Video ChatBot: Triggering Live Social Interactions by Automatic Video Commenting

Yehao Li †, Ting Yao ‡, Rui Hu †, Tao Mei ‡, Yong Rui ‡
† Sun Yat-Sen University, Guangzhou, China
‡ Microsoft Research, Beijing, China
yehaoli.sysu@gmail.com; {tiyao, ruhu, tmei, yongrui}@microsoft.com

ABSTRACT

We demonstrate a video chatbot, which can generate human-level emotional comments referring to the videos shared by users and trigger a conversation with users. Our video chatbot performs a large-scale similar video search to find visually similar videos w.r.t. a given video using approximate nearest-neighbor search. Then, the comments associated with the searched similar videos are ranked by learning a deep multi-view embedding space for modeling video content, visual sentiment and textual comments. The top ranked comments are selected as responses to the given video and trigger the succeeding text-based chat between users and the chatbot. The demonstration is conducted on a newly collected dataset with over 102K videos and 10.6M comments. Moreover, our video chatbot has great potential to increase live social interactions.

Keywords

Video Commenting; Multi-View Embedding; Deep Convolutional Neural Networks.

1. INTRODUCTION

The advent of video sharing sites and rapid development of video technologies have led to the unprecedented delivery of online video content. Nowadays, millions of daily users are broadcasting, consuming and communicating via videos. Video has become a predominant media for the growing live social interactions, e.g., Vine, Instagram, and Periscope. Automatic generation of emotional comments to a video has great potential to significantly increase user engagement in many socio-video applications (e.g., chatbot).

A comment is usually a natural sentence that expresses the emotional reaction and feelings after viewing a video. We demonstrate a video chatbot, which is capable of generating comments for videos. Figure 1 shows two use cases of our video chatbot. Specifically, users first share their videos in any social platform. Our video chatbot will then generate comments to the videos for the purpose of social engagement. Take the first video as an example, a comment is presented, e.g., “What a cute and beautiful baby. How old is he?” The generated comment can further trigger the succeeding text-based chat between users and our chatbot, e.g., “He was 2 years old in this video.” Therefore, it is of great potential to increase social interactions and thus enhance user engagement.

Our video chatbot is novel by enabling automatic commenting on videos. To the best of our knowledge, the work represents the first effort towards this target in the multimedia research community. Furthermore, a text-based chatbot is integrated into our video chatbot, which is capable of supporting the succeeding text-based chat with users triggered by video commenting.

2. VIDEO COMMENTING SYSTEM

The key technology of our video chatbot is automatic video commenting. The system overview is given in Figure 2, which is composed of two components: similar video search (VS) and comment dynamic ranking (DR). The former efficiently finds the visually similar videos, while the latter effectively ranks the comments associated with the searched similar videos in a learnt deep multi-view embedding space.
The comment with the highest relevance score will be taken as the final comment in our video commenting system, which can further trigger the succeeding text-based interactions/chat between the user and our chatbot.

## 3. SYSTEM AND USER INTERFACE

Our video chatbot can be accessed through a software operated on a PC client. As shown in Figure 1, an end user first uploads a video to our system and then the video chatbot will generate the comment w.r.t the uploaded video automatically. Based on the comment presented by video chatbot, the user could further chat with our chatbot, which is supported by integrating a real text-based chatbot.

The whole system is currently run on a regular PC (Intel dual-core 3.39GHz CPU and 16GB RAM). Given a five seconds’ video, the time of feature extraction of the video is about 4.3 sec, the similar video search basically completes a search within 0.02 sec, and the dynamic ranking takes about 0.50 sec. Overall, it takes around 4.82 sec to complete the comment generation by our video chatbot, which is less than the duration of the video.

## 4. EVALUATIONS

To evaluate the performance of our video chatbot, we collected 4,000 and 1,000 videos from Vine as validation and testing samples, respectively. Eight evaluators from different education backgrounds are invited to annotate the comments for testing samples generated by our video chatbot. Each comment was annotated on a five-point ordinal scale: 5-Very Good; 4-Good; 3-Normal; 2-Bad; 1-Very Bad. Note that only the comments with their scores more than 3 points are regarded as relevant and satisfying ones for evaluation. The rate of satisfying result for our video chatbot is 54.9%.

Furthermore, to verify the merit of video commenting by search, we further conducted the experiments on this task when it is treated as a problem of sentence generation. Following [1], a LSTM-based model is learnt over all the video-comment pairs in training set. When we apply the model to our testing set, most of the predicted comments to the videos are very general-purpose phrases, e.g., “cute” and “it is amazing.” This is expected, as these phrases often appear in the comments of different videos, making the predicted probabilities of these phrases very high for any input videos. This comparison further validates our proposal of video commenting by search.

## 5. REFERENCES