Visual-Texual Emotion Analysis With Deep Coupled Video and Danmu Neural Networks

Chenchen Li, Jialin Wang, Hongwei Wang, Miao Zhao, Wenjie Li, and Xiaotie Deng

***Abstract*—User emotion analysis toward videos is to automatically recognize the general emotional status of viewers from the multimedia content embedded in the online video stream. Existing works fall into two categories: 1) *visual-based* methods, which focus on visual content and extract a specific set of features of videos. However, it is generally hard to learn a mapping function from low-level video pixels to high-level emotion space due to great intra-class variance. 2) *textual-based* methods, which focus on the investigation of user-generated comments associated with videos. The learned word representations by traditional linguistic approaches typically lack emotion information and the global comments usually reflect viewers' high-level understandings rather than instantaneous emotions. To address these limitations, in this paper, we propose to jointly utilize video content and user-generated texts simultaneously for emotion analysis. In particular, we introduce exploiting a new type of user-generated texts, i.e., "danmu," which are real-time comments floating on the video and contain rich information to convey viewers' emotional opinions. To enhance the emotion discriminativeness of words in textual feature extraction, we propose *Emotional Word Embedding* (EWE) to learn text representations by jointly considering their semantics and emotions. Afterward, we propose a novel visual-textual emotion analysis model with *Deep Coupled Video and Danmu Neural networks* (DCVDN), in which visual and textual features are synchronously extracted and fused to form a comprehensive representation by deep-canonically-correlated-autoencoder-based multi-view learning. Through extensive experiments on a self-crawled real-world video-danmu dataset, we prove that DCVDN significantly outperforms the state-of-the-art baselines.**

***Index Terms*—Danmu, deep multimodal learning, emotion analysis.**

I. Introduction

IN SOME online video platforms, such as Bilibili[1] and Youku,[2] overlaying moving subtitles on video playback streams have become a featured function on the websites, through which users can share feelings and express attitudes towards the content of videos when they are watching. Given an online video clip as well as its associated textual comments, visual-textual emotion analysis is to automatically recognize the general emotional status of viewers towards the video with the help of visual information and embedded comments. A precise visual-textual emotion analytical method will promote in-depth understanding of viewers' experience, and benefit a broad range of applications such as opinion mining [1], [2], affective computing [3]–[5], recommendation systems [6], and trailer production [7].

Existing methods for emotion analysis of online videos can be divided into two categories according to the types of input data. The first class of methods is *visual-based*, i.e., they take the visual content in videos as input, and perform emotion analysis based on the visual information. Typically, in visual-based methods, a specific set of visual features are extracted from video frames to reveal its underlying emotion, such as [8], [9] using the low-level features, [10] using the mid-level features and [11]–[13] using the deep features. However, visual-based methods exhibit the following limitations: 1) It is generally hard to learn a mapping function solely from low-level video/image pixels to high-level emotion space due to the great intra-class variance [11], [14]. 2) It is only feasible to directly apply visual-based methods to images and short videos, as the features of video would increase explosively with its length. Otherwise, visual features need to be periodically sampled. 3) In visual-based methods [2], the well-selected visual features are more relevant to the emotion of the video content than the emotion of viewers, which could inevitably dampen their performance in user emotion analysis scenarios.

As opposed to visual-based methods, the second class of methods is *textual-based*, which utilize user-generated textual comments as input, and extract linguistic or semantic information as features for emotion analysis [6]. Based on their methodologies, existing textual-based methods can further be classified into lexicon-based methods and embedding-based methods.

Manuscript received February 5, 2019; revised July 10, 2019; accepted September 29, 2019. Date of publication October 9, 2019; date of current version May 21, 2020. This work was supported by the National Key Research and Development Program of China under Grant 2017YFB0701900. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Elisa Ricci. *(Corresponding author: Chenchen Li.)*

C. Li and H. Wang are with the Department of Computer Science, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: lcc1992@sjtu.edu.cn; wanghongwei55@gmail.com).

J. Wang, M. Zhao, and W. Li are with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong (e-mail: wangjialin@hust.edu.cn; mzhao.ny@gmail.com; cswjli@comp.polyu.edu.hk).

X. Deng is with the School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: xiaotie@pku.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2019.2946477

[1]https://www.bilibili.com

[2]http://www.youku.com

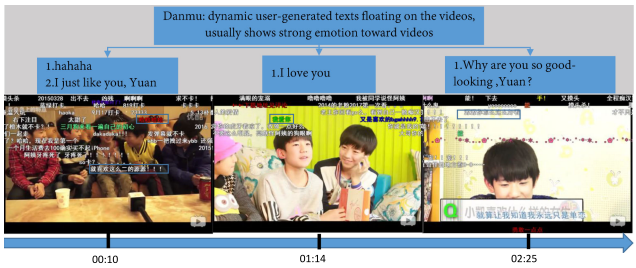


Fig. 1. Illustration of danmus associating with a video clip on Bilibili website.

Traditional lexicon-based approaches [15], [16] lack considering the syntactic and semantic information of words, hence unable to achieve satisfactory performance in practice. Recently, word2vec [17], as a typical example of embedding-based methods, provides an effective way of modeling semantic context of words. However, word2vec can only extract semantic proximity of words from texts, while the contextual emotional information is ignored. As a result, words with different emotions, such as happy and anger, are mapped to close vectors [18]. Moreover, it is worth noticing that, most of the textual-based methods are based on the global comments for videos (comments that are attached to the videos below), which, unfortunately, can only reflect viewers' high-level understandings on the content rather than their emotion development towards the video.

To address the aforementioned limitations, in this paper, we consider analyzing the viewer's emotion towards online videos by utilizing a new type of textual data, known as "danmu". Unlike the traditional global comments gathered in a comment section below the videos, danmu is the real-time comments floating on the video in the snapshot, moving along with video playback. Viewers can watch the video while sending comments and reading other viewers' comments simultaneously. An example of danmu screenshot is illustrated in Fig. 1. Generally, as viewers can express their emotion without any delay, danmus are real-time commentary subtitles and play an important role in conveying emotional opinion from the commentator to other viewers. Compared with global comments, danmus have two distinguishing characteristics: 1) Danmus are highly correlated with the specific moments in video clips. 2) Danmus are generally not distributed uniformly over the whole video. In fact, the distribution pattern of danmus reflects the development of the viewers' emotion, e.g., emotion burst, which could greatly facilitate emotion analysis tasks.

Given danmu as the new source of data, we propose a novel visual-textual emotion analysis model, named *Deep Coupled Video and Danmu Neural networks* (DCVDN). DCVDN takes both video frames and associated danmus as input data and aims to construct a joint emotion-oriented representation for viewers' emotion analysis. Specifically, for each video clip, we first perform clustering on all of its danmus according to their burst pattern. Each set of clustered danmus is aggregated into one danmu document as nearby danmus express viewers' attitudes towards similar video content at a specific moment. In DCVDN, to overcome the limitation of emotion-unaware textual-based methods, we propose a novel textual representation learning method, called *Emotional Word Embedding* (EWE), to learn textual features from danmu documents. The key idea of EWE is to encode emotional information along with the semantics into each word for joint word representation learning, which is proved to be able to effectively preserve the original emotion information in texts during the learning process. In addition, we also extract video features from video frames synchronized with the burst points of danmu. As viewer's emotion can be reflected as a joint expression of both video content and danmu texts, in this work, we intend to explore the learning of highly non-linear relationships that exist among the visual and textual features. In DCVDN, joint emotion-oriented representation is developed over the space of video and danmu, by utilizing a Deep Canonically Correlated Auto-Encoder (DCCAE) to achieve multi-view learning for emotion analysis.

To evaluate our proposed DCVDN, we collect video clips and their associated danmus from Bilibili, one of the most popular online video websites in China. Our video-danmu dataset consists of 4,056 video clips and 371,177 danmus, in which each example is associated with one of seven emotion classes: Happy, Love, Anger, Sad, Fear, Disgust, and Surprise. We compare our DCVDN with 14 state-of-the-art baselines by conducting extensive experiments on the video-danmu dataset, and the results demonstrate that DCVDN achieves substantial gains over other visual-based or textual-based methods. Specifically, DCVDN outperforms visual-based baselines by 54.87% to 78.54% on *Accuracy* and outperforms textual-based baselines by 9.27% to 241.59% on *Accuracy*.

