MIMOQA: Multimodal Input Multimodal Output Question Answering

Anonymous NAACL submission

Abstract

Research on multimodal understanding has picked up 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 output, predominantly text. In this paper, we propose a novel task - **MIMOQA** - **M**ultimodal **I**nput **M**ultimodal **O**utput **Q**uestion **A**nswering where the output is also multimodal. Through human experiments, we empirically establish the better cognitive understanding of the answers provided by such multimodal outputs and propose a novel multimodal question-answering framework, **MExBERT**, that jointly incorporates textual and visual attention towards producing such a multimodal output. Our method relies on a novel way to curate a multimodal dataset for this problem from publicly available unimodal datasets. We show superior performance of MExBERT against strong baselines on both automatic and 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 nature of the content 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 (Silberer and Lapata, 2014; Srivastava and Salakhutdinov, 2012; Feng and Lapata, 2010), various works have started going beyond unimodal machine comprehension (Hermann et al., 2015; Kočiskỳ et al., 2018; Nguyen et al., 2016; Kwiatkowski et al., 2019) towards a holistic multimodal machine 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) and have shown significant cognitive improvements.

While the focus on multimodal understanding has increased, all these explorations, specifically in question answering, have limited their focus to unimodal outputs even with multimodal inputs. For example - the task of *Visual Question Answering (VQA)* 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 lines of work, thus implicitly assume that irrespective of the question, the textual answers can satisfy the needs of the query across multiple input modalities. We posit that such an assumption is not true; while textual answer can address several queries, a multimodal answer can almost always enhance the cognitive understanding of the end user.

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 contributions can be summarized as below:

- We introduce the problem of multimodal input multimodal output question answering. We establish the importance of such multimodal outputs in question-answering for cognitive understanding via human experiments.
- We propose **MExBERT**, a novel multimodal setup for *extracting* multimodal answers to

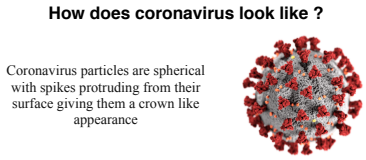

Figure 1: *Here, the text alone technically answers the query, but the image provides a better cognitive understanding and grounding of various ambiguous terms in the text*

a given question and compare it against relevant baselines. Our proposed method also includes a novel pretraining mechanism to effectively bootstrap our framework in limited data scenario. We also propose a *proxy* supervision technique for the image selection in our framework and conduct ablations to show its importance and utility.

- We curate a large dataset for the introduced problem by extending the publicly available datasets for textual question answering. 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 example, in Figure 1, the multimodal output provides a clearer *picture* for understanding the question and also provides grounding to words like 'spherical', 'crown-like' enabling a better communication of the textual answer.

In some cases, textual answer might be even 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. To verify this hypothesis, we collated 200 Question-Answer pairs (as discussed later); for each pair, we created its unimodal as well as multimodal counterparts. We conducted a human experiment where each question-answer pair was judged by 5 annotators, each annotator asked to rate if the textual answer is sufficient for the input query. Irrespective of whether it is sufficient, the annotators were also asked whether the image in the multimodal variant enhances the understanding of the answer and add value to it. In such human experiments, often it is possible that the annotators begin getting biased towards a richer multimodal response. To avoid this, we had explicitly inserted a few irrelevant images (20%) to avoid and assess any unintended bias of the annotators towards multimodal answers.

Out of 80.27% of the total responses by annotators who felt that textual answers were sufficient, 87.5% responded that 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 total 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 cognitive 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 the unimodal responses are insufficient to answer a given query in many cases and the multimodal answer 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 Question Answering

We formally define our problem as - *given a piece of input text* ***T*** *along with a set of related images* ***I*** *and a query* ***Q****, extract a multimodal answer* ***M*** *from* $\{$***I, T***$\}$ *consisting of both text and images.* 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, irre-

spective of the sufficiency of the textual answers. Thus, our multimodal answer **M** consists of a text $\mathbf{M_T}$ and an image $\mathbf{M_I}$.

3.0.1 Dataset Curation

Since multimodal output question answering is a new problem, there are no existing datasets which fulfill the needs of our problem. We, therefore, curate a dataset by utilizing the existing datasets. Many 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), where some answers are extracted from Wikipedia. From MS-MARCO, we retain only those questions whose answer source is Wikipedia. We retained the entire NQ dataset, since all answers come from Wikipedia. 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. We scraped all the images from the original Wikipedia articles corresponding to the filtered questions.

