Video Emotion Recognition with Transferred Deep Feature Encodings

Baohan Xu[1], Yanwei Fu[*] [23], Yu-Gang Jiang[1], Boyang Li[3] and Leonid Sigal[3]
[1]School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, China
[2]School of Data Science, Fudan University, China
[3]Disney Research, USA
{bhxu14, ygj}@fudan.edu.cn, y.fu@qmul.ac.uk, {albert.li, lsigal}@disneyresearch.com

ABSTRACT

Despite growing research interest, emotion understanding for user-generated videos remains a challenging problem. Major obstacles include the diversity and complexity of video content, as well as the sparsity of expressed emotions. For the first time, we systematically study large-scale video emotion recognition by transferring deep feature encodings. In addition to the traditional, supervised recognition, we study the problem of zero-shot emotion recognition, where emotions in the test set are unseen during training. To cope with this task, we utilize knowledge transferred from auxiliary image and text corpora. A novel auxiliary Image Transfer Encoding (ITE) process is proposed to efficiently encode and generate video representation. We also thoroughly investigate different configurations of convolutional neural networks. Comprehensive experiments on multiple datasets demonstrate the effectiveness of our framework.

1. INTRODUCTION

Recognizing implicitly conveyed emotions in user-generated videos is an important yet often overlooked dimension of dimension of video understanding. Computational understanding of such emotions has many applications. For example, video recommendation services can benefit from matching users' interests with video emotion. An accurate understanding of video emotion can maintain consistency between emotions expressed in the main video and advertisements accompanying it, avoid social inappropriateness such as placing a funny advertisement alongside a funeral video.

Recognizing emotion from video, especially user-generated video, is challenging for the following reasons. First, due to close interaction between cognitive processes and emotional appraisals [29, 13, 14], human emotions are rich and complex. Recent research [2, 25] suggests that basic emotion categories, such as proposed by Ekman [11] and Plutchik [34], are merely modal responses, which cannot capture the full range of human emotion. Second, emotional expressions are sparse in videos. Of all frames in a video, only a small subset directly depict emotions. The rest of frames are needed, for instance, to set up the situation and introduce the context. Finally, user-generated videos are highly diverse. Compared to commercial content like movies and sports, user-generated videos cover a broader set of content and exhibit highly variable quality. Such intra-class variability creates difficulties for emotion recognition.

For the first time, we systematically study emotion recognition in user-generated video, specifically *supervised* [20] and *zero-shot* emotion recognition [42]. The zero-shot emotion recognition, where emotions in the test set are completely unseen during training time, is directly motivated by the variability of real-world emotions and insufficiency of basic emotion categories [2, 25]. To solve these tasks, a unified deep convolutional neural network (CNN) architecture is introduced to enable our encoding-based multi-instance learning framework, which transfers knowledge from auxiliary image and text data to better understand testing video data.

Our contributions are three-fold: (1) a novel auxiliary Image Transfer Encoding (ITE) process is proposed to efficiently encode and generate video representations; (2) we, for the first time, systematically and comprehensively investigate the effectiveness of features from different CNN architectures and layers in the task of video emotion recognition and knowledge transfer; and (3) we also explore the complementarity of deep features with the existing visual and audio hand-crafted features. The results show that our framework can significantly improve upon the previous state-of-the-art results [20] by 7.7% absolute percentage points on YouTube dataset. To our best knowledge, this is the first large-scale systematic study of video emotion recognition conducted by transferring deep feature encodings.

2. RELATED WORK

2.1 Psychological Theories of Emotion

Basic emotion theories claim existence of few universal emotion categories, each of which is associated with a set of prototypical facial expressions, physiological measurements, behaviors, and external causes. Ekman [11], for example, proposed six basic emotions including happiness, sadness, disgust, surprise, anger, and fear. However, recent find-

*Corresponding Author.

