Emotion Recognition With Sequential Multi-task Learning Technique

Phan Tran Dac Thinh*
Chonnam National University
phantrandacthinh2382@gmail.com

Hoang Manh Hung*
Chonnam National University
hung.hoangmanh96@gmail.com

Hyung-Jeong Yang
Chonnam National University
hjyang@jnu.ac.kr

Soo-Hyung Kim
Chonnam National University
shkim@jnu.ac.kr

Guee-Sang Lee†
Chonnam National University
gslee@jnu.ac.kr

Abstract

The task of predicting affective information in the wild such as seven basic emotions or action units from human faces has gradually become more interesting due to the accessibility and availability of massive annotated datasets. In this study, we propose a method that utilizes the association between seven basic emotions and twelve action units from the AffWild2 dataset. The method based on the architecture of ResNet50 involves the multi-task learning technique for the incomplete labels of the two tasks. By combining the knowledge for two correlated tasks, both performances are improved by a large margin compared to those with the model employing only one kind of label.

1. Introduction

Affective computing aims to transfer the understanding of human feelings to computers, so they could recognize humans' emotional states and be applied to multiple advanced areas such as education or health service. Affective states can be decided by a wide range of sources in three main categories, namely visual, auditory and biological signals. Visual information, especially facial clues, is the most important and most adopted data due to high availability, interpretability and strong pertinence to emotional states.

To analyze the affective states, Ekman [2] introduced the six basic emotions, i.e., anger, disgust, fear, happiness, sadness and surprise. Those categorical values are extensively applicable to human beings but this is not the only way to perceive the emotional states. In terms of dimensional model, they can be represented as continuous values, namely valence and arousal. Valence shows how positive or negative the emotion is while arousal measures the agitation level which is from non-active to ready to act. Furthermore, according to the Facial Action Coding System (FACS) [3], facial movements which are defined as Action Units (AUs) are recorded to interpret emotions.

Affective Behavior Analysis In-The-Wild (ABAW) 2021 [6] is a competition with the primary goal of improving the machines' capability of understanding human feelings, emotions and behaviors. The competition provides the massive dataset called AffWild2 [7–12, 18] about emotions in the wild with annotations for seven basic emotions, AUs and valence/arousal. There are three tasks corresponding to the three types of annotations and the dataset contains the videos and also the cropped and aligned images extracted from those videos.

In this paper, since the dataset does not have complete labels for seven basic expression classification and facial action unit detection, we propose a modified multi-task learning technique for ResNet50 as the main model for implementing both tasks. Moreover, the data for seven emotions is highly imbalanced towards more common emotions such as neutral and happy, so we employ the Focal Loss to counter this effect. The detail of our work will be described in the next section.

2. Proposed Method

2.1. Preprocessing

The AffWild2 dataset provides the cropped and aligned images that are extracted from the videos. We use them for both training and validation stages and did not use other tools to acquire the images from the videos. The input size for the model is 112 x 112 and RGB color space is applied. The images are normalized before inputting to the model and no augmentation techniques are used to enlarge the dataset. The audios are not adopted for training in our method since not all videos contain sounds and the sounds in some cases are noise from the environment or human activities. Without proper processing or an adequate mecha-

*Contributed equally
†Corresponding author

nism to analyze the audios, they probably cause ambiguity and drop of performance to the main model.

2.2. Model Structure

ResNet50 [4] is the backbone of our deep learning network. Commonly, the pretrained weights on the ImageNet [14] are used to accelerate and enhance the training performance. In the field of emotion recognition, thanks to the existing works on emotion recognition, we opt for the pretrained weights on the VGGFace2 dataset [1]. The VGGFace2 dataset is not only large in the amount of images but also varied in the number of subjects and covers a large range of pose, age and ethnicity too. This coincides with the concept of AffWild2 dataset which is not dependent on the context nor the age, gender, ethnicity, social status, etc. The fully connected layer is cut and then the pretrained weights are loaded on the main backbone. A new dense layer with 512 and two dense layers of 7 or 12 neurons are sequentially added according to the specific task.

