Visual Question Answering on Video and Text

Aaron Honculada, Aisha Urooj Khan, Mubarak Shah
University of Central Florida
Center for Research in Computer Vision
achonculada@knights.ucf.edu aishaurooj@gmail.com shah@crcv.ucf.edu

Abstract

In this paper we discuss methods for training LSTM-based architectures to perform visual question answering. LSTM have been shown to perform well on sequential data, such as text or audio. Given the multi-modal structure of VQA tasks, LSTMs are a common baseline in which to test novel models against. Building upon our LSTM-based architectures we explore the performance of our models with the inclusion of both self-attention and guided-attention modules. We have found that while the inclusion of attention assists our model to perform with a similar accuracy as those trained without attention, we were not able to exceed the accuracy of our best performing baseline model which did not use attention.

1. Introduction

Visual Question Answering (VQA) lies at the intersection of Computer Vision and Natural Language Processing. Both are open areas of research in the field of artificial intelligence and computer understanding and reasoning. Given an image or video and a question on the visual information, VQA systems attempt to correctly answer the question. Computer vision systems are able to recognize objects in an image, utilize action recognition, and perform image segmentation. Language models can both parse and learn from text leading to impressive tasks like word prediction. In order for successful visual question answering, both of these fields need to be used as well as expanded upon towards a form of visual understanding. Applications of VQA include assistance for the blind and visually impaired, filtering inappropriate and offensive content on social media, augment image retrieval systems. We investigate how to make the visual information essential to VQA, while it has been shown that past approaches have mainly leveraged the information in the text alone.

2. Related Works

Works which relate to our investigation of Visual Question Answering include the Neural-Symbolic VQA and Neural-Symbolic Concept Learner which proposed a way to perform VQA without the use of deep learning [3] [4]. The Transformer network introduced attention-based approach in the form of associating weights with each input of the sequence [1].

3. Approach

3.1 Dataset

We use the TVQA dataset which includes over feature for the model to learn on Visual concepts, video features from ResNet, subtitles, clips from scene with timestamps. The TVQA dataset contains clips from popular television series which are more similar to real-world scenarios. Other datasets for VQA such as DAQUAR or CLEVR are solely image-based while other video-based datasets such as PororoQA and MarioQA do not provide realistic scenarios. The TVQA dataset provides a range of scenarios centered around human activity while the questions and answers of the dataset are generated by humans [6]. Each question contains a grounding section, in which the answer to the question is temporally localized in the scene, and a main section whose answer is one of the five multiple choice answers.

3.2 Methodology

After replicating the results of the state-of the-art model, we train the LSTM module and analyze the results. We compare both a unidirectional and bidirectional [1] LSTM and build our novel network

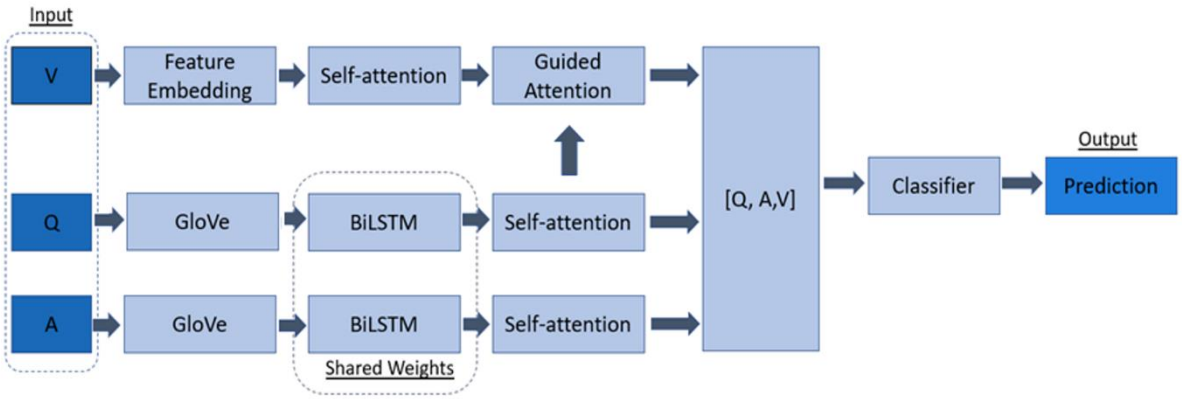

Figure 2: Attention-based CNN+LSTM

upon the modules which give us the best results, using a cross-entropy loss the function and validation accuracy as our metric to compare models.

We look at the total amount of questions our best model answered correctly, comparing our model trained only on the text in the form of question and answers against a model trained on questions, answers, and video features. This comparison allows us to see the amount of questions that each model generated a correct prediction and more importantly the amount of questions each model exclusively guessed correctly. By looking at this exclusivity, we were able to analyze the effects of training on text versus training on video and text.

We then parse the questions input into question families and subqueries and categorize them according to what is being asked in the question. The question families are as follows: Who, What, Where, Why, and How. We further classify the What question family into subqueries which focus on if the answer to what is being asked is an Object, Action, or a third class we call Abstract. The answers to the Abstract questions are found in the dialog and usually do not strictly pertain to the visual information in the scene.

Taking our highest performing network, we add attention-based modules to investigate the changes in the generated predictions. We used self-attention in the form of producing a weighted score from a query matrix, a key matrix, and a value matrix [2].