ICMR'16, June 06 - 09, 2016, New York, NY, USA

ISBN 978-1-4503-4359-6/16/06…$15.00

DOI: http://dx.doi.org/10.1145/2911996.2912006

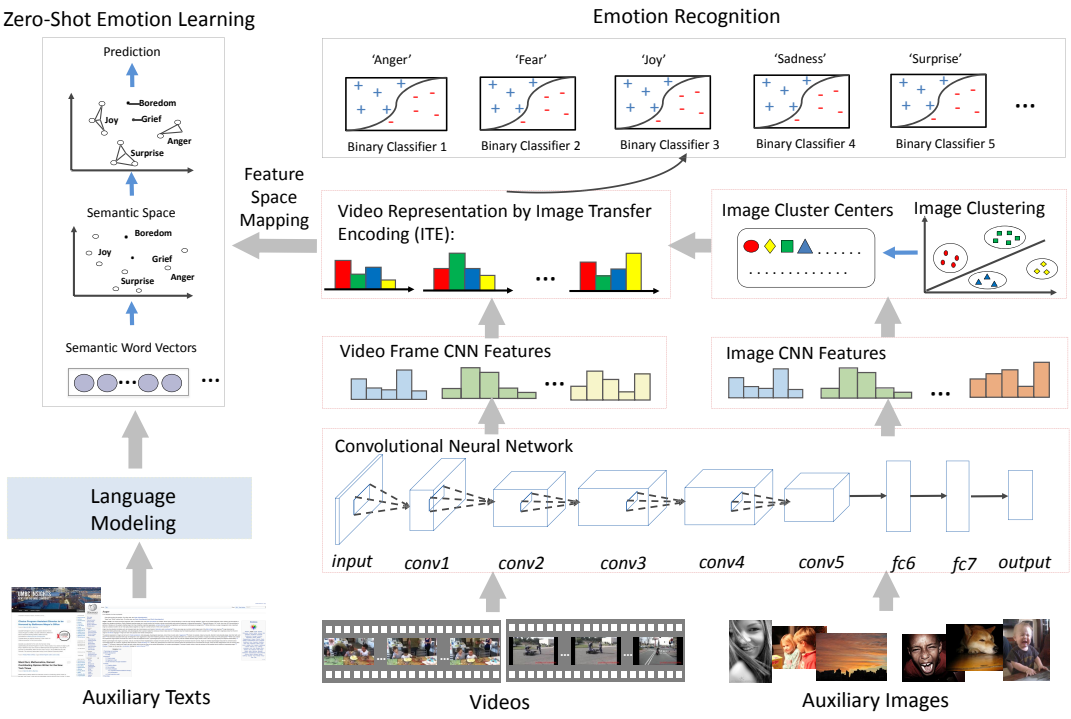

Figure 1: An overview of our framework. Information from a large text corpus is utilized for zero-shot emotion recognition as illustrated on the left. Auxiliary images (bottom right) are used to extract an emotion-centric dictionary, which help subsequently encode video (bottom middle) and recognize supervised emotion recognition (top right) and also enable zero-shot emotion recognition (top left).

ings [5, 2, 25] dispute whether these emotion categories are exhaustive, and suggest that among the diverse emotional landscape, the basic emotions are merely prototypical responses. Cognitive processes (which are needed for processing context) and emotional appraisal closely interact to create a diverse sets of emotions and affects [13, 14, 29], potentially leading to difficulties in labeling non-protypical emotions. Besides the traditional supervised recognition, we consider zero-shot emotion recognition, which allows us to recognize a large variety of emotion categories at test time without training examples.

2.2 Deep Visual Sentiment Analysis

In recent years, features from deep neural networks have been widely used for variety of tasks in computer vision and multimedia, *e.g.*, image categorization [6, 23] and object detection [35]. Promising results of such architectures in other domains inspired us to evaluate deep feature representations for video emotion recognition task. Further, we utilize auxiliary image information to the improve the effectiveness of the resulting recognition model.

Existing works explored emotion recognition from commercial movies [22, 39], animated images [21] and, to a lesser extent, user-generated videos [20]. Recently, several works on emotion recognition [7, 43, 47] also explored deep features extracted from CNNs, such as AlexNet [23]. Such deep features were shown to outperform hand-crafted low-level features and features from SentiBank [3]. In this paper, we perform a systematic layer-wise study of features from deep CNN architecture, and complementarity of such representations with hand-crafted features, in the setting of knowledge transfer and zero-shot learning.

2.3 Multi-Instance Learning

Multi-instance learning (MIL) is a particular form of learning where each input is a bag of multiple data vectors and only one class label is observable for all vectors. Most of early MIL approaches adapt single-instance supervised learning algorithms directly to multi-instance bags; examples include miSVM [1], MIBoosting [44], Citation-kNN [40], and MI-Kernel [15, 36]. Such approaches achieve satisfactory results in small or moderate-sized datasets but have difficulties with large-scale video data-sets due to the high computational cost. More recent algorithms (*e.g.*, CCE [48], Mi-FV [41], and MILES [8]) explore encoding multi-instance bags into single-instance representations to cluster the instances of all the bags to several groups. Inspired by these works, we encode multi-instance bags into video-level emotion-related representations. Different from the other methods we employ auxiliary sentiment image data to help the encoding procedure. Particularly, we study the role of various deep feature representations in such a MIL framework, as well as combination of such representations with other features (*e.g.*, audio) to improve performance.

3. PROBLEM FORMULATION

Figure 1 shows an overview of our framework. In this section, we formally define the video emotion recognition problem. We define a training video dataset as

$$Tr = \{(V_i, X_i, \boldsymbol{s}_i, z_i)\}_{i=1,\cdots,n_{Tr}},$$

where the i^{th} video V_i is a set of n_i frames $\{\mathbf{f}_{i,1}, \cdots, \mathbf{f}_{i,n_i}\}$, and each frame $\mathbf{f}_{i,j}$ has a feature vector $\mathbf{x}_{i,j}$. X_i is the set $\{\mathbf{x}_{i,1}, \cdots, \mathbf{x}_{i,n_i}\}$. $\boldsymbol{s}_i$ denotes a video-level feature of video V_i, obtained using auxiliary image transfer encoding, which is introduced in Sec. 3.1. $z_i \in Z_{Tr}$ is the class label from the set of training labels Z_{Tr} and n_{Tr} is the total number of training videos. The testing data set is likewise defined as

$$Te = \{(V_i, X_i, \boldsymbol{s}_i, \tilde{z}_i)\}_{i=1,\cdots,n_{Te}},$$

where n_{Te} is the total number of testing videos. For the purpose of knowledge transfer, we introduce a large-scale auxiliary image set, denoted as $A = \{(a_i, \boldsymbol{\phi}_i)\}_{i=1,\cdots,|A|}$, where $\boldsymbol{\phi}_i$ is the feature vector for an image a_i. Deep CNNs are used to extract both $\mathbf{x}_{i,j}$ and $\boldsymbol{\phi}_i$ from video frames and images.

