

MuMuQA: Multimedia Multi-Hop News Question Answering via Cross-Media Knowledge Extraction and Grounding

Revanth Gangi Reddy¹, Xilin Rui², Manling Li¹, Xudong Lin³, Haoyang Wen¹, Jaemin Cho⁴, Lifu Huang⁵, Mohit Bansal⁴, Avirup Sil⁶, Shih-Fu Chang³, Alexander Schwing¹, Heng Ji¹

¹ University of Illinois Urbana-Champaign ² Tsinghua University ³ Columbia University

⁴ University of North Carolina at Chapel Hill ⁵ Virginia Tech ⁶ IBM Research AI

{revanth3, hengji}@illinois.edu

Abstract

Recently, there has been an increasing interest in building question answering (QA) models that reason across multiple modalities, such as text and images. However, QA using images is often limited to just picking the answer from a pre-defined set of options. In addition, images in the real world, especially in news, have objects that are co-referential to the text, with complementary information from both modalities. In this paper, we present a new QA evaluation benchmark with 1,384 questions over news articles that require *cross-media grounding* of objects in images onto text. Specifically, the task involves multi-hop questions that require reasoning over image-caption pairs to identify the grounded visual object being referred to and then predicting a span from the news body text to answer the question. In addition, we introduce a novel multimedia data augmentation framework, based on cross-media knowledge extraction and synthetic question-answer generation, to automatically augment data that can provide weak supervision for this task. We evaluate both pipeline-based and end-to-end pretraining-based multimedia QA models on our benchmark, and show that they achieve promising performance, while considerably lagging behind human performance hence leaving large room for future work on this challenging new task.¹

1 Introduction

To answer questions, humans seamlessly combine context provided in the form of images, text, background knowledge or structured data such as graphs and tables. Availability of a similarly capable question answering (QA) model could increase information accessibility and allow quick understanding of news or scientific articles, where the main narrative is enriched by images, charts and graphs with captions.

However, most QA research to date focuses on extracting information from a single data modality, such as text (TextQA) (Rajpurkar et al. 2016; Kwiatkowski et al. 2019), images (VQA) (Antol et al. 2015; Goyal et al. 2017) or videos (VideoQA) (Yang et al. 2003). Only recently, there have been initial attempts to combine information from multiple modalities (Kembhavi et al. 2017; Lei et al. 2018), which often use questions in a multiple-choice setting. However,

multiple choice questions in VQA have been shown to have explicit bias (Kazemi and Elqursh 2017; Manjunatha, Saini, and Davis 2019), which can be taken advantage of by using question understanding or answer option analysis.


To address this concern, more recently, Talmor et al. (2021) introduce an extractive multi-hop QA dataset that involves reasoning across tables, text and images. However, each image is associated with a Wikipedia entity, therefore the reasoning over images essentially boils down to ranking the images based on the question and using the entities corresponding to the top-ranked images. In contrast, questions about images in real life require cross-media reasoning between images and text. Current multimedia benchmarks cannot be used to solve this as they do not require to ground images onto text when answering questions about images.

To bridge this gap, we introduce a new benchmark QA evaluation task called ‘Multimedia Multi-hop Question Answering’ (MUMUQA) along with a corresponding dataset of news articles. To succeed in this new MUMUQA task, methods need to excel at both identifying objects being referred to in the question and leveraging news body text to answer the question. Specifically, the task is formulated as multi-hop extractive QA where questions focus on objects grounded in image-caption pairs. To enforce multi-hop reasoning, we ask information-seeking questions about objects in the image by referring to them via their visual attributes. The requirement for grounding between image and text along with multi-hop reasoning makes MUMUQA much more challenging than the tasks which are commonly explored in the QA community so far.

Figure 1 illustrates the task and the data. Answering the question in Figure 1a requires first identifying that “the person with the blue tie” is on the right in the image, which needs to be grounded to “Benjamin Netanyahu” (in red) in the caption. Subsequently, the news body text is needed to extract the answer. Grounding between image-caption pairs along with textual understanding and coreference resolution is hence crucial. Note that the question cannot be answered by a text-only model, since the text mentions two parties (green and yellow highlight) that correspond to the two people in the image-caption pair (names in red and brown). Figure 1b is another example, with the image necessary to disambiguate that “people in the image” refers to *opposition supporters* and not *security forces*.

Image - Caption	Body Text
	A dispute between Israeli Prime Minister Benjamin Netanyahu and his finance minister over broadcast regulation sparked speculation on Sunday that Netanyahu could seek an election two years ahead of schedule.
Israeli Prime Minister Benjamin Netanyahu (R) speaks with Finance Minister Moshe Kahlon during the weekly cabinet meeting in Jerusalem	... The Israeli media quoted Netanyahu as telling ministers from his Likud party that he would dissolve the government if Kahlon didn't fall into line. Kahlon heads the Kulanu party, a center-right partner in Netanyahu's ...
Question: What party does the person with the blue tie in the image belong to? Answer: Likud	

(a)

Image - Caption	Body Text
	Venezuelan security forces fired scores of tear gas volleys and turned water cannons on rock-throwing protesters on a bridge in Caracas on Wednesday as the death toll from this month's anti-government unrest hit at least 29.
Opposition supporters clash with security forces during a rally against Venezuela's President Nicolas Maduro in Caracas, Venezuela, April 26, 2017.	Red-shirted supporters of Maduro, the 54-year-old former bus driver who succeeded Hugo Chavez in 2013, also rallied on the streets of the capital, punching their fists in the air and denouncing opposition "terrorists."
Question: What are the people in the image accused of behaving like? Answer: terrorists	

(b)

Figure 1: Two examples from our evaluation benchmark with the question-answer pairs and their corresponding news articles. The bridge item, which needs to be grounded from image onto text, is shown in red and the final answer is marked in green in the news body text.

