Share-and-Chat: Achieving Human-Level Video Commenting by Search and Multi-View Embedding[*]

Yehao Li [†], Ting Yao [‡], Tao Mei [‡], Hongyang Chao [†], Yong Rui [‡]
[†] Sun Yat-Sen University, Guangzhou, China
[‡] Microsoft Research, Beijing, China
yehaoli.sysu@gmail.com; {tiyao, tmei, yongrui}@microsoft.com; isschhy@mail.sysu.edu.cn

ABSTRACT

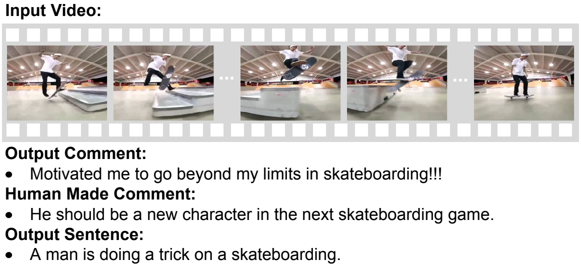
Video has become a predominant social media for the booming live interactions. Automatic generation of emotional comments to a video has great potential to significantly increase user engagement in many socio-video applications (e.g., chat bot). Nevertheless, the problem of video commenting has been overlooked by the research community. The major challenges are that the generated comments are to be not only as natural as those from human beings, but also relevant to the video content. We present in this paper a novel two-stage deep learning-based approach to automatic video commenting. Our approach consists of two components. The first component, similar video search, efficiently finds the visually similar videos w.r.t. a given video using approximate nearest-neighbor search based on the learned deep video representations, while the second dynamic ranking effectively ranks the comments associated with the searched similar videos by learning a deep multi-view embedding space. For modeling the emotional view of videos, we incorporate visual sentiment, video content, and text comments into the learning of the embedding space. On a newly collected dataset with over 102K videos and 10.6M comments, we demonstrate that our approach outperforms several state-of-the-art methods and achieves human-level video commenting.

Keywords

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

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., chat bot). Therefore, video commenting has become a crucial technique yet a challenge for real-world applications.



Output Comment:
- Motivated me to go beyond my limits in skateboarding!!!

Human Made Comment:
- He should be a new character in the next skateboarding game.

Output Sentence:
- A man is doing a trick on a skateboarding.

Figure 1: Examples of video comment and caption (sentence). The input is a short video, while the output of comment is a text response to this video, usually a natural sentence expressing emotional views of the video.

Although there have already been emerging research on video tagging [24][36], video captioning [8][20][33], and visual question answering [1], video comments have their own characteristics and thus are different from tags, captions and answers. A video tag is usually the name of a specific object, action, or event in a video (e.g., "man," "skateboarding" in Figure 1). A caption goes beyond tags by describing a video with a natural sentence, where the basic structure is a triplet of (subject, verb, object). An answer could be a response to a specific question related to a video (e.g., "skateboard" to the question of "what is the man playing with" in Figure 1). Nevertheless, a comment is a text response to a video for the purpose of social engagement. Usually, a comment is a natural sentence with emotional perspective yet relevant to the video content. Figure 1 shows the examples of comment and caption (description) of a video. To our best knowledge, there has been little research on automatic video commenting in both vision and multimedia communities.

The problem of automatic generation of video comments provides challenges to the research communities. The need to understand not only video content but also emotional reaction makes the task very challenging. Moreover, the generated comments should be as natural as those from human beings, in other words, the comments are in the form of natural sentences. Despite the difficulty of this problem, there are two possible solutions: by generation and by

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

MM '16, October 15-19, 2016, Amsterdam, Netherlands
© 2016 ACM. ISBN 978-1-4503-3603-1/16/10. . . $15.00
DOI: http://dx.doi.org/10.1145/2964284.2964320

search. The former attempts to generate a sentence that describes video content, which are mainly inspired by recent advances in machine translation using Recurrent Neural Networks (RNNs) [2][27]. Among these successful works, most of them use Long Short-Term Memory (LSTM) [13], a variant of RNN, which can capture sequential information by mapping sequences to sequences. Nevertheless, this category of approaches, either explicitly or implicitly, assumes the existence of a translatable mapping between the input sequence and the output one. However, video commenting, which expresses an emotional state on a video, often responds with diverse reactions from varying angles. In other words, there are multiple possible outputs with very various statements to a given video, making it difficult to train RNN models. Commenting by search, in contrast, solves this problem by leveraging "crowdsourcing" human intelligence which is the underlying user-generated comments in online interactions. More specifically, the idea is to search visually similar videos on social platform and then exploit the real user comments associated with these similar videos to respond to the given video. A fundamental issue that underlies the success of search is video representation learning, which recently achieves promising progresses thanks to the development of deep convolutional neural networks. Therefore, we follow this elegant recipe to produce video comments by search.

In the direction of video commenting by search, there is also the need to measure the relevance between comment and video. As textual comment and video content are of different views, they cannot be directly compared. Inspired by the success of multi-view embedding, this paper studies the problem by learning a common embedding space that allows direct comparison of text comments and videos.

By consolidating the idea of search and multi-view embedding, we propose a novel framework for video commenting as illustrated in Figure 2, which is composed out of two major components: similar video search (VS) and comment dynamic ranking (DR). Given a video, a 2-D and/or 3-D convolution neural networks is utilized to extract visual features of selected video frames/clips, while the video representations are produced by mean pooling over these visual features. Then, approximate nearest neighbors (ANNs) of the given video are identified from the pool of social videos based on the video representations by using neighborhood graph search. The comments associated with ANNs are regarded as candidates to express the feelings on the given video. Next, a deep multi-view embedding model learnt on three views of video content, visual sentiment and textual comment is exploited to measure the relevance between these candidate comments and the given video. Finally, the comments are ranked by sorting their relevance scores and the top-K comments will be selected as responses to the video.

In summary, this paper makes the following contributions:

- We study the problem of automatic video commenting to express emotional response to a video. It has great potential for increasing social interaction and enhancing user engagement in many socio-video applications. To the best of our knowledge, this paper represents the first effort towards this target in the multimedia research community.

- We propose a novel two-stage framework for achieving human-level video commenting. The framework is grounded on the similar video search paradigm by using approximate nearest-neighbor search. Additionally, the framework also tackles the problem of relevance measurement between textual comment and video by learning an embedding space between three views of video content, visual sentiment and textual comment.

- The proposed methods are evaluated on a newly created dataset including about 102K videos and 10.6 million comments crawled from Vine, which is so far the first dataset for video commenting.

The remaining sections are organized as follows. Section 2 describes the related work. Section 3 presents our proposed two-stage approach including similar video search and comment dynamic ranking, while Section 4 introduces our newly created dataset for video commenting. In Section 5, we provide empirical evaluations, followed by the discussions and conclusions in Section 6.

