Multi-modal Feature Fusion for YouMakeup Video Question Answering Challenge 2020

Yajie Zhang
University of Chinese Academy of Sciences
Beijing, China
zhangyajie19@mails.ucas.ac.cn

Li Su
University of Chinese Academy of Sciences
Beijing, China
suli@ucas.ac.cn

Abstract

YouMakeup VQA Challenge 2020 is introduced for the fine-grained action understanding in makeup videos. This paper presents the details of our models used in the this challenge. There are two tasks, facial image ordering task and step ordering task, and there are both image data and text data. For image ordering task, we propose to use the makeup action descriptions as guidance for better classification the order of each image pairs. Our model achieves 69.067 accuracy score on the testing set which ranks the first place. For step ordering task, we propose to use both sentences and their corresponding images to get a better pairwise comparison model. Our model achieves 73.50 accuracy score on testing set and ranks the second place.

1. Introduction

The large-scale multimodal instructional video dataset, YouMakeup [5], is introduced to support fine-grained semantic comprehension research in makeup domain. Two tasks, facial image ordering and step ordering, are proposed in YouMakeup VQA Challenge 2020 [2]. The task of facial image ordering is to get the correct order of five shuffled facial images according to the ordered step descriptions. And another task, step ordering, is to sort the step descriptions according to the corresponding videos.

As a matter of fact, for each question from both tasks four alternative answers are provided with one of them is the true answer. So after our models predict an answer we will calculate the distance between the predicted answer and the four alternative, and then choose the best alternative as our prediction.

2. Facial Image Ordering Task

In facial image ordering task, there are five images extracted from each videos at different steps and step descriptions, our model needs to sort the five facial images into correct order according to the given step descriptions. The ordering task can be formulated as a classification task of judging whether an image is before another image. The given action described in natural language will cause to the face changes. And the changes of different action descriptions to the face vary greatly.

Table 1. The performance of baseline model and our model, the image pairwise comparison model without curriculum learning [2] is selected as the baseline.

	validate accuracy	test accuracy
baseline model [2]	65.70	67.90
our model	68.58	69.07

2.1. Our Model

To solve this facial image ordering task, our model need to decide the order of some given images. We simplify this task as a classification task of getting the relative order of image pairs. Then, we can get all the relative orders of each pair to construct the predicted order. And, there are also makeup step descriptions of the entire video which can provide additional guidance for our model, so we introduce another branch to utilize the descriptions.

The overall network architecture of our model is shown as Figure 1. The input of the model consists of two facial images (I_i, I_j) which belong to two different makeup steps and step descriptions (S) which is the makeup action descriptions of the corresponding videos.

Imagenet-pretrained resnet-18 [3] is applied as feature extractor for I_i and I_j and BiGRU is applied as feature extractor for S. Then the generated feature embeddings of images and texts are concatenated as input of the binary classifier.

2.2. Experimental Results

Data Preprocessing. Due to the large size of original video dataset, we directly use the provided ResNet-18 [3] embedding for train/dev images and for images in task ques-

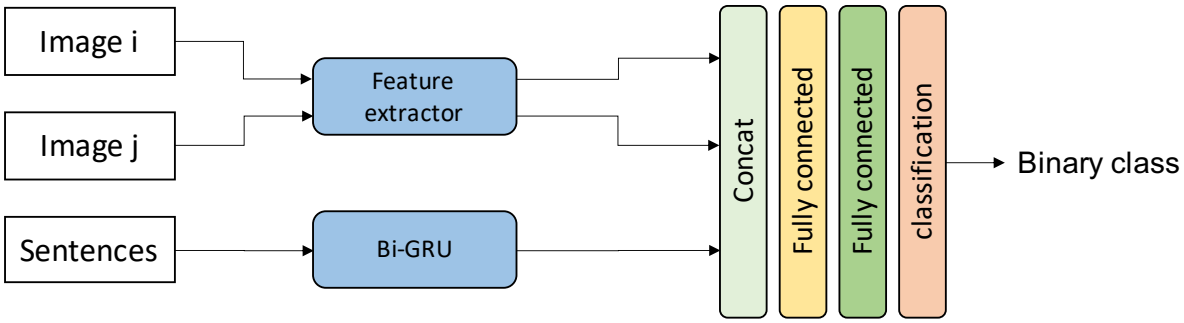

Figure 1. The overview of our image+text classifier for step ordering task.