To study MUMUQA, we release an evaluation set with 263 development and 1,121 test examples. Annotators were asked to identify objects grounded in image-caption pairs and come up with questions about them which can be answered by the news body text. Given the high cost of annotating such examples, we use human-curated examples only as the evaluation set and develop a novel multimedia data augmentation approach to automatically generate silver-standard training data for this task, which can then be used to train or fine-tune state-of-the-art (SOTA) vision-language models (Tan and Bansal 2019; Li et al. 2020c). Specifically, we generate silver training data by leveraging multimedia knowledge extraction, visual scene understanding and language generation. In short, we first run a state-of-the-art multimedia knowledge extraction system (Li et al. 2020a) to capture the entities that are grounded in image-caption pairs, such as “Benjamin Netanyahu”. Next, we apply a question generation approach (Shakeri et al. 2020) to automatically generate questions from the news body text, that are conditioned on one of the grounded entities, such as “What party does Benjamin Netanyahu belong to?”. Then, we use a visual attribute recognition model (Lin et al. 2019) to edit these questions to refer to the grounded entity by its visual attributes, such as replacing “Benjamin Netanyahu” with “the person with the blue tie” in the question. Finally, we filter out questions that are answerable by a single-hop text-only QA model (Chakravarti et al. 2020). This pipeline automatically generates training data for our task.

We evaluate both text-only QA and multimedia QA models on our MUMUQA benchmark. Particularly, we explore both pipeline-based and end-to-end multimedia QA approaches. The pipeline-based model first decomposes the task into image-question and text-question. Then, it uses a SOTA multimedia knowledge extraction model (Li et al. 2020a) for visual entity grounding and attribute matching to answer the image-question and uses a single-hop text QA model (Chakravarti et al. 2020) to extract the final answer. For the end-to-end multimedia QA system, we directly fine-tune a SOTA visual-language pretrained model (Li et al.

2020c) using auto-generated silver-standard training data.

The contributions of this work are as follows:

- We release a new extractive QA evaluation benchmark, MUMUQA. It is based on multi-hop reasoning and cross-media grounding of information present in news articles. To the best of our knowledge, our work is the first to attempt using information grounded in the image in an extractive QA setting (see Section 2 and Section 3).
- To automatically generate silver-standard training data for this task, we introduce a novel pipeline that incorporates cross-media knowledge extraction, visual understanding and synthetic question generation (Section 4).
- We measure the impact of images in our benchmark by evaluating competitive text-only QA models on our task, and demonstrate the benefit of using multimedia information (see Section 5).

2 MUMUQA Task

In this section, we present the details of our Multimedia Multi-hop Question Answering (MUMUQA) task. As illustrated in Figure 1, given a news article with an image-caption pair and a question, a system needs to answer the question by extracting a short span from the body text. Importantly, answering the questions requires multi-hop reasoning: the first hop, referred to as *image entity grounding*, requires cross-media grounding between the image and caption to obtain an intermediate answer, named *bridge item*, for the image-related question; and the second hop requires reasoning over the news body text by using the bridge item to extract a span of text as the final answer. For example, in Figure 1a, with the first hop, we need to ground “person with the blue tie in the image” to the particular entity “Benjamin Netanyahu” in the caption. Taking “Benjamin Netanyahu” as the bridge item to the second hop, we further extract the final answer as “Likud” from the news body text. The questions require using information present in the image for entity disambiguation, thereby needing cross-media grounding.

Our benchmark reflects questions that news readers might

have after looking at the visual information in the news article, without having read the relatively longer body text. News articles usually have objects that are mentioned in both images and text, thereby requiring cross-media grounding to answer questions about them. In our task, we follow the ordering of using the visual information first, and then pick answer from the news body text so as to allow for a wide range of answers. In contrast, following the other reasoning order would require answers to come from images, which previous VQA work has shown to be restricted to a pre-defined vocabulary. We use multi-hop questions to enforce the constraint of using information from different modalities, which also follows recent work on multimedia QA (Talmor et al. 2021), that however does not require any cross-media grounding. Prior work (Yang et al. 2018; Welbl, Stenetorp, and Riedel 2018) has also used multi-hop questions as a means to evaluate such complex reasoning chains.

