Embedding Attention Blocks for the VizWiz Answer Grounding Challenge

Seyedalireza Khoshsirat Chandra Kambhamettu
VIMS Lab, University of Delaware
{alireza, chandrak}@udel.edu

Abstract

Answer grounding is the task of locating relevant visual evidence for the Visual Question Answering (VQA) task. In this work, we propose a cross-attention block, which we term Embedding Attention, that re-calibrates channel-wise image feature-maps by explicitly modeling inter-dependencies between the image feature-maps and the image-question-answer embedding. We build upon the current best practices of attention methods to design this block. The flexibility of our method makes it easy to use different pre-trained backbone networks. We demonstrate the effectiveness of our method on the VizWiz-VQA-Grounding dataset. Our method holds first place on the 2023 VizWiz-VQA-Grounding challenge leaderboard.

1. Introduction

The answer grounding task is defined as detecting the pixels that can provide evidence for the answer to a given question regarding an image.

Attention Mechanism. In deep learning, attention is a mechanism that mimics cognitive attention. The goal is to enhance the important parts of the input data and fade out the rest. Attention methods can be classified into two classes based on their inputs: self-attention and cross-attention. Self-attention is a type of attention that quantifies the interdependence within the elements of a single input, and cross-attention finds the interdependence across two or more inputs [5]. Usually, cross-attention methods are used for multimodal inputs.

The Squeeze-and-Excitation method [2] is a channel-wise self-attention mechanism widely used in classification networks. It consists of a global average pooling of the input, followed by two linear layers with an interleaved non-linearity, and a sigmoid function. Concretely, the output of this method is:

$$\sigma(FC(RELU(FC(g\_avg\_pool(\mathbf{X}))))) \times \mathbf{X} \quad (1)$$

We adopt this method to build an attention module for the answer grounding task.

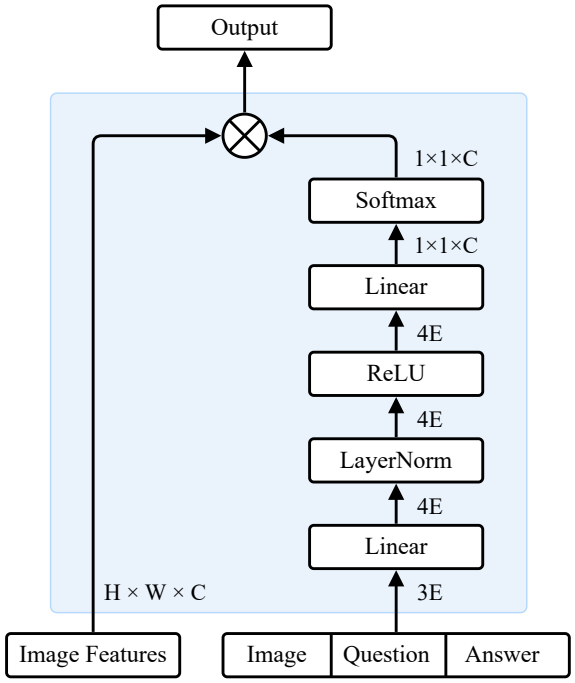

Figure 1. Our proposed embedding attention block. C denotes the number of channels in image feature-maps, and E represents the embedding size. Image, question, and answer embeddings are concatenated into one vector.

2. Method

In order to achieve maximum accuracy, we prioritize the usage of pre-trained models. We use CLIP [3], which is a neural network trained on a variety of image-text pairs. The model outputs embeddings for an image and text such that the similarity of the embeddings correlates with the correspondence of the image and text. We aim to predict answer groundings by processing the image features using the embeddings for an image, question, and answer set (Figure 2). Towards this goal, we design an attention block based on the Squeeze-and-Excitation (SE) block [2] as follows. We start with an SE block and make the following three modifications:

