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.

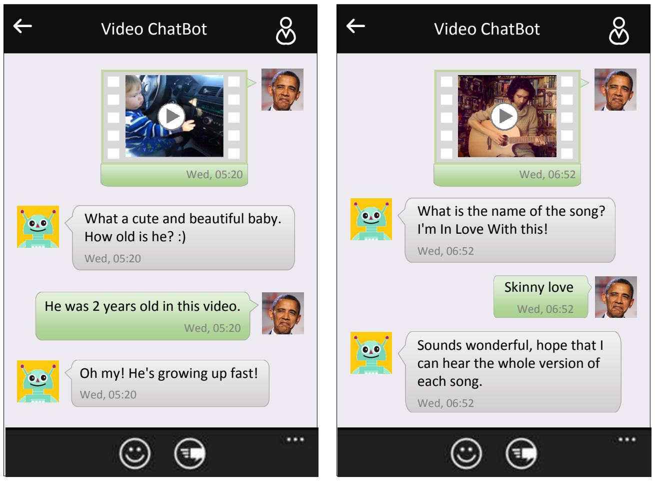
Keywords

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 booming 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." and "Oh my! He's growing up fast!" Therefore, it is of great potential to increase social interactions and thus enhance user engagement.



Wed, 05:20

What a cute and beautiful baby. How old is he? :)
Wed, 05:20

He was 2 years old in this video.
Wed, 05:20

Oh my! He's growing up fast!
Wed, 05:20

Wed, 06:52

What is the name of the song? I'm In Love With this!
Wed, 06:52

Skinny love
Wed, 06:52

Sounds wonderful, hope that I can hear the whole version of each song.
Wed, 06:52

Figure 1: The use cases of our video chatbot.

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.

[*] This work was performed when Yehao Li was visiting Microsoft Research Asia as a research intern.

MM '16 October 15-19, 2016, Amsterdam, Netherlands

ACM ISBN 978-1-4503-3603-1/16/10.
DOI: http://dx.doi.org/10.1145/2964284.2973835

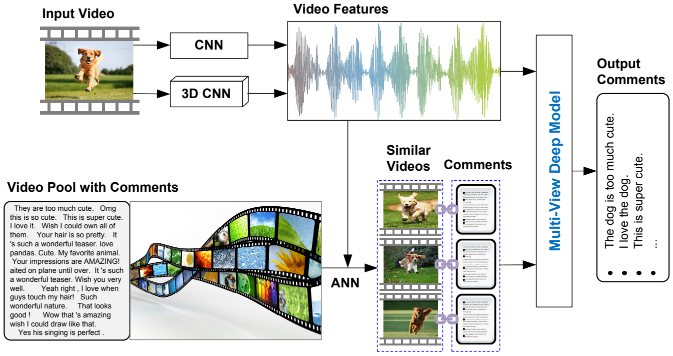

Input Video
CNN
3D CNN
Video Features
Multi-View Deep Model
Output Comments
The dog is too much cute.
I love the dog.
This is super cute.
...
Video Pool with Comments
They are too much cute. Omg this is so cute. This is super cute. I love it. Wish I could own all of them. Your hair is so pretty. It 's such a wonderful teaser. love pandas. Cute. My favorite animal. Your impressions are AMAZING! aited on plane until over. It 's such a wonderful teaser. Wish you very well. Yeah right , I love when guys touch my hair! Such wonderful nature. That looks good ! Wow that 's amazing wish I could draw like that. Yes his singing is perfect .
ANN
Similar Videos
Comments

Figure 2: An architecture overview of video commenting system in our video chatbot.

2.1 Similar Video Search

In this component, given the video uploaded by an end user, the 2-D [2] and/or 3-D [3] convolutional neural networks (CNNs) are utilized to extract visual features of video frames/clips, while deep video representations are produced by mean pooling over these visual features. Next, approximate nearest neighbor (ANN) search is employed to search the visually similar videos from our video pool and the comments associated with these similar videos are regarded as candidates. Specifically, for the purpose of efficiently similar video search with fast computation, we follow [4] and apply the query-driven iterated neighborhood graph search model for ANN search, which contains two stages. A video neighborhood graph on video visual representations is first built and then a query-driven iterated neighborhood graph search approach is exploited to locate the ANNs. Note that our video pool here consists of 96,752 videos with their corresponding human-generated comments collected from Vine[1]. It covers a wide variety of video categories (12 general categories, e.g., animal, entertainment and sports). Hence, with this large-scale video pool, we can leverage the "crowdsourcing" human intelligence to easily collect the comment candidates consisting of real user comments associated with these similar videos.

2.2 Comment Dynamic Ranking

After we obtain the comment candidates in VS stage, how to select the most relevant one in DR stage? As textual comment and video content are of different views, they cannot be directly compared. Inspired by the success of multi-view embedding [5], we utilize our deep multi-view embedding model to learn a latent space where the direct comparison of text comments and videos are allowed. Specifically, our deep multi-view embedding model is learnt based on three views: (1) video content view which depicts objects and temporal dynamics in video, (2) visual sentiment view which represents the visual emotions evoked on video, and (3) textual comment view which expressed the reaction and feelings from users. With the learnt three mappings for the above three views, we can directly calculate the distance between each pair of comment and video, which reflects how relevant the candidate comment could be presented to the given video. Thus, given the uploaded video, a rank list of candidate comments is produced by sorting the relevance scores of video-comment pairs.

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

[1] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui. Jointly modeling embedding and translation to bridge video and language. In *CVPR*, 2016.

[2] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[3] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. C3d: generic features for video analysis. In *ICCV*, 2015.

[4] J. Wang and S. Li. Query-driven iterated neighborhood graph search for large scale indexing. In *ACM MM*, 2012.

[5] T. Yao, T. Mei, and C.-W. Ngo. Learning query and image similarities with ranking canonical correlation analysis. In *ICCV*, 2015.

[1] https://vine.co/