Midterm Report

Haofei Yu Jiyang Tang Ruiyi Wang Ziang Zhou

Abstract

It is important to examine multi-hop web-based VQA datasets such as WebQA because such benchmarks require information retrieval on both image and text sources, better aggregation and summary of knowledge, and higher reasoning ability on open-domain questions. By implementing and analyzing the baseline models on WebQA, we found that the image resources are not fully extracted and understood and the model is heavily dependent on text. In addition, the large pre-trained VLP model is very time- and memory-consuming. Therefore, our goal is to find a better way to align text and image via image-text matching loss and multimomdal cross-attention module. We also aim at reducing the model size and improving the training speed by applying a lighter model via detector-free visual encoders and knowledge distillation.

1. Introduction

Visual Question Answering in the open domain is an important topic in multimodal machine learning. Compared to closed-domain settings where the main focus is object detection with a fixed vocabulary of responses (Zhu et al., 2015), we are more interested in studying the challenges of visual representation learning, knowledge retrieval, and effective language generation in the open domain. WebQA is a novel multi-hop and open-domain VQA benchmark for large-scale text-image alignment, knowledge retrieval in both modalities, and question answering in multiple categories (Chang et al., 2021). The fine-tuned model applied on WebQA is pre-trained on VLP, a unified encoder-decoder transformer largely used in VQA tasks (Zhou et al., 2019). However, the results show that the model fine-tuned on VLP is more capable of pure-text questions rather than image-based ones, indicating the failure to extract enough visual information and fully utilize the interconnections between text and images. In addition, the increasing scale of the VLP pre-trained model leads to parameter inefficiency in fine-tuning (Chen et al., 2022), urging us to break the traditional paradigm and improve the training and inference efficiency of the pre-trained model. Based on the challenges posed by the benchmark, we are particularly interested in the following three research questions: 1) apply image-text matching loss to source information retrieval 2) replace inefficient VLP encoders with detector-free visual encoders, 3) improve text-image alignment by applying the cross-attention module in late fusion, and 4) augment the multimodal model performance by distilling knowledge from the pure-text model to the multimodal model.

WebQA differs from other open-domain VQA benchmarks such as OK-VQA, which requires knowledge aggregation but the images are only used as part of the query rather than the knowledge source (Marino et al., 2019). As a step further to previous multimodal VQA datasets, one big challenge posed by the multi-hop nature of WebQA is bringing visual representation learning into the information retrieval stage and aggregating text and image knowledge for response generation. There are few state-of-the-art models giving a satisfactory performance on WebQA as the dataset is very new and different. The fine-tuned approach is based on VLP while the few-shot approach is built upon PICa (Yang et al., 2021), the currently strongest model on OK-VQA. Both approaches are not specifically tailored to multihop information retrieval and thus not performing well. In addition, the top model (ITL) on the WebQA leaderboard is heavily dependent on transforming the images into text and generating responses based on the combined text embeddings, reducing the model to unimodal settings. Therefore, it is worth researching whether the proposed models extracted useful visual representations or were just dependent on text resources and discovering better ways to integrate text and image context using the cross-attention mechanism and knowledge distillation.

In this paper, our proposed work will shed light on the problem of image-text integration and efficient fine-tuning scheme on the multihop WebQA dataset. By applying image-text matching loss to information retrieval, we are able to reduce the size of the baseline VLP model and make the training process more efficient. By replacing VLP encoders with detector-free visual encoders, we will allow for quicker visual feature extraction suitable for our current hardware environment and configuration. By applying knowledge distillation in our multimodal settings, we are able to make our model lighter while rigorously monitoring the loss in different parts of the pipeline. Moreover, we will

further discover the actual effect of cross-modal attention on the multi-hop dataset and do a comparison between unified attention and cross-attention.

We will first present an overview of the field related to our research questions as well as the highlights of our proposed work. Then we will give a rigorous mathematical formulation of the problem we aim to solve. One important aspect of this paper is the analysis of baseline results, so we will formulate the baseline models mathematically and provide the experimental methodology we applied to implement the baseline models. After that, we will give a concrete error analysis of the baseline models and discuss our hypotheses for the possible reasons for the baseline model's pitfall. Finally, we will present our four new research ideas.

2. Related Work

2.1. Image-Text Matching Loss

Image-text matching is at the core of cross-modal information retrieval. The motivation for image-text matching is that the words in a sentence correspond to some region in the image and the whole sentence is a weak annotation. Prior work on image-text matching demonstrates the effectiveness of inferring the latent region-word correspondences (Karpathy & Fei-Fei, 2014) and Lee et al. (2018) proposed a stacked cross attention strategy that differentially attends to the alignment between image and text. Another method of cross-attention matching is to utilize visual semantic reasoning, where text and images are mapped to the same semantic space and then the alignment is optimized based on a hinge-based triplet ranking loss (Li et al., 2019). As a step further, some lines of research sampled negatives offline from the entire training set together with adaptive quintuplet loss instead of using triplet loss to find hard negatives (Chen et al., 2020). For BERT-based models, the hidden output of [CLS] is fed into a binary classifier to optimize the binary cross-entropy loss (Gao et al., 2020). Instead of sticking to unified encoder-decoder models, we aim to use a multimodal encoder enabling cross-attention between text and image where ITM loss will be applied to optimize the binary classification of hard negatives (Li et al., 2021).