II. Related Work

Among textual-based methods for emotion analysis, lexicon [15], [16] has been widely used due to its simplicity. In [15], they leverage lexicon emotion dictionaries and POS tags to extract linguistic features from the textual documents. In [16], they propose to predict the sentiment of tweets with the help of the positive and negative emoticons. However, lexicon-based methods cannot exploit the relationship between words. Recently, distributed representations of words have emerged and successfully proliferated in language models and NLP tasks [17], [19], which can encode both syntactic and semantic information of words into low-dimensional vectors to quantify and categorize semantic similarities between words. Most word embedding models typically represent each word using a single vector, making them indiscriminative under different emotion circumstances. Aware of this limitation, some multi-prototype vector space models have been proposed [20]–[22]. [20] uses latent topic models to discriminate word representations by jointly considering words and their contexts. [22] uses a mix of unsupervised and supervised techniques to learn word vectors capturing semantic term-document information as well as rich sentiment content. Distinguishable from existing works, our EWE first uses Latent Dirichlet Allocation (LDA) [23] to infer emotion labels and then incorporates them along with word context in representation learning to differentiate words under different emotional and semantic context. There are also various topic models on sentiment analysis [24], [25]. [24] proposes

a novel probabilistic modeling framework based on LDA to detect sentiment and topic simultaneously from the text. [25] observes that sentiments are dependent on the local context, and relaxes the sentiment independent assumption. It considers the words of the sentiment as a Markov chain. A probabilistic graphical model can also help to learn discriminative multimodal descriptors and infer the confidence of label noise when it's difficult to collect a sufficient amount of training labels [4].

There are also quite a lot of works on visual sentiment analysis. For example, [8], [26] use low-level image properties, including pixel-level color histogram and Scale-Invariant Feature Transform (SIFT), as the features to predict the emotion of images. [1], [10] employ middle-level features, such as visual entities and attributes, as the features for emotion analysis. Besides, kernelized features [13] are also used to recognizing emotions in user-generated videos by reformulating the equation of the discrete Fourier transform and constructing a polynomial kernel function based on the linear kernel. It shows superior discriminative capability than spatial features. In [27], they focus on the object and the background, and found there may be slight differences in emotion due to different backgrounds even with the same object. The output emotion values in their framework are in two-dimensional space valence and arousal, which are more effective than using a small number of emotion categories. The continuous probability distribution prediction of visual emotions (valence and arousal) is also considered in [5]. They observed that the emotion distribution can be well-modeled by a Gaussian mixture model, and formalize the emotion distribution prediction task as a shared sparse regression (SSR) problem and extend it to multitask settings finally. [9], [11] utilize Convolutional Neural Networks (CNNs) to extract high-level features through a series of nonlinear transform, which has been proved surpassing other models with low-level and mid-level features [9]. [28] think that the local areas are pretty relevant to human's emotional response to the whole image and proposed a model to utilize the recent studies attention mechanism to jointly discover relevant local regions and build a sentiment classifier on top of these local regions. [29] presented a new deep visual-semantic embedding model trained to identify visual objects using both labeled image data as well as semantic information gleaned from the unannotated text. Inspired by the observation that the whole image and local regions convey significant sentiment information, [2] propose a framework to leverage affective regions to remove redundant and noisy proposals, and aggregate the output of local images and whole images to produce the final predictions. The method for video classification [30] can also be applied for visual sentiment analysis while they usually perform worse the visual sentiment methods.

To combine visual and textual information, recent years have witnessed some preliminary effort on multimodal models. For example, [19], [26] employ both text and images for sentiment analysis. [26] employs Deep Boltzmann Machine (DBM) to fuse features from audio-visual and textual modalities, while [19] employs cross-modality consistent regression. Moreover, Deep Neural Network (DNN) based approaches [26] are generally used for multi-view representation fusion. Prior works have shown the benefits of multi-view method on emotion analysis [19]. One step advanced in our work, we employ DCCAE [31], which combines autoencoder and canonical correlation to obtain unsupervised representation learning by jointly optimizing the reconstruction errors minus canonical correlation between extracted features in multiple views. Autoencoder is a useful tool for representation learning, in which the objective is to learn a compact representation that best reconstructs the inputs [32] via unsupervised learning. In [33], it introduced an encoder-decoder pipeline that learns (a): a multimodal joint embedding space with images and text and (b): a novel language model for decoding distributed representations from our space. Canonical correlation analysis (CCA) [34] can maximize the mutual information between different modalities and has been justified in many previous works [35]–[37]. In [36], it makes CCA to learn the correlations between visual features and textual features for image retrieval. The discriminative multiple canonical correlation analysis (DMCCA) for multimodal information analysis and fusion was proposed in [38]. It demonstrated that the optimally projected dimension by DMCCA can be quite accurately predicted, leading to both superior performance and less computational cost. In [37], it uses a variant CCA to learn a mapping between textual words and visual words. Although multi-view methods have been studied extensively, there only exist few works on emotional analysis [19], [39], [40]. For example, [39] proposed a novel Cross-media Bag-of-words Model (CBM) for Microblog sentiment analysis. It represented the text and image of a Weibo tweet as a unified Bag-of-words representation. In [6], they propose a sentiment-based rating prediction method, which not only considers a user's own sentimental attributes but also take interpersonal sentimental influence into consideration, to improve prediction accuracy in recommender systems. A deep and bidirectional representation learning model is proposed in [41] to address the issue of image-text cross-modal retrieval.

III. Visual-Textual Emotion Analysis

In this section, we discuss the proposed DCVDN with details. We first provide a model overview and then introduce video and danmu preprocessing, EWE, DCCAE, and classification in the subsequent subsections, respectively. Each video is with one label, thus we are solving a one-label classification problem.

A. DCVDN Overview

Fig. 2 depicts the framework of DCVDN, which consists of three modules: preprocessing and feature extraction, multi-view representations learning, and classification.

The first module is committed to preprocess the inputs and extract visual and textual features. It is observed that, for each video clip, danmus are likely to burst around some key video frames. The distributions of danmus usually reflect the emotion evolution of the viewers and nearby danmus are more likely to express emotions towards the same video content. Therefore, for each video clip, we cluster all of its associated danmus according to their burst pattern. Utilizing the results, we aggregate the danmus in each cluster into one document, since it is more effective to analyze longer document rather than shorter ones, which are typically semantic and emotion-ambiguous. Afterward, we

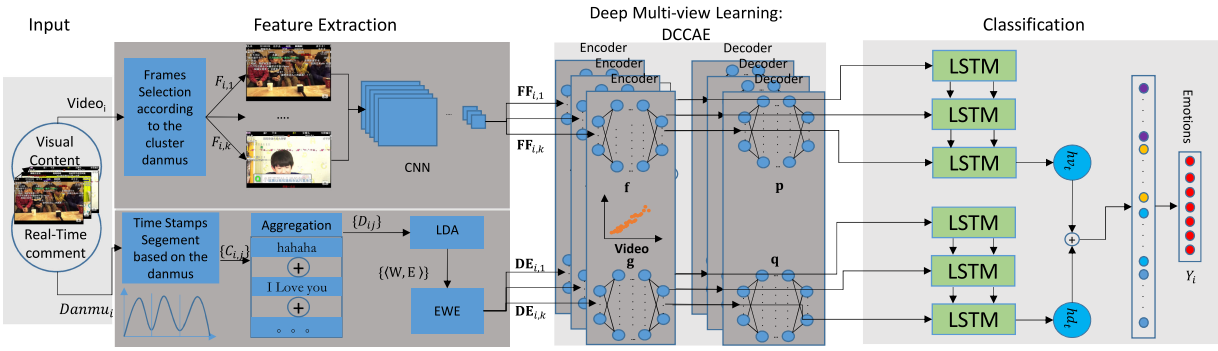


Fig. 2. The framework of DCVDN: the video and associated danmus are clustered based on danmus' burst pattern, video segments and danmu documents are synchronized in time, visual and textual features are extracted respectively by CNN and EWE, and finally the joint representations are learned by DCCAE for the subsequent classification.

aim to learn emotion-aware word embedding and document embedding for each word and each danmu document, respectively. Correspondingly, we propose EWE, which combines semantic and emotional information of each word to enhance emotion discriminativeness. For videos, we synchronize the selection of the frames corresponding to the burst points of danmus and focus on feature extraction from those selected frames, which are more important and relevant than others to invoke viewers' emotion burst. We apply pre-trained Convolutional Neural Networks (CNN) to extract features of the video frames as CNN has been proved to achieve the state-of-the-art performance on sentiment analysis [9]. These danmu document embeddings and CNN features will be fed into following DCCAE for further joint representation learning.

The second module is the multi-view representation learning for information fusion between video and danmu. The documents of danmu have highly direct correlations with viewers' emotion and video frames can provide robust background information with appropriate guidance. In DCVDN, for each pair of danmu document and corresponding video frame, we employ DCCAE to learn a multi-view representation in an unsupervised way. A set of obtained multi-view representations will be fed into the following classification module as the input features. From the implementation point of view, unsupervised joint representation learning ahead of supervised classification helps avoid complicated end-to-end model training, which effectively facilitates the convergence of the training process in practice.