3 Benchmark Construction

To facilitate progress on this multi-hop extractive MU-MUQA task, we collect a new dataset. Our dataset consists of an evaluation set that is human-annotated and a silver-standard training set that is automatically generated (described later in Section 4). We choose to manually construct an evaluation set of high quality, to mimic the information-seeking process of real newsreaders. We first provide a brief description of the news articles in our dataset (Section 3.1) and then detail how the evaluation set (Section 3.2) was created, along with a brief analysis (Section 3.3).

3.1 News Data Acquisition

We take 108,693 news articles (2009-2017) from the Voice of America (VOA) website², covering a wide array of topics such as military, economy and health. We use articles from 2017 for annotating the evaluation set and articles from 2009 to 2016 for creating the training set.

3.2 Evaluation Set Construction

To evaluate models, we collect data via a newly developed and specifically tailored annotation interface³. News articles are shown in the interface along with their images and corresponding captions and annotators are asked to come up with questions. The annotation process requires the annotator to first look at the image-caption pair to identify which objects in the image are grounded in the caption and to choose one of them as the bridge item. The annotators then study the news body text to look for mentions of the bridge item and pick a sentence for which the question will be created. Finally, annotators create a multi-hop question with the answer coming from the news body text and the bridge item being referred to in the question by its visual attributes.

To ensure a real need for multimedia reasoning, we provide annotators with access to a publicly available single-hop text-only QA model (Chakravarti et al. 2020). Specifi-

cally, annotators are asked to ensure that the text-only model cannot pick up the answer by looking at only the question and the news body text. We also ensure that the annotators do not attempt to ‘beat’ the text-only model by formulating questions that are too complex or convoluted. For this, we require the answer to appear in the top-5 text-only model answers when the image reference is removed from the question, by replacing it with the *bridge item*. Additionally, note that the annotators are specifically asked to not use their own background knowledge of the people in the images when trying to ground them into the caption, i.e., grounding is intended to be purely based on visual or spatial attributes, without any need for explicit face recognition.

Our annotators are native English speakers and have had extensive experience annotating for other NLP tasks. On average, they took 3.5 minutes to come up with the question and annotated one in every three news articles they saw (i.e., they chose to skip the other two).

For quality control, we ask a different set of senior annotators to validate all examples in the evaluation set. Specifically, we ask to check that the image reference is grounded in the caption, questions are answerable and unambiguous. Kwiatkowski et al. (2019) have shown that aggregating answers from multiple annotators is more robust than relying on a single annotator. Following them, we use an Amazon Mechanical Turk (Buhrmester, Kwang, and Gosling 2016) task where crowd-workers were asked to answer the question and also provide the bridge item, given the news article and the corresponding image-caption pair. We obtain one more answer per question this way, which we combine with the original answer to get 2-way annotations for each example in the evaluation set. The Turkers on average needed 2.5 minutes to answer a question.

3.3 Dataset Analysis

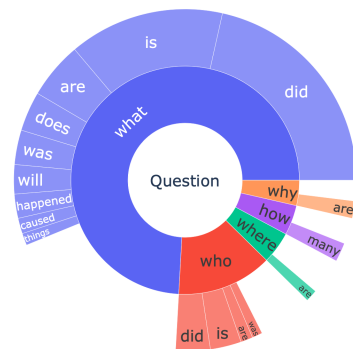


Figure 2: Distribution of questions for 5 most common first words and the subsequent second words.

The evaluation benchmark contains 1384 human-annotated instances, with 263 instances in the development set and the remaining in the test set. Figure 2 shows the sunburst plots from the analysis of the most common words in each question. We see that “what” is the most common question type, which is similar to other QA datasets (Lewis et al. 2020b; Hannan, Jain, and Bansal 2020).

²www.voanews.com

³Screenshot of the interface and annotation instructions are available in the appendix. The annotation tool will be made publicly available.

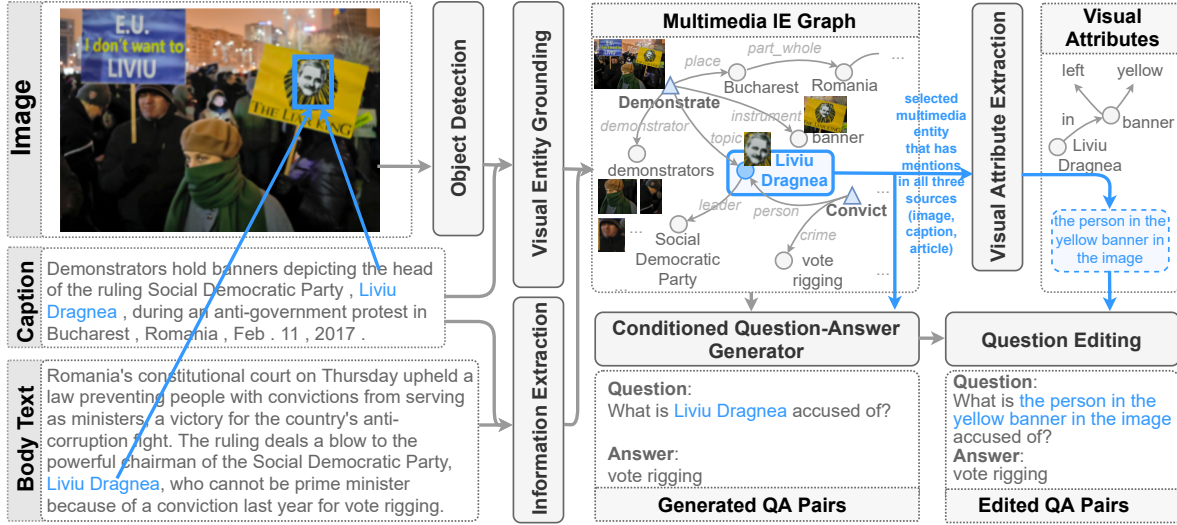


Figure 3: The workflow of synthetic question generation for creating the training set.