2.2. Detection-free Visual Encoder

VLP is the baseline model for WebQA, which jointly learns visual and text representations within a unified multimodal encoder (Zhou et al., 2019). However, the image representations are extracted from faster RCNN, which leads to a higher cost of memory consumption even though the time consumption at the proposal stage is improved (Ren et al., 2015). In addition to the existing problem of extracting RCNN features, object detection can be both annotation-expensive and compute-expensive because it requires bounding box annotations during pre-training and high-resolution images during inference (Li et al., 2021). To solve the potential expensiveness and inefficiency of the VLP pre-trained model, detector-free visual encoders are proposed. ALBEF uses a detector-free visual encoder and a text encoder and fuses both features via a multimodal encoder (Li et al., 2021). Their proposed detector-free visual encoder is based on ViT-Base/16 because the scaling of Vision Transformer is likely to lead to improved performance (Dosovitskiy et al., 2020). In our proposed work, we aim at replacing the fast RCNN feature extraction with detector-free visual encoders to improve the speed while maintaining the performance.

2.3. Cross-modal Attention

The concept of cross-modal attention is proposed in the vanilla transformer (Vaswani et al., 2017), where two separate embedding sequences of the same dimension are combined. The idea of cross-attention has been extended to multimodality settings. The multimodal transformer (MulT) proposed by Tsai et al. (2019) is an end-to-end transformer that is able to learn representations from unaligned modalities. Following the paradigm of the cross-attention module of the vanilla transformer, MulT uses the pivot modality as queries and the dependent modality as keys and values and generates cross-modal attention which is then passed to a self-attention transformer. There are other lines of research applying multimodal cross-attention for image-text matching by jointly modeling the intra-modality relationship and intermodality relationship of image regions and sentence words in a unified deep model (Wei et al., 2020).

2.4. Knowledge Distillation

Knowledge Distillation is widely applied in language models. Kim & Rush (2016) adopted knowledge distillation for sequential model compression. (Izacard & Grave, 2020) proposed a reader-retriever system for weakly-supervised documents in question answering. For transformer-based models such as DistillBERT, which applied a cosine embedding loss on the basis of hidden embedding in the transformer block and a soft-target probability loss to reduce the model size by 40% (Sanh et al., 2019). The method of knowledge distillation will greatly reduce the size of the pre-trained model and improve the efficiency of the entire pipeline. There is a new line of research that applies knowledge distillation to visual-linguistic models. Fang et al. (2021) used a new compressing scheme where both the Teacher and Student use the same lightweight object detector and several loss terms are enforced on the Teacher and Student network to align their attention weights as well as the classification correctness during fine-tuning. Taking a step further, the work proposed by Wang et al. (2022) improves the previous DistillVLM by adaptively distilling useful knowledge from pre-trained encoders to cross-modal VL encoders. However,

this approach is very complicated and we will not integrate this approach into our research plan.

To our knowledge, this paper is the first to thoroughly implement and examine the pitfall of the baseline models on the newly released dataset WebQA. Our research questions are focusing on building lightweight pre-trained models for visual question-answering tasks, as well as discovering the boundary of unified attention compared with cross-attention.

3. Problem Statement

The high-level research objective of WebQA is to simulate the multimodal and multi-hop searching behaviors of humans. It is publicly agreed that VQA consists of two stages: query and generation (Chang et al., 2021). In this work, our group focuses on building two baseline systems for the retrieval task and presents ideas that are practical for improvements.

Note that the RCNN feature files provided by the WebQA official repository have some image feature missing, creating obstacles for training. Since it is extremely difficult and expensive to set up an environment to extract RCNN features, we removed 24 sources with missing RCNN features in total.

The retrieval problem, from a very low level, is to use a query to retrieve resources in the order of relevance. Therefore, this problem can be abstracted into two steps, the form a query and to compare the matching degree. For the query construction period, multiple pretrained models are applied. For the textual information, including, questions, answers, textual facts, and image captions, textual embeddings are tokenized using `bert-base-cased` tokenizer(Devlin et al., 2018). Image representations consisted of 100 bounding boxes, are extracted with the first fully-connected layer from ResNeXt-101 FPN backbone pretrained on Visual Genome(Krishna et al., 2017). The query of each pass is formatted as `<[CLS], si, [SEP], Q, [SEP]>`, and the outcome p_i is the confidence that the source is selected.

There is a slight difference between the two baselines we proposed. In the RoBERTa baseline, the `si` does not contain image representations. Instead, it has extra textual information generated from image captions. In the VLP baseline, both image and textual representations are bounded in `si`.

The objective function for the retrieval task can be written as the following,

$$Loss_{retrieval} = \sum_{s_i \in \mathcal{G}} logp_{s_i} + \sum_{s_i \in \mathcal{D}} log(1 - p_{s_i}) \quad (1)$$

where $\mathcal{G}$ and $\mathcal{D}$ denotes the gold sources and distractor sources respectively (Chang et al., 2021).

4. Multimodal Baseline Models

4.1. VLP Baseline

The official WebQA baseline uses a unified Visual-Language Pretraining (VLP) architecture for source retrieval and question answering (Zhou et al., 2019). This architecture is essential to a BERT-base (Devlin et al., 2018) model with a different pretraining method. The input of the model consists of a sequence of image region embeddings followed by text embeddings.

