Frame-Transformer Emotion Classification Network

Jiarui Gao[1], Yanwei Fu[2], Yu-Gang Jiang[1], and Xiangyang Xue[12*]
[1]School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing,
Fudan University, China
[2]School of Data Science, Fudan University, China
{jrgao14,yanweifu,ygj,xyxue}@fudan.edu.cn

ABSTRACT

Emotional content is a key ingredient in user-generated videos. However, due to the emotion sparsely expressed in the user-generated video, it is very difficult to analayze emotions in videos. In this paper, we propose a new architecture–Frame-Transformer Emotion Classification Network (FT-EC-net) to solve three highly correlated emotion analysis tasks: emotion recognition, emotion attribution and emotion-oriented summarization. We also contribute a new dataset for emotion attribution task by annotating the ground-truth labels of attribution segments. A comprehensive set of experiments on two datasets demonstrate the effectiveness of our framework.

KEYWORDS

Video emotion recognition; video emotion attribution; video emotion-oriented summarization and spatial-transformer network

ACM Reference format:
Jiarui Gao[1], Yanwei Fu[2], Yu-Gang Jiang[1], and Xiangyang Xue[12]. 2017. Frame-Transformer Emotion Classification Network. In *Proceedings of ICMR '17, Bucharest, Romania, June 06-09, 2017,* 6 pages.
DOI: http://dx.doi.org/10.1145/3078971.3079030

1 INTRODUCTION

The explosive growth of user-generated videos creates a great demand for computational understanding of visual media data. Great efforts and success have been made on the video content understanding, such as video actions and activities. Sentimental analysis on text data [10] and image data [4] have been studied recently. However, the ability to understand emotions from videos, to a large extent, remains an unaddressed problem, despite the fact that video content can convey strong emotional information to their viewers. Computational understanding of the emotions aroused by video content nevertheless has many real-world applications. For example, video recommendation services can benefit from matching users' interests with the emotions of video content.

The challenges of video emotion understanding come from three aspects. Firstly, the emotions are often sparsely expressed by a subset of the video, while the other parts of the video perform as the story context for video emotion. Secondly, there are several different types of emotions in one video, despite a single dominant emotion exists. It is essential to know, and yet difficult to answer which video segment contributes the most to the video's overall emotion, which is defined as video emotion attribution [24]. Thirdly, the user-generated videos are often captured in an uncontrolled environment and of high diverse content. The unconstrained space of objects, scenes, and events in user-generated videos, not only makes their content very complex to be analyzed; but also gets the user-generated videos more likely to suffer from the problems of occlusions of objects and illumination conditions than the commercial videos (e.g. movies, news and sports).

Previous efforts of understanding video emotion aim at solving the three tasks of emotion recognition, emotion attribution and emotion-oriented summarization, by either combining various of low-level and middle-level features [7]; or by using auxiliary image sentiment dataset to re-encode the features of video frames [23, 24]. These works still have to design a specific algorithm for each task, rather than a single framework solving all these three highly correlated tasks jointly.

In this paper, we propose our new architecture – Frame-Transformer Emotion Classification Network (FT-EC-net), which facilitates solving emotion recognition, emotion attribution and emotion-oriented summarization jointly. FT-EC-net is composed of FT-net and EC-net. Particularly, the FT-net is a variant of spatial transform networks (ST-net) [6]. It can learn to detect the emotional video segments; and thus facilitates both emotion attribution and emotion-oriented summarization. The EC-net is a classification network which further processes the results of FT-net for emotion recognition. Our architecture is thoroughly evaluated on Ekman6 and Emotion6 video dataset.

Contributions: (1) Our newly proposed FT-EC-net can solve the emotion recognition, emotion attribution and emotion-oriented summarization simultaneously. As a variant of ST-net, FT-net is firstly introduced in this paper to enable video emotion attribution and emotion-oriented summarization; (2) In establishing a good benchmark for emotion attribution task, we re-annotate the Ekman6 dataset with the most emotion-oriented segments which can be used as the ground-truth for emotion attribution task. (3) We also introduce a new evaluation metric to evaluate the video segments detected in attribution tasks.

