MIMOQA: Multimodal Input Multimodal Output Question Answering

Anonymous NAACL submission

Abstract

Multimodal research has picked up significantly in the space of question answering with the task being extended to visual question answering, charts question answering as well as multimodal *input* question answering. However, all these explorations produce a *unimodal* textual output as the answer. In this paper, we propose a novel task - **MIMOQA** - **M**ultimodal **I**nput **M**ultimodal **O**utput **Q**uestion **A**nswering in which the output is also multimodal. Through human experiments, we empirically show that such multimodal outputs provide better cognitive understanding of the answers. We also propose a novel multimodal question-answering framework, **MExBERT**, that incorporates a joint textual and visual attention towards producing such a multimodal output. Our method relies on a novel multimodal dataset curated for this problem from publicly available unimodal datasets. We show the superior performance of **MExBERT** against strong baselines on both the automatic as well as human metrics.

1 Introduction

Multimodal content is at the heart of digital revolution happening around the world. While the term *modality* has multiple connotations, one of its common usage is to indicate the *content modality* i.e. images, text, audio etc. It has been shown that multimodal content is more engaging and provides better cognitive understanding to the end user (Dale, 1969; Moreno and Mayer, 2007; Sankey et al., 2010). With recent improvements in vision-language grounding and multimodal understanding (Bisk et al., 2020; Luo et al., 2020; Sanabria et al., 2018; Das et al., 2018), several works have explored beyond unimodal machine comprehension (Hermann et al., 2015; Kočiskỳ et al., 2018; Nguyen et al., 2016; Kwiatkowski et al., 2019) towards a holistic multimodal comprehension (Antol et al., 2015; Das et al., 2017; Anderson et al., 2018; Zhu et al., 2018; Goyal et al., 2017; Fayek and Johnson, 2020) with significant improvements.

However, all these explorations on multimodal understanding, question answering in particular, have limited their focus to *unimodal* outputs even with multimodal inputs. For example - *Visual Question Answering (VQA)* task takes a textual query and an image to produce a *textual* answer. The *multimodal* question answering tasks (Antol et al., 2015; Kafle et al., 2018; Lei et al., 2018) take *multiple input modalities*, but the output is limited to *text* only. Even the recently proposed *ManyModalQA* (Hannan et al., 2020) relies on multimodal understanding to produce a textual answer. These works implicitly assume that the textual answers can satisfy the needs of the query across multiple input modalities. We posit that such an assumption is not always true; while textual answer can address several queries, a multimodal answer almost always enhances the cognitive understanding of the end user; understanding the answer through visuals is faster and provides enhanced user satisfaction.

In this paper, we propose a new task, ***Multimodal Input Multimodal Output Question Answering*** (**MIMOQA**), which not only takes multimodal input but also answers the question with a multimodal output. Our key contributions are:
1) We introduce the problem of *multimodal input multimodal output question answering*. We establish the importance of such multimodal outputs in question-answering for enhanced cognitive understanding via human experiments.
2) We propose **MExBERT**, a novel multimodal framework for *extracting* multimodal answers to a given question and compare it against relevant strong baselines. Our proposed method includes a novel pretraining methodology and uses a *proxy* supervision technique for the image selection.
3) We curate a large dataset for the introduced prob-

lem by extending the MS-MARCO (Nguyen et al., 2016) and Natural Question (Kwiatkowski et al., 2019) datasets for multimodal outputs. We propose the use of different automatic metrics and conduct human experiments to show their effectiveness.

2 Multimodal Output

Multimodal output not only provides better understanding to the end user but also provides *grounding* to the actual answer. For e.g., the multimodal output for the question in Figure 1**(a)** aids in better comprehension of the answer, while also providing grounding to words like 'stick', 'knob'. In some cases, textual answer might even be insufficient, especially, for questions which seek explicit visual understanding (questions about colors, structures, etc). In such cases, existing systems apply image understanding on top of the images to arrive at a 'textual description' of the desired answer. While this might suffice in some cases, a multimodal output can almost always enhance the quality of such answers. In Fig. 1**(b)**, the textual answer is insufficient and gets completed only with the help of the final image-text combination.

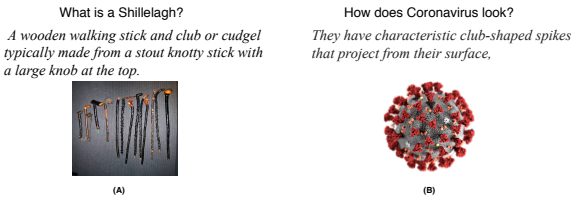


Figure 1: ***(a)**: The textual answer is sufficient but images provides better understanding, **(b)** The textual answer is insufficient and is completed by an image*

To verify the hypothesis, we collated 200 Question-Answer pairs (refer to supplementary for details); for each pair, we created its unimodal and multimodal answers. We conducted a human experiment where each question-answer pair was judged by 5 annotators; each annotator rating if the textual answer is sufficient for the input query. Irrespective of its sufficiency, the annotators were also asked whether the image in the multimodal variant enhances the understanding of the answer and adds value to it. To avoid the natural bias towards richer multimodal response in such experiments, we had explicitly inserted a few questions with irrelevant images (20%) and only considered the annotations which did not exhibit any bias in such questions.

Out of 80.27% of the total responses where the annotators felt that textual answers were sufficient, 87.5% felt the image enhanced their understanding even with such *sufficient* textual answer validating the importance of a multimodal answer. However, only 22.2% of the annotators felt the same when an irrelevant image was shown, indicating the absence of a strong bias towards richer responses. When the text was insufficient (19.73% of the responses), the relevant image boosted the understanding in 90.62% of the cases, further indicating that text only answers are not always sufficient and in such cases, an appropriate image can aid in better understanding. Here again, only 27.65% felt that an irrelevant image will add such a value, again indicating the lack of a strong bias towards multimodal answers just because they are richer. This experiment establishes that multimodal answers almost always improves the overall understanding irrespective of the sufficiency of textual answer. Motivated by this, we propose the novel problem of **multimodal input, multimodal output (MIMO) QA** - which attends to multiple modalities and provides responses in multiple modalities.

3 Multimodal Output QA

Formally, *given a piece of input text **T** along with a set of related images **I** and a query **Q**, our problem is to extract a multimodal answer **M** from {**I, T**}*. In an ideal case, multimodal answer does not have to be ***multi***-modal, especially when there is no relevant image in the input. However, for the sake of simplicity, we assume that there is at least one image in the input that can *complement* the textual answer even if it is not extremely critical for the textual answer to make sense. This follows our human experiments which showed that image adds value to the response over 90% of the time, irrespective of the sufficiency of the textual answers. Thus, our multimodal answer **M** consists of a text $\mathbf{M}_\mathrm{T}$ and an accompanying image $\mathbf{M}_\mathrm{I}$.