The image region embeddings $\{M_1, \ldots, M_N\}$ are calculated using the result of a pretrained object detector. Each patch is composed of region features $R = [R_1, \ldots, R_N]$, region object labels $C = [C_1, \ldots, C_N]$, region geometric information (bounding box coordinates) as $G = [G_1, \ldots, G_N]$, where N is the number of regions.

$$M_i = W_r R_i + W_p \left[\text{LN}(W_c C_i) | \text{LN}(W_g G_i)\right] \quad (2)$$

where W is the weight of each feature, $\text{LN}(x)$ is the layer normalization of x, $|$ means concatenation, and the bias and nonlinearity terms are omitted.

The embeddings of text tokens $\{T_1, \ldots, T_K\}$ are extracted using a pretrained tokenizer, where K is the number of text tokens.

The overall model input X is the concatenation of the three embedding sequences, separated by special tokens.

$$X = \texttt{[CLS]}, M_1, \ldots, M_N, \\ \texttt{[SEP]}, T_1, \ldots, T_K, \texttt{[SEP]}$$

where $E(x)$ is the embedding of token x returned by BERT.

The training object of VLP is the same as the ones used in BERT (Devlin et al., 2018), masked language modeling objective. After pretraining, the model can be finetuned to perform source retrieval and visual language answering. During finetuning, a multi-layer perceptron (MLP) on top of the element-wise product of the last hidden states of `[CLS]` token and `[SEP]` token is trained, same as (Lu et al., 2019). The input sequence for finetuning stays the same as before, but the training objective is VQA-specific.

In the context of the WebQA baseline, the input sequence is slightly different.

For source retrieval, if the input is an image source, the model input is

$$X = \texttt{[CLS]}, M_1, \ldots, M_N, I_1, \ldots, I_N, \\ \texttt{[SEP]}, Q_1, \ldots, Q_K, \texttt{[SEP]}$$

where $\{I_1, \ldots, I_N\}$ is the embeddings of the image caption, and $\{Q_1, \ldots, Q_K\}$ is the embeddings of the question sentence with length K.

If the input is a text source, the model input is

$$X = \texttt{[CLS]}, T_1, \ldots, T_J, \texttt{[SEP]}, Q_1, \ldots, Q_K, \texttt{[SEP]}$$

where $\{T_1, \ldots, T_J\}$ is the embeddings of the text fact.

The training objective is the binary cross entropy between the predicted probability x generated by the MLP and the true label y:

$$CE(x, y) = y \cdot \log x + (1 - y) \cdot \log(1 - x) \quad (3)$$

During inference, the output of the MLP is converted to a binary discrete value using a threshold H, 1 meaning that the source is selected and vice versa.

For question answering, the input of the image fact or the text fact sample is

$$X = \texttt{[CLS]}, M_1, \ldots, M_N, I_1, \ldots, I_N, \texttt{[SEP]}, Q_1, \ldots, Q_K, A_1, \ldots, A_K, \texttt{[SEP]}$$

or

$$X = \texttt{[CLS]}, T_1, \ldots, T_J, \texttt{[SEP]}, Q_1, \ldots, Q_K, A_1, \ldots, A_K, \texttt{[SEP]}$$

where $\{A_1, \ldots, A_L\}$ is the embeddings of the answer sentence with length L. The training objective is masked language modeling. The answer is generated by repeatedly appending a `[MASK]` token to the end of the input and replacing it with a predicted token and appending a new `[MASK]` for the next step. The generation stops when the output is `[SEP]`, `[PAD]`, or the maximum length is reached. Beam Search is performed to find the best output sequence.

Meanwhile, M_i is generated by a faster RCNN (x101fpn) (Ren et al., 2015) trained on VisualGenome (Krishna et al., 2017) dataset, text embeddings are extracted using a pretrained Bert-base-uncased model, G is normalized to $[0, 1]$, and N is set to 100.

During the training process of either of the tasks, a mini-batch contains both samples with an image source and samples with a text source.

4.2. RoBERTa baseline

When considering the retrieval task in the WebQA pipeline, one way is to use one model like Vision-Language Pretraining and rely on the pretrained multimodal model to fuse information in different modalities. However, there is another way to think of doing the multimodal retrieval task. In order to reach better multimodal retrieval performance, we rely on the assumption that large-scale pretrained language models are much better than pretrained visual-language models. As a result, we separate the retrieval pipeline into two separate stages. In the first stage, image caption models are used to help us convert images into useful text. In the second stage, a text-only retriever like RoBERTa (Liu et al., 2019) is used to retrieve the related source.

For the first stage of our pipeline, instead of using image region embeddings $\{M_1, \ldots, M_N\}$ calculated by a pretrained object detector, one image is directly encoded using Vision Transformer (Dosovitskiy et al., 2020) and the encoded features can be defined as $\{f_1, \ldots, f_N\}$. These encoded features are sent into one pretrained language model like GPT-2 (Radford et al., 2019) to generate text that is related to the image. Mathematically speaking, the generated text can be defined as $\{T_1, \ldots, T_N\}$. The semantic information that is included in the corresponding image and is helpful for retrieval and question generation task is represented in concrete text form instead of features.

For the second stage of our pipeline, we use separate encoders to encode each image or text fact in our source and predict whether it is a positive fact or a negative fact. With the help of $\{T_1, \ldots, T_N\}$ obtained from the previous stage of the pipeline, the input tokens of the text-only retriever for QA pairs can be formally written as:

$$X = \texttt{[CLS]}, T_1, \ldots, T_N, T_{N+1}, \ldots T_{N+M}, \texttt{[SEP]}, Q_1, \ldots, Q_K, \texttt{[SEP]}$$

For image-based questions, $\{T_1, \ldots, T_N\}$ stands for the generated text from the image and $\{T_{N+1}, \ldots T_{N+M}\}$ stands for the provided image caption from the dataset. For text-based questions, $\{T_1, \ldots, T_N\}$ stands for the text fact in the dataset and $\{T_{N+1}, \ldots T_{N+M}\}$ stands for the title of the text fact from the dataset.

