A Multimodal video-text embedding for video question answering

Anonymous CVPR submission

Paper ID ****

Abstract

In this work, we focus on video Question Answering task. We draw inspirations from visual question answering model for tackling video question answering problem. Specifically, we first build a multimodal representation for videos and questions, and use embedding architectures to learn semantics across different modes of input. We conduct experiments on VideoQA dataset [6] to analyze how our method deal with this problem.

1. Introduction

Visual Question and Answering is a challenging computer vision task, that has gained a lot of attention in the recent past. Addressing this task requires one to design algorithms that not just effectively integrate information available through visual and linguistic inputs, but also to extract discriminative information that is useful to successfully answer the question.

The major difference between other contemporary vision tasks and visual question answering is that the output can be determined only during the run time. Therefore, it is imperative that the challenge here might involve understanding and learning aspects of the input pertinent to the question. These aspects exclusive to visual question answering make the task quite challenging [8].

In this work, we deal with video question answering. It has one more dimension of complexity, compared to image question and answering in that it has much larger input, and the algorithms need to learn to focus on regions only pertinent to answering the question. In case of videos, it is much more challenging, and focusing on wrong temporal regions could prove counterproductive.

Embedding based approaches are popular way of dealing with visual question answering in images [5, 4, 7, 1]. These methods are inspired by the success of embedding approaches in other applications, such as image captioning [3], dense image descriptions [2], to name a few. They involve constructing robust representations for questions and images, and jointly embedding them to finally infer the output.

We propose a multimodal embedding based approach to tackle this problem. Specifically, we use recursive neural network representation for sentences and videos, and employ it in a multimodal learning based approach to learn robust information present in the input pertinent to question answering representations. We finally perform a joint embedding, to be fed to a simple classifier that learns to predict the correct option.

Through this work, we aim to analyze our embedding framework for video question question answering. We compare our approach against other recent methods to show effectiveness of our model. The salient contributions of our work are

- A new multimodal embedding based approach for learning video question answering representation.
- End to end classification style model for video question and answering.

References

[1] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016. 1

[2] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015. 1

[3] A. Karpathy, A. Joulin, and F. F. F. Li. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pages 1889–1897, 2014. 1

[4] J.-H. Kim, S.-W. Lee, D.-H. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang. Multimodal residual learning for visual qa. *arXiv preprint arXiv:1606.01455*, 2016. 1

[5] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE Inter-*

national Conference on Computer Vision, pages 1–9, 2015. 1

[6] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler. MovieQA: Understanding Stories in Movies through Question-Answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016. 1

[7] Q. Wu, C. Shen, L. Liu, A. Dick, and A. v. d. Hengel. What value do explicit high level concepts have in vision to language problems? *arXiv preprint arXiv:1506.01144*, 2015. 1

[8] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. v. d. Hengel. Visual question answering: A survey of methods and datasets. *arXiv preprint arXiv:1607.05910*, 2016. 1