Multiple Transfer Learning and Multi-label Balanced Training Strategies for Facial AU Detection In the Wild

Sijie Ji*[1], Kai Wang*[2], Xiaojiang Peng†[2], Jianfei Yang[1], Zhaoyang Zeng[3], and Yu Qiao[2]

[1] Nanyang Technological University Singapore
[2]ShenZhen Key Lab of Computer Vision and Pattern Recognition, SIAT-SenseTime Joint Lab,Shenzhen Institutes of Advanced Technology, Chinese Academy of Science
[3]Sun Yat-Sen University, China

Abstract

This paper[1] presents SIAT-NTU solution and results of facial action unit (AU) detection in the EmotiNet Challenge 2020. The task aims to detect 23 AUs from facial images in the wild, and its main difficulties lie in the imbalanced AU distribution and discriminative feature learning. We tackle these difficulties from the following aspects. First, to address the unconstrained heterogeneity of in-the-wild images, we detect and align faces with multi-task convolutional neural networks (MTCNN). Second, by using multiple transfer strategies, we pre-train large CNNs on multiple related datasets, e.g. face recognition datasets and facial expression datasets, and fine-tune them on the EmotiNetdataset. Third, we employ a multi-label balanced sampling strategy and a weighted loss to mitigate the imbalance problem. Last but not the least, to further improve performance, we ensemble multiple models and optimize the thresholds for each AU. Our proposed solution achieves an accuracy of 90.13% and F1 of 44.10% in the final test phase. Our Code is available at: https://github.com/kaiwang960112/enc2020_ au_ detection

1. Introduction

Facial action unit means visible muscle group changes on facial images. AU detection facilitates face perception and emotion theory research including emotion recognition [13], student engagement prediction [14], mental health diagnoses, deception detection and so on [17]. While the state-of-the-art AU detection methods [3, 20] yield good performance in the controlled scenarios, namely, images and videos collected in the laboratory, AU detection in the wild still faces challenges. AU have significant subject-dependent variations and too subtle to annotate, thus, realistic AU detection should require algorithms that are robust to intra-class variability, illumination conditions changes, plane rotations, low-resolution images and variations in pose and point of view.

Automated facial AU detection has been a vital research field for objectively describing facial actions, which has been studied for decades and many approaches have been proposed. Conventional methods were focused on extracting representative features and fine-tuning more robust classifiers [12, 10, 16]. With the development of deep learning, recent CNN-based or LSTM-based AU detection works frequently achieve new state-of-the-art performance in the literature. Zhao *et al*. [22] divided the input image to sub-regions trained by CNN independently and then merged back into one network followed by fully connected layers. Jaiswal *et al*. employed CNN to learn static region features and used LSTM to extract temporal dynamic features. Despite the current improvements, in-the-wild emotion recognition and AU detection remain an open problem for the computer vision community. To this end, EmotiW Challenge [5] consists of group emotion recognition [15] and predicting person engagement [19] in the wild, and EmotioNet Challenge [6] offers opportunity to obtain in-the-wild AU detection benchmark. The performance decline ascribes to those models learn from encodes expert prior knowledge instead of learning true AU features. In summary, AU detection is challenging because various AU make subtle facial appearance change over various regions at different scales. While some method achieved great results in terms of those constrained AU detection dataset [3],

*Equally-contributed first authors
†Corresponding author (xj.peng@siat.ac.cn)
[1]This work is partially supported by National Natural Science Foundation of China (U1813218,U1713208).

in-the-wild AU detection makes problem more difficult by introducing various of uncertainty and ambiguity, which remains exploited. Furthermore, previous methods typically focus on specifying challenge of AU detection, either exploring better features or address data imbalance problem.

Unlike previous methods that mainly leverage image-level feature, we transfer knowledge from face related tasks that give sub-level details. Face landmarks are coherently related to the muscle that may change to form AU. Thus, knowledge of face landmark can be transferred to AU detection to provide more precise AU locations and lead to better performance. AU is the basic component of emotion, and some AU are likely to appear in pairs. For example, AU 12 (i.e. lip corner puller) often presents happiness, which may occur together with AU 10 (i.e. upper lip raiser). Hence, transferring knowledge from facial expression recognition may offer underlying AU correlation features to help detect those AU with tiny muscle changes. We address the two tasks in parallel to train two models, in order to obtain multi-view features with conditionally independent so that enlarge the containing information by each view of futures. Besides, considering data imbalance problem, we embed the image distribution information in our model and jointly consider accuracy and recall performance when training.

