Prompt Tuning for Facial Action Unit Detection in the Wild

VU NGOC TU, HUYNH VAN THONG, 김애라, 김수형
전남대학교 인공지능융합학과

tu369@jnu.ac.kr, vthuynh@jnu.ac.kr, arkim@jnu.ac.kr, shkim@jnu.ac.kr

Prompt Tuning for Facial Action Unit Detection in the Wild

Vu Ngoc Tu, Huynh Van Thong, Aera Kim, Soo-Hyung Kim
Dept. of Artificial Intelligence Convergence, Chonnam National University

Abstract

Facial Action Units Detection (FAUs) problem focuses on identifying various detail units expressing on the human face, as defined by the Facial Action Coding System, which constitutes a fine-grained classification problem. This is a challenging task in computer vision. In this study, we propose a Prompt Tuning approach to address this problem, involving a 2-step training process. Our method demonstrates its effectiveness on the Affective in the Wild dataset, surpassing other existing methods in terms of both accuracy and efficiency.

1. Introduction

The growing concern over mental health issues in society and the advancements in human-machine interaction systems have created a need for a better understanding of human emotions. Although significant progress has been made by researchers in tasks such as facial expression recognition, sentiment analysis, and speech emotion recognition, some challenges remain. This research focuses on facial action unit detection, a fine-grained multi-label classification task which recognizes different action units on the human face.

Recently, the pursuit of Large Language Models is on trend with different techniques to tune the model for different applications. Currently, due to the lack of the model's source code accessible and huge resource requirement, one of the most effective techniques is Prompting. Since first applied in GPT-3 [21], it has become the most popular tuning technique. Recent work proved that prompting is not only efficient for Natural Language Processing tasks but also can be utilized for Vision problems [1], which outperform traditional end-to-end Fine-tuning method.

In this work, we propose a novel approach for detecting Action Units, which involves the utilization of a Prompt tuning method. Our method draws inspiration from the Deep Visual Prompt Tuning technique introduced in [1] and aims to address the prevalent issue of imbalanced datasets. Specifically, our main contribution lies in the application of this approach to action unit detection, which has not been explored in previous research.

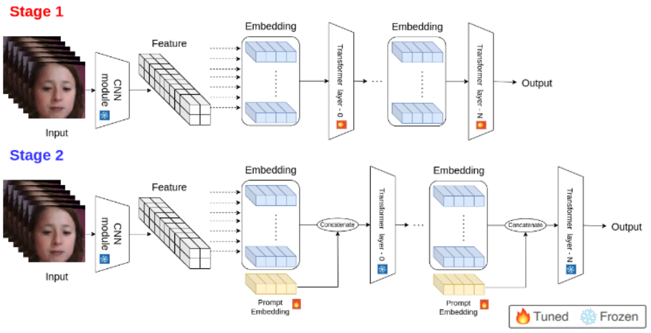
2. Related Works

Action Units Detection in the Wild: Action Units Detection has long been researched in the field of computer vision. This problem stems from the fact that different facial action unit combinations can represent different expressions. Each of these units describes the apparent changes caused by facial muscle movements. Facial Action Coding System (FACS) [20] is introduced to formally define these units. Based on this system, many benchmarks have been proposed to evaluate models' performance. However, these benchmarks are mostly conducted in the controlled laboratory environment, which limit performance validation in real-world situations. In 2017, Kollias et al. [9] introduce a new dataset which contains videos capturing human emotion in the wild. Since then, Aff-wild has become a standard benchmark for many Facial Action Units Detection methods. In this study, we specifically focus on this dataset to prove the effectiveness of our methods for practical usage.

Prompt Tuning for Vision model: In the computer vision domain, the emergence of the Transformer model in 2020 has resulted in the widespread adoption of the Vision Transformer (ViT) [8] as a fundamental or pivotal element in various methodologies. Nonetheless, the efficacy of these approaches has yet to achieve satisfactory levels. Despite the potential of the parameter-tuning technique in enhancing the performance of many ViT-based approaches, its effectiveness in facial action unit detection remains inadequate compared to training-from-scratch setting. Thus, we proposed using Prompt Tuning – a Transformer oriented tuning technique for the Action Units Detection problem.