4. Experiments

Baseline models common to VQA tasks were trained on the TVQA dataset and the accuracy was recorded. Both LSTM and BiLSTM were trained on the question and answers, visual concepts, video features from ResNet. Our proposed CNN+LSTM model was trained on questions, answers, and both the features from the final layer of I3D and the features from the convolutional layers of I3D.

The CNN+LSTM architect was trained with attention in the forms of self-attention and guided attention (Fig. 2). The model was trained on 1, 2, 4, 6, 8, 12, and 15 attention heads, with 4 heads giving the best results.

5. Results

The results of the various baseline models and attention-based model were compared against the state-of-the-art model's performance. (Fig. 3) The best results which came from the CNN+LSTM architecture which was trained on the fully connected features from I3D.

Model	Feature	Accuracy
LSTM	--------	42.48%
	VCPT	42.67%
CNN+LSTM	2D CNN (FC)	42.73%
	I3D (FC)	42.86%
	I3D (Conv)	42.59%
Attention	I3D (FC)	42.73%
TVQA	2D CNN (FC)	43.78%

Table 1. Comparison of baseline and attention-based models with the TVQA model.

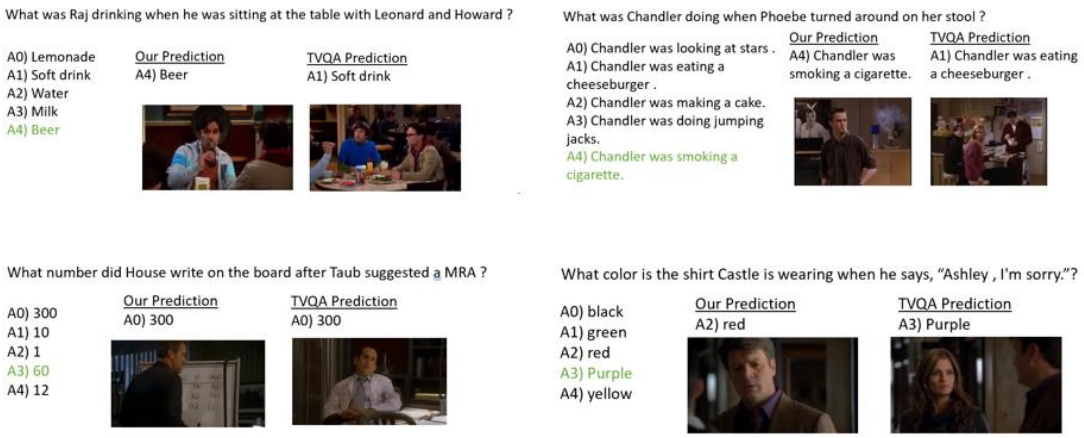

Figure 3. Results of the CNN+LSTM using fully connected features of I3D.

6. Discussion

In this report we propose an attention-based network alongside a CNN+LSTM network to investigate these essential components to successful VQA models. While we were not able to exceed the accuracy of the state-of-the-art model, we were able to investigate if the addition of attention modules to common baseline models for VQA would improve performance. There will be additional work on different iterations of the inclusion of attention to our existing network.

7. Future Work

Future work includes training different configuration of attention-based networks. While maintaining the underlying CNN+LSTM architecture, we will experiment with the placement of the attention modules. Other variation of self-attention and guided attention will be used. In the case of guided attention, we will attend videos with question as both the query and key.

We believe that it is possible to achieve higher accuracy scores on our attention-based models with further fine tuning of the hyper parameters of our model, such as varying the number of attention heads as well as increasing the layers of attention in the network.

We also intend to implement full transformer network for the TVQA dataset which replace the recurrent components of our proposed model. Other attention-based models to be implemented are the self-attention GAN (SAGAN) [10] and the Video Action Transformer Network [11].

8. Acknowledgements

We would like to thank the National Science Foundation for funding the Research Experience for Undergraduates program at the Center for Research in Computer Vision at UCF. We would also like to thank Dr. Shah and Dr. Lobo for their help and support throughout this program.

References

[1] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, Hannaneh Hajishirzi. Bidirectional Attention Flow for Machine Composition. arXiv preprint arXiv:1611.01603

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Lones, Aidan N. Gomez, Lukasz Keiser, Illia Polosukhin. Attention is all you need. arXiv preprint arXiv:1706.03762

[3] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, Joshua B. Tenenbaum. Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. arXiv preprint arXiv:1810.02338

[4] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, Jiajun Wu. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. arXiv preprint arXiv:1904.12584

[5] Drew A. Hudson, Christopher D. Manning. Compositional Attention Networks for Machine Reasoning. arXiv preprint arXiv:1803.03067

[6] Jie Lei, Licheng Yu, Mohit Bansal, Tamara L. Berg. TVQA: Localized, Compositional Video Question Answering. arXiv preprint arXiv:1809.01696

[7] Jie Lei, Licheng Yu, Tamara L. Berg, Mohit Bansal. TVQA+: Spatio-Temporal Grounding for Video Question Answering. . arXiv preprint arXiv:1904.11574

[8] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, Devi Parikh. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. arXiv preprint arXiv:1612.00837

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv preprint arXiv:1311.2524

[10] Han Zhang, Ian Goodfellow, Dimitris Metaxas, Augustus Odena. Self-Attention Generative Adversarial Networks. arXiv preprint arXiv:1805.08318

[11] Rohit Girdhar, João Carreira, Carl Doersch, Andrew Zisserman. Video Action Transformer Network. arXiv preprint arXiv:1812.02707