An auxiliary text sentiment dataset is introduced here for zero-shot emotion recognition; particularly, textual data are represented as a sequence of words $W = (w_0, \ldots, w_{|W|})$, $w_j \in \mathcal{V}$ where the vocabulary $\mathcal{V}$ is the set of unique words. A $\mathcal{K}$-dimensional distributed word embedding $\boldsymbol{\psi}_w$ is learned for each $w \in \mathcal{V}$ by the skip-gram model [30].

3.1 Auxiliary Image Transfer Encoding (ITE)

We treat a video as a bag of video frames (in the MIL sense) and introduce Image Transfer Encoding for encoding videos as a BoW representation obtained using auxiliary image sentiment data. Note we do not use the clustering of instances from all training bags, since the video emotions are typically very sparsely expressed in only a few key frames.

We first compute D clusters from the auxiliary images by performing a spherical k-means clustering [18] on the auxiliary image dataset, which amounts to solving:

$$\min \sum_{i=1}^{|A|} (1 - \gamma_{i,d} \cos(\boldsymbol{\phi}_i, \boldsymbol{c}_d)), \tag{1}$$

where $\cos(\boldsymbol{\phi}_i, \boldsymbol{c}_d)$ is the cosine similarity between $\boldsymbol{\phi}_i$ and $\boldsymbol{c}_d$. The goal is to find D spherical cluster centers $\boldsymbol{c}_1 \ldots, \boldsymbol{c}_D$. The *responsibility* $\gamma_{i,d} = 1$, if an image a_i is assigned to the closet cluster center d (*i.e.*, $d = \text{argmax}_j \cos(\boldsymbol{\phi}_i, \boldsymbol{c}_j)$).

The cluster centers are then used to encode each video V_i into a single vector. Our BoW scheme translates the feature set X_i into a D-dimensional vector $\boldsymbol{s}_i = (s_{i,1}, \ldots, s_{i,d}, \ldots s_{i,D})$. Given the cluster centers $\{\boldsymbol{c}_1, ..., \boldsymbol{c}_D\}$, we identify K nearest cluster centers for each frame $\mathbf{f}_{i,j}$. The assignments $\nu_{i,j,d}$ are thus defined as

$$\nu_{i,j,d} = \begin{cases} 1 & if\ \boldsymbol{c}_d \in \text{KNN}(\mathbf{x}_{i,j}), \\ 0 & otherwise, \end{cases} \tag{2}$$

where $\text{KNN}(\mathbf{x}_{i,j})$ denotes the spherical K nearest neighbours[1] to $\mathbf{x}_{i,j}$ from the cluster centers. The feature vector $\boldsymbol{s}_i$ is computed as $s_{i,d} = \sum_{j=1}^{n_i} \nu_{i,j,d} \cdot \cos(\mathbf{x}_{i,j}, \boldsymbol{c}_d)$.

3.2 Supervised Emotion Recognition

The encoding scheme from the frame-level deep features to the video-level emotional representation helps the standard video emotion recognition task. Given a test video $V_k \in Te$, its class label can be estimated using

$$\hat{z}_k = \text{argmax}_z \mathcal{L}(\boldsymbol{s}_k \mid \boldsymbol{s}_{Tr}, z_{Tr}), \tag{3}$$

[1]Generally, we require that $K \gg 1$, since video frames can express much more 'versatile' emotions compared to images.

where $\mathcal{L}(\cdot)$ is the predictor trained from the video-level feature set S_{Tr} of the training video set Tr. We use the support vector machines (SVM) [10] classifier with chi-square kernel as the predictor $\mathcal{L}(\cdot)$.

3.3 Zero-Shot Emotion Recognition

Old emotion theories (*e.g.*, of Ekman [12]) only analyze a fixed number of prototypical emotions with relatively detailed textual explanations. In contrast, some very recent research [2, 25] questioned the validity of basic emotional categories and implied high variances of emotions far beyond several fixed basic categories. This naturally raises an interesting question: when the variances of emotions are big enough to be separated into a sub-emotion class, whether we can identify those emotions purely from their definitions.

Zero-shot emotion recognition predicts the emotions not observed in the training set. We relate the class labels in the training set, w_{Tr}, and those in the test set, w_{Te}, through an embedding space that (partially) captures the meaning of the label words. The embedding space maps a label word to a feature vector, and is obtained by training the *word2vec* model [30] on a large-scale textural corpus with significant emotion descriptions. Thus the emotion class z can be represented by a word vector $\boldsymbol{\psi}_z$. We use the embedding space as an intermediary between video features and emotion classes by training a regressor from the video feature space to the word embedding space:

$$g : \boldsymbol{s}_{Tr} \rightarrow \boldsymbol{\psi}_{w_{Tr}}, \tag{4}$$

where g is a support vector regressor with a linear kernel for each dimension of the word vector $\boldsymbol{\psi}_{w_{Tr}}$, similar to [24].

