A Modern Language-Only Baseline for Visual Question Answering

Andriamarolahy Rabetokotany
(arabetokotany@aimsammi.org)
African Master for Machine Intelligence (AMMI)
Rwanda

Supervised by: Pr. Stefan Lee

August 31, 2021

Submitted in Partial Fulfillment of a Masters II at AMMI

Abstract

Recent developments in both computer vision and natural language processes have led many research to examine multimodal vision-and-language applications. A canonical problem in this space is Visual Question Answering (VQA). VQA defined as the task that a system accepts an image and natural language questions about that image as inputs to produce an accurate answer. To verify that models are actually using both visual and linguistic information, researchers use single-modality baselines. When the VQA task was originally released in 2015, deep sequence models (RNNs, LSTMs, GRUs, etc) over pretrained word-embeddings (word2vec, GLoVE, etc) were used to design question-only baselines. Since then, many modern VQA models involved in using a large-scale pre-trained language model like Bidirectional Encoder Representation from Transformers. However, it is unclear how much the improvement in performance comes from this better language encoding. The aim of this study is to retrain a language-only baseline VQA model using $BERT$ and evaluate the performance that a better language model could bring into the VQA task.

Declaration

I, the undersigned, hereby declare that the work contained in this essay is my original work, and that any work done by others or by myself previously has been acknowledged and referenced accordingly.

Andriamarolahy Rabetokotany, August 31, 2021

Contents

1. Introduction

In recent years, tasks that require reasoning about both vision and language have garnered significant attention from Computer Vision (CV) and Natural Language Processing (NLP) communities. One such task which has become very popular is Visual Question Answering (VQA) [1, 2].The VQA task is a task that a system can answer natural language questions about a given image. For example, given the image in Figure 1.1 and the question "How many horses are in this image?", a VQA system should be able to give the right answer which is 2.

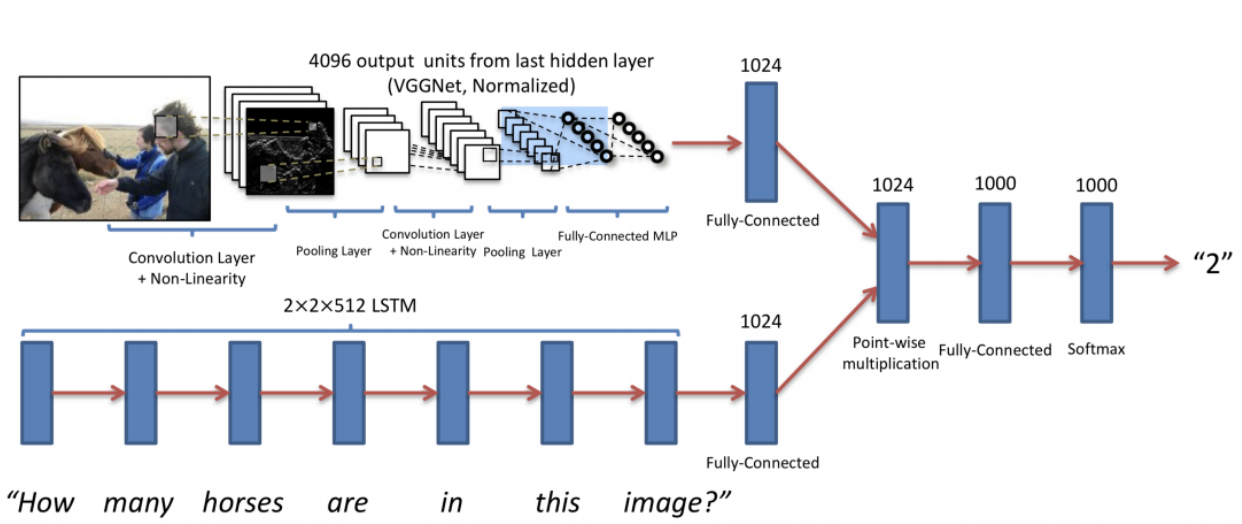

Figure 1.1: An example of VQA system. Figure reproduced from [1]

This is by nature a multimodal task as it seems to require both language understanding and visual processing. For instance, in the above example the system must understand that this is a "how many" question, so the answer must be a number, and then detect and recognize a horse based on its appearence. To validate that hypothesis, single-modality baselines are typically released with new datasets in order to know how well can a model learn to answer questions from just vision or from just language.