While natural questions consists of only *extractive* answers, MS-MARCO also includes free form answers. However, MS-MARCO additionally provides information about the exact passage from which the answer has been retrieved. 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.

Table 1 shows different statistics about the dataset. The dataset includes a large number of images in the input corpus making the task of selecting appropriate image non-trivial. The large variety of images also necessitates a robust visual and language understanding by our model. The passages have been formed by combining the answer

	# 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*

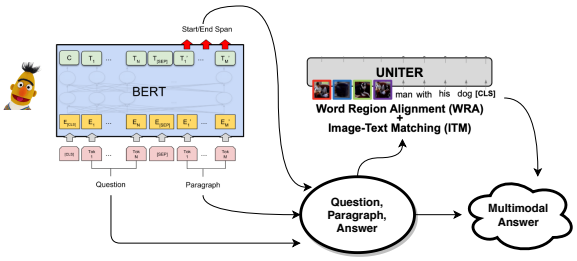

Figure 2: ***E&M** Baseline: **E**xtract a textual answer and utilize UNITER to **M**atch the appropriate image to the text.*

source passage and randomly chosen $2-3$ different 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 as input) considering the focus of our experiments is on multimodal output and not article-passage ranking. At inference time, however, we test with both the mixed passages and the entire input article (ranking the passage with BERT-based ranker (Nogueira and Cho, 2019) followed by question-answering).

3.0.2 Extract & Match (E&M)

In the absence of any existing frameworks that output a multimodal response, we develop a strong baseline by combining the existing state-of-the frameworks in question answering and visuolingual understanding. We first extract the textual answer and then match the appropriate image utilizing the extracted textual answer or query or input passage. We refer to this variant as **Extract** and **Match** (**E&M**). Given the input query (**Q**) and the input text (**T**) and images (**I**), we either utilize a BERT based passage ranker (Nogueira and Cho, 2019) to rank the different paragraphs in the input text (**T** before selecting top 3 paragraphs and concatenating them or we select the paragraph as curated in our dataset. We use this concatenated piece of text along with the query and extract the textual answer using BERT-QA (Devlin et al., 2018) framework.

We use this extracted answer, query, and input text to selected an image from the input images. For our image selection network, we use 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. Due to strong pretraining, UNITER has achieved SOTA performance on a variety of vision and language task. So, we use this as our baseline for image selection. We, then, use the classification confidence as predicted by image-text matching head of UNITER and select the image which receives the highest confidence score corresponding to a given text (which can be a combination of query, answer, and source paragraph as discussed in Experiments).

3.0.3 Multimodal Extractive BERT (MExBERT)

A major problem with extracting and then matching multimodal answers 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 3 shows the overall architecture of our proposed MExBERT framework. Inspired by the recent visuo-lingual pretraining frameworks (Tan and Bansal, 2019; Lu et al., 2019; Chen et al., 2019), our framework 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 based transformer (Vaswani et al., 2017) layers. The input to our textual stream as shown in Figure 3 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 embeddding to provide the positional information of tokens and segment IDs to help distinguish between query and passage.

Unlike the the canonical BERT-QA, our textual stream employs two different types of layers - regular self-attention layers and layers with an additional cross-attention block. 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 block. If we represent the attention computation in query-key-value format, the cross-attention block works by using textual token as a query and images' features' representations from the visual stream as the keys and values. This is different from self-attention where (query, keys and values) are all input textual tokens of the textual stream.

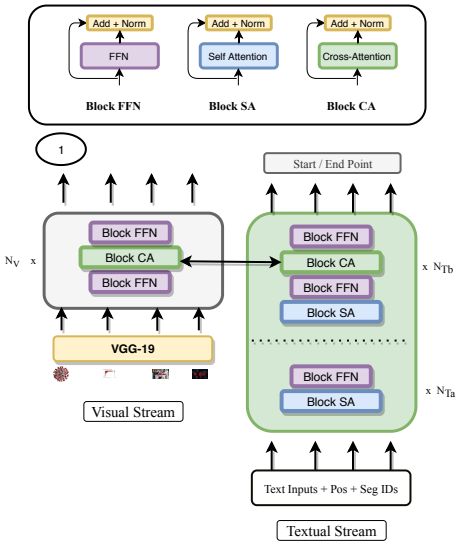

Figure 3: ***MExBERT***. *Details of the three blocks in the visual and textual streams is illustrated on the top. The visual stream takes the output of VGG-19 as input while the textual stream takes BERT Embeddings as input*

If the i^{th} textual token's features and j^{th} image's features being used as input for k^{th} layer in textual stream and $(k - N_{T_a})^{th}$ layer in the visual stream (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 given by $attn(q, k, v)$ then the self-attention and cross-attention for textual stream 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 then used to calculate the start and end position of the answer. The setup is similar to the original BERT-QA (Devlin et al., 2018) where one linear layer predicts the starting token through softmax applied over all tokens while another layer predicts ending token in a similar manner. The goal is to optimize the cross entropy loss over both the token position predictions.