2 RELATED WORK

Image and Video Emotion Recognition. Recently, inspired by the psychological theory, such as Ekman's six pan-cultural basic emotions and Plutchik's wheel of emotions, researchers have studied the problem of image recognition. Various features have been explored, such as the features inspired by psychology and art theory[15], and the shape features [13].

[*] Yanwei Fu is the corresponding authour.

ICMR '17, Bucharest, Romania
© 2017 ACM. 978-1-4503-4701-3/17/06…$15.00
DOI: http://dx.doi.org/10.1145/3078971.3079030

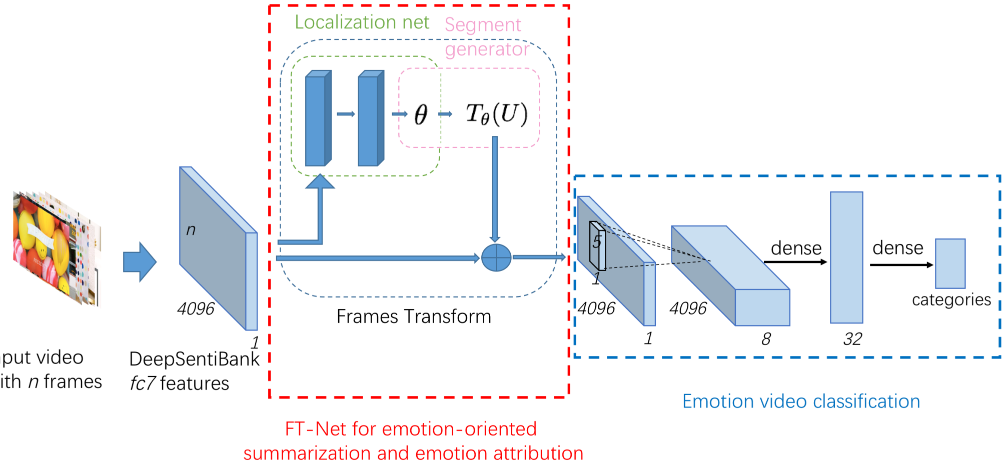

Localization net
Segment generator
θ
T_θ(U)
Input video with n frames
n
4096
1
DeepSentiBank fc7 features
Frames Transform
5
1
4096
1
4096
8
dense
32
dense
categories
Emotion video classification
FT-Net for emotion-oriented summarization and emotion attribution

Figure 1: An overview of our framework. We use the DeepSentiBank to extract the features of each frame. With the extracted deep features, the localization network is thus regressed towards the parameters θ. The FT-EC-Net uses the extracted $fc7$ features as inputs, and is capable of emotion recognition (right part), emotion-oriented summarization and emotion attribution (middle part).

Recently with the renaissance of convolution neural networks, deep convolutional neural networks have been used for visual sentiment analysis [2, 25]. A large scale of visual sentiment dataset was proposed in Sentibank [2] and DeepSentiBank [4]. It contains a set of 1,533 adjective-noun pairs, such as "cute dog" and "happy wedding".

Video emotion recognition has been investigated recently and still mainly focused on different types of features, such as the low-level visual and audio features as well as attribute features in [7]; the mid-level audio-visual features in [1]. Emotions have also been analyzed in GIF files [8] which can be taken as one type of short videos. The facial expressions have also been investigated on videos [5]. For a more recent survey, we refer to [21].

Most these existing works still formulate emotion understanding as a classification task. In our previous work [23, 24], we propose another two tasks – video emotion attribution and video emotion-oriented video summarization. Thus different from all previous works, we propose a new architecture that is able to solve these three tasks jointly.

Emotion attribution and emotion-oriented summarization. Video summarization has been explored for more than two decades [19] and the detailed review is beyond our scope. In general, the video summary has two forms, i.e., keyframes extraction and video skims; and to generate video summary, a set of features have been exploited, such as visual saliency[14], motion cues [12] and mid-level features [20], and semantic recognition [22].