3. Method.

Baseline model: Since Action Units Detection problem trying to detect the Human Facial Units in video. We employ Video Vision Transformer [8] as baseline. We use ViVit variation that comprises two main modules: spatial module and temporal module. However, as the pretrained ViT models are mainly trained on a general multi-class classification dataset. Hence, it is not feasible to directly use the whole pretrained model for the multi-label classification problem.



(Fig. 1) The training process of our proposed method.

In this case, we custom the ViT into multi-label classification model by altering the classification head of the model. Besides, we replaced some early layers of ViT backbone with a CNN based model. We adopt this configuration from previous methods because of its proven effectiveness. Our architecture is present in Fig. 1.

To be more specific, after getting the extracted feature F from the CNN backbone, with N transformer encoder layers, the feature is divided into m fixed-sized patches. These feature patch acts as the d dimensional embeddings. Then, we add them with the positional encoding vectors.

$$e_0^j = f_j + \text{Pos\_embed}$$

We denote the collection of feature patch embeddings. This *E* collection is defined as:

$$\mathrm{E}_i = \{e_i^j \in R^d | j \in \mathrm{N}, 1 \le j \le m\}$$

This collection will be the inputs to the (i+1)-th Transformer layer – L(i+1). Alongside the classification token ([CLS]) the ViT is formulated as:

$$[\mathrm{c_i}, \mathrm{E_i}] = \mathrm{L}_i(c_{i-1}, \mathrm{E}_{i-1});\ i = 1,2,\dots,\mathrm{N}$$

$$\mathrm{y} = \mathrm{Head}(\mathrm{c}_i, \mathrm{E}_i)$$

Where c is the [CLS] embedding, [., .] indicates stacking and concatenation on the sequence length dimension. Each layer L includes Multiheaded Self-Attention (MSA) and Feed-Forward Networks (FFN) together with LayerNorm and residual connections. Then, after the encoder layer, the embedding is fed to neural classification head to map it into a predicted class probability distribution.

Prompt Tuning for Action Units Model: Inspiring from Visual Prompt Tuning [2], we employ the prompt tuning into our baseline model. The set of prompt embeddings is denoted as p with same d dimension of original layers input embeddings. As the Deep Visual Prompt tuning shows its robustness in the original paper [2], we utilize this technique for our approach, i.e., we inserted prompt into the input of every Transformer block layer. For each layer, the collection of input learnable prompts is denoted as P. From there, the layer in the model become:

$$[x_i, \_, E_i] = L_i([x_{i-1}, P_i, E_{i-1}]); i = 1,2,3,\dots,\mathrm{N}$$

Training strategy: During the experiment phase, we noticed that since the two problems are trained for different problems, and the face image is much more detailed than the general training data. It requires so much time and resource to fully fine-tune the ViT model. Hence, we apply a 2-stage training phase for our model: End-to-end training phase and Prompt-tuning phase.

- End-to-end training phase: To reduce the training time while still achieving decent performance, we remove some layers of ViT model and train from scratch on the remaining layers while still freeze the CNN backbone.
- Prompt-tuning phase: The trained model is prompt-tuned on the smaller set of the training data. We freeze every layer in the stage 1 model and attach the prompt embeddings into the input of every Transformer layer. These embedding parameters are learnable.

4. Experiments

Dataset: The Affective Analysis in the Wild (Aff-Wild) dataset is first introduced in 2017 [19]. Until now, this dataset is still constantly updated by the authors [9-19]. Aff-Wild contains around 541 videos having 2.7 million frames with 438 subjects. There are 12 types of AU for detecting, namely AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU15, AU23, AU24, AU25, AU26. These identifier numbers are stipulated by Facial Action Coding System (FACS)[20].

