Multimodal Representation for Automatic Video Description

CSC 400 - Yapeng Tian

Introduction

With the rapid growth of consumer electronics, like the smartphone, online videos quickly have been a key means for people to satisfy their information and entertainment needs. Everyday people watch hundreds of millions of hours on YouTube and generate billions of views. In addition, there are more than 400 hours of video contents uploaded to the Youtube each and every minute. It is impossible to watch all of these videos and then annotate them, therefore, intelligent techniques to analyze videos are desired. Video description is the such a technique, which can automatically generate a natural language description of the content of a video. It has various applications such as assistance to a visually impaired person and improving the quality of online video search [1]. Benefiting from its increasing demand, video description has attracted a lot of attentions from both computer vision and natural language processing communities.

The incipient research on video descriptions usually utilize template-based statistical methods [2, 3]. These approaches adopt the graphical model to learn semantic sentence representation, and then map the representation to high-level concepts, such as the actors, actions, and objects, to generate language description of video contents. Unlike these elegant hand-craft works, modern video description approaches design end-to-end deep networks to learn the mappings between videos and its language descriptions directly. Donahue *et al.* [4] use convolutional neural networks (CNN) [5] to learn the single frame representation as the input to the long-term recurrent convolutional networks to output sentences. In [6], Venugopalan *et al.* introduce a CNN-RNN model, which uses CNN to extract video frame features and then feed features into Long Short Term Memory (LSTM) Recurrent Neural Networks (RNN) [7] to translate videos into sentences. To fully exploit video contents, a temporal structure is introduced into the CNN-RNN framework in [8]. Besides, Li *et al.* propose to utilize the attention mechanism to better exploit temporal structure [9]. Although these approaches have explored the spatial and temporal information of videos, the audio contents are not been used.

Recently, several multimodal video description methods are developed in [10, 11]. However, these methods just simply concatenate audio and visual features, which are extracted from audio and vision contents separately. How to learn powerful joint multimodal representation is not been fully investigated. What's more, there is no a public dataset that is designed for multimodal video description task. Therefore, creating a new database is desired.

To this end, this proposal studies on how to learn powerful multimodal representation from vision and audio content in videos towards video description. To encourage more researchers to dive into the multimodal modeling field, PI Tian will build a new video description database to facilitate further researches.

Proposed Study

The objective of this proposal is to construct a Multimodal Video Description (MVD) dataset and develop algorithms to learn powerful multimodal representation towards accurate video description.

MVD Dataset

videos can be very noisy, there may be no audio information contained in the video, and the audio and visual content can be completely unrelated. Here, we hope that the vision and audio content of videos in MVD dataset are related. So, we introduce a correspondence classifier in [12], which can verify whether visual and audio information in the video is relevant, to select desired videos from Youtube. Furthermore, we annotate the extracted videos and each video sample will be given several language sentences to describe it.

Multimodal Representation

The success of machine learning algorithms generally depends on data representation [13]. Therefore, multimodal representation learning is the most important part of the multimodal video description system.

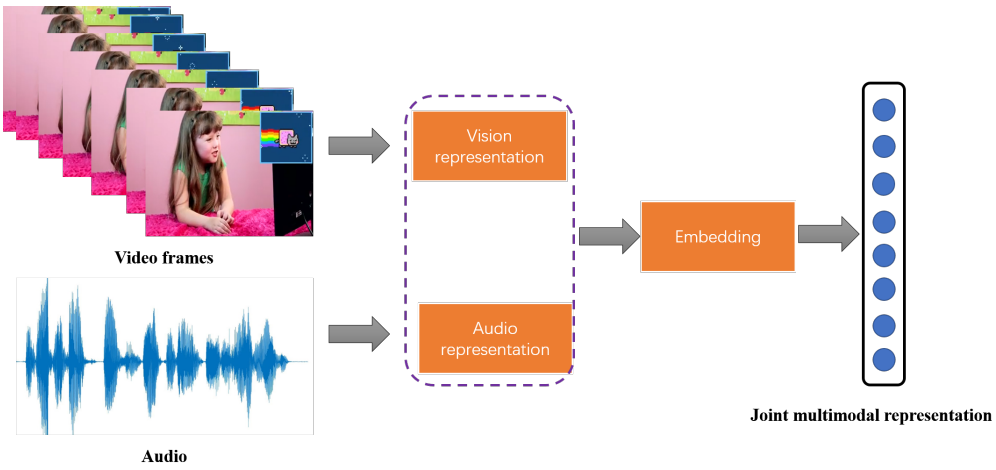

Figure 1: The pipeline of multimodal representation learning.

We propose a multimodal representation framework in Fig. 1. The framework contains three modules: vision representation, audio representation, and representation embedding. We learn vision and audio representations based on CNN, and then utilize fully connected neural networks to embed them into a unified feature space. The joint multimodal representation is capable of capturing both visual and audio semantics from raw videos. I believe that it can be transferred into other video understanding tasks, like action recognition.

Feeding the joint multimodal representation into an LSTM decoder, we can obtain a language sentence to describe the video.

Work Plan

The work will be finished in two years:

Year 1: design audio-visual correspondence classifier and create a large multimodal video description dataset;

Year 2: develop algorithms in learning multimodal representation and methods for video description.

Broader Impacts of the Proposed Work

The proposed work is an important technology that can greatly benefit the society, including healthcare (assistance to a visually impaired person) and multimedia (video searching). The created dataset and proposed multimodal representation algorithm will also make contributions to the advancement of research on multimodal video understanding.

References Cited

[1] R. Pasunuru and M. Bansal, "Multi-task video captioning with video and entailment generation," in *Proceedings of ACL*, 2017.

[2] A. Barbu, A. Bridge, Z. Burchill, D. Coroian, S. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanaswamy, D. Salvi *et al.*, "Video in sentences out," in *Proceedings of UAI*, 2012.

[3] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele, "Translating video content to natural language descriptions," in *Proceedings of ICCV*, 2013.

[4] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of CVPR*, 2015, pp. 2625–2634.

[5] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[6] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, "Translating videos to natural language using deep recurrent neural networks," in *Proceedings of ACL*, 2015, pp. 2625–2634.

[7] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proceedings of NIPS*, 2014, pp. 3104–3112.

[8] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence - video to text," in *Proceedings of ICCV*, 2015.

[9] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, "Describing videos by exploiting temporal structure," in *Proceedings of ICCV*, 2015, pp. 4507–4515.

[10] Q. Jin, J. Chen, S. Chen, Y. Xiong, and A. Hauptmann, "Describing videos using multi-modal fusion," in *Proceedings of ACM MM*. ACM, 2016, pp. 1087–1091.

[11] V. Ramanishka, A. Das, D. H. Park, S. Venugopalan, L. A. Hendricks, M. Rohrbach, and K. Saenko, "Multimodal video description," in *Proceedings of ACM MM*. ACM, 2016, pp. 1092–1096.

[12] R. Arandjelović and A. Zisserman, "Look, listen and learn," in *Proceedings of ICCV*, 2017.

[13] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.