant scale, we resorted to synthetic sample generation. As evidenced by our experiments, both the quality and difficulty of synthetic samples are lower when compared to human annotated samples.

In terms of benchmarking, while we followed the ideas of state-of-the-art approaches in closed-ended multimodal QA that leverage large closed-source pre-trained models (Gui et al., 2022) (GPT-3), it would be interesting to evaluate state-of-the-art multimodal foundational backbones such as FLAVA (Singh et al., 2022). Since this requires adapting these models for span-based extraction, we leave this for future work.

References

Imen Akermi, Johannes Heinecke, and Frédéric Herledan. 2020. [Transformer based natural language generation for question-answering](#). In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 349–359, Dublin, Ireland. Association for Computational Linguistics.

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. [Synthetic QA corpora generation with roundtrip consistency](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173, Florence, Italy. Association for Computational Linguistics.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Cláudio Bartolomeu, Rui Nóbrega, and David Semedo. 2022. [Understanding news text and images connection with context-enriched multimodal transformers](#). In *Proceedings of the 30th ACM International Conference on Multimedia*, MM ’22, page 5821–5832, New York, NY, USA. Association for Computing Machinery.

Soravit Changpinyo, Doron Kukliansy, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. 2022. [All you may need for VQA are image captions](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages

1947–1963, Seattle, United States. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. [Reading Wikipedia to answer open-domain questions](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. [Uniter: Universal image-text representation learning](#).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. [An image is worth 16x16 words: Transformers for image recognition at scale](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2022. [KAT: A knowledge augmented transformer for vision-and-language](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 956–968, Seattle, United States. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. [Span-BERT: Improving pre-training by representing and predicting spans](#). *Transactions of the Association for Computational Linguistics*, 8:64–77.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. [Visualbert: A simple and performant baseline for vision and language](#). *CoRR*, abs/1908.03557.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. [Oscar: Object-semantics aligned pre-training for vision-language tasks](#). In *ECCV*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar,

712	and Larry Zitnick. 2014. Microsoft coco: Common objects in context . In <i>ECCV</i> . European Conference on Computer Vision.	768
713		769
714		770
715	Fangyu Liu, Rémi Lebre, Didier Orel, Philippe Sordet, and Karl Aberer. 2020. Upgrading the newsroom: An automated image selection system for news articles . <i>ACM Trans. Multimedia Comput. Commun. Appl.</i> , 16(3).	771
716		772
717		
718		773
719		774
720	Andreas Lommatzsch, Benjamin Kille, Özlem Özgöbek, Yuxiao Zhou, Jelena Tešić, Cláudio Bartolomeu, David Semedo, Lidia Pivovarov, Mingliang Liang, and Martha Larson. 2022. Newsimages: Addressing the depiction gap with an online news dataset for text-image rematching . In <i>Proceedings of the 13th ACM Multimedia Systems Conference, MMSys '22</i> , page 227–233, New York, NY, USA. Association for Computing Machinery.	775
721		776
722		777
723		
724		778
725		779
726		780
727		781
728		782
729	Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks . Curran Associates Inc., Red Hook, NY, USA.	783
730		784
731		
732		
733	Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge . In <i>2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)</i> , pages 3190–3199.	785
734		786
735		787
736		788
737		789
738		790
739		791
740	Paul Messaris and Linus Abraham. 2001. <i>The role of images in framing news stories.</i> , LEA's communication series., pages 215–226. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US.	792
741		793
742		794
743	NLP Connect. 2022. vit-gpt2-image-captioning (revision 0e334c7) .	795
744		796
745	Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2022. Exploring the limits of transfer learning with a unified text-to-text transformer. <i>J. Mach. Learn. Res.</i> , 21(1).	797
746		798
747		799
748		800
749		801
750	Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD . In <i>Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)</i> , pages 784–789, Melbourne, Australia. Association for Computational Linguistics.	802
751		
752		
753		
754		
755		
756		
757	Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text . In <i>Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing</i> , pages 2383–2392, Austin, Texas. Association for Computational Linguistics.	811
758		812
759		813
760		814
761		815
762		
763	Revant Gangi Reddy, Xilin Rui, Manling Li, Xudong Lin, Haoyang Wen, Jaemin Cho, Lifu Huang, Mohit Bansal, Avirup Sil, Shih-Fu Chang, Alexander Schwing, and Heng Ji. 2022. Mumuqa: Multimedia multi-hop news question answering via cross-media knowledge extraction and grounding . <i>Proceedings of the AAAI Conference on Artificial Intelligence</i> , 36(10):11200–11208.	816
764		817
765		818
766		819
767		820
		821
		822
	John Schwartz. 2014. Study gauges plastic levels in oceans .	
	Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering . <i>Proceedings of the AAAI Conference on Artificial Intelligence</i> , 33(01):8876–8884.	
	Shurong Sheng, Luc Van Gool, and Marie-Francine Moens. 2016. A dataset for multimodal question answering in the cultural heritage domain . In <i>Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)</i> , pages 10–17, Osaka, Japan. The COLING 2016 Organizing Committee.	
	Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model . In <i>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)</i> , pages 15638–15650.	
	Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vi-bert: Pre-training of generic visual-linguistic representations .	
	Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers . In <i>Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)</i> , pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.	
	Peter Tolmie, Rob Procter, David William Randall, Mark Rouncefield, Christian Burger, Geraldine Wong Sak Hoi, Arkaitz Zubiaga, and Maria Liakata. 2017. Supporting the use of user generated content in journalistic practice . In <i>Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI '17</i> , page 3632–3644, New York, NY, USA. Association for Computing Machinery.	
	Alasdair Tran, Alexander Mathews, and Lexing Xie. 2020. Transform and tell: Entity-aware news image captioning . In <i>2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)</i> , pages 13032–13042.	
	Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordani, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset . In <i>Proceedings of the 2nd Workshop on Representation Learning for NLP</i> , pages 191–200, Vancouver, Canada. Association for Computational Linguistics.	