2. Problem Definition

In this challenge of AU detection task, denote label 1 an occurrence of an AU and label 0 no occurrence of an AU. Those AU occluded are denoted by label 999. In this challenge, a total of 23 categories of AU are provided. Some AU might simultaneously occur in the same image while a lot of AU may not occur. Therefore, we define the problem to be a multi-label classification task with data imbalanced problem, which guides us to use Weighted Binary Cross Entropy loss:

$$L = -\frac{1}{C}\sum_{i=1}^{C} w_i \left[y_i \cdot \log \sigma(f_i(x)) + (1 - y_i) \cdot \log(1 - \sigma(f_i(x)))\right] \tag{1}$$

where C is the class of AU, y_i is ground-truth label of the occurrence for the ith AU, x_i denotes the probability predicted for the ith AU, σ denotes the sigmoid function and w_c is a weighted parameter to better cope with the data imbalance of AU by using the selective learning strategy [9].

In addition, since different emotion cannot happen at the same time, for one image, most of the Action Unit is labeled negative. Such characteristic requires our approach to truly be able to discriminate images with a certain AU present. In other words, instead of detecting non-present AU correctly, those present AU should be sensitively reported. False positives are not as important as true positives in this task. Therefore, accuracy (average of true positives and true negatives) metric is not enough to evaluate the performance of this task, F1-score balancing relationship between precision(true positive to false negative) and recall(true positive to false positive) should be taken as the evaluation metric.

3. Methodology

3.1. Overview

After defining the problem and seizing the main challenge, we design our framework mainly in three-fold, image data pre-processing, multi-view feature learning, and post-optimization threshold searching. Figure 1 gives the overall pipe-line of our framework. First, we use an algorithm to detect human face in the images and further align the face to cope with in-the-wild image heterogeneity of brightness, angles, and resolution. Second, after weighted random sampler, we conducted face recognition and facial expression recognition to generate multi-view representative deep features. Then, ensembling multi-model performs better generalization. Last but not least, instead of using a fix hard threshold to mapping the Sigmoid output probability vector to 0 and 1 for classification result, we adapt soft threshold searching in the validation phase which reconstructs data distribution for more reliable mapping and helps refine the final classification result.

3.2. Image Data Pre-processing

As mention above, the key point of in-the-wild AU detection is the unconstrained heterogeneity images. To prevent diversity scale, illumination and poses slowing down our model convergence, we use a well-known multi-task cascaded CNNs (MTCNN) framework [21] for jointly face detection and alignment to pre-process the huge in-the-wild images. This framework first employs several scales of images to build an image pyramid, then use Proposal Network to obtain multiple face region proposals, followed by Refine Network that rejects those overlapping and false bounding box. Finally, it go through Output Net to obtain facial landmarks positions. Figure 1 shows that this network help to remove irrelevant background and also deal with various pose problem.

3.3. Multi-view feature generation

To obtain multi-view feature, we use co-training approach. Consider tiny face muscle change leads to different AU, AU detection task need sub-face level feature, so the network should be powerful enough to learn deep fine-grained representative face features. Hence, we choose IR-152[4] as our backbone, which utilizes the power of residual nets to build deep models. To learn sub-face level feature, instead of learning from the ground up, we conduct face-

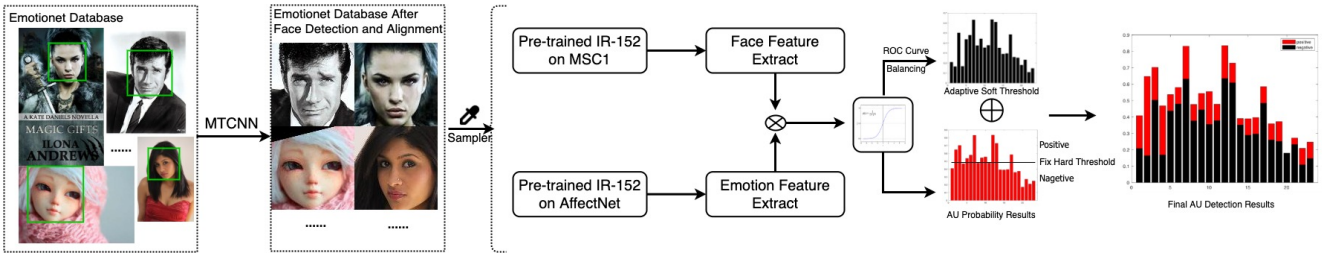

Figure 1. The pipeline of our approach.

level tasks first then fine-tuning based on face level tasks to obtain deep fine-grained sub-face level features. For the task selection, face recognition task is conducted because the most location of AU happens on the face landmark position, the facial expression recognition task is conducted because of some AU will occur simultaneously to form a specific emotion. The dataset we use for the face recognition task is MS-Celeb-1M [8] which contains 10,000,000 face images. For facial expression recognition task we use AffectNet [11], which contains more than 1,000,000 facial images. Affectnet also is in-the-wild dataset thus may have similar underlying ambiguity of Emotinet Database that we can take advantage of. Here, we parallel train these two task models instead of cascading these two tasks as part of our AU detection. The reason is that separately training will guarantee two-view features conditionally independent and enlarge the information provided by different views [1].