Same with the VLP baseline mentioned above, the training objective for the text-only retriever is also the binary cross entropy between the predicted logits x coming from the [CLS] classifier and the pre-defined ground truth label y:

$$CE(x, y) = y \cdot \log x + (1 - y) \cdot \log(1 - x) \quad (4)$$

For the inference process, each source image is first sent into the ViT+GPT-2 image caption pipeline to get its text form and is concatenated with its title. For text facts, each fact is concatenated with its corresponding title. After that process, both image-based facts and text-based facts are concatenated with the question to get its final input format. Therefore, based on the input format, prediction results are obtained using text-only retriever.

For optimization, both the image caption model and text-based retrieval model are optimized using the AdamW optimizer.

5. Experimental Methodology

5.1. WebQA Dataset

WebQA data (Chang et al., 2021) is composed of a large list of questions. Each question contains several "facts", or sources, some of them must be used to correctly answer the question (gold sources, or positive facts), while others are not relevant to the question (distractors, or negative facts). The facts are either images with their caption text or text snippets. The answers to these questions are in natural language form. Compared to other QA datasets that only have single-word or double-word answers, WebQA requires more powerful answer generation. The data contains 36,766 training questions, 4,966 validation questions, and 7,540 test questions. Based on which modality the gold sources have, the whole dataset can be classified into two types of questions: image-based and text-based.

Train	#Img(+)	#Img(-)	#Txt(+)	#Txt(-)
Image-based QA	1.44	15.85	0	15.35
Text-based QA	0	11.61	2.03	14.62
Dev	#Img(+)	#Img(-)	#Txt(+)	#Txt(-)
Image-based QA	1.44	15.88	0	15.29
Text-based QA	0	11.78	2.04	14.63

Table 1. Statistical information about both image-based and text-based negatives and positives. Each number in this table represents the average number of text-based and image-based positives and negatives that one sample has.

It is worth noticing that for image-based question-answering pairs, text positives are not provided while for text-based question-answering pairs, image positives are not provided. As a result, we calculate statistical information about positive samples and negative samples separately on image-based question-answering pairs and text-based question-answering pairs. Table 1 shows how many average positives and negatives belong to each sample in the dataset.

5.2. Evaluation Metrics

The evaluation metrics we used are the same as the ones used in the WebQA benchmark. For source retrieval, a given number of facts are fed into the model one at a time, and the model outputs the probability of selecting that fact as relevant to the question. On *restricted* setting, the model is given only about 40 facts, which are within the general topic of the question. On *full* setting, the model is given all of the facts in the WebQA dataset. We only test our model on the *restricted* setting. The metric is the F1 score of correctly selected positive sources. The evaluation metric for question answering is the multiplication of fluency score, calculated from BARTScore (Yuan et al., 2021), and a score **acc** of keyword occurrences.

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

$$\mathbf{acc} = \begin{cases} \text{F1} & \text{if QType} \in \text{[color,shape,number,Y/N]} \\ \text{Recall} & \text{otherwise} \end{cases}$$

The keywords are estimated from standard answers depending on the specific question categories. For example, for *Color* category a list of colors, such as "orange" and "yellow", is searched in the standard answer to obtain the keywords.

In this report, we mainly focus on the source retrieval baselines so that future models can have a clear comparison subject to evaluate their multimodal information processing capabilities.

5.3. VLP Baseline Setting

Same as the official baseline, we find that the best performing threshold H on the validation set is 0.2, so we set $H = 0.2$ in all testing procedures. The max number of facts trained in a batch is set to 32 during training, and 40 during inference. Excessive facts are randomly truncated. We use the AdamW (Loshchilov & Hutter, 2017) optimizer with default parameters ($\beta_1 = 0.9$, $\beta_2 = 0.99$), setting the initial learning rate as $3e-5$, with a cosine learning rate scheduler with warmup steps being 0. And gradient accumulation step is 128.

5.4. RoBERTa Baseline Setting

Concerning the hyper-parameter setting for the RoBERTa baseline, in the image caption model, we set max length = 32, and beam size = 4. For the retrieval model, we set classifier threshold = 0.3, use max choice number = 16 for positive and negative training, and use the same optimizer as above but with an initial learning rate of $5e-5$. The learning rate scheduler is also the same.

6. Results and Discussion

In this section, we show the retrieval results of our two baselines. We submitted our inference results on the test dataset to WebQA's official test server to produce a convincing result. Moreover, we do error analysis and compare model performance based on the validation dataset (on both full-size and image-based only sub-part).

6.1. Baseline Results

The VLP baseline achieved an F1 score of 67.96 on the test set using $H = 0.2$, which is slightly lower than the results from (Chang et al., 2021). The F1 score of text-based questions in the validation set is 70.07, and the F1 of

image-based questions is 67.88, as shown in Table 2

The RoBERTa baseline achieved a better F1 score of 75.73 on the test set using $H = 0.3$. However, when tested on the validation set, the F1 score of the RoBERTa performs badly on image-based questions and well on text-based questions. Therefore, the RoBERTa baseline achieves an overall better F1 score on the entire test set.