The model takes the input as static images. Because a lot of images do not have both labels for seven emotions and AUs, we need to have a particular training scheme to better learn the shared knowledge between two tasks. The training scheme is shown as below:

Algorithm 1: The training strategy for proposed method

Input:
Set images which has 7 expression as target E
Set images which has 12 action units as target A
Set images which has both types of labels B
Number of epochs T
Output:
One label for 7 expression e
Multi-labels for 12 action units a
for $t = 1, t \leq T$ **do**
  for $i = 1, i \leq len(E)$ **do**
    $e = Model(E_i)$
    $Loss = F(e, \theta_i)$
    Update weight θ_i
  end
  for $i = 1, i \leq len(A)$ **do**
    $a = Model(A_i)$
    $Loss = F(a, \theta_i)$
    Update weight θ_i
  end
  for $i = 1, i \leq len(B)$ **do**
    $e, a = Model(A_i)$
    $Loss = F(e, a, \theta_i)$
    Update weight θ_i
  end
end

2.3. Loss Function

For seven basic emotion classification, we use Focal loss [13]. The dataset for expression classification is imbalanced so we apply this loss to deal with this problem. For facial action unit detection, we use binary cross entropy loss for each action unit. The total loss is the summation of the two above losses with the same weight.

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise} \end{cases} \quad (1)$$

where p is probability for the class with label y

$$FLoss(p_t) = -\alpha_t (1 - p_t)^{\gamma} log(p_t) \quad (2)$$

The value of α is set as 2 and the value of γ is set as 0.25.

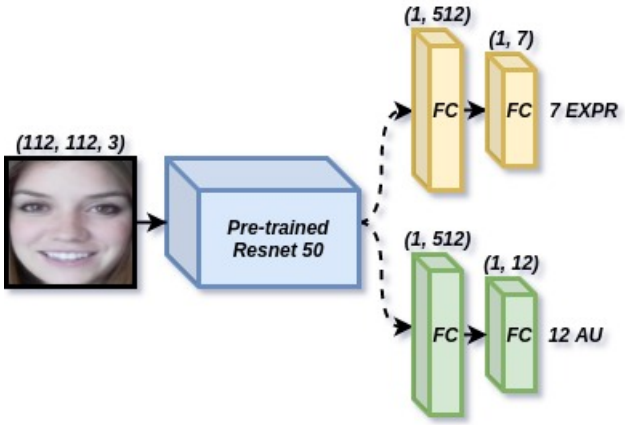

Figure 1. Overview system of proposed method

3. Experimental Results

3.1. Training Setup

The training process is optimized by the Adam optimizer. We use GPU RTX 2080Ti as the hardware and Pytorch framework as the software. The mini-batch has size 256. To regularize the training process and accelerate the convergence of the model, we use the Cosine Annealing as the learning rate scheduler with the starting learning rate of 0.001. The testing models for the most time achieve the best results only after 10 epochs on the validation set.

3.2. Results

Table 1 shows the results from our experiment on the validation set of emotion classification. By using the pretrained weight from the EmotionNet [16] dataset, the model when only trains with seven emotion label gets the performance metric of 0.462. After apply the multi-task learning

Table 1. Seven basic emotion classification result on the validation set

Expression	F1_Score	Accuracy	0.67* F1 + 0.33* Acc
Baseline	0.3	0.5	0.366
ResNet50 (EmotionNet)	0.395	0.598	0.462
ResNet50 (VGG-Face2) (Multitasking)	0.494	0.684	0.556
ResNet50 (VGG-Face2) (Shared backbone)	0.675	0.791	0.713
ResNet50 (VGG-Face2) (Multitasking) (Focal Loss)	0.724	0.826	0.757

Table 2. Facial action unit detection result on the validation dataset

Action Units	F1_Score	Accuracy	0.5* F1 + 0.5* Acc
Baseline	0.22	0.4	0.31
ResNet50 (EmotionNet)	0.439	0.878	0.659
ResNet50 (VGG-Face2) (Shared backbone)	0.427	0.883	0.655
ResNet50 (VGG-Face2) (Multitasking)	0.566	0.895	0.731

