Deep Neural Networks for Video-level Emotion Analysis

Xunbo Ji
Supervisor: Philipp Paukner
Lehrstuhl für Datenverarbeitung
Technische Universität München

***Abstract*—Deep neural networks have shown strong performance in video action recognition tasks. The recently proposed network architectures can learn spatiotemporal features by fusing convolutional networks spatially and temporally. Motivated by that, in this paper we intend to transfer the end-to-end video-level representation learning approaches to conduct video emotion analysis. Four of them including 3D convolutional neural network(C3D) [1], 2plus1D convolution neural network(R2+1D) [2], Temporal Segment Network(TSN) [3] and an Efficient Convolutional Network(ECO) [4] are modified to analyse video-level emotions in LIRIS-ACCEDE dataset. Specifically, all networks are trained to analyse video-level emotions by aggregating the frame-level features which consist of spatial and temporal cues. The key difference among these networks is temporal pooling method. Experimental results show that the four networks are not performing well as we expected, though ECO model can analyse the emotions in training dataset quite accurately. All experiments are implemented with PyTorch and codes are available at https://gitlab.ldv.ei.tum.de/EmoVid/playground/tree/master/DPL.**

I. Introduction

Video-level emotion analysis is a challenging task which has drawn a significant amount of attention in computer vision and artificial intelligence research community. As streaming media services grow, more and more people communicate and satisfy their certain emotional needs by watching videos. To give people assistance in finding the videos, video contents should be analysed and classified according to emotions automatically. Not only that, analysing emotions has a lot of potential applications like building natural human-computer interfaces.

Emotion can be described using discrete approach or the continuous dimensional approach. The most frequently used discrete descriptor in the field of video affective content analysis is Ekman's six basic emotion categorical classes [5], including happiness, sadness, anger, disgust, fear, and surprise. The dimensional descriptor divides emotion into continuous spaces, e.g. arousal and valence, which first introduced by Wundt [6]. The various continuous values in each space can represent effective emotion features. For example, a relaxed state relates to low arousal, while anger relates to high arousal. Positive valence relates to a happy state, while negative valence relates to a depressed or angry state [7]. In this work, we focus on the problem where the goal is to analyse emotions in the valence and arousal space.

Inspired by the advanced performance of deep neural networks in image recognition [8] [9] [10] [11], researchers are using deep learning methods to solve video related problems like emotion and activity recognition. In comparison with image classification, the temporal context of videos provides additional and important information for content analysis. Recently, deep neural networks especially convolution neural networks have shown their superior ability to extract spatiotemporal features within videos for action recognition [12] [13] [14] [15]. The way to capture temporal relations between frames is one of the most important parts of video-level analysis. The temporal neighbourhood of a single frame comprises mostly redundant information and is almost useless for improving the belief about what happens in that frame. On the other hand, the contextual relationship between distant frames is meaningless, a simple aggregation of these frames is unwise. The researchers have proposed several long-term spatiotemporal architectures to solve this problem and each of them has its advantages and disadvantages. Most of these architectures are designed and evaluated for action recognition, but it is possible that the architectures may also valuable for other problems. Therefore, in this paper, we propose to exploit the state-of-the-art deep neural networks, introduced for action recognition, to analyse emotions in videos. The intended network architectures are all end-to-end video-level representation learning methods, including 3D convolution neural network(C3D) [1], R2plus1D convolution neural network(R2+1D) [2], Temporal Segment Network(TSN) [16], Efficient Convolutional Network(ECO) [4].

II. Related Work

A. Traditional machine learning methods

Before deep learning approach emerges, earlier methods first extract visual and audio features to characterize the video content using certain approaches. Visual features including shot-related features [17], motion-related features [18], camera distance-related features [19], lighting-related features [20], color features [21], [22] etc.. Since shot and motion control the tempo of videos, which reflect excitement of videos, while color saturation, layout, heat, domination, are also important emotion to affect viewer's emotion. Audio features including speech energy, pitch, duration, fundamental frequency, Log Frequency Power Coefficients (LFPC) [19], Mel-frequency Cepstrum Coefficients (MFCC) [23], For example, happy speech has a high energy at high-frequency range, and segmentation of music, speech and environmental sound is also a critical part of audio feature extraction. After feature

extraction, machine learning methods are applied to map video features and emotional descriptors, e.g. support vector machines (SVMs) [24], Adaboost [25], Gaussian Mixture Models (GMMs) [26], K-Nearest Neighbor (KNN) [27] etc.. The advantage of the traditional machine learning methods are interpretable, lower computing power consumption, but they highly rely on distinguishing feature extraction.

B. Deep learning methods

Most of the deep learning approaches are applied for action recognition. Simonyan et al. [15] introduced a well known two-stream ConvNets, which utilize pre-trained ImageNet [9] for the spatial stream and optical flow to capture short-term motion cues for the temporal stream. It samples a fixed number of RGB images to extract appearance features and a stack of optical flow frames to capture temporal motion information. Feichtenhofer et al. enhanced this two-stream networks with ResNet architecture [28] and additional connections between streams [29].

