Model Level Ensemble for Facial Action Unit Recognition at the 3rd ABAW Challenge

Wenqiang Jiang*, Yannan Wu*, Fengsheng Qiao*
Liyu Meng, Yuanyuan Deng, Chuanhe Liu
Beijing Seek Truth Data Technology Co., Ltd.

Abstract

In this paper, we present our latest work on Action Unit Detection, which is a part of the Affective Behavior Analysis in-the-wild (ABAW) 2022 Competition [15]. Our proposed network is based on the IResnet100 [6]. First of all, we utilize feature pyramid networks (FPN) [25] and single stage headless (SSH) [29] to enlarge the receptive field and extract more facial texture features. Then we employ the ML-ROS data balancing [4] and the BCE Loss plus Multi-label Loss to solve the multi-label imbalance problem. We also use three different models as the base model to fine-tune the Aff-Wild2 dataset. The pre-train backbones are the AU detection model, expression model and face recognition model. Finally, we adopt an ensemble methodology to get the final result. Our f1 score achieved 49.82 on the AU test set and ranked second in this challenge with a very small difference from the first team 49.89.

1. Introduction

As an important part of Artificial Intelligence and Human Interaction, affective computing has arisen more and more attention. Meanwhile, it has lots of applications in many fields, such as customer satisfaction surveys, financial anti-fraud, psychological analysis, etc.

The 3th ABAW Competition 2022 is large-scale in the wild emotion database which is held by Dimitrios Kollias [19] [21], etc. It provides Aff-Wild2 which consists of three kinds of emotional databases including categorical expression (such as happy, angry, sad), valence arousal, and 12 facial action units. Aff-Wild2 has 564 videos downloaded from YouTube. There is variety in ethnics, poses, ages, etc. [40] [19] [20] [17] [16]

Different from seven basic categorical expressions and valence arousal, action units (AU) describe facial muscle movements developed by Paul Ekman in the 1970s [7]. Action units usually have concurrence. For example, AU25 (lips part) and AU26 (jaw drop) often occur at the same time.

In this paper, we address the AU task in ABAW 2022. We analyze prior methods of facial action recognition in section 3. In section 4, we present our approach of data balancing, model structure, loss function, and post-processing. Details about the dataset, evaluation metric, experiment settings, and ablation study are illustrated in section 5. We conclude our work in Section 6.

2. Related Works

In this section, we briefly review the latest studies of facial action unit (AU) recognition including some prior methods in the ABAW 2021 competition.

Since much effort and time is needed to annotate AU, Li Yong et al. [24] and Niu Xuesong et al. [31] try to learn the representation of AU without lots of AU annotations. They proposed a self-supervised framework with an Autoencoder structure. Although they only need a small amount of AU annotations, their results can not outperform the supervised learning method.

Fan Y et al. [8] and Jacob G et all. [13] try to learn the relationships of different AUs. Fan Y et al. use Graph Neural Network to learn the concurrence of AUs, while Jacob G et al. firstly add Transformers to their network to the relationships of AUs.

There are several papers about ABAW 2022 Challenge. Yue Jin et all. [14] proposed a method for Action Unit and Expression Recognition utilizing audio and visual information of the Aff-Wild2. Researchers from Netease Fuxi AI Lab designed a multi-task streaming network [41] which can learn the intrinsic relations among three emotion tasks.

Although both of them belong to multi-task emotion recognition which is not allowed in the uni-task AU challenge this year, their approaches demonstrated that different emotion tasks including categorical emotions, valence arousal, and action units can promote each other. In our work, we get pre-train models from the training action unit,
expression, and face recognition dataset. These pre-trained models can provide good weights when fine-tuned with AFF-Wild2.

Yue Jin et al. [14] use a sequence model to tackle the visual and audio information. However, we find that sequence model such as LSTM and Transformer [36] leads to a bad result in our experiments. We analyze that AU labels in Aff-Wild2 are discrete (0 or 1), not continuous as in the valance arousal task, so we do not adopt the sequence model in our work.

3. Method

3.1. Overview

Our workflow is shown in Figure 1. The model can extracts face features effectively by the pre-trained models, which is trained on the emotion and face recognition datasets. Firstly, we get three pre-trained models for each dataset. Then we fine-tune the models on the Aff-Wild2 dataset. Also we choose the best checkpoint for each AU prediction. Finally, we ensemble these 12 checkpoints to get the final result [38]. Since participants need to predict each frame in the AU test set (videos format), as the post processing, we smooth the output logits with a mean filter using a sliding window over the video sequence.

