Visual Question Answering

Yibin Hu

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. Given the dataset, the result of this report shows that complex attention model don't simple but effective baseline model, which doesn't necessarily implies that use of attention is incorrect, instead it suggests that VQA problem shouldn't be formulated as classification problem due to its infeasiblity to scale as answer space increases.

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 Background

The task of Visual Question Answering (VQA) represents a complex challenge at the intersection of computer vision and natural language processing. The goal is to enable a model to provide accurate answers to questions about the contents of an image. This requires a nuanced understanding of visual elements and their semantic relationships within the context of the query. Pioneering works in the field, such as that by Antol et al. (2015) with the VQA dataset, have laid the groundwork for this interdisciplinary endeavor. Subsequently, various approaches have been proposed to improve the accuracy and reliability of VQA systems. Among these, attention mechanisms have emerged as a significant advancement, allowing

models to focus on relevant parts of the image in relation to the question asked. The Hierarchical Co-Attention model proposed by Lu et al. (2016) introduces a method of attending simultaneously to both the visual and textual inputs, creating a synergistic effect that enhances the model's comprehension.

3 Approach

VQA problem is formulated as a classification problem, so each distinct answer appeared in the training set will correspond to a class. All possible answers forms the answer space, from which the model make selection.

The Bag of Words baseline model leverages pre-trained GoogleNet features alongside one-hot encoded textual data, followed by a softmax layer to predict the outcome. The CLIP baseline extracts features without training on the dataset, relying on its robust pre-existing embeddings. In contrast, the Hierarchical Co-Attention model builds a complex structure that correlates image and text features at multiple levels of granularity, from individual words to entire questions.

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

3.1 Bag of Words + Images

The bag of words-image model concatenate image features extracted from pretrained visual model GoogleNet and one-hot encoded text featue, and feed this feature vector into a softmax layer to get prediction. In detail, the question text is first tokenized and then sent to a embedding layer in which model learns the embedding vector of the question text, then the embedding vector is concatenated with image feature vector and then send into a fully connected classifer layer for final prediction.

3.2 CLIP baseline

Instead of learning the embedding vector from dataset, which may be constrained from sparsity of text feature present in the dataset. Another baseline is considered by using matured CLIP model for embedding vector extraction, since CLIP also has bulit-in image feaute extractor, the CLIP baseline just extracts text and image feature and send them into fc classifer.

3.3 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.

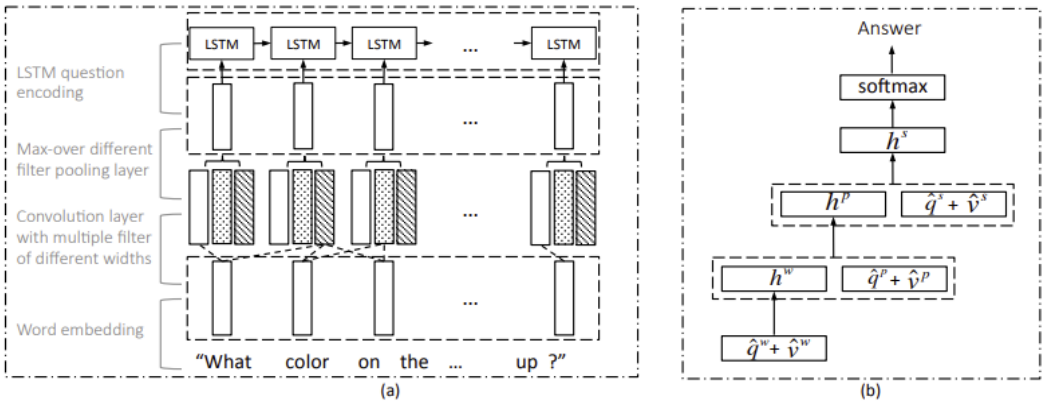

Figure 1: a) Hierarchical question encoding b) Encoding for predicting answers

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

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

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

3.4 Dataset

The model will be trained and evaluated on a small Visual Question Answering (VQA) dataset available on Kaggle. This dataset[4] consists of images paired with corresponding questions and answers. Each image-question pair is annotated with the correct answer, providing a rich resource for training and evaluating VQA models.. It contains 9974 training samples and 2494 testing samples. The answer space size is 582, so random guesser has around 0.002 accuracy.

4 Experiments and Results

All three models are trained for 60 epochs and then tested every 10 epochs starting from 10 epochs. For an given input image-question pair, the model outputs a vector of logits of the size 582 same as the size of the answer space, and loss is calculated by cross entropy loss, so the model is meant to choose the best match answer from all possible answers. Due to the large size of answer space, using argmax to choose answer with largest logit doesn't give good performance, instead the performance is measured by recall and precision which is affected threshold. Threshold is set so that answers with logits above that are chosen. The result shows that attention model is beaten by baseline models, while CLIP baseline is far better than bag of words baseline that tries to learn and use question embedding vector from current dataset. The precision and recall result for bag of words and CLIP baseline is shown in figures below.

