Facial Expression Classification using Fusion of Deep Neural Network in Video

Kim Ngan Phan
Dept. of Artificial Intelligence Convergence
Chonnam National University
Gwangju, South Korea
kimngan260997@gmail.com

Hong-Hai Nguyen
Dept. of Artificial Intelligence Convergence
Chonnam National University
Gwangju, South Korea
honghaik14@gmail.com

Van-Thong Huynh
Dept. of Artificial Intelligence Convergence
Chonnam National University
Gwangju, South Korea
vthuynh@jnu.ac.kr

Soo-Hyung Kim*
Dept. of Artificial Intelligence Convergence
Chonnam National University
Gwangju, South Korea
shkim@jnu.ac.kr

Abstract

For computers to recognize human emotions, expression classification is an equally important problem in the human-computer interaction area. In the 3rd Affective Behavior Analysis In-The-Wild competition, the task of expression classification includes eight classes with six basic expressions of human faces from videos. In this paper, we employ a transformer mechanism to encode the robust representation from the backbone. Fusion of the robust representations plays an important role in the expression classification task. Our approach achieves 30.35% and 28.60% for the F_1 score on the validation set and the test set, respectively. This result shows the effectiveness of the proposed architecture based on the Aff-Wild2 dataset and our team archives 5^{th} for the expression classification task in the 3rd Affective Behavior Analysis In-The-Wild competition.

1. INTRODUCTION

Understanding Affective Behavior is playing an essential role in the interaction between computers and humans [16]. This interaction makes it possible for computers to understand human behaviors and emotions and feelings. For many years, scientists have been working to build an intelligent and automated machine that can understand and serve humans in many fields of health, education, and services. The emotion recognition system allows for receiving many different data sources such as biological signals, visuals, or documents. Visual data directly depicts interpretations of emotions through facial expressions, thus playing an important role in emotion classification. In 1969, Ekman proposed six basic emotions in [3] including anger, disgust, fear, happiness, sadness, and surprise. They are used popular but are not sufficient to express complex human emotional states. In 2021, the 2nd Affective Behavior Analysis In-The-Wild (ABAW2) Competition solves the problem of affective behavior analysis in-the-wild, the target is to create machines and robots that are capable of understanding people's feelings, emotions and behaviors [9–16, 25]. The ABAW2 competition has three challenges-tracks are: valence-arousal estimation, basic expression classification, and facial action unit detection. In 2022, the 3rd Affective Behavior Analysis In-The-Wild (ABAW3) Competition is organized with the major goal of building a system and improving the emotional recognition ability of machines [9–16, 25]. The competition consists of 4 challenges: valence-arousal estimation, expression classification, action unit (AU) detection, and multi-task-learning. Each task corresponds to huge datasets with different sizes from Aff-Wild2 Database [9–16, 25]. They include videos and cropped and aligned frames that contain annotations in terms. We only perform the expression classification task in this paper.

In this study, we propose the combination of representative features from deep learning networks for the expression classification task. We opt RegNet as the backbone of our network. The transformer encoder plays a role as the embedding layer to extract robust representations from the backbone. We employ multi-head attention with the space of the attention heads being expanded whose embed dimensions are flexible. Our model archives better performance than the baseline on the validation set. Section 2 describes the proposed method. The training details and results are

*Corresponding author

reported at the last of the paper.

2. RELATED WORKS

There are many previous studies on expression recognition in ABAW2 Competition. Zhang et al. [26] propose a multi-task streaming network by the hierarchical relationships between different emotion representations. Jin et al. [6] employ visual and audio modalities for multi-task learning. They fuse uni-modal features from the visual model and audio model and then feed into the encoder network. Thinh et al. [21] propose the multi-task learning technique and ResNet50 model for two tasks of emotion classification and action unit detection. Wang et al. [23] propose a mean teacher framework to solve incomplete labeled datasets for the multi-task multi-modal model. Deng et al. [1] improve emotion uncertainty estimation by deep ensemble models using a multi-generational self-distillation algorithm.

Following that, the studies on expression recognition are developed through the ABAW3 competition. Zhang et al. [27] fuse the static vision features and the dynamic multi-modal features and feed them into a transformer-based fusion module. Jeong et al. [5] perform fine-tuning ResNet architecture to extract features. They propose a DAN network by a combination of a spatial attention unit and channel attention. Xue et al. [24] propose Coarse Net and Negative Net for two groups of emotions divided into two stages. Savchenko et al. [20] concatenate the embeddings and scores from EfficientNet that is implemented in an Android mobile application. Kim et al. [7] apply Swin transformer [18] to develop a three-stream network consisting of a visual stream, a temporal stream and an audio stream for facial expression recognition.