Given test video V_j, its class label $\tilde{z}_j$ can be estimated as

$$\hat{\tilde{z}}_j = \text{argmax}_{z \in Z_{Te}} \cos(g(\boldsymbol{s}_j), \boldsymbol{\psi}_z). \tag{5}$$

Note that Eq (5) intrinsically solves the problem of vector space classification; and $\boldsymbol{\psi}_z$ ($z \in Z_{Te}$) is the only available information for recognition. Thus to further improve the results, we propose Transductive 1-Step Self-Training (T1S) to adjust the word vector of new emotion classes. This strategy is a variant of Rocchio algorithm in information retrieval [28], which is a method for relevance feedback that works by using more relevant instances to update the query instances for better recall and possibly precision in vector space. Specifically, for a class $z \in Z_{Te}$ and the corresponding word vector $\boldsymbol{\psi}_z$, we compute a smoothed version $\bar{\boldsymbol{\psi}}_z$:

$$\bar{\boldsymbol{\psi}}_z = \frac{1}{K} \sum_{g(\boldsymbol{s}_i) \in \text{KNN}(\boldsymbol{\psi}_z), V_i \in Te}^{K} g(\boldsymbol{s}_i), \tag{6}$$

using a set of spherical K nearest neighbors to $\boldsymbol{\psi}_z$.

We empirically verify the semantic word vectors using emotion-based vector-oriented reasoning. Interestingly, we find that such reasoning is compatible with emotion theories such as [32]. For example, Vec("surprise") +Vec("sadness") is closest to Vec("disappointment"); Vec("joy") is very far from Vec("sadness").

4. FEATURES FROM DEEP NETWORKS

While convolutional neural networks gained popularity in emotion recognition, existing studies do not attempt to quantify or systematically study how CNN features affect the performance. For the problem of image categorization,

on the other hand, several works studied architecture design [23] and how to combine features across CNN layers [6]. Findings suggest that for image categorization deeper architectures tend to perform better [6] and that combining features across layers further improves the performance [17]. Yet, for some tasks, like texture recognition, deep learning features are not as effective and custom designed features or combinations are more effective [9]; for pose estimation [16] the $5th$ layer features tend to be more invariant to pose. These results indicate that studying the architectures and features within specific vision problem is important. In this section we conduct exhaustive and comprehensive study of various CNN architectures, feature combinations from various levels and combination of CNN features with hand constructed counterparts for the problems of supervised and zero-shot video emotion recognition.

4.1 Different Deep Architectures

Several popular deep convolutional architectures have been proposed for large-scale image classification tasks, including AlexNet [23], VGG-16, VGG-19 [6], and GooLeNet-22 [38]. **AlexNet** has seven layers where the first five are convolutional ($conv1 - conv5$) followed by 2 fully connected layers ($fc6 - fc7$). The fully connected layers can be represented by 4096 dimension features after ReLU, while convolutional layers ($conv1 - conv5$) are multidimensional arrays that represent convolution of the image with a learned filter; in practice they can be flattened in to d-dimensional feature vectors. Since filter sizes change with the layer, the dimensions of the feature representations at ($conv1 - conv5$) change as well. **VGG-16 and VGG-19** models [6] extend the AlexNet by expanding convolutional layers and have 16 and 19 layers respectively. **GoogleLeNet-22** is inspired by Hebbian principle with multi-scale processing and it has 22 layers. Nevertheless, these layers are still designed and optimized for image (esp. ImageNet) classification tasks; but not necessarily good for video emotion recognition tasks.

In the experimental section, we study the results of using these different deep convolutional architectures for video emotion recognition. Interestingly, while GoogleLeNet-22 is shown to be very effective for image recognition [38] and store-front classification [31], we find that it performs poorly on the emotion recognition problem.

4.2 Layer-wise Features of Deep Architecture

Rather than giving us a single feature representation, deep neural network is inherently a stacked structure which gives us a feature representation from each layer. One interesting phenomenon is that, from bottom to top layers of deep architectures, the features learned are from *general* to *specific*. For example, the first layer is known to learn the features that are similar to Gabor filters and color blobs. Such types of features are shown to be agnostic to the task, *i.e.*, they are *general*. In contrast, the higher-level layers are usually well trained for *specific* tasks, *e.g.*, image classification [46].

Most previous work on image sentiment analysis [7], [47] and [43] by default directly use the feature outputs of high-level layers, since the high-level semantics expressed in these high-level layers potentially are more related to image sentiment. Recently, [4] explored the layer-wise features on image sentiment dataset. However, video emotion is different from image sentiment analysis due to more diverse video content and more sparsely expressed video emotions. No previous work discussed how deep features should be used for video emotion recognition; not to mention the effects of layer-wise features and combinations.

We explore these questions in this paper. Particularly, we evaluate using $conv1 - conv5$ and $fc6 - fc7$ features from AlexNet [23]. The output of each layer is considered as visual descriptor of each frame. These experiments enable us to measure the difference in accuracy between layers and get intuition on their suitability for video emotion analysis.

4.3 Complementarity of Deep Features

As mentioned above, the deep architecture (from bottom to top) learns the features from *general* to *specific* with respect to a supervised classification objective. This notion raises another important question: the complementarity of deep features from various layers. To simplify discussion and isolate confounding factors, we evaluate these properties by using the direct concatenation of different layer features for video emotion recognition.

We also discuss complementarity of CNN with hand-crafted features. This is inspired by the recent study of using hand-crafted features for video emotion understanding [20]. Particularly, we use denseSIFT [27] as visual hand-crafted features. DenseSIFT method densely samples local frame patches rather than only use interest points in original SIFT. Then dense extracted SIFT descriptors are further encoded into a bag-of-words representation.