The last module refers to the classification task. It is clear that for each video clip, the multi-view representations output from the second module are still in time series, each corresponding to a clustered time period in the video. Hence, Long Short-Term Memory (LSTM) is adopted to address the time dependency across those features. The output of LSTM is treated as the ultimate emotion-aware embedding for each video clip, and eventually fed into softmax function to obtain the target emotion prediction.

B. Preprocessing and Feature Extraction

In this subsection, we discuss the preprocessing and feature extraction methods for danmus and videos in detail. The whole process is also shown in Algorithm 1.

Algorithm 1: Preprocessing and Feature Extraction

Input: Videos = $\{ video_1, \ldots, video_N \}$, Danmus = $\{ danmu_1, \ldots, danmu_N \}$.
Output: **DEs**, **FFs**.
for $i \leftarrow 1$ **to** N **do**
  $C_{i,1}, \ldots, C_{i,K}$ = K-means($danmu_i$);
  for $j \leftarrow 1$ **to** K **do**
    $D_{i,j}$ = Aggregate($C_{i,j}$);
$\langle \mathbf{W}, \mathbf{E} \rangle$ = LDA($\{D_{i,j}\}_{1 \le i \le N, 1 \le j \le K}$)
$\{\mathbf{DE_{i,j}}\}_{1 \le i \le N, 1 \le j \le K}$ = EWE($\langle \mathbf{W}, \mathbf{E} \rangle$)
for $i \leftarrow 1$ **to** N **do**
  for $j \leftarrow 1$ **to** K **do**
    $F_{i,j}$ = FrameSelect($video_i, C_{i,j}$);
    $\mathbf{FF_{i,j}}$ = CNN($F_{i,j}$);
DEs = $\{\mathbf{DE}_{1,1}; \ldots; \mathbf{DE}_{N,K}\}$;
FFs = $\{\mathbf{FF}_{1,1}; \ldots; \mathbf{FF}_{N,K}\}$;

1) Preprocessing and Feature Extraction on Danmu: As aforementioned, danmu is a kind of timely user-generated comment with non-uniform distribution over the entire video. The distribution of danmus reflects user engagement to the video content and the video content at burst points of danmus is typically more attractive to viewers than other parts. Aware of this phenomenon, we apply the K-means algorithm to segment danmus into a set of clusters according to their burst pattern and aggregate all danmus in the same cluster into a danmu document. Formally speaking, consider a dataset with a total of N videos, denoted by $V = \{video_1, \ldots, video_N\}$. Each $video_i$ is associated with a collection of danmus, denoted by $danmu_i = \{(s_{i,1}, offset_{i,1}), \ldots, (s_{i,n_i}, offset_{i,n_i})\}$, where $s_{i,j}$ represents the text of j-th danmu for $video_i$, $offset_{i,j}$ represents the emergence moment of $s_{i,j}$ relative to the beginning of $video_i$ ($offset_{i,j} < offset_{i,h}$, if $j < h$), and n_i is the total number of danmus in $video_i$. For each $video_i$, we aim to find a K-partition $\{C_{i,1}, ..., C_{i,K}\}$ satisfying

$$\underset{C_{i,1},\ldots,C_{i,K}}{\operatorname{argmin}} \sum_{j=1}^{K} \sum_{h=1}^{|C_{i,j}|} \left| offset_{i,\sum_{t=1}^{j-1}|C_{i,t}|+h} - \mu_{i,j} \right|^2, \quad (1)$$

where $\cap_{1\le j\le K} C_{i,j} = \emptyset$ and for each cluster j, we have $\mu_{i,j} = \frac{\sum_{h=1}^{|C_{i,j}|} offset_{i,\sum_{t=1}^{j-1}|C_{i,t}|+h}}{|C_{i,j}|}$, which is the centroid of cluster j and is also treated as the burst point of cluster $C_{i,j}$. Once clusters are formed, we obtain the danmu document set for $video_i$ by aggregating all danmus in same clusters, i.e., $D_i = \{D_{i,1}, ..., D_{i,K}\}$, where each $D_{i,j}$ corresponds to a danmu document, i.e., $D_{i,j} = \oplus_{1\le h\le|C_{i,j}|} s_{i,\sum_{t=1}^{j-1}|C_{i,t}|+h}$, and the document set includes all danmu documents associated with $video_i$, i.e., $D_i = \cup_{1\le j\le K} D_{i,j}$.

To extract the textual features from danmu to enhance emotion discriminativeness, we correspondingly propose emotion-based Latent Dirichlet Allocation (eLDA) [23] and EWE to first learn the emotional embedding of each word and then derive the emotional document embedding for each danmu document $D_{i,j}$. We will discuss the details in the next subsection.

2) Preprocessing and Feature Extraction on Video: For videos, we exploit the clustering information in danmus to select frames for visual feature extraction. Specifically, we draw out the video frames corresponding to the burst points of danmu clusters in each video clip as they are more attractive to the viewers. In this way, the video frames and danmu documents are synchronized in time. Formally, for each video $video_i$, based on danmu cluster partition $danmu_i = \{C_{i,1}, ..., C_{i,K}\}$, we select the keyframe $F_{i,j}$ at the time moment of burst point $\mu_{i,j}$ to represent the basic visual content of cluster j. As the result, we would get a set of frames for $video_i$, i.e., $\{F_{i,1}, ..., F_{i,K}\}$, in one-to-one correspondence to the danmu clusters $\{C_{i,1}, \ldots, C_{i,K}\}$.

Previous work [9] has shown that visual features extracted by CNN networks can achieve satisfactory performance for emotion analysis. Therefore, in this work, we employ the pre-trained CNN, i.e., VGG-Net fc-7 [42], for visual feature extraction from each video frame $F_{i,j}$. Basically, danmu texts could explicitly deliver viewers' opinions and video frames would provide supportive background information of emotion-relevant content.

C. Danmu Document Embedding Learning

In this subsection, we discuss the emotion-oriented embedding learning for word and danmu documents. We first introduce eLDA method to estimate the emotion label of each word, and then discuss the details of proposed EWE model, which aims to combine emotion and semantic information to learn a comprehensive and effective word embedding to facilitate viewers' emotion analysis.

1) eLDA: LDA [23] is an unsupervised model and is commonly used to infer the topic label for words and documents. Inspired by LDA for topic analysis, in this work, we exploit it to infer the emotion labels by considering danmu documents as a random mixture over latent emotions and each emotion is characterized by a distribution over words. Particularly, each danmu document $D_{i,j}$ is represented as a multinomial distribution $\boldsymbol{\Theta}_D$ from Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\alpha})$ over a set of emotions, each emotion is usually represented as a multinomial distribution $\boldsymbol{\Pi}_l$ over a set of vocabulary from $\mathrm{Dir}(\boldsymbol{\beta})$. The generative process is defined formally as follow:

- For each danmu document $D_{i,j}$, choose a multinomial distribution $\boldsymbol{\Theta}_D$ over the emotions from $\mathrm{Dir}(\boldsymbol{\alpha})$;
- For each emotion l, choose a multinomial distribution $\boldsymbol{\Pi}_l$ over the words from $\mathrm{Dir}(\boldsymbol{\beta})$;
- For each word position t in document $D_{i,j}$,
  - Choose an emotion l_t from $\mathrm{Multinomial}(\boldsymbol{\Theta}_D)$;
  - Choose a word w_t from the $\mathrm{Multinomial}(\boldsymbol{\Pi}_{l_t})$.

From the implementation perspective, in order to effectively infer emotions, we need prior knowledge of the emotional ground truths of some words. When determining the emotion of a word, if the word exists in our emotion lexicon, we choose to use its corresponding emotion in the lexicon, otherwise, we choose the emotion according to the probabilities of $\mathrm{Multinomial}(\boldsymbol{\Theta}_D)$. Considering that danmu culture (sometimes called manga and anime) is mainly popular among youngsters, the authentic word emotions are somewhat different from common sense in the existing lexicon. Therefore, it is desired to build a new lexicon specifically for the manga and anime culture. We spend great effort to construct such kind of lexicon, which consists of 1,592 network-popular words and 1,670 emoticons. Emoticons are the textual portrayals of a user's moods or facial expressions in the form of icons. For example, ^_^ represents happiness and ⊤_⊤ stands for crying. We select these focused words and emoticons according to their occurrence frequency in our dataset.