3.4. Multilabel Balancing Strategies

AU suffer from serious data imbalanced problems. We address this problem from two aspects, the angle of data and the angle of optimization.

Balance Sampling From data point of view, we weighted sampling the training samples by over-samples minority classes and under-samples majority classes. We use the MLROS algorithm[2] which independent of the classification algorithms used once the datasets have been preprocessed.

Selective Learning In multi-label problem, sampling one label changes the distribution for the other labels. So we use Selective Learning [9] method to adjust labels in batch-wise. First, the whole distribution of labels can be obtained by calculating all the labels across the database. For each label in each batch, if the batch distribution is equal to true distribution, selective learning does nothing. If the label is over-presented, selective learning will weight the negative samples by the ratio of positive sample to negative sample so that the negative samples effectively match the balanced target distribution. If the label is under-presented, reverse the above process, sampling from negative instances and weighting the positive instances by the ratio of negative sample to positive sample. The weights is denoted w_i in equation 1.

3.5. Model Ensemble

Recall that we simultaneously train two models to generate multi-view features. Model one pre-trained on MS-1M fine-tune by AU detection task which expects to learn sub-face level features. Model two pre-trained on Affectnet fine-tune by AU detection task which expects to learn emotion feature and underlying AU correlation feature. We further ensemble the two networks with an average of their predicted probabilities according to 23 AU, then use the sigmoid layer normalizing the final result mapping to 0 to 1. We demonstrate ensemble effective on ablation experiments.

3.6. Soft Thresholding

Although selective learning was being used at the training phase to handle data imbalance problem, which is allowing a deep network to learn a true representation of the data, rather than just the the bias of the training. We further consider the class skew problems on the final stage to refine the result. The output of the network is a discrete probability vector that shows the probability of an AU present. Typically, a threshold of 0.5 is used by default to convert predicted probabilities into class predictions, for example, those higher than 0.5, the classifier may report positive. However, the prediction probability distribution may suffer from skew distribution. The threshold can be adjusted in a reasonable way to increase model sensitivity or specificity while sensitivity and specificity have an inverse relationship need to trade-off. Here, we use ROC curving analysis [7] in the validation phase. The basic idea is to set each batch output probability as a threshold reference, calculate corresponding accuracy and F1 base on that threshold and select the one with the best result as a soft threshold. By utilizing this soft threshold, we further deal with the class skew problems in AU detection task. To take full advantage of $\sim$25K manual precisely annotated AU images, essentially, we reconstructing AU distribution, searching probability threshold start from near the ground truth distribution which helps speed up finding the best soft threshold.

4. Experiments

4.1. Dataset

Emotionet dataset is the first very large scale (∼950k) in-the-wild image dataset for AU detection. This year challenge focus on 23 different AU detection task which annotated by Emotionet Algorithm [6], some of those images contains small local occluders with different scale resolution of images. The accuracy of these annotations is about 81%. In this challenge, 325K images after augmentation(scaling and random occlusion) are being used for validation and final testing.

4.2. Implementation Details

For the model configurations, the initial learning rate of the network is 0.01 and is divided by 10 every 10 epoch. The weight decay is 0.00001 and the momentum is 0.9. We resize aligned face image to 224×224 as our network input with 16 picture as a batch. We use weighted random sampler choose data from dataset with shuffle. To better leverage manual precisely annotated data, we split it to cross validate our model performance.

4.3. Evaluation Criterion

To take full advantage of manual precisely annotated 25K image data, we split it by 5-fold cross-validation. We do this strategy because the Emotionet Dataset is annotated by algorithm, which has 19% error rate. After training by large scale dataset we fine-tune by expert data and use cross-validation to help our model reliable estimate the out-of-sample performance by reducing the variance. This strategy also gives us information on whether our model truly learns AU association feature and can adapt to unreliability across the database. Besides, accuracy is calculated to know overall classifies correction, F1 score is calculated to focus on when the actual value is positive, how often is the prediction correct. Final reference metric should be the average score of accuracy and F1.

4.4. Experiment Results

4.4.1 Backbone Selection

Table 1 demonstrate our experiment of backbone selection results follow our implementation details and evaluation strategy. Besides, we remove the 7x7 pooling layer in our task. This trick has been proved to improve the original ResNet in regards to AU detection task [18]. In the beginning, ResNet-18 shows satisfying accuracy result, however, the output result dominates by false positive, which shows that the model only learns prior knowledge and not sensitive enough when true AU occurrence. Therefore, the deeper network needs to be used. Compare to 50 layers and 152 layers, results shows that 152 layer gives a better balance between precision and recall, finally, ResNet-152 set as our backbone.

