Cross-Attention with Self-Attention for VizWiz VQA

Rachana Jayaram
PES University
Bangalore, India
rachana.jayaram@gmail.com

Shreya Maheshwari
PES University
Bangalore, India
shreya.2506@gmail.com

Hemanth C
PES University
Bangalore, India
hemanthvenki910@gmail.com

Sathvik N Jois
PES University
Bangalore, India
sathvik.njois@gmail.com

Dr. Mamatha H.R.
PES University
Bangalore, India
mamathahr@pes.edu

Abstract

A self-evident application of the VQA task is to design systems that aid blind people with sight reliant queries. The VizWiz VQA dataset originates from images and questions compiled by members of the visually impaired community and as such, highlights some of the challenges presented by this particular use case. We propose a model that addresses these challenges by employing a hybrid attention mechanism that involves both intermodal attention and intramodal attention, resulting in state of the art performances. Cross attention between the modalities effectively captures visual-semantic relationships between the given image and question, making the model more adept at recognizing unanswerable image-question pairs. Self attention helps the model glean more intramodal information within the question space - inconsequential conversational phrases are given less attention, thereby improving question comprehension.

1. Introduction

Typical VQA datasets only include well captured images. The questions posed in these types of data are well framed and to the point. However such data is not representative of the input that will likely be given by people with sight impairments: As can be seen in the VizWiz dataset [1], images captured by blind photographers are often too blurry or too dark, making them incomprehensible to models trained on generic data. The object of interest is frequently out of frame of the image, causing the question to seem unrelated to the image presented. The questions asked tend to be incomplete or conversational in nature as they are transcriptions of recorded vocal questions.

Our proposed deep network addresses these key points by jointly modelling the intramodal and intermodal relations between key visual objects in the image and key words in the question via a novel cross attention and self attention mechanism. Cross attention has previously been employed to gauge image-question similarity for multi-modal matching tasks as in [2]. The same idea also uniquely applies to visual question answering for the VizWiz dataset, given that a subset of the task is to detect unanswerable image-question inputs where the image seems to bear no relation to the question posed. Furthermore, the use of self attention on the input question greatly increases question understanding: the model pays less attention to conversational phrases like "Could you please" and "thank you", thereby abating the effects of prefix and suffix attacks.

2. Methodology

1. A Faster-RCNN model [3] pre-trained on the Visual Genome dataset is used to obtain the image feature vector $\boldsymbol{v}$ and a unidirectional single layer LSTM network is used to obtain the question feature vector $\boldsymbol{q}$.

2. Multimodal fusion is done via a deep network implementing self attention and cross attention networks.

   (a) Self attention is applied only on the question feature vector to obtain $\boldsymbol{q}_{self\_attended}$.

   (b) Cross attention is applied on a matrix that encodes the similarity between every object in the image and every word in the question, in-order to model their inter-relationships. The said mechanism yields $\boldsymbol{v}_{attended}$ and $\boldsymbol{q}_{attended}$.

   The answer representation $\boldsymbol{r}$ is just the concatenation of the attended feature vectors.

$$\boldsymbol{r} = \boldsymbol{v}_{attended} + \boldsymbol{q}_{attended} + \boldsymbol{q}_{self\_attended}$$

