BUPTMM Submission for YouMakeUp VQA Challenge in Step Ordering Task

Kun Liu[1], Huadong Ma[1] and Tat-seng Chua[2]
[1]Beijing University of Posts and Telecommunications, [2]National University of Singapore
liu_kun,mhd@bupt.edu.cn, chuats@comp.nus.edu.sg

Abstract

The first YouMakeUp VQA Challenge, hosted at CVPR 2020 Workshop on Language & Vision with applications to Video Understanding (LVVU), focuses on fine-grained action understanding in makeup instructional videos. This challenge presents two question-answering tasks: image ordering and step ordering. In this paper, we propose an overview and comparative analysis of our system designed for the step ordering task. First, we regard the step ordering task as the moment retrieval problem. Then, we compute the Levenshtein distance between the predicted order and each candidate. We select the candidate with the shortest distance as the final answer. Specifically, we adopt two techniques to improve performance. To retrieve the moments accurately, a late guidance module is designed to further integrate the videos and sentences tightly. We then combine multiple moment retrieval models to improve classification accuracy. Finally, our approach obtains the 85.19% accuracy on the testing set.

1. Introduction

The step ordering task aims to evaluate models' fine-grained semantic alignment abilities between textual descriptions and video clips. In essence, we define the step ordering as the moment retrieval task, which aims to temporally localize activities from unconstrained videos via language [2].

Current approaches for moment retrieval tasks mainly consist of the following stages: feature extraction stage for encoding the videos and sentences into embedding, cross-modal fusion stage for integrating the information from both modalities, and moment localization stage for identifying the start and end time of the described activity inside the video.

More specifically, these methods usually first extract powerful video features using 3D convolutional neural network (CNN) (e.g. the I3D [1]) or an ImageNet pre-trained CNN (e.g. VGG16 [4]), and obtain a sentence embedding via a LSTM over word embeddings or an off-the-shelf sentence encoder. Then video features and sentence embedding are fused in the cross-modal feature fusion stage, such as element-wise addition used in [2] and vector concatenation adopted in [5]. Finally, these methods predict the temporal position in the localization stage, which is usually composed of MLP [3], temporal convolutional networks [7] or 2D-CNN [8].

Despite the promising performance, the majority of proposed methods do not integrate the video features and language tightly. Some state-of-the-art systems ignore the sentence embedding in the localizer stage. Although these methods combine the language and video in the fusion stage, it is far from enough for moment localization due to its high demand for the proper semantic alignment between vision and linguistic domain.

To solve the above challenges, unlike previous work that only combines features in the fusion stage, we propose a framework that leverages sentence embedding in the localizer stages. We utilize the sentence embedding to guide the localizer to determine the start and end time of moments, which can further bridge the vision and language domain. Specifically, the sentence embedding is linearly transformed to generate the attention weight for each channel of feature maps.

2. Approach

We utilize the I3D features provided by the YouMakeUp organizer as the input of moment retrieval networks. We adopt one of the state-of-the-art models, 2D Temporal Adjacent Network (2D-TAN) [8], as our baseline model for the moment retrieval task. Then, we equip the late guidance module with 2D-TAN to localize the moments more precisely.

Late Guidance Module. After obtaining the multi-modal representation, we further guide the process of moment localization through the sentence embedding. Following 2D-TAN [8], we adopt the 2D-CNN as the backbone of

the moment localization. More specially, we first shift the length of sentence embedding into the channel numbers of the feature map of 2D-CNN. Then, we employ the sentence embedding as channel attention for the outputs of 2D-CNN. Specifically, we update each channel by multiplying the corresponding scalar α_i from the sentence embedding. Finally, a normalization operation is conducted to avoid over-fitting.

This module is achieved by following equation:

$$C_i^{'} = \frac{\alpha_i \times C_i}{\|\alpha_i \times C_i\|_2}, \alpha = W^{M'} f_s, \quad (1)$$

where the α_i is the i_{th} of learned scalars α, C_i is i_{th} channel of the feature map of 2D-CNN, f_s is sentence embedding, and $W^{M'}$ is the learned parameters of MLP.

The late guidance module can regard as a channel attention mechanism designed for the outputs of localization networks. Besides, various α_i can modulate individual feature maps in a variety of ways. More importantly, α can offer valuable guidance to the moment retrieval since it is derived from the sentence embedding. Moreover, as we share the same $W^{M'}$ in the moment localization stage, the late guidance module is a computationally efficient method and only involves a handful of parameters.

3. Experiments

To evaluate our approach, we conduct experiments on the YouMakeUp dataset. In this section, we first describe evaluation metrics. Then, we compare the performance of our method with the state-of-the-art models.