In another direction,3D ConvNets(C3D) [1] [30] extends 2D ConvNets [9] [10] [11] [28] to a to learn spatiotemporal features from a short video clip with consecutive frames. Later, Tran et al. [31] applied ResNet architecture with 3D convolutions and showed the improvements over their earlier C3D architecture. Recently, they are focusing on decomposing the 3D convolution filters into separate 2D spatial and 1D temporal filters [2], so-called R2+1D network, which obtains comparable and superior results compared to state-of-the-art action recognition methods. Carreira et al. [32] presents a two-stream Inflated 3D ConvNet (I3D) that inflates the 2D filters and pooling kernels (and optionally their parameters) into 3D, the very deep network achieves high performance after pre-training on Kinetics dataset [33].

However, the methods mentioned above extract clip-level features instead of video-level representations, they do not sufficiently capture the comprehensive information from the whole video, that suffers from the confusion caused by partial observation. Diba et al. [34] designed Temporal Linear Encoding (TLE), which aggregate temporal features sparsely sampled over the entire video with bilinear coding, but it only samples three frames in a video. Another solution to capture entire video-level features is Temporal Segment Network(TSN) [3], [16]. TSN combines a fixed number of frame-level predictions to make a global video prediction during training, typically eight and sixteen frames. Built upon TSN, an Efficient Convolutional Network(ECO) [4] was proposed to take advantage of 3D convolutional kernels to learn the temporal context between the frames that are chosen and processed by the same method as TSN. The ECO architecture also achieves favorable performance with a superior runtime-accuracy trade-off. Table I compares the performance of different network architectures for action recognition.

Besides, Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) [37] [38] [39] are widely used to learn the features of temporal sequence. Donahue et al. [40] and Ng et al. [41] employed a LSTM incorporated with a 2D CNN to aggregate spatiotemporal features. Yet LSTM has not shown its capacity in action recognition.

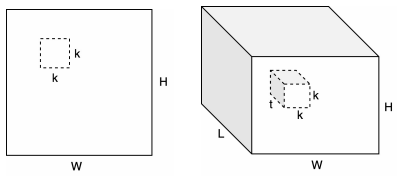
TABLE I
COMPARISON OF STATE-OF-THE-ART METHODS ON THE UCF101 [35] AND HMDB51 [36] ACTION RECOGNITION DATASETS. FOR FAIR COMPARISON, WE CONSIDER METHODS THAT USE ONLY RGB INPUT.

Model	Pretrain dataset	UCF101	HMDB51
Two Stream-RGB [15]	ImageNet	73.0%	40.5%
C3D [1]	ImageNet	82.3%	52.6%
I3D-RGB [32]	ImageNet	84.5%	49.2%
TSN-RGB [16]	ImageNet	86.4%	53.8%
ECO Lite [4]	-	90.2%	63.3%
ECO Full [4]	-	91.7%	65.6%
R(2+1)D-RGB [2]	Kinetics	96.8%	74.6%

III. NEURAL NETWORK ARCHITECTURES

A. 3D Convolution Neural Network(C3D)

C3D [1] is the initial study to utilize 3D convolutional filters to capture spatiotemporal features in videos, proposed in the year 2015. Since 2D convolutional filter only preserves spatial features, the 3D convolutional filter can simultaneously capture temporal information and propagate it through the layers of the network, which is indispensable for video-level analysis.



(a) 2D convolution (b) 3D convolution

Fig. 1. 2D and 3D convolution kernel, redrawn from [1]

As shown in Figure 1, a stack of L consecutive frames stands for the video and then feeds to network, the 3D convolutional kernel of size $t \times k \times k$ where t donates the temporal continuous extent of the kernel. Depends on computing resource, the structure of the network can be varied. The C3D architecture we used is illustrated in Figure 2.

B. R2+1D Convolution Neural Network(R2+1D)

R2+1D [2] was introduced in the year 2018 to take advantage of both 2D and 3D ConvNets, which decompose full 3D convolution filter into a 2D spatial convolution followed by a 1D temporal convolution(see the 2+1D filter in Figure 3). Compared to full 3D convolution, the decomposition form is easier to optimize. Furthermore, R2+1D utilize residual block in ResNet instead of convolution. Figure 4 shows the architecture of R2+1D and notation is the same as in C3D. Besides each residual or convolution block also include Batch Normalization and ReLU activation.

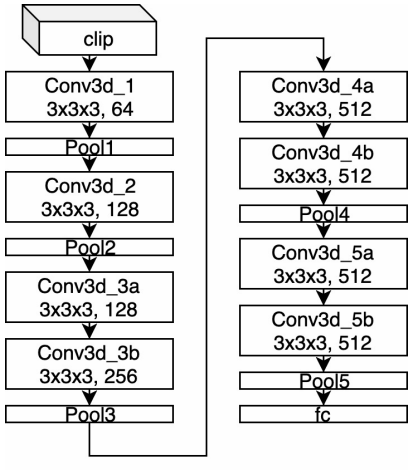


Fig. 2. C3D Architecture, redrawn from [1]

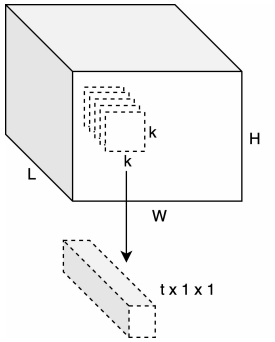


Fig. 3. R2+1D convolution, redrawn from [2].

