# Emotion Recognition for In-the-wild Videos

Hanyu Liu*‡, Jiabei Zeng* , Shiguang Shan*† and Xilin Chen*†

*Key Lab of Intelligent Information Processing, Chinese Academy of Sciences (CAS)
Institute of Computing Technology, CAS, Beijing 100190, China
{jiabei.zeng, sgshan, xlchen}@ict.ac.cn
†School of Computer Science and Technology
University of Chinese Academy of Sciences, Beijing 100049, China
‡School of Computer Science
Beijing University of Posts and Telecommunications, Beijing 100876, China
liuhanyu@bupt.edu.cn

*Abstract*—This paper is a brief introduction to our submission to the seven basic expression classification track of Affective Behavior Analysis in-the-wild Competition held in conjunction with the IEEE International Conference on Automatic Face and Gesture Recognition (FG) 2020. Our method combines Deep Residual Network (ResNet) and Bidirectional Long Short-Term Memory Network (BLSTM), achieving 64.3% accuracy and 43.4% final metric on the validation set.

*Keywords*-facial expression recognition

## I. INTRODUCTION

Automated facial expression recognition (FER) in-the-wild is a long-standing problem in affective computing and human-computer interaction. To analyze facial expression, psychologists and computer scientists have classified the facial expression into a list of emotion-related categories, such as six basic emotions, i.e., anger, disgust, fear, happiness, sadness, and surprise. Ekman et al.[1] have shown that the six basic emotional expressions are universal among human beings. There has been an encouraging progress on facial expression recognition and during the past decades.

In the the Affective Behavior Analysis in-the-wild (ABAW) 2020 Competition[2], the holders provide a large scale in-the-wild database called Aff-Wild2[3], [4], [5], [6], [7], including videos annotated with emotion categories, facial action unit[8], and valence and arousal[9] dimension. In this paper, we present our method used in the expression track in ABAW. In this track, the task is to distinguish seven basic facial expressions (i.e., neutral, anger, disgust, fear, happiness, sadness, surprise) of the person in the given videos. Our method adopts a 101-layer ResNet[10] with convolutional block attention module (CBAM)[11] to extract frame-by-frame features. Then, the features are fed into a bidirectional recurrent neural network with long short-term memory (BLSTM)[12] units.

## II. RELATED WORK

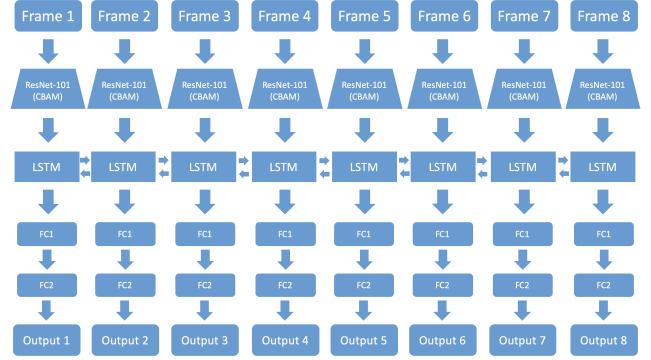Deep neural network-based algorithms are widely used in image and video analysis in recent years. Convolutional neural networks (CNN), for example, residual neural network (ResNet), VGGNet[13], AlexNet[14] are shown effective in image classification and image feature extraction. Long short-term memory network (LSTM)[12], a specific improvement of recurrent neural network (RNN), which is capable of capturing serial information, is used in natural language processing as well as video analysis. Meanwhile, the architecture of combination of CNN and RNN is proved to have excellent performance on emotion related tasks. Woo et al. proposed convolutional block attention module (CBAM)[11], a lightweight and general attention module boosting the performance of all kinds of CNNs. D. Kollias et al. collected Aff-Wild dataset[15], the first dataset with annual annotations for each frame of the videos for facial action unit, facial expression and valence-arousal research.



Figure 1. Architecture of the proposed method.

## III. METHOD

Our method consists of three part: ResNet-101 with CBAM that extracts features for each frame, BLSTM that captures the dynamic features of continuous frames, classification module that makes the decisions. Fig. 1 illustrate the framework of our method. Below, we present the three parts in details.
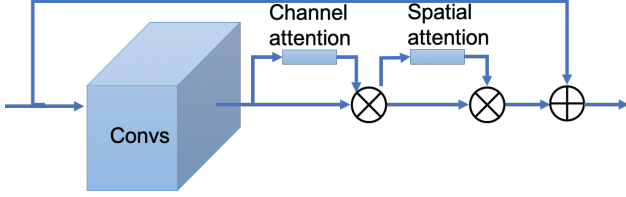
Figure 2. Structure of the residual block with CBAM

### A. ResNet-101 with CBAM

Since ResNet has achieved considerable performance in a lot of computer vision tasks[10], we adopt a 101-layer ResNet (ResNet101) to extract visual features from each frame. Considering facial expression appears in particular location in the image, we add a convolutional block attention module (CBAM)[11] after each residual block of ResNet101 to introduce channel attention and spatial attention. Fig. 2 illustrates the structure of a residual block with CBAM.

### B. BLSTM

Since facial expressions is continuous in the time dimension, we use a Long Short-Term Memory Network (LSTM) to process timing information. Considering that we need to select features for each frame, including the starting frame of the eight frame video clip, we use a bidirectional LSTM here.

### C. Classification module

Lastly, fully connected layers are applied to classify the features into seven classes based on the features extracted and selected by previous layers.

## IV. EXPERIMENT

### A. Dataset

Other than the emotion-annotated part in the provided Aff-Wild2 dataset, we used several internal facial expression datasets (AffectNet[16], RAF-DB[17][18]) and a self-collected datasets with 300,000 images to pre-train our model.

