Improving Multi-Modal Multi-Hop Web Question Answering

Nikhil Yadala [*1] **Paritosh Mittal** [*1] **Saloni Mittal** [*1] **Shubham Gupta** [*1]

1. Introduction

Humans can learn to gather information from different modalities and collate them effectively to answer complex questions. However, current AI agents cannot still effectively perform this task. The advent of deep learning and large scale data processing, has led to some progress in combining inputs from two modalities. The success of neural models on Visual Question Answering (V-QA) (Antol et al., 2015; Goyal et al., 2017) is a testament to this. However, the problem setting here ensures that sufficient information is available in visual and textual cues. This is not a very strong approximation for real world applications.

In real world the information is often scattered across multiple sources. Also, it is possible that to answer certain questions, we need context from multiple sources. The domain of multi-hop reasoning aims to tackle this problem. Here the AI agent is tasked to gather information from different sources and use the collated information to solve down-stream tasks such as QA.

A challenging and much more natural extension to multi-hop QA is multimodal multi-hop QA. Let us take the example of web search. Since web has information in the form of text, vision, speech etc. any query answering solution needs to be multi-modal. On top of that complete information about a query is rarely found in a single source and ideal system should (1) go through multiple sources; (2) look for similar patterns and (3) aggregate information based on reasoning. We believe that in order to achieve human-like question answering abilities for AI agents, one needs to transcend Visual Dialog and aim to achieve multi-modal multi-hop question answering.

In this project, our team (mmmlX) aims to target the problem of multimodal multihop question answering. The inclusion of multi-modal inputs in multi-hop question answering makes this a relatively nascent and unexplored domain. Key

Figure 1. Top: Sample query; Mid: Possible Sources; Bottom: Desired response

technical challenges in this problem include: (1) Extracting information from different modalities; (2) Selecting the relevant sources (from a set of sources); (3) Aggregating information from relevant sources; (4) producing a natural answer to the question based on the aggregated information.

As part of this project, our team will mainly aim to boost the performance over existing dataset (WebQA - (Chang et al., 2021)). We have identified certain research directions which are directed towards improving the performance and are influenced by the domain knowledge. On a high level these include: developing a class conditioned neural model which can leverage the semantic label of each question (whether it's a choice,YesNo or number etc.); using high level image representations (bounding boxes or segmentation maps) or compressed latent representations which can capture global information across inputs. Further details and more directions are included in Section 4.

We will use this Github repository for the project: `https://github.com/shubham-gupta-iitr/mmmlX`

2. Experimental Setup

WebQA (Chang et al., 2021) tries to capture the web search flow by providing the dataset with the intersection of text and vision modalities, although the way dataset is constructed is not truly multimodal since sources can either be text or images with captions but not both. Also, note that for true resemblance, we should ideally have other modalities like speech, tables - WebQA contains only text and

[*]Equal contribution [1]Carnegie Mellon University. Correspondence to: Nikhil Yadala <nyadala@andrew.cmu.edu>, Paritosh Mittal <paritosm@andrew.cmu.edu>, Saloni Mittal <salonim@andrew.cmu.edu>, Shubham Gupta <shubham2@andrew.cmu.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

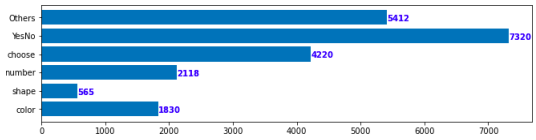

Figure 2. Data Distribution for different image based queries

image modalities. But that said, this is the closest dataset available resembling the web search behavior. Hence we propose to use this dataset.

Most of the older datasets for visual question answering have been too simplistic in terms of the task- classification over a fixed vocabulary of answers. OK-VQA (Marino et al., 2019) broadens this task by having open-ended questions but the images are only a part of the query and not the answers. This makes the task less similar to the fundamental problem of web search. There were several attempts for multihop datasets- QAngaroo (Welbl et al., 2017), HotpotQA (Yang et al., 2018b), and ComplexWebQuestions (Talmor & Berant, 2018). But all of these are text-based. MultiModalQA (Talmor et al., 2021b) does a great job in capturing the text, tables, and images as part of sources but the questions in this dataset are generated from pre-defined templates. This oversimplifies the model's task to determine the question template.