3. PROPOSED METHOD

In this paper, we employ a transformer encoder with multi-head attention as the embedded layer to generate sequence representations. In the multi-head attention, the heads are expanded with flexible embed dimensions to enhance the information on the heads. The transformer helps encode the robust representations for the backbone of the model. We also employ the pre-trained model RegNet [19] as the backbone for the proposed network. We use the weight of RegNetY with 1.6GF architecture on ImageNet dataset [2] to extract feature of images. The pre-trained model takes the input size of the image as 112x112x3. The backbone is extracted with 888 features by the flattened layer of the pre-trained model. We reshape the backbone to (batch size, sequence, feature) and fed it into the transformer encoder. In this work, we opt for the length of the sequence to be 64. To improve the performance of the classification task, we consider the fusion of the robust representations before entering the classification model. We combine the backbones and the representation of the transformer encoder mechanism to get the final representation features for the expression classification task. Dropout layer with 50% information and dense layers of 8 neurons are used for final output corresponding to 8 expressions of human. Fig 3 describe detail of our architecture.

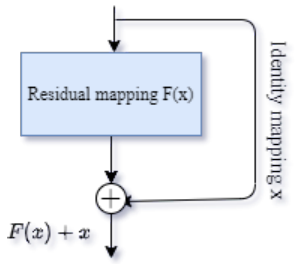

Figure 1. A building block in residual learning.

3.1. Residual Learning

In [4], the authors propose the residual nets (ResNets) as an effective solution in a deep learning network. The ResNet performs a deep residual learning framework to solve degradation problems when depth increases. It not only helps the model have a compact structure but also achieves state-of-the-art performance for classification tasks. A building block of residual learning includes the shortcut connection and the element-wise addition. There is no computational complexity in shortcut connection because of skipping one or more layers. We consider x as input, the output y of the building block is defined as:

$$y = F(x, \{W_i\}) + x$$

where F represents the residual mapping that stores the stacked nonlinear layers. In there, x is identity mapping performed shortcut connection and W_i is the weight parameters of the stacked nonlinear layers in the residual block. Fig 1 depicts overview of deep residual learning. The residual learning keeps information of the previous information and connects to the stacked nonlinear layer. In this work, we set F as the massive transformer encoder structure and x as the representation of the RegNet backbone.

3.2. RegNet Backbone

In [19], the authors start AnyNet design space. The input is fed into the simple stem, followed by the body that contains 4 stages operating at progressively reduced resolution, and then head to predict the classes. The AnyNet design space creates models that are experimented with the combination of different parameters based on the ImageNet dataset. They narrow down the design space and arrive

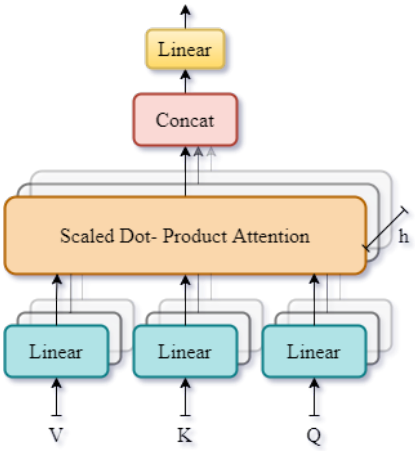

Figure 2. Overview of Multi-Head attention Module.

at the optimized RegNet design space with simplicity and regular networks. To generate RegNet design space, the authors introduce parameters of network structure contain: depth d, initial width w_0, slope w_a, width parameter w_m and bottleneck b and group g. These parameters are different to obtain different RegNets with different properties. In this work, we employ RegNetY with better performance than Efficientnet for most flop regimes.

3.3. Transformer Encoder

In [22], the authors introduce multi-head attention has several attention layers in parallel. In the attention function, a query maps the key-value pairs to an output. The output of attention is a weighted sum of the value, where weight is the specific computations of a query with the corresponding key. In multi-head attention, the attention layer runs independently and their outputs are concatenated and linearly transformed into the new space with the expected dimension. Fig 2 is overview of multi-head attention. In multi-head attention, the attention heads are enlarged with flexible dimensions for our study. The authors recommend the transformer architecture that has the encoder-decoder structure to build global dependencies for the connection of input and output. The transformer encoder including a stack of multi-head attention mechanisms and feed-forward networks can map an input sequence to representation features. In this study, we employ the transformer encoder as the embedding layer to extract robust representations from the RegNet backbone.