**Aff-Wild2:** Aff-Wild2 annotated in total 539 videos consisting of 2,595,572 frames with 431 subjects, 265 of which are male and 166 female. The dataset are split into train/validation/test parts in a subject-independent manner, with 253, 71, 233 subjects in each.

**AffectNet:** AffectNet contains about 440,000 manually annotated facial images collected from Internet search engines. We only used the images with neutral and 6 basic emotions in the training part, including around 280,000 images.

**RAF-DB:** Real-world Affective Faces Database (RAF-DB) contains around 30,000 facial images annotated with basic or compound expressions. We only used the 12,271 ones in the training part annotated with basic emotions.

### B. Preprocessing

The original videos were first divided into frames. These images were later applied on RetinaFace detector[19] to detect all the faced to be analyzed, aligned and cropped into size of $256 \times 256$. In order to make the external dataset perform better, all the procedures during preprocessing are similar to the official preprocessing except the tools used. For some frames in which human face were not able to be detected by the detector, the corresponding images were removed from the training dataset.

### C. Training

We implemented our model using PyTorch[20], on a server with four Nvidia GeForce GTX Titan X GPUs, each with 12GB memory. The model is trained with stochastic gradient descent (SGD) with learning rate 0.0001 and momentum 0.9. Loss function is cross entropy loss. The training batch size is set as 4. At each step during the training, one video from all the videos in the training dataset is selected with equal probability. And then a continuous 8 frame video clip (i.e., without any frame from which faces are unable to be detected) is randomly selected from this video as a batch. The model makes to its best performance usually within 200,000 batches. After every 1000 iteration, we recorded the temporary parameters of the model as a checkpoint.

### D. Evaluation

All the video frames in validation set are arranged into 8 frame clips to be calculated collectively. If the length of a video is not divisible by 8, the last several frames are padding with zeros. The BLSTM part outputs the features for each time step in the clip so that these 8 frames are classified and labeled in a single round. We counted the number of successfully predicted frame as well as the total number of frames processed by our model.

The final metric $\mathcal{S}$ is a combination of accuracy and $F_1$ formulated as:

$$\mathcal{S} = 0.33 Acc + 0.67 F_1, \qquad (1)$$

where $Acc$ is the accuracy which is computed as the ration total number of correctly predicted frames over the total frames. $F_1$ of computed as unweighted mean of all $F_1$ of seven categories. The $F_1$ of a single category is computed as:

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \qquad (2)$$

We manually select the parameter with best performance on validation set from all the checkpoints.

### E. Result

We evaluated our method on the validation set of Aff-Wild2 and reported the result of our method in Table I. The baseline method is MobileNetV2. **ResNet+BLSTM**

#### Table I
#### RESULT ON THE VALIDATION SET

| Method | Acc | F1 | $\mathcal{S}$ |
|---|---|---|---|
| baseline[2] | - | - | 0.36 |
| ResNet+BLSTM | 0.647 | 0.281 | 0.402 |
| ResNet+BLSTM+CBAM | 0.640 | 0.333 | 0.434 |

$$\mathcal{S} = 0.33Acc + 0.67F_1$$

is the combination of vanilla ResNet101 and BLSTM. **ResNet+CBAM+BLSTM** added the CBAM after each layer of ResNet101. As can be seen in Table I, **ResNet+CBAM+BLSTM** achieves higher final metric $\mathcal{S}$.

### V. CONCLUSION

Our proposed method reaches 64.65% accuracy on the validation set, and 43.43% final metric on the validation set, 7.43% higher than the 36% baseline proposed in the competition announcement.

### ACKNOWLEDGMENT

### REFERENCES

[1] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[2] D. Kollias, A. Schulc, E. Hajiyev, and S. Zafeiriou, "Analysing affective behavior in the first ABAW 2020 competition," 2020.

[3] D. Kollias and S. Zafeiriou, "Aff-Wild2: Extending the Aff-Wild database for affect recognition," *CoRR*, vol. abs/1811.07770, 2018. [Online]. Available: http://arxiv.org/abs/1811.07770

[4] ——, "Expression, affect, action unit recognition: Aff-Wild2, multi-task learning and ArcFace," *CoRR*, vol. abs/1910.04855, 2019. [Online]. Available: http://arxiv.org/abs/1910.04855

[5] ——, "A multi-task learning & generation framework: Valence-arousal, action units & primary expressions," *CoRR*, vol. abs/1811.07771, 2018. [Online]. Available: http://arxiv.org/abs/1811.07771

[6] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. W. Schuller, I. Kotsia, and S. Zafeiriou, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *International Journal of Computer Vision*, vol. 127, no. 6-7, pp. 907–929, 2019.

[7] D. Kollias, M. A. Nicolaou, I. Kotsia, G. Zhao, and S. Zafeiriou, "Recognition of affect in the wild using deep neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 1972–1979.

[8] E. Friesen and P. Ekman, "Facial action coding system: a technique for the measurement of facial movement," *Palo Alto*, vol. 3, 1978.

[9] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[11] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

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

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

[15] S. Zafeiriou, D. Kollias, M. A. Nicolaou, A. Papaioannou, G. Zhao, and I. Kotsia, "Aff-Wild: Valence and arousal 'in-the-wild' challenge," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 1980–1987.

[16] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[17] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 2584–2593.

[18] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, 2019.

[19] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-stage dense face localisation in the wild," *CoRR*, vol. abs/1905.00641, 2019.

[20] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Annual Conference on Neural Information Processing Systems*, 2019, pp. 8024–8035.