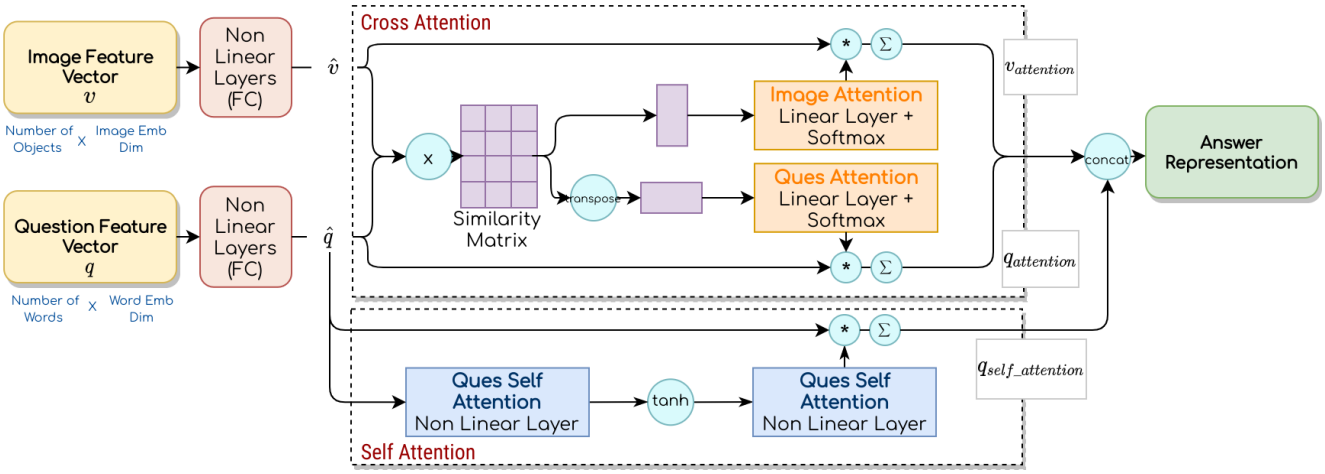

Figure 1. Architecture of the Deep Attention Network.

Here $\boldsymbol{r} \in \mathbb{R}^{3*h_{dim}}$

3. A simple two layer MLP is used to classify the unified answer representation into one of the answers in the candidate answer set.

2.1. Cross Attention

The cross attention mechanism employed in the multi-modal fusion module is implemented as follows:

1. The image and question feature vectors are projected onto the same dimension h_{dim} to obtain $\hat{\boldsymbol{v}}$ and $\hat{\boldsymbol{q}}$.

2. A similarity matrix M which encodes the distance of each object word pair is calculated:

$$M = \hat{\boldsymbol{v}} \times \hat{\boldsymbol{q}}^T$$

Where, $\hat{\boldsymbol{v}} \in \mathbb{R}^{K \times h_{dim}}$, $\hat{\boldsymbol{q}} \in \mathbb{R}^{L \times h_{dim}}$ and $M \in \mathbb{R}^{K \times L}$

3. Attention is then applied on the similarity matrix:

$$a_{vi} = \boldsymbol{w}_v M_i$$

$$a_{qi} = \boldsymbol{w}_q (M^T)_i$$

$$\boldsymbol{\alpha}_v = softmax(\boldsymbol{a}_v)$$

$$\boldsymbol{\alpha}_q = softmax(\boldsymbol{a}_q)$$

Where, $\boldsymbol{\alpha}_v \in \mathbb{R}^K$, $\boldsymbol{\alpha}_q \in \mathbb{R}^L$ and $\boldsymbol{w}_v$, $\boldsymbol{w}_q$ are learned parameters.

4. Cross attended image and question feature vectors are calculated:

$$\boldsymbol{v}_{attended} = \Sigma \alpha_{vi} \hat{\boldsymbol{v}}_i$$

$$\boldsymbol{q}_{attended} = \Sigma \alpha_{qi} \hat{\boldsymbol{q}}_i$$

3. Results

The Cross-Attention + Self-Attention model achieved an overall accuracy of 53.19% on the test-dev split of the Vizwiz dataset. We can deduce from the results obtained that employing both the attention mechanisms performs better than either one in isolation.

Model Name	Overall
Cross Attention	52.53
Self Attention	52.86
Cross Att. + Self Att.	**53.19**

Table 1. Results breakdown of accuracies (in %) into categories of various models for the Vizwiz test-dev split.

References

[1] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people, 2018. 1

[2] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. 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. 1

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. doi: 10.1109/TPAMI.2016.2577031. 1