We also analyze the entity type of the bridge item, which is the answer to the part of the multi-hop question corresponding to the image, i.e., the *image-question*. We see that the majority of questions are about people (69%), with significant questions referring to locations (10%), organizations (10%), nationalities or political groups (7%) and other types (4%). Also, we find that the bridge item can be found in the caption 89% of the time and it is present in the news body text in the remaining cases.

We compare the total answer vocabulary for MUMUQA with VQA and also vocabulary size per question to account for difference in dataset sizes. We see that VQA, with 214k questions and 1.5 words per answer on average, has a vocabulary size of 26k (0.12 per question). MUMUQA, with 1384 questions and 6 words per answer on average, has a vocabulary size of 3k (2.18 per question).

4 Silver Training Set Generation

Given the cost and complexity associated with creating question-answer pairs required for MUMUQA, we use human-curated data only for the evaluation set. However, to support end-to-end training of models for this task, we create silver-standard training data using a novel multimedia data augmentation pipeline to generate multi-hop questions that require cross-media grounding. An overview of our generation pipeline is detailed in Section 4.1, which consists of using multimedia entity grounding (Section 4.2), visual understanding (Section 4.3) and conditioned question generation (Section 4.4) with question editing and filtering (Section 4.5) to obtain questions that are multi-hop and require cross-media grounding to answer them.

4.1 Training Data Generation Overview

Figure 3 shows our training data generation framework. At a high level, we intend to automatically generate training data that contains questions about entities that are grounded in the

image-caption pairs (e.g., “Liviu Dragnea”), with the answer coming from the news body text, such as “vote rigging”.

First, we perform multimedia entity grounding on the image-caption pairs to identify objects in the image which are grounded in the captions, to obtain the grounded entities, such as “Liviu Dragnea” in Figure 3. We extract the visual attributes for the grounded entities by running the visual attribute extraction system on their bounding boxes, e.g., “yellow, banner” in Figure 3. It enables us to generate a description “the person in the yellow banner in the image” for the grounded entity “Liviu Dragnea”.

Next, we generate the questions for the grounded entities, such as “What is Liviu Dragnea accused of?”. We first run a state-of-the-art knowledge extraction system (Li et al. 2020a) on the caption and body text to identify mentions of the grounded entities in the body text, such as “Liviu Dragnea” and “chairman” in the body text. It enables us to find candidate context for question generation, which we feed with the grounded entity e into the synthetic question generator to get a question-answer pair (q, a) . We ensure that the generated question has a mention of the grounded entity e in its text. Then, we edit these questions to replace the grounded mention by its corresponding visual attributes, to produce the final multi-hop question such as “What is the person in the yellow banner in the image accused of?”.

4.2 Multimedia Entity Grounding

We ground each entity in the text to a bounding box in the image using a multi-layer attention mechanism following Akbari et al. (2019). The system⁴ extracts a multi-level visual feature map for each image in a document, where each visual feature level represents a certain semantic granularity, such as *word*, *phrase*, *sentence*, etc. In addition, each entity is represented using contextualized embeddings (Peters et al. 2018), and we compute an attention map to every visual fea-

⁴<https://hub.docker.com/r/gaiaaaida/grounding-merging>

ture level and every location of the feature map. In this way, we can choose the visual feature level that most strongly matches the semantics of the entity, and the attention map can be used to localize the entity. The model is fine-tuned on the VOA news that we collected.

4.3 Visual Attribute Extraction

By analyzing the object detection results, we observe that on the development set, about 75% of images contain at least one person that can be grounded to its text mentions. Therefore, describing the attributes of a person is of great importance. Given the object detection results from the grounding step, we explore three types of attributes to comprehensively describe people in images: spatial attributes, personal accessories, and personal attributes. The spatial attribute is the relative position of the person. We associate personal accessories from object detection results to the person bounding boxes. To obtain personal attributes, we use a widely-used person attribute recognition model⁵ (Lin et al. 2019). More details are provided in the supplementary material.

4.4 Conditioned Question Generation

Given a passage in the news body text and an entity present in that passage, we aim to generate a synthetic question about that entity using the passage. Specifically, we train a synthetic example generator to take a passage p , an entity e and generate a question q and its corresponding answer a . To achieve this, we fine-tune BART (Lewis et al. 2020a), an encoder-decoder based model, using Natural Questions (NQ) (Kwiatkowski et al. 2019), an existing machine reading comprehension dataset. This dataset appears in the form of (q, p, a) triples. We identify the entity e in the question, that also appears in the passage. This entity is passed as input along with the passage to condition the question generation.