In section 3.2 we proposed our method of data balance for the AU dataset in Aff-Wild2. In section 3.3, the model structure is presented. As in section 3.4, BCE loss plus Multi-label loss are adopted in our training stage. At last, in section 3.5 the post-processing of smoothing is illustrated in detail.

3.2. Data Balancing

Facial action unit recognition is a multi-label visual task in deep learning. There usually exist label imbalance problems in the multi-label task. Table 1 shows that the numbers of 12 AU in Aff-Wild2 vary in a wide range. Data balancing is very difficult because of label concurrence. Several papers are proposed to solve this problem. Wu Tong designed loss functions [39] to solve this problem. Other scientists use data sampling to alleviate data unbalance. In this paper, we make the Aff-Wild2 AU dataset more balanced with ML-ROS method. We can see the promotion to model performance of data balance in Table 1.

3.3. Model Structure

We use IResnet100 as the backbone. IResnet was chosen because it offers three major improvements over Resnet [9]: the flow of information through the network layers, the residual building block, and the projection shortcut. To improve the flow of information through the network, each stage is divided into three parts: one Start ResBlock, a number of Middle ResBlocks, and one End ResBlock. Without increasing the computational overhead, the residual blocks are reconstructed using grouped convolution instead of 1×1 convolution in order to improve the accuracy. The projection shortcut reduces the information loss, improving the overall recognition performance of the network and a combination of 3×3 convolution with a step size of 2 and 3×3 max pooling with a step size of 2 is used for sampling.

During the experiment, we found that increasing the texture feature information of the face is helpful for the classification of AU. Due to the increase of the network depth, the semantic features are more abundant but the texture features will be lost, so we added the feature pyramid networks (FPN) and single stage headless (SSH) modules [5] to increase the texture information and receptive field of the face. see Figure 2.

At the same time, we flatten the features of each layer passing through feature pyramid networks (FPN) and single stage headless (SSH) modules and splicing the features of each layer to output 512 dimensions of features through a
fully connected layer. And in order to make the network pay more attention to a certain part, we added the Coordinate Attention module [11] to the shallow and deep layers of the network. The experimental results (in section 4) show that the feature information obtained in this way contains more texture features, and the classification effect is better than the previous AU classification.

3.4. Loss Function

For the AU dataset in Aff-Wild2, the distribution of each AU in the training set and validation set is shown in Table 1.

We can see from the chart that the training set and validation set is extremely imbalanced, especially for AU15, AU23 and AU24. Because Action Unit Detection is a multi-label problem, Data Augment cannot solve the data imbalance problem. Therefore, we try to solve this problem from the loss function.

\[
l_{i} = -w_n[y_i \cdot \log \sigma(x_i) + (1 - y_i) \cdot \log(1 - \sigma(x_i))]
\]

\[
W = \{w_1, ..., w_N\}
\]

\[
\sigma(x_i) = \frac{1}{1 + e^{-x_i}}
\]

Equation 2 is a binary classification loss function, where \(W\) represents the loss weight of each AU, in our method, \(W = [1, 2, 1, 1, 1, 1, 1, 6, 6, 5, 1, 5]\), where \(x\) represents softmax output and the value range of \(x\) is \([0, 1]\), where \(y\) represents the target and takes either 0 or 1.

\[
m_{l}(x, y) = -w_n \sum y[i] \log((1 + \exp(-x[i]))^{-1})
\]

\[
+ (1 - y[i]) \log \left( \frac{\exp(-x[i])}{1 + \exp(-x[i])} \right)
\]

Equation 5 is a multi-label loss function, where \(x\) represents softmax output and the value range of \(x\) is \([0, 1]\), where \(y\) represents the target and takes either 0 or 1.

\[
\text{total loss} = \text{multi label loss}(x, y) + \text{ce loss}(x, y)
\]

Finally, we add BCE loss and Multi-label together, as Equation 6.

3.5. Post Processing

Considering all participants need to predict each frame in the AU Test Set (sequence format), we smooth the logits generated by the last layer of the network with a mean filter using a sliding window on the sequences. In detail, for the \(j\)-th frame in the \(v\)-th video, its \(i\)-th AU predict value is \(p_v^{i,j}\), we replace it with a new predict value \(\hat{p}_v^{i,j}\) by averaging the values of its neighbors in the window (its width is \(w\)), which is centered as it:

\[
\hat{p}_v^{i,j} = \sum_{s=-w/2}^{w/2} p_v^{i,j+s}
\]

4. Experiments

4.1. Dataset

Three types of datasets are used to get the pre-trained models:

Aff-Wild2 dataset: For this Challenge, the Aff-Wild2 database will be used by all participants. In total, 547 videos of around 2.7M frames will be used that contain annotations in terms of 12 action units, namely AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU15, AU23, AU24, AU25, AU26.