Dataset Description: WebQA dataset contains the text snippets and the images along with captions as the source. The queries are text-based. The responses are expected to be aggregated from multiple sources. Figure 1 shows the sample query, expected answer and the input sources.

The total dataset comprises of 41732 queries out of which training set has 36766 queries and validation set has 4966 queries. Each query can be either based on text-based sources or image based sources. The total text based queries are 20267 and remaining 21465 are image based queries. The image based queries can be further classified into following 6 categories- color, shape, number, choose, YesNo and Others. Figure 2 shows the data distribution for the different categories of image based queries.

Task Formulation: WebQA has 2 fold tasks- Given a question Q, and a list of sources S = s1, s2, ..., a system must a) identify the sources from which to derive the answer, and b) generate an answer as a complete sentence. Note, each source s can be either a snippet or an image with a caption. A caption is necessary to accompany an image because object names or geographic information are not usually written on the image itself, but they serve as critical links between entities mentioned in the question and the visual entities. Although is not exactly how data is present on the web, we need this to simplify the task.

Evaluation: We follow the standard evaluation metrics proposed by the WebQA paper. Since there are 2 tasks, we have separate metrics for each task. For task 1, where we need to identify the sources (Source retrieval) we use the F1 score as the evaluation metric. For task 2, the answer quality is measured using Fluency BARTScore(Yuan et al., 2021) and Accuracy(keywords overlap). The authors have modified the BARTScore to have a bounded normalization as shown

$$\mathbf{FL}(c, R) = \max \left\{ \min \left(1, \frac{BARTScore(r, c)}{BARTScore(r, r)} \right) \right\}_{r \in R} \quad (1)$$

3. Related Work

3.1. Multi-modal visual Q/A

Visual Question Answering (VQA) refers to predicting a solution to a question using cues from visual and textual modalities. The seminal work of (Antol et al., 2015) released a largescale VQA dataset and used deep learning to solve open-ended VQA. This work highlighted that VQA systems often need a more detailed understanding of the visual signals (images, videos etc.) to predict correct solutions. This work combined both vision and text modalities and used CNNs for images and LSTMs for text to extract uni-modal feature representations. These features were then fused (point wise multiplication) followed by fully connected layers to predict the correct 'class' of solution. This is one of the first approaches to use deep learning for large scale VQA. While VQA aims to answer a single question, natural spoken language is more of a dialog.

The work of (Das et al., 2017) introduced the task of visual dialog which is a more natural extension to VQA. Aim of this problem is to develop a conversational AI agent which can also process inputs from visual modality. The authors released VisDial - a large scale dataset for visual dialog, and proposed neural visual dialog models. However even with large scale datasets, there is a gap between the quantitative (reported) performance and the actual ability of models to generate realistic and diverse inputs.

The seminal work from (Murahari et al., 2019) aims to improve the generative ability of models for real life scenarios. They aim to answer diverse questions using a Q-BOT-A-BOT approach wherein Q-BOT is trained to ask diverse questions using a known description. This forces A-BOT to explore a larger state space and jointly answer more informatively. Even with this progress, the performance was still limited. Often, the models fail to optimally utilize the information in the input modalities and the results are significantly influenced by the biases in our data.

The work of (Goyal et al., 2017) was aimed to balance the VQA dataset by having almost twice the original image-question pairs. One interesting modification was that this dataset contained almost similar images with opposing answers to same questions. This pushes the neural models to

learn better relationships across inputs. In particular, this work aims to boost the importance of visual cues. Existing baselines often performed poorly on the modified dataset.