4.5 Question Editing and Filtering

We edit the questions by replacing the grounded entity’s text mention in the question with their visual attributes. This step aims to incorporate the need for using the image to do entity disambiguation. Finally, to filter out questions that are answerable directly using the text, we apply a cycle-consistency filter similar to Alberti et al. (2019), by using a single-hop text-only QA model (Chakravarti et al. 2020). This step discards questions that can be answered using text-only cues, so as to ensure that information from the image is required for answering the question. Since synthetic questions can be noisy, we also discard questions for which the answer does not appear in the top-5 answers to ensure quality of questions.

5 Question Answering Models

In this section, we describe various competitive baseline methods for evaluation on our benchmark. We begin with a multi-hop text-only extractive QA baseline and then proceed to use both pipeline-based and end-to-end multimedia QA baselines.

⁵<https://github.com/hyk1996/Person-Attribute-Recognition-MarketDuke>

5.1 Multi-hop Text-only QA

The multi-hop text-only QA model takes the question, caption and body text as input, and processes each paragraph in the body text independently, similar to Alberti, Lee, and Collins (2019). The QA model has an extractive answer predictor that predicts the indices of answer spans on top of the output representations \mathbf{H} from pre-trained language models (LM) (Devlin et al. 2019). Specifically, the text-only QA model trains an extractive answer predictor for the beginning b and ending e of the answer span as follows: $\alpha_b = \text{softmax}(\mathbf{W}_1 \mathbf{H})$ and $\alpha_e = \text{softmax}(\mathbf{W}_2 \mathbf{H})$, where $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{1 \times D}$ and D is the LM’s output embedding dimension and T is length of input text. The loss function is the averaged cross entropy on the two answer predictors:

$$\mathcal{L} = -\frac{1}{2} \sum_{t=1}^T \{1(\mathbf{b}_t) \log \alpha_b^t + 1(\mathbf{e}_t) \log \alpha_e^t\}.$$

5.2 End-to-End Multimedia QA

Following recent progress in building pre-trained multimedia models (Tan and Bansal 2019; Chen et al. 2020b; Li et al. 2020c), we experiment with finetuning an end-to-end multimedia QA model for our task. Specifically, we use OSCAR (Li et al. 2020c) which learns cross-media representations of image-text pairs with object tags from images added as anchor points to significantly ease the learning of alignments. OSCAR has shown SOTA performance in multiple vision-language tasks like VQA (Goyal et al. 2017), image-text retrieval, image captioning (You et al. 2016; Agrawal et al. 2019) and visual reasoning (Suhr et al. 2019).

We finetune OSCAR using the synthetic data that we generated in Section 4. We add an extractive answer predictor on top of the outputs from OSCAR to predict the start and end offset of the final answer from the body text. The classifier is trained with cross-entropy loss.

To obtain the image features and the object labels, we use a publicly available implementation⁶ of Faster-RCNN feature extraction tool (Ren et al. 2015; Anderson et al. 2018).

5.3 Pipeline-based Multimedia QA

An overview of our pipeline-based baseline for multi-hop multimedia QA is shown in Figure 4. First, we split a multi-hop question into a question referencing the image, referred to as *image-question*, and a question about the text, referred to as *text-question*. To achieve this, we use a multi-hop question decomposition model (Min et al. 2019). For example, in Figure 4, the question “What did the person with the red coat in the image talk about?” is decomposed to “Who is the person with the red coat in the image” and “What did [ANSWER] talk about”, where [ANSWER] denotes the answer of the first question. We take the first question as *image-question* and the second one as *text-question*.

Next, we find a bounding box that can answer the *image-question*, such as the blue bounding box in Figure 4. In detail, we first obtain the bounding boxes and identify the visual attributes for these bounding boxes using the approach