Audio hand-crafted features are also investigated, since human perception often relies on the use of multiple senses [37]: for example, videos that convey "joy" mostly contain laughter and "fear" may co-occur with screaming in the audio track. We utilize the well-known Mel-frequency cepstral coefficients (MFCC) as audio representation. An MFCC descriptor is computed over every 32 *ms* time-window with 50% overlap. The descriptors from the entire soundtrack of a video are converted to a bag-of-words representation using vector quantization.

5. EXPERIMENTS

In this section, we first introduce the experimental settings in Section 5.1, and then validate the effectiveness of our framework on supervised and zero-shot emotion recognition using the features from $fc7 - 7^{th}$ fully-connected layer in Section 5.2. Finally, the details of different deep architectures as well as the complementarity with hand-crafted features are investigated and compared for supervised video emotion recognition in Section 5.3.

5.1 Datasets and Settings

We utilize two video emotion datasets for evaluation: YouTube and Ekman. The Ekman dataset was collected from social mediate platform by us and will be made available to the community.

5.1.1 The YouTube emotion dataset

YouTube emotion dataset contains 1101 videos annotated with 8 basic emotions from the Plutchik's Wheel of Emotions [20]. To validate the zero-shot emotion recognition, we re-annotate the dataset with 'fine-grained' emotions. We create these more diverse emotion categories by adding 3 variations to each original emotion (24 emotions in total). For example, *anger* class is split into *annoyance*, *anger*, and *rage* along the arousal dimension according to Plutchik's wheel of emotions

dataset	MaxP	AvgP	Mi-FV	CCE	ITE($fc7$)
Y	34.5	41.1	38.4	30.2	**43.8**
E	39.0	48.4	36.4	31.5	**50.9**

Table 1: Supervised learning results reported on emotion recognition datasets. We use 2000 bins and $fc7$ features for our method. The two baselines use both linear and chi-square kernels.

[33]. We use **Y-8** (or just **Y**) and **Y-24** to indicate the original and re-annotated datasets respectively. Specifically, **Y-24** has 36 *anger*, 33 *annoyance*, 32 *rage*, 44 *anticipation*, 32 *interest*, 25 *vigilance*, 42 *boredom*, 64 *disgust*, 9 *loathing*, 12 *apprehension*, 79 *feat*, 76 *terror*, 23 *ecstasy*, 76 *joy*, 81 *serenity*, 27 *grief*, 11 *pensiveness*, 63 *sadness*, 29 *amazement*, 59 *distraction*, 148 *surprise*, 39 *acceptance*, 26 *admiration*, and 35 *trust* videos.

5.1.2 The Ekman-6 emotion dataset

According to Ekman's there are 6 basic emotions. The dataset is collected from social video-sharing websites (*e.g.*, YouTube and Flickr), resulting in 1637 videos for which those 6 emotions are annotated, with a minimum of 221 videos per class. The labels are annotated by 10 different volunteers who are unaware of the goals of the project. Each video was labelled by the majority voting result from at least 3 annotators.

5.1.3 Auxiliary images and texts

We use an auxiliary image dataset, a subset of $110K$ images of Adjective-Noun Pairs (ANPs) in Flickr image dataset [3] that have the top ranks (440 ANPs) with respect to the emotions[2]. The auxiliary text data has 7 billion words[3]. Most of the documents are about scientific articles and professional reports which have very strict definitions, descriptions and usage of the emotion and sentiment related words. To facilitate the efficient training of such large-scale corpus, we employ the *word2vec* model [30] which results in 4 million element vocabulary semantic space.

5.1.4 Experimental settings

Each video is uniformly sampled every 5 frame to reduce the computational cost. Our AlexNet model [23] is trained using 2600 ImageNet classes with the Caffe toolkit [19]. The auxiliary image data are clustered into 2000 clusters ($D = 2000$). The number of nearest neighbors K in Eq (2) is empirically set to 10% of the image clusters, which balances the computational cost with a good representation. For presentation simplicity, we use **Y**, **E** to represent the YouTube and Ekman-6 datasets respectively.

5.2 Video emotion recognition by $fc7$

In this subsection, we use the $fc7$ features of AlexNet for video emotion recognition, since $fc7$ is the most widely used deep feature (*e.g.*, the top layer feature) in most of the other computer vision tasks [23, 35].

[2]Please refer to Table 2 in [3].

[3]Composed of the UMBC WebBase data (3 billion words), the latest Wikipedia articles (3 billion words) and some other documents (1 billion words).

5.2.1 Supervised emotion recognition

To evaluate our encoding algorithm, we compare different video emotion recognition methods by using $fc7$ with the following baselines. **(1) MaxP.** Instance-level classifiers are trained to recognize instance labels of every testing video. The video class label is majority-voted by predicted instance labels [26]. **(2) AvgP.** It is a standard approach of aggregating, using an average, frame-level features into video-level descriptions (*e.g.*, [45]). **(3) Mi-FV.** It maps MIL bags of training videos into a new bag-level Fisher Vector (FV) representation, which efficiently deals with large-scale of data such as emotion datasets[41]. **(4) CCE.** [48] clustered the instances of all training videos into b groups. Each bag is re-represented by b binary features: assigning 1 to the i^{th} feature if one bag has instances falling into the i^{th} group and 0 otherwise. Linear kernel is used for Mi-FV and MaxP due to the large number of samples/dimensions, and the Chi-square kernel is used for others. We use 1-Vs-All strategy for multi-class classification.