3.4. Focal Loss

Focal loss is introduced in [17] reshaping the standard cross entropy loss to solve class imbalance.

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}log(p_t)$$

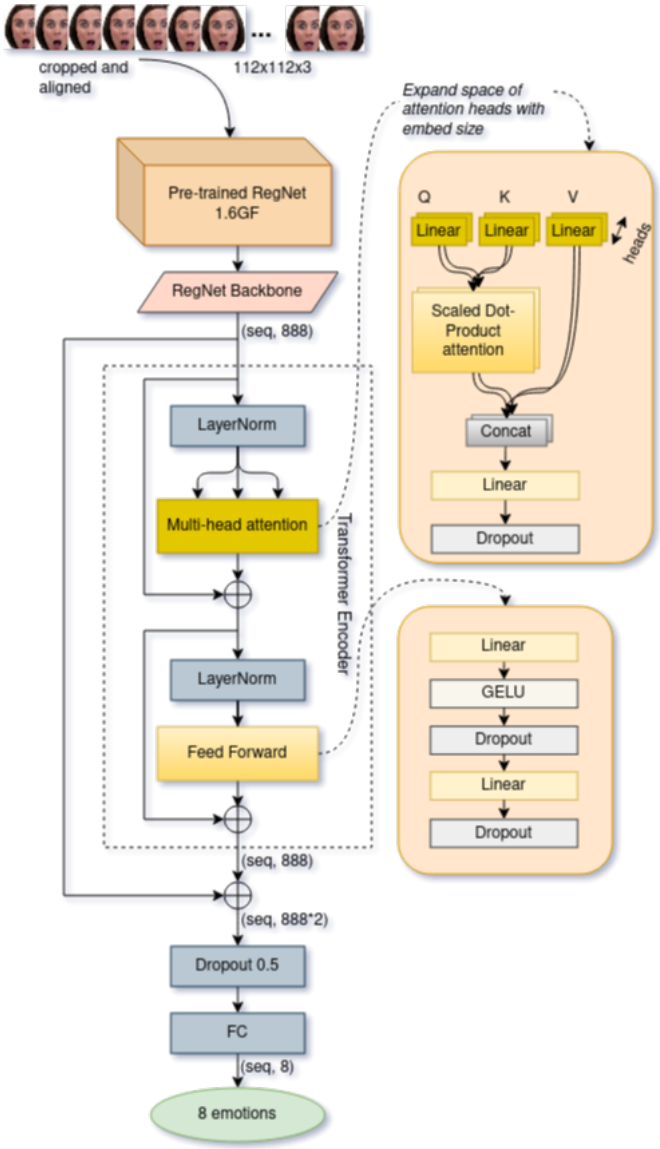

Figure 3. Detail of our architecture for facial expression classification.

where parameter $\alpha_t \geq 0$ and $\gamma \in [0, 5]$

4. EXPERIMENTS

4.1. Dataset

The large-scale in-the-wild Aff-Wild2 database contains 548 videos with approximately 2.7 million frames. The training process is conducted on the Aff-Wild2 database. Data is used containing annotations on 6 basic expressions including Anger, Disgust, Fear, Happiness, Sadness, Surprise, plus Neutral state, and Other which denotes emotional expressions other than the 6 basic states. The par-

Parameters	Transformer Encoder
Depth	1
Head Number	2
Embedded Dimension	64
Feedforward Dimension	512
Dropout Value	0.

Table 1. Hyperparameter of our architecture.

Model	Validation	Test
Baseline (VGG16) [9]	23	20.50
Our Model (Attention)	29.11	26.32
Our Model (Transformer)	**30.35**	**28.60**

Table 2. Expression classification results of our model on the validation set and the test set.

Team	F_1
Netease Fuxi Virtual Human [27]	35.87
IXLAB [5]	33.77
AlphaAff [24]	32.17
HSE-NN [20]	30.25
PRL (our method)	**28.6**
dgu [7]	27.2
USTC-NELSLIP	21.91
Baseline [9]	20.50

Table 3. Comparison of Expression classification results on the test set in ABAW3 Competition.

ticipants are provided with frames in RGB color space from the database. The images are cropped and aligned with an input size of 112x112 from the video.

4.2. Training Details

The network is trained on the Pytorch framework. We use Adam optimization [8] to update the weights. We opt Focal Loss [17] for the classification task of eight emotions with $\alpha = 0.9$ and $\gamma = 2.0$. The training process automatically finds the best learning rate. The batch size of 16 is trained during the training process. Our model learns the epoch of 30 and saves the best performance on the validation set. In this work, the final result is evaluated across the average F_1 score of 8 emotion categories:

$F_1^{final} = \frac{\sum F_1^{expr}}{8}$ where F_1^{expr} is F_1 score of each expression.