⁶<https://github.com/airsplay/py-bottom-up-attention>

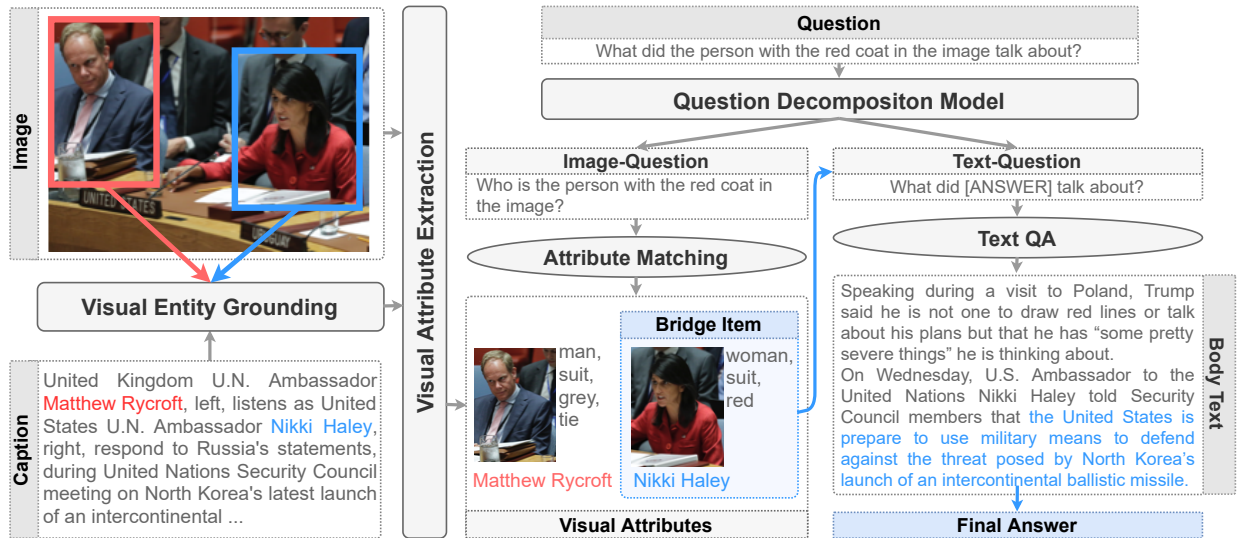


Figure 4: Overview diagram of the pipeline based multimedia QA approach.

described in section 4.3. We then match the image-question to the bounding boxes based on the similarity of their embeddings. The bounding box is represented as bag-of-words over its visual attribute classes, such as “woman, suit, red”; the question embedding is also represented as bag-of-words over the tokens in the image-question with stop words removed. We average the embeddings from FastText (Mikolov et al. 2018) and compute cosine similarity.

Then, we obtain a text span associated with the selected bounding box, such as “Nikki Haley” for the blue bounding box. Specifically, we use the cross-modal attentions between the bounding boxes and caption text spans from the grounding approach in section 4.2. We call this text span as *bridge item*. Finally, we insert the bridge item into the text-question and run it against the single-hop text-only QA model (Chakravarti et al. 2020), to get the final answer.

6 Experiments

6.1 Settings

The multi-hop text-only QA model is trained on HotpotQA (Yang et al. 2018), which is a multi-hop extractive QA dataset over Wikipedia passages. For the single-hop text-only QA model, which is used in filtering and the pipeline-based multimedia baseline, we follow (Chakravarti et al. 2020) by training first on SQuAD 2.0 (Rajpurkar, Jia, and Liang 2018) and then on Natural Questions (Kwiatkowski et al. 2019). Both models use bert-large-uncased (Devlin et al. 2019).

For the end-to-end multimedia QA model, we use OSCAR-large⁷ as the underlying architecture. We start with the pre-trained model provided by Li et al. (2020c). The model is then trained on 20k silver-standard training examples from Section 4.

⁷<https://github.com/microsoft/Oscar>

6.2 Results and Analysis

Table 1 provides the results of various baselines on the development and test sets of our MUMUQA benchmark. The human baseline performance is evaluated over 70 randomly sampled examples. We use the macro-averaged F_1 score of the final answer for evaluation. The benefit of incorporating the multimedia knowledge extraction can be seen from the strong performance of the pipeline-based multimedia QA system. Interestingly, we see that the end-to-end multimedia QA baseline underperforms the multi-hop text-only system. This could be due to the fact that OSCAR is pre-trained with image-caption pairs, which makes it potentially not suited for reasoning over larger text input (news body text in this case).

Model	Dev	Test
Multi-hop Text-only QA	18.5	16.5
End-to-end Multimedia QA	12.1	11.5
Pipeline-based Multimedia QA	33.9	30.8
Human Baseline	-	66.5

Table 1: F_1 Performance (%) of different baselines on the MUMUQA evaluation benchmark.

For the pipeline based multimedia baseline, we analyze performance of the two steps (i.e computing the bridge answer and final answer). We first compute the F_1 score for the intermediate answer, which we call bridge F_1 . We see that the pipeline-based multimedia QA system has a bridge F_1 of 33.5% and 29.8% on the dev and test sets respectively, whereas the human baseline has a considerably higher bridge F_1 of 78.8%. This shows that the models are considerably behind humans in identifying the grounded object being referred to in the question, thereby justifying our task. Next, we observe that when the correct bridge item is provided to the text question in the second step (i.e., when bridge F_1 =

100%), the system has a final F1 score of 51.1% on the dev set. This suggests that even when the right bridge item is provided, finding the final answer is not straightforward.

We analyze performance of individual components of the pipeline based QA baseline, specifically the grounding and attribute matching systems. First, we evaluate the coverage of the bridge item in the output of the grounding system. We see that in 45% of the cases, the bridge item is present in the canonical mention of the grounded entity. Figure 5 shows one such example where the grounding system was unable to capture the bridge item. Next, whenever the bridge item is captured in the grounding, we observe that the attribute matching system is able to pick the correct bridge item in 60% of the cases.



Figure 5: An example where the grounding system failed to capture the ground-truth bridge item (in green). The grounded entity is in blue in the caption and its corresponding bounding box is shown in blue in the image.