Recently, we [24] introduced the tasks of emotion-oriented summarization which extracting video summarization according to more general video emotion content. Inspired by the task of semantic attribution in text analysis, [24] also defined the emotion attribution as attributing the video's overall emotion to its individual segments. However, [24] still treated the emotion recognition, summarization and attribution tasks as several separate tasks. Intrinsically, the emotion recognition can greatly help emotion attribution and emotion-oriented summarization; and the emotion-oriented summary can be selected from the results of emotion attribution. Thus, our framework can solve these three tasks simultaneously.

Spatial transform networks. Spatial transform networks (ST-net) are firstly proposed in [6] for image classification. ST-Net provides the spatial transformation capabilities[6], which enables a wide variety of tasks such as co-localization [18], and spatial attention [17]. Our FT-net component is a variant of ST-net and it enables selectively learning to segment the emotional video segments.

3 FRAME-TRANSFORMER EMOTION CLASSIFICATION NETWORK

As illustrated in Fig. 1, this section introduces our framework which is composed of three parts: DeepSentibank, Frame-Transformer subnetwork (FT-net) and Emotion Classification subnetwork (EC-net).

Suppose we have n frames extracted from each video. The DeepSentiBank [4] is utilized to extract the features of each frame. It contains five convolutional layers and three fully-connected layers. Specifically, we use as features the $4096-dim$ output of $fc7$ layer of DeepSentiBank. The DeepSentiBank is trained on 2089 Adjective Noun Pairs(ANPs) (such as "sad eyes") with 867919 images. In practice, a particular number of frames are equally sampled from each video to formulate the input of the network.

The FT-net aims at learning to detect emotional video segments. It has the localization sub-network regressing the transformer parameter θ and normalized segment generator of partitioning the video segments from the data stream generated by DeepSentiBank.

The segments generated by FT-net can be used for emotion attribution and emotion-oriented summarization.The EC-net is constructed for video emotion recognition, with one convolutional layer and two fully connected layers. We will explain each part in the next subsections.

3.1 FT-net

The FT-Net can be further processed into localization sub-network and normalized segment generator.

Localization sub-Network. It has two fully connected layers to further project the extracted features of each frame into a non-linear representation. The localization network in Fig. 1 aims at regressing the sampling parameter $\theta = [\theta_1, \theta_2]$ for normalized segment generator. The parameter θ is calculated through backpropagation from the EC-net; thus no supervision is provided to the localization network.

Normalized Segment Generator. To represent the specific output frames, we set x_i^t as the target coordinate of each frame in the regular output segment and x_i^s as the source input coordinate of each frame along the input frames n. The output segment S is formed by L frames, *i.e.*, $S = \left[x_i^t\right]_{i=1}^{L}$. The segment generator enables projecting the emotional frames (with coordinate x_i^t, $i = 1, \cdots, L$) to the corresponding source frames x_i^s by the 2D affine transformation,

$$x_i^s = \begin{bmatrix}\theta_1 & \theta_2\end{bmatrix} \begin{pmatrix} x_i^t \\ 1 \end{pmatrix} \tag{1}$$

where θ_1 is the scale parameter which controls the length ratio in the input video projected from output segments; and θ_2 is the translation parameter which indicates the offset of the segment over the video. The reason why the computation above is backward (from x_i^t to x_i^s) is that it enables the length of the source segment from the original video to be unfixed. But this network can only detect one main segment due to its structure. In practice, to avoid the numerical issues, the coordinates x_i^t and x_i^s are normalized to $(-1, 1)$ $(-1 \leq x_i^t, x_i^s \leq 1)$.

This parameter θ is computed through the localization subnetwork according to the same input features as it is applied, so the output segment S can be taken as the attribution results of the original video. We can further generate the summary by the attribution results.

3.2 EC-net

The output of normalized segment generator is further processed by classification sub-network.