Multimodal Extractive BERT (MExBERT): As we show later, a major problem with independently extracting the textual answer and matching an image is the absence of joint understanding of visual and textual requirements for the query. We, therefore, propose a joint attention **M**ultimodal **Ex**tractive **BERT** based framework (**MExBERT**) using query **Q** over both input text **T** and input images **I**. Figure 2 shows the overall architecture of our proposed MExBERT framework. Inspired by the recent visuo-lingual models (Tan and Bansal, 2019; Lu et al., 2019; Chen et al., 2019), our frame-

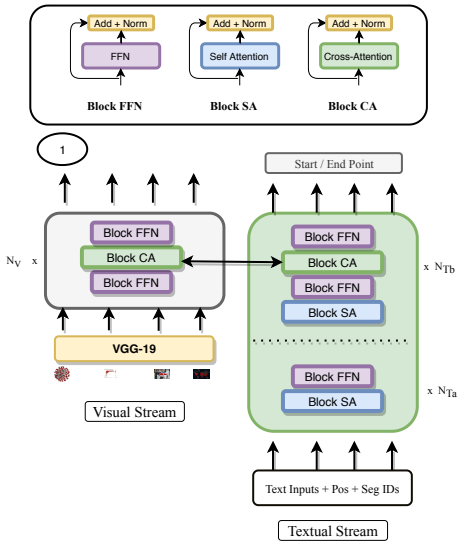


Figure 2: ***MExBERT***. *Details of the three blocks in the visual and textual streams is illustrated on the top. The visual stream takes the output of VGG-19 as input while the textual stream takes BERT Embeddings as input*

work has two separate streams - textual and visual stream; textual stream takes the query and input passage as input while visual stream takes the images as input.

The **textual stream** is extended from the BERT-QA framework (Devlin et al., 2018) and consists of self-attention transformer (Vaswani et al., 2017) layers. The input to the textual stream as shown in Figure 2 is tokenized BERT embedding of words in both passage and query. We also use the standard [CLS] and [SEP] tokens - the former prepended in the beginning and the latter embedded between query and the input passage. We use positional embedding to additionally provide positional and segment information for the MExBERT to better distinguish between query and passage. Unlike the the canonical BERT-QA, our textual stream employs two types of layers - regular *self-attention* layers and additional *cross-attention* layers. The initial layers of the textual stream include N_{T_a} regular self-attention based transformer layers similar to the canonical BERT-QA. The latter half of the textual stream is composed of N_{T_b} layers each of which consists of an additional cross-attention block along with the regular self-attention. Representing the attention computation in query-key-value format, the cross-attention block uses textual tokens as query and image representation from the visual stream as keys and values. This is different from self-attention where (query, keys and values) are all input textual tokens of the textual stream. The cross-attention block enables the framework to choose spans that are also coherent with the the visual stream. If the i^{th} textual token's features and j^{th} image's features used as input for k^{th} textual stream layer and $(k - N_{T_a})^{th}$ visual stream layer (as discussed later) are given by T^i_{k-1} and V^j_{k-1}; attention with q query, k keys, and v values is $attn(q, k, v)$, the self-attention and cross-attention is given by,

$$T^i_{k_{self}} = attn(T^i_{k-1}, T_{k-1}, T_{k-1}), \quad (1)$$

$$T^i_{k_{cross}} = attn(T^i_{k_{self}}, V_{k-1}, V_{k-1}) \quad (2)$$

where $T_k : \{T^0_k, ..., T^n_k\}$ and $V_k : \{V^0_k, ..., V^m_k\}$. Here, n is the number of textual tokens and m is the number of input images. The final layer of the textual stream is used to calculate the start and end position of the answer, similar to the canonical BERT-QA (Devlin et al., 2018) where one linear layer predicts the starting token and another layer predicts ending token through softmax applied over all tokens. The goal is to optimize the cross entropy loss over both the token position predictions.

The **visual stream** is similar to the textual stream with two key differences - **(i)** There is only one type of layer in the network and the number of layers $N_V = N_{T_b}$ and **(ii)** All the layers consist of only cross-attention blocks (along with feed-forward layers and residual connections) and do not contain self-attention block as shown in Figure 2. The self-attention was not used as the images mostly derive their relevance/context from the textual counterparts (powered by the cross-attention) in the input passage or query rather than other input images. The cross-attention is similar to the textual stream except that query is an image feature vector and the keys and values are textual tokens' representation from the corresponding textual stream layer. The input to the visual stream is the global VGG-19 (Simonyan and Zisserman, 2014) features of each of the images. We do not use positional/segment encodings in the visual stream. We use a linear head on top of visual features to predict whether a particular image should be in the output answer and use weighted binary cross-entropy for training where the weights w and $1 - w$ come from the proxy supervision values (as discussed later). The image with the highest confidence score on inclusion in the answer is regarded as the predicted image during inference.

Extract & Match: A natural framework to output a multimodal response would be to combine exist-

ing state-of-the frameworks in question answering and visuo-lingual understanding. To illustrate the shortcomings of such an assembled framework and motivate the need for a holistic framework, we implement such a framework using existing models as our *Extract & Match* baseline. Given the input query (**Q**) and the input text (**T**) and images (**I**), we first extract the textual answer using unimodal BERT-QA. (Devlin et al., 2018). We use this extracted answer, query, and input text to select an image from the input images using UNITER (Chen et al., 2019) to rank the images. UNITER has been trained on millions of image-text pairs for image-text matching task - the task of identifying whether a given image-text pair are actually the image and its caption. We provide each image along with the text (answer, query and input) to UNITER and use the classification confidence predicted by image-text matching head to rank the images. The image which receives the highest confidence score for a given text is taken as the matched output.

4 Dataset & Pretraining

Since there is no existing dataset which satisfies the requirements of the task, we curate a new dataset (refer to supplementary for details on curation strategy and data samples) by utilizing the existing public datasets. We observe that several QA datasets contain answers that come from a *Wikipedia article*. Since most Wikipedia articles come with a set of related images, such images could feature as the input **I** in our setup. Extending this heuristic, we use two QA datasets - MS-MARCO (Nguyen et al., 2016) and Natural Question (NQ) (Kwiatkowski et al., 2019), to extract those question-answer pairs which are originally extracted from Wikipedia and scrape all images from the original article.

Table 1 shows various statistics about the dataset. The dataset includes large number of images making the task of selecting appropriate image non-trivial. The variety of images also necessitates a robust visual and language understanding by our model. The passages have been formed by combining the answer source passage and randomly chosen $2-3$ 'distractor' passages from the original Wikipedia article. This allows the model to learn to find the right answer in unseen conditions also. The # of tokens in our input passages are large enough to be regarded to as a full input (instead of using the entire article) considering the focus here is on multimodal output and not article-passage ranking.