tions [2] as our input. The YouMakeup VQA Challenge 2020 dataset contains 1680 training videos and 280 validation videos, we extract 10 frames at the end of each clip follow. And finally 177,390 images are generated as our training set.

Results. Accuracy is used in this task to evaluate the models. Table 1 presents the performances of the image pairwise comparison model without curriculum learning [2] and our model. Comparing with the image pairwise comparison model, our model has additional step descriptions guidance. The results show that step descriptions are helpful for this task.

3. Step Ordering Task

Different from image ordering task, the step ordering task is to sort five given step descriptions for a video. The I3D [1] and C3D [4] video features and images extracted from each given steps are provided. Also, we formulated this task as a classification task of judging whether a step is before another step.

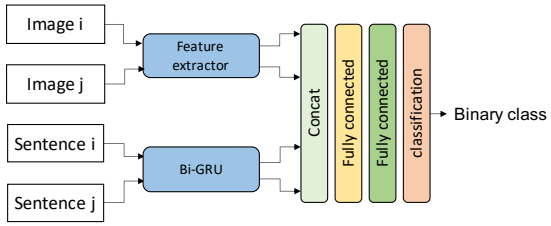

Figure 2. The overview of our image+text classifier for step ordering task.

3.1. Our Model

Our model is required to recognize the relative order of step pairs. For each step we have both image and text features, and there are also video features which can be a global guidance. In our network, we utilize the image and text features to train a binary classifier to get the relative order.

The overall network architecture of our image-text classifier shown as Figure 2. The input of the model consists of two facial images (I_i, I_j) which belong to the two different makeup steps and two corresponding step descriptions (S_i, S_j). Imagenet-pretrained resnet-18 [3] is applied as feature extractor for I_i and I_j and BiGRU is applied as feature extractor for S_i and S_j.

Table 2. The performance of four models, including text classifier, SCDM+, image-text classifier, and our final predicition.

	val acc	test acc
Text Classifier	70.22	69.19
SCDM+	68.41	71.72
Our Image-Text Classifier	70.94	71.37
Our Model	73.33	73.50

Text classifier [2], image-text classifier, SCDM+ [2] are selected for the final ensemble and we use the averaged predictions as the final prediction.

3.2. Experimental Results

Data Preprocessing. We directly use the provided ResNet-18 [3] embedding for train/dev images and we use the provided ResNet-18 [3] to get embeddings for images in task questions. And the provided I3D [1] features are used for training SCDM+ [2].

Results. The results on validation set and testing set are shown in Table 2. The results indicate that image features can improve the performance.

4. Conclusion

In the YouMakeup VQA Challenge 2020, we formulate the two tasks as pairwise comparison tasks and utilize both images and texts to build our models. And, there is still many things can be done for further improvement. First of all, a better feature extractor like ResNet-152 may lead to better performance. Besides, some other features, like audio, may be helpful for the step ordering task. Also, it's crucial to learn the interactions between the texts and images.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. 2017.

[2] Shizhe Chen, Weiying Wang, Ludan Ruan, Linli Yao, and Qin Jin. Youmakeup vqa challenge: Towards fine-grained action understanding in domain-specific videos. 2019.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] Zhuolin Jiang, Viktor Rozgic, and Sancar Adali. Learning spatiotemporal features for infrared action recognition with 3d convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

[5] Weiying Wang, Yongcheng Wang, Shizhe Chen, and Qin Jin. Youmakeup: A large-scale domain-specific multimodal dataset for fine-grained semantic comprehension. 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 5136–5146, 2019.