2. RELATED WORK

We briefly group the related works into two categories: video tagging and vision to sentence. The former draws upon research in automatically assigning tags (annotations) to videos, and the later mainly focuses on sentence generation of visual content.

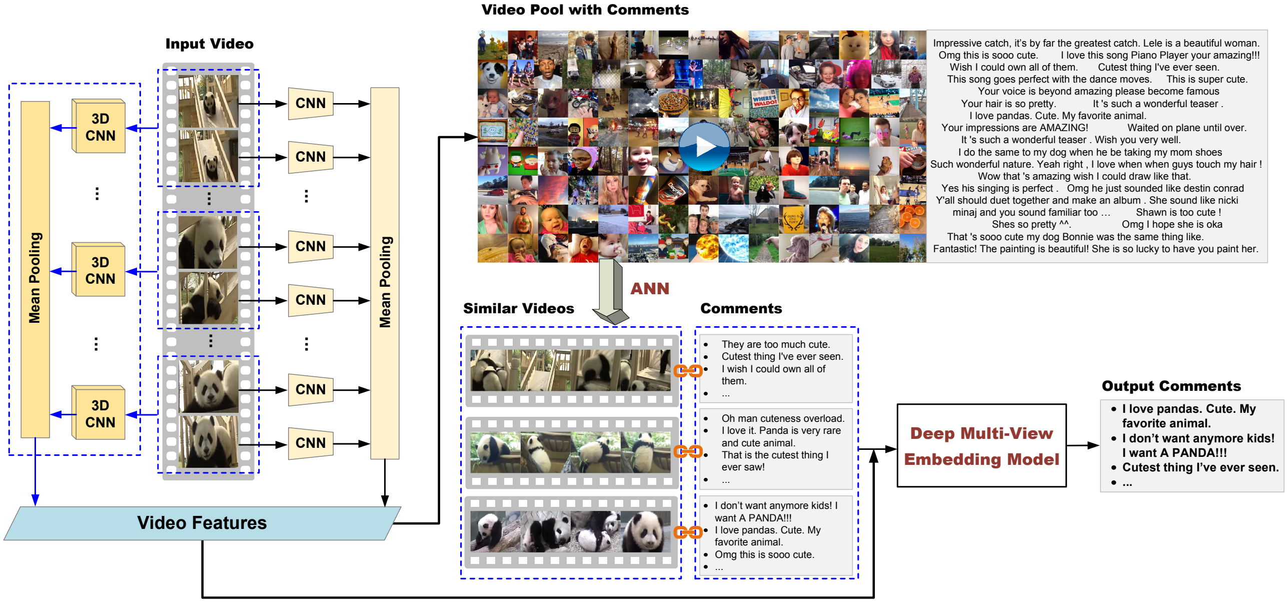
2.1 Video Tagging

Video tagging problem has received intensive research attention and the significance of the topic can be partly reflected from the huge volume of published papers. The research in this direction has proceeded along two different dimensions: model-based methods [6][21][37] and data driven approaches [24][32][36].

Model-based methods assume that a training set of videos along with keyword annotations is provided for developing concept classifiers. Cristianini *et al.* employed SVM with one-against-the-other strategy to learn a set of detectors, each of which independently models the presence/absence of a certain concept in [6]. Later in [21], Qi *et al.* proposed correlative multi-label method to simultaneously model both the individual concepts and their correlations in a unifying formulation and the principle of least commitment was obeyed. Recently, EventNet [37], a large scale structural concept library, is proposed for video tagging by utilizing large margin concept classifiers trained with deep representations on YouTube videos.

Different from model-based methods, data-driven approaches construct video similarity for annotation. In [24], the overlapping or duplicated content of videos was exploited for measuring video similarity. With this, tags associated with similar videos were exchanged for generating new tag assignments. In another work by Wang [32], both distance between samples and the difference of their surrounding neighborhood sample distributions were taken into account for video similarity estimation. Recently, Yao *et al.* [36] combined click-through and video document features for deriving a latent subspace, in which the dot product of the mapping can be taken as the video similarity. The learnt video similarity is then applied for video tagging tasks.

In short, the target of video tagging is to annotate video content by individual semantic tags (words). Our work in this paper contributes by not only extending the natural



Impressive catch, it’s by far the greatest catch. Lele is a beautiful woman.
Omg this is sooo cute. I love this song Piano Player your amazing!!!
Wish I could own all of them. Cutest thing I've ever seen.
This song goes perfect with the dance moves. This is super cute.
Your voice is beyond amazing please become famous
Your hair is so pretty. It 's such a wonderful teaser .
I love pandas. Cute. My favorite animal.
Your impressions are AMAZING! Waited on plane until over.
It 's such a wonderful teaser . Wish you very well.
I do the same to my dog when he be taking my mom shoes
Such wonderful nature. Yeah right , I love when when guys touch my hair !
Wow that 's amazing wish I could draw like that.
Yes his singing is perfect . Omg he just sounded like destin conrad
Y'all should duet together and make an album . She sound like nicki
minaj and you sound familiar too … Shawn is too cute !
Shes so pretty ^^. Omg I hope she is oka
That 's sooo cute my dog Bonnie was the same thing like.
Fantastic! The painting is beautiful! She is so lucky to have you paint her.

- They are too much cute.
- Cutest thing I've ever seen.
- I wish I could own all of them.
- ...

- Oh man cuteness overload.
- I love it. Panda is very rare and cute animal.
- That is the cutest thing I ever saw!
- ...

- I don’t want anymore kids! I want A PANDA!!!
- I love pandas. Cute. My favorite animal.
- Omg this is sooo cute.
- ...

- **I love pandas. Cute. My favorite animal.**
- **I don’t want anymore kids! I want A PANDA!!!**
- **Cutest thing I’ve ever seen.**
- **...**

Figure 2: An overview of our video commenting framework mainly including similar video search (VS) and comment dynamic ranking (DR) (better viewed in color). In VS, the video representations are first produced by mean pooling over the visual features of frames/clips, extracted by a 2-D/3-D CNN. Next, approximate nearest neighbor (ANN) search is employed to search the visually similar videos in our video pool and the comments associated with these similar videos are regarded as candidates. In DR, a deep multi-view embedding model is exploited to measure the relevance between comment candidates and the input video. The comments with highest relevance scores are finally presented to the input video.

descriptions from individual tags to complete sentences, but also experiencing the emotions and feelings on videos.

2.2 Vision to Sentence

Image to sentence. For image sentence generation task, there have been several techniques [7][9], being proposed to generate image captions through utilizing captions from visually similar images. Farhadi *et al.* [9] annotated images with sentences by using nearest neighbors in an intermediate meaning space. Later in [7], Devlin *et al.* provided a detailed exploration on nearest neighbors approaches based on different representations and caption selection through different similarity functions. In another work by Chen *et al.* [5], viewer affect after viewing an image is predicted based on sentiment concepts defined in SentiBank [3]. An alternative scheme on image to sentence was recently proposed in [8][30], which utilize RNN to generate more flexible captions inspired by recent advances in machine translation. The basic idea is to model the probability of generating a word given previous words and image in RNN architecture.