823 Jiexin Wang, Adam Jatowt, and Masatoshi Yoshikawa.
824 2021. [Archivalqa: A large-scale benchmark dataset](#)
825 [for open domain question answering over archival](#)
826 [news collections](#). *CoRR*, abs/2109.03438.

827 Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and
828 Anton van den Hengel. 2018. [Fvqa: Fact-based vi-](#)
829 [sual question answering](#). *IEEE Transactions on Pat-*
830 *tern Analysis and Machine Intelligence*, 40(10):2413–
831 2427.

A Appendix

A.1 Crowdsourcing Details

A.1.1 Our Mechanical Turk Annotators Interface

Figure 6 displays the task interface that the workers interacted with. We present a brief introduction of our project, explain what we aim to accomplish with the task’s result and list the criteria the worker must consider when classifying the given sample. We also give the worker a code created particularly for that pair ("Your code:"). The worker must copy this code and paste it into the right place at the end of the form before submitting it. This way, we filter out workers who actually read the task introduction and instructions, from workers that don’t do the task correctly. For the given image-passage pair and question-answer pair the worker has to choose one of the four options presented.

When launching the task for the public, we counted two qualification requirements regarding the workers that could respond. The HIT Approval Rate (%) for all Requesters’ HITs must be greater than 97, and the number of HITs Approved greater than 1000 (where a HIT is a response from the worker considered accepted). Upon receiving responses from workers, we automatically rejected any worker who did not use the code provided for this sample and those who always responded with the same answer. Each time a response was rejected, the respective sample became available again for new workers to respond correctly. In the end, 10 707 assignments were done, 707 were rejected and 10 000 were approved. 257 distinct workers participated in the task and provided an average of 42 responses each. The average response time for each response was 56.46 seconds. Each worker received \$0.03 for each accepted answer. We spent a total of 400\$ (300\$ on rewards and 100\$ on Fees to Mechanical Turk). Besides the time spent and the answer they gave to the task, we did not require or save any other personal information about the workers.

A.1.2 The rationale of our AMT interface

In a first attempt, we experimented with launching an AMT task to collect question-answer pairs instead of generating candidates. In this task, we asked the workers to directly generate two questions and corresponding answers given an image-passage pair. However, when testing this approach with a small batch of samples and with the rewards

we aim to employ, we notice the answers were not promising. We concluded that the workers look for tasks that are quick to understand and respond to, and that have easy access to the information needed to correctly complete it. Coming up with a relevant question, aligned with the news story, does not fall into this type of tasks as it is in fact cognitively demanding.