ITE > MaxP/AvgP/Mi-FV/CCE. The result is reported in Table 1, which shows that the ITE method significantly outperforms the four methods on both datasets. The improvement of ITE over CCE and Mi-FV shows that using auxiliary image dataset to achieve knowledge transfer can create better video-level feature representations. This also support our hypothesis that most of frames are not closely related to video emotions. The worst performance comes from CCE. This might be because re-encoding process of CCE loses discriminative information gained from the deep learning network. The same training/testing split are used as in [20] on YouTube dataset. AvgP and ITE have much better results than Mi-FV and MaxP and thus we employ AvgP and ITE as the main comparison methods in following experiments. The AvgP result is comparable with the 41.9% reported in [20] of using all visual features, while our ITE results are much better. Note that the result of 41.9%±2.2% combines different types of hand-crafted visual features with the state-of-the-art multi-kernel strategy. In contrast, AvgP simply averages frame-level image features. This means that the performance of the $fc7$ features is comparable to those of multi-kernel combination of visual hand-crafted features.

Some qualitative results of supervised emotion predictions are shown in Figure 2. In the successful cases, testing videos share the common visual characteristics with auxiliary image dataset like the *bright light* and *smile face* in the "joy" category. The "anger" videos are wrongly classified as "fear". Comparing with "anger", the "fear" category is more highly correlated with *dark lightning* and *screaming faces* which are visually dominated in the failed case.

5.2.2 Zero-shot emotion recognition

Since Ekman dataset lacks sufficient variants (only 6 classes) of emotions, we conducted zero-shot emotion recognition on **Y-8** and **Y-24** dataset, which has more diverse emotion categories. **Y-8** uses *fear* and *sadness* as the testing classes. For **Y-24**, we randomly split dataset into 18 training and 6 testing classes with 5-fold repeated experiments. No testing class video instances are seen during training in zero-shot recognition tasks.

T1S > DAP. As a baseline for zero-shot recognition, we compare with Direct Attribution Prediction (DAP) which is

		DAP	Ours
	Chance	$fc7/fc6/conv5/conv4$	$fc7/fc6/conv5/conv4$
Y-8	50	51.5/ 53.04 / 48.05 /50.37	56.3 / **56.44** / 43.32 /53.55
Y-24	16.7	23.3 / 27.59 / 21.45 / 22.28	**32.6** /32.14 / 16.22 /27.76

Table 2: Zero-shot Learning on emotion dataset analysis. Video are only encoded by ITE since AvgP method can get very weak results which are slightly higher than chance and thus not considered here.

Figure 2: Qualitative results on supervised emotion prediction. The experiment uses $fc7$ features on Ekman dataset. The ground truth categories are at the top of each column; red labels indicate wrong predictions.

	VGG-16	VGG-19	GoogLeNet-22	AleNet
Y	44.7	44.0	35.6	41.1
E	49.3	48.8	38.3	48.4

Table 3: VGG and GoogLeNet architecture comparisons. The AvgP is used for reported results here.

proposed in [24] and is the most canonical algorithm used for zero-shot learning. For DAP, at test time each dimension of the word vectors of each test sample is predicted, from which the test class labels are inferred. DAP can be formulated as directly using Eq (5) without the word vector smoothing. Table 2 shows the results of each layer of deep architecture. We find that our method is much better than DAP when using the features of fully connected layers. The results improve DAP baseline by 4.9 and 9.3 absolute percentage points on $fc7$, which validate the effectiveness of our method.

fc6/fc7 > conv5/conv4. We further validate the zero-shot emotion prediction by using different types of features (*e.g.*, $fc6$, $conv5$ and $conv4$) as compared in Tab. 2. The results show that the features of fully connected layers ($fc6$ and $fc7$) are generally more favorable for zero-shot emotion recognition than those of convolutional layers ($conv4$, $conv5$). And the results of using convolutional features are only slightly higher than chance. If we compare the results

Figure 3: Key frames of two successful cases of zero-shot emotion recognition ($fc7$ features on Y-24): top row is a video about a bored boy walking and lying on the couch; the bottom row illustrated video of grief fans feel when their football team loses a game.

of the two dataset, we find that the results on Y-24 have a larger margin improvement than those on Y-8 for the same type of features. This means that finer-grained variant set of auxiliary emotions can enable better zero-shot learning.

In Figure 3, we show some successful examples of zero-shot emotion prediction. We highlight that even without any training examples of these categories, our method can still classify these video successfully using the encoded features. Considering the difficulty of zero-shot emotion prediction, our results are very promising.

5.3 Results of Validating Deep Architecture

5.3.1 *Different deep architecture*

VGG-16/VGG-19/AlexNet > GoogLeNet. While previous experiments showed satisfactory results on emotion analysis task by using AlexNet architecture, we want to compare different architectures to better understand deep feature encodings. VGG-16 and VGG-19 [6] and GoogLeNet-22 [38] achieved state-of-the-art performance for image classification on ImageNet challenge. Thus we conducted video emotion recognition using high layer features extracted from the two architecture as descriptors. Table 3 illustrates experimental results. We use $fc7$ of 16 and 19 layers VGG and $inception-5b$ of GoogLeNet. AvgP is used for all the deep architectures. The results of VGG-16 and VGG-19 are comparable to AlexNet, and outperform that of GoogLeNet-22. The result of VGG-19 is a little lower than VGG-16, which demonstrates that deeper networks may not be appropriate for the emotion recognition task. Although GoogLeNet gets promising results on image classification task, the lower results in Table 3 imply that GoogLeNet may not be the best choice for video emotion recognition.