Video to sentence. There are also two major directions for translation from videos. One is template-based methods [11][28], which first identify semantic content (subject, verb, object) in the video and then produce sentences on the predefined templates. This direction is highly depended on the sentence templates and can only generate sentences with syntactical structure. Similar to recent breakthroughs in image to sentence, another direction on video to sentence is to utilize RNN model to generate sentences for videos in [8][20][33].

The aforementioned vision to sentence approaches focus on generating sentence, which describes video content objectively. Our work is different in the way that we do not explicitly describe video content, but instead express the emotional reaction and feelings after viewing the video like a human being.

3. VIDEO COMMENTING APPROACH

The basic idea of this work is to automatic comment on videos by searching visually similar videos and ranking the comments associated with the searched similar videos. The similar video search is performed by using ANN search on powerful spatio-temporal video representations learnt by 2-D and 3-D CNN, while dynamic ranking ranks the candidate comments by learning a deep multi-view embedding model. In this way, the original incomparable views of video content, visual sentiment and textual comment could be directly compared in this embedding space. Moreover, the learning of the embedding space is integrated into an optimization to better reflect the preference relations in the training triplets. The approach overview is demonstrated in Figure 2.

In this section, we first present our adopted ANN search in VS stage, followed by three views for video commenting and our proposed deep multi-view embedding model in DR.

3.1 Notation

Let $\mathcal{V}$ and $\mathcal{C}$ denote the collections of videos and comments respectively. For each video $v_k \in \mathcal{V}$, its n associated comments are denoted as $\{c_{k1}, c_{k2}, ..., c_{kn}\} \in \mathcal{C}$. The goal of our deep multi-view embedding model is to construct three mappings $f_v : v \to \mathbb{R}^d$, $f_s : v \to \mathbb{R}^d$ and $f_c : c \to \mathbb{R}^d$ for video content view, visual sentiment view and textual comment view, such that all the three views can be mapped into a d-

Algorithm 1 Query-driven iterated neighborhood graph search

```
$P_0 \leftarrow$ GenerateInitialSolution($q$, $Tr$)
$R^* \leftarrow$ LocalNGSearch($P_0$, $G$)
repeat
    $P^{'} \leftarrow$ Perturbation($R^*$, $q$, $Tr$, $history$)
    $R^{*^{'}} \leftarrow$ LocalNGSearch($P^{'}$, $G$, $history$)
    $R^* \leftarrow$ AcceptanceCriterion($R^*$, $R^{*^{'}}$)
until termination condition met
```

dimensional embedding space, which are represented as F_1, F_2 and F_3, respectively.

3.2 ANN Search Model

Nearest neighbor (NN) search has been a fundamental research topic and widely applied in computer vision, machine learning, and information retrieval. Given a query q, one straightforward solution to the NN search is linear scan. Unfortunately, such NN search is often impractical for the large-scale high-dimensional database. To deal with the problem, we follow [31] and apply the query-driven iterated neighborhood graph search model for ANN search.

The ANN search model contains two stages. A video neighborhood graph on video visual representations is firstly built and then a query-driven iterated neighborhood graph search approach is exploited to locate the ANNs. The basic procedure is outlined in Algorithm 1.

GenerateInitialSolution(q, Tr) searches over trees Tr, which are constructed to index the reference videos. The initial solution contains a small amount of initial NN candidates that have high probabilities to be near true NNs. Following the implementation in [31], we use kd-trees in our experiments. LocalNGSearch(P_0, G) starts from a set of seeds P_0 and searches over G by conducting neighborhood expansions in a best-first manner. Perturbation(R^*, q, Tr, $history$) generates new seeds from trees Tr according to the search history and previously selected NNs (R^*), to avoid unnecessary neighborhood expansions. LocalNGSearch($P^{'}$, G, $history$) is slightly different from LocalNGSearch(P_0, G) as the search history, i.e., the NNs discovered up to the current iteration are considered in neighborhood expansion. Readers can refer to [31] for technical details.

3.3 Three Views for Video Commenting

Video can be naturally decomposed into spatial, temporal and emotional components, which are related to ventral, dorsal and limbic system for human perception respectively [16][25]. The ventral and dorsal systems play the major roles in the recognition of objects and visually guided actions which are both focusing on the video content, while the limbic view mainly supports for the neurological regulation of emotion about the visual sentiment to the given video. In addition, human-generated textual comment on shared video can be regarded as another view in natural language for video. Therefore, we input all the three views of video content, visual sentiment and textual comment into a following deep multi-view embedding model for video commenting, as shown in Figure 3. The video content view depicts objects and temporal dynamics by frame appearance and video clip consisting of multiple frames, while the visual sentiment view conveys the emotion of the whole video and the textual comment view expresses the observation and feelings from users' viewpoint.

Video content view. The video content view consists of both spatial and temporal structures underlying video. For the spatial component, VGG-19 [26], which is the recent superior 2-D CNN architecture for image classification, is exploited for extracting frame-level spatial representation. The VGG-19 is pre-trained on 1.2 million images of ImageNet challenge dataset [23] and its 4,096-way fc6 layer output is taken as frame representation. Then, mean pooling is performed over all the sampled frames to get a 4,096 dimensional video-level spatial representation (VGG). For the temporal component, to capture the temporal dynamics between multiple frames, 3-D CNN is utilized for each video clip consisting of continuous frames. Different from traditional 2-D CNN, 3-D CNN architecture takes video clip as the input and consists of alternating 3-D convolutional and 3-D pooling layers, which are further topped by a few fully connected layers. Specifically, C3D [29], which is pre-trained on Sports-1M video dataset [15], is used and we regard the output of the 4,096-way fc6 fully-connected layer as representation for each video clip. Similar to spatial component, we adopt mean pooling over all the C3D outputs of sampled clips to generate video-level temporal representation (C3D). The final representation for video content view is the concatenation of C3D and VGG (C3D+VGG) which models both spatial and temporal information in video.

Visual sentiment view. To represent the visual emotions evoked on video, we use ANP-Net [4], which is a CNN architecture pre-trained with 867,919 images from Visual Sentiment Ontology [3] to classify images into 2,089 Adjective Noun Pairs (ANPs). The 2,089-dimensional softmax scores of ANP-Net is adopted as the visual sentiment representation for each frame. Then, max pooling is performed over all the visual sentiment representations of sampled frames to generate the video-level representation from visual sentiment view.