It is worth pointing out that we cluster the emotion distribution into a certain number of classes and treat the result of clustering as the final emotion label for each word, rather than directly use the emotion with the maximal probability as adopted in TWE [20]. Specifically, suppose we obtain the emotion distribution $\boldsymbol{\Theta}_{i,j}$ of each word $w_{i,j}$ (the j-th word in the danmu aggregation of $video_i$) after the interference of eLDA model. Then we use the K-means algorithm to cluster these emotion distributions, which aims to find a KE-partition $\{CE_1, \ldots, CE_{KE}\}$ satisfying

$$\underset{CE_1,\ldots,CE_{KE}}{\operatorname{argmin}} \sum_{k=1}^{KE} \sum_{\boldsymbol{\Theta}_{i,j}\in CE_k} |\boldsymbol{\Theta}_{i,j} - \boldsymbol{\eta}_k|^2 \tag{2}$$

where $\cap_{1\le k\le KE} C_k = \emptyset$ and η_k is the centroid of cluster k. The new emotion label $l_{i,j}$ of $w_{i,j}$ is k if $\boldsymbol{\Theta}_{i,j} \in CE_k$. The reason for the clustering is that the number of emotion labels is generally small (7 for emotion classification tasks) and we can't fully explore the information hidden in the distributions with such a few labels. To avoid the dilemma, we recluster the distributions into more classes in order to make the new labels more discriminative. The new labels would be used to learn EWE at a later time.

2) Emotional Word and Document Embeddings: Word embedding, which represents each word using a vector, is widely used to capture semantic information of words. Skip-Gram model [17] is a well-known framework for word embedding, which learns word representation that is useful for predicting context words in a sliding window when given the target word. Formally, given a word sequence $\{w_1, \ldots, w_T\}$, the objective

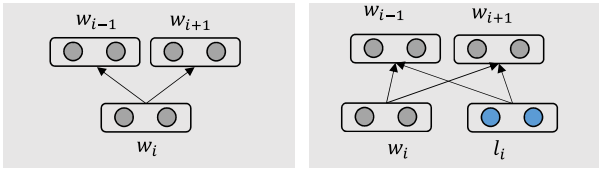


Fig. 3. Skip-Gram and EWE models. Gray and blue circles represent word and emotion embeddings, respectively.

of Skip-Gram is to maximize the average log probability

$$\mathcal{L}(D) = \frac{1}{T}\sum_{t=1}^{T}\sum_{-k\le c\le k, c\neq 0} \log p(w_{t+c}|w_t), \tag{3}$$

where k is the context window size for the target word, which can be a function of the centered word w_t. Probability $p(w_c|w_t)$ is defined as a softmax function as follows:

$$P(w_c|w_t) = \frac{\exp\left(\mathbf{w}_c^\top \mathbf{w}_t\right)}{\sum_{\mathbf{w}\in W}\exp\left(\mathbf{w}^\top \mathbf{w}_t\right)}, \tag{4}$$

where $\mathbf{w}_t$ and $\mathbf{w}_c$ are word vectors of target word w_t and context word w_c, respectively, and W is word vocabulary.

It is noticed that Skip-Gram model for word embedding focuses on the semantic context and assumes that each word always preserves a single vector, which sometimes is indiscriminate under different emotion circumstances. These facts motivate us to propose a joint emotion and semantics learning model, named EWE. The basic idea of EWE is to preserve emotion information of words when measuring the interaction between target word w_t and context word w_c. In this way, a word with different associated emotions would correspond to different embeddings so as to effectively enhance the emotion discriminativeness of each word.

Specifically, rather than solely using the target word w_t to predict context words in Skip-Gram, inspired by [20], EWE jointly utilizes l_t's as well, i.e., the emotions of the words in the danmu documents. EWE aims to learn the representations for words and emotions separately and simultaneously. In particular, it regards each emotion as a pseudo word and considers the occurrence of this pseudoword in all the positions wherever the positioned words are assigned with this emotion. EWE uses both the target word w_t and its associated emotion l_t to predict context words, as shown in Fig. 3. For each target word with its emotion $\langle w_t, l_t\rangle$, the objective of EWE is to maximize the following average log probability

$$\mathcal{L}(D) = \frac{1}{T}\sum_{t=1}^{T}\sum_{-k\le c\le k}^{c\neq 0}\Big(\log p(w_{t+c}|w_t) + \log p(w_{t+c}|l_t)\Big), \tag{5}$$

where $P(w_c|l_t)$ is similar with $P(w_c|w_t)$

$$P(w_c|l_t) = \frac{\exp\left(\mathbf{w}_c^\top \mathbf{l}_t\right)}{\sum_{\mathbf{w}\in W}\exp\left(\mathbf{w}^\top \mathbf{l}_t\right)}, \tag{6}$$

and $\mathbf{w}$ and $\mathbf{l}$ are the represention vectors of words and emotions respectively. When we minimize the log loss $\log p(w_{t+c}|l_t)$, we consider that w_t is l_t. The other learning process of the emotional word is the same as the textual word. The process is shown in Algorithm 2. N_{seq} is the length of the documents, N_w is the number of vocabulary and N_l is the number of emotions, which are the results of clustering of eLDA. c is the size of contextual window, m is the user-defined size of representation and μ is the learning rate. V_w and V_l are the representation vectors of words and emotions respectively.

Algorithm 2: EWE

Input: $W = [\langle w_1, l_1\rangle, \ldots, \langle w_{N_{seq}}, l_{N_{seq}}\rangle]$, c, m, μ
Output:
$\mathbf{V}_w = [\mathbf{v}_w(w_1), \ldots, \mathbf{v}_w(w_{N_w})]$,
$\mathbf{V}_l = [\mathbf{v}_l(l_1), \ldots, \mathbf{v}_l(l_{N_l})]$
Initialize randomly a matrix $\mathbf{V}_w \in R^{N_w\times m}$
Initialize randomly a matrix $\mathbf{V}_l \in R^{N_l\times m}$
for $i \leftarrow 1$ **to** n **do**
  /* n is the length of the sentence. */
  $L = 0$
  Forward Propagation:
  for $j \leftarrow i-c$ **to** $i+c$ **do**
    if $j < 0$ *or* $j > n$ **then**
      w_j = #
    if $j \neq i$ **then**
      $L = L + \log P(w_j|w_i) + \log P(w_j|l_i)$
  Backward Propagation:
  $\mathbf{v}_w(w_i) = \mathbf{v}_w(w_i) - \mu\frac{\partial L}{\partial \mathbf{v}_w(w_i)}$
  $\mathbf{v}_l(l_i) = \mathbf{v}_l(l_i) - \mu\frac{\partial L}{\partial \mathbf{v}_l(l_i)}$

Emotional word embedding of word w in emotion l is obtained by concatenating the embeddings of $\mathbf{w}$ and $\mathbf{l}$, i.e., $\mathbf{w}^l = \mathbf{w}\oplus\mathbf{l}$, where $\oplus$ is the concatenation operation and the dimension of $\mathbf{w}^l$ is double of $\mathbf{w}$ and $\mathbf{l}$. Correspondingly, the document embedding in EWE is to aggregate emotional word embeddings of the words in a danmu document. The document embedding of $D_{i,j}$ is defined as $\mathbf{d} = \sum_{w\in D_{i,j}} \mathrm{P}(w|D_{i,j})\mathbf{w}^l$, where $\mathrm{P}(w|D_{i,j})$ can be the term frequency-inverse document frequency of word w in $D_{i,j}$.

D. Deep Multi-View Representation Learning

In this subsection, we introduce the multi-view representation learning method in DCVDN, which aims to simultaneously utilize video and danmu information to learn a joint representation based on the extracted visual and textual features. Inspired by canonical correlation analysis (CCA) and reconstruction-based objectives, we employ deep canonically correlated autoencoders to fuse the latent features from video and danmu points of views. In particular, DCCAE consists of two autoencoders and optimizes the combination of canonical correlation between the learned textual and visual representations and the reconstruction errors of the autoencoders. The structure of DCCAE is shown in the module of "Deep Multi-view Learning: DCCAE" in Fig. 2

and its optimization objective is as follows

$$\min_{\mathrm{W_f},\mathrm{W_g},\mathrm{W_p},\mathrm{W_q},\mathrm{U},\mathrm{V}} -\frac{1}{M}\,\mathrm{tr}(\mathrm{U}^\top \mathrm{f}(\mathrm{X})\mathrm{g}(\mathrm{Y})^\top \mathrm{V}) + \frac{\lambda}{M}\sum_{i=1}^{M}(\|\mathrm{x}_i - \mathrm{p}(\mathrm{f}(\mathrm{x}_i))\| + \|\mathrm{y}_i - \mathrm{q}(\mathrm{g}(\mathrm{y}_i))\|), \quad \mathbf{s.t.}\ \ \mathrm{u_i}^\top \mathrm{f}(\mathrm{X})\mathrm{f}(\mathrm{X})^\top \mathrm{u_i} = \mathrm{v_i}^\top \mathrm{g}(\mathrm{Y})\mathrm{g}(\mathrm{Y})^\top \mathrm{v_i} = M, 1 \le i \le L, \tag{7}$$