3.1. Evaluation Metrics

Following previous work [6], we adopt Rank n@ tIoU = m metric to evaluate moment retrieval method. It is defined as the percentage of sentence queries having at least one correct moment retrieval in the top n retrieved clips. Specifically, we use n $\in \{1, 5\}$ with m $\in \{0.1, 0.3, 0.5, 0.7\}$ for YouMakeUp dataset. Besides, we adopt the common classification accuracy for question answer task.

3.2. Results and Analysis

To improve the performance, we conduct experiments according to the following three directions: sentence pair-wise comparison only using the NLP model, moment retrieval, and combination of multiple moment retrieval models.

First, Following the [6], we adopt the Siamese network structure for textual embedding and binary classification. We adopt a deeper network with more fully-connected layers and ReLU layers. We also utilize different pre-trained word embedding models. However, both tricks do not bring significant gain.

Second, we formulate the step ordering as the moment retrieval task and adopt 2D Temporal Adjacent Network (2D-TAN) [8] as our baseline model for the moment retrieval task. We verify the effect of late guidance module and list the result in Table 1. We can see that the late guidance module boosts the Rank1@IoU0.5 from 37.83% to 41.74%. Besides, our method also outperforms one of the state-of-the-art models, SCDM [7] (34.39%) and SCDM trained with more supervision (37.33%) with the same I3D [1] visual feature[1].

Table 2. Classification accuracy for the step ordering task.

Methods	accuracy
SCDM	69.18
SCDM+	71.72
2D-TAN	68.18
Ours	70.34
Ours and SCDM+	79.56
Ours, SCDM+ and 2D-TAN	83.09
Ours, SCDM+, 2D-TAN, and SCDM	85.19

Third, we combine multiple moment retrieval models to obtain higher classification accuracy. After obtaining the position of target activity from moment retrieval models, we calculate the Levenshtein distance between the predicted order and each candidate. Next, we fuse the distance of multiple models and select the candidate with the shortest distance.

Table 2 shows the accuracy of single model and multiple models. Interestingly, although our model can localize moments more accurately, it obtains a lower accuracy than the SCDM+ model. Beside, fusing multiple models can improve performance significantly. We compute the sum of multiple models' distance and select the candidate with the shortest distance. Specifically, combining 10 models (e.g., 1 SCDM model, 1 SCDM+ model, 4 2D-TAN models, and 4 our models) achieve the 85.19% classification accuracy.

4. Conclusion

In this paper, we introduce our solution to the step ordering task. First, we design the late guidance module to further integrate the videos and sentence tightly. Then, we fuse multiple moment models to achieve higher classification accuracy. In the future, we plan to utilize more powerful visual features to solve the task of moment retrieval.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceed-*

[1] https://github.com/AIM3-RUC/Youmakeup_Baseline

Table 1. Comparison of state-of-the-art methods on YouMakeUp validation dataset using same I3D features.

Methods	Rank1@ IOU0.1(%)	Rank1@ IOU0.3(%)	Rank1@ IOU0.5(%)	Rank1@ IOU0.5(%)	Rank5@ IOU0.1(%)	Rank5@ IOU0.3(%)	Rank5@ IOU0.5(%)	Rank5@ IOU0.3(%)	mIoU
SCDM [7]	57.74	47.53	34.39	18.93	77.02	67.70	52.37	29.47	32.43
SCDM+ [7]	59.86	50.47	37.33	19.91	83.50	76.20	62.17	32.71	34.36
2D-TAN [8]	59.33	51.34	37.83	19.89	82.25	75.09	63.21	34.58	34.34
Ours	**62.01**	**53.52**	**41.74**	**23.11**	**84.24**	**76.44**	**66.06**	**36.98**	**36.92**

ings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6299–6308, 2017.

[2] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *IEEE International Conference on Computer Vision*, pages 5267–5275, 2017.

[3] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. Attentive moment retrieval in videos. In *the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 15–24, 2018.

[4] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, pages 1–8, 2015.

[5] Weining Wang, Yan Huang, and Liang Wang. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 334–343, 2019.

[6] Weiying Wang, Yongcheng Wang, Shizhe Chen, and Qin Jin. Youmakeup: A large-scale domain-specific multimodal dataset for fine-grained semantic comprehension. In *the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5136–5146, 2019.

[7] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In *Advances in Neural Information Processing Systems*, pages 534–544, 2019.

[8] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks formoment localization with natural language. In *AAAI Conference on Artificial Intelligence*, 2020.