The **visual stream** is similar to the textual 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 3. The self-attention was not used as the images mostly derive their relevance/context from the textual counterparts (powered by the cross-attention block) in the input passage or query unlike textual tokens which derive their contextual meaning from other tokens in the sentence. 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' feature representation in the corresponding layer of the textual stream. The input to the visual stream is the VGG-19 (Simonyan and Zisserman, 2014) features of each of the images. We don't use positional/segment encodings since we do not want to provide any positional information to MExBERT. We use a linear head on top of visual features to predict whether a particular image should be part of the multimodal output answer. The image with the highest confidence score on inclusion in the answer is regarded as the predicted image.

3.0.4 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 utilizing two types of information about the image - its position in the original input article and its caption information. Note that 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 MExBERT would be able to infer the correct multimodal response irrespective of the availability of such information at inference time.

Since MS-MARCO and Natural Questions both 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 score of the caption with the Query, the Answer and the source passage as shown in Figure 4. The overall supervision score of the image is then calculated as a weighted sum of these four scores where proximity score is calculated as $1 - P$. Once we have the supervision scores, we use these normalized

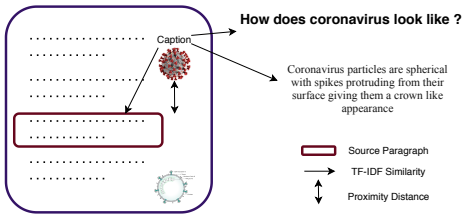

Figure 4: Calculation of proxy supervision scores

(between $0 - 1$) scores as targets for linear layer on top of the visual features. We use weighted binary cross-entropy where the weights w and $1 - w$ come from the proxy supervision values for a particular image.

3.0.5 Pretraining

Vision and Language Tasks have relied on pretraining to address the complexities associated in building visuo-lingual relationships (Tan and Bansal, 2019; Lu et al., 2019; Chen et al., 2019). Along the same vein, we leverage pretraining to better initialize our model. Another motivation in our case is that our signals (even after we include proxy supervision) are relatively sparse for a visuo-lingual task, calling for a strong 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 retrieved from the visual information also, visual stream is also fine-tuned in this process. Since, our final model is also going to use segment IDs, we randomly assign a segment ID of either query or passage to each caption during training runtime in order to imbibe language understanding for both type of tokens.

For **pretraining 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 initializa-

tion, the textual stream is also fine-tuned due to the cross-attention layers between the two streams.

4 Experiments

We conduct extensive experiments and ablations for the E&M baseline and the proposed MExBERT framework. We collated our curated dataset into three subsets - train, development and test after applying all our filters and Table 1 shows the different statistics for the dataset. Note that the passages in our datasets have been formed from the original wikipedia articles by first selecting the answer source passage and then combining it with $2-3$ *negative/distractor* passages. The *distractor* passages have been added or concatenated to construct a larger input which prepares the model to deal with complex inputs at inference time when the exact source passage information is not available.

As mentioned before, we used the 3.2 million Image-Caption pairs from Conceptual Captions dataset (Sharma et al., 2018) for pretraining MExBERT layer. For proxy supervision, we manually curated 50 multimodal answers and fit our weights over all four (proximity scores, and query, answer, passage scores) to maximize the overlap with these 50 answers. Rounding the weights yielded proximity weight $w_{px} = 0.4$, passage weight $w_p = 0.3$, query weights $w_q = 0.15$ and answer weight $w_a = 0.15$

We tested the E&M baseline for text extraction by pretraining it on the SQUAD dataset (Rajpurkar et al., 2016) (**E&M + PT**). Alternatively, we directly finetuned the BERT model (which has not finetuned on SQuAD) on our dataset (Table 1) as the **E&M + FT**. We also finetuned the SQUAD pre-trained model on our dataset as another variant (**E&M + PT + FT**). For the UNITER based image matching, we tested variants of image ranking using the input query (**Q**), input passage **P** and the extracted input answer **A**. We also tested with different combinations of these 3 inputs for our image matching baseline variants. For MExBERT, we tested different variants with and without proxy supervision (**PS**). We also tested different pre-training setups - where the textual stream alone was pretrained, visual stream alone was pre-trained and both the streams were pretrained - 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.