C. Temporal Segment Network(TSN)

As the two ConvNets above capture temporal features of short-term consecutive frames, normally 1 - 2 seconds. For instance, the sampled frames only occupy a small portion of a 10-second video. TSN was designed to incorporate long-range information in the year 2016. To tackle the redundancy from consecutive frames, as shown in Figure 5, TSN first divides a video into K segments of equal duration, typically 8 segments. Then a snippet is randomly sampled from each segment. Thus, no matter how long a video is, TSN operates on the certain number of frames that evenly distributed in the video.ConvNets take the snippets as input and process them independently. The segmental consensus fuses the output feature representations from multiple snippets to yield a video-level prediction. Normally TSN uses weighted average pooling to aggregate the predictions of different snippets. When K equals to 1, TSN degenerates to the plain two-stream ConvNets. The 2D ConvNets in TSN here apply BN-Inception network [42] as the backbone.

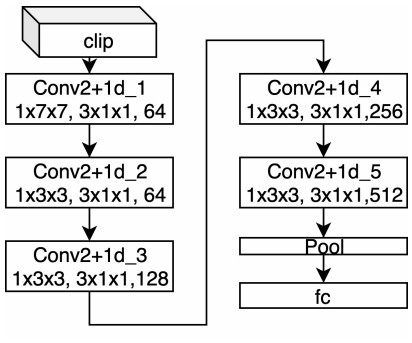


Fig. 4. R2+1D Architecture, redrawn from [2].

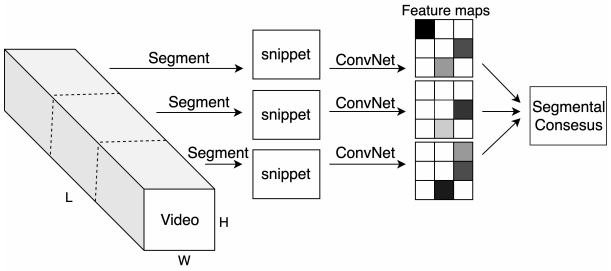


Fig. 5. TSN Architecture, redrawn from [3]. Each video is divided into K segments of equal duration. Then a snippet is randomly sampled from each segment. 2D ConvNets take the snippets as input and process them independently. The segmental consensus fuses the output snippet-level feature representations from multiple snippets to yield a video-level prediction.

D. Efficient Convolutional Network(ECO)

Motivated by the concept of TSN, Zolfaghari et al. devised the ECO model [4] last year. The early parts of ECO are similar to TSN, a fixed number of frames are selected from corresponding segments in the entire video. After 2D ConvNets, the sequences of extracted 2D feature maps are fed into a 3D ConvNet in order to catch temporal information. The 3D feature maps are obtained from 3D ConvNet and pass through fully-connected layers to get the final prediction. This architecture is illustrated in Figure 6 and called ECO Lite. There is another variant known as ECO Full. In ECO Full design, the 3D feature map and 2D feature maps are concatenated at the second last step and used to make the final decision. Note that the backbone of ECO is the same as TSN.

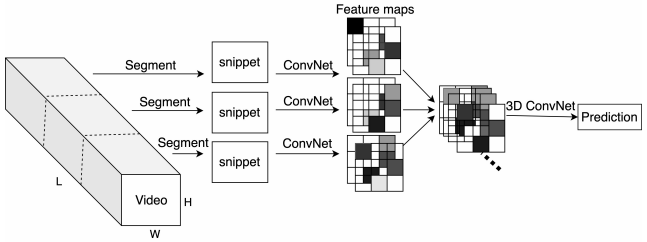


Fig. 6. ECO Architecture, redrawn from [4]. Each video is divided into K segments of equal duration. Then a frame is randomly sampled from each segment. 2D ConvNets take the frames as input and process them independently. Then the output frame-level feature representations from multiple frames are fed into 3D ConvNet to yield a video-level prediction.

IV. Experiments

In this section, we first introduce the evaluation dataset. Second, we present implementation details of the above four architectures C3D, R2+1D, TSN, ECO. Third, quantitative and qualitative results of applying different networks on the video emotion analysis dataset are reported.

A. Dataset

Experiments are conducted on video emotion dataset LIRIS-ACCEDE [43], which consists of 9800 video clips extracted from 160 movies. They are annotated by 1517 people from 89 different countries. All clips have a fixed frame rate and resolution of 25 FPS and 280x390 respectively. The length of all excerpts ranges from 8 to 12 seconds. The emotion in each video clip is annotated by a valence value and an arousal value with variance, which caused by different feelings of annotators. Valence values range from 1.3 to 3.6 while arousal values from 1.3 to 4.55. The dataset is split into training set 4900 clips, validation set 2450 clips and test set 2450 clips, in which the authors marked training set and test set are switched. The distributions of valence and arousal labels on training, validation, test set are drawn in Figure 7,8 and 9.