4.3. Results

We report results by F_1 score in Table 2 on both the validation set and the test set. In the baseline [9], the authors perform the pre-trained VGG16 network on the VGGFACE dataset and get softmax probabilities for the 8 expression predictions. They archive the F_1 score of 23% and 20.50% on the validation set and the test set, respectively. In the proposed model, we not only use the backbone representation but also combine the additional representation of the transformer encoder. As a result, this fusion has a lot of information so that the dropout layer is applied immediately after. Our model performs better than the baseline. We try combining the backbone representation and representation using multi-head attention mechanism. This fusion doesn't perform better than the transformer mechanism. This shows the effectiveness of the transformer as an embedding layer to encode the salient information of the backbone. In addition, we also show comparison with other participating methods in Table 3 for expression classification task. Our team is bold line has a better result than baseline result and ranks 5^{th} in the rankings.

5. CONCLUSION

In the 3rd Affective Behavior Analysis In-The-Wild (ABAW3) Competition, we have the opportunity to contribute research results in the field of human-computer interaction. Our proposed model performs the expression classification task based on videos of the Aff-Wild2 database. We recommend the fusion of the robust representative features from deep neural layer branches including the pre-trained RegNet model and transformer encoder. Results show the effectiveness of fusion for facial expression classification task.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2020R1A4A1019191) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1I1A3A04036408). This work was also supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub).

References

[1] Didan Deng, Liang Wu, and Bertram E Shi. Iterative distillation for better uncertainty estimates in multitask emotion recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3557–3566, 2021. 2

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[3] Paul Ekman, E Richard Sorenson, and Wallace V Friesen. Pan-cultural elements in facial displays of emotion. *Science*, 164(3875):86–88, 1969. 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[5] Jae-Yeop Jeong, Yeong-Gi Hong, Daun Kim, Yuchul Jung, and Jin-Woo Jeong. Facial expression recognition based on multi-head cross attention network. *arXiv preprint arXiv:2203.13235*, 2022. 2, 4

[6] Yue Jin, Tianqing Zheng, Chao Gao, and Guoqiang Xu. A multi-modal and multi-task learning method for action unit and expression recognition. *arXiv preprint arXiv:2107.04187*, 2021. 2

[7] Jun-Hwa Kim, Namho Kim, and Chee Sun Won. Facial expression recognition with swin transformer. *arXiv preprint arXiv:2203.13472*, 2022. 2, 4

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[9] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. *arXiv preprint arXiv:2202.10659*, 2022. 1, 4

[10] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 794–800. 1

[11] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019. 1

[12] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multitask learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021. 1

[13] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, pages 1–23, 2019. 1

[14] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 1

[15] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021. 1

[16] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3652–3660, 2021. 1

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3, 4

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2

[19] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10428–10436, 2020. 2

[20] Andrey V Savchenko. Frame-level prediction of facial expressions, valence, arousal and action units for mobile devices. *arXiv preprint arXiv:2203.13436*, 2022. 2, 4

[21] Phan Tran Dac Thinh, Hoang Manh Hung, Hyung-Jeong Yang, Soo-Hyung Kim, and Guee-Sang Lee. Emotion recognition with incomplete labels using modified multi-task learning technique. *arXiv preprint arXiv:2107.04192*, 2021. 2

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[23] Lingfeng Wang, Shisen Wang, Jin Qi, and Kenji Suzuki. A multi-task mean teacher for semi-supervised facial affective behavior analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3603–3608, 2021. 2

[24] Fanglei Xue, Zichang Tan, Yu Zhu, Zhongsong Ma, and Guodong Guo. Coarse-to-fine cascaded networks with smooth predicting for video facial expression recognition. *arXiv preprint arXiv:2203.13052*, 2022. 2, 4

[25] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild'challenge. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1980–1987. IEEE, 2017. 1

[26] Wei Zhang, Zunhu Guo, Keyu Chen, Lincheng Li, Zhimeng Zhang, Yu Ding, Runze Wu, Tangjie Lv, and Changjie Fan. Prior aided streaming network for multi-task affective analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3539–3549, 2021. 2

[27] Wei Zhang, Zhimeng Zhang, Feng Qiu, Suzhen Wang, Bowen Ma, Hao Zeng, Rudong An, and Yu Ding. Transformer-based multimodal information fusion for facial expression analysis. *arXiv preprint arXiv:2203.12367*, 2022. 2, 4