6.3 Training Set Analysis

In this work, we bypass the high cost of manually annotating examples by using a carefully designed pipeline to generate a silver-standard training set for our MUMUQA task. While such automatically generated training data is free and can be scaled to much bigger sizes at no cost, we cannot expect the silver-standard training data to match the quality of human-annotated data. We performed a human evaluation study on 100 examples randomly sampled from the automatically generated training set. We observed that 80% of questions required using image information, 59% had the correct bridge answer and had an average score of 2.6 (range of 1-3) for grammaticality. Out of the questions with the correct bridge answer, 64% had the correct final answer. Upon manual inspection of the incorrect cases, we attributed 46% of errors to the grounding system, 4% to the visual-attribute extraction and 33% to the conditioned-question generator, with the remaining from incorrectly identifying the image co-reference while editing the questions. From this study, we observe that the automatic training data has some noise, with errors arising from different components of the generation pipeline. While it is not economically feasible to annotate thousands of such examples from scratch, one potential direction is to have annotators correct these automatically generated examples to help improve the quality.

7 Related Work

Visual Question Answering (Antol et al. 2015) aims to find a natural language answer given a natural language ques-

tion and an image. Several datasets have been proposed, such as VQA (Agrawal et al. 2017; Goyal et al. 2017), DAQUAR (Malinowski and Fritz 2014), COCO-QA (Ren, Kiros, and Zemel 2015), VCR (Zellers et al. 2019), and PlotQA (Methani et al. 2020), all of which require answering questions about images. However, the questions are multiple-choice or the answers are from a predefined vocabulary with only the information in the image being sufficient to get the correct answer. Simple baseline methods that only use question understanding (Kazemi and Elqursh 2017) or answer option sentiment analysis (Manjunatha, Saini, and Davis 2019) have been proven to perform surprisingly well on datasets such as VQA (Antol et al. 2015) and VQA2.0 (Goyal et al. 2017). But they are unlikely to provide good answers for understanding complex events. Current QA datasets over news involve using just the news body text (Trischler et al. 2017). In contrast, our benchmark focuses on questions with informative answers that require reasoning across multiple data modalities.

Recently, there has been some interest in using information from multiple modalities for answering questions. The first step in this direction is ManyModalQA (Hannan, Jain, and Bansal 2020), which requires figuring out which modality to use when answering the question. However, all questions are still answerable using a single modality, without any need for cross-media reasoning. MultimodalQA (Talmor et al. 2021) extends it by using multi-hop questions that require cross-media reasoning, with each image linked to a Wikipedia entity. In contrast, our dataset requires grounding between the image-caption pairs to identify which objects in the image are being referred to in the question. Moreover, the images in MUMUQA, which are from news articles, capture real-world events and are hence more realistic.

Grounding text mentions to image regions has previously been explored via cross-media attention (Yeh et al. 2017; Tan and Bansal 2019; Li et al. 2020b,c) or learning of an optimal transport plan (Chen et al. 2020b,a). Different from general text phrases, cross-media entity coreference (Akbari et al. 2019; Li et al. 2020a) takes text knowledge extraction graph as input and ground the entity with graph context. We are the first to explore cross-media grounding in an extractive QA setting.

8 Conclusions and Future Work

We present a new QA task, MUMUQA, along with an evaluation benchmark for multimedia news understanding. The task is challenging in the requirement of cross-media grounding over images, captions, and news body text. We demonstrate the benefit of using multimedia knowledge extraction, both for generating silver-standard training data and for a pipeline-based multimedia QA system. The multimedia baselines are still considerably behind human performance, suggesting ample room for improvement. Future work will incorporate other forms of media in news, such as video and audio, to facilitate information seeking from more comprehensive data sources. Another direction is to infuse the end-to-end multimedia QA system with additional input from the grounding and visual attribute extraction systems.

Acknowledgement

We would like to thank Sean Kosman, Rebecca Lee, Kathryn Conger and Martha Palmer for their help on data annotations, and thank Prof. Ernest Davis (NYU) for insightful advice and feedback on our data set and paper. This research is based upon work supported in part by U.S. DARPA AIDA Program No. FA8750-18-2-0014 and U.S. DARPA KAIROS Program No. FA8750-19-2-1004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