Even with these extensive related works, one important limitation is that visual cues is often limited to a single image. However, the context needed to answer many real life questions can span across multiple images. Hence advanced AI conversational agents need to embody multi-hop reasoning with ability to aggregate information from different sources.

3.2. Multi-hop QA

There has been substantial work in the recent years on building Question Answering (QA) models that can reason over multiple sources of evidence. The most early works in this domain focussed on a single modality mainly answering text questions from a set of textual sources. In 2018, (Yang et al., 2018a) introduced a QA dataset, HotPotQA that required reasoning over multiple supporting documents. The paper re-implemented the model proposed by (Clark & Gardner, 2018) to benchmark results on this dataset. They used an RNN-based architecture that combined character-level models, self-attention and bi-attention. Given a question and a pool of paragraphs as input, the model was designed to answer "yes"/"no" or span-based answers. They also showed significant performance improvements on the QA task by using strong supervision in the form of a joint learning objective on a supporting fact prediction task. Looking for answers in the wild from 5,000,000+ wiki paragraphs requires retrieval of a subset of relevant paragraphs that can be fed in the trained model. They use a cascade of two filters, first an inverted-index based filtering strategy and then a final pool of 10 candidates using tf-idf based retrieval.

More recently, this line of work has advanced into building multi-hop multimodal QA systems that can answer complex multi-hop questions by reasoning over multiple modalities like images, text and tables. In one such work, MultiModalQA, (Talmor et al., 2021a) proposed a Multi-hop decomposition (ImplicitDecomp) model that first uses a RoBERTa-based question-type classifier on the question, where question-type pertains to a program specifying the modalities and the order in which to approach them in a multi-hop setting. In each hop, the ImplicitDecomp model is fed the question, predicted question type, the hop number and the context of the corresponding modality. The model automatically identifies which part of the question is relevant at the current hop and activates the unimodal QA module of the corresponding modality to generate an answer. This intermediate answer is fed as input to the next hop so that it can leverage from this information as well before generating the final answer.

MIMOQ (Singh et al., 2021) have taken multimodal QA one step further by introducing the ability of the QA system to not just reason but also respond in multiple modalities. They propose a novel multimodal framework called MExBERT (Multimodal Extractive BERT) that uses joint attention over input textual and visual streams for extracting multimodal answers given a question. The authors highlight the importance of a generic joint understanding framework of multimodal input (vision and language in this paper). They show noticeable improvements using MExBERT that leverages cross-modal learning over other baselines that independently extract answers from unimodal QA modules after identifying the answer modalities.

WebQA differs from MIMOQ, where the task introduces an additional challenge of information aggregation and summarization before producing the final natural language answer while MIMOQ is limited to producing a multimodal answer (like a pair of an image and text snippet) without the answer being necessarily interpretable.

3.3. Cross modality representations

All the SOTA models for the tasks discussed so far draws their power by initializing the representations from pretrained models of various tasks involving image and language modalities. There are two major directions of pretraining; 1. Parallel streams of encoders one for each modality followed by fusion (Tan & Bansal, 2019)(Lu et al., 2019) 2. Unified encoder-decoder representations that can take both the language or image modalities (Zhou et al., 2019). WebQnA baseline is implemented from the VLP pretrained model (Zhou et al., 2019) which trains unified representations for both language and Image. For the text inputs, the sentence is tokenized using word tokenizers like is the case with all the BERT models. However, for the image modality, traditionally, the features of ImageNet or ResNet backbone are used to encode the features. However, (Anderson et al., 2017) proposed a new technique of extracting the objects from the image, using Faster-RCNN, and the image is represented as a set of embeddings of the Regions of Interest (ROI). Along with the object features, the positional features (spatially) are also added to the encodings - representing the positions of the bounding boxes.