4.0.1 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 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 our case by conducting a human experiment and calculating their correlations with human judgments.

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. This way, for each of 200 pairs, we have both its predicted as well as ground truth counterpart. We, then, 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 (whether from ground truth or predicted), 5 different annotators are made to rate the answer. This way, we end up calculating the rating for both predicted and ground truth answer. We calculate the quality of 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, Precision@1 was strongly correlated with human judgement (0.5421). While the expectation might be that such a metric should have a perfect correlation but please note that the user judgement is also being biased by the corresponding textual answer and hence, the scores are different even if the image is same in actual and predicted answer.

We report our results on both type of inputs - ranker output being used as input paragraph for the models and curated input passage (with one source and $2-3$ *distractor* passages) being used to answer the questions. We report ROUGE and BLEU scores for textual evaluation of answers in Table 2. For evaluating the image outputs, we rank the images in the test set on the basis of our proxy scores. Please note that using proxy scores for both evaluation as well as training does not result in any 'leak' since we are not passing the proxy information of the images to the model, and use it only for evaluation. Hence, the only way for model to perform well on both training and testing datasets is if there is an intrinsic meaning in such scores which is generalizable. After ranking the images in the test set, we find the top predicted image by our model and depending upon if it is present in top-1, 2 or 3 *proxy* supervision ranked images, we calculate the Precision@1, Precision@2 and Precision@3 scores for different models as shown in Table 3. Additionally, the results while using ranker output as input to final model are also shown in brackets to the right of each score.

4.0.2 Evaluating Textual Outputs

Table 2 shows the performances of different baseline setups against MExBERT on extracting the right textual part of the multimodal answer. We observe that even though E&M+ PT (i.e. BERT Finetuned over SQuAD) achieves a decent performance, finetuning the pretrained BERT (which has not been finetuned over SQuAD) over our corpus (i.e. **E&M+ FT**) outperforms the **E&M+ PT**. However, we observe that the BERT-QA, pre-trained on the SQuaD and finetuned on our corpus (**E&M+ PT + FT**) performs the best among the baseline variants in extracting the textual part of the answer.

For the MExBERT, we test a variant, where we remove the visual input by passing all image feature values as zero to the visual stream. We conduct this experiment to test if, even without supervision, visual input can enhance the textual answer. While not using supervision, we use average attention weights over a particular image to asses its fitness for the 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.

MODEL	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
E&M+ PT	39.95 (35.45)	38.61 (33.01)	40.43 (35.78	22.11 (20.07)
E&M+ FT	44.28 (39.23)	42.04 (34.96)	45.01 (39.88)	24.74 (22.88)
E&M+ PT + FT	46.77 (42.53)	43.26 (40.95)	47.22 (42.89)	25.17 (23.39)
MExBERT + Zeroed Visual Input	44.10 (38.67)	41.90 (34.82)	44.91 (39.47)	24.28 (22.33)
MExBERT	45.13 (37.79)	43.02 (33.93)	45.77 (38.12)	24.96 (22.87)
MExBERT + PS	45.67 (38.12)	43.59 (34.17)	46.17 (38.97)	25.04 (22.93)
MExBERT + PS + L PT	48.12 (42.06)	46.22 (40.27)	48.82 (42.91)	28.01 (23.99)
MExBERT + PS + V PT	46.18 (38.93)	44.11 (35.23)	47.24 (39.81)	25.89 (23.73)
MExBERT + PS + V+ L PT	**48.88** (42.36)	**47.02** (40.55)	**49.03** (43.01)	**28.50** (24.10)

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

We noticed a marginal improvement with 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 large datasets used usually in recent works, we considered pretraining to be a natural choice to improve our model further. Evidently, pretraining improves the model performance significantly. While the improvements with visual training are marginal for the scores (which is expected since this training is directed at visual stream), language pretraining yields reasonable improvements as shown in Table 2.