	# of pairs	Avg # of tokens	# of Images
Train	52,466	242.31	373,230
Development	722	180.62	3,563
Test	3,505	242.58	24,389

Table 1: *Statistics for the all three different splits of the curated MIMO Question Answering Dataset*

Proxy Supervision: Although we have scraped the images from the original articles, we do not have any supervision for these images in our dataset. We, therefore, develop *proxy* targets by using two types of information about the image - its position in the original article and its caption. We use the caption and position information only to obtain the target scores during training and not as an explicit input to our model since such information is not always readily available. Thus, our model is able to infer the correct multimodal response irrespective of the availability of such information at inference time. Since MS-MARCO and Natural Questions provide information about the original source passage for the final answer, we know the position of the source passage. We calculate the *proximity* distance **P** between the first token of source passage of answer and an image with number of tokens chosen as the distance unit. We, further, normalize this with the total number of tokens present in the entire article. We calculate the TF-IDF similarity of the caption against the Query, Answer and source passage (Figure 3). The overall supervision score is calculated as a weighted sum of these 4 scores where proximity score is calculated as $1 - P$. The normalized supervision scores (between $0 - 1$) are used as targets for linear layer of the visual stream.

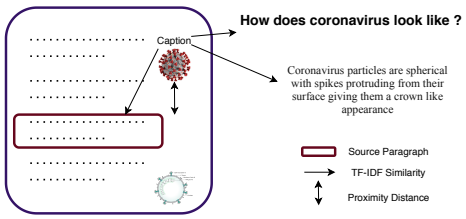


Figure 3: Calculation of proxy supervision scores

Pretraining: Vision and Language Tasks have relied on pretraining to address the complexities in building visuo-lingual relationships (Tan and Bansal, 2019; Lu et al., 2019; Chen et al., 2019). Following this, we leverage pretraining to better initialize our model. Further, our signals (even after including proxy supervision) are relatively sparse for a visuo-lingual task, calling for a stronger model initialization. We use Conceptual Captions

(Sharma et al., 2018) as it has been shown to impart a generic V-L understanding (Chen et al., 2019). We use the standard Masked Language Modelling (**MLM**) task over the Conceptual Captions to **pre-train the textual stream** and employ the cross entropy loss over the masked tokens. While the task is intended to train the textual stream, since the entire caption is generated from the visual information through the cross-attention mechanism, visual stream is also fine-tuned in this process. Since, our final model uses segment IDs, we randomly assign a segment ID of either query or passage to each caption during pretraining in order to imbibe language understanding for both type of tokens. For **pre-training the visual stream**, we modify the Conceptual Captions (Sharma et al., 2018) by choosing a random number between $(3-10)$ (**N**) for each caption followed by selecting **N-1** negative images (i.e. those images which have different captions) along with the image that is associated with the caption. We provide the caption as input to the textual stream and these **N** images as input to the visual stream. We train the model to predict the image corresponding to the caption by using binary cross entropy loss over images. Again, while this tasks is focused majorly on visual stream initialization, the textual stream is also fine-tuned due to the cross-attention layers between the two streams.

5 Experiments

We conduct extensive experiments and ablations for the proposed **MExBERT** framework and compare it against the **E&M** baseline. We divide our curated dataset into train, development and test sets as shown in Table 1. As mentioned before, we used the 3.2 million Image-Caption pairs from Conceptual Captions dataset (Sharma et al., 2018) for pretraining MExBERT layers. For proxy supervision, we empirically determine the weights: the proximity weight $w_{px} = 0.4$, passage weight $w_p = 0.3$, query weights $w_q = 0.15$ and answer weight $w_a = 0.15$ after analyzing the manually selected images in the dev set (as discussed later).

For the E&M baseline, we pretrain the text extraction with the SQUAD dataset (Rajpurkar et al., 2016) and finetune it on our dataset. For the image matching, we use image ranking using the input query (**Q**), input passage **P** and the extracted input answer **A** all concatenated together. For MExBERT, we tested different variants with and without proxy supervision (**PS**); with different pre-training setups - pretraining the textual stream alone, visual stream alone and both - to test the independent value of different pre-training.

Except pretraining experiments and baseline experiments, all our experiments on MExBERT have been conducted with 3 random seeds and the reported scores have been averaged over the 3 seeds. We use BERT pretrained embeddings for the textual stream of MExBERT and use $N_{T_a} = N_{T_b} = N_V = 6$. For finetuning MExBERT, we use Adam optimizer initialized with a learning rate of 0.0001 and train it till the validation loss saturates. The model was trained over 4 V100 machines using a batch size of 8 for finetuning and 64 for pretraining. For pretraining, we use an Adam optimizer with a learning rate of 0.0001 for 2 Epochs over 3.2 million Image-Text pairs for all our ablations during pretraining stage. We use 768 dimensional textual embeddings with a vocabulary size of $30,522$ and intermediate hidden embedding size 3072 for both textual and visual features. We project 4096 dimensional VGG-19 image features into 2048 dimensions and use it as input to the visual stream.

Evaluation Metrics: We independently evaluate the text and image part of the extracted answer using various metrics. For the text, we considered standard metrics like ROUGE, BLEU popularly used in the literature for textual question answering task. For images, we use the precision @1,2 and 3 in which we measure if the predicted image is in top-1,2 or 3 images as selected in the ground truth. Although these metrics are standard, we verify their utility in the multi-modal case by conducting a human experiment and calculating their correlations with human judgments.

To further validate the choice of our metrics, we collated a subset of 200 examples which have their ground truth available (collected as discussed later). We, then, apply our best performing model for these examples and generate the multimodal answers. For each of 200 pairs, we have both its predicted as well as ground truth counterparts. We conduct a human experiment where the annotators are asked to rate the quality of both textual and image part of the answer on relevance **R** and user satisfaction **S**. The overall quality of the answer is high if it is both relevant and provides high user satisfaction. For each pair, 5 different annotators rate the answers resulting in independent ratings for both predicted and ground truth answers. We calculate the overall quality of a predicted answer Q_a

with respect to the ground truth by calculating the ratio between the quality (which we represent by R*S) of predicted answer and the ground truth answer, $Q_a = \frac{R*S \text{ for predicted}}{R*S \text{ for ground truth}}$. We compute the pearson correlation between different metrics and Q_a. We observe that Rouge-1, Rouge-2, Rouge-L and BLEU yielded a correlation scores of 0.2899, 0.2716, 0.2918 and 0.2132 - indicating a moderate correlation and reassuring their viability for evaluating textual answer even in our multimodal setup. For image metrics, we found precision@1 to be most strongly correlated with human judgement (0.5421). While the expectation might be that such a metric has a perfect correlation, the user judgement is also biased by the corresponding textual answer leading to different scores even if the image is same in actual and predicted answer.

