Tell Me the Evidence? Dual Visual-Linguistic Interaction for Answer Grounding

Junwen Pan[1], Guanlin Chen[2], Yi Liu[2], Jiexiang Wang[1], Cheng Bian[1], Pengfei Zhu[2], Zhicheng Zhang[1]
[1]ByteDance
[2]Tianjin University
panjunwen@bytedance.com

Abstract

*Answer grounding aims to reveal the visual evidence for visual question answering (VQA), which entails highlighting relevant positions in the image when answering questions about images. Previous attempts typically tackle this problem using pretrained object detectors, but without the flexibility for objects not in the predefined vocabulary. Recent visual-linguistic models have made remarkable advances by leveraging powerful Transformers to enable visual-linguistic interaction. However, these black-box methods solely concentrate on the linguistic generation, ignoring the visual interpretability. In this paper, we propose **Dual Visual-Linguistic Interaction (DaVI)**, a novel unified framework with the capability for both linguistic answering and visual grounding. DaVI innovatively introduces two visual-linguistic interaction mechanisms: 1) visual-based linguistic encoder that understands questions incorporated with visual features and produces linguistic-oriented evidence for further answer decoding, and 2) linguistic-based visual decoder that focuses visual features on the evidence-related regions for answer grounding. This way, our approach ranked the 1st place in the public answer grounding track of 2022 VizWiz Grand Challenge.*

1. Introduction

Visual question answering (VQA) is a fundamental visual-linguistic task in various real-life applications such as assisting visually impaired people to answer questions about images [1]. Although the VQA community has made significant progress, the best-performing systems are complicated black-box models, raising concerns about whether their answer reasoning is based on correct visual evidence. By understanding the reasoning mechanism of the model, we can evaluate the quality of answers, improve model performance, and provide explanations for end-users.

To address this problem, answer grounding has been introduced into VQA systems, which requires the model to locate relevant image regions as well as answer visual questions. Most previous efforts usually rely on pretrained detection models to provide object features, entailing low answering flexibility for objects not in the predefined vocabulary [2]. Recently, a few works try to get rid of these limitations in a weakly supervised fashion [2], but neither the answering accuracy nor grounding accuracy achieved is satisfactory [1]. In contrast, visual language pre-training has tremendously advanced VQA performance using large-scale multi-modal Transformers and web-scale datasets [3], eliminating the need for detectors. However, these text-generation oriented schemes concentrate on the linguistic modeling based on off-the-shelf visual features, and are therefore not capable of visual-oriented prediction in the grounding task.

To overcome the challenges of answer grounding, a fundamental but critical perspective is how to reconcile the visual-linguistic interaction for both visual-oriented and linguistic-oriented predictions. As shown in Fig. 1, state-of-the-art works inject features from the other modality into the task-oriented modeling, effectively exploiting rich

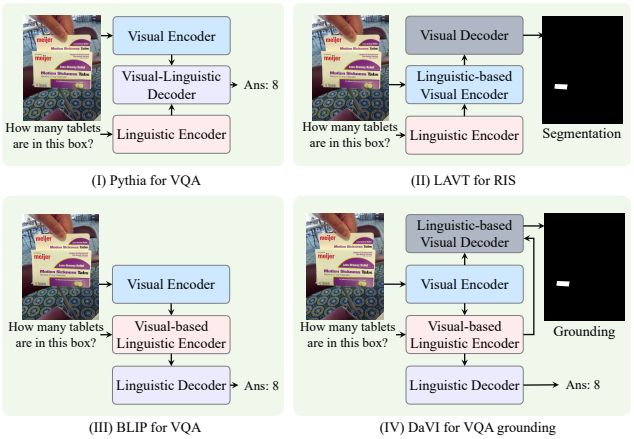

Figure 1. Overview of visual-linguistic interactions: I) Visual-linguistic decoder [4] for VQA, II) Linguistic-based visual encoder [5] for refer image segmentation (RIS), III) Visual-based linguistic encoder (VLE) [3] for VQA, and IV) our DaVI with VLE and LVE for answer grounding. It predicts an answer along with a mask while taking as inputs an image and a question.