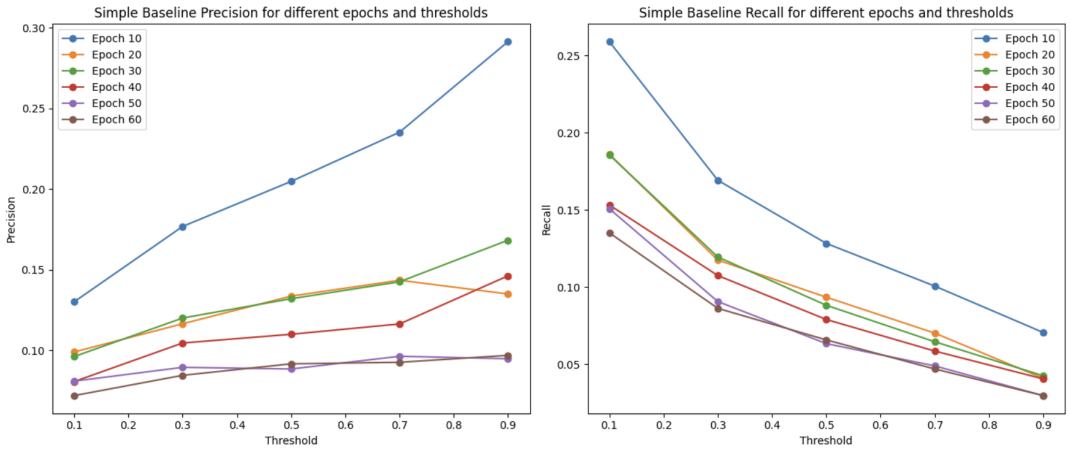

Figure 2: Results of bag of words baseline

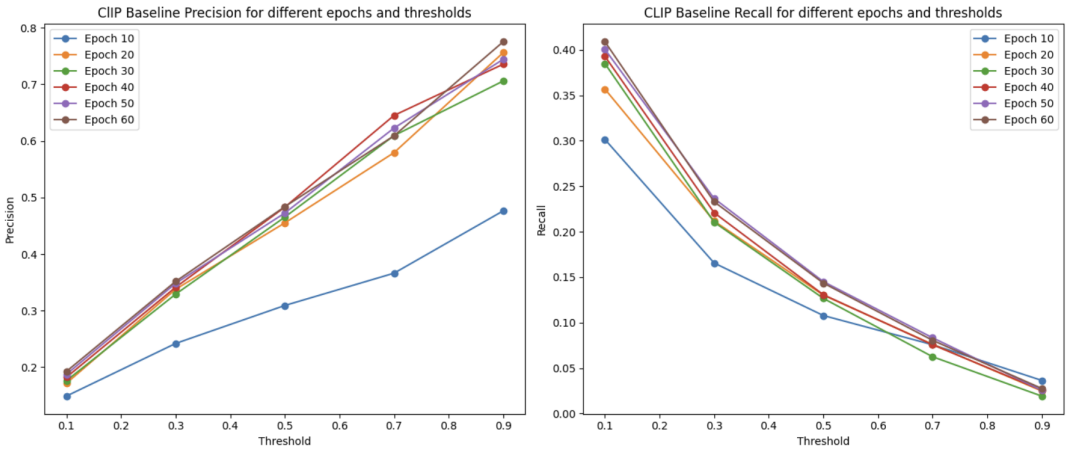

Figure 3: Results of CLIP baseline

It is expected that as threshold increases, precision increases and recall drops. The recall and precision plot for Co-Attention model is missing because, Co-Attention model consistently gives low logits for all answers such that threshold needs to be very low for non-zero

recall. Using argmax to select the answer with max logit, Co-Attention model gets to testing accuracy around 0.03 to 0.04, which is still lower than baseline's 0.08 to 0.16.

The observed disconnect between expected and actual performance raises questions about the underpinnings of VQA as a classification problem. The expansive nature of potential answers appears to dilute the Co-Attention model's effectiveness, indicating that a sheer increase in model complexity does not necessarily translate to better performance. The baselines' relative success, particularly the CLIP model, underscores the effectiveness of transfer learning and pre-trained models when faced with a broad and diverse answer space.

5 Conclusion

The experimentation with the VQA models yielded insightful results. Contrary to our expectations, the more sophisticated Hierarchical Co-Attention model did not outperform the simpler baseline models. Particularly, the CLIP baseline exhibited best performance, suggesting that the pre-trained embeddings it utilizes are robust even when not fine-tuned on the dataset.

These outcomes suggest that while attention mechanisms are intuitively appealing for VQA tasks, their practical efficacy may depend heavily on how the VQA problem is formulated. The classification approach to VQA, as demonstrated in this study, faces scalability challenges as the answer space expands. Future research might explore alternative formulations or hybrid models that can leverage the strengths of attention mechanisms without being constrained by a classification framework.

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).

[4] https://www.kaggle.com/datasets/bhavikardeshna/visual-question-answering-computer-vision-nlp