Table 3. Seven basic emotion classification result on the testing dataset

Method	F1_Score	Accuracy	0.67* F1 + 0.33* Acc
[20]	**0.763**	**0.806**	**0.777**
[5]	0.683	0.771	0.712
[15]	0.475	0.732	0.560
[17]	0.407	0.673	0.495
Ours	0.604	0.729	0.646

Table 4. Action units result on the testing dataset

Method	F1_Score	Accuracy	0.5* F1 + 0.5* Acc
[20]	**0.506**	0.888	**0.697**
[5]	0.489	**0.892**	0.690
[19]	0.435	0.888	0.662
[15]	0.394	0.874	0.634
Ours	0.461	0.877	0.669

technique and using the pretrained weight from the VGGFace2 dataset, the metric improves by nearly 0.1. We also implement the shared backbone architecture between the two models of two tasks and get the performance metric of 0.713. The Focal loss which is used to counter the effect of imbalance dataset of seven emotions achieved the best results of 0.757. Table 2 displays the results from our experiments on the validation set of action unit detection. The application of shared backbone architecture does not improve the performance from using only annotation from the twelve action units in this case. However, the multi-task learning technique helps us attain the performance metric of 0.731. Table 3 and Table 4 compare our results and the results of other teams on the test dataset. We got the third places on both tasks. To achieve those results, we combine the training and validation dataset into one dataset and use K-Fold validation to get the final prediction. This technique is better than learning only the training dataset, which may not be generalized.

4. Conclusion

In this paper, we present our experiments on the two tasks of emotion classification and action unit detection. ResNet50 with pretrained weight on the VGGFace2 dataset produces good results on the AffWild dataset and proposed training scheme with the application of multi-task learning enhances the performance by a considerable margin on both tasks. Moreover, Focal loss is suitable for solving the imbalance problem on the dataset of seven emotions.

5. Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (NRF-2020R1A4A1019191) and also supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A3B05049058).

References

[1] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[2] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[3] E Friesen and Paul Ekman. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 3(2):5, 1978.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceed-*

ings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[5] 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.

[6] Dimitrios Kollias, Irene Kotsia, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. *arXiv preprint arXiv:2106.15318*, 2021.

[7] Dimitrios Kollias, Attila Schulc, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 637–643. IEEE, 2020.

[8] 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.

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

[10] 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*, 127(6):907–929, 2019.

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

[12] 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.

[13] 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.

[14] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[15] Lingfeng Wang and Shisen Wang. A multi-task mean teacher for semi-supervised facial affective behavior analysis. *arXiv preprint arXiv:2107.04225*, 2021.

[16] Zijun Wei, Jianming Zhang, Zhe Lin, Joon-Young Lee, Niranjan Balasubramanian, Minh Hoai, and Dimitris Samaras. Learning visual emotion representations from web data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13106–13115, 2020.

[17] Sachihiro Youoku, Takahisa Yamamoto, Junya Saito, Akiyoshi Uchida, Xiaoyu Mi, Ziqiang Shi, Liu Liu, and Zhongling Liu. Multi-modal affect analysis using standardized data within subjects in the wild. *arXiv preprint arXiv:2107.03009*, 2021.

[18] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: valence and arousal'in-the-wild'challenge. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 34–41, 2017.

[19] Chenggong Zhang, Juan Song, Qingyang Zhang, Weilong Dong, Ruomeng Ding, and Zhilei Liu. Action unit detection with joint adaptive attention and graph relation. *arXiv preprint arXiv:2107.04389*, 2021.

[20] Wei Zhang, Zunhu Guo, Keyu Chen, Lincheng Li, Zhimeng Zhang, and Yu Ding. Prior aided streaming network for multi-task affective recognitionat the 2nd abaw2 competition. *arXiv preprint arXiv:2107.03708*, 2021.