One Convolutional layer has 8 convolutional filters with the size of 5 × 1. Intuitively, the $fc7$ features extracted by DeepSentiBank can be taken as a generic extractor for sentimental related features. Nevertheless, the FT-net does not have a principle way to model the contextual information among the consecutive frames, which however is extremely important for our task of emotion analysis. Thus this layer is utilized to compute contextual information among each feature channel.

Two fully connected layers have 32 neurons individually. The final softmax layer is followed to classify the emotions. Due to the difficulty of emotion analysis task, the fully connected layers are added to increase the non-linearity of the classification network.

Note that the convolutional layer follows the fully connected layers in DeepSentiBank and FT-net, which however do not have negative effects on learning the parameters of convolutional layers. (1) The DeepSentiBank is pre-trained beforehand and its parameters are fixed when we extract the $fc7$ features of frames. (2) The FT-net has two fully connected layers to regress the parameter θ in its localization subnetwork. Once the θ is learned to segment the input features, our convolutional layer still processes the data extracted by DeepSentiBank as shown in Fig. 1.

3.3 Our tasks

A very nice property of our network is that it can enable the emotion recognition, emotion attribution and emotion-oriented summarization simultaneously. Specifically,

Emotion recognition. The final softmax layer of EC-net is to identify the emotion. It thus outputs the likelihood of one video belonging to which type of emotion.

Emotion attribution. The video segments of high emotion scores, i.e. emotion attribution, can be directly computed by using the θ values computed by FT-net. With the well trained FT-EC-net and θ computed, the emotion attribution segment is computed as follows. Suppose the length of candidate video is l, the selected segment starts at t_s and ends at t_e;

$$t_s = \frac{l}{2} \cdot (\theta_2 - \theta_1 + 1) \tag{2}$$

$$t_e = \frac{l}{2} \cdot (\theta_1 + \theta_2 + 1) \tag{3}$$

The Eq (2) and Eq (3) are derived from Eq (1).

Emotion-oriented summarization. By utilizing the results of video attribution, the summary of video skims can be directly generated from the selected segments. We also uniformly sample the frames/video clips from selected segments as the keyframes summary/video skims.

4 EXPERIMENTS

4.1 Dataset

We conduct experiments on two video emotion datasets.

Emotion6 video dataset. Inspired by [6], we augment the Emotion6 dataset [16] and create Emotion6 video dataset as the testbed for our tasks. Emotion6 dataset consists of 6 basic emotions. Each image has been labelled with a distribution of all these emotions as well as one domaint emotion. To construct the dataset, we randomly crawled an auxiliary image dataset online which has no strong evoked emotions and uses as the noise set. The frames of each video are selected either from Emotion6 dataset with the dominant evoked emotion, or from the auxiliary image set. The corresponding emotion labels from Emotion6 dataset are used as the ground-truth labels for the generated videos.

Ekman6 dataset. Ekman6 dataset is firstly collected by [24]. It contains 6 different types of emotions; totally 1637 videos with around 220 videos per class. In this paper, two largest video classes, namely, anger and surprise, are employed for the evaluation. To further evaluate the tasks of emotion-oriented video summarization and attribution, the videos of these two classes are annotated with the most emotion-oriented segments. Specifically, for each video, three annotators are invited to label the key segments which contribute the most to the overall emotion label of this video. The overlapped segment labels beween two annotators are thus saved as ground-truth segments. Finally, we obtain $1 - 3$ emotion related segments of each video. We will release these annotations upon the acceptance.

Features. DeepSentibank [4] is utilized to extract the features of the $fc7$ layer of each frame. DeepSentibank is initialized with the weights trained from ImageNet and fine-tuned on the Sentibank dataset.

4.2 Competitors

Our model is compared against the following models.

SVM by Majority Voting (SVM) . The $fc7$ features of Emotion6 training set are used to train a linear SVM classifier. The emotion scores can be thus predicted on each individual frame of testing videos. The final classification labels are voted majoritively; and the attribution/summarization results are obtained by selecting the segments with the highest emotion scores.