Textual comment view. For the representations of textual comment, we utilize two popular methods: (1) Term Frequency (tf). Each comment is represented as a vector of words, and each word is weighed by term frequency. (2) Word2vec ($w2v$). We represent each word with 300-dimensional vector generated from word2vector neural network [17] pre-trained on Google News dataset (about 100 billion words) and then the sentence-level representation for textual comment view is generated by mean pooling over all the word vectors in the comment.

It is worth noticing that although the three views used here are video content, visual sentiment and textual comment, our approach is applicable to include any other view, e.g., audio view.

3.4 Deep Multi-View Embedding Model

To learn the whole architecture of our deep multi-view embedding, the objective is to maximize the correlations among video content, visual sentiment and textual comment. One natural choice is to minimize the distances between each pair of views in the embedding space. However, without any carefully designed constraints, the optimization will easily result in meaningless solution that all three views map to the identical representation in the embedding space.

Deriving from the idea of exploring relative relationship through ranking [19][34][35], we develop a triplet deep rank-

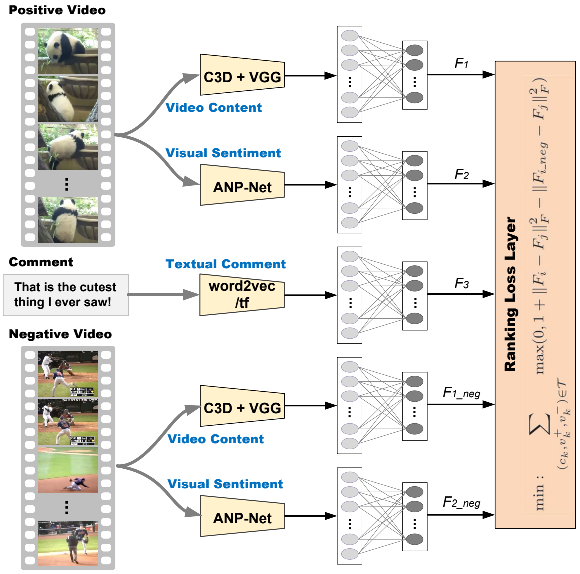


Figure 3: The training of deep multi-view embedding model. The inputs are a set of triplets: textual comment, positive video and negative video. For both videos, video representations are learnt by C3D plus VGG architectures while visual sentiment is modeled by ANP-Net, followed by an embedding layer for each of the two views, respectively. Note that the embedding layers of positive and negative videos share parameters. For textual comment, word2vec ($w2v$) or term frequency (tf) is extracted and then feed into a textual embedding layer. A loss layer is on the top to evaluate the ranking loss of the triplet.

ing model to learn our deep multi-view embedding model (DE) for dynamic ranking. Figure 3 shows the training of DE architecture with triplet deep ranking model. Given a triplet of textual comment, positive video which is associated with the textual comment and negative video which is irrelevant to the textual comment, we aim to optimize our DE architecture, which makes the mapping of textual comment closer to positive video than negative video in the embedding space. Formally, suppose we have a set of triplets $\mathcal{T}$, where each triplet $(c_i, {v_i}^+, {v_i}^-)$ consists of a textual comment c_i, a positive video ${v_i}^+$ and a negative video ${v_i}^-$. Note that c_i is selected from textual comments associated with positive video ${v_i}^+$ while the negative video ${v_i}^-$ is randomly sampled from a different category and does not contain comment c_i. The positive and negative videos in the triplet are fed separately into two identical embedding layers with shared architecture and parameters. Our goal is to learn the DE architecture that makes one view F_j in the embedding space is closer to another view F_i of positive video than that view F_{i_neg} of negative video, which can be expressed as

$$\begin{aligned} &\|F_i - F_j\|_F^2 \prec \|F_{i\_neg} - F_j\|_F^2, \quad \forall (c_k, v_k^+, v_k^-) \in \mathcal{T} \\ &s.t. \quad i, j = 1, ..., 3, \quad i \neq j, \quad i \neq 3 \; . \end{aligned} \tag{1}$$

As the distances between two views in embedding space exhibit a relative ranking order, a ranking loss layer on the top is employed to evaluate the margin ranking loss of each triplet, which is a convex approximation to the 0-1 ranking error loss and has been used in several information retrieval methods. Specifically, it can be given by

$$\begin{aligned} \min: \sum_{(c_k, v_k^+, v_k^-) \in \mathcal{T}} & \max(0, 1 + \|F_i - F_j\|_F^2 - \|F_{i\_neg} - F_j\|_F^2) \\ s.t. \quad & i, j = 1, ..., 3, \quad i \neq j, \quad i \neq 3 \; . \end{aligned} \tag{2}$$

The ranking loss layer does not have any parameters. During learning, it evaluates the model's violation of the ranking order, and back-propagates the gradients to the lower layers so that the lower layers can adjust their parameters to minimize the ranking loss. Please also note that each view is equipped with a hyperbolic tangent function plus a L_2 normalization before fed into the ranking loss layer.

After optimization, we can obtain the mappings f_v, f_s and f_c for the three views, respectively. Next, given a test video and candidate comment pair $(\hat{v}, \hat{c})$, we compute the distance value between the pair as

$$r(\hat{v}, \hat{c}) = \|F_1(\hat{v}) - F_3(\hat{c})\|_F^2 + \|F_2(\hat{v}) - F_3(\hat{c})\|_F^2 \, . \tag{3}$$

This value reflects how relevant the candidate comment could be presented for the given video, with lower numbers indicating higher relevance. Thus, given a test video, a rank list of candidate comments is produced by sorting the values of video-comment pairs.

4. VIDEO COMMENTING DATASET

To the best of our knowledge, our work is the first attempt to automatically generate comments for videos. To substantially evaluate our approach, we collect a new dataset for video commenting. The dataset is characterized by the unique properties including the large scale video-comment pairs, comprehensive video categories, diverse video content and comments. We next describe how we collect representative videos, select high-quality comments, and split the dataset.

4.1 Collection of Videos

Our dataset focuses on general videos in our life and is collected from Vine[1]. Vine is the entertainment network where the world's stories are captured, created, remixed and shared. The users can follow your favorite creators, watch their stories and post your comments as well.

To collect representative videos, we obtain 2,278 popular tags from TagsForLikes website[2], corresponding to 12 categories[3]. Then we crawl the top 100 video search results for each tag and collect at most 500 comments for each video. We remove duplication, as well as the videos with low-quality comments, to maintain the data quality. As a result, we have 101,752 representative videos. All the videos were downloaded with high quality and audio channel.