where $\lambda > 0$ is the trade-off parameter, M is the sample size, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_M]$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_M]$ are the feature matrices of visual and textual viewpoints, each $\mathbf{x}$ and y referring to the visual and textual features extracted from a damu document and corresponding video frame, respectively. Moreover, $\mathbf{f}$, $\mathbf{g}$, $\mathbf{p}$ and $\mathbf{q}$ denote mapping functions implemented in autoencoders. The encoder-decoder pair ($\mathbf{f}$, $\mathbf{p}$) and ($\mathbf{g}$, $\mathbf{q}$) constitute two autoencoders, each for one of two viewpoints. The corresponding parameters in encoding functions $\mathbf{f}$ and $\mathbf{g}$ and decoding functions $\mathbf{p}$ and $\mathbf{q}$ are denoted by $\mathbf{W_f}$, $\mathbf{W_g}$, $\mathbf{W_p}$, and $\mathbf{W_q}$, respectively. $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_L]$ and $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_L]$ are the CCA directions that project the outputs of $\mathbf{f}$ and $\mathbf{g}$, where L is the dimensionality of input features to autoencoders.

Mathematically, the first term of Eq. (7) is the objective of CCA, while the second and third terms are the losses of autoencoders, which can be understood as adding autoencoders as regularization terms to CCA objective. The constraint is for CCA to ensure the objective is invariant to the scale of $\mathbf{U}$ and $\mathbf{V}$. CCA aims to maximize the mutual information between videos and danmus, while autoencoders aim to minimize the reconstruction errors of two views. In this way, DCCAE tries to explore an optimal trade-off between the information captured from the reconstruction of each view and the information captured from learning the mutual information of two views. Therefore, it can achieve better representations. The output features of two views are $\mathrm{U}^\top \mathrm{f}(\mathrm{X})$ and $\mathrm{V}^\top \mathrm{g}(\mathrm{Y})$, respectively, which would be as the inputs to two seperated LSTM for the later classification.

E. Classification

In this module, the outputs of DCCAE are utilized to facilitate the emotion classification task. As aforementioned, the output representations from DCCAE are still in time series. To address time dependency across the representations, we feed the representations from two modalities into two separated LSTMs and get the final outputs of two LSTMs, i.e., h_v from the video part and h_t from the text part. Then we simply concatenate h_v and h_t to form a unified and comprehensive representation, i.e., $h_a = [h_v, h_t]$. The obtained representation h_a would eventually be fed into the following fully-connected neural network with a softmax function to obtain the target emotion prediction. It is worth pointing out that the output of the classifier would be the probability distribution over seven emotion classes under our investigation. This is consistent with the fact that video clips typically involve a mixture of different emotions. However, since our ultimate goal is to indicate the main and dominant emotion in this work, each video clip would be finally associated with only one emotion label as the output. The classification network is depicted as the rightmost module in Fig. 2.

TABLE I
THE BASIC STATISTICS OF THE VIDEO-DANMU DATASET

Number of Videos			Avg. Length		Length Range	
4,056			82.89s		1.44s - 514.13s	
Happy	Love	Anger	Sad	Fear	Disgust	Surprise
620	877	290	631	647	669	322

TABLE II
NUMBER OF WORDS AND EMOTICONS OF DIFFERENT EMOTION CLASSES IN SELF-BUILT EMOTION LEXICON

Number of Words						
Happy	Love	Anger	Sad	Fear	Digust	Surprise
90	784	25	132	146	348	67
Total	1592	Ave. Occurrence Freq.			1550.39	
Number of Emoticons						
Happy	Love	Anger	Sad	Fear	Digust	Surprise
417	522	41	192	148	265	85
Total	1670	Ave. Occurrence Freq.			7.90	

IV. EXPERIMENTS

In this section, we carry out extensive experiments to evaluate the performance of DCVDN. We first introduce the datasets used for experiments and then compare our model with other 14 baselines, which address on visual, textual and joint features for emotion analysis and videos classification, respectively.

A. Datasets

1) Datasets For Video-Danmu: Considering the lack of existing danmu-related dataset, we put great effort to self construct a new dataset, called Video-Danmu.[3] This dataset includes videos and their associated danmus directly crawled from the Bilibili website, which is one of the most popular websites providing danmu services in China.

We recruited four students to label the dataset. They label them according to their feeling. Danmu is popular among young people, thus they should be familiar with them. In the first round, each video and each word is labelled by two persons independently. In the second round, for the videos or the words whose labels are not the same from the two persons, the final labels are determined by the discussions of them.

There are 4,056 videos in the dataset, which last ranging from 1.44 to 514.13 seconds and average at 82.89 seconds. We labelled the videos into 7 emotion classes, i.e., Happy, Love, Anger, Sad, Fear, Disgust and Surprise, with the help of a group of student helpers in our university. Table I shows the basic statistics of the dataset. The number of videos falling in each emotion category is relatively balanced, ranging from 290 to 669 pieces. Table II lists the number of words and emoticons belong to each emotion class in self-built emotion lexicon. Emoticon is a kind of text expression, like ($\wedge\ \wedge$) shows the happiness, $\top\_\top$ shows the crying. They usually directly express the emotion of viewers.

[3]We are happy to share this dataset to the public after the paper gets published.

The average occurrence frequency of the words in our dataset is about 1,550, which strongly validates their popularity in practice. And the average occurrence of emoticons in the dataset is about 8 times.

2) Datasets For Textual Analysis: In order to show that our EWE can be applied to other text-based emotion applications as well, we also use two additional text datasets for comparison.

- Incident reports dataset (ISEAR) [43]: ISEAR contains 7,000 incident reports obtained from an international survey on emotion reactions. A number of psychologists and non-psychologists, were asked to report situations in which they had experienced all of 7 major emotions (joy, fear, anger, sadness, disgust, shame, and guilt).
- Multi-Domain Sentiment Dataset [44]: This dataset contains four-domain (books, dvd, electronic and kitchen & housewares) reviews of productions from Amazons. It consists of 8,000 reviews, including 4,000 positive reviews and 4,000 negative reviews.

B. Baselines

We compare proposed DCVDN with other 14 baselines, which can be divided into four categories, i.e., visual-based, textual-based, multi-view learning and video classification methods. We also compare the proposed EWE with other textual emotion analysis baselines.

Visual-based baselines:

- GCH/LCH: use low-level features (64-bin global color histogram features (GCH) and 64-bin local color histogram (LCH)) as defined in [8].
- Caffenet: An ImageNet pre-trained AlexNet [45] followed by fine tuning.
- PCNN: Progressive CNN [11].
- VGGNet: We use two ImageNet pretrained VGGNets [42], VGGNet-16 and VGGNet-19, which have 16 layers and 19 layers respectively. We use the output of the layer fc-7 as the features.
- DenseNet: An ImageNet pre-trained DenseNet [46].
- ResNet: An ImageNet pre-trained ResNet [47]. We use the outputs of the last layer before the softmax as the features.
- GoogleNet: which is also called Inception Net [48]. We use the outputs of the last layer before the softmax as the features. It's also pretrained on the ImageNet.

Note that in all the above approaches, the image features of selected frames will be fed into LSTM for final classification.

Textual-based baselines:

- Lexicon method: We count the number of words belonging to each emotional class in each document. Then we choose the emotion class with the largest count as the result.
- eLDA: Aggregate all danmus of a video into one document and infer the emotion distribution of the document. Choose the emotion class with the largest probability as a result.
- Word embedding: Learn word representations by Skip-Gram model [17].
- Topical word embedding (TWE): Learn word representations by TWE model [20], which jointly utilizes the target word and its topic to predict context words.

Fig. 4. An example of the relation between burst points of danmus and selected key frames. The above frame sequence is achieved by even choosing, and the below frame sequence is achieved by our clustering method. The middle chart shows the change in the amount of danmus appearing in each second.

- SSWE [18]: Sentiment-Specific word embedding for sentiment classification.

Multi-view learning baselines:

- Simple-Con: Concatenate the features from different views.
- DistAE: A joint learning [31], the objective of which is a combination of reconstruction errors of autoencoders of different views and the average discrepancy across the learned projections of multiple views.

Video classification baselines:

- Conv3D [49]: 3D Convolutional neural networks.
- Temporal [50]: Optical flows of the frames in the video, widely used for actions recognition.
- Temporal + Spatial [50]: Use CNN to extract the spatial features, and average the temporal features and spatial features.