Image Transfer Encoding (ITE). Our model is compared with the state-of-the-art method – ITE [24]. Particularly, we use the same setting of ITE as [24]: we cluster 2000 centers from the emotion-rich auxiliary image dataset; and by using these 2000 centers as bins, we encode the frames of one video into video-level representation; a linear SVM model is trained to solve the emotion recognition task. The attribution/summarization can also be solved by selecting the frames that have the smallest distances to video-level representation.

Convolutional neural networks (CNN). In video emotion classification, our model is compared with convolutional neural networks. We remove the FT-net from our framework and get a pure CNN structure capable for emotion recognition task.

4.3 Settings and Evaluation

Emotion recognition. The predicted labels are compared against the ground truth labels to calculate the classification accuracy.

Emotion Attribution. We employ the evaluation metrics – mean Average Precision (mAP) for video detection [3].

Emotion-oriented summarization. User study is employed to quantitatively compare our results against the competitors.

Parameter Settings. We empirically set the initial value of $\theta = [0.2, 0]$; Specifically we set the bias of the second fully connected layer in localization network to be $[0.2, 0]$. The models are trained for 1000 epochs with the batch size of 40 on Emotion6 video dataset, and the batch size of 20 on Ekman6 dataset. Dropout is employed here on all fully connected layers and the ratio is set as 0.75. A softmax layer is utilized to calculate loss and Adam[9] is used to accelerate learning. Our codes are implemented on tensorflow. For each dataset, the model is repeatedly trained from 5 times to reduce the variance; and the averaged performance is reported.

Preprocessing of dataset. On Emotion6 video dataset, we uniformly sample 30 frames for each video. Due to the fact that videos in Ekman6 dataset are slightly longer, 100 frames are uniformly sampled for each video. Zero-value frames are added if the total frames of one video is less than 100 .

Table 1: Emotion recognition results.

Dataset	Chance	SVM	ITE	CNN	Ours
Emotion6 video	16.7%	80.0%	77.5%	81.3%	**82.2%**
Ekman6	50.0%	60.1%	67.0%	74.0%	**74.4%**

4.4 Results on Emotion Recognition

We perform the emotion recognition on two datasets. The results are reported in Tab. 1. We compare with SVM, ITE, and CNN methods. The experimental results show that our framework outperforms all the other baselines on both datasets. This validates the effectiveness of our methods.

We also observe that, (1) comparing with CNN methods, our model has 0.9 and 0.4 absolute percentage improvement thanks to the FT-net which can attribute the emotional segments. (2) The improvement of our model over ITE and SVM shows that by integrating the attribution results, our EC-net can get better emotion classification results. (3) The worst performance comes from SVM. This might be due to the fact that emotion is intrinsically sparsely expressed in the video; and thus the trained linear SVM classifier is not reliable to predict confident emotion scores of frames.

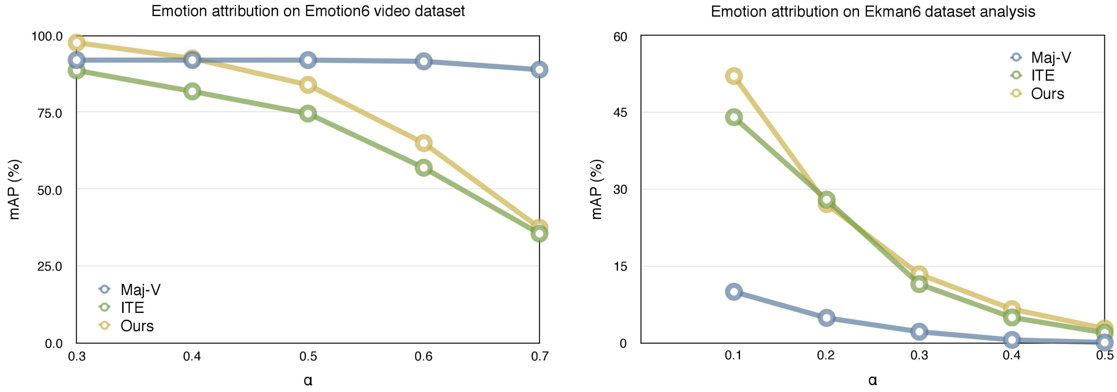
4.5 Results on Emotion Attribution

We conduct the experiments on emotion attribution. Specifically we compare with the methods of ITE and SVM. For SVM method, the emotion segments can be sampled from the longest majority-voted video emotion segments.