A.2 Synthetic Annotation Model Details

The methodology employed in the assembly of the dataset utilizes two distinct models, both of which can be accessed through the Hugging Face library. For the generation of the question-answer candidate pairs, we used a multi-task T5-Base⁵ model trained for question-answering and answer-aware question generation tasks. For the second answer reinforcement, we used a T5-Base⁶ model fine-tuned for question answering in the SquAD v1 dataset.

A.3 Additional Examples

To gain a deeper insight into the NewsMQA dataset, we provide additional examples of the Human Validated and Synthetic subsets in Figure 7 and Figure 8, respectively.

A.4 Dataset details

Tables 7 and 8 show further statistics about the number of questions per topic in each subset.

A.5 Training details

Fine-tuning on the full NewsMQA dataset took approximately 2 hours for the span-based models (BERT and SpanBERT) and approximately 5 hours for T5, on an NVIDIA A100 GPU with 40GB RAM. The three models were fine-tuned with a batch size of 8 for both training and validation. We used a learning rate of 1e-6, 1e-4 and a weight decay of 0.01 and 1e-2 for the span-based models and T5, respectively.

⁵<https://huggingface.co/valhalla/t5-base-qa-qg-hl>

⁶<https://huggingface.co/valhalla/t5-base-squad>

Help Us Validate the Multimodal Questions About Articles

We are building a system capable of answering multimodal questions regarding news.
Given this image and passage extracted from a news article, tell us if this question and corresponding answer verify our criteria.

----- Consider these **CRITERIA** when validating the Question-Answer pair -----

- Question must be relevant regarding **both visual and textual elements**;
 - Question must have one singular answer;
 - Answer must be a **span** (section of the given passage);
 - Answer must respond **correctly** the question;
- (check the instructions to better understand our hit criteria)

You must select how relevant the [Question] is **AND** if the [Answer] is valid or not.

Your code:
QA-15353-CL

Instructions
Shortcuts

Image



Passage
Most cities around the world denounce, or grudgingly tolerate, painting on public property. But on the Mediterranean island of Malta, the process is encouraged. In the shade of a pedestrian bridge, where old men and women sit on the concrete benches, staring out to sea, a wall has been splashed with color and the spray-painted words NO WAR.

Question
What is the name of the mural on Malta?

Answer
Mediterranean island

Insert your code here:

Select an option

[Question] Relevant + [Answer] Valid	1
[Question] Relevant + [Answer] Invalid	2
[Question] Irrelevant + [Answer] Valid	3
[Question] Irrelevant + [Answer] Invalid	4

Figure 6: Our Mechanical Turk interface for annotating the generated candidates.

Table 7: Number of unique articles and questions per topic of the Human Validated set.

	Human Validated subset					
	Train		Validation		Test	
	#N	#Q	#N	#Q	#N	#Q
arts	423	484	110	112	112	114
books	419	511	120	125	98	101
food	254	295	68	70	71	74
movies	478	548	139	142	110	111
politics	359	392	80	83	86	87
science	255	287	64	69	55	58
sports	312	350	65	67	68	71
tech	286	346	74	76	73	76
Total	2066	2309	481	492	484	492

Table 8: Number of unique articles and questions per topic of the Synthetic set.

	Synthetic subset					
	Train		Validation		Test	
	#N	#Q	#N	#Q	#N	#Q
arts	779	854	179	179	166	166
books	794	879	186	187	182	183
food	391	436	101	101	98	100
movies	1471	1569	355	356	326	328
politics	962	1018	203	203	224	225
science	493	519	111	112	127	130
sports	909	987	218	222	188	190
tech	505	541	101	103	123	125
Total	5146	5519	1175	1182	1172	1182



Passage: In Five Feet Apart, Stella (Haley Lu Richardson) is a teenager with cystic fibrosis. Will (Cole Sprouse), a newcomer to the hospital where she resides, also has the disease and is undergoing a drug trial for an antibiotic-resistant infection. They don't mesh, at first. She is obsessive-compulsive and effusive; she dutifully takes her meds and keeps a YouTube diary about her life. He is a brooding artist and rule-breaker convinced that they are breathing borrowed air. Because of the risk of cross-infection, they must stay six feet apart. Five feet, as measured by the pool cue they carry between them, will be their eventual concession to love and living dangerously.