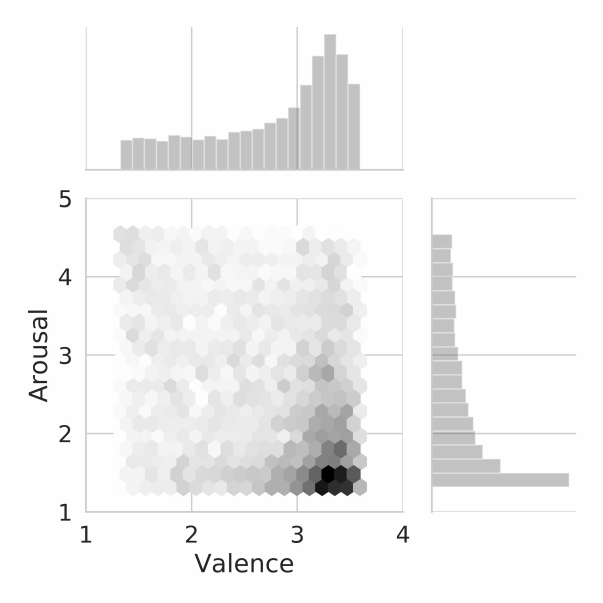


Fig. 7. Distribution of valence and arousal value on trainning set.

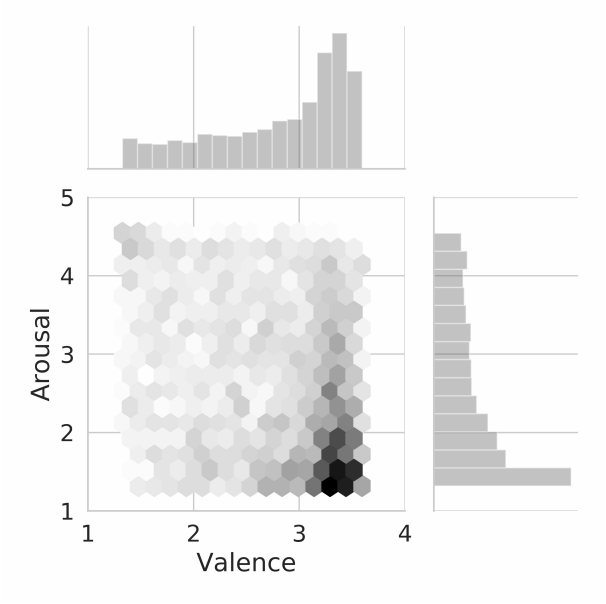


Fig. 8. Distribution of valence and arousal value on validation set.

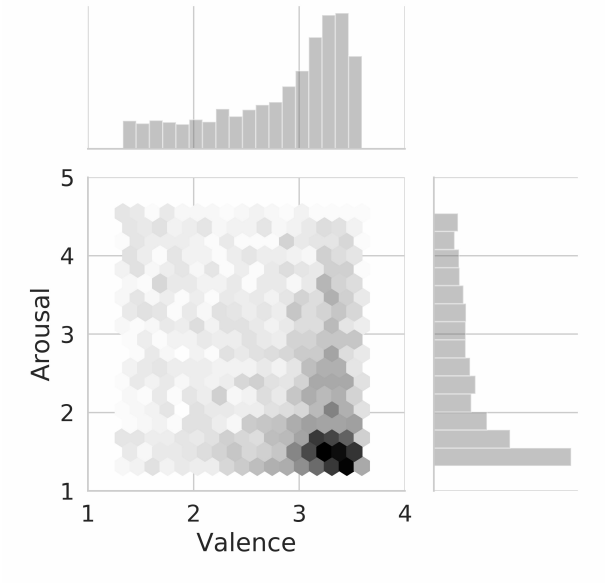


Fig. 9. Distribution of valence and arousal value on test set.

V. Experiments

A. Implementation Details

All networks use the mini-batch stochastic gradient descent algorithm to learn the network parameters. To ensure the fairness of experiments, every network takes eight RGB frames as input. The reason is that the average clip length of the mostly used action recognition dataset UCF101 [35] is 7.21 seconds, which is similar as LIRIS-ACCEDE dataset, and the paper [1] [3] [4] have proven that eight frames as input can achieve best runtime-accuracy trade-off. We also campared the effects of different numbers of input frames. The frames are resized to 224×224 using BILINEAR interpolation. In addition, the video clips in the original dataset are downsampled into 5 FPS. Since all the networks are originally designed for action recognition, and action recognition is a classification problem, we have to modify the last layer of network structure to transfer classification to regression. Meanwhile, Mean Square Error(MSE) is used as a loss function. All the neural networks are implemented using PyTorch and trained on a NVIDIA

TITAN X GPU. If not specifically noted, the experiments in the section are trained from scratch.

To evaluate the effect of different networks, we adopt the Pearson correlation coefficient [44] with p-value to show the relationships between predictions and ground truth of emotion. Pearson correlation coefficient(marked as pearsonr) is a measure of the linear correlation between two variables X and Y. It ranges from -1 to 1. A value of 1 implies that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line for which Y increases as X increases. A value of -1 implies that all data points lie on a line for which Y decreases as X increases. A value of 0 implies that there is no linear correlation between the variables. And p-value tells you whether the correlation is statistically significant. If p-value less than the a significance level(typically 0.05), which indicates that the correlation coefficients are significant.

B. Results

1) C3D: In C3D experiment, eight temporal consecutive frames(224×224) of each video clip are selected and fed into the C3D for training, then the network outputs a predicted valence or arousal value through fully-connected layers. The network settings include the size and number of the convolutional kernel, the order of different layers are presented in Figure 2. Batch normalization is applied to all convolutional layers. The learning rate starts at 0.0001 and is decreased by a factor of 10 every 4 epoch with in total 20 epochs.