Face recognition dataset: We use the Glint360K [1], which is the largest and cleanest face recognition dataset which contains 170M images of 360k IDs. The pre-trained models provide us base features of human faces, which is very important for AU recognition.

EXPR dataset: The facial expression model is pre-trained on the FER+ [2], the RAF-DB [23], and the AffectNet [28] dataset. The FER+ dataset is an extension of the original FER dataset, where the images have been relabelled into one of 8 emotion types: neutral, happiness, surprise, sadness, anger, disgust, fear, and contempt. AffectNet is a large facial expression dataset with around 0.4 million images manually labeled for the presence of eight facial expressions along with the intensity of valence and arousal. The Real-world Affective Faces Database (RAF-DB) is a dataset for facial expression. It contains 29672 facial images tagged with basic or compound expressions by 40 independent taggers.

AU dataset: We use authorized commercial dataset, which contains 7K high definition images. The data is labeled into 15 face action unit categories (AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU11, AU12, AU15, AU17, AU20, AU24, and AU26).

4.2. Evaluation Metric

There are two types of annotations for each AU in the Aff-Wild2 AU dataset: the absence of AU is annotated as 0 and the presence of AU is annotated as 1. We consider these two labels equally important in the AU metric, so the
Table 1. Numbers of positive and negative AU in both training set and validation set.

<table>
<thead>
<tr>
<th>tag</th>
<th>AU1</th>
<th>AU2</th>
<th>AU4</th>
<th>AU6</th>
<th>AU7</th>
<th>AU10</th>
<th>AU12</th>
<th>AU15</th>
<th>AU23</th>
<th>AU24</th>
<th>AU25</th>
<th>AU26</th>
</tr>
</thead>
<tbody>
<tr>
<td>neg</td>
<td>160K</td>
<td>68K</td>
<td>210K</td>
<td>363K</td>
<td>540K</td>
<td>468K</td>
<td>333K</td>
<td>38K</td>
<td>43K</td>
<td>35K</td>
<td>853K</td>
<td>101K</td>
</tr>
<tr>
<td>pos</td>
<td>62K</td>
<td>41K</td>
<td>69K</td>
<td>112K</td>
<td>178K</td>
<td>156K</td>
<td>113K</td>
<td>13K</td>
<td>11K</td>
<td>13K</td>
<td>289K</td>
<td>47K</td>
</tr>
</tbody>
</table>

F1 score we evaluate on the validation set is calculated as the F1 scores of the two labels, then we average them. Our F1 score is defined as:

\[
F_{\text{our}} = \frac{1}{2} \sum_{i=0}^{1} \frac{2 \times p_i \times r_i}{p_i + r_i}.
\]  

\[
F_{\text{AU, our}} = \frac{\sum_{\text{all}} F_{\text{our}}}{12}.
\]

\[
F_{\text{AU, our}} = \frac{1}{2} \sum_{i=0}^{11} \frac{2 \times p_i \times r_i}{p_i + r_i}.
\]

\[
F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.
\]

Among them, \(p_i\) means precision of the i-th label, and \(r_i\) means recall of the i-th label. In this case, the F1 score is calculated using the \(f_\text{score}\) function in scikit-learn and the average parameter is macro [5].

Different from us, the official evaluation method in the test set only calculates the F1 score of the positive sample. The F1 score is defined as:

\[
F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.
\]
The evaluation metric of ABAW 2022 was the average F1 score (i.e., macro F1 score) of all 12 AUs:

$$F_{AU} = \frac{\sum_{au} F_{au}}{12}.$$  \hspace{1cm} (11)

In our paper, the evaluation on the validation set is $F_{AU,our}$.

The submit evaluation result on the test set is $F_{AU}$.

4.3. Training and Testing

We use Iresnet100 as the backbone. Our framework input size is 112x112. The SGD [32] optimizer is used with a learning rate of 0.001, the momentum is 0.9, and weight decay is 5e-4, and with a batch size of 256. The total training epoch is set as 15 in the ABAW training dataset. The learning rate is divided 10 when training to the 4/6/8 epoch. We implement color jitter (30% chance of brightness, 30% chance of contrast, 30% chance of saturation, and 30% chance of hue) and random horizontal flip for data augmentation. The dropout rate is 0.6. All frameworks are implemented in PyTorch and the training environment is 4 * RTX-3090 GPUs.

4.4. Ablation Study

<table>
<thead>
<tr>
<th>structures</th>
<th>glint</th>
<th>$F_{AU,our}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>iresnet34</td>
<td>✓</td>
<td>0.507</td>
</tr>
<tr>
<td>iresnet50</td>
<td>✓</td>
<td>0.508</td>
</tr>
<tr>
<td>iresnet100</td>
<td>✓</td>
<td>0.534</td>
</tr>
</tbody>
</table>

Table 3. Results of different network on the validation set.