The mAP score is introduced here to evaluate the performance of emotion attribution. To compare the mAP, we have to decide whether one video detection is true positive or not. In particular, we firstly calculate the overlap between the prediction of temporal video segment and the ground-truth segment. This overlap is measured by the temporal intersection over union ($tIou$) score of predicted and ground-truth segment. The prediction is marked as positive if the overlapping is greater than a threshold α. The mAP score can thus be computed.

We compare the emotion attribution results on Ekman6 and Emotion6 video dataset in Fig. 2. We vary the threshold α on these two datasets. On Ekman6 dataset, our results are comparable or better than those of ITE method. This validates that our FT-net network can actually help identify the video segments that contribute the most to the overall emotion of one video. Furthermore, note that the good results of FT-net also benefit greatly from the recognition task of EC-net, since FT-net is basically learning to sample the video segments by the classification results. On the Emotion6 video dataset, our results still consistently outperform those of ITE method. On this dataset, we also notice that the SVM method has very good and stable performance; this is because our SVM classifier is exclusively trained on Emotion6 images. However, when we empirically validate training a SVM predictor which is

Figure 2: Emotion attribution results. We report the mAP scores for each dataset.



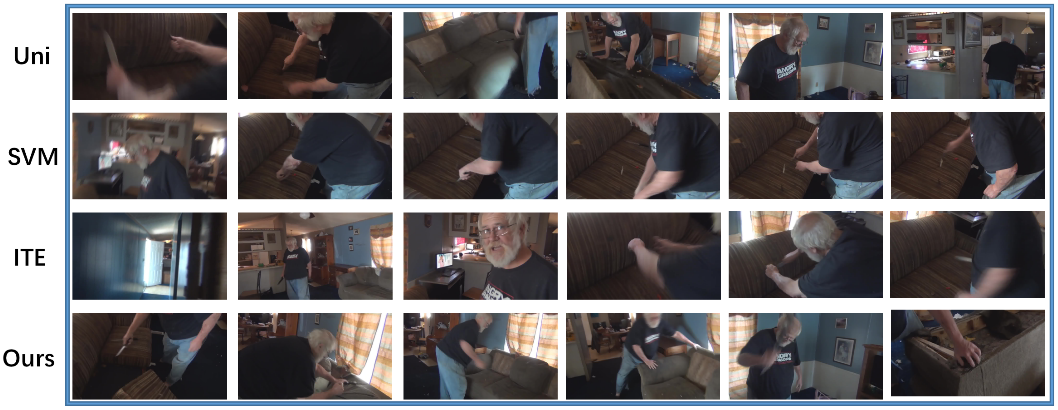


Figure 3: The qualitative results of emotion-oriented video summarization. We compare several baselines.

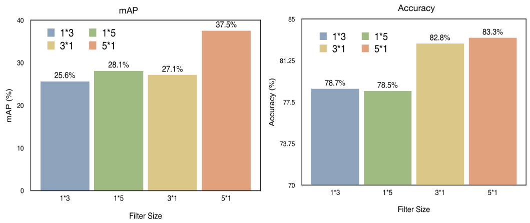


Figure 4: The classification accuracy and mAP scores of ablation study on the filter size of convolutional layer. We use $\alpha = 0.7$ to compute the mAP score.