Evaluating Textual Outputs: Table 2 shows the performances of E&M against MExBERT (and its ablations) on extracting the right **textual part of the multimodal answer**. In order to test whether the visual attention on it's own makes any difference to the text answer quality, we also compare two variants of MExBERT - one where the visual input is zeroed out and another where the images are given as input without any supervision on the image selection. In the latter case we use the average attention weights of an image to determine its relevance to an answer. While not drastically large, we observed noticeable improvements with the visual input as compared to zero visual input, affirming our understanding about the value of utilizing multimodal input and cross-modal learning. We notice a marginal improvement in the text scores if we use proxy supervision scores during training. Intuitively, this is because of better focus of query on the target image which further enhances its attention over the correct part of the answer in the input. Due to relatively smaller corpus as compared to text only QA datasets used usually in recent works, we considered pretraining to be a natural choice to improve our model further. While the improvements in text scores with the visual training are marginal (which is expected since this training is directed at visual stream), language pretraining yields reasonable improvements as shown in Table 2.

Evaluating Image Output: We rank images in test set using our proxy supervision scores. We also select the image with the highest score as predicted by the respective model. We deem this image as

MODEL	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
E&M	46.77	43.26	47.22	25.17
MExBERT + Zeroed Visual Input	44.10	41.90	44.91	24.28
MExBERT	45.13	43.02	45.77	24.96
MExBERT + PS	45.67	43.59	46.17	25.04
MExBERT + PS + L PT	48.12	46.22	48.82	28.01
MExBERT + PS + V PT	46.18	44.11	47.24	25.89
MExBERT + PS + V+ L PT	**48.88**	**47.02**	**49.03**	**28.50**

Table 2: Results showing the performance of E&M and MExBERT over various textual metrics for test set. The results in bracket indicate that they have been obtained after ranking the passages in input article

MODEL	PRECISION@1	PRECISION@2	PRECISION@3
Random	0.139	0.258	0.381
E&M	0.255	0.444	0.541
MExBERT	0.211	0.421	0.528
MExBERT + PS	0.268	0.449	0.544
MExBERT + PS + L PT	0.271	0.453	0.546
MExBERT + PS + V PT	0.288	0.459	0.549
MExBERT + PS + V+ L PT	**0.291**	**0.459**	**0.549**

Table 3: Results showing the performance of E&M and MExBERT over the image modality of the multimodal answer as measured using proxy scores over test set

Precise @1,2 or 3 depending upon if it is present in top-1, top-2 or top-3 images as ranked by our proxy-supervision mechanism. While conducting evaluation, we skip those data points which have no-image or only a single image in the input to avoid any bias in the evaluation. After removing such datapoints, there were $2,800$ test datapoints with 2 or more images. As mentioned before, in the E&M, we retrieve the highest scoring image matched based on concatenation of **Q**, Passage **P**, and the extracted Answer **A** as the matching text, so that model has access to the whole textual input. Evidently, the results obtained are better than random but are still far from accurate. In fact, they are just more than half as good as those obtained with our heuristically created proxy scores when compared with human preferences as shown in Table 4. This shows that the problem is much harder than just using image retrieval models calling for a joint attention to understand the relevance of question, passage and answer. Using questions and answers as input text for UNITER were either poorer or similar, and hence not reported due to space limitation.

The power of joint multimodal attention is strongly evident as even without any visuo-lingual pretraining, we obtain meaningful (better than random) scores with *just* the averaged attention. The assumption, while using the highest average attention weights for selection the image, is that the model learns to focus on relevant images while be-

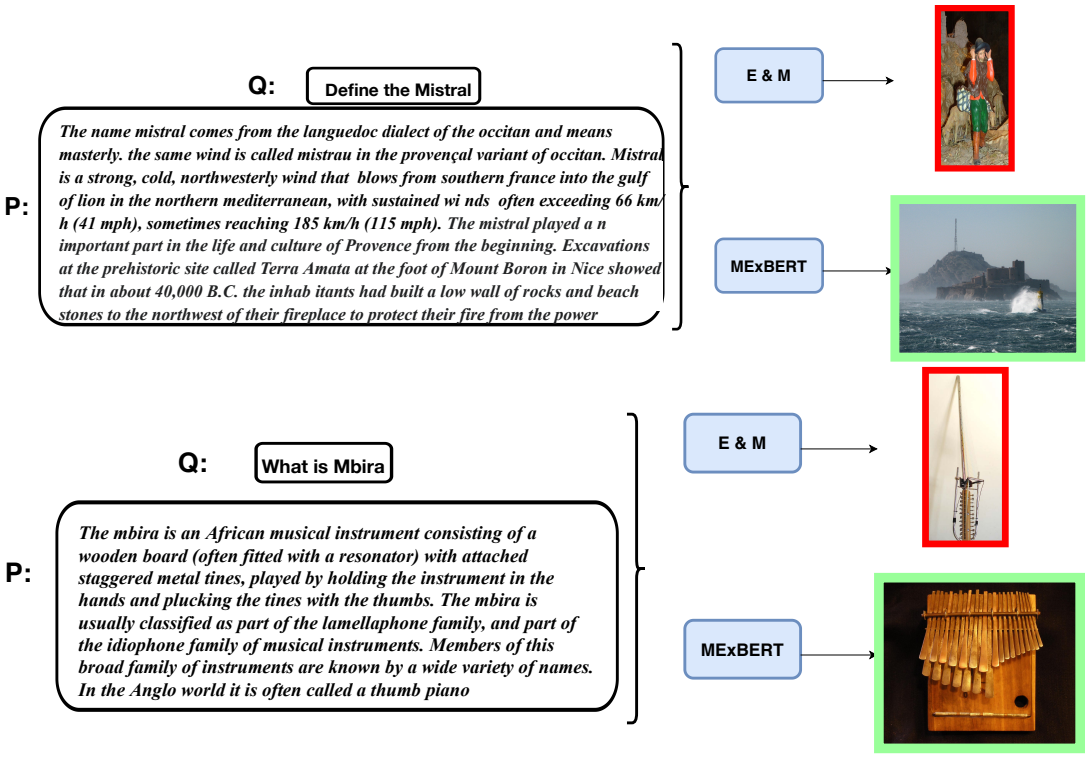


Figure 4: *The use of joint attention provide better understanding enabling models to focus better for retrieval of both image and textual answer from the input*

ing trained to optimize for better textual answer generation. Applying our proxy supervision mechanism while training the model, we find a very significant improvement specially in PRECISION @ 1 scores. PRECISION @ 2,3 scores are however similar to what we obtained with E&M. That is perhaps due to the fact that UNITER is good at estabilishing the relationships between text and images resulting in good PRECISION@2,3 scores but it fails at deciding the top image with high confidence due to lack of explicit understanding about where to focus on the text. Such a joint understanding is the main strength of **MExBERT**. Visual pretraining yields larger improvements on PRECISION@1 metric, while the language pretraining provides marginal improvements.