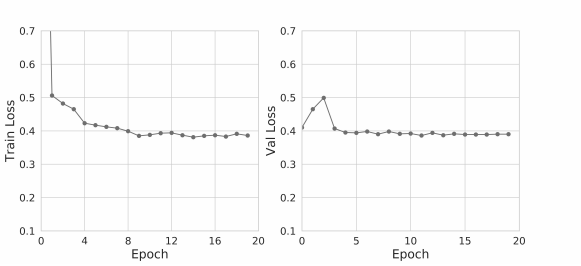
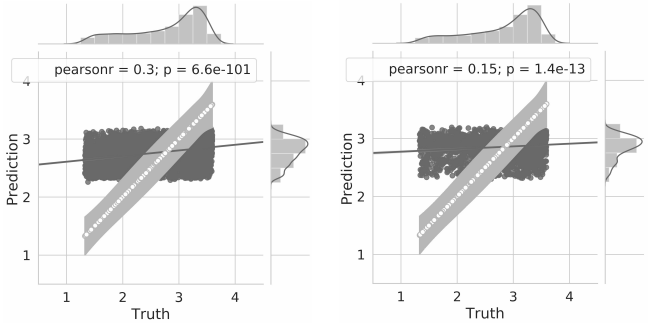


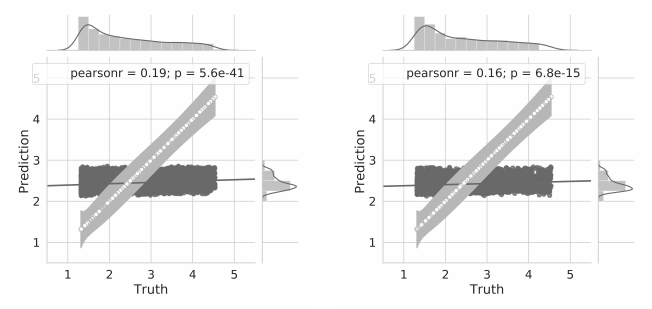
Fig. 11. Evolution of loss during training C3D.

Figure 10 reports the difference between network predictions and targets of valence and arousal value on training set and test set, where pearsonr refers to Pearson correlation coefficient and p refers to p-value. The white circles with light gray region indicate ground truth with corresponding annotated standard deviation. The best performance should be a straight line with a slope of 1. The predictions on training set and test set mostly distributed around 2.8 and are not related to ground truth. Thus we can say, that C3D doesn't perform qualified for emotion analysis and can't extract useful information. The line chart of loss value during training and test is plotted in Figure 11 and shows that the network doesn't overfit. Notably, that loss value is Mean Square Error(MSE).

2) R2+1D: R2+1D is a improved version of C3D achieved better performance than C3D for action recognition. Except those 3D convolutions are replaced with 2+1D convolutions, the other parts of R2+1D are similar to C3D, see Figure 4.

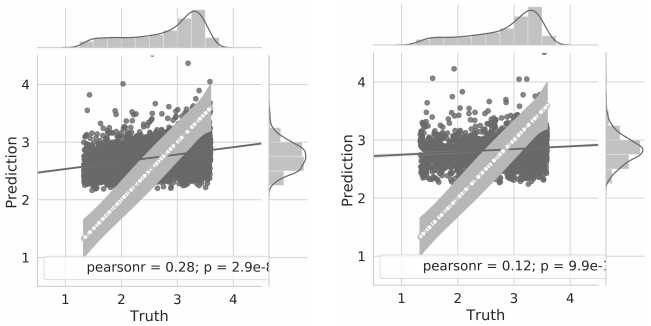


(a) Predictions and truth of Valence on training set (b) Predictions and truth of Valence on test set

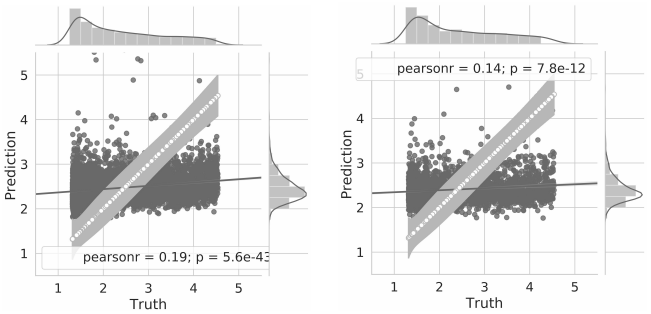


(c) Predictions and truth of Arousal on training set (d) Predictions and truth of Arousal on test set

Fig. 10. Results of C3D



(a) Predictions and truth of Valence on training set (b) Predictions and truth of Valence on test set



(c) Predictions and truth of Arousal on training set (d) Predictions and truth of Arousal on test set

Fig. 12. Results of R2+1D

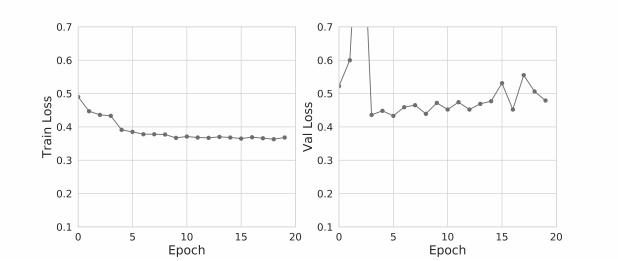


Fig. 13. Evolution of loss during training R2+1D.