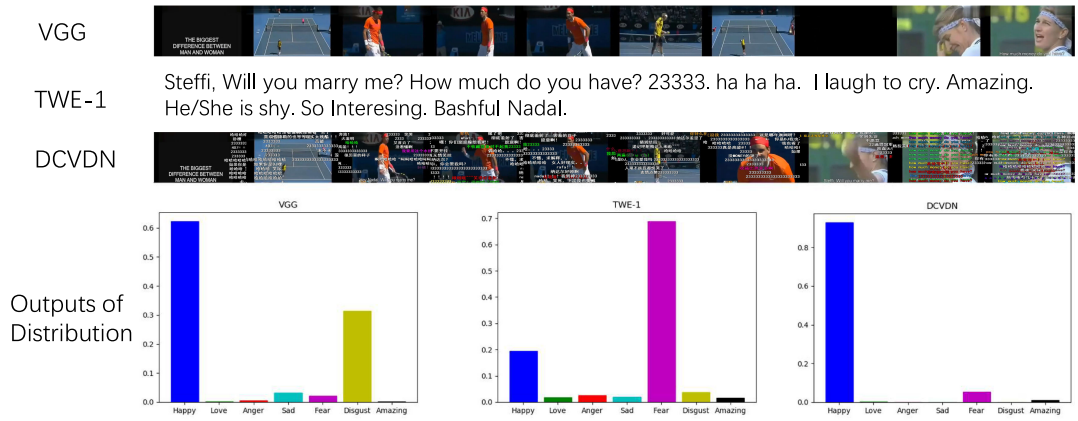
C. Parameter Settings

For the set of danmus in each video, we divide it into 10 clusters and aggregate each part into one document. While the number of clusters is a user-defined parameter, and 10 performs well in our experiments, thus we recommend it. In EWE, the dimensions of word vectors and emotion vectors are 128, therefore the dimensions of an emotional word and document embeddings are 256. The visual features are extracted by VGGNet fc-7, which results in 4,096 features. In multi-view learning module, the two autoencoders in the DCCAE are 3-layer, in which the size of the middle layer is 256 and the size of other layers is equal to the inputs. In the classification module, We use LayerNorm LSTM [51] here. The length of LSTM is set to 10, forget-bias is 0.1, and hidden layer size is 2,048. The following fully-connected network has 2 layers, with the size of the hidden layer as 4,096. We focus on *Accuracy* and *Precision* as the evaluation metrics in our experiments. The ratio of the training set to the validation set to testing set is $6:2:2$. For each experiment, we first split the dataset randomly. Then we train our model on the training set, stop the training according to its performance on the validation set and evaluate it on the testing set.

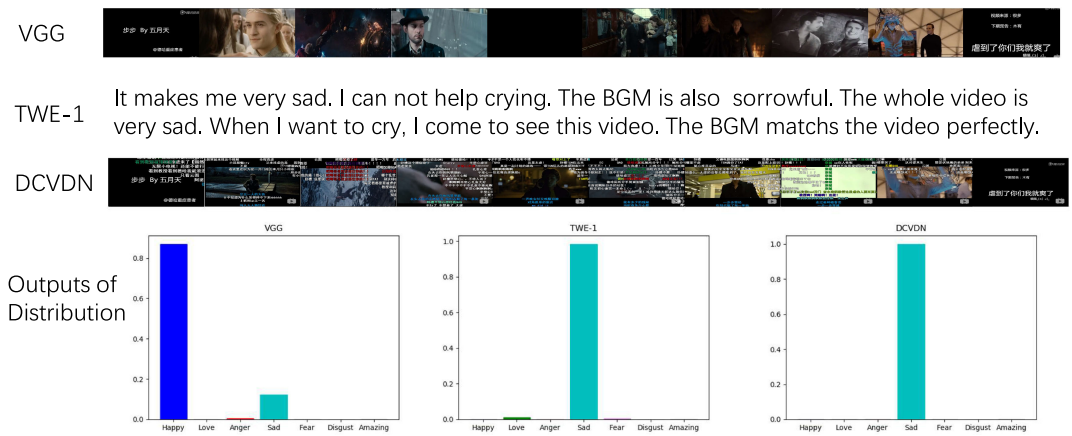
D. Case Studies

In this subsection, we first present one example to show that the keyframes are strongly related to the burst points of

(a) A wacky video with common visual content and scaring BGM. The label is "Disgust". VGG predicts Love, EWE predicts Fear and DCVDN predicts correctly.



(b) Two audiences made proposes to the tennis stars and got different responses. The label is "Happy". VGG and DCVDN predict correctly, but EWE considers it's "Fear".



(c) A combination of several movie clips with sad BGM. The label is "Sad". EWE and DCVDN predict correctly, but VGG considers it's "Happy".

Fig. 5. Three prediction examples to illustrate performance comparison between VGG, EWE and DCVDN.

danmus, then we present three prediction examples to validate the superiority of DCVDN over other baselines.

1) An Example of Key Frames and Danmus: We provide an example of the relationship between the burst points of danmus and the selected keyframes as shown in Fig. 4, in order to show that our clustering approach can select more important frames with the help of danmus. The most famous clip of this video is that an ancient minister just abused his opponent. "I've never seen anyone so brazen!" his opponent angrily said and died then. The above frame sequence is achieved by even chosen from the video, and the lower frame sequence is achieved by our clustering method based on danmu burst pattern. The middle

chart shows the change in the amount of danmus appearing in each second. Our method successfully finds the keyframe, which more comprehensively reflects the content background. It is also evident in the middle chart that the amount of danmus always change over time and the changes are strongly related to the audiences' interest. Our clustering method nearly selects the frame with most danmus in each time interval. Moreover, our method selects the frame with the most important words, "I've never seen anyone so brazen!". And as shown in the chart, the selected frame corresponds to the time moment with most danmus. By contrast, the uniform selection misses the frame, which is not always effective in practice.

2) Three Examples of Prediction Results: In the subsubsection, we provide three prediction examples to illustrate the performance comparison between VGG (visual method only), EWE (textual method only) and DCVDN (our method, jointly combining visual and textual information), as shown in Fig. 5.

Fig. 5(a) is a wacky music video with common visual content, however, the background music (BGM) sounds scaring. The ground truth of this video is "Disgust". DCVDN gives the right answer. VGG considers it's "Love" for it looks like a music video and most music video make us feel Love. The result of EWE is "Fear," the reason of which is that the audiences sometimes express their disgust via the adjective of fear in texts, such as "Help me! I died as soon as he sang. I choose to die". Fig. 5(b) is a video about two tennis stars, Nadal and Steffi. At the beginning of the video, one audience asked Nadal, "Will you marry me?". Nadal refused her shyly. Then, another audience asked Steffi, "Will you marry me?". Steffi asked back, "How much do you have?". Other audiences in the stands laughed loudly. The ground truth of example (b) is "Happy," as the audiences propose to the stars are very funny. VGG and DCVDN all give the right answer, probably because the video is about sports stars, while the probability of Disgust ranks second high in the result of VGG maybe due to the poor quality of videos. EWE considers the emotion of the video "Fear". This probably because word "shy" appears frequently in danmu texts and "Shy" is the subclass of "Fear" in our dataset. Fig. 5(c) is a music video combination clips from several movies, such as Harry Potter and the Lord of the Rings. These clips with sorrowful BGM are more about the parting or about death, and the ground truth is "Sad". Both of EWE and DCVDN give the right answer with almost 100% probability, while VGG considers it's "Happy" with high probability and it's "Sad" with low probability. These results are reasonable because the visual content is mainly about the movie stars, which looks "Happy" in most time. However, in these cases, danmu would give us more information about the true feeling of audiences, which is beyond the visual content.

