Vision-Language Model-based PolyFormer for Recognizing Visual Questions with Multiple Answer Groundings

Dai Quoc Tran
Global Frontiers of Resilient EcoSmart City,
Sungkyunkwan University, South Korea
daitran@skku.edu

Armstrong Aboah
Depart. of Civil, Construction & Env. Engineering
North Dakota State University, United States
armstrong.aboah@ndsu.edu

Yuntae Jeon
Department of Global Smart City
Sungkyunkwan University, South Korea
jyt0131@g.skku.edu

Minsoo Park
Sungkyun AI Research Institute
Sungkyunkwan University, South Korea
pms5343@skku.edu

Seunghee Park *
School of Civil, Arch Eng. and Landscape Architecture
Sungkyunkwan University, South Korea
shparkpc@skku.edu

Abstract

This paper presents a new method that utilizes the capabilities of Vision-and-Language Transformers (ViLT) and the advanced PolyFormer model to tackle the Single Answer Grounding Challenge in the VQA-Therapy dataset. The initial step of our approach involves employing the ViLT model to predict the possible count of unique responses by considering the input question and image. The PolyFormer model subsequently examines the output from ViLT in conjunction with the image to produce visual answer masks that correspond to the input. The presence of overlap between these masks determines whether the answers have a common grounding. If there is no overlap, it indicates the existence of multiple groundings. Our approach achieved an F1 score of 81.71 on the test-dev set and 80.72 on the VizWiz Grand Challenge test set, positioning our team among the top three submissions in the competition. Code is available at https://github.com/daitranskku/VizWiz2024-VQA-AnswerTherapy

*Corresponding author.

Acknowledgements: This research was supported by a grant [2022-MOIS38-002 (RS-2022-ND630021)] from the Ministry of Interior and Safety (MOIS)'s project for proactive technology development safety accidents for vulnerable groups and facilities, and this research was supported by a grant from the Korean Government (MSIT) to the NRF [RS-2023-00250166]. This work is financially supported by Korea Ministry of Land, Infrastructure and Transport(MOLIT) as Innovative Talent Education Program for Smart City.

1. Introduction

Visual question answering (VQA) is a dynamic area of study that combines computer vision and natural language processing. It involves developing models that can accurately answer questions about specific images. Conventional VQA tasks typically assume that there is ***only one correct answer*** for each question, without considering the inherent variation in human interpretations and responses. This assumption can result in the creation of AI systems that lack inclusivity and are unable to accurately portray the complete range of human perspectives. The VQA-Therapy dataset [1] addresses the need to consider diversity by introducing a new task: ***determining if different annotators' answers to a visual question are based on the same elements of an image***. In order to tackle this challenge, our research presents a novel methodology that integrates two state-of-the-art models: the Vision-and-Language Transformer (ViLT) [2] and the PolyFormer [3]. As shown in Figure 1, the proposed approach initially employs ViLT to estimate the number of unique answers that a question-image pair can generate, taking into account both the visual content and the context of the question. Subsequently, the PolyFormer is utilized to predict and examine visual grounding masks using the ViLT outputs and the image itself. The ultimate determination of whether an answer set has a sin-

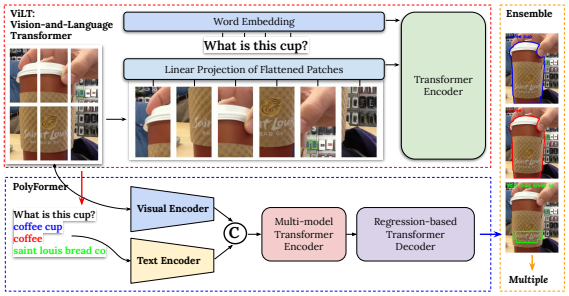

Figure 1. The proposed approach utilized two SOTA models: 1) ViLT: Vision-and Language Transformer [2], 2) PolyFormer [3]

gular grounding or multiple groundings is contingent upon the analysis of the overlap between these masks.

2. Methodology

2.1. Vision-and-Language Transformer (ViLT)

ViLT [2] is a transformer-based model designed to process combined inputs of visual content and textual data without the need for object detection or region proposal networks that are common in other vision-language models. The primary output of this stage is ***a set of potential answers***, predicted based on the relevance and presence of visual and textual cues in the image-question pair.

Training settings. The ViLT model was retrained using a curated selection of 3,794 images from the VQA-Therapy [1] dataset, which is specifically designed to explore and understand the diversity of annotator responses. For validation purposes, a separate set of 646 images was used to evaluate the model's performance and tuning. We utilized the **vilt-b32-finetuned** configuration, a variant of the ViLT model that has been specifically fine-tuned for VQA tasks. This model configuration is optimized to handle the complexities of integrating visual and textual data, making it well-suited for our grounding task. The proposed approach is inferenced on Intel Core i9, and NVIDIA 4090 24GB and 64GB RAM. Models are trained on Intel Xeon Silver 4210R, and 2 NVIDIA RTX A6000 48GB and 126GB RAM.

2.2. PolyFormer

Building upon the predictions made by ViLT, the PolyFormer [3]—a multi-modal transformer model—takes over to analyze these potential answers in the context of the same image. The PolyFormer comprises separate encoders for visual and textual data and a multi-modal transformer encoder that fuses these inputs to better understand the context and significance of each element. Each predicted answer from ViLT is treated as a separate input sequence alongside the image to the Polyformer. In this research, we utilized a pretrained weight from **PolyFormer-L**, and adopted Swin-L as the visual backbone with 12 transformer encoder and decoder layers. For each potential answer, the ***PolyFormer generates a visual grounding mask*** that highlights areas of the image most relevant to that answer. Finally, the presence of overlap between these masks determines whether the answers have a common grounding. If there is no overlap, it indicates the existence of multiple groundings. Table 1 compares the performance metrics of two configurations of our proposed approach evaluated on the VizWiz Grand Challenge test-dev set. The table lists the F1 Score, Precision, and Recall for each model configuration. The first row presents the results for the ViLT model fine-tuned exclusively on VizWiz data, achieving an F1 Score of 77.22, a Precision of 77.62, and a Recall of 77.22. The second row details the performance of the ViLT model finetuned on an aggregated dataset, further enhanced with the PolyFormer, which shows improved outcomes with an F1 Score of 81.71, a Precision of 78.85, and a Recall of 84.81. These results proved the effectiveness of integrating PolyFormer into the training process, particularly in enhancing the recall metric, thereby indicating a more comprehensive coverage in identifying relevant visual groundings across diverse data inputs. For the VizWiz Grand Challenge test set, our approach achieved an F1 score of 80.72.

Model	F1 Score	Precision	Recall
Our approach 1	77.22	77.62	77.22
Our approach 2	81.71	78.85	84.81

Table 1. Proposed approach on VizWiz Grand Challenge test-dev set. Our approach 1 means ViLT fine-tuned on VizWiz data, and our approach 2 means ViLT fine-tuned on VizWiz + VQA data.

3. Conclusion

Our approach to the VQA challenge, which integrates the ViLT and the PolyFormer, has demonstrated capability in predicting the grounding of answers based on visual and textual inputs. By employing these models, we successfully predicted whether the answers to a given visual question share the same grounding, achieving a robust F1 score of 80.72 and securing a top-3 position in the VizWiz Grand Challenge.

References

[1] Chongyan Chen, Samreen Anjum, and Danna Gurari. Vqa therapy: Exploring answer differences by visually grounding answers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15315–15325, 2023. 1, 2

[2] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021. 1, 2

[3] Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18663, 2023. 1, 2