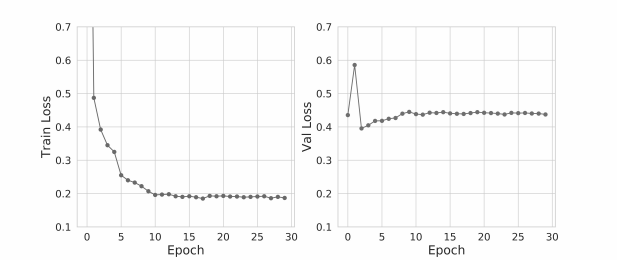
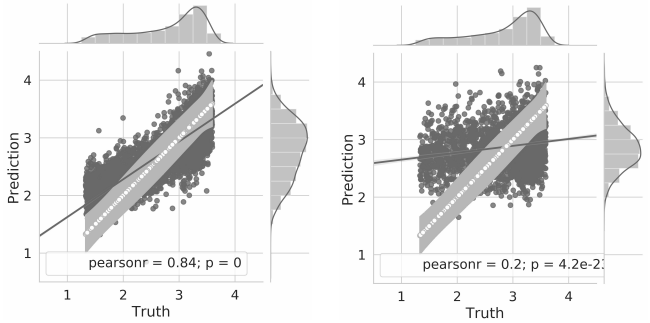


Fig. 15. Evolution of loss during training TSN.

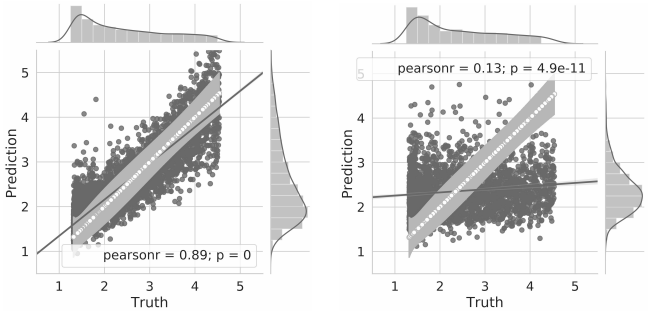
The initial learning rate is set to 0.0001 and divided by 10 every 4 epochs. The results in Figure 12 show that R2+1D doesn't perform better than C3D in our task.

3) TSN: In original TSN settings, the network takes two-stream RGB and optical flow as input. But in our task we use only RGB frames. Every video is averagely split into eight segments. The concatenated feature maps through BN-Inception and consensus function are utilized to predict emotion value. We trained the network for 30 epochs and reducing the learning rate every 5 epochs by a factor of 10, which is initialized as 0.001. Figure 14 displays that on train set TSN can roughly predict the emotion value with small error. The Pearson correlation coefficient 0.84 s shows that the predictions and truth are obvious related. Nonetheless, TSN doesn't make much improvement on test set. Thus, TSN is fit well on training set, but if it faces the videos never seen before, it can't show its recognition ability.

4) ECO: Encouraged by the results of TSN, we focus on the newly introduced ECO model. ECO additionally applies 3D convolutions to the sparsely sampled frames. The other settings including learning rate remain the same as TSN. As mentioned before, ECO has two variants, ECO Lite and ECO Full. Due to complexity, ECO Full consumes more computing power and more time for training. However, from Figure 16 and 18 we can observe that ECO Lite outperforms ECO Full on training set, which is contrary to the results in action recognition. And ECO Lite shows the best capacity to analyse emotion on training set, while unfortunately, the performance on test set remains not good. Furthermore, some adjustments including different optimizers and numbers of fully-connected layers are applied for evaluation, as the results are shown in Table II. Not only that, we also compared the effects of different numbers of frames as the input of network in Table III.

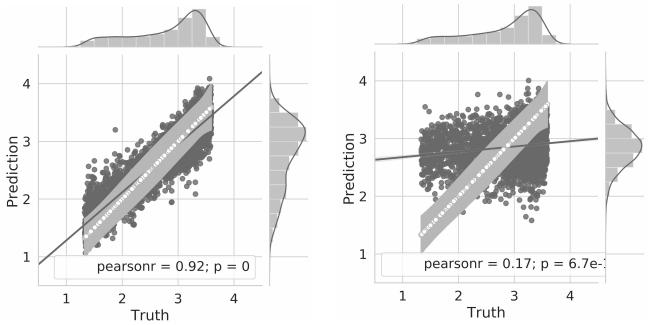


(a) Predictions and truth of Valence on training set (b) Predictions and truth of Valence on test set

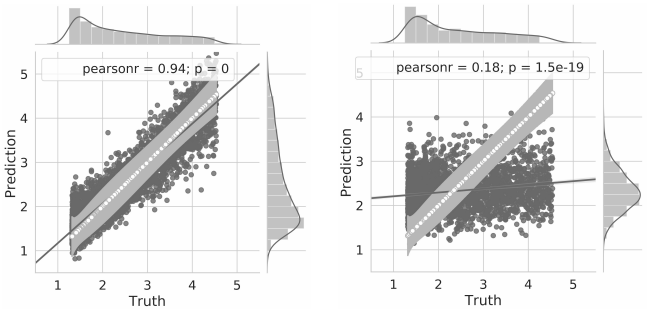


(c) Predictions and truth of Arousal on training set (d) Predictions and truth of Arousal on test set

