Enhancing Textual Cues in Multi-modal Transformers for VQA

Yu Liu, Lianghua Huang, Liuyihang Song, Bin Wang, Yingya Zhang, Pan Pan
Machine Intelligence Technology Lab, Alibaba Group
ly103369,xuangen.hlh,liuyihan.slyh,ganfu.wb,yingya.zyy,panpan.pp@alibaba-inc.com

Abstract

Visual Question Answering (VQA) requires a comprehensive understanding of both visual (image) and textual (question) contents. Existing methods usually emphasize more on the improvements of visual branches by fusing grid and ROI features, while ignoring the improvements of textual cues that are usually more directly correlated to the answers. In this paper, we propose to enrich the textual cues in the VQA task. Specifically, with a multi-modal transformer architecture, we incorporate representations of OCR tokens, detected object categories, as well as the question and answer tokens of the nearest reference image. Experiments on VizWiz VQA dataset demonstrate the effectiveness of our approach, and we achieve an overall accuracy of 57.21% on the test-dev set.

1. Introduction

Visual Question Answering (VQA) aims at learning a multi-modal model to answer the question according to a given image. The task requires a comprehensive interpretation of both visual and textual information. To achieve this, existing VQA methods usually follow a uniform paradigm, *i.e.*, using a multi-modal transformer to fuse language cues that are extracted by BERT-style models, and visual cues that are extracted by recognition or detection models.

Several existing works focus on the improvements of visual encoding as well as multi-modal fusion mechanisms. However, the textual information have not been fully explored. Considering that a number of answers have high correlation with the linguistic cues of images, the performance of these existing methods could be sub-optimal.

To address this issue, in this paper, we propose to enrich the textual cues by exploring three external tags. First, we include the category names of the detected objects in the image. Second, we introduce OCR tokens extracted from images with a pretrained OCR model. Third, in order to fully explore the knowledge in the training data, we search for the closest training image for each query image, and include its question-answer pairs as external tags. For each of the three tags, we train a separate multi-modal transformer. During test, we ensemble their predictions to vote for the final answer. We achieve an overall accuracy of 57.21% on the test-dev subset on VizWiz VQA 2021 challenge.

2. Method

2.1. Overview

We take Oscar [2] as our baseline, and we extend it to support three external tags: object categories, OCR tokens, as well as the question-answer pairs of the closest reference image. Figure 1 shows our multi-modal transformer architecture with three inputs: (1) the image features that are extracted by Faster RCNN [3] pretrained with VinVL [4]; (2) the question tokens; and (3) the external tags. Three options are considered as the external tag: the categories of detected objects, the OCR tokens, as well as the question-answer pairs of the closest reference image. We concatenate question tokens and the external tags, and perform word embedding to get the language encoding. Finally, through several attention layers, the multi-modal transformer outputs a softmax score prediction over a pool of candidate answers. We ensemble the predictions of the three transformers with different external tags, and choose the answer with the highest score as the final prediction.

2.2. External Tags

2.2.1 Object Categories

Most existing methods only consider the RoI features extracted with a detection model as the object representation. However, we find some of the answers in VQA have high correlation with the object categories. Therefore, similar to Oscar [2], we also include the predicted object categories as the external tag. With this modification, we are able to achieve an overall accuracy of 54.25% on the test-dev subset on VizWiz VQA 2021 challenge (baseline is 53.62%).

2.2.2 OCR Tokens

Many existing methods (*e.g.*, BUTD-based [1]) overlook the textual information in images. However, we find that

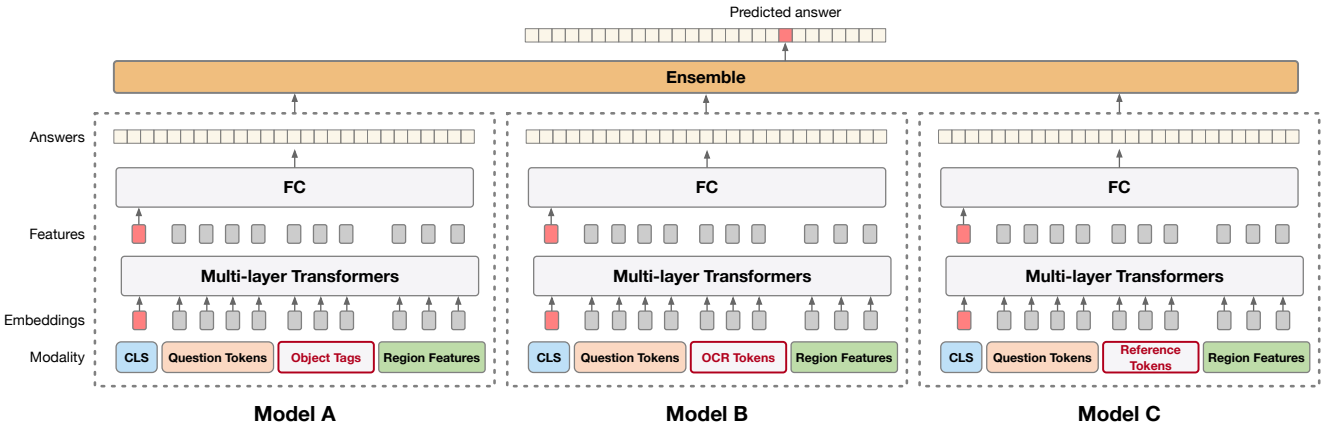

Figure 1: The overview of our approach. For each kind of external tags, we train three models (Model A, Model B and Model C) respectively. During test, we ensemble their predictions to vote for the final answer.

Table 1: Ablation study on the effectiveness of different components in our method. We use the smaller VinVL_B [4] in the ablation study for fast experiments. However, when submitted to the VizWiz challenge, we use the larger VinVL_L for higher accuracy.

Object Tags	OCR Tokens	Reference Tokens	Accuracy (%)
			53.62
✓			54.25
	✓		55.20
		✓	54.49
✓	✓	✓	57.21

many questions (*e.g.*, "what does it say?") require the model to explicitly read texts within the image. Therefore, we use a pretrained OCR model to extract OCR tokens from images, and feed them to the multi-modal transformer. With this modification, we can improve the performance from 53.62% (baseline) to 55.20%.

2.2.3 Reference Image Tokens

A natural way for a human to answer a new question is to refer to existing answers of similar questions asked under similar contexts. We introduce this mechanism to our model, where we search for the closest training data for each query image by using cosine feature similarities with a pretrained ResNet50 model. If the cosine similarity is greater than 0.65, we take the question-answer tokens of this reference image as the side information and feed them to the transformer. Otherwise, we ignore them in all transformer attention layers. With this modification, we can improve the performance from 53.62% (baseline) to 54.49%.

3. Conclusion

In this paper, we focus on how to enrich textual cues in the VQA task. To achieve this, we explore three external tags. First, we include the category names of the detected objects in the image. Second, we introduce OCR tokens extracted from images with a pretrained OCR model. Third, we search for the closest training image for each query image, and include its question-answer pairs as external tags. For every kind of external tags, we train three models respectively, and ensemble their predictions to vote for the final answer. We achieve an overall accuracy of 57.21% on the test-dev subset on VizWiz VQA 2021 challenge.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018. 1

[2] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, pages 121–137. Springer, 2020. 1

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015. 1

[4] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. *arXiv*, 2021. 1, 2