Human Evaluation: While our proxy scores have been intuitively designed, they are error prone. We therefore collected human annotations over the entire test corpus to further validate our model's performance. We conduct a Mechanical Turk experiment where the turkers were asked to select an image from a given set of input images for (question, answer, source passage) triplet which embellishes the textual response. Every question-answer pair was annotated by 5 annotators, with each annotator annotating 5 such pairs; we pay $0.2 for every such annotation. We also provide an option of selecting 'no image' since some inputs might not have any relevant image that could go well with answer. We find an agreement rate of over 50 % for the selected image in over 90 % of the cases. We, therefore, use the average number of votes per image as a 'preference' score for the image, and use this to compute the precision values in Table 4. The performance of MExBERT against such human annotations is better than its performance when calculated over proxy scores indicate that the proposed MExBERT is robust to the noise that might have crept in the proxy-supervision and generalizes well. This also explains why the precision is lower in the noisy setting of proxy supervision than the low-noise setting based on the human annotations. High precision values of proxy scores over the human preference scores demonstrate the effectiveness of our proposed heuristic for preparing proxy training targets.

6 Related Works

Machine reading comprehension and question-answering have been explored for a while, with the earliest works dating back to 1999 (Hirschman et al., 1999). Most of these works dealt with sin-

Model	Precision@1	Precision@2	Precision@3
Random	0.144	0.275	0.396
E&M	0.284	0.492	0.612
MExBERT	0.196	0.385	0.498
MExBERT + PS	0.316	0.505	0.608
MExBERT + PS + L PT	0.321	0.511	0.612
MExBERT + PS + V PT	0.381	0.535	0.616
MExBERT + PS + V+ L PT	**0.386**	**0.538**	**0.618**
Proxy Scores	**0.422**	**0.631**	**0.753**

Table 4: Results comparing performance of E&M and MExBERT over the image modality of the multimodal answer based on Human Evaluation over test set

gle modality at a time until recently. While earlier datasets were small, beginning with SQuAD (Rajpurkar et al., 2016) several large datasets (Rajpurkar et al., 2018; Yang et al., 2018; Choi et al., 2018; Reddy et al., 2019; Kwiatkowski et al., 2019) have been proposed. Though many of these are *extractive* in nature, there are a few multiple-choice datasets (Mihaylov et al., 2018; Richardson et al., 2013). Datasets like QAngaroo and HotpotQA (Welbl et al., 2018; Yang et al., 2018) enable reasoning across multiple documents. Recently, several Table-QA datasets have also been proposed, aimed at providing a natural language answer by reasoning over tables. While some datasets like WikiTableQuestions (Pasupat and Liang, 2015) and MLB (Cho et al., 2018) have natural language questions, others like TabMCQ (Jauhar et al., 2016) have multiple choice questions.

A popular exploration in **multimodal question answering** is Visual Question Answering or VQA (Antol et al., 2015; Goyal et al., 2017; Anderson et al., 2018; Lu et al., 2016, 2019; Tan and Bansal, 2019) where the input is a textual query along with an image and the output is a text answer. Another variant of this, Charts Question Answering (Kafle et al., 2020, 2018; Kahou et al., 2017; Chaudhry et al., 2020), allows for the input to be a chart instead of a natural image. While both of these problems involve multimodality (image + question or chart + question), the output is still textual (specifically an answer class since this is modelled as a classification problem usually). While the question is received as a text in these problems, the reasoning is performed over a single modality only. In our work, we reason out across multimodal input by simultaneously attending to images and text in the input to arrive at our target output.

To overcome unimodal reasoning, there are attempts at truly **multimodal reasoning** with the datasets such as ManyModalQA (Hannan et al., 2020), RecipeQA(Yagcioglu et al., 2018), and TVQA (Lei et al., 2018). While RecipeQA aims reasoning over recipes and the associated pictures, TVQA involves multimodal comprehension over videos and their subtitles. The recently proposed ManyModalQA goes a step further by adding tables to the multimodal reasoning as well. However, these datasets provide responses in a single modality only, either an MCQ or textual response. With the rate at which multimodal consumption is taking place in our lives, it is important that the answering systems also enable multimodal output which, as discussed, already can provide better cognitive understanding when combined with textual modality.

7 Conclusion

We presented one of the first exploration, to the best of our knowledge, of multimodal output question answering from multimodal inputs and proposed usage of publicly available textual datasets for it. We proposed strong baselines by utilizing the existing frameworks for extract textual answers and independently match them with an appppropriate answer. We demonstrate the value of a joint-multimodal understanding for multimodal outputs in our problem setup by developing a multimodal framework **MExBERT** which outperformed the baselines significantly on several metrics. We also developed a proxy supervision technique in absence of labelled outputs and showed its effectiveness for improved multimodal question answering. We used some existing metrics to compare the different models and justified the usage of these metrics based on a human experiment.

While it is an interesting and challenging task even in its current shape, we believe there are several limitations in our proposed framework. While our datasets had multimodal elements, modeling multimodal reasoning from multimodal inputs and using it to arrive at a multimodal answer calls for a more careful question curation that includes these challenges. Recently proposed datasets such as MultimodalQA have created questions explicitly aimed at reasoning across multimodal input, but however, lack the multimodal output component. Future works could include questions which specifically aim for a visual elements making the output requirement multimodal. Also, free form answer generation in the multimodal input/output context is another interesting subject of further research.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 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*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. 2020. Experience grounds language. *arXiv preprint arXiv:2004.10151*.

Ritwick Chaudhry, Sumit Shekhar, Utkarsh Gupta, Pranav Maneriker, Prann Bansal, and Ajay Joshi. 2020. Leaf-qa: Locate, encode & attend for figure question answering. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 3512–3521.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*.

Minseok Cho, Reinald Kim Amplayo, Seung-won Hwang, and Jonghyuck Park. 2018. Adversarial tableqa: Attention supervision for question answering on tables. *arXiv preprint arXiv:1810.08113*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*.

Edgar Dale. 1969. Audiovisual methods in teaching.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2054–2063.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Haytham M Fayek and Justin Johnson. 2020. Temporal reasoning via audio question answering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 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*, pages 6904–6913.

Darryl Hannan, Akshay Jain, and Mohit Bansal. 2020. Manymodalqa: Modality disambiguation and qa over diverse inputs. In *AAAI*, pages 7879–7886.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Lynette Hirschman, Marc Light, Eric Breck, and John D Burger. 1999. Deep read: A reading comprehension system. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, pages 325–332.

Sujay Kumar Jauhar, Peter Turney, and Eduard Hovy. 2016. Tables as semi-structured knowledge for question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 474–483.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5648–5656.

Kushal Kafle, Robik Shrestha, Scott Cohen, Brian Price, and Christopher Kanan. 2020. Answering questions about data visualizations using efficient bimodal fusion. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1498–1507.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in neural information processing systems*, pages 289–297.

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Xilin Chen, and Ming Zhou. 2020. Univilm: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Roxana Moreno and Richard Mayer. 2007. Interactive multimodal learning environments. *Educational psychology review*, 19(3):309–326.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. 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*, pages 193–203.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347*.

Michael Sankey, Dawn Birch, and Michael Gardiner. 2010. Engaging students through multimodal learning environments: The journey continues. In *Proceedings ASCILITE 2010: 27th annual conference of the Australasian Society for Computers in Learning in Tertiary Education: curriculum, technology and transformation for an unknown future*, pages 852–863. University of Queensland.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes. *arXiv preprint arXiv:1809.00812*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. Msmo: Multimodal summarization with multimodal output. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4154–4164.