5.3.2 *Layer-wise features of deep architecture*

fc6/fc7 > conv4/conv5. The results of the experiments on layer-wise features are reported in Table 5. Clearly features of fully connected layers significantly outperform those

	denseSIFT	MFCC	ITE($fc7$)	[ITE($fc7$), denseSIFT]	[ITE($fc7$), MFCC]	[ITE($fc7$), denseSIFT, MFCC]
Y	35.6	44.0	43.8	43.8	**52.6**	46.7
E	38.6	39.0	50.9	48.8	**51.2**	50.4

Table 4: Concatenated results of hand-crafted feature and deep features. ITE is computed from $fc7$.

Methods	ITE		AvgP	
Features	$fc7$	$fc6$	$fc7$	$fc6$
Y	43.8	**45.6**	41.1	42.0
E	**50.9**	49.4	48.4	48.7

Table 5: Layer by layer analysis results on emotion datasets. We use AvgP as the default video emotion recognition method. The results for convolutional layers $conv5-conv1$ are $22.5\pm2\%$ which are significant lower than those of fully connected layers.

Methods	ITE		AvgP	
Dataset	Y	E	Y	E
$[fc6, fc7]$	44.7	49.1	42.2	48.7
$[conv5, fc6, fc7]$	**45.1**	50.2	42.4	48.8
$[conv4, conv5, fc6, fc7]$	44.9	**51.2**	42.0	48.9

Table 6: Concatenation results of different layers of deep features in supervised learning setting.

of convolutional layers (which is $22.5 \pm 2\%$) by a large margin. This means that the features of convolutional layers ($conv5-conv1$) are too general to be discriminative enough for video emotion recognition; at the same time indicating that features of high-level layers contain more semantic information which can benefit video emotion understanding.

ITE>AvgP and $fc6 \sim fc7$. Inspired by the good performance of fully connected layers, we further report the results of using ITE encoding on $fc6$ and $fc7$ layers. And it also clearly shows that the ITE results are better than the AvgP of corresponding layer, which also validate the effectiveness of our framework. Nevertheless, the results of using $fc6$ features are generally comparable to those of using $fc7$ features in our experiments: the results of YouTube dataset are more favorable on $fc6$ features, while those results of Ekman dataset have better performance on $fc7$.

5.3.3 Feature Complementarity

We investigate the concatenation of different layer features in the deep architecture in Table 6. Specifically, we notice that (1) fully connected layers ($fc6$ and $fc7$) are generally complementary to each other. Both the concatenated features of $[fc6, fc7]$ for AvgP and ITE methods have better performance than those of only $fc6$ and $fc7$ respectively. (2) Fully connected layers are complementary to convolutional layers. This is shown by the results of $[conv5, fc6, fc7]$ of AvgP and ITE methods, which are better than those of $[fc6, fc7]$. (3) The results of convolutional layers are comparatively less complementary to each other. There is no significant improvement in accuracy when adding the features of $conv4$: the ITE result of $[conv4, conv5, fc6, fc7]$ is slightly worse than that of $[conv5, fc6, fc7]$ on YouTube dataset, due to the increased dimensionality (from less complementary $conv4$ layer features).

Table 4 reports concatenated results using ITE encoding and hand-crafted features. We normalized the different sets of features before concatenation. We find that (1) the concatenated results of visual features (denseSIFT) are still comparable to those of ITE on two dataset. This shows that deep features are less complementary to visual hand-crafted features. (2) The methods of using audio features can achieve very high accuracy for video emotion analysis. This means that audio track is very useful for video emotion recognition; (3) The audio hand-crafted features (MFCC) are very complementary to deep video features, since they come from different "sensors". (4) Concatenating all features has worse results than that of $[ITE(fc7), \text{MFCC}]$ due to the increased dimensions from weaker visual hand-crafted features.

5.3.4 Fine-tuning

We tried to fine-tune the networks to further improve results of video emotion recognition. The tuning data came from both training video frames or auxiliary image dataset. However, our experimental results suggested fine-tuning does not work well for video emotion recognition tasks. Even after 1 million iterations, the loss function still did not significantly decrease, and the deep features only marginally improve the final results ($\pm0.5\%$). Our fine-tuning does not work due to (1) The images of the same category may be in different emotion class, *e.g.*, we have 'adorable cat', 'crazy cat', 'lonely cat', 'ugly cat', *etc.*, which will confuse deep network which is trained from ImageNet classification data. (2) The noisy images further confuse the deep network. For example, 'terrible fire' class has both images of fierce fire and images of some fire trucks.

6. CONCLUSIONS

This paper, for the first time, provides the study of knowledge transfer for both supervised and zero-shot emotion recognition. Image Transfer Encoding (ITE) framework facilitates the creation of a representation conducive to the tasks of video emotion understanding. Deep architectures are also systematically explored for video emotion recognition tasks. We validate how different CNN architectures and layers are related to video emotion understanding, which can set the foundation for future research on video emotion analysis using deep features. Furthermore, we investigate the concatenation of CNN feature encodings and other hand-crafted features. Comprehensive set of experiments shows the effectiveness of deep features and their complementarity among layers and with audio features. Future work will address advanced fusion strategies on different deep features to further improve the recognition results.

