Visual Question Answering with Question-Image Hierachical Co-Attention

Yibin Hu

March 22, 2024

Tulane University

Abstract

This report analyses and reproduces the result in the paper Hierarchical Question-Image Co-Attention for Visual Question Answering[1] and compare the co-attention model with a bag of words-image baseline model, and co-attention model is then fine-tuned to do a harder classification task and be compared with pure image model. The VQA problem requires the model to reason about the question based on given image and choose the answer from a list of possible choices. The result is currently missing and to be reported.

1 Introduction

Visual question answering[2] is problem that incorporate computer vision and natural language processing. A model for such problem need to understand the relation between image and text question, and this makes attention necessary to this type of question, as attention will identify which and how any part of image or text affect the result and recognize variants of the equivalent question formulated in different way to have the same meaning. The co-attention model has two main components, co-attention which correlate and attends the image and question feature, and question hierarchy which separate the text into three levels as word, phrase, and question level, with each upper level being convoluted from lower level, this tries to capture information contained at different scales.

2 Method

VQA problem is formulated as a classification problem, so each distinct answer appeared in the training set will correspond to a class.

In this section, the baseline bag of words-image model[3] and co-attention model is described.

2.1 Bag of Words + Images

The bag of words-image model concatenate image features extracted from pretrained visual model GoogleNet and one-hot encoded word featue, and feed this feature vector into a softmax layer to get prediction.

2.2 Hierarchical Co-attention

Given a question of T words, its one-hot encoding is $Q = \{q_1, ..., q_T\}$, and from this the embedding of each word forms word-level feature $Q^w = \{q_1^w, ..., q_T^w\}$. From word-level feature, the phrase-level feature Q^p is formed by computing the inner product of the word vectors with filters of three window sizes: unigram, bigram and trigram as

$$\hat{\boldsymbol{q}}_{s,t}^p = \tanh\left(\boldsymbol{W}_c^s \boldsymbol{q}_{t:t+s-1}^w\right), \quad s \in \{1, 2, 3\}$$

and then max-pooled as

$$\boldsymbol{q}_t^p = \max\left(\hat{\boldsymbol{q}}_{1,t}^p, \hat{\boldsymbol{q}}_{2,t}^p, \hat{\boldsymbol{q}}_{3,t}^p\right), \quad t \in \{1, 2, \ldots, T\}$$

From the phrase-level feature Q^p, the question-level feature $q_t^s \in Q^s$ is encoded as LSTM hidden state at t. The hierarchical representation is illustrated in figure 1.a.

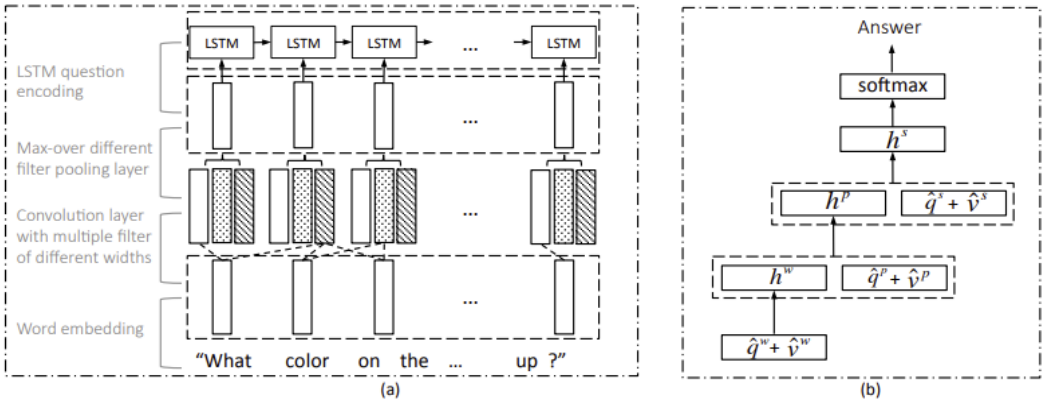

Figure 1: a) Hierarchical question encoding b) Encoding for predicting answers

For each level of word feature $Q^r \in R^{d_r \times T}$ $(r \in w, p, s)$ and image feature $V \in R^{d_2 \times T}$, affinity matrix $C \in R^{N \times T}$ is computed as

$$C = \tanh\left(Q^T W_c V\right)$$

then the attention vector for image and feature is trained as

$$\boldsymbol{H}^v = \tanh\left(\boldsymbol{W}_v \boldsymbol{V} + \left(\boldsymbol{W}_q \boldsymbol{Q}\right) \boldsymbol{C}\right), \boldsymbol{H}^q = \tanh\left(\boldsymbol{W}_q \boldsymbol{Q} + \left(\boldsymbol{W}_v \boldsymbol{V}\right) \boldsymbol{C}^T\right)$$
$$\boldsymbol{a}^v = \text{softmax}\left(\boldsymbol{w}_{hv}^T \boldsymbol{H}^v\right), \boldsymbol{a}^q = \text{softmax}\left(\boldsymbol{w}_{hq}^T \boldsymbol{H}^q\right)$$

which are used to form weight-summed image and question feature as

$$\hat{\boldsymbol{v}} = \sum_{n=1}^{N} a_n^v \boldsymbol{v}_n, \quad \hat{\boldsymbol{q}} = \sum_{t=1}^{T} a_t^q \boldsymbol{q}_t$$

The parallel co-attention is done at each level in the hierarchy, leading to $\hat{\boldsymbol{v}}^r$ and $\hat{\boldsymbol{q}}^r$ where $r \in \{w, p, s\}$. Illustrated in figure 1.b, the final prediction is recursively encoded by a MLP from

hierarchical features as

$$\begin{aligned}
\boldsymbol{h}^w &= \tanh\left(\boldsymbol{W}_w\left(\hat{\boldsymbol{q}}^{\boldsymbol{w}} + \hat{\boldsymbol{v}}^{\boldsymbol{w}}\right)\right) \\
\boldsymbol{h}^p &= \tanh\left(\boldsymbol{W}_p\left[\left(\hat{\boldsymbol{q}}^{\boldsymbol{p}} + \hat{\boldsymbol{v}}^{\boldsymbol{p}}\right), \boldsymbol{h}^w\right]\right) \\
\boldsymbol{h}^s &= \tanh\left(\boldsymbol{W}_s\left[\left(\hat{\boldsymbol{q}}^{\boldsymbol{s}} + \hat{\boldsymbol{v}}^{\boldsymbol{s}}\right), \boldsymbol{h}^p\right]\right) \\
\boldsymbol{p} &= \mathrm{softmax}\left(\boldsymbol{W}_h \boldsymbol{h}^s\right)
\end{aligned}$$

with $\boldsymbol{p}$ being used to classify the most likely answer.

3 Dataset

The model will be trained and evaluated on VQA dataset.

4 Experiments

5 Discussion

6 Conclusion

References

[1] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in Advances In Neural Information Processing Systems, pp. 289–297, 2016.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In ICCV, 2015.

[3] Zhou, Bolei, et al. "Simple baseline for visual question answering." arXiv preprint arXiv:1512.02167 (2015).