Supplementary Material

Anonymous NAACL submission

1 Implementation details

Our models were trained on 4 V100 machines and takes just 1 sec for the whole people in such setting. As mentioned, Except pretraining experiments and baseline ex-periments, all our experiments on MExBERT havebeen conducted with 3 random seeds and the re-ported scores have been averaged over the 3 seeds. For pretraining, we use an Adam optimizer with alearning rate of 0.0001 for 2 Epochs over 3.2 mil-lion Image-Text pairs for all our ablations duringpretraining stage. We use768dimensional tex-tual embeddings with a vocabulary size of30,522and intermediate hidden embedding size3072forboth textual and visual features. We project4096dimensional VGG-19 image features into2048di-mensions and use it as input to the visual stream

2 Human Evaluation

We conduct elaborate human experiments for analyzing the performance of our models as well as the utility of the task. As mentioned, we perform an experiment to establish the need of such a task and how multimodal outputs provide enhanced understanding to the end user. Before that, however, we perform a human experiment to label the relevant image for each question in the test set as discussed in the section on human evaluation. The interface for the experiment is as shown in Fig. 1. For each HIT, we provide the turkers with 5 (Question, Answer, Passage) triplets and multiple choice options where they select the most relevant image corresponding to the question-answer pair. We demonstrate the what relevance means with the help of an example as shown. We also provide them with an option to select the option 'None of these' as in some cases, no image might be relevant. In order to ensure the quality of responses (to accept or reject turkers' responses), out of 5 questions, we insert random images in one random question. Ideally, a turker paying attention while providing responses is expected to select 'None of these' for the question. We find more than 90% acceptance ratio in the first event indicating the high quality annotation.

After creating the test set (over 3.5k examples), we randomly select 200 examples from the test set (ensuring there are atleast 2 images in the selected examples) and provide a unimodal as well multimodal answer for the annotators to analyze in another experiment. As shown in Figure 2, we ask the annotators a set of overall 6 questions. We have already discussed the outcomes of the experiment in the main paper. We, here, highlight how we maintain the quality of the responses. In some random inputs to the annotator, we make text-image pair incompatible while in some cases we make the answer non-recoverable from the input passage. A turker paying appropriate attention to the task will be easily able to identify the answer - 'No' to the two additional questions given at the end. The answers to those two questions determine whether a particular HIT is accepted or rejected. Since, we provided reasonable amount for annotation, we find ¿95% acceptance ratio indicating that the evaluation so performed is pure and can be reliable used to make conclusions.

3 Dataset

In this section, we describe the dataset collection process and present some statistics about the dataset.

As already described in the main paper, we create our dataset by curating and subsampling a set of questions with images from MS-Marco and Natural Questions dataset. Fig. 3 shows the distribution of different types of tokens in the dataset. We have

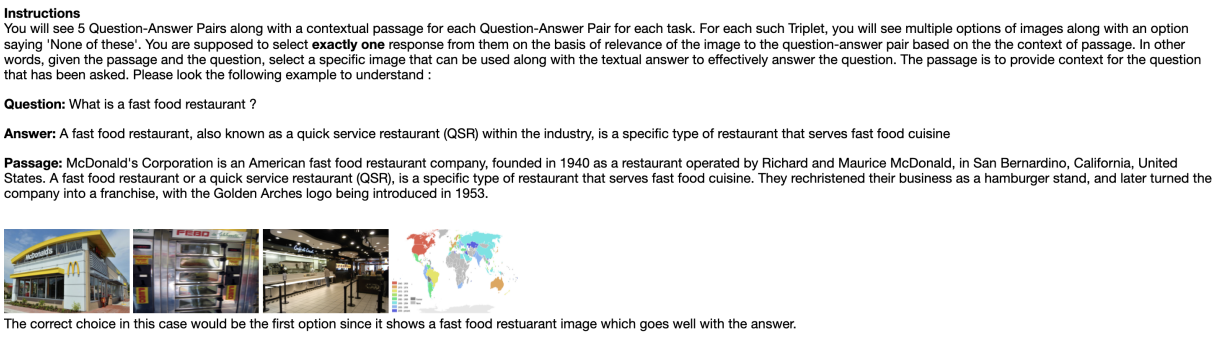


Figure 1: *Instructions provided to the human annotators for labelling the relevant image for the triplet*