E. Evaluations

1) EWE on the Emotional Analysis: We first compare our EWE model with textual-based baselines on our own dataset and two public datasets. Table III shows the comparison results with the texts in our own dataset. EWE outperforms all other textual-based baselines under investigation by 2.09% to 219.16% on *Accuracy*. The performance of the lexicon-based method and eLDA is pretty poor, which indicates that the relation between the number of emotional words and the emotion label of videos are not that strong. The embedding-based methods can perform much better, which can effectively capture the upper-level features in emotion space through highly non-linear transformations. EWE achieves the best performance, which is due to the fact that the emotional word embeddings are more informative and could provide more hints for emotion analysis. Table IV shows the comparison results on the ISEAR dataset. EWE performs more steadily than all other textual-based baselines with 10.92% to 167.57% improvement on *Accuracy*. EWE may perform even better if the training examples in the dataset is with more balanced distribution across different classes. Table V demonstrates the comparison results on the Multi-Domain Sentiment dataset, which only contains positive and negative labels. Besides the baselines investigated with the previous two datasets, we also include SSWE [18] as the baseline for the sentiment classification task. The performance of EWE is also the best one, with 1.40% to 28.91% improvement on *Accuracy*, although it's relatively not that outstanding like the performance on the previous two datasets. We notice that the

TABLE III
COMPARISON OF *PRECISION* AND *ACCURACY* BETWEEN EWE AND TEXTUAL BASELINES ON THE VIDEO-DAMU DATASET

Precision	Lexcion	eLDA	WE	TWE	EWE
Happy	0.106	0.493	0.568	0.644	0.636
Love	0.757	0.051	0.737	0.749	0.777
Anger	0.0	0.0	1.0	1.0	1.0
Sad	0.067	0.547	0.803	0.826	0.811
Fear	0.049	0.223	0.384	0.624	0.504
Disgust	0.094	0.659	0.554	0.647	0.630
Surprise	0.12	0.062	0.299	0.688	0.403
Accuracy	0.214	0.321	0.624	0.669	**0.683**

TABLE IV
COMPARISON OF *PRECISION* AND *ACCURACY* BETWEEN EWE AND TEXTUAL BASELINES ON ISEAR

Precision	Lexcion	eLDA	WE	TWE	EWE
Anger	0.131	0.274	0.258	0.234	0.309
Disgust	0.139	0.308	0.339	0.313	0.346
Joy	0.154	0.186	0.428	0.451	0.472
Shame	0.148	0.165	0.244	0.309	0.244
Fear	0.150	0.445	0.414	0.391	0.244
Sadness	0.161	0.359	0.437	0.515	0.488
Guilt	0.152	0.0	0.312	0.304	0.488
Accuracy	0.148	0.272	0.354	0.357	**0.396**

TABLE V
COMPARISON OF *PRECISION* AND *ACCURACY* BETWEEN EWE AND TEXTUAL BASELINES ON MULTI-DOMAIN SENTIMENT DATASET

Precision	positive	negative	*Accuracy*
Lexcion	0.510	0.516	0.512
eLDA	0.505	0.505	0.505
SSWE	0.560	0.624	0.613
WE	0.569	0.680	0.639
TWE	0.560	0.678	0.642
EWE	0.580	0.680	0.651

TABLE VI
COMPARISON OF *PRECISION* AND *ACCURACY* BETWEEN DCVDN-V AND VISUAL BASELINES ON THE VIDEO-DANMU DATASET

Precision	GCH	LCH	PCNN	CaffeNet	VGGNet16	VGGNet19	DenseNet	GoogleNet	ResNet	DCVDN-V
Happy	0.283	0.170	0.271	0.174	0.341	0.379	0.327	0.302	0.363	0.323
Love	0.361	0.355	0.355	0.405	0.577	0.491	0.429	0.622	0.667	0.463
Anger	0.833	0.573	0.938	0.963	0.209	0.346	0.926	0.650	0.960	1.0
Sad	0.364	0.384	0.440	0.541	0.465	0.575	0.466	0.445	0.345	0.616
Fear	0.359	0.438	0.438	0.481	0.552	0.481	0.316	0.395	0.298	0.609
Disgust	0.448	0.452	0.411	0.518	0.571	0.555	0.460	0.530	0.503	0.642
Surprise	0.219	0.343	0.346	0.272	0.0	0.343	0.320	0.0	0.233	0.338
Accuracy	0.410	0.394	0.423	0.455	0.438	0.454	0.438	0.445	0.472	**0.532**

TABLE VII
COMPARISON OF *PRECISION* AND *ACCURACY* BETWEEN DCVDN-V AND VIDEO CLASSIFICATION BASELINES ON THE VIDEO-DANMU DATASET

Precision	Conv3D	Temporal	T + S	DCVDN-V
Happy	0.275	0.313	0.308	0.323
Love	0.366	0.446	0.338	0.463
Anger	0.941	0.940	0.968	1.0
Sad	0.472	0.489	0.667	0.616
Fear	0.430	0.407	0.387	0.609
Disgust	0.417	0.433	0.524	0.642
Surprise	0.439	0.345	0.232	0.338
Accuracy	0.436	0.423	0.427	**0.532**

TABLE VIII
COMPARISON OF *PRECISION* AND *ACCURACY* BETWEEN DCVDN AND MULTI-VIEW LEARNING BASELINES ON THE VIDEO-DANMU DATASET

Precision	Simple-Con	DistAE	DCVDN
Happy	0.729	0.622	0.732
Love	0.816	0.782	0.754
Anger	1.0	1.0	1.0
Sad	0.795	0.805	0.814
Fear	0.632	0.571	0.716
Disgust	0.601	0.627	0.628
Surprise	0.442	0.652	0.450
Accuracy	0.720	0.713	**0.731**

performance of lexicon method and eLDA are bad, which may indicate that the quality of the sentiment dictionary is not good. This could adversely dampen the performance of EWE, which could be the possible reason for EWE not prominent as with the previous two datasets.

2) DCVDN-V on the Video-Danmu Dataset: We compare the visual part of our model, DCVDN-V, with other visual-based baselines and video classification methods on the Video-Danmu dataset. DCVDN-V is the reduced version of DCVDN solely considering visual input and using VGGNet-16 and autoencoder for feature extraction. Table VI shows the comparison results between visual-based baselines and DCVDN-V. Similarly, the *Precision* is counted based on each respective emotion class and the *Accuracy* is the overall average across all emotion classes. Basically, DCVDN-V outperforms other visual-based baselines by 12.71% to 35.03% on *Accuracy*. Moreover, the deep learning based methods generally achieve more or fewer improvements compared with the low-level feature-based approach. It is also worth pointing out that the *Precision* of "Happy" and "Love" predicted by visual-based methods is relatively lower than other classes compared with textual-based methods. The reason may be that the visual characteristics of "Happy" and "Love" videos are quite similar to each other so that the features may lead to great intra-class variance. This phenomenon strongly verifies that it is hard to learn a clear mapping function solely from visual features to high-level emotions. Therefore, with the enhancement by the interactive information from user-generated texts, joint features may achieve remarkable improvements compared with pure visual-based methods. As our dataset is based on videos, we also compare DCVDN-V with other video classification baselines, with the result shown in Table VII. The results show that the performance of different video classification methods are very close to each other, while DCVDN-V outperforms them significantly by 22.03% to 25.77% enhancement on *Accuracy*. This demonstrates that our method can learn more information related to emotion analysis rather than video classification methods. What's more, the outperformance of DCVDN-V in these experiments can show that the features if the videos are learned well with the help of the mutual information from the text part.

3) DCVDN on the Video-Danmu Dataset: Table VIII shows the comparison results between DCVDN and multi-view learning baselines, where the "T + S" means "Temporal + Spatial" in the first line of the fourth column. The proposed DCVDN with DCCAE surpasses other multi-view learning methods by 1.53% to 2.52% on *Accuracy*. The performance of DistAE sometimes is even worse than Simple-Con. This is because DistAE aims to minimize the distance between visual and textual views, however, they are not the same although they are somewhat correlated. By contrast, DCCAE provides the flexibility to dig deep about the relationship between different views so as to effectively facilitate joint representation learning. The whole results are shown in Table VIII can also justify that CCA is able to maximize the mutual information between videos and danmus.

4) Impact of the Size of Dataset: In this subsection, we show that the size of our dataset is large enough to learn a good model. We test the accuracy with different ratios of the size of our dataset. The accuracy is 67.8%, 71.3%, 72.7% respectively when the ratio of size is 0.4, 0.6 and 0.8, which can show that our dataset is large enough to prove the performance of our models.

V. CONCLUSIONS

In this paper, we studied user emotion analysis toward online videos by jointly utilizing video frames and danmu texts

simultaneously. To encode emotion into the learned word embeddings, we proposed EWE to learn text representations by jointly considering their semantics and emotions. Afterward, we proposed a novel visual-textual emotion analysis approach with deep coupled video and danmu neural networks, in which visual and textual features are synchronously extracted and fused to form a comprehensive representation by deep-canonically-correlated-autoencoder-based multi-view learning. To evaluate the performance of EWE and DCVDN, we conducted extensive experiments on public datasets and self-crawled video-danmu dataset. The experimental results strongly validated the superiority of EWE and the overall DCVDN over other state-of-the-art baselines.

REFERENCES

[1] J. Yuan, S. Mcdonough, Q. You, and J. Luo, "Sentribute: Image sentiment analysis from a mid-level perspective," in *Proc. 2nd Int. Workshop Issues Sentiment Discovery Opinion Mining*, 2013, Art. no. 10.

[2] J. Yang *et al.*, "Visual sentiment prediction based on automatic discovery of affective regions," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2513–2525, Sep. 2018.