4.2 Selection of High Quality Comments

The original comments associated with each video contain too much noise, e.g., random responses, bad words and negative induced emotion. In order to make the comments in our dataset clean and high quality, we select the comments from three aspects: 1) length, 2) relevance, and 3) sentiment. In between, the length of comment is generally an important

[1] https://vine.co/

[2] http://vine.tagsforlikes.com/

[3] The categories include animal, art, celebration, entertainment, family, fashion, food, sports, urban, vehicle, weather, other.

factor based on the observation that very short comments are always general-purpose while long comments are often too specific. Furthermore, the use of relevance measurement between comment and video is to filter out the random responses. As the comments are posted as social interactions, those with very negative sentiment should be also removed. In addition, the comments containing bad or dirty words are removed directly from our dataset. Specifically, the quality measurement on each criterion is as below.

Length. Let *min_len* and *max_len* denote the expected minimal and maximal length of comment, respectively. Denote *length* as the length of comment after removing punctuation and emoji. Then, the quality score of comment in terms of length is defined as

$$\begin{aligned} S_{len} = \omega_1 * max((min\_len - length), 0) \\ + \omega_2 * max((length - max\_len), 0) \end{aligned} \quad , \tag{4}$$

where ω_1 and ω_2 are tradeoff parameters between punishing very short and too long comments.

Relevance. We estimate the relevance between video and comment by measuring the semantic similarity between the tag submitted for searching this video and words in the comment. Specifically, each comment c is represented as a sequence of words. Words are stemmed and stop words are removed. Then, the similarity $sim(tag, w_i)$ between the tag and each word w_i in the comment is computed by WordNet [18]. Finally, the quality score of the comment on relevance is given by

$$S_{rel} = \max_{w_i \in c} sim(tag, w_i) \quad . \tag{5}$$

Sentiment. Following [14], we apply VADER for comment sentiment analysis. VADER (Valence Aware Dictionary for Sentiment Reasoning) is a rule-based model for general-purpose sentiment analysis, which is shown to be powerful and outperforms individual human raters on some specific social media text, e.g., tweets. For each comment c, the sentiment score $senti(c)$ is taken as its quality score and is formally computed as

$$S_{senti} = senti(c) \quad . \tag{6}$$

The final quality score of each comment is obtained by linearly fusing the three scores as

$$S_{qua} = \alpha_1 * S_{rel} + \alpha_2 * S_{senti} - \alpha_3 * S_{len} \quad , \tag{7}$$

where α_1, α_2 and α_3 are tradeoff parameters, which weight the importance of the three criteria. If the final quality score is smaller than a threshold, the comment will be regarded as a low-quality one and thus removed from our dataset. After selection, there are about 10.6 millions of comments.

4.3 Statistics and Dataset Splits

Our dataset consists of 185 hours videos and 10.6 millions comments (corresponding to 90 millions words and one million unique words). It is derived from a wide variety of video categories (101,752 videos from 12 general categories) and thus this can benefit the generalization capability of model learning on this dataset. Figure 4 further details the category distribution of all the videos. Moreover, there are more than 100 comments for each video on average, leading to a better chance of finding more natural and diverse comments. In summary, we present a comprehensive, diverse, and complex dataset for video commenting.

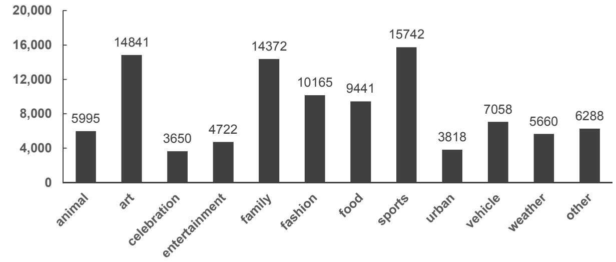


Figure 4: The distribution of video categories in our dataset. This distribution well aligns with the real data statistics in a social video platform.

To split the dataset to training, validation and testing set, we separate the videos according to the corresponding categories. The videos from the same category will appear in all the three sets to avoid overfitting. Finally, we sample 96,752, 4,000, 1,000 videos in the training, validation, and testing set, respectively.

5. EXPERIMENTS

We conducted our experiments on the aforementioned newly crawled video commenting data and evaluated our proposed framework on video commenting task.

5.1 Experimental Settings

Compared Approaches: We compare the following approaches for performance evaluation:

- Random Selection (RS). We search visually similar videos of the input video by utilizing ANN search, and then randomly select the comments from all the associated comments of these searched similar videos as the final output comments. No comment ranking model is learned in this baseline. We name this run as *RS*.

- Canonical Correlation Analysis (CCA) [10][12]. CCA is a popular and successful approach for mapping two or more input views into a latent embedding space where the correlation between different views are maximized. We employ the classical CCA method in DR stage to build a multi-view embedding space based on video content (V), textual comment (T) and/or visual sentiment (S) to measure the relevance between textual comment and video. Two slightly different variants of this run are named as *CCA-VT* and *CCA-VST*, which rank the candidate comments for the input video in the learnt two-view (V and T) embedding space and three-view (V, S and T) embedding space, respectively.

- Deep Multi-view Embedding model (DE). We design two runs, *DE-VT* and *DE-VST*, for our proposed two-stage framework. For the dynamic ranking of candidate comments, *DE-VT* learns the embedding space based on two views of video content and textual comment (V and T), while *DE-VST* exploits all the three views by additionally incorporating the view of visual sentiment (V, S and T) to learn the embedding space.

Parameter Settings: In the VS stage, the k nearest neighbors for ANN search is chosen within {1, 5, 10, 15, 20}. For each nearest neighbor video, we randomly sample at most 50 associated comments to generate the candidate comments. In the next stage, for each input view, the sample rate is set as 1 frame/clip per second. Specifically, for the tf representation in textual comment view, we use the top 35,000 most frequent words as the word vocabulary. Moreover, for the four embedding models, the dimensionality of the embedding space is in the range of {32, 64, 128, 256}. Please note that for each compared method, top 5 comments are finally presented to the input video. We set the parameters in high quality comment selection as $\mathit{min\_len} = 3$, $\mathit{max\_len} = 30$, $\omega_1 = 10$, $\omega_2 = 1$, $\alpha_1 = 0.5$, $\alpha_2 = 0.3$ and $\alpha_3 = 0.01$ empirically.

Ground Truth: To facilitate the evaluation and comparison with other approaches, we randomly selected 1,000 videos as testing samples. For each testing video, the top 5 comments obtained by different methods are annotated during the final evaluation. Eight evaluators from different education backgrounds, including linguistics, business, information management, international news, electrical engineering, and computer science are invited. All evaluators are familiar with video sharing sites. 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 ground truth relevant ones for evaluation. The evaluators are requested to watch the whole video before evaluating comments.

Evaluation Metrics: For the evaluation of video commenting, we adopted average precision (AP) as the performance metric and mean Average Precision (mAP) over all the videos in testing set is finally reported. Given a comment ranked list, the AP score at the depth of m in the ranked list is defined by:

$$AP@m = \frac{1}{m}\sum\nolimits_{j=1}^{m}\frac{R_j}{j} \times I_j, \tag{8}$$

where R_j is the total number of relevant comments in the top-j list, I_j is set as 1 if the jth comment is relevant and 0 otherwise.