7. ACKNOWLEDGEMENT

This work was supported in part by a National 863 Program (#2014AA015101) and a grant from the NSF China (#61572134).

8. REFERENCES

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, 2003.

[2] L. F. Barrett. Are emotions natural kinds? *Perspectives on Psychological Science*, 1(1):28–58, 2006.

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

[4] V. Campos, A. Salvador, X. Giro-i Nieto, and B. Jou. Diving deep into sentiment: Understanding fine-tuned cnns for visual sentiment prediction. In *ACM ASM*, 2015.

[5] J. M. Carroll and J. A. Russell. Do facial expressions signal specific emotions? Judging emotion from the face in context. *Journal of Personality and Social Psychology*, 70(2):205–218, 1996.

[6] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*, 2014.

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

[8] Y. Chen, J. Bi, and J. Z. Wang. Miles: Multiple-instance learning via embedded instance selection. *IEEE TPAMI*, 28(1):1931–1947, 2006.

[9] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

[10] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[11] P. Ekman. Universals and cultural differences in facial expressions of emotion. *Nebraska Symposium on Motivation*, 19:207–284, 1972.

[12] P. Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200, 1992.

[13] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Learning multimodal latent attributes. *IEEE TPAMI*, 36(2):303–316, 2014.

[14] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Transductive multi-view zero-shot learning. *IEEE TPAMI*, 37(11):2332–2345, 2015.

[15] T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola. Multi-instance kernels. In *ICML*. Morgan Kaufmann, 2002.

[16] A. Ghodrati, M. Pedersoli, and T. Tuytelaars. Is 2D information enough for viewpoint estimation? In *BMVC*, 2014.

[17] B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *CVPR*, 2015.

[18] J. A. Hartigan and M. A. Wong. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *CoRR*, 2014.

[20] Y.-G. Jiang, B. Xu, and X. Xue. Predicting emotions in user-generated videos. In *AAAI*, 2014.

[21] B. Jou, S. Bhattacharya, and S.-F. Chang. Predicting viewer perceived emotions in animated gifs. In *ACM MM*, 2014.

[22] H.-B. Kang. Affective content detection using hmms. In *ACM MM*, 2003.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[24] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE TPAMI*, 36(3):453–465, 2013.

[25] K. A. Lindquist, T. D. Wager, H. Kober, E. Bliss-Moreau, and L. F. Barrett. The brain basis of emotion: a meta-analytic review. *Trends in Cognitive Sciences*, 35(3):121–143, 2012.

[26] G. Liu, J. Wu, and Z. Zhou. Key instance detection in multi-instance learning. In *ACML*, 2012.

[27] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[28] C. D. Manning, P. Raghavan, and H. Schutze. *Introduction to Information Retrieval*. Cambridge University Press, 2009.

[29] S. Marsella and J. Gratch. EMA: A process model of appraisal dynamics. *Journal of Cognitive Systems Research*, 10(1):70–90, 2009.

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

[31] Y. Movshovitz-Attias, Q. Yu, M. C. Stumpe, V. Shet, S. Arnoud, and L. Yatziv. Ontological supervision for fine grained classification of street view storefronts. In *CVPR*, 2015.

[32] A. Ortony, G. Clore, and A. Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, 1988.

[33] R. Plutchik and H. Kellerman. *Emotion: Theory, research and experience. Vol. 1, Theories of emotion*. 1980.

[34] R. Plutchik, editor. *The Emotions*. University Press of America, 1991.

[35] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *ICLR*, 2014.

[36] K. Sikka, A. Dhall, and M. Bartlett, Weakly supervised pain localization using multiple instance learning. In *IEEE FG*, 2013.

[37] B. E. Stein and T. R. Stanford. Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neuroscience*, 9(4):255–266, 2008.

[38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CoRR*, 2014.

[39] H.-L. Wang and L.-F. Cheong. Affective understanding in film. *IEEE TCSVT*, 16(6):689–704, 2006.

[40] J. Wang and J.-D. Zucker. Solving the multiple-instance problem: A lazy learning approach. In *ICML*, 2000.

[41] X.-S. Wei, J. Wu, and Z.-H. Zhou. Scalable multi-instance learning. In *ICDM*, 2014.

[42] B. Xu, Y. Fu, Y.-G. Jiang, B. Li, and L. Sigal. Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization. *CoRR*, 2015.

[43] C. Xu, S. Cetintas, K.-C. Lee, and L.-J. Li. Visual sentiment prediction with deep convolutional neural networks. *CoRR*, 2014.

[44] X. Xu and E. Frank. Logistic regression and boosting for labeled bags of instances. In *PAKDD*, 2004.

[45] Z. Xu, Y. Yang, and A. G. Hauptmann. A discriminative CNN video representation for event detection. *CoRR*, 2014.

[46] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *NIPS*, 2014.

[47] Q. You, J. Luo, H. Jin, and J. Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI*, 2015.

[48] Z.-H. Zhou and M.-L. Zhang. Solving multi-instance problems with classifier ensemble based on constructive clustering. *Knowledge and Information Systems*, 11(2):155–170, 2007.