Other works like ViLBERT (Lu et al., 2019), LXMERT (Tan & Bansal, 2019) train separate encoders for each modality-followed by cross modality encoders. LXMERT is trained on 5 pretraining tasks; 1) Masked cross modality Language modelling 2) Masked object prediction via ROI feature regression, 3)Masked object prediction via detected-label, 4)Classification, 5) Cross-modality matching, 6) Image question answering. The first 2 tasks encourage the unimodal representations to be as representative of the modality as possible. The latter tasks are used to train the cross modality encoders. ViLBert takes a similar approach too, but is trained on only two pre-training tasks; predicting

semantics of masked words and image regions given the unmasked inputs, predicting whether an image and text segment correspond to each other. However, ViLBERT is extensively validated a better performance on 4 downstream tasks - VQA, Visual commonsense reasoning, Referring expressions and caption based image retrieval., Faster-RCNN with Resnet Backbone is used to encode visual representations. The results described stand as a strong evidence that the models trained by transfer learning from sibling-tasks provide a better way of solving webQnA e2e problem.

3.4. Evaluation

The conventional metrics for text-based modality BLEU, METEOR, ROUGE (Papineni et al., 2002; Lavie & Agarwal, 2007; Lin, 2004) are known to correlate poorly with human judgement in evaluating dialog responses. The VQA tasks have been retrieval based for single word answer accordingly the metrics used for them have been accuracy based metrics like F1-score, precision, recall etc. Later on the task of VQA was extended to more generalized task of visual dialog. These methods have more generalized longer responses to the queries and hence proposed to use new class of metrics namely normalized discounted cumulative gain (NDCG) (Järvelin & Kekäläinen, 2000) and Mean Reciprocal Rank (MRR) metrics. But again, these were retrieval based and hence biased towards discriminative modeling.

FlipDial (Massiceti et al., 2018) proposed to use a generative approach for Visdial and in concordance they also proposed to use the metric of KL divergence(KLD) and cross entropy. The fundamental problem with this metric is again that they tend to use only the word matching and no semantic meaning of the word. Later, BERTScore (Zhang* et al., 2020) was proposed which was based on semantic proximity of the generated response to the ground truth. But, this also has problem since it indiscriminately treats all colors or all shapes as nearly identical.

4. Research Ideas

Since the WebQA dataset is recent, we plan to explore multiple research directions.

Exploiting question category information : The dataset has following labels attached to each query: 'color', 'shape', 'number', 'choose', 'YesNo', 'text' and 'Others' as covered in 2. One way is to have a joint encoder followed by separate classification (or generation) heads for each category. This can improve the BARTScore as we expect model to not confuse between categories. During inference we first use clustering to find question class followed by generation.

Jointly optimizing source selection: Existing models only provide positive sources as input to the QA model. However, this too introduces a bias in the training pipeline as the answering module is not discarding negative information. We propose to jointly train the model with both hard positive and negative sources as input. The model is tasked to also predict (classify) the positive sources. We expect that this will enforce the model to solve a broad problem and we can also eliminate the first task (of explicit source selection).

Pre-training on related datasets and tasks:

The unified VLP model 1 performance degrades when compared to the case in which Image modality related questions are trained with only the image modality sources. (VLP$^{\mathcal{I}}$).

	Query Type	Image	Text
Limited	Lexical overlap	57.62	45.17
	VLP	68.46	69.70
	VLP$^{\mathcal{I}}$/VLP$^{\mathcal{T}}$	73.16	70.13
Full	Lexical overlap	44.83	33.78
	VLP	68.13	69.48

Table 1. Source Retrieval (F1 ↑) when the model is provided sources only in the correct modality versus the full set.

We can look into training two separate unified representation models. Further, the cross modality representations can be computed by attending to other modalities key value pairs during the finetuning stage. This approach allows us to tune the representations for each modality individually for the downstream tasks. We may have to pre-train or finetune the models on other tasks as described in (Lu et al., 2019) (Tan & Bansal, 2019) to improve accuracy.

Using Compressed Latent Space: Transformers with multi-modal inputs are often used to learn cross modal representations and perform well in QA tasks. However, these models are greatly limited by their input lengths. Hence for visual inputs we cannot simultaneously process using their current representations. One idea is to learn a compressed and discrete latent representation (similar work done by (Esser et al., 2021) as Taming-transformers) and use that for source retrieval and multi-hop reasoning. Team believes that F1 score for source retrieval should improve majorly using this approach.