Transformer layers in the task-oriented encoder. However, this visual-linguistic interaction is specific to a single-modal output such as linguistic-oriented answering or visual-oriented segmentation.

Accordingly, this paper proposes DaVI which unifies linguistic-oriented answering and visual-oriented grounding into an end-to-end framework. The two key components of DaVI are Visual-based Linguistic Encoder (VLE) and Linguistic-based Visual Decoder (LVD) which will be detailed in the next section.

2. Method

As shown in Fig. 1 (IV), DaVI ingests a question along with the corresponding image as input and outputs an evidence mask while predicting the answer. For simplicity, we employ a Vision Transformer (ViT), instead of an object detector, to extract image features. Then, VLE encodes questions with the image references and produces linguistic-oriented evidence features. The Linguistic Decoder (LD), with the same structure as the BERT, answers visual questions on the basis of linguistic-oriented evidence. Different from LD, to predict evidence masks, LVD performs another interaction between the visual features from VE and the linguistic-oriented evidence.

Visual-based Linguistic Encoder (VLE). We interleave cross-attention (CA) layers in between the self-attention layers and the feed-forward networks of BERT to form the VLE module. The CA layers attend to visual references using linguistic features as queries. These visual-linguistic interactions in VLE provide multi-modal evidence for answer decoding in LD. However, due to the language-oriented modeling, these features need to be remapped into visual grids by LVD to decode visual evidence.

Linguistic-based Visual Decoder (LVD). LVD can be regarded as an "identical" structure of the CA layer in VLE but with different inputs: it uses visual features as queries to attend to linguistic cues from VLE. Following the best practices in computer vision, we collect multi-level visual features generated by VE and feed them into LVD. Finally, a simple convolutional segmentor is exploited to predict the visual mask for answer grounding.

3. Experiment

Dataset. The dataset of VizWiz answer grounding contains 6494, 1131 and 2373 samples in the *train*, *val* and *test* set, respectively. Evaluations are conducted by the challenge server on the *test* set.

Implementation Details. In our experiments, VE, VLE, and LD are pretrained on the web-scale dataset [3]. In the training phase, we keep an exponential moving average copy of parameters for inference. We also ensemble the results from models trained with different hyperparameters.

Table 1. Ablation study on *test* set of VizWiz Grounding track.

Variant	IoU (%)	Δ (%)
LE + LVE + VD (w/o answer)	64.28	–
VE+VLE+LVD+LD	68.52	+4.24
VE+VLE+LVD+LD (w/ ensemble)	**70.57**	+2.05

Table 2. Results on *test* set of VizWiz Grounding track.

Rank	Team	IoU (%)
1	**Aurora (ours)**	**70.57**
2	hsslab	70.15
3	MGTV	69.70
4	UD VIMS Lab	67.55
5	MindX	66.92
6	boostboom	64.49
7	Black	58.44

Results. As shown in Tab 1, compared to the single interaction paradigm, DaVI successfully achieved a 4.24% IoU improvement, thanks to the design of dual visual-linguistic interactions. Tab. 2 shows public results on VizWiz answer grounding track, where DaVI ranked the 1^{st} place, substantially outperforming other methods.

4. Conclusion

We propose DaVI for answer grounding, which utilizes the VLE and LVD for visual-linguistic interaction to unify the linguistic-oriented answering and visual-oriented grounding. DaVI ranked the 1^{st} place on the public track of VizWiz answer grounding Challenge.

References

[1] Chongyan Chen, Samreen Anjum, and Danna Gurari. Grounding answers for visual questions asked by visually impaired people. *arXiv preprint arXiv:2202.01993*, 2022. 1

[2] Aisha Urooj Khan, Hilde Kuehne, Kevin Duarte, Chuang Gan, Niels da Vitoria Lobo, and Mubarak Shah. Found a reason for me? weakly-supervised grounded visual question answering using capsules. In *CVPR*, pages 8465–8474, 2021. 1

[3] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 1, 2

[4] Amanpreet Singh and et al. Mmf: A multimodal framework for vision and language research. https://github.com/facebookresearch/mmf, 2020. 1

[5] Zhao Yang, Jiaqi Wang, Yansong Tang, and et al. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, 2022. 1