Question: What separates them?
Groundtruth: cystic fibrosis



Passage: He grins impishly out of lithographs or woodcuts or gets translated rather literally into wood, as in White Gloves, 4 Wheels (2007), where the figure perches on a pallet outfitted with wheels and spattered with paint, his arms thrust outward. In The Crying Sand (2006), Pinocchio's neutral coloring links him bluntly with his fictional wooden origins, while in another wooden piece, Pinocchio at Night (2004), the artist attempts to turn him into a darker character, literally, by coloring the sculpture with dark enamel paint. It should be noted that the prints made to accompany a 2006 translation of Pinocchio, which Mr. Dine illustrated, are more interesting. There are 36 hand-painted lithographs on view, along with layouts of the text; the book itself is for sale in the museums. Here, Mr. Dine's images serve as expressive accompaniments to a translated version of the original text, in which Pinocchio fights off knife-wielding assassins, does prison time and gets eaten by a fish before making it to the promised land of boyhood.

Question: What is the name of the character in which Mr. Dines inspired?
Groundtruth: Pinocchio



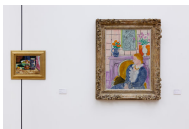
Passage: LONDON Leonardo da Vinci's Bust of a Warrior, a pyrotechnic graphic performance by the young artist, is now one of his most famous drawings. It was inspired by a lost cast metal relief by Leonard's master, Verrocchio, and the lines of the drawing are inscribed on the paper in metal for this is one of the artist's works in silverpoint. The drawing is one of the treasures of the British Museum, which has the most important collection in the world of drawings in metalpoint in silver, gold, lead and other metals and which is now staging, in association with the National Gallery of Art in Washington, its first exhibition of nearly 100 works, stretching from the early Renaissance to the present day.

Question: What is the name of the character in which Mr. Dines inspired?
Groundtruth: Pinocchio



Passage: WHAT about an audiobook? Pharrell Williams asked, sitting at the head of a conference table at the Park Avenue South offices of Rizzoli as he looked at the nearly finished galleys for an October release called Pharrell: Places and Spaces I've Been. Here was a lavish coffee-table book filled with images of the many products he has designed in collaboration with other artists and fashion designers, and interviews between Mr. Williams and the likes of Jay-Z, Anna Wintour and Zaha Hadid, which do not exactly lend themselves to the narrative treatment. But why not?

Question: Who sat on the chair designed by Domeau & Pérès?
Groundtruth: Pharrell Williams



Passage: OSLO When the Henie Onstad art museum near here received an unexpected call in June 2012 on behalf of the heirs of a French art dealer, the director, Tone Hansen, had no idea that it was the beginning of an odyssey that would end with the museums giving away one of its prize paintings and opening a major exhibition built around that work's absence. Visitors to the exhibition, In Search of Matisse, which runs until Dec. 13, will not be seeing Blue Dress in a Yellow Armchair, which the artist painted in 1937 and had been in the museum's possession. In March 2014, the painting was returned to the heirs of the French art dealer, Paul Rosenberg.

Question: What was the name of this painting that was returned to Paul Rosenberg's heirs in 2014?
Groundtruth: Blue Dress in a Yellow Armchair



Passage: I didn't install controls for light switches, blinds or fireplace because it can get pricey (Control4 switches cost \$180 each), but they are also on a totally different level of sophistication. But I did set up a Control4 entertainment system, which proved to be quite useful, though not cheap. The grand total for this setup, including a Control4 hub (\$750), the Control4 Door Station (\$800) and a touch-screen controller (\$900), was about \$3,000 including installation. Was it worth it? Well, a few weeks ago my mother-in-law was babysitting and couldn't figure out how to find her television show, so I logged in with my phone, turned on Homeland and set the volume so it wasn't too loud. And guess where I was? Sitting on an airplane.