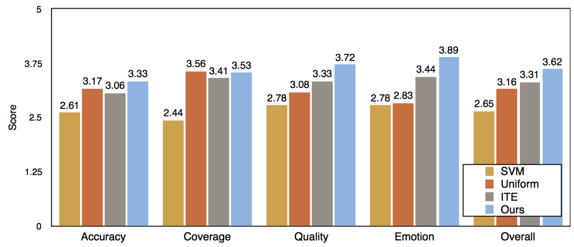


Figure 5: The qualitative results of emotion-oriented video summarization. "Overall" scores are the averaged sum of all the other four scores.

weakly supervised on Ekman6 dataset, the performance is very poor.

4.6 Video Emotion-oriented Summarization

The emotion-oriented video summarization is evaluated finally. Several different baselines are compared against our results: (1) Uniform (Uni): uniformly sample the frames/clips from the videos; (2) SVM: the video is summarized by the scores of SVM prediction. The frames with top-6 scores of each label are selected in practice. (3) ITE: we use the method based on ITE [24] for video summarization. To make the results more comparable, the length of summary is fixed to the same length.

User study is utilized to evaluate the results of summarization. Five students kept unknown from our project participate in the study. We show the results of all methods to each participant who rate the quality of summary on five-point scale by answering the following four questions [24]: (1) Accuracy: whether the summary accurately describes the main content of original video?(2) Coverage: how much visual content has been covered in the summary?(3) Quality: how is the overall subjective quality of summary? (4) Emotion: how many percentage of the same emotion has been conveyed from original video?

The user-study results are shown in Fig. 5. Our framework achieves the best performance on all the scores. This shows the good summary results that our model can generate. Figure 3 gives the qualitative illustration of the summary generated by different methods. The example is an "angry" video, which capture the process of an angry man smashing everything in the room. The Uniform and SVM methods are not good enough to grasp the video summary of angry emotion; and it makes the viewers think that the man is finding something. In contrast, both ITE and ours can help identify emotional segments. Especially, our summary has the shot of the man smashing the sofa with a hammer.

4.7 Ablation Study of the Convolutional Layer

The convolutional layer in EC-net performs the role of connecting FT-net and EC-net in our architecture. This layer processes the output of FT-net for the classification in EC-net. Previous work [11] used symmetric convolutional filters. In contrast, with the $fc7$ input features, our convolutional layer aims at modeling the contextual information within consecutive frames of each feature dimension. Considering the importance of the convolutional layer, we conduct the ablation study to verify different filter sizes of this layer. We compare the tasks of emotion recognition and emotion attribution with the filter sizes of 1×3, 1×5, 3×1, and 5×1 respectively on Emotion6 video dataset.

The results are reported in Fig.4. The results with the filter size of 5×1 can beat all the other choices on both classification accuracy and mAP score. This suggests that our choice of 5×1 filter size can successfully capture the contextual emotion information which can finally help both emotion classification and attribution tasks. In general, the convolutional filter sizes of 3×1 and 5×1 capture the information on the same feature channel (totally 4096 channels) along consecutive frames. In contrast, the filter sizes of 1×3 and 1×5 extract information from the same frame of neighboring feature channels, which however have limited contextual information usable for EC-net. That explains the low results of emotion recognition accuracy of filter sizes 1×3 and 1×5.

5 CONCLUSION

In this paper, we present a new architecture– Frame-Transformer Emotion Classification Network (FT-EC-net), which can solve the emotion recognition, emotion attribution and emotion-oriented summarization tasks simultaneously. A new labelled emotion attribution dataset is contributed. We also introduce an objective evaluation metric of emotion attribution task. The experiments on Emotion6 video and Ekman6 dataset have validated the effectiveness of our framework.

6 ACKNOWLEDGEMENT

This work was supported in part by three NSFC projects (#U1611461, #U1509206, #61572134) and a grant from STCSM, Shanghai, China (#16JC1420401), and Shanghai Sailing Program (17$YF$1427500).

