Global Fusion Attention for Vision and Language Understanding (Student Abstract)

Zixin Guo[1*†], **Chen Liang**[2*], **Ziyu Wan**[2], **Yang Bai**[3]
[1]Department of Computer Science, Aalto University
[2]Department of Computer Science, City University of Hong Kong
[3]Department of Computer Science, China University of Geosciences, Beijing
zixin.guo@aalto.fi, {cliang23-c, ziyuwan2-c}@my.cityu.edu.hk, 2104200014@cugb.edu.cn

Abstract

We extend the popular transformer architecture to a multi-modal model, processing both visual and textual inputs. We propose a new attention mechanism on Transformer-based architecture for the joint vision and language understanding tasks. Our model fuses multi-level comprehension between images and texts in a weighted manner, which could better curve the internal relationships. Experiments on benchmark VQA dataset CLEVR demonstrate the effectiveness of the proposed attention mechanism. We also observe the improvements in sample efficiency of reinforcement learning through the experiments on grounded language understanding tasks of BabyAI platform.

Introduction

Vision and language understanding is a challenging area where visual and textual co-relations are modelled in a structured process. Several tasks are built for multi-modal learning. In this paper, we focus on modelling relations in two tasks: visual questions answering (VQA) and grounded language understanding.

Recent works have shown that more accurate predictions result from hierarchical co-attention of the visual and textual modalities. However, the precision of correlation between image regions and question words is limited by learned coarse interactions of multimodal instances.

Compared to previous co-attention models, a Transformer-based architecture has proved the effectiveness of representing correlations between vision and language on multimodal learning tasks (Yu et al. 2019). However, the decoder has little idea of how well the obtained attention representations related to the query, misleading the decoder into giving erroneous results. Inspired by this, we consider a refining approach to generate more precise features with a full exploration of the multi-layer decoders.

Here we design a Global Fusion Attention (GFA) through assigning learnable weights to different layers of decoder outputs. Experimental results in Table 1 show that transformer-based architecture with GFA achieves comparable performance on CLEVR, which verifies the effectiveness of the proposed attention module. Furthermore, we obtain experimental improvements on grounded language understanding tasks (BabyAI) with GFA, which is demonstrated in supplementary file.

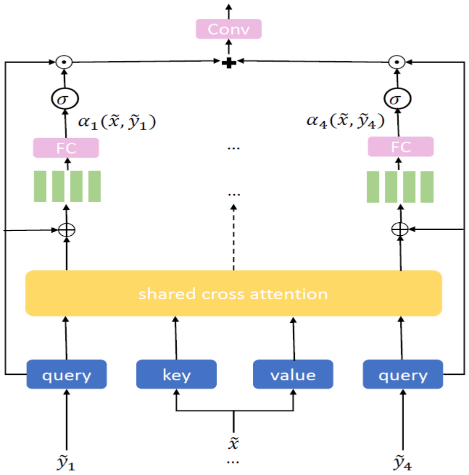

Figure 1: Proposed global fusion attention. $\tilde{X}$ denotes the visual feature from the encoder, while $\tilde{Y}_i$ denotes the text feature from different decoders.

Our Methodology

Problem Definition

We are given a fixed set of answers or action choices $C = \{1, \ldots, S\}$ and visual regions or observations x. We are also given corresponding questions q. The goal of VQA is building a classification model for predicting answers given visual regions and questions. While for grounded language understanding, the model sequentially obtains observations based on the last predicted actions according to instructions until the end of the episode.

Global Fusion Attention Method

Our attention mechanism is illustrated in Figure 1. Given visual representations $\tilde{Y}$ from multiple decoders and textual

[*] The first two authors contributed equally to this work.
[†] Zixin Guo is the corresponding author.

representations $\tilde{X}$ from the encoder, the global fusion attention operator performs weighted summation on decoder outputs and fuses information of each layer in a fine-grained approach. During cross attention, the similarity between vision and language is modelled and weighted feature maps are generated. Here, we assign weights via attending these maps with textual elements in the same cross attention module and obtain how similar these generated representations are related to texts. The procedure is defined as

$$\mathrm{Attn}(\tilde{Y}, \tilde{X}) = \sum_{i} \alpha_i(\tilde{X}, \tilde{Y}_i) \cdot \tilde{Y}_i \tag{1}$$

$$\alpha_i(\tilde{X}, \tilde{Y}_i) = \sigma(W_i[\mathrm{CA}(\tilde{X}, \tilde{Y}_i), \tilde{Y}_i] + b_i) \tag{2}$$

where $\alpha_i(\cdot, \cdot)$ is an operation that generates a weight matrix with the same dimension to $\tilde{Y}_i$, $[\cdot, \cdot]$ stands for concatenation, σ is a sigmoid function, W_i is a weight matrix $\in \mathbb{R}^{2d \times d}$, b_i is a learnable bias vector and CA is the cross attention operator. Details of architectures are described in supplementary file.