Question: What is the cost of this device?
Groundtruth: \$900



Passage: Beyond this North-South bifurcation, the earliest paintings here also reflect the extraordinary variety of artistic heritages Bearden was exposed to as a young man. Paintings like Untitled (Harvesting Tobacco) (circa 1940) and Untitled (Husband and Wife) (circa 1941) feature large, full-bodied figures like those of George Grosz, the German-born migrant painter who was Bearden's mentor at the Art Students League in the early 1930s and who encouraged Bearden to look at Dutch and Flemish painting. The Family (circa 1941), with its horizontal orientation and multiple figures, is reminiscent of the Mexican muralists, who were influential in New York in the 1930s. Presage (1944) displays the fragmentation of forms that was the hallmark of Cubism. Bearden started working as a political cartoonist; it was Grosz who persuaded him to switch to fine art and the paintings here were created around the same time that he wrote essays like The Negro Artist and Modern Art (1934) and The Negro Artists Dilemma (1946).

Question: What is the title of this piece at the Newark Museum?
Groundtruth: Untitled (Harvesting Tobacco)



Passage: MONTREAL When Germany, the world's top team, meets the United States, the world's second-ranked squad, in a Women's World Cup semifinal game Tuesday, there will be a matchup of goalkeepers befitting the stakes. On one end will be Hope Solo, whom many, particularly those on her team, consider the best keeper in the world. On the other side will be Nadine Angerer, who two years ago became the first goalkeeper to win FIFA's World Player of the Year award.

Question: Who is she?
Groundtruth: the first goalkeeper to win FIFA's World Player of the Year award



Passage: What's on TV EUPHORIA (2019) 10 p.m. on HBO. Drugs, sex, love and social media fuel this modern, anxiety-ridden teenage drama centered around Rue Bennet (Zendaya), a 17-year-old who is searching for connection after leaving rehab. Rue and her orbit of young friends and classmates struggle to find their identities and sense of purpose within a haze of substance-fueled parties, preparation for college and a hefty amount of family and personal trauma. Sam Levinson (The Wizard of Lies) wrote the eight episodes of the American adaptation of the Israeli series, which includes not only Zendaya, a former Disney Channel star, as the cast lead, but also Drake as an executive producer.

Question: What is the name of the model and actress who stars in the American version of Euphoria?
Groundtruth: Zendaya



Passage: Donald J. Sobol (his parents gave him the middle initial, but it did not stand for anything, his son said) was born on Oct. 4, 1924, in the Bronx. His father owned gas stations, which he later sold to Standard Oil. Mr. Sobol graduated from the Ethical Culture School in Manhattan in 1942 and enrolled at Oberlin College. In the middle of his freshman year he enlisted in the Army and served during World War II as a sergeant in a combat engineer battalion in the Pacific. He returned to Oberlin in 1946 and later gave much of the credit for his career to an English professor there, John Singleton, who gave him a personal course in advanced creative writing.

Question: What was the name of the creator?
Groundtruth: Donald J. Sobol

Figure 7: Manually annotated NewsMQA Samples.



Passage: "The Terminator" (directed by James Cameron, 1984)"This was my favorite movie for ages. It let the characters and emotion lead the story rather than the tech, which was so rare in sci-fi. The film is also a great primer on Cameron's cutting, action and mood." "Cyborg" (directed by Albert Pyun, 1989)"I saw this as a kid and it was just nuts. It had a really snappy percussive tempo to the fight sequences — that was the Asian influence coming in. They also were creative with how to make the action cool without a huge effects budget."

Question: What movie did Vincent Klyn star in?

Groundtruth: Cyborg



Passage: The reality-show villain has evolved. Ladies and gentlemen, meet the YouTube villain. Look no further than Jake Paul, a YouTube-generation successor to platinum-blond miscreants like Spencer Pratt of "The Hills" and Jonny Fairplay of "Survivor," mixed with a little bit of the "Jackass" crew.

Question: What is the name of the YouTuber who has taken a heel turn in recent months?

Groundtruth: Jake Paul



Passage: In some cases, more radical action might be necessary. One top auto analyst recommended on Thursday that General Motors, which has lost money in Europe for 12 straight years, sell or even close down its European operations before they undermine the company's overall comeback from its government bailout and bankruptcy. "One of the worst things in the auto industry is owning a cash burning, resource consuming business," the analyst, Adam Jonas of Morgan Stanley, wrote in a lengthy research report about G.M.'s Opel unit. "One of the best things is getting rid of that business."

Question: What brand did G.M. want to leave Europe?