4.0.3 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, then, deem this image as 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 all

MODEL	PRECISION@1	PRECISION@2	PRECISION@3
Random	0.139 (0.139)	0.258 (0.258)	0.381 (0.381)
E&M + A	0.243 (0.244)	0.438 (0.438)	0.528 (0.528)
E&M + P	0.248 (0.239)	0.442 (0.436)	0.531 (0.531)
E&M + Q	0.259 (0.248)	0.448 (0.441)	0.542 (0.534)
E&M + Q + A	0.252 (0.243)	0.445 (0.445)	0.547 (0.547)
E&M + Q + P	0.251 (0.246)	0.441 (0.448)	0.532 (0.532)
E&M + P + A	0.244 (0.233)	0.434 (0.434)	0.526 (0.523)
E&M + Q + P + A	0.255 (0.246)	0.444 (0.439)	0.541 (0.542)
MExBERT	0.211 (0.209)	0.421 (0.417)	0.528 (0.511)
MExBERT + PS	0.268 (0.258)	0.449 (0.447)	0.544 (0.544)
MExBERT + PS + L PT	0.271 (0.260)	0.453 (0.449)	0.546 (0.546)
MExBERT + PS + V PT	0.288 (0.280)	0.459 (0.459)	0.549 (0.549)
MExBERT + PS + V+ L PT	**0.291** (0.286)	**0.459** (0.459)	**0.549** (0.547)

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

different combinations of Query **Q**, Passage **P**, and the extracted Answer **A** as the matching text. As seen in Table 3, paragraph and answer don't help as much in such retrieval as query does. We posit that this is due to the additional noise at paragraph level with no real understanding about the answer as well. In fact, surprisingly, we found the best performance among all such combinations while using the query alone. This proves that the query is the most important part in ensuring correct image retrieval even though other information (from passage and answer) can also be helpful. Please note that for all the **E&M** results in Table 3 and 4, we use the answer from its best performing model from Table 2 i.e. **E&M + PT + FT.**

The power of joint multimodal attention and understanding comes out very strongly while using MExBERT. 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 being trained to optimize for better textual answer generation. Applying our proxy supervision mechanism while training the model, we find a very significant improvement, even better than a large pretrained model like UNITER, specially in PRECISION @ 1 scores. PRECISION @ 2,3 scores are however similar to what we obtained with UNITER. 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 that much confidence due to lack of explicit understanding about where to focus on the text. Such an understanding is, however, the main strength of **MExBERT**. Pretraining yields very significant inprovments in scores especially on PRECISION@1 metric. The language pretraining also provides some noticeable marginal improvements.

4.0.4 Human Evaluation

While our proxy scores have been intuitively designed, it is 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 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. We notice significant improvement on human scores while performing joint input understanding with MExBERT. It is natural to expect the model to perform on the proxy scores on which it was trained on (albeit on a different dataset). The performance on the human annotations bettering that of proxy scores indicates 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. The visual pretraining task also provides more significant boost on human scores as compared to the boost it provides on proxy scores based metric for the same reason.

5 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 single 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.,

Model	Precision@1	Precision@2	Precision@3
Random	0.144 (0.144)	0.275 (0.275)	0.396 (0.396)
E&M + A	0.283 (0.271)	0.485 (0.478)	0.606 (0.606)
E&M + P	0.283 (0.271)	0.485 (0.478)	0.606 (0.606)
E&M + Q	0.285 (0.277)	0.492 (0.492)	0.612 (0.612)
E&M + Q + A	0.283 (0.274)	0.485 (0.478)	0.610 (0.610)
E&M + Q + P	0.283 (0.274)	0.487 (0.481)	0.610 (0.610)
E&M + P + A	0.283 (0.271)	0.484 (0.478)	0.606 (0.606)
E&M + Q + P + A	0.284 (0.277)	0.492 (0.492)	0.612 (0.612)
MExBERT	0.196 (0.185)	0.385 (0.382)	0.498 (0.498)
MExBERT + PS	0.316 (0.310)	0.505 (0.505)	0.608 (0.608)
MExBERT + PS + L PT	0.321 (0.320)	0.511 (0.511)	0.610 (0.610)
MExBERT + PS + V PT	0.381 (0.367)	0.535 (0.528)	0.616 (0.612)
MExBERT + PS + V+ L PT	**0.386** (0.372)	**0.538** (0.530)	**0.618** (0.612)
Proxy Scores	**0.422**	**0.631**	**0.753**

Table 4: Results showing the performance of E&M and MExBERT over the image modality of the multimodal answer as measured using Human Evaluation over test set

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.

6 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 apppropriate 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 to 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 element making the output requirement totally 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.

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, 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*.

Yansong Feng and Mirella Lapata. 2010. Visual information in semantic representation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 91–99.

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.

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.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

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.

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.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 721–732, Baltimore, Maryland. Association for Computational Linguistics.

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

Nitish Srivastava and Russ R Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230.

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.