VLP Baseline	F1	Precision	Recall
Val text-only	70.07	-	-
Val image-only	**67.88**	**65.20**	78.48
Test	67.96	-	-
RoBERTa Baseline	**F1**	**Precision**	**Recall**
Val text-only	**76.22**	**67.86**	**86.95**
Val image-only	65.75	52.50	**87.92**
Test	**75.73**	-	-

Table 2. Overall Classification Metrics of the Baselines

Note that image-based questions have image caption data and negative text facts, so this subset contains two modalities. However, the text-based questions have only text facts, which is a unimodal subset.

We also performed some more fine-grained tests. Since we have no control over the testing procedure hosted on the WebQA evaluation server, we did these tests on the validation subset. For the following analysis, we focus on image-based questions.

Overall, the VLP baseline achieved full retrieval accuracy on 41.38% of questions. In other words, the predictions of these questions have the same number of sources and every source matches the canonical annotation. The model did not select enough sources on 18.52% of the questions, while selected too many sources on 34.63% of the questions. For the RoBERTa baseline, full retrieval accuracy is 43.89%, 9.32% of the questions have too few predicted sources, and 44.48% has too many.

As shown in Table 3, lots of questions are predicted to have either 0 sources or too many sources. For image-based questions, WebQA has at most two sources (Chang et al., 2021). Table 3 shows the frequency of the number of sources per question predicted by the model.

#Sources	Freq (VLP)	Freq (RoBERTa)
0	258	150
1	915	868
2	675	652
3	302	325
4	127	208
>= 5	137	308

Table 3. Number of selected sources in each image-based question in the validation set

Both VLP and RoBERTa retrieved an incorrect number of sources on many occasions. But the RoBERTa baseline tends to retrieve more. This also corresponds to RoBERTa's high recall on the validation dataset.

Table 5 shows the percentage of fully retrieved questions with regard to several question attributes. The RoBERTa has more fully correctly retrieved questions across all question categories except *Shape*. Single-source questions are the ones that require only one source fact to answer, while multi-source questions require two or more sources to be answered. We believe we can measure multi-hop capability by the percentage of correctly retrieved questions among all questions that require two or more sources.

Category	VLP(%)	RoBERTa(%)
Choose	40.50	**42.23**
Shape	**50.00**	45.95
YesNo	33.67	**35.99**
Others	43.12	**46.19**
Color	44.12	**49.16**
Number	59.27	**62.16**

Table 4. Percentage of fully correct question predictions in different question categories

Sources	VLP(%)	RoBERTa(%)
single	51.49	**54.91**
multi	28.73	**30.11**

Table 5. Percentage of fully correct question predictions in multi-source questions and single-source questions

6.2. Discussion

We noticed a pattern that a model that performs better on image-based questions performs worse on text-based questions. We believe that there should be a way to modulate and encourage the model to learn from both modalities. Hence, we proposed the ITM loss research idea in Section 7.1.

The results show that the RoBERTa baseline has a better multi-hop capability. This is probably because the attention between multiple modalities is more difficult to train compared to unimodal attention. Because of this, we proposed our multimodal cross-attention research idea mentioned in Section 7.3.

Note that both the VLP and the RoBERTa baseline use the exact same BERT architecture, but the significant difference in the performance suggests that maybe the BERT model itself cannot capture multimodal information well. Therefore, we want to test the Image-Text Matching loss and multimodal cross-attention module research ideas in our future work.

Meanwhile, we are not fully certain if the image regional fea-

ture produced by RCNN is suitable for the VQA task, plus the network itself takes a long time to run. We believe that there is a way to reduce the overall architectural complexity while keeping the performance. Therefore, we proposed Detection-free visual encoder and multimodal knowledge distillation research ideas in Section 7.2 and Section 7.4.

7. New Research Ideas

7.1. Image-Text Match Loss

As mentioned in the Introduction section, we will be focusing on the retrieval part of WebQA for now. For multimodal sources, $s_1, \cdots, s_i$, the loss of retrieval in WebQA adopts Cross-entropy loss. Since s_i symbols the confidence that a certain source will be selected or not, thus the cross-entropy loss is also within the scope of Binary Cross-entropy Loss (BCE).

$$Loss_{retrieval} = \sum_{s_i \in \mathcal{G}} log p_{s_i} + \sum_{s_i \in \mathcal{D}} log(1 - p_{s_i}) \quad (5)$$

In the official leaderboard of WebQA[1], the current SOTA on the retrieval task, the ITL team, utilizes unimodality only. They first caption text from the image facts and then encode the caption information. Surprisingly, they achieved better performance compared with other multimodal approaches. This gives out a message that text (questions, context, image captions, etc.) are actually playing a bigger part in source retrieval compared to images. Our first baseline model also shed light on this analysis.

With only a cross-entropy objective, we are not able to further encourage a joint understanding of image and text. Plus, the retrieval problem is more than a binary classification problem. Therefore, our idea is to introduce the Image-Text Matching Objective to the retrieval task. In fact, the retrieval problem is more like the definition of the Image-text Matching task using attention-based methods (Lee et al., 2018). Image-text matching predicts whether a pair of images and text is matched or not (Li et al., 2021). In the cases of WebQA retrieval, image-text matching can be thought of as whether a Question matches the image facts or not. Image text pairs are formed using image facts and multiple text sources, including questions, captions, and context. With ITM loss introduced, we expect to achieve a higher F1 score compared to the baseline model. We will perform experiments to validate our hypothesis before the next milestone.