[3] E. Ko and E. Y. Kim, "Recognizing the sentiments of web images using hand-designed features," in *Proc. IEEE 14th Int. Conf. Cogn. Inform. Cogn. Comput.*, 2015, pp. 156–161.

[4] F. Chen, R. Ji, J. Su, D. Cao, and Y. Gao, "Predicting microblog sentiments via weakly supervised multimodal deep learning," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 997–1007, Apr. 2018.

[5] S. Zhao, H. Yao, Y. Gao, R. Ji, and G. Ding, "Continuous probability distribution prediction of image emotions via multitask shared sparse regression," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 632–645, Mar. 2017.

[6] X. Lei, X. Qian, and G. Zhao, "Rating prediction based on social sentiment from textual reviews," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1910–1921, Sep. 2016.

[7] S. Liu *et al.*, "What makes a good movie trailer?: Interpretation from simultaneous EEG and eyetracker recording," in *Proc. ACM Multimedia Conf.*, 2016, pp. 82–86.

[8] S. Siersdorfer, E. Minack, F. Deng, and J. Hare, "Analyzing and predicting sentiment of images on the social web," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 715–718.

[9] Q. You, J. Luo, H. Jin, and J. Yang, "Building a large scale dataset for image emotion recognition: The fine print and the benchmark," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 308–314.

[10] D. Borth, T. Chen, R. Ji, and S.-F. Chang, "Sentibank: Large-scale ontology and classifiers for detecting sentiment and emotions in visual content," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 459–460.

[11] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 381–388.

[12] V. Campos, B. Jou, and X. Giro-i Nieto, "From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction," *Image Vis. Comput.*, vol. 65, pp. 15–22, 2017.

[13] H. Zhang and M. Xu, "Recognition of emotions in user-generated videos with kernelized features," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2824–2835, Oct. 2018.

[14] J. Wang, J. Fu, Y. Xu, and T. Mei, "Beyond object recognition: Visual sentiment analysis with deep coupled adjective and noun neural networks," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 3484–3490.

[15] E. Kouloumpis, T. Wilson, and J. D. Moore, "Twitter sentiment analysis: The good the bad and the omg!" in *Proc. Int. Conf. Weblogs Social Media*, 2011, pp. 538–541.

[16] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," in *Proc. Workshop Lang. Social Media (LSM 2011)*, 2011, pp. 30–38.

[17] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[18] D. Tang *et al.*, "Learning sentiment-specific word embedding for twitter sentiment classification," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 1555–1565.

[19] Q. You, J. Luo, H. Jin, and J. Yang, "Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia," in *Proc. 9th ACM Int. Conf. Web Search Data Mining*, 2016, pp. 13–22.

[20] Y. Liu, Z. Liu, T.-S. Chua, and M. Sun, "Topical word embeddings," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 2418–2424.

[21] H. Wang, J. Wang, M. Zhao, J. Cao, and M. Guo, "Joint topic-semantic-aware social recommendation for online voting," in *Proc. ACM Conf. Inf. Knowl. Manage.*, 2017, pp. 347–356.

[22] J. Reisinger and R. J. Mooney, "Multi-prototype vector-space models of word meaning," in *Proc. Human Lang. Technol., Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2010, pp. 109–117.

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[24] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," in *Proc. 18th ACM Conf. Inf. Knowl. Manag.*, 2009, pp. 375–384.

[25] F. Li, M. Huang, and X. Zhu, "Sentiment analysis with global topics and local dependency," in *Proc. 24th AAAI Conf. Artif. Intell.*, 2010, pp. 1371–1376.

[26] L. Pang, S. Zhu, and C. W. Ngo, "Deep multimodal learning for affective analysis and retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2008–2020, Nov. 2015.

[27] H.-R. Kim, Y.-S. Kim, S. J. Kim, and I.-K. Lee, "Building emotional machines: Recognizing image emotions through deep neural networks," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 2980–2992, Nov. 2018.

[28] Q. You, H. Jin, and J. Luo, "Visual sentiment analysis by attending on local image regions," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 231–237.

[29] A. Frome *et al.*, "Devise: A deep visual-semantic embedding model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2121–2129.

[30] J. Wang, W. Wang, and W. Gao, "Multiscale deep alternative neural network for large-scale video classification," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2578–2592, Oct. 2018.

[31] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *Proc. 32nd Int. Conf. Int. Conf. Mach. Learn.*, 2015, pp. 1083–1092.

[32] J. Ngiam *et al.*, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 689–696.

[33] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," 2014, *arXiv:1411.2539*.

[34] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[35] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.

[36] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 251–260.

[37] R. Socher and L. Fei-Fei, "Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 966–973.

[38] L. Gao, L. Qi, E. Chen, and L. Guan, "Discriminative multiple canonical correlation analysis for information fusion," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1951–1965, Apr. 2018.

[39] M. Wang, D. Cao, L. Li, S. Li, and R. Ji, "Microblog sentiment analysis based on cross-media bag-of-words model," in *Proc. Int. Conf. Internet Multimedia Comput. Service*, 2014, Art. no. 76.

[40] D. Cao, R. Ji, D. Lin, and S. Li, "A cross-media public sentiment analysis system for microblog," *Multimedia Syst.*, vol. 22, no. 4, pp. 479–486, 2016.

[41] Y. He, S. Xiang, C. Kang, J. Wang, and C. Pan, "Cross-modal retrieval via deep and bidirectional representation learning," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1363–1377, Jul. 2016.

[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*, 2014.

[43] W. G. Parrott, *Emotions in Social Psychology: Essential Readings*. Hove, U.K.: Psychology Press, 2001.

[44] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification," in *Proc. 45th Annu. Meeting Assoc. Comput. Linguistics*, 2007, pp. 440–447.

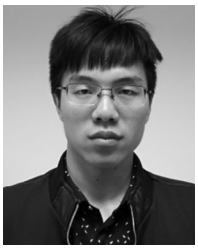
[45] V. Campos, A. Salvador, X. Giro-i Nieto, and B. Jou, "Diving deep into sentiment: Understanding fine-tuned CNNs for visual sentiment prediction," in *Proc. 1st Int. Workshop Affect Sentiment Multimedia*, 2015, pp. 57–62.

[46] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

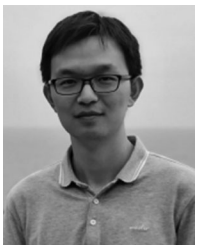
[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[48] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, San Francisco, CA, USA, 2017, pp. 4278–4284.
[49] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4489–4497.
[50] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.
[51] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

Chenchen Li received the B.E. degree from the Department of Software Engineering, East China Normal University, Shanghai, China, in 2014. He is currently working toward the Ph.D degree with the Department of Computer Science, Shanghai Jiao Tong University, Shanghai, China. His research interests include machine learning, data mining, and the data-driven methods for algorithmic game theory.



Jialin Wang received the B.E. and M.S. degrees from the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China, in 2015 and 2018, respectively. He is currently working as an Algorithm Engineer with Jingdong Group, Beijing, China. His current research interests include recommender system and representation learning.



Hongwei Wang received the Ph.D. degree from the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2018, and the B.E. degree from ACM Class, Shanghai Jiao Tong University, in 2014. He is currently a Postdoctoral Researcher with Stanford University, Stanford, CA, USA. His research interests include machine learning and data mining, particularly in graph representation learning mechanisms, algorithms, and their applications in real-world data mining scenarios.

Miao Zhao received the bachelor's and master's degrees from the Department of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan, China, and the Ph.D. degree from the Department of Electrical and Computer Engineering, State University of New York, Stony Brook, NY, USA, in 2010. Her current research interests include big data analytics, artificial intelligence, recommender systems, social networks, and multimedia analytics and networking.

Wenjie Li received the B.Sc. and M.Sc. degrees from Tianjin University, Tianjin, China, and the Ph.D. degree from the Department of Systems Engineering and Engineering Management, Chinese University of Hong Kong, Hong Kong. Before joining the Hong Kong Polytechnic University in 2001, she was a Postdoctoral Researcher with the Department of Systems Engineering and Engineering Management, Chinese University of Hong Kong. She is currently an Associate Professor with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong. Her research interests include natural language processing, text mining, social media analysis, information retrieval, extraction, and summarization.

Xiaotie Deng received the B.Sc. degree from Tsinghua University, Beijing, China, the M.Sc. degree from the Chinese Academy of Sciences, Beijing, China, and the Ph.D. degree from Stanford University, Stanford, CA, USA. He was with the University of Liverpool, City University of Hong Kong, and York University. He was an NSERC International Fellow with Simon Fraser University. He is cuurently a Professor of Computer Science with Peking University of China, Beijing, China. His current research focuses on algorithmic game theory with applications to Internet economics.