5.2 Performance Comparison

Table 1 shows the mAP performances of five runs averaged over 1,000 test videos. It is worth noting that *RS* selects comments randomly from candidates without learning any ranking model and for other four runs, the performances are given by choosing 64 as the dimensionality of the latent embedding space and taking tf as comment representations.

Overall, our proposed *DE-VST* consistently outperforms the other runs across different depths of mAP. In particular, the mAP@1 of *DE-VST* can achieve 0.549, making the relative improvements over *CCA-VST* by 9.6%. Though both *DE-VST* and *CCA-VST* involve utilization of multi-view embedding, different strategies are used for learning the embedding space. *CCA-VST* aims to learn an embedding space which maximizes the correlations between the heterogeneous representations across three original views, while *DE-VST* additionally incorporates the preference relations in the training triplets to further adjust the embedding space. The results indicate that our *DE-VST* is benefited from the utilization of relative preference, and capable of finding more desired comments in response to the video. *RS*, which even selects the comments randomly from the comments associated with visually similar videos, still shows encouraging performance. This somewhat reveals the effectiveness of ANN search on our powerful spatio-temporal video representations. Compared to *DE-VT* (*CCA-VT*), *DE-VST* (*CCA-VST*) exhibits better performance by considering visual sentiment as another view of video representations for multi-view embedding learning. This is expected, as for video commenting the emotions or feelings on videos are highly related to visual sentiment reflected in videos.

Table 1: Performance comparison of different approaches for video commenting.

App.	mAP@1	mAP@2	mAP@3	mAP@4	mAP@5
RS	0.259	0.244	0.219	0.203	0.191
CCA-VT[12]	0.458	0.421	0.399	0.389	0.382
CCA-VST[10]	0.501	0.465	0.439	0.429	0.419
DE-VT	0.504	0.469	0.447	0.433	0.422
DE-VST	0.549	0.513	0.486	0.471	0.459

To verify that the performance of different approaches is not by chance, we conducted significance test using the randomization test [22]. The number of iterations used in the randomization is 100,000 and at 0.05 significance level. *DE-VST* is found to be significantly better than others.

Figure 5 further details the mAP@1 performance for all the 12 categories. Among all the categories, *DE-VST* achieves the best performance for 10 categories, followed by *DE-VT* and *CCA-VST* for one category, respectively. The improvements can be generally expected by further taking the view of visual sentiment into account in both CCA and our deep multi-view embedding model learning on all the categories except "food" category. That is not surprise because the perceived sentiment or emotion signals from videos in "food" category are often neutral and thus not easy to be predicted precisely.

Figure 6 shows a few comment examples produced by different approaches. From these exemplar results, it is easy to see that all of these automatic methods can find somewhat relevant comments. With embedding model, the ranks of more precise comments are likely to be moved up and responded to the videos. Moreover, our *DE-VST* can offer more accurate emotional comments. For instance, the returned comment "wow this is amazing! I love to draw and to photograph." to the first video experiences the video content more desirably.

5.3 Search vs Generation

To empirically verify the merit of video commenting by search, we further conducted the experiments to examine how it works when the task is treated as a problem of sentence generation. Following [8], an end-to-end 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," "cool," and "it is amazing." This is arisen from the fact that these phrases often appear in the comments of different videos, making the predicted probabilities of these phrases very high for any input videos. On the other hand, for each video, the comments exploited in training are very diverse with various subjective descriptions. Therefore, it is very difficult to establish a translatable mapping from video to comment for generation. This comparison basically validates our analysis and proposal of video commenting by search.

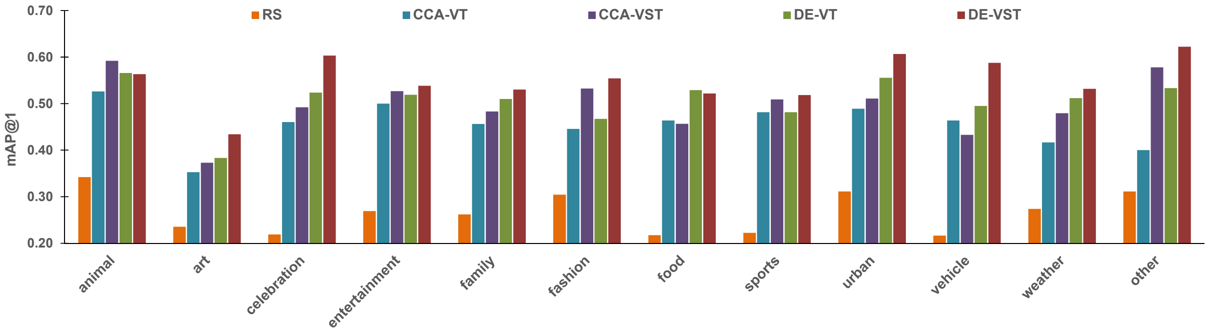
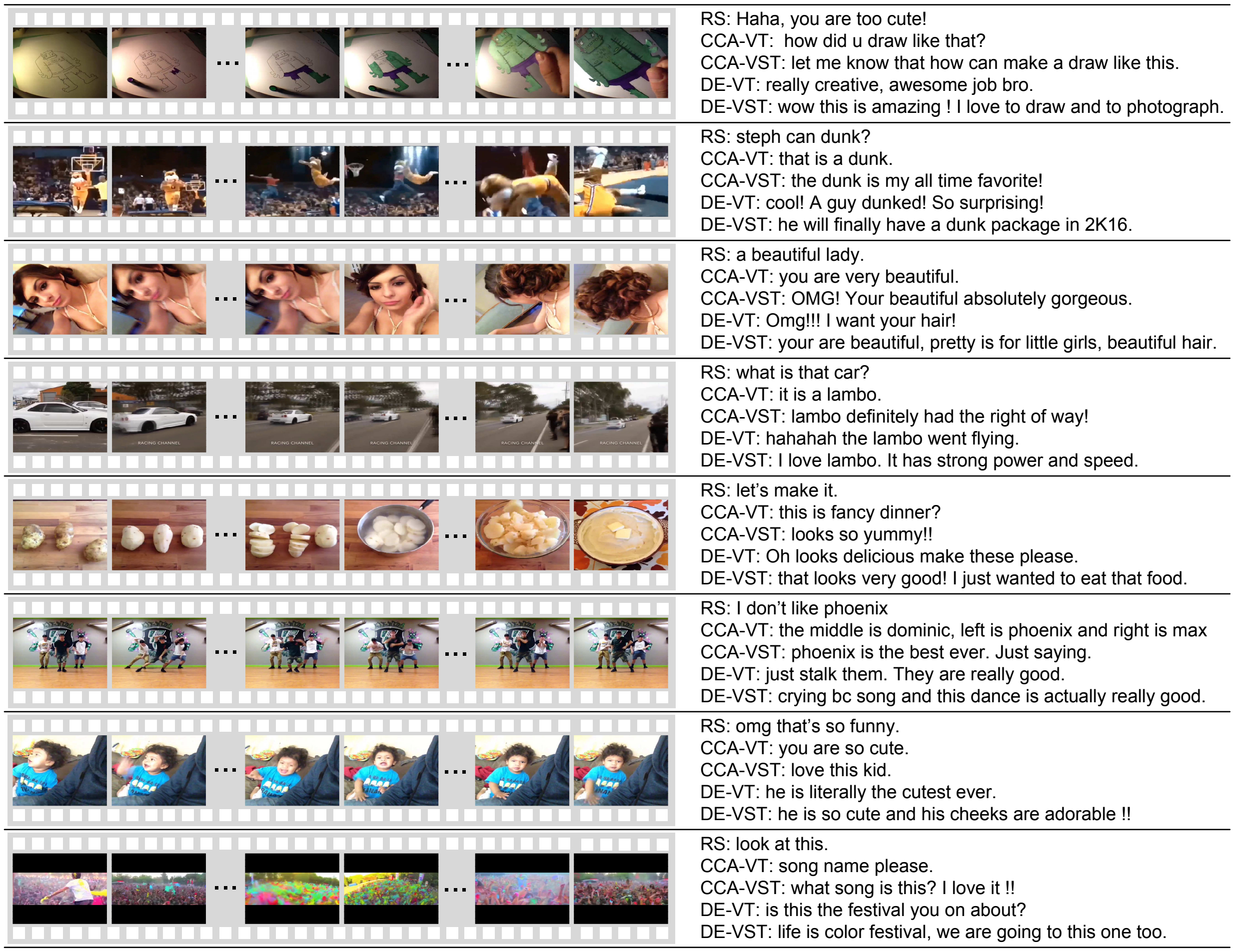