When the VQA dataset was released, Recurrent Neural Networks (RNN) like Long Short Term Memory ($LSTM$) [5] over pretrained word-embeddings (word2vec, GLoVE, etc) were the dominant way language was processed so the language-only baselines are based on $LSTM$s. Since then, the field of NLP has seen many breakthroughs, especially the shift to transfer learning. Large-scale pretrained language models like Bidirectional Encoder Representations from Transformers ($BERT$) [8] are now available and can be reused and fine-tuned for smaller tasks, rather than training a new model from scratch. $BERT$ is a deep learning based model that has given state-of-the-art results on a wide variety of NLP tasks from sentiment analysis to question answering.

Many VQA models have incorporated these advanced techniques. However, it is not yet determined how much is the performance from these techniques. The purpose of this project is to retrain a language-only baseline VQA model using a pre-trained $BERT$ model and evaluate how much improvements in VQA come from better language models.

This work is composed of four chapters with the following contents: chapter 2 provides a general literature survey about VQA task, chapter 3 gives some backgrounds study of $BERT$ model, chapter 4 describes the methods used for this project, chapter 5 discusses the experimental results and finally gives conclusion and some perspectives.

2. Literature survey

In their seminal paper of 2016, Antol et al. [1] introduced the task of Visual Question Answering (VQA) with a large dataset containing over $250K$ images, $760K$ questions, and around $10M$ answers. Different baseline models were benchmarked with this dataset in which the authors reported that the language-alone model outperformed the vision-alone model. Further analysis suggested that this is due to the language bias in the dataset. This hypothesis is confirmed by Goyal et al. in [2] when they tried to aleviate the problem of language biases in the original dataset.

In general, most of the existing methods consist of 1) extracting features from image and question then combine the obtained features to produce an answer and 2) framing VQA as a classification problem in such way the image and question features are the inputs to the classification system and each answer represents a class. Pre-trained CNNs model like $VGGNet$ [3] and $ResNet$ [4] are the most common image features extractor while there are a wide variety of question features extractor such as bag-of-words (BOW), long short term memory ($LSTM$) encoders [5], gated recurrent units (GRU) [6], and skip-thought vectors [7]. Thus the main difference between them is how they integrate the question and image features before feeding them to a classification network. Several techniques ranging from trivial like concatenation, elementwise multiplication, or elementwise addition [10] to the more sophisticated like using approach that maps questions and images to vector representations in a common feature space [9] have been used to combine the question and image features.

Several researchers have highlighted the performance and limitation of the existing datasets and techniques. For instance C. Kanan et al. found that $ResNet$ outperforms $VGGNet$ or $GoogLeNet$ across multiple algorithms. [11] [12] stated that attention based mechanisms, modular and memory-augmented architectures improve model's performance. However existing datasets and models have major shortcoming: evaluating answers generated by algorithms is a challenging problem, and existing datasets suffer from biases and other problems [11]. Various approaches have been hypothesized to solve these issues for future research. [11] [13] [14] suggest to develop more sophisticated deep learning models as well as more challenging datasets for VQA. C. Kanan et al. [11] underline that evaluating different kinds of questions individually in a nuanced manner, rather than using naive accuracy alone, will significantly help. Q. Wu et al. [12] propose to scale up the inclusion of additional external knowledge from structured knowledge bases, as well as a continued exploration of the potential of NLP tools.

The focus of this study will be based on the suggestion of Q. Wu et al. to explore the potential of NLP tools by getting the power of $BERT$ inolve into the VQA workflow.

3. Background: BERT

The shift to transfer learning is one of the major breakthrough in the field of NLP. Instead of training new models from scratch, pre-trained models like $ELMo$ [15] and OpenAI's GPT [16] are now available and can be fine-tuned on smaller task-specic datasets such as question answering and sentiment analysis. $BERT$ is one of the best among these techniques that has given state-of-the-art results on a wide variety NLP tasks [8].

3.1 The core idea behind BERT

BERT model was released by the Google's researchers at the end of 2018. It has been pre-trained on the data extracted from the BooksCorpus and English Wikipedia [8].