Fig. 14. Results of TSN



(a) Predictions and truth of Valence on training set (b) Predictions and truth of Valence on test set



(c) Predictions and truth of Arousal on training set (d) Predictions and truth of Arousal on test set

Fig. 16. Results of ECO Lite

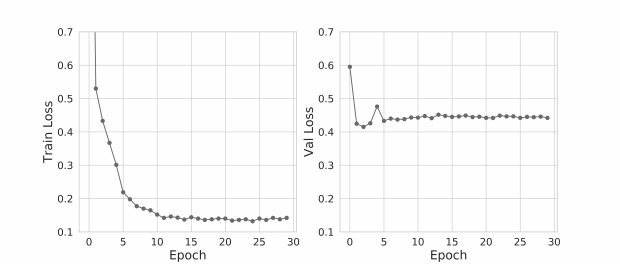
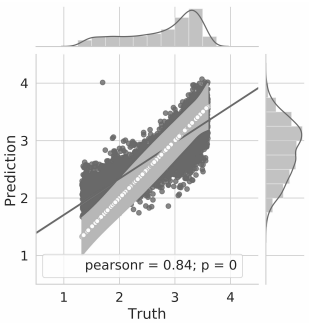
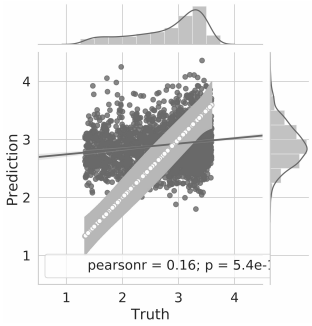


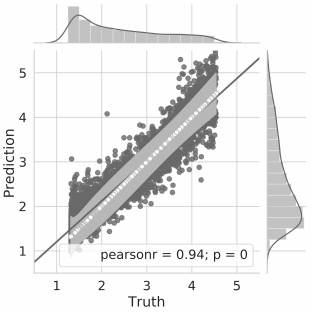
Fig. 17. Evolution of loss during training ECO Lite.



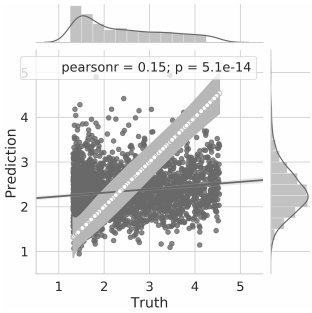
(a) Predictions and truth of Valence on training set



(b) Predictions and truth of Valence on test set



(c) Predictions and truth of Arousal on training set



(d) Predictions and truth of Arousal on test set

Fig. 18. Results of ECO Full

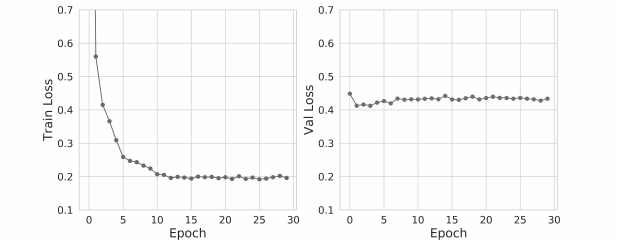


Fig. 19. Evolution of loss during training ECO Full.

Table IV summarizes all the implementation results of C3D, R2+1D, TSN and ECO network models for valence and arousal value prediction separately. Despite great fitting effects on training set, TSN and ECO did not show the desired effect on test test. C3D and R2+1D can not even learn the features from training set properly.

TABLE II
PERFORMANCE OF DIFFERENT HYPER-PARAMETERS IN ECO MODEL ON VALENCE TRAINING DATASET

Model	Optimizer	FC layer	MSE[a]
ECO Lite	SGD	1	0.412
	SGD	3	0.532
	ADAM	1	0.136
	ADAM	3	0.317
ECO Full	SGD	1	0.375
	SGD	3	0.432
	ADAM	1	0.183
	ADAM	3	0.336

[a]Mean Square Error.

TABLE III
EFFECT OF DIFFERENT NUMBERS OF INPUT FRAMES FOR AROUSAL

Model	Frames	Trainning Set		Test Set	
		Pearson[a]	MSE[b]	Pearson[a]	MSE[b]
ECO Lite	4	0.860	0.221	0.160	0.999
	8	0.939	0.109	0.182	0.963
	12	0.917	0.123	0.199	0.964

[a]Pearson Correlation Coefficient.
[b]Mean Square Error.

TABLE IV
SUMMARY OF INPLEMENTATION RESULTS

	Model	Trainning Set		Test Set	
		Pearson[a]	MSE[b]	Pearson[a]	MSE[b]
Valence	C3D	0.298	0.372	0.149	0.380
	R2+1D	0.277	0.390	0.123	0.400
	TSN	0.838	0.123	0.195	0.429
	ECO Lite	0.924	0.061	0.172	0.420
	ECO Full	0.843	0.124	0.162	0.437
Arousal	C3D	0.190	0.871	0.156	0.884
	R2+1D	0.194	0.901	0.138	0.920
	TSN	0.891	0.192	0.132	1.047
	ECO Lite	0.939	0.109	0.182	0.963
	ECO Full	0.942	0.109	0.150	1.029

[a]Pearson Correlation Coefficient.
[b]Mean Square Error.