Figure 5: Per-category mAP@1 of different approaches for all the 12 categories.



RS: Haha, you are too cute!
CCA-VT: how did u draw like that?
CCA-VST: let me know that how can make a draw like this.
DE-VT: really creative, awesome job bro.
DE-VST: wow this is amazing ! I love to draw and to photograph.

RS: steph can dunk?
CCA-VT: that is a dunk.
CCA-VST: the dunk is my all time favorite!
DE-VT: cool! A guy dunked! So surprising!
DE-VST: he will finally have a dunk package in 2K16.

RS: a beautiful lady.
CCA-VT: you are very beautiful.
CCA-VST: OMG! Your beautiful absolutely gorgeous.
DE-VT: Omg!!! I want your hair!
DE-VST: your are beautiful, pretty is for little girls, beautiful hair.

RS: what is that car?
CCA-VT: it is a lambo.
CCA-VST: lambo definitely had the right of way!
DE-VT: hahahah the lambo went flying.
DE-VST: I love lambo. It has strong power and speed.

RS: let's make it.
CCA-VT: this is fancy dinner?
CCA-VST: looks so yummy!!
DE-VT: Oh looks delicious make these please.
DE-VST: that looks very good! I just wanted to eat that food.

RS: I don't like phoenix
CCA-VT: the middle is dominic, left is phoenix and right is max
CCA-VST: phoenix is the best ever. Just saying.
DE-VT: just stalk them. They are really good.
DE-VST: crying bc song and this dance is actually really good.

RS: omg that's so funny.
CCA-VT: you are so cute.
CCA-VST: love this kid.
DE-VT: he is literally the cutest ever.
DE-VST: he is so cute and his cheeks are adorable !!

RS: look at this.
CCA-VT: song name please.
CCA-VST: what song is this? I love it !!
DE-VT: is this the festival you on about?
DE-VST: life is color festival, we are going to this one too.

Figure 6: Examples of top-1 ranked comment on each video produced by different approaches. The videos are represented by sampled frames.

5.4 Comparison on Different Representations

We designed six runs to compare the performances of our multi-view embedding model as a result of exploiting different video and comment representations, while the representations of visual sentiment are always extracted by ANP-Net. The results are shown in Figure 7.

There is a performance gap between using $w2v$ and tf representations. Though both representations are learnt on a large corpus, $w2v$ is predicted by Skip-gram models which are pre-trained on news articles, and tf vector of each comment is weighted by term frequency of each word in the vocabulary built on our created video commenting dataset. As indicated by our results, tf representations can constantly lead to better performance. We speculate that this may be caused by the big difference of wording between news (formal) and comments (social). Moreover, video representations learnt by C3D which has a better ability in encapsulating temporal information leads to better performance

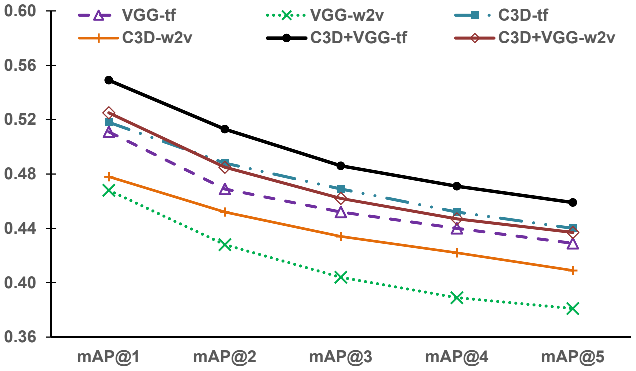


Figure 7: Performance comparison on different representations of video and comment.

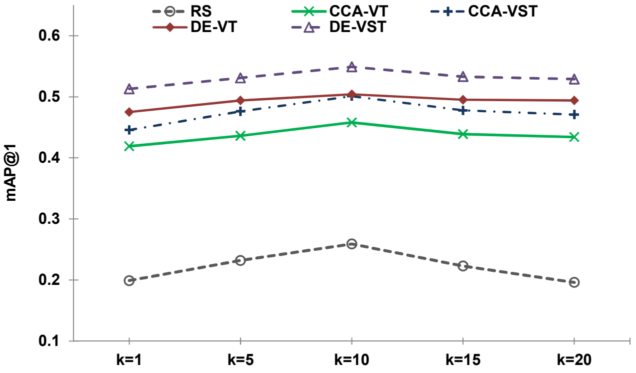


Figure 8: The mAP@1 performance curve with different number of nearest neighbors.

than those by VGG. When combining the representations from VGG and C3D, C3D+VGG further increases the performance gains.

5.5 Effect of the Number of Nearest Neighbors

The number of nearest neighbors is an important parameter in similar video search. In the previous experiments, the number was fixed to 10. Next, we conducted experiments to evaluate the performances of all the compared approaches with the number of nearest neighbors in range of {1, 5, 10, 15, 20}.