Joint Optimization on Positive Evidence Regeneration: Another potential approach is to enforce the QA model to not only generate the answer, but also generate the compact image representations of the positive sources. This allows the NLG model to learn to attend to the correct visual clues in the images, while discarding the distractor objects in the images. This also helps in interpretability of the answer generation model.

In addition to these primary ideas, the team also has some other interesting thoughts of exploring graph-based methods that are popular in common-sense reasoning literature. We find all these directions very promising and will test their effectiveness on the WebQA dataset.

References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. Bottom-up and top-down attention for image captioning and VQA. *CoRR*, abs/1707.07998, 2017. URL http://arxiv.org/abs/1707.07998.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

Chang, Y., Narang, M., Suzuki, H., Cao, G., Gao, J., and Bisk, Y. WebQA: Multihop and Multimodal QA. 2021. URL https://arxiv.org/abs/2109.00590.

Clark, C. and Gardner, M. Simple and effective multi-paragraph reading comprehension. *ArXiv*, abs/1710.10723, 2018.

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M., Parikh, D., and Batra, D. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Esser, P., Rombach, R., and Ommer, B. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Järvelin, K. and Kekäläinen, J. Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pp. 41–48, New York, NY, USA, 2000. Association for Computing Machinery. ISBN 1581132263. doi: 10.1145/345508.345545. URL https://doi.org/10.1145/345508.345545.

Lavie, A. and Agarwal, A. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pp. 228–231, USA, 2007. Association for Computational Linguistics.

Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Lu, J., Batra, D., Parikh, D., and Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *CoRR*, abs/1908.02265, 2019. URL http://arxiv.org/abs/1908.02265.

Marino, K., Rastegari, M., Farhadi, A., and Mottaghi, R. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Massiceti, D., Siddharth, N., Dokania, P. K., and Torr, P. H. Flipdial: A generative model for two-way visual dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Murahari, V., Chattopadhyay, P., Batra, D., Parikh, D., and Das, A. Improving generative visual dialog by answering diverse questions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pp. 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

Singh, H., Nasery, A., Mehta, D., Agarwal, A., Lamba, J., and Srinivasan, B. V. Mimoqa: Multimodal input multimodal output question answering. In *NAACL*, 2021.

Talmor, A. and Berant, J. The web as a knowledge-base for answering complex questions. In *North American Association for Computational Linguistics (NAACL)*, 2018.

Talmor, A., Yoran, O., Catav, A., Lahav, D., Wang, Y., Asai, A., Ilharco, G., Hajishirzi, H., and Berant, J. Multimodalqa: Complex question answering over text, tables and images. *ArXiv*, abs/2104.06039, 2021a.

Talmor, A., Yoran, O., Catav, A., Lahav, D., Wang, Y., Asai, A., Ilharco, G., Hajishirzi, H., and Berant, J. Multimodal{qa}: complex question answering over text, tables and images. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=ee6W5UgQLa.

Tan, H. and Bansal, M. LXMERT: learning cross-modality encoder representations from transformers. *CoRR*, abs/1908.07490, 2019. URL http://arxiv.org/abs/1908.07490.

Welbl, J., Stenetorp, P., and Riedel, S. Constructing datasets for multi-hop reading comprehension across documents. *CoRR*, abs/1710.06481, 2017. URL http://arxiv.org/abs/1710.06481.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., and Manning, C. D. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*, 2018a.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., and Manning, C. D. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018b.

Yuan, W., Neubig, G., and Liu, P. Bartscore: Evaluating generated text as text generation, 2021.

Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K. Q., and Artzi, Y. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J. J., and Gao, J. Unified vision-language pre-training for image captioning and VQA. *CoRR*, abs/1909.11059, 2019. URL http://arxiv.org/abs/1909.11059.