Generate Answer to Visual Questions with Pre-trained Vision-and-Language Embeddings

Hadi Sheikhi, Maryam Hashemi, Sauleh Eetemadi
Iran University of Science and Technology
{ha_sheikhi, m_hashemi94@comp.iust.ac.ir, sauleh@iust.ac.ir}

Abstract

Visual Question Answering is a multi-modal task under the consideration of both the Vision and Language communities. Present VQA models are limited to classification answers and cannot provide answers for reasoning questions. In this work, we introduce an encoder-decoder model using vision-and-language pretrained embedding, which delivers multi-word generated sentences as answers. We utilise LXMERT and VisualBERT embedding space with three different generative decoder heads, including RNNs, Attention RNNs and Transformers. Extensive experiments show competitive performance on the FSVQA dataset through qualitative and quantitative evaluation and a Human Error Analysis.

1 Introduction

Over the past few years, most end-to-end VQA models seek to learn joint representations using visual and textual content and perform classification over a predetermined set of candidate answers based on the joint representations instead of generating an answer. As a result, the output answer is usually a single word. However, many questions that require reasoning cannot be answered in one word.

Picking the answer from a set of candidate answers turns question answering into pattern matching. However, considering VQA as an answer generation task instead of an answer selection task can generate more natural and richer answers and prevent model from overfitting to biases in the dataset.

Hence, we propose a generative model with two major components. The first is the vision-and-language pretrained model as an encoder encodes visual and linguistic information in joint embeddings. The second is the language sequence decoder, which uses encoded information to decode answers.

2 Methodology

Our method consists of implementing an encoder-decoder architecture and experimenting with various models to determine the combination with the best performance. We train our model on the Full-Sentence Visual Question Answering (FSVQA) dataset (Shin et al., 2016).

2.1 Encoders

The encoder part of our architecture extracts features from the image and question. In this study, we utilize two different pretrained models, including LXMERT (Tan and Bansal, 2019) as a dual-stream VLP and VisualBERT (Li et al., 2019) as a single-stream VLP.

2.2 Decoders

The decoder part of our architecture extracts the answer of the input image and question. We employ RNN-based and Transformer-based networks as decoders. To investigate the applicability of RNNs and Transformers to answer generation, we evaluate the performance of several variations of our proposed architecture, such as changing RNN cell types, adding attention to RNNs and using different attention mechanisms, and changing the number of RNNs and Transformer layers.

3 Experiments and Results

Evaluating natural language generation models is challenging. The similarity between reference and predicted answers should be measured by some means. We use BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and RougeL (Lin, 2004) as our n-gram-based metrics. Embedding-based metrics can be Average score and BERTScore.

We report the results in Table 1. LXMERT-3Transformer achieves the best performance in all metrics and shows marginal improvement over

Method		Word-based			Embedding-based	
Encoder	Decoder	BLEU	METEOR	ROUGE-L	Average Score	BERT Score
LSTM Q+I(Shin et al., 2016)	LSTM(Shin et al., 2016)	23.9	23.3	-	-	-
LXMERT	3-BiLSTM	43.54	66.39	64.74	90.58	84.02
	1-BiGRU	40.89	64.15	62.66	89.88	82.46
	1-LSTM+Bahdanau attention	79.03	86.43	85.49	95.94	91.84
	1-LSTM+Luong(general) attention	79.54	86.96	86.25	96.11	91.90
	1-GRU+Bahdanau attention	71.40	83.26	82.62	95.42	89.32
	1-GRU+Luong(dot) attention	73.92	84.71	83.84	95.73	90.37
	3-Transformer Decoder	**86.73**	**91.18**	**90.60**	90.20	**95.01**
VisualBERT	3-BiLSTM	20.46	41.51	41.19	87.21	73.43
	1-BiGRU	22.72	45.66	45.26	88.42	74.32
	1-LSTM+Bahdanau attention	84.27	88.07	87.28	97.11	93.50
	1-LSTM+Luong(concat) attention	82.90	87.71	86.90	**97.17**	93.11
	1-GRU+Bahdanau attention	72.20	82.87	82.40	96.10	89.26
	1-GRU+Luong(dot) attention	79.65	86.17	85.23	96.93	91.81
	3-Transformer Decoder	85.95	89.76	89.09	91.94	94.44

Table 1: Results of our proposed models on FSVQA dataset. The numbers in the decoder names mean the number of layers. Bold indicates best overall performance, while an underline indicates best in encoder category performance.

LSTM Q+I as the baseline. Comparing different encoders together demonstrates that when we utilize RNN and Transformer as a decoder, LXMERT outperforms VisualBERT. However, when we use Attention RNNs, VisualBERT produces better results. By comparing performance across the decoders, Attention RNNs surpass RNNs by a large margin. This is due to Global Attention's ability to focus more on the critical parts of the input data to generate answers and solve the forgetting problem of RNNs dependencies between words in the sequence. Furthermore, using Transformers as a decoder shows the best performance compared to RNNs and Attention RNNs.

4 Error Analysis

This section investigates why Transformers as decoders perform better than RNNs and Attention RNNs. We randomly select 100 instances from the test set based on each question category's ("yes/no", "counting", "color detection", and "others") distribution overall and analyze the answers generated by each decoder's models trained on the FSVQA training set. We evaluate the best models LXMERT-3BiLSTM, VisualBERT-BahdanauLSTM and LXMERT-3Transformer. The generated answers when examined manually, broadly fell within the following categories.

- **EM** (Exact Match)
- **WA** (Wrong Answer): The model only generates incorrect VQA single/multi word answers while the context and description are correct.
- **GE** (Grammatical Error)
- **WD** (Wrong Description): When the model generates the correct VQA single/multi word answer, but the description is wrong.
- **E** (Error): The generated answer is completely wrong.
- **AA** (Alternative Answers): These answers are correct but are not an exact match.

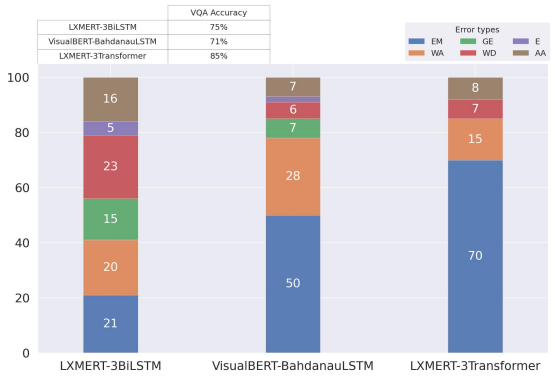

Figure 1: Number of errors different generative VQA models make, split by error category.

In Figure 1, we show the number of errors that different models make, aggregated by error category. We note that LXMERT-3Transformer tend to produce fewer errors across all error categories, especially not generating any sentence with grammatical issues. We also conclude from Figure 1 that Transformers outperform RNNs and Attention RNNs to generate answers. Moreover, they had results better than RNNs and Attention RNNs in VQA singular answers.

5 Conclusion

We have presented a novel encoder-decoder model using vision-and-language pretrained embedding, a generative solution for the visual question answering task. We implemented different decoder heads, including RNNs, attention RNNs and Transformers on some popular VLP such as LXMERT and VisualBERT. Empirical results on the FSVQA dataset show that our model is comparable with the classification settings of VQA. We also show the effectiveness of several model components and training methods via detailed analysis.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Andrew Shin, Yoshitaka Ushiku, and Tatsuya Harada. 2016. The color of the cat is gray: 1 million full-sentences visual question answering (FSVQA). *CoRR*, abs/1609.06657.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.