ITM Loss has been applied together with Masked Language Model (MLM) Loss in the multimodal stage in ALBEF(Li et al., 2021). The ITM loss can be addressed as

$$\mathcal{L}_{itm} = \mathbb{E}_{(I,T)\sim D} H(y^{itm}, p^{itm}(I,T)) \quad (6)$$

where the joint representation of the image-text pair can be directly taken from the `[CLS]` token, followed by a fully connected layer (FC) layer and softmax function to predict a binary probability p^{itm} (Li et al., 2021). The y^{itm} is a one-hot vector of 2 dimensions, hosting the ground truth label.

With ITM loss integrated, the retrieval objective can now be written as

$$Loss_{new} = Loss_{retrieval} + L_{itm} \quad (7)$$

Note, not all sources in WebQA have image facts. `text` category without image facts will be using the objective function with BCE Loss only. This may result in inconsistent loss scale during retrieval tuning, therefore, additional scaling on the $Loss_{new}$ will be needed.

7.2. Detection-free Visual Encoder

In WebQA, both retrieval and QA tasks utilize the VLP part of the model backbone (Chang et al., 2021). Although VLP uses a unified multimodal encoder to jointly encode textual and visual representation, they are still generated from different feature spaces. The image representations are extracted using the top 100 regional features from faster RCNN (Ren et al., 2015). It is true that faster RCNN has avoided repetitive calculation, however, that comes with the cost of higher memory consumption. Moreover, despite the faster RCNN innovatively proposed Region proposal network (RPN) to improve the time consumption of proposal stage (Ren et al., 2015), the proposal generation and the detection are still in different stages, thus the speed cannot meet the real-time requirements.

Moreover, in real practice, when we try to extract RCNN features for a new image, it is very difficult to set up a proper environment that has both functional hardware and configurations. We felt that impedance in the environment setup process will be a non-trivial factor that affects the practicality of our model. Plus, object detection is both annotation-expensive and compute-expensive (Li et al., 2021). Furthermore, the expressive ability is limited due to the predefined visual dictionary (Kim et al., 2021). Thus, we are looking for an alternative approach that is able to address these concerns.

Thus, our second research idea is to replace the expensive and inefficient VLP model with detector-free visual encoders. ALBEF employs a 12-layer visual transformer ViT-B/16 (Dosovitskiy et al., 2020) as their image encoder, while ViLT encodes images without convolution and region supervisions (Kim et al., 2021).

To validate our hypothesis that a detector-free visual encoder is an efficient and cheaper alternative to faster RCNN, we will analyze retrieval performance together with parameter

[1] https://eval.ai/web/challenges/challenge-page/1255/leaderboard/3168

number and computation flops in future experiments.

7.3. Cross vs. Unified Attention Modules

One of the baseline models of WebQA is VLP, a unified encoder-decoder framework jointly encoding the visual and text representations (Zhou et al., 2019). Meanwhile, "two-tower" models are also popular choices. These models encode multimodal information using separate encoders and fuse them using a multimodal cross-attention module. The authors of VLP (Zhou et al., 2019) compared it with several "two-tower" models, such as ViLBERT and LXMERT. However, there is no significant improvement between the unified model and the "two-tower" models. The "two-twoer" models achieved better performance in some cases (Zhou et al., 2019). Therefore, we are interested in understanding the corresponding effect of unified attention modules and cross-modal attention modules for multi-hop VQA.

The cross-attention mechanism is proposed in the early transformer structure, which combines two separate embedding sequences of the same dimension, in contrast to self-attention where the input is a single embedding sequence (Vaswani et al., 2017). The cross-attention mechanism based on the transformer structure can be extended to multiple modalities. The multimodal transformer is an end-to-end extension of the vanilla transformer framework that learns representations from unaligned modalities (Tsai et al., 2019). We will largely follow the mathematical formulation proposed in this paper.

For the VQA case, we only need to align two modalities denoted as L (language) and I (image), which can be seen as a simplified case of the tri-modality transformer. Suppose our learned representations of two modalities are $X_\alpha \in \mathbb{R}^{T_\alpha \times d_\alpha}$ and $X_\beta \in \mathbb{R}^{T_\beta \times d_\beta}$ respectively. We pivot the α as the Querys: $Q_\alpha = X_\alpha W_{Q_\alpha}$, and image features as Keys and Values: $K_\beta = X_\beta W_{K_\beta}$, $V_\beta = X_\beta W_{V_\beta}$, where $W_{Q_\alpha} \in \mathbb{R}^{d_\alpha \times d_k}, W_{K_\beta} \in \mathbb{R}^{d_\beta \times d_k}$ and $W_{V_\beta} \in \mathbb{R}^{d_\beta \times d_v}$ are weights. Following the same latent adaptation in (Tsai et al., 2019), we formulate the cross-modal attention as:

$$\begin{aligned} Y_\alpha &= \mathbf{CM}_{\beta\to\alpha}\left(X_\alpha, X_\beta\right) \\ &= \text{softmax}\left(\frac{Q_\alpha K_\beta^\top}{\sqrt{d_k}}\right) V_\beta \\ &= \text{softmax}\left(\frac{X_\alpha W_{Q_\alpha} W_{K_\beta}^\top X_\beta^\top}{\sqrt{d_k}}\right) X_\beta W_{V_\beta} \end{aligned}$$

The i-th time step of Y_α is the weighted summary of values of β, which can be considered as a single-head cross-attention module between modalities.