REFERENCES

[1] Esra Acar, Frank Hopfgartner, and Sahin Albayrak. 2016. A comprehensive study on mid-level representation and ensemble learning for emotional analysis of video material. *Multimedia Tools and Applications* (2016), 1–29.

[2] Damian Borth, Rongrong Ji, Tao Chen, Thomas M. Breuel, and Shih-Fu Chang. 2013. Large-scale Visual Sentiment Ontology and Detectors using Adjective Noun Pairs. In *ACM MM*.

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 961–970.

[4] Tao Chen, Damian Borth, Darrell, and Shih-Fu Chang. 2014. DeepSentiBank: Visual Sentiment Concept Classification with Deep Convolutional Neural Networks. *CoRR* (2014).

[5] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. 2016. Emotiw 2016: Video and group-level emotion recognition challenges. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 427–432.

[6] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and others. 2015. Spatial transformer networks. In *Advances in Neural Information Processing Systems*. 2017–2025.

[7] Yu-Gang Jiang, Baohan Xu, and Xiangyang Xue. 2014. Predicting Emotions in User-Generated Videos. In *AAAI*.

[8] Brendan Jou, Subhabrata Bhattacharya, and Shih-Fu Chang. 2014. Predicting Viewer Perceived Emotions in Animated GIFs. In *ACM MM*.

[9] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[10] Dimitrios Kotzias, Misha Denil, Phil Blunsom, and Nando de Freitas. 2014. Deep Multi-Instance Transfer Learning. *CoRR* (2014).

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*.

[12] Jie-Ling Lai and Yang Yi. 2012. Key frame extraction based on visual attention model. *Journal of Visual Communication and Image Representation* 23, 1 (2012), 114–125.

[13] X. Lu, P. Suryanarayan, R. B. Adams, J. Li, M. G. Newman, and J. Z Wang. 2012. On shape and the computability of emotions. In *ACM MM*.

[14] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li. 2002. A User Attention Model for Video Summarization. In *ACM MM*.

[15] J. Machajdik and A. Hanbury. 2010. Affective image classication using features inspired by psychology and art theory. In *ACM MM*.

[16] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew Gallagher. 2015. A Mixed Bag of Emotions: Model, Predict, and Transfer Emotion Distributions. In *CVPR*.

[17] Attend Show. 2015. Tell: Neural Image Caption Generation with Visual Attention. *Kelvin Xu et. al.. arXiv Pre-Print* (2015).

[18] Krishna Kumar Singh and Yong Jae Lee. 2016. End-to-end localization and ranking for relative attributes. In *European Conference on Computer Vision*. Springer, 753–769.

[19] Ba Tu Truong and Svetha Venkatesh. 2007. Video Abstraction: A Systematic Review and Classification. *ACM TOMM* 3, 1 (2007), 79–82.

[20] Meng Wang, R. Hong, Guangda Li, Zheng-Jun Zha, Shuicheng Yan, and Tat-Seng Chua. 2012. Event Driven Web Video Summarization by Tag Localization and Key-Shot Identification. *IEEE TMM* 14, 4 (2012), 975–985.

[21] S. Wang and Q. Ji. 2015. Video affective content analysis: a survey of state of the art methods. *IEEE TAC* PP, 99 (2015), 1–1. DOI:https://doi.org/10.1109/TAFFC.2015.2432791

[22] Xi Wang, Yu-Gang Jiang, Zhenhua Chai, Zichen Gu, Xinyu Du, and Dong Wang. 2014. Real-time summarization of user-generated videos based on semantic recognition. In *ACM MM*.

[23] Baohan Xu, Yanwei Fu, Yu gang Jiang, Boyang Li, and Leonid Sigal. 2016. Video Emotion Recognition with Transferred Deep Feature Encodings. In *ICMR*.

[24] Baohan Xu, Yanwei Fu, Yu gang Jiang, Boyang Li, and Leonid Sigal. 2017. Heterogeneous Knowledge Transfer in Video Emotion Recognition, Attribution and Summarization. *IEEE TAC* (2017).

[25] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2015. Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks. In *AAAI*.