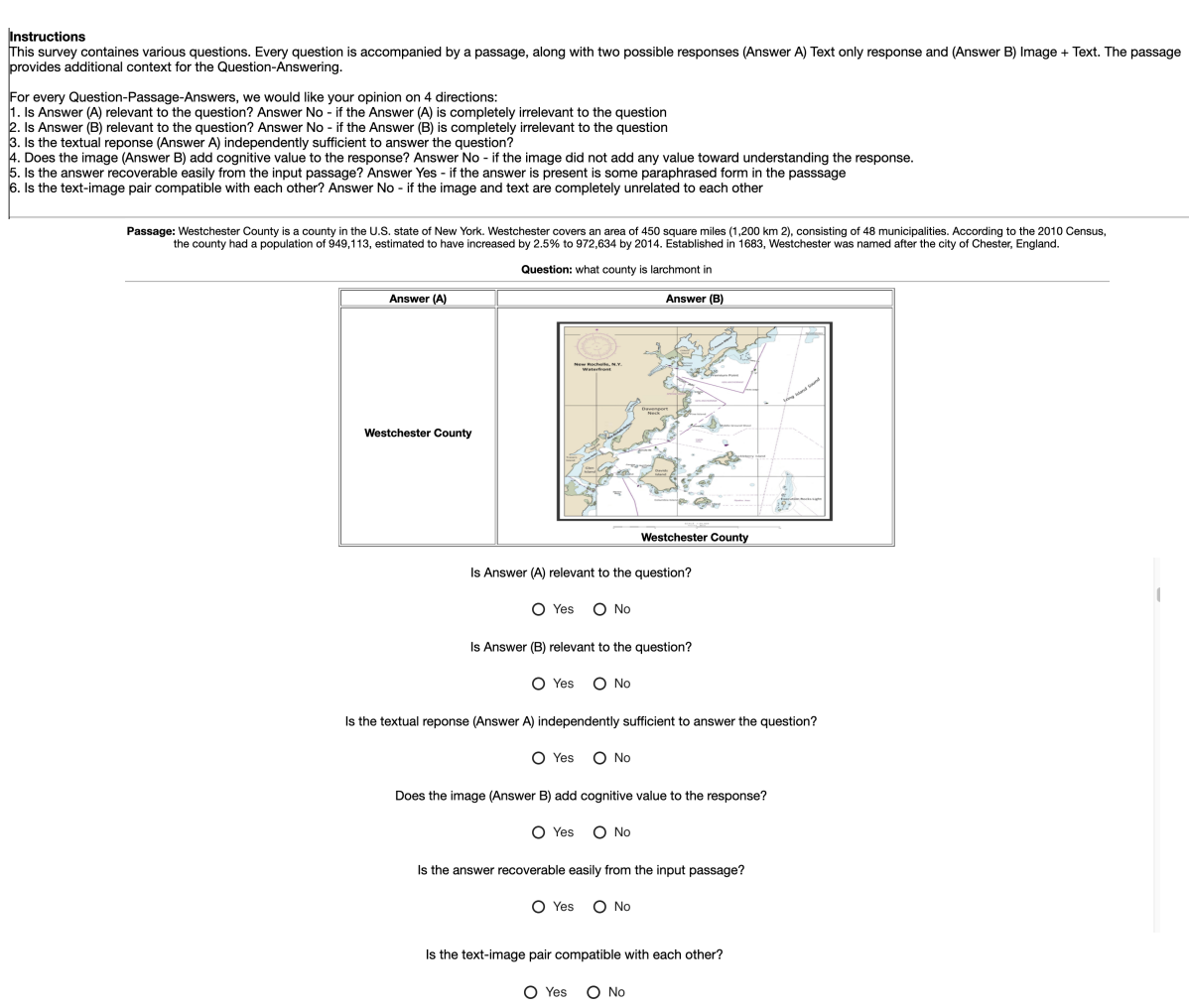


Figure 2: *Interface shown to the human annotators for the task of identifying the need of the multimodal output*

only retained those frequently occuring tokens (for both levels) which have more than 5% of the total frequency for their category for the simplicity of representation.

Filtering for MS-MARCO From the MS-MARCO dataset we filter out the entries which do not have a Wikipedia page as a source for the answer paragraph. Since, we are focusing on extractive multimodal outputs in this paper, we further eliminate all those question-answer pairs where the answer does not *appear* in the selected passages. Instead of eliminating answers without an exact match, we use edit distance to retain answers that include minor edits (e.g. removal of parenthesis) in our dataset.

Filtering for Natural Questions For the Natural Questions dataset all answers are guaranteed to be grounded in Wikipedia entries. We use the short

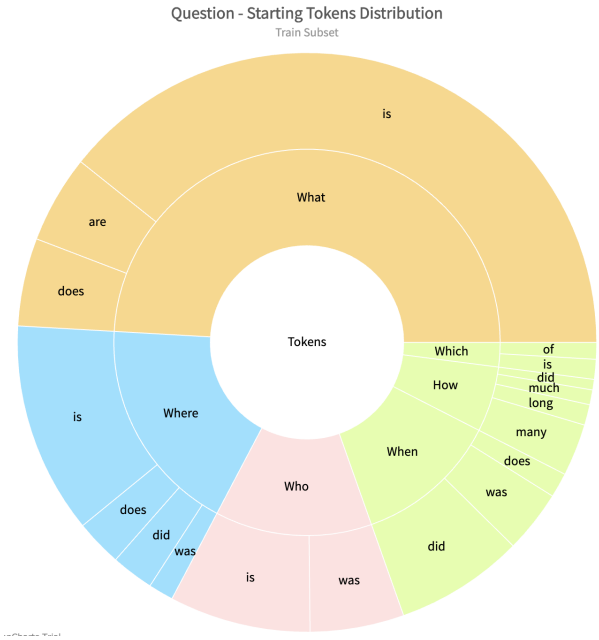


Figure 3: *Starting Token Distribution for the train set*

answer provided by the authors as our target answer, and use the long answer along with distractor passages as the input to our model. However, to reduce the noise from NQ, we removed questions with a *single-word* answer and questions where the original Wikipedia article had no images.

Scraping images from Wikipedia: Our main motivation of using answers grounded in Wikipedia articles for our corpus was to exploit the structure of such articles to scrape images and get proxy supervision. To this end we prepend the title of the article provided in the `url` field of MS-MARCO with `http://en.wikipedia.com/wiki/` to get the URL of the appropriate Wikipedia article. We use the `BeautifulSoup` package to find all objects of the `img` class from the HTML page and scrape the largest available resolution of the image (found from the `srcset` property). Further, we only scrape images which are of the `.png` or `.jpeg` formats to avoid other media. Finally we only retain a maximum of 20 images from each page in order to avoid a large number of irrelevant images.

Providing proxy supervision: Our proxy supervision score broadly consists of a proximity score and a caption relevance score. To compute the former we determine the position of each image in the HTML source of the Wikipedia page by finding the paragraph directly below the image. In case the image does not have any text below it we consider the paragraph directly above it. We then compute the number of tokens between the first (or last) word of the found paragraph and the first word of the answer paragraph and normalize it. For the case of multiple images together in the HTML we assign the same proximity score to all of them.

To compute the caption relevance we use the `thumbcaption` attribute of the image from the HTML source. In case an image does not have this property in the HTML we consider the text below the image in place of the caption. We then compute the TF-IDF scores of the "caption" with the answer, query and answer passage to get the caption relevance scores.

Image distribution We also show the distribution of the number of input images per question in Fig. 4. Evidently more than 80 % of the dataset has more than two images while a significant proportion (more than 30%) has more than 6 images making the task fairly difficult. This has also been demonstrated by the large difference between the UNITER accuracy and MExBERT's accuracy.

We show below some randomly chosen samples from the dataset (which were also correctly chosen by our model MExBERT) to provide reader with an idea about the variety of inputs and input images. The question is shown at the top of the box while the input passage and the set of images have been shown inside the box. The red boundary over one box one of the images denote the image which was annotated as the selected image during annotation and was also predicted correctly by the MExBERT

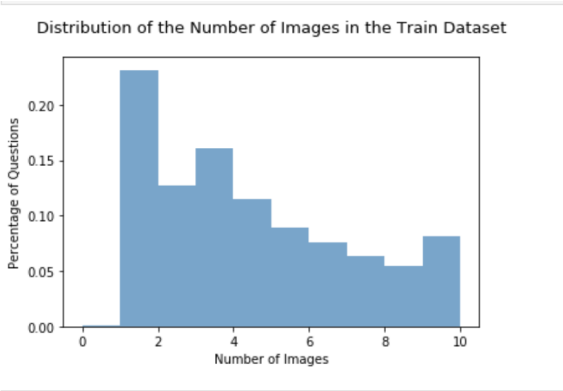


Figure 4: *Distribution of the of Input Images Per Question for the Train Dataset*

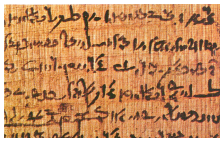
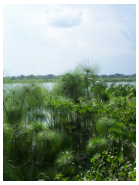
framework.

Question: What is Papyrus ?