<table>
<thead>
<tr>
<th>post-processing</th>
<th>$F_{AU,our}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ensemble</td>
<td>0.731</td>
</tr>
<tr>
<td>ensemble+smooth</td>
<td>0.735</td>
</tr>
</tbody>
</table>

Table 4. Results of smooth on the validation set.

<table>
<thead>
<tr>
<th>cross validation</th>
<th>$F_{AU,our}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>fold 1</td>
<td>0.721</td>
</tr>
<tr>
<td>fold 2</td>
<td>0.705</td>
</tr>
<tr>
<td>fold 3</td>
<td>0.724</td>
</tr>
<tr>
<td>fold 4</td>
<td>0.689</td>
</tr>
<tr>
<td>fold 5</td>
<td>0.717</td>
</tr>
</tbody>
</table>

Table 5. Results of 5-fold cross on the validation set.

We use glint360 (glint) pre-trained models to improve by 14.4%. We use data balance (bal) to improve by 1.5%. We use Multi-label loss (mll) to improve by 2.1%. We add coordinate attention (ca) [12] module improve 4.4%. We use label smooth (ls) [34] to improve 5.9%. We use BCE loss + Multi-label loss (b+m) to improve 1.1%. We use feature pyramid networks (FPN) to improve 1.9%. We use bigger batch size 256 (bs256) to improve 0.3%. We use Single Stage Headless (SSH) to improve 0.3%. We add additional training dataset improve by 0.6%.

Comparation of base model structure Table 3 shows the results of different network results on the validation set, Iresnet100 performs better than the other two models achieving 0.534.

Post processing As shown in Table 4, the results of logits smooth can improve the result about 0.4%.

Cross validation As shown in Table 5, the results of 5-fold cross-validated on the validation set. During the testing phase, the 5-fold cross-validation achieved the best results.

Test result As shown in Table 7, we achieved 49.82 ($F_{AU}$) on the AU test set and ranked second in this challenge with a very small difference from the first team 49.89.

4.5. Ensemble Model

Previous work [27] [26] has demonstrated the effectiveness of model ensembling. In numerous experiments, we also adopted a model ensemble strategy, which ensembled the model with the highest F1 score for each AU, and obtained the final result on the validation set. As shown in Table 6, from column 1 to column 12, each column indicates the best model with the highest F1 score for each AU. It is chosen from different checkpoints of different models. For example, in column 2, 'AU2' indicates the best model for AU2. Value 0.735 is the $F_{1,our}$ score of this model on validation set. The last column is the $F_{AU,our}$, which
Table 6. Numbers of positive and negative AU in both training set and validation set.

<table>
<thead>
<tr>
<th>Team</th>
<th>( F_{AU} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Netease Fuxi Virtual Human</td>
<td>0.4989</td>
</tr>
<tr>
<td>SituTech</td>
<td>0.4982</td>
</tr>
<tr>
<td>PRL [30]</td>
<td>0.4904</td>
</tr>
<tr>
<td>STAR-2022 [37]</td>
<td>0.4883</td>
</tr>
<tr>
<td>HSE-NN [33]</td>
<td>0.4731</td>
</tr>
<tr>
<td>ISIR DL [35]</td>
<td>0.4432</td>
</tr>
<tr>
<td>SCPRLab@CNU [10]</td>
<td>0.4206</td>
</tr>
<tr>
<td>USTC-AC</td>
<td>0.4157</td>
</tr>
<tr>
<td>baseline</td>
<td>0.3650</td>
</tr>
</tbody>
</table>

Table 7. The overall results on the test dataset means the average of 12 AU \( F_{1,our} \) scores.

5. Conclusion

For the AU task in ABAW Competition 2022, we design our backbone with IResnet100 adding FPN and SSH. To train a high-performance model, we first utilize three different datasets (au, expression, and face recognition dataset) to get pre-trained models. Then we fine-tune these models on Aff-Wild2. The problem caused by data imbalance is alleviated by using BCE loss, Multi-label loss, and ML-ROS. Finally, the best checkpoint for each AU is chosen and then ensembled. As for the post-processing method, the predicted logits is smoothed using a mean filter by sliding a window over frames in the video. We achieved second place on the AU challenge with an F1 score of 49.82, which demonstrates the effectiveness of our method.

References


[33] Andrey V. Savchenko. Facial expression and attributes recognition based on multi-task learning of


[37] Lingfeng Wang, Shisen Wang, and Jin Qi. Multimodal multi-label facial action unit detection with transformer, 2022.