Groundtruth: Opel



Passage: ANCHORAGE — An independent candidate, Bill Walker, has won the Alaska governor's race after an aggressive challenge to Gov. Sean Parnell, a Republican. Mr. Walker, a lawyer and former mayor of Valdez, was ahead on election night, but the race was too close to call until thousands of absentee and questioned ballots were counted this week. He is the first candidate unaffiliated with a party since statehood to win the office.

Question: What is the name of the candidate who won the governor's race in Alaska?

Groundtruth: Bill Walker



Passage: Mr. Goodman, 61, opened Trash and Vaudeville in 1975, in what was formerly Limbo, a hippie clothing shop. He was a recent graduate of the Fashion Institute of Technology and borrowed \$1,000 from his parents for the down payment. He named it Trash and Vaudeville because people often asked, "What's that trash you're wearing?" For years, Mr. Goodman spent his days "rag picking," searching through warehouses of clothes from previous decades for music-inspired looks: Hawaiian shirts, gabardine shirts, go-go boots.

Question: What was the name of the clothing store that left its East Village home?

Groundtruth: Trash and Vaudeville



Passage: The Golden State Warriors followed up their second consecutive N.B.A. title — and third in four years — by adding DeMarcus Cousins, the four-time All-Star. It became commonplace to declare that they had ruined the league by removing any suspense to the season. A starting lineup of five All-Stars, two of which are former recipients of the Most Valuable Player Award, was just too much. There was palace intrigue, however. Kevin Durant and Draymond Green feuded in public, Cousins was routinely exposed on defense, Klay Thompson had a rough start and Stephen Curry's shot seemed to fail him before he began wearing contacts on the court.

Question: What is the name of the greatest 3-point shooter in N.B.A. history?

Groundtruth: Stephen Curry



Passage: "Fast Color" is many things — a dreamlike dystopian drama; a warm celebration of family and female power; a teasing hint of superhero-origin story — none of which fully explains its gentle grip on our attention. What it is not, is speedy, with a narrative that moves carefully, even languidly, and visuals so beautifully patient that we have time to memorize their texture and import. Special effects are all the more so for being used sparingly, and the movie's few characters feel no need to be constantly explaining themselves. For fairly long stretches, there are no major incidents — which is not the same as saying that nothing happens. Set in a near-future America parched and wilting from eight years of drought, the movie drops us into a car beside an anxiously fleeing Ruth

Question: What is the name of the character in Fast Color?

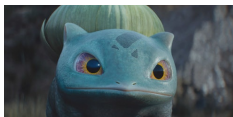
Groundtruth: Gugu Mbatha-Raw



Passage: These supporters argue machine intelligence will produce so much economic surplus that we could collectively afford to liberate much of humanity from both labor and suffering. The most idealistic thinkers see the plan as a way to foster the sort of quasi-utopian future we've only encountered in science fiction universes like that of "Star Trek." As computers perform more of our work, we'd all be free to become artists, scholars, entrepreneurs or otherwise engage our passions in a society no longer centered on the drudgery of daily labor.

Question: What is the Atlas Robot?

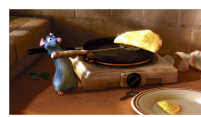
Groundtruth: machine intelligence



Passage: A gaggle of these friendly, bulbous creatures come to Pikachu's rescue, falling all over one another. "We took a series of clips of bulldog puppies jumping on top of each other and copied their movements with a roscope," Nordby said. "The playfulness and top-heaviness you have with bulldogs, where they don't quite seem like they're ever quite comfortable on their feet, worked really well with the body shape of the bulbasaur." Snubbull

Question: What was the inspiration for Sonic the Hedgehog?

Groundtruth: bulldog puppies



Passage: He is also, at least implicitly, a severe critic of the laziness and mediocrity that characterize so much popular culture. He criticizes partly by example, by avoiding the usual kid-movie clichés and demonstrating that a clear, accessible story can also be thoughtful and unpredictable. "Ratatouille" features no annoying sidekick and no obtrusive celebrity voice-work, and while Remy is cute, he can also be prickly, demanding and insecure. Moreover, his basic moral conflict — between family obligation and individual ambition — is handled with unusual subtlety and complexity, so that the reassurances and resolutions of the movie's end feel earned rather than predetermined.

Question: What is the rat with the palate in "Ratatouille"?

Groundtruth: Remy

Figure 8: Synthetic NewsMQA samples.