Based on the cross-attention module, we construct our multimodal transformer as follows. We denote the feature representations $X_{\{L,I\}} \in \mathbb{R}^{T_{\{L,I\}} \times d_{\{L,I\}}}$. Then we pass the input through a 1D temporal convolutional layer:

$$\hat{X}_{\{L,I\}} = \text{Conv 1D}\left(X_{\{L,I\}}, k_{\{L,I\}}\right) \in \mathbb{R}^{T_{\{L,I\}} \times d} \quad (8)$$

We then augment the output of temporal convolutions with the positional embedding:

$$Z_{\{L,I\}}^{[0]} = \hat{X}_{\{L,I\}} + \text{PE}\left(T_{\{L,I\}}, d\right) \quad (9)$$

where $\text{PE}\left(T_{\{L,I\}}, d\right) \in \mathbb{R}^{T_{\{L,I\}} \times d}$ is the computes the fixed embeddings for each position index, and $Z_{\{L,I\}}^{[0]}$ are the low-level position-aware features for different modalities.

Let's take L to I cross-attention as the example, the forward-computation of the cross-attention transformer for $i = 1, 2, ..., D$ layers is:

$$\begin{aligned} Z_{L\to I}^{[0]} &= Z_I^{[0]} \\ \hat{Z}_{L\to I}^{[i]} &= \text{CM}_{L\to I}^{[i],\text{mul}}\left(\text{LN}\left(Z_{L\to I}^{[i-1]}\right), \text{LN}\left(Z_I^{[0]}\right)\right) \\ &\quad + \text{LN}\left(Z_{L\to I}^{[i-1]}\right) \\ Z_{L\to I}^{[i]} &= f_{\theta_{L\to I}^{[i]}}\left(\text{LN}\left(\hat{Z}_{L\to I}^{[i]}\right)\right) + \text{LN}\left(\hat{Z}_{L\to I}^{[i]}\right) \end{aligned}$$

where f_θ is a position-wise feed-forward sublayer.

Since only two modalities are considered, there is no need to stack the outputs from the cross-attention transformer that share the same modality. The $Z_{L\to I}$ and $Z_{I\to L}$ are passed through the self-attention transformer, and then the last elements are passed through a fully-connected layer to make predictions.

In this research question, We will compare the unified attention model with the cross-attention model and discuss the places where the unified version fails.

7.4. Multimodal Knowledge Distillation

Knowledge distillation is widely used in text question answering. Some models follow the retriever-reader systems, with the retriever retrieving documents from a large source of knowledge and the reader processing the support documents to solve the task (Izacard & Grave, 2020). Distillation enables us to use a lighter model for downstream tasks, which effectively compresses the pre-trained model. Now there are a few lines of research focusing on applying knowledge distillation in multimodal settings. Fang et al. (2021) proposed a compressing scheme for the visual-linguistic models (DistillVLM). Instead of using pre-trained object detectors such as fast RCNN, they aim to use a lightweight detector for faster inference. To solve the problem of the unalignment of attention distributions between the Teacher and Student's visual tokens, they utilized the same object proposals obtained from Student's lightweight detector. Then

they used a loss term to enforce the Student to mimic the Teacher's self-attention distribution at the last transformer layer. And at last, they distilled the knowledge from the outputs of the transformer layers (Fang et al., 2021). Following the similar procedure proposed in DistillVLM, we formulate our new research idea as follows.

First, we will apply a lightweight object detector for both the Teacher and Student modules. We will choose the same object detector as DistillVLM, which is TEE used by MiniVLM (Wang et al., 2020a) whose backbone is replaced with EfficientNet (Tan & Le, 2019) and a BiFPN module (Tan et al., 2019). Both the Teacher and the Student use the same object tags obtained from the lightweight object detectors during distillation.

Second, we will impose the distillation loss for attention distributions for the Teacher and Student. For transformer-based distillation using the last transformer block's attention map yields equivalent results (Wang et al., 2020b). So we can simply formulate the distillation loss by minimizing the divergence between the self-attention matrices of the last layer of the Teacher and the Student:

$$\mathcal{L}_{\text{ATT}} = \frac{1}{T \cdot H} \sum_{i=1}^{T} \sum_{j=1}^{H} \text{MSE}\left(\mathbf{A}_{i,j}^{S}, \mathbf{A}_{i,j}^{T}\right) \tag{10}$$

where $\mathbf{A} = \text{softmax}\left(\mathbf{Q} \cdot \mathbf{K}/\sqrt{d_k}\right)$.

Third, in order to align the hidden representations for the Teacher and the Student, we want to minimize the divergence of the hidden embedding ($\mathbf{H} \in \mathbb{R}^{T \times d}$) of every Transformer block. Ths distillation for hidden representations is calculated as:

$$\mathcal{L}_{\text{HID}-\text{MSE}} = \frac{1}{T \cdot L} \sum_{i=1}^{T} \sum_{j=1}^{L} \text{MSE}\left(\mathbf{H}_{i,j}^{S}\mathbf{W}_h, \mathbf{H}_{i,j}^{T}\right) \tag{11}$$

where L denotes the number of transformer blocks and $\mathbf{W}_h$ is a learnable linear transformation that maps the Student hidden embedding into the identical dimension of Teacher embedding. Apart from the MSE loss for the layer-to-layer method, we also use noise contrastive estimation (NCE) loss to align the Teacher & Student's hidden representations. Following the paper, we employ a pre-defined instance queue $\left[\mathbf{h}_0^T, \mathbf{h}_1^T \cdots \mathbf{h}_K^T\right]$ to store K random sampled embeddings and one positive embedding from the Teacher network. The loss is formulated as:

$$\mathcal{L}_{\text{HID}} = -\log \frac{\exp\left(\mathbf{h}_i^S \cdot \mathbf{h}_i^T/\tau\right)}{\sum_{j=0}^{K} \exp\left(\mathbf{h}_i^S \cdot \mathbf{h}_j/\tau\right)} \tag{12}$$

where τ denotes the temperature hyper-parameter.