Input Passage:

Egyptians used papyrus for 4000 years until other plants and trees were used to make paper for economical reasons. Papyrus is still made, but normally only as a tourist attraction. Cyperus papyrus (papyrus sedge, paper reed, Indian matting plant, nile grass) is a species of aquatic flowering plant belonging to the sedge family cyperaceae. it is a tender herbaceous perennial, native to africa, and forms tall stands of reed-like swamp vegetation in shallow water. papyrus is made from a plant that grows on the banks of the nile river in egypt. the aquatic plant, cyperus papyrus, grows up to 15 feet (4.5 meters) high. its green, triangular stem has long, sharp leaves and flower clusters 10 to 20 inches (25 to 50 cms) long. these flowers bloom at the tip. 1 the climate of egypt and certain parts of mesopotamia preserved papyri in the ruins of ancient towns and cemeteries. 2 egyptians used papyrus for 4000 years until other plants and trees were used to make paper for economical reasons. 3 papyrus is still made, but normally only as a tourist attraction

Input Images

Question: What is a Shillelagh?

Input Passage:

An Irish word for a cudgel made of blackthorn, oak, or other hardwoods. usually slightly smaller than a walking stick or cane. good for quick repetitious beating of individuals. there in the village of shillelagh you'll find the namesake shillelagh sticks, stout cubs or cudgels made from the wood of oak or blackthorn. the wood is fashioned into walking sticks, clubs, cudgels, fighting sticks, staffs, and even good luck charms. a shillelagh , willow or blackthorn stick is a wooden walking stick and club or, cudgel typically made from a stout knotty stick with a large knob at the, top that is associated with ireland and irish.
folklore

Input Images


Question: What does fawn mean in Dogs?

Input Passage:

terriers and hounds. tan dog with a black saddle and white markings (to any extent). trim. various-general term. a small amount of white on the chest, muzzle, toes and/or tail tip. trindle. various-general term. brindle tricolour (i.e. black with brindle points and white markings. a fawn great dane. fawn is a light yellowish tan colour. it is usually used in reference to clothing, soft furnishings and bedding, as well as to a dog 's coat colour. it occurs in varying shades, ranging between pale tan to pale fawn to dark deer-red. the first recorded use of fawn as a colour name in english was in 1789. this can be a bit of a barrier when it comes to working out the genetics of particular breeds, so to make things easier, here's a list of some of the terms you'll find (either on breed standards or being used by breeders), and what they actually mean in terms of the genetics we've studied on this site

Input Images


Question: What is Altitude training?

Input Passage: altitude training is the practice by some endurance athletes of training for several weeks at high altitude, preferably over 2,400 metres (8,000 ft) above sea level, though more commonly at intermediate altitudes due to the shortage of suitable high-altitude locations.ltitude training works because of the difference in atmospheric pressure between sea level and high altitude. at sea level, air is denser and there are more molecules of gas per litre of air. altitude training works because of the difference in atmospheric pressure between sea level and high altitude. at sea level, air is denser and there are more molecules of gas per litre of air.regardless of altitude, air is composed of 21% oxygen and 78% nitrogen.ltitude training works because of the difference in atmospheric pressure between sea level and high altitude . at sea level, air is denser and there are more molecules of gas per litre of air. altitude training can be simulated through use of an altitude simulation tent, altitude simulation room, or mask-based hypoxicator system where the barometric pressure is kept the same, but the oxygen content is reduced which also reduces the partial pressure of oxygen.ltitude training works because of the difference in atmospheric pressure between sea level and high altitude.

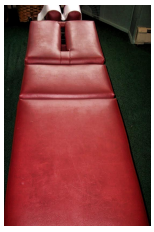
Input Images



Question: What is chiropractic treatment

Input Passage: . chiropractic is a health care profession dedicated to the non-surgical treatment of disorders of the nervous system and/or musculoskeletal system. generally, chiropractors maintain a unique focus on spinal manipulation and treatment of surrounding structures.iew chiropracticvideos. chiropractic is a health care profession dedicated to the non-surgical treatment of disorders of the nervous system and/or musculoskeletal system. generally, chiropractors maintain a unique focus on spinal manipulation and treatment of surrounding structures. chiropractic is a form of alternative medicine that focuses on diagnosis and treatment of mechanical disorders of the musculoskeletal system, especially the spine, under the belief that these disorders affect general health via the nervous system.he specific focus of chiropractic practice is chiropractic subluxation. traditional chiropractic assumes that a vertebral subluxation or spinal joint dysfunction interferes with the body's function and its innate intelligence. spinal adjustment/manipulation is a core treatment in chiropractic care, but it is not synonymous with chiropractic. chiropractors commonly use other treatments in addition to spinal manipulation, and other health care providers (e.g., physical therapists or some osteopathic physicians) may use spinal manipulation.top.ands-on therapy—especially adjustment of the spine—is central to chiropractic care. chiropractic is based on the notion that the relationship between the body's structure (primarily that of the spine)

Input Images



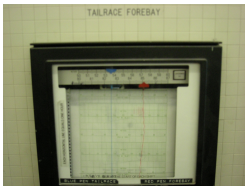
Question: What does the three gorges dam produce?

Input Passage: the three gorges dam area is rich in archaeological and cultural heritage. many different cultures have inhabited the areas that are now underwater, including the daxi (circa 5000-3200 b.c.e), which are earliest neolithic culture in the region, and its successors, the chujialing (circa. the itaipu dam opened in 1984 in south america as the largest, producing 14,000 mw but was surpassed in 2008 by the three gorges dam in china at 22,500 mw. hydroelectricity would eventually supply some countries, including norway, democratic republic of the congo, paraguay and brazil, with over 85% of their electricity. construction on the three gorges dam was completed in 2008. the dam stands 185m high and 2,309m wide, making it the world's largest hydro plant, well ahead of brazil's 12,600mw itaipu installation.",

Input Images





Question: What caused august 6 1945

Input Passage: In this aug. 6, 1945, file photo, smoke rises around 20,000 feet above hiroshima, japan, after the first atomic bomb was dropped. on two days in august 1945, u.s. planes dropped two atomic bombs, one on hiroshima, one on nagasaki, the first and only time nuclear weapons have been used.igh-angle view of a section of the city of hiroshima after the us atomic bombing on august 6, 1945. (photo by keystone/getty images). sacred trees stand bare and broken near fallen tombstones at the temple of kokutaiji, following the us atomic bombing of hiroshima, japan on august 6, 1945. the united states dropped atomic bombs on the japanese cities of hiroshima and nagasaki in august 1945, during the final stage of the second world war.the two bombings, which killed at least 129,000 people, remain the only use of nuclear weapons for warfare in history.he united states dropped atomic bombs on the japanese cities of hiroshima and nagasaki in august 1945, during the final stage of the second world war. the u.s. attacked japan on august 6, 1945 using a gigantic, atomic bomb, codename "little boy", that was equivalent to 20,000 tons of tnt. the bomb was dropped in hiroshima and destroyed the city, killing thousands of civilians.he u.s. attacked japan on august 6, 1945 using a gigantic, atomic bomb, codename "little boy", that was equivalent to 20,000 tons of tnt. the bomb was dropped in hiroshima and destroyed the city, killing thousands of civilians

Input Images