The key difference of $BERT$ from other popular embedding models, such as word2vec, is that it is a context-based embedding model. This means it generates a representation of each word based on the other words in the sentence. For example, consider the two following sentences: *"I am learning python."* and *"I am scared of the big python."*. The word *python* which refers to a programming language will have the same representation as the word *python* which is a snake using the context-free models. In contrast, $BERT$ will understand the context and generate the representation of the word accordingly. So for the two example sentences, the word *python* will have different representations based on the context.

Instead of predicting the next word in a sentence, BERT model makes use of a new technique called Masked Language Modeling (MLM) to perform this contextual encoding. MLM randomly masks words in the sentence and then tries to predict them. Thus the model must looks in both directions and uses the full context of the sentence in order to predict the masked words. This implies that BERT is bidirectionally trained.

3.2 BERT architecture

As the name implies, BERT is a Transformer-based model [17]. A transformer consists of encoding and decoding the input text to produce a prediction by learning contextual relationships between words from text. But $BERT$ needs only the encoder part since its goal is to generate a language representation model. Thus it is a multi-layer bidirectional Transformer encoder.

There are two types of pre-trained versions of $BERT$ depending on the scale of the model architecture [8]: $BERT$-base contains 12 transformer blocks (layers), 12 attention heads and 110 million parameters and $BERT$-large contains 24 layers, 16 attention heads and 340 million parameters.

3.2.1 BERT functionality. The input to the encoder for $BERT$ is a sequence of tokens, which are first converted into vectors and then processed in the neural network. But before that, the input needs to be in the same format as required by $BERT$:

- Token embeddings: a $[CLS]$ token is added to the input word tokens at the beginning of the first sentence and a $[SEP]$ is inserted at the end of each sentence.

- Segment embeddings: a marker indicating Sentence A or Sentence B is added to each token to allow the encoder to distinguish between sentences.
- Position embeddings: Position embeddings is added to each token to indicate its position in the sentence.

So the input representation for BERT is then the sum of the token embedding, the segmentation embedding and the position embedding.

A word starts with its embedding representation from the embedding layer. Every layer does some multi-headed attention computation on the word representation of the previous layer to create a new intermediate representation. All these intermediate representations are of the same size.

In this project we are taking advantage of the power of $BERT$ to train a language-alone baseline for the VQA task.

4. Methods

To evaluate how the adoption of pre-trained language encoders affects question-only VQA performance, we will finetune a BERT model as a uni-modal baseline.

4.1 Data

For the purpose of training a language-alone baseline, our model accepts only the questions as inputs and the top $K = 1000$ most frequent answers as possible outputs.

4.2 Methods

In attempt to measure the performance that a better language model can provide to VQA, we essentially have the same method as described by Antol et al. [1] and Goyal et al.[2] when training a language-alone model. Specifically, in our procedure a pre-trained $BERT$-base model is used instead of $LSTM$ to obtain 1024-dim embeddings with the same hidden size for the question. We then extracted the last hidden layer of the [CLS] token and fed into a feed-forward classifier with 2 fully connected layers and $ReLU$ non-linearity followed by a softmax layer to obtain a distribution over 1000 answers. And finally, the answer with the highest activation among them is selected as the predicted answer.

Our model was trained on the training set and tested on the validation set from the balanced version of the VQA dataset [2] using Pytorch and Hugging Face's Transformers. The model is then learned end-to-end with a binary cross entropy with logits loss and optimized using AdamW with careful chosen set of hyper-parameters.

For testing, the dataset papers for VQA suggest to report results on test-standard accuracies and compare their methods' accuracies. However, for the sake of time restriction our results are reported on validation only.

4.3 Evaluation metric

To evaluate the performance of our model, we use the standard evaluation metrics [1] defined in the equation 4.1. As each VQA question has 10 human answers, there may be multiple correct solutions. To account for this, the VQA accuracy metric considers an answer fully correct if it matches at least three of the ground truth answers for the associated question.

$$accuracy \quad = \quad min\left(\frac{\#\ Matching\ \ Answers}{3}, 1\right) \tag{4.1}$$