- Agrawal, A.; Lu, J.; Antol, S.; Mitchell, M.; Zitnick, C. L.; Parikh, D.; and Batra, D. 2017. Vqa: Visual question answering. *International Journal of Computer Vision*, 123(1): 4–31.
- Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; and Anderson, P. 2019. nocaps: novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8948–8957.
- Akbari, H.; Karaman, S.; Bhargava, S.; Chen, B.; Vondrick, C.; and Chang, S.-F. 2019. Multi-level Multimodal Common Semantic Space for Image-Phrase Grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12476–12486.
- Alberti, C.; Andor, D.; Pitler, E.; Devlin, J.; and Collins, M. 2019. Synthetic QA Corpora Generation with Roundtrip Consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6168–6173.
- Alberti, C.; Lee, K.; and Collins, M. 2019. A bert baseline for the natural questions. *arXiv preprint arXiv:1901.08634*.
- Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6077–6086.
- Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. VQA: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.
- Buhrmester, M.; Kwang, T.; and Gosling, S. D. 2016. Amazon’s Mechanical Turk: A new source of inexpensive, yet high-quality data?
- Chakravarti, R.; Ferritto, A.; Iyer, B.; Pan, L.; Florian, R.; Roukos, S.; and Sil, A. 2020. Towards building a Robust Industry-scale Question Answering System. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, 90–101.
- Chen, L.; Gan, Z.; Cheng, Y.; Li, L.; Carin, L.; and Liu, J. 2020a. Graph optimal transport for cross-domain alignment. In *International Conference on Machine Learning*, 1542–1553. PMLR.
- Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020b. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, 104–120. Springer.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186.
- Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6904–6913.
- Hannan, D.; Jain, A.; and Bansal, M. 2020. ManyModalQA: Modality Disambiguation and QA over Diverse Inputs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 7879–7886.
- Kazemi, V.; and Elqursh, A. 2017. Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv preprint arXiv:1704.03162*.
- Kembhavi, A.; Seo, M.; Schwenk, D.; Choi, J.; Farhadi, A.; and Hajishirzi, H. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, 4999–5007.
- Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7: 453–466.
- Lei, J.; Yu, L.; Bansal, M.; and Berg, T. 2018. TVQA: Localized, Compositional Video Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1369–1379.
- Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020a. 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*, 7871–7880.
- Lewis, P.; Oguz, B.; Rinott, R.; Riedel, S.; and Schwenk, H. 2020b. MLQA: Evaluating Cross-lingual Extractive Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7315–7330.
- Li, M.; Zareian, A.; Lin, Y.; Pan, X.; Whitehead, S.; Chen, B.; Wu, B.; Ji, H.; Chang, S.-F.; Voss, C.; et al. 2020a. Gaia: A fine-grained multimedia knowledge extraction system. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 77–86.
- Li, M.; Zareian, A.; Zeng, Q.; Whitehead, S.; Lu, D.; Ji, H.; and Chang, S.-F. 2020b. Cross-media Structured Common Space for Multimedia Event Extraction. In *Proceedings of*

the 58th Annual Meeting of the Association for Computational Linguistics, 2557–2568.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020c. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, 121–137. Springer.

Lin, Y.; Zheng, L.; Zheng, Z.; Wu, Y.; Hu, Z.; Yan, C.; and Yang, Y. 2019. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 95: 151–161.

Malinowski, M.; and Fritz, M. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in neural information processing systems*, 1682–1690.

Manjunatha, V.; Saini, N.; and Davis, L. S. 2019. Explicit Bias Discovery in Visual Question Answering Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9562–9571.

Methani, N.; Ganguly, P.; Khapra, M. M.; and Kumar, P. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1527–1536.

Mikolov, T.; Grave, É.; Bojanowski, P.; Puhersch, C.; and Joulin, A. 2018. Advances in Pre-Training Distributed Word Representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Min, S.; Zhong, V.; Zettlemoyer, L.; and Hajishirzi, H. 2019. Multi-hop Reading Comprehension through Question Decomposition and Rescoring. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6097–6109.

Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. 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)*, 2227–2237.

Rajpurkar, P.; Jia, R.; and Liang, P. 2018. 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)*, 784–789.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392.

Ren, M.; Kiros, R.; and Zemel, R. S. 2015. Exploring models and data for image question answering. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, 2953–2961.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

Shakeri, S.; dos Santos, C.; Zhu, H.; Ng, P.; Nan, F.; Wang, Z.; Nallapati, R.; and Xiang, B. 2020. End-to-End Synthetic Data Generation for Domain Adaptation of Question Answering Systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5445–5460.

Suhr, A.; Zhou, S.; Zhang, A.; Zhang, I.; Bai, H.; and Artzi, Y. 2019. A Corpus for Reasoning about Natural Language Grounded in Photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6418–6428.

Talmor, A.; Yoran, O.; Catav, A.; Lahav, D.; Wang, Y.; Asai, A.; Ilharco, G.; Hajishirzi, H.; and Berant, J. 2021. Multi-ModalQA: Complex Question Answering over Text, Tables and Images. *arXiv preprint arXiv:2104.06039*.

Tan, H.; and Bansal, M. 2019. 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)*, 5103–5114.

Trischler, A.; Wang, T.; Yuan, X.; Harris, J.; Sordoni, A.; Bachman, P.; and Suleman, K. 2017. NewsQA: A Machine Comprehension Dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, 191–200.

Welbl, J.; Stenetorp, P.; and Riedel, S. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6: 287–302.

Yang, H.; Chaisorn, L.; Zhao, Y.; Neo, S.-Y.; and Chua, T.-S. 2003. VideoQA: question answering on news video. In *Proceedings of the eleventh ACM international conference on Multimedia*, 632–641. ACM.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2369–2380.

Yeh, R.; Xiong, J.; mei Hwu, W.; Do, M.; and Schwing, A. 2017. Interpretable and Globally Optimal Prediction for Textual Grounding using Image Concepts. In *Proc. Neural Information Processing Systems*.

You, Q.; Jin, H.; Wang, Z.; Fang, C.; and Luo, J. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4651–4659.

Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6720–6731.