WebQA Team 6, TBD

Haofei Yu [*1] **Jiyang Tang** [*1] **Ruiyi Wang** [*1] **Ziang Zhou** [*1]

1. Introduction

Remember to mention our github repository https://github.com/tjysdsg/MMML-Fall22

2. Experimental Setup

2.1. WebQA Benchmark

WebQA (Chang et al., 2021) is a multimodal open-domain question answering benchmark. It focuses on the ability to extract and aggregate information from text and images. Its data is extracted from the internet and carefully designed so that the answers cannot be directly copied from an existing questions or images, and that both the vision and text modalities must be used to correctly answer the questions.

In most existing VQA benchmarks, a question is about a pair of images, thus making the image itself a query. However, images are a knowledge source based on which the machine learning models reason about the questions. This can encourage the models to learn common sense from the data and answer questions better. In addition, WebQA uses a new evaluation metric to encourage the answers to be in the form of natural language sentences, instead of a word, a short sentence, or a choice from possible answers. In other words, instead of producing a simple "yes" or "no" answer, the model needs to return a full a fluent sentence that answers the question logically.

Its task is formulated in two stages, source retrieval, and question answering. The model identifies the data sources to derive the answers during source retrieval. This includes images and their descriptions. During question answering, the model derives its answer from retrieved sources.

2.2. Data

WebQA's data is crowdsourced. Each annotator are presented with six distinct but related images and they produce three question-answer pairs from these images. Each pair requires one or two images out of the six to be answered correctly. Meanwhile, annotators are instructed to avoid questions that are simple facts, easily answered by a text-only or image-only search, or tied to a specific image. Then Hard Negative Mining is used to produce a set of hard negatives. The questions are categorized into open and closed classes. Closed classes include colors, shapes, numbers, and yes or no questions. Open classes include open-ended questions.

In total, the data contains 34 thousand question-answer pairs with 390 thousand images. The average length of questions is about 17.5 words while the average length of standard answers is around 12.5 words.

An example question-answer pair is listed below.

- Question: Are both the National Museum of the American Indian in Washington, D.C., and the Xanadu House in Kissimmee, Florida the same color?
- Standard Answer: Yes, both the National Museum of the American Indian in Washington, D.C., and the Xanadu House in Kissimmee, Florida is beige.
- Topic: Strange architecture
- Question Category: Yes/No
- 2 positive images and 16 negative images with relevant text

However, the images are not directly used in the baseline models, as discussed in the next section.

2.3. Feature Extraction

Text input in the questions, answers, textual sources and image captions are tokenized by the **Bert-base-cased** (Devlin et al., 2018) tokenizer.

Images are represented with 100 regions produced by an object detection model. The object detection model used is a variant of Faster RCNN with a ResNeXt-101 FPN backbone, the same one used in the VLP VQA task (Zhou et al., 2019). The authors also experimented with the latest state-of-the-art image representations from VinVL (Zhang et al., 2021).

2.4. Baselines

Baselines

2.5. Results

Metrics + results

3. Related Work

3.1. Multimodal Datasets

3.2. Pretraining Methods

3.3. VQA

4. Research Ideas

4.1. 1234

4.2.

References

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

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. Vinvl: Making visual representations matter in vision-language models. *CoRR*, abs/2101.00529, 2021. URL https://arxiv.org/abs/2101.00529.

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.