5. Results

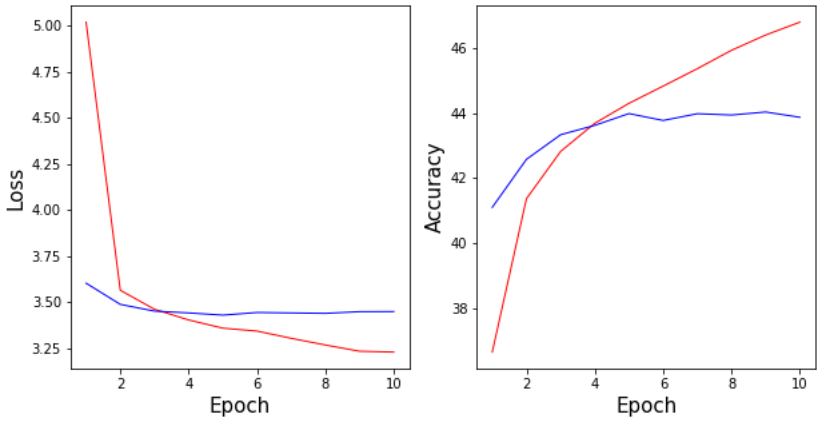

Figure 5.1: **Left**: plot of the training (red) and the validation (blue) loss, **Right**: plot of the training (red) and the validation (blue) accuracy with 10 epochs

Figure 5.1 shows our experimental results. We can see from the plot that the validation start saturated around 43 and stop increasing at 43.9. Even if it is slightly lower than we expected, this value is already an improvement of the one obtained from Deeper $LSTM$ (d-$LSTM$) as reported in [2]. Table 5.1 summarizes the comparison of our experimental result to that of d-$LSTM$.

	d-$LSTM$	$BERT$-base
accuracy	43.01	43.9

Table 5.1: Accuracy of $BERT$ compared to d-$LSTM$ on VQA $v2.0$ val

Given that our findings are based on a limited tweaks of the hyper-parameters, due to the time and resource restriction, we believe that there is certainly room for improvement. Nevertheless, there is evidence to support the hypothesis that the VQA task benefits from having a better language model like $BERT$.

6. Conclusion

Our work has led us to conclude that $BERT$ outperforms the d-$LSTM$. Despite the fact that there is a restriction from having better performance, these results suggest that a better language model like $BERT$ could bring additional performance to VQA task.

Our results are encouraging and should be validated by a further better methods. For instance, our results should reported on test-standard as the original papers suggested. Additionally, tuning the hyper-parameters should be done from further analysis.

Acknowledgements

I gratefully acknowledge the help provided by Pr. Stefan Lee for supervising me along this work.

Bibliography

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. *VQA: Visual Question Answering*, 2016.

[2] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra and D. Parikh. *Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering*, 2017.

[3] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, in International Conference on Learning Representations (ICLR), 2015.

[4] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[5] S. Hochreiter and J. Schmidhuber, *Long short-term memory*, 1997.

[6] K. Cho, B. Van Merri¨enboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, *Learning phrase representations using rnn encoder-decoder for statistical machine translation*, in Conference on Empirical Methods on Natural Language Processing (EMNLP), 2014.

[7] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler, *Skip-thought vectors*, in Advances in Neural Information Processing Systems (NIPS), 2015.

[8] J. Devlin, M. Chang, K. Lee and K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2019.

[9] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola, *Stacked attention networks for image question answering*, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[10] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, *Simple baseline for visual question answering*, arXiv preprint arXiv:1512.02167, 2015.

[11] K. Kafle and C. Kanan. *Visual Question Answering: Datasets, Algorithms, and Future Challenges*, 2017.

[12] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick and A. van den Hengel. *Visual Question Answering: A Survey of Methods and Datasets*, 2016.

[13] A. K. Gupta. *Survey of Visual Question Answering: Datasets and Techniques*, 2017.

[14] Y. Srivastava, V. Murali, S. R. Dubey, and S.Mukherjee. *Visual Question Answering using Deep Learning: A Survey and Performance Analysis*, 2020.

[15] Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power, *Semi-supervised sequence tagging with bidirectional language models*, 2017.

[16] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever, *Improving language understanding with unsupervised learning*. Technical report, OpenAI, 2018.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, Aidan N. Gomez, L. Kaiser and I. Polosukhin *Attention Is All You Need*, arXiv:1706.03762 [cs.CL], 2017.