Table 1. Result on Preliminary Backbone Selection

Backbone	Acc	F1	Avg
ResNet-18	0.901	0.364	0.633
ResNet-50	0.905	0.409	0.657
ResNet-152	0.905	0.474	0.690

4.4.2 Ablation Experiment

Instead of directly using the residual convolution network to learn image scale features, knowledge from face-level feature like face skeleton point will help the network pay attention to more subtle features that essential to AU detection task. We separately pre-trained our model on MS-Celeb-1M [8] and AffectNet [11] to focus on face feature and emotion feature. Next, we fusion this two-view features together by ensemble the two models to get a better result. To address unconstrained images in-the-wild, MTCNN used beforehand to filter out noise introduced by various poses, different illuminations, small occlusions, and none face data. After MTCNN, the input image only contains an aligned face. In the end, we design adaptive threshold searching corresponding to different AU categories in the validation phase. Such algorithm reconstructs probability distribution on large scale image data in terms of different AU intensity. Table 2 demonstrate our method effectiveness and step by step improvement result.

Table 2. Ablation Results

Method	Acc	F1	Avg
F_face	0.921	0.384	0.652
F_emotion	0.925	0.429	0.677
F_face + F_emotion	0.925	0.494	0.710
Align + F_face + F_emotion	0.927	0.527	0.727
Align + F_ face + F_emotion + balancing	0.915	0.552	0.734

5. Conclusion

In this paper, we elaborate our solution for Emotionet Challenge 2020. We first detect and align face in large scale images database to tackle uncertainty of in-the-wild images. Second, we use selective learning and jointly co-regularization to handle typical data imbalance problem in AU detection. We further propose a scheme that focuses on learning sub-face level multi-view features and ensemble multi-model for better performance. Finally, we use adaptive thresholding algorithm to refine the results. Experimental results shows the effectiveness of our approach in terms of detecting large scale AU in-the-wild and we eventually obtain 0.6711 final score in test phase.

References

[1] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998. 3

[2] F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera. Addressing imbalance in multilabel classification: Measures and random resampling algorithms. *Neurocomputing*, 163:3–16, 2015. 3

[3] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE transactions on pattern analysis and machine intelligence*, 38(8):1548–1568, 2016. 1

[4] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 2

[5] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 423–426, 2015. 1

[6] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5562–5570, 2016. 1, 4

[7] T. Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006. 3

[8] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016. 3, 4

[9] E. M. Hand, C. Castillo, and R. Chellappa. Doing the best we can with what we have: Multi-label balancing with selective learning for attribute prediction. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2, 3

[10] S. Koelstra, M. Pantic, and I. Patras. A dynamic texture-based approach to recognition of facial actions and their temporal models. *IEEE transactions on pattern analysis and machine intelligence*, 32(11):1940–1954, 2010. 1

[11] 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*, 10(1):18–31, 2017. 3, 4

[12] M. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, pages 149–149. IEEE, 2006. 1

[13] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020. 1

[14] K. Wang, J. Yang, D. Guo, K. Zhang, X. Peng, and Y. Qiao. Bootstrap model ensemble and rank loss for engagement intensity regression. In *2019 International Conference on Multimodal Interaction*, pages 551–556, 2019. 1

[15] K. Wang, X. Zeng, J. Yang, D. Meng, K. Zhang, X. Peng, and Y. Qiao. Cascade attention networks for group emotion recognition with face, body and image cues. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 640–645, 2018. 1

[16] Z. Wang, Y. Li, S. Wang, and Q. Ji. Capturing global semantic relationships for facial action unit recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3304–3311, 2013. 1

[17] G. Warren, E. Schertler, and P. Bull. Detecting deception from emotional and unemotional cues. *Journal of Nonverbal Behavior*, 33(1):59–69, 2009. 1

[18] L. Xiong, J. Karlekar, J. Zhao, Y. Cheng, Y. Xu, J. Feng, S. Pranata, and S. Shen. A good practice towards top performance of face recognition: Transferred deep feature fusion. *arXiv preprint arXiv:1704.00438*, 2017. 4

[19] J. Yang, K. Wang, X. Peng, and Y. Qiao. Deep recurrent multi-instance learning with spatio-temporal features for engagement intensity prediction. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 594–598, 2018. 1

[20] S. Zafeiriou, A. Papaioannou, I. Kotsia, M. Nicolaou, and G. Zhao. Facial affect"in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 36–47, 2016. 1

[21] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 2

[22] K. Zhao, W.-S. Chu, and H. Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3391–3399, 2016. 1