Experiments

Datasets

We evaluate the effectiveness of the proposed model on visual question answering (VQA) benchmark CLEVR (Johnson et al. 2017) and grounded language understanding platform BabyAI (Chevalier-Boisvert et al. 2018). CLEVR is a benchmark dataset of 700K training, 150K validation and 15K test tuples, consisted of image, question, answer, program. 3D-rendered objects with multiple colours, shapes, sizes and materials compose the synthetic image. Questions have five types, ranging from counting to comparison, and may have more than 40 words. Answers are selected from a 28-choice answer set. Programs are additional supervisory instructions for answering the question. Accuracy is utilized as the metric, measuring the extent of understanding image contents and questions. BabyAI is introduced in supplementary file.

Experimental Results

We summarize the experimental results in Table 1. Following work as in Mao et al. (2019), we report results on the CLEVR validation split. Our method is comparable with multiple baselines using zero program annotation. Compared with the strong baseline model using 700k programs as extra supervision, our method still could obtain a good performance, which does not involve any extra information.

NS-CL (Mao et al. 2019) utilized image-based features together with extra region-based features from pre-trained Mask RCNN, and we utilized image-based features, which follows work as in Johnson et al. (2017). As a novel attentional method, MAC (Hudson and Manning 2018) achieves higher scores under a few metrics, however, its training requires about twice more learnable parameters than ours.

We also conduct experiments on BabyAI where sample efficiency is evaluated. Besides, we visualize the attention masks of BabyAI and demonstrate training settings of the two tasks. All of them are included in supplementary file.

Model	Prog. Anno.	Over-all	Cnt.	Cmp. Num.	Exist	Query Attr.	Cmp. Attr.	Param.
Human	N/A	92.6	86.7	86.4	96.6	95.0	96.0	
NMN(2017)	700k	72.1	52.5	72.7	79.3	79.0	78.0	
N2NMN(2017)	700k	88.8	68.5	84.9	85.7	90.0	88.8	
IEP(2017)	700k	96.9	92.7	98.7	97.1	98.1	98.9	
DDRprog(2018)	700k	98.3	96.5	98.4	98.8	99.1	99.0	
TbD(2018)	700k	99.1	97.6	99.4	99.2	99.5	99.6	
RN(2017)	0	95.5	90.1	93.6	97.8	97.1	97.9	0.37M
FiLM(2018)	0	97.6	94.5	93.8	99.2	99.2	99.0	2.75M
LCGN(2019)	0	97.9	–	–	–	–	–	15.68M
MAC(2018)	0	**98.9**	97.2	**99.4**	**99.5**	**99.3**	**99.5**	12.20M
NSCL(2019)	0	**98.9**	98.2	99.0	98.8	**99.3**	99.1	–
MCAN(2019)	0	98.3	98.5	99.0	96.2	98.9	99.0	3.71M
Ours	0	**98.9**	**99.0**	**99.4**	97.7	**99.3**	99.3	4.12M

Table 1: Performance on CLEVR val split. The model is described in the groups of whether utilizing programs or not. Human evaluation is obtained via taking a majority vote among collected responses (Johnson et al. 2017).

Conclusion

In this paper, we have proposed Transformer-based global fusion attention for vision and language understanding. Experimental results demonstrate that, without extra annotations and stronger feature representations, our approach is still comparable with multiple strong baselines on CLEVR main dataset. Additionally, BabyAI platform results show that the sample efficiency of our model in reinforcement learning tasks gains consistent improvements.

References

Chevalier-Boisvert, M.; Bahdanau, D.; Lahlou, S.; Willems, L.; Saharia, C.; Nguyen, T. H.; and Bengio, Y. 2018. Babyai: A platform to study the sample efficiency of grounded language learning. In *International Conference on Learning Representations*.

Hudson, D. A.; and Manning, C. D. 2018. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067* .

Johnson, J.; Hariharan, B.; van der Maaten, L.; Fei-Fei, L.; Lawrence Zitnick, C.; and Girshick, R. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2901–2910.

Mao, J.; Gan, C.; Kohli, P.; Tenenbaum, J. B.; and Wu, J. 2019. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv preprint arXiv:1904.12584* .

Yu, Z.; Yu, J.; Cui, Y.; Tao, D.; and Tian, Q. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6281–6290.