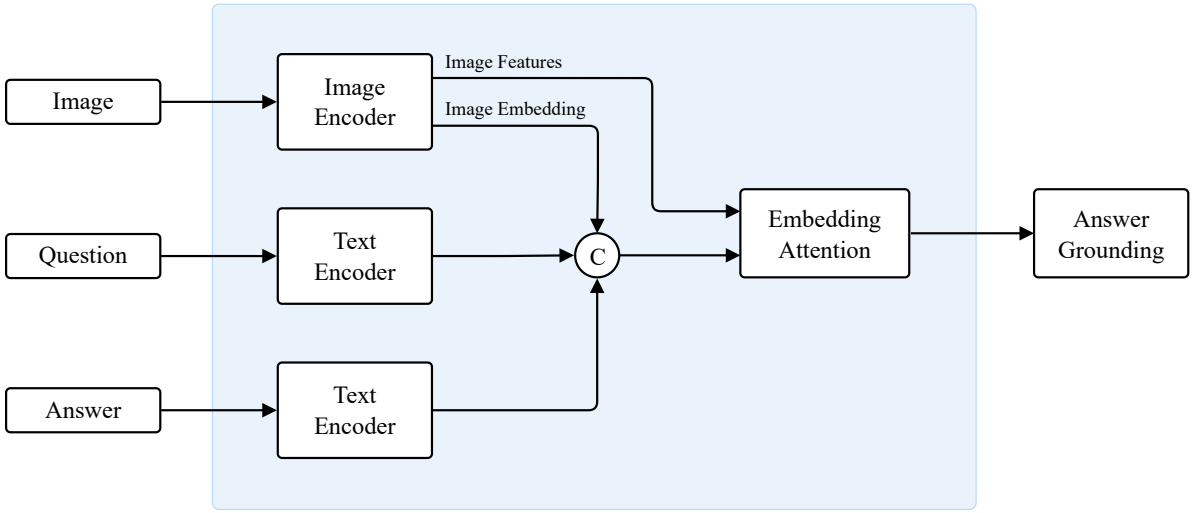

Figure 2. Our proposed network architecture. Our embedding attention block processes the image features using the image, question, and answer embeddings. We use the image and text encoders from CLIP [3]. *C* denotes the concatenation operation.

1. Changing from self-attention to cross-attention.
2. Adding a normalization layer.
3. Replacing the sigmoid function with softmax.

Therefore, our proposed attention block is as follows:

$$Softmax(FC(ReLU(LN(FC(\mathbf{IQA}))))) \times \mathbf{I}_i \quad (2)$$

where FC is a single-layer perceptron, LN is LayerNorm, $\mathbf{IQA}$ is the image-question-answer embedding vector and $\mathbf{I}_i$ is the image features at scale i. Figure 1 depicts this design.

3. Experiment

Setup. We use AdamW optimizer with a weight decay of 0.05 and batch size of 16. We apply the "polynomial" learning rate policy with a poly exponent of 0.9 and an initial learning rate of 0.0001. Synchronized batch normalization is used across multiple GPUs. We use RandAugment for data augmentation.

Dataset. The VizWiz-VQA-Grounding [1] dataset is a subset of the VizWiz-VQA dataset and contains a total of 9,998 VQAs that are divided into 6,494/1,131/2,373 VQAs for training, validation, and testing. We augment the training set of the VizWiz-VQA-Grounding dataset using samples from TextVQA-X [4]. TextVQA-X is a subset of the TextVQA dataset where the images are annotated by humans. It has 18,096 questions and 11,681 unique images.

Results. Table 1 shows the top-performing five teams of the 2023 VizWiz-VQA-Grounding challenge. Our proposed method achieves a mean IoU accuracy of 74.1% and holds the first place in this ranking.

Rank	Team	mean IoU
1	**UD VIMS Lab (Ours)**	**74.1**
2	MGTV_Baseline	72.4
3	USTC-IAT-United	70.6
4	pangzihei	70.3
5	DeepBlue_AI	69.2

Table 1. The top-five teams from the 2023 VizWiz-VQA-Grounding challenge leaderboard [1].

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. 2

[2] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 1

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2

[4] Varun Nagaraj Rao, Xingjian Zhen, Karen Hovsepian, and Mingwei Shen. A first look: Towards explainable textvqa models via visual and textual explanations. *arXiv preprint arXiv:2105.02626*, 2021. 2

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1