Moreover, during the fine-tuning stage, we will calculate the classification loss of both the Teacher and the Student via cross-entropy loss. The loss is written as:

$$\mathcal{L}_{\text{CLS}} = \text{CE}\left(\mathbf{z}^S/\tau_d, \mathbf{z}^T/\tau_d\right) \tag{13}$$

where $\mathbf{z}^T, \mathbf{z}^S$ are the soft label outputs from the Teacher and the Student network.

During training, we will minimize the combined loss:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{CLS}} + \alpha\mathcal{L}_{\text{ATT}} + \beta\mathcal{L}_{\text{HID}} \tag{14}$$

where α and β are the weights of the loss terms, and $\mathcal{L}_{\text{CE}}$ is the original classification task in the specific downstream-task.

References

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.

Chen, F., Zhang, D., Han, M., Chen, X., Shi, J., Xu, S., and Xu, B. Vlp: A survey on vision-language pre-training, 2022. URL https://arxiv.org/abs/2202.09061.

Chen, T., Deng, J., and Luo, J. Adaptive offline quintuplet loss for image-text matching, 2020. URL https://arxiv.org/abs/2003.03669.

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.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. URL https://arxiv.org/abs/2010.11929.

Fang, Z., Wang, J., Hu, X., Wang, L., Yang, Y., and Liu, Z. Compressing visual-linguistic model via knowledge distillation, 2021. URL https://arxiv.org/abs/2104.02096.

Gao, D., Jin, L., Chen, B., Qiu, M., Li, P., Wei, Y., Hu, Y., and Wang, H. Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval, 2020. URL https://arxiv.org/abs/2005.09801.

Izacard, G. and Grave, E. Distilling knowledge from reader to retriever for question answering, 2020. URL https://arxiv.org/abs/2012.04584.

Karpathy, A. and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions, 2014. URL https://arxiv.org/abs/1412.2306.

Kim, W., Son, B., and Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pp. 5583–5594. PMLR, 2021.

Kim, Y. and Rush, A. M. Sequence-level knowledge distillation, 2016. URL https://arxiv.org/abs/1606.07947.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D. A., Bernstein, M. S., and Fei-Fei, L. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017. doi: 10.1007/s11263-016-0981-7. URL https://doi.org/10.1007/s11263-016-0981-7.

Lee, K.-H., Chen, X., Hua, G., Hu, H., and He, X. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., and Hoi, S. C. H. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705, 2021.

Li, K., Zhang, Y., Li, K., Li, Y., and Fu, Y. Visual semantic reasoning for image-text matching, 2019. URL https://arxiv.org/abs/1909.02701.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Loshchilov, I. and Hutter, F. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017. URL http://arxiv.org/abs/1711.05101.

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, 2019. URL https://arxiv.org/abs/1906.00067.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Ren, S., He, K., Girshick, R. B., and Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015. URL http://arxiv.org/abs/1506.01497.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2019. URL https://arxiv.org/abs/1910.01108.

Tan, M. and Le, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. 2019. doi: 10.48550/ARXIV.1905.11946. URL https://arxiv.org/abs/1905.11946.

Tan, M., Pang, R., and Le, Q. V. Efficientdet: Scalable and efficient object detection. 2019. doi: 10.48550/ARXIV.1911.09070. URL https://arxiv.org/abs/1911.09070.

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., and Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences, 2019. URL https://arxiv.org/abs/1906.00295.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, J., Hu, X., Zhang, P., Li, X., Wang, L., Zhang, L., Gao, J., and Liu, Z. Minivlm: A smaller and faster vision-language model, 2020a. URL https://arxiv.org/abs/2012.06946.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020b. URL https://arxiv.org/abs/2002.10957.

Wang, Z., Codella, N., Chen, Y.-C., Zhou, L., Dai, X., Xiao, B., Yang, J., You, H., Chang, K.-W., Chang, S.-f., and Yuan, L. Multimodal adaptive distillation for leveraging unimodal encoders for vision-language tasks, 2022. URL https://arxiv.org/abs/2204.10496.

Wei, X., Zhang, T., Li, Y., Zhang, Y., and Wu, F. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Yang, Z., Gan, Z., Wang, J., Hu, X., Lu, Y., Liu, Z., and Wang, L. An empirical study of gpt-3 for few-shot knowledge-based vqa, 2021. URL https://arxiv.org/abs/2109.05014.

Yuan, W., Neubig, G., and Liu, P. Bartscore: Evaluating generated text as text generation. *CoRR*, abs/2106.11520, 2021. URL https://arxiv.org/abs/2106.11520.

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

Zhu, Y., Groth, O., Bernstein, M., and Fei-Fei, L. Visual7w: Grounded question answering in images, 2015. URL https://arxiv.org/abs/1511.03416.