VI. CONCLUSION

In this paper, we have investigated different deep neural networks that originally proposed for video-level action recognition, and evaluated the performances for emotion analysis. The network architectures look only at a small stack of frames from a video and learn to capture spatiotemporal information of the video. Four network architectures including C3D, R2+1D, TSN and ECO are modified to analyse emotion in video and adapt LIRIS-ACCEDE dataset. Though the ECO model fits

the training dataset very well, the performances of all four network on test dataset are not as satisfied as we expected due to a poor generalization ability. This time the transfer learning didn't achieve expectable results on the dataset but we hope that this work could make some contributions to future research in video-level emotion analysis.

REFERENCES

[1] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 4489–4497, 2015.

[2] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," *2018 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, 2018.

[3] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European Conference on Computer Vision*, 2016.

[4] M. Zolfaghari, K. Singh, and T. Brox, "Eco: Efficient convolutional network for online video understanding," in *European Conference on Computer Vision*, 2018.

[5] P. Ekman, "Basic emotions," *Handbook of Cognition and Emotion*, pp. 45–60, 1999.

[6] W. M. Wundt, *Grundzüge der physiologischen Psychologie*. W. Engelman, 1905.

[7] S. Wang and Q. Ji, "Video affective content analysis: A survey of state-of-the-art methods," *IEEE Transactions on Affective Computing*, vol. 6, pp. 410–430, 2015.

[8] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

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

[11] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2014.

[12] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1725–1732, 2014.

[13] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[14] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 4597–4605, 2015.

[15] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, pp. 568–576, 2014.

[16] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool, "Temporal segment networks for action recognition in videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, pp. 2740–2755, 2017.

[17] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa, "Affective audio-visual words and latent topic driving model for realizing movie affective scene classification," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 523–535, 2010.

[18] Y. Cui, S. Luo, Q. Tian, S. Zhang, Y. Peng, L. Jiang, and J. S. Jin, "Mutual information-based emotion recognition," in *The Era of Interactive Media*, pp. 471–479, Springer, 2013.

[19] H. L. Wang and L.-F. Cheong, "Affective understanding in film," *IEEE Transactions on circuits and systems for video technology*, vol. 16, no. 6, pp. 689–704, 2006.

[20] L. Canini, S. Benini, and R. Leonardi, "Affective recommendation of movies based on selected connotative features," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 4, pp. 636–647, 2012.

[21] S. Arifin and P. Y. Cheung, "A computation method for video segmentation utilizing the pleasure-arousal-dominance emotional information," in *Proceedings of the 15th ACM international conference on Multimedia*, pp. 68–77, ACM, 2007.

[22] S. Arifin and P. Y. Cheung, "Affective level video segmentation by utilizing the pleasure-arousal-dominance information," *IEEE Transactions on Multimedia*, vol. 10, no. 7, pp. 1325–1341, 2008.

[23] D. Wu, T. D. Parsons, and S. S. Narayanan, "Acoustic feature analysis in speech emotion primitives estimation," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[24] C.-Y. Wei, N. Dimitrova, and S.-F. Chang, "Color-mood analysis of films based on syntactic and psychological models," *2004 IEEE International Conference on Multimedia and Expo (ICME))*, vol. 2, pp. 831–834 Vol.2, 2004.

[25] H.-B. Kang, "Affective content detection using hmms," in *Proceedings of the eleventh ACM international conference on Multimedia*, pp. 259–262, ACM, 2003.

[26] A. Yazdani, E. Skodras, N. Fakotakis, and T. Ebrahimi, "Multimedia content analysis for emotional characterization of music video clips," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, p. 26, 2013.

[27] A. Yazdani, K. Kappeler, and T. Ebrahimi, "Affective content analysis of music video clips," in *Proceedings of the 1st international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, pp. 7–12, ACM, 2011.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[29] C. Feichtenhofer, A. Pinz, and R. Wildes, "Spatiotemporal residual networks for video action recognition," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 3468–3476, 2016.

[30] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," in *European conference on computer vision*, pp. 140–153, Springer, 2010.

[31] D. Tran, J. Ray, Z. Shou, S.-F. Chang, and M. Paluri, "Convnet architecture search for spatiotemporal feature learning," *arXiv preprint arXiv1708.05038*, 2017.

[32] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4724–4733, 2017.

[33] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, A. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[34] A. Diba, V. Sharma, and L. V. Gool, "Deep temporal linear encoding networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1541–1550, 2016.

[35] K. Soomro, A. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, 12 2012.

[36] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb51: A large video database for human motion recognition," 2011.

[37] A. Graves, N. Jaitly, and A. rahman Mohamed, "Hybrid speech recognition with deep bidirectional lstm," *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 273–278, 2013.

[38] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems 27* (Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds.), pp. 3104–3112, 2014.

[39] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015.

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

[41] J. Y.-H. Ng, M. J. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4694–4702, 2015.

[42] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
[43] Y. Baveye, J.-N. Bettinelli, E. Dellandréa, L. Chen, and C. Chamaret, "A large video database for computational models of induced emotion," *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 13–18, 2013.
[44] Wikipedia, "Pearson correlation coefficient."