The mAP@1 of all the approaches with different number of nearest neighbors are shown in Figure 8. As illustrated in the figure, the proposed $DE-VST$ consistently outperforms others when returning each number of nearest neighbors and the optimal k is happened at 10 for all the compared methods on our dataset. This is also expected, as for few neighbors their associate comments may not be enough to find right ones while too many neighbors will include imperfect similar videos.

5.6 Effect of the Dimensionality of the Embedding Space

In order to show the relationship between the performance and the dimensionality of the embedding space, we compared the results of the dimension in the range of 32, 64, 128, and 256. Note that only the four runs with embedding space learning are included in this comparison.

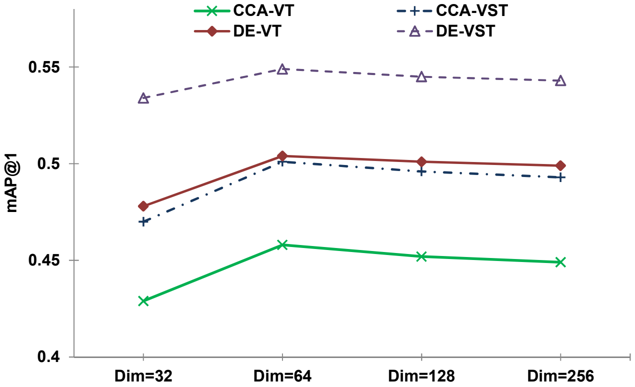


Figure 9: The mAP@1 performance curve with different dimensionality of the embedding space.

Table 2: Speed efficiency of our proposed framework for commenting on a five seconds' video.

Key step	Speed
Representation extraction by VGG	3.6 sec
Representation extraction by C3D	4.3 sec
Representation extraction by ANP-Net	0.62 sec
Similar video search	0.02 sec
Comment dynamic ranking	0.5 sec

Figure 9 shows the mAP@1 performance with respect to different dimensionality of the embedding space. Compared to the other three runs, performance improvements are consistently observed at each dimensionality of the embedding space by our proposed *DE-VST*. Furthermore, *DE-VST* achieves the best result at the dimensionality of 64 and the results at other dimensionality higher than 64 are close to the best one. In addition, the performance trends of mAP at other depths are similar with that of mAP@1.

5.7 Run Time

The experiments were conducted on a regular PC (Intel dual-core 3.39GHz CPU and 16GB RAM). Table 2 details the running time for commenting on a five seconds' video. The major portion of the processing time is consumed by video representation extraction, which takes 3.60, 4.30 and 0.62 seconds by VGG, C3D and ANP-Net, respectively. As we can run each representation extraction in parallel, the total extraction time can be very close to 4.3 sec. Similar video search basically completes a search within 0.02 sec which is very fast. By processing all the comments associated with the top-10 ranked similar videos, dynamic ranking takes about 0.50 sec. Overall, the current implementation finishes in 4.82 sec, which is less than the duration of the video. The run time will be further reduced to 1.20 sec when testing on a single NVIDIA K40 GPU.

6. CONCLUSIONS

We have presented our framework for automatic commenting videos, along with the proposal of techniques for spatio-temporal video representation learning, similar video search and dynamic ranking of candidate comments. Particularly, we analyze the advantages of addressing the commenting problem, from the viewpoint of search rather than generation, by leveraging huge amounts of real user-generated comments on the social platforms. To verify our claim, we have presented approaches based on the existing works in the literature for searching visually similar videos to obtain the comments associated with these similar videos as candidates. In comment dynamic ranking, our deep multi-view embedding model, which projects video content, visual sentiment and textual comment into a common space, offers apparent improvements over other methods, in terms of boosting the chance of finding the right comment from candidates. Experimental results demonstrate that, by searching from a newly created dataset of over 0.1 million videos and 10.6 million comments, top-K comments can be returned to an input video within the duration of that video, with a good chance of finding the desired comments.

Our work can still be improved on several aspects. First, more spatio-temporal video representations will be explored especially for modeling temporal dynamics, e.g., by using alternative solution like recurrent neural networks. Second, our multi-view embedding model can be enhanced by further considering audio information. The audio features can be exploited together with video content and visual sentiment for a more comprehensive manner of characterizing videos.

7. REFERENCES

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, , D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *ICCV*, 2015.

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[3] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM MM*, 2013.

[4] T. Chen, D. Borth, T. Darrell, and S.-F. Chang. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*, 2014.

[5] Y.-Y. Chen, T. Chen, W. H. Hsu, H.-Y. M. Liao, and S.-F. Chang. Predicting viewer affective comments based on image content in social media. In *ICMR*, 2014.

[6] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[7] J. Devlin, S. Gupta, R. Girshick, M. Mitchell, and C. L. Zitnick. Exploring nearest neighbor approaches for image captioning. *arXiv preprint arXiv:1505.04467*, 2015.

[8] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[9] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*. 2010.

[10] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 106(2):210–233, 2014.

[11] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*, 2013.

[12] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[13] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[14] C. J. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *AAAI Conference on Weblogs and Social Media*, 2014.

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[16] M. A. Kheirbek and R. Hen. Dorsal vs ventral hippocampal neurogenesis: implications for cognition and mood. *Neuropsychopharmacology*, 36(1):373, 2011.

[17] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[18] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[19] Y. Pan, Y. Li, T. Yao, T. Mei, H. Li, and Y. Rui. Learning deep intrinsic video representation by exploring temporal coherence and graph structure. In *IJCAI*, 2016.

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

[21] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang. Correlative multi-label video annotation. In *ACM MM*, 2007.

[22] J. P. Romano. On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association*, 85(411):686–692, 1990.

[23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[24] S. Siersdorfer, J. S. Pedro, and M. Sanderson. Automatic video tagging using content redundancy. In *SIGIR*, 2009.

[25] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014.

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

[27] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.

[28] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R. J. Mooney. Integrating language and vision to generate natural language descriptions of videos in the wild. In *COLING*, volume 2, page 9, 2014.

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

[30] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

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

[32] M. Wang, X.-S. Hua, J. Tang, and R. Hong. Beyond distance measurement: constructing neighborhood similarity for video annotation. *IEEE Trans. on MM*, 11(3):465–476, 2009.

[33] J. Xu, T. Mei, T. Yao, and Y. Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, 2016.

[34] T. Yao, F. Long, T. Mei, and Y. Rui. Deep semantic-preserving and ranking-based hashing for image retrieval. In *IJCAI*, 2016.

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

[36] T. Yao, T. Mei, C.-W. Ngo, and S. Li. Annotation for free: Video tagging by mining user search behavior. In *ACM MM*, 2013.

[37] G. Ye, Y. Li, H. Xu, D. Liu, and S.-F. Chang. Eventnet: A large scale structured concept library for complex event detection in video. In *ACM MM*, 2015.