Experiment setup: The whole experiment process is fully conducted with PyTorch framework. The CNN backbone module of the model is Regnet model. The network is optimized with the sigmoid focal loss function and SGD optimizer. There are 256 consecutive frames taken for each sample and the size of these frames is 112x112. For the CNN extractor, we set the number of dimensions for each feature is 440 and each time, the module will return 4 feature embedding. The number of encoder layers for Transformer is 8.

Metrics: One of the most frequently used metrics for Action Units Detection is Macro F1-score. This actually is the average of the F1 Score of all 12 AUs:

$$P_{AU} = \frac{(\sum_{au} F_1^a u)}{12}$$

<Table 1> Comparison between our approach and SOTA.

Method	Feature	Validation Score
Kollias et al. [9]	VGG-16	0.390
Hoai Le et al. [3]	Custom CNN	0.480
Wang et al. [4]	ResNet-18	0.501
Savchenko et al. [5]	EfficientNet	0.508
Baseline model	RegNet	0.497
Prompt-tuned	**RegNet**	**0.511**

Results: We compare our methods with other State-of-the-art in Tab. 1. From the table, we can observe that the highest validation score is achieved by the Prompt-tuned method, with a score of 0.511 – higher than the baseline 0.14. Moreover, we can see that Savchenko et al. [5] achieved the second-best validation score of 0.508 using the EfficientNet feature, which is slightly better than the ResNet-18 feature used by Wang et al. [4] with a validation score of 0.501. However, it is worth noting that both methods have more parameters than RegNet. This difference is reflected in the score of baseline methods with those ones.

We also demonstrate the effectiveness of our Prompt-tuning technique in Tab. 2. We apply the Prompt-tuning technique on Base Vision Transformer pretrained on 21k ImageNet [6]. The result shows that Prompt Tuning not only increases the performance of good score model, but also raises the score of weak perform model, for ViT it increases the result by 0.2. Despite the promising result, our model still needs the whole model fine-tuning stage (First stage) on the target dataset. Hence, we will try to develop a more efficient tuning method without fine-tuning stage.

<Table 2> Experimenting with and without prompt tuning of base ViT and Baseline model.

Method	W/o prompt tune	Prompt tune
Base-ViT-21k ImageNet [6]	0.224	0.248
Baseline model	0.497	0.512

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1I1A3A04036408) and 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] Jia, Menglin, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. "Visual prompt tuning." In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII, pp. 709-727. Cham: Springer Nature Switzerland, 2022.

[2] Tallec, Gauthier, Edouard Yvinec, Arnaud Dapogny, and Kevin Bailly. "Multi-label transformer for action unit detection." arXiv preprint arXiv:2203.12531 (2022).

[3] Le Hoai, Duy, Eunchae Lim, Eunbin Choi, Sieun Kim, Sudarshan Pant, Guee-Sang Lee, Soo-Huyng Kim, and Hyung-Jeong Yang. "An attention-based method for multi-label facial action unit detection." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2454-2459. 2022.

[4] Wang, Lingfeng, Jin Qi, Jian Cheng, and Kenji Suzuki. "Action unit detection by exploiting spatial-temporal and label-wise attention with transformer." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2470-2475. 2022.

[5] Savchenko, Andrey V. "Video-based frame-level facial analysis of affective behavior on mobile devices using EfficientNets." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2359-2366. 2022.

[6] Ridnik, Tal, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. "Imagenet-21k pretraining for the masses." arXiv preprint arXiv:2104.10972 (2021).

[7] Arnab, Anurag, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. "Vivit: A video vision transformer." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 6836-6846. 2021.

[8] Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

[9] D.Kollias, et. al.: "ABAW: Valence-Arousal Estimation, Expression Recognition, Action Unit Detection & Emotional Reaction Intensity Estimation Challenges", 2023

[10]D.Kollias: "ABAW: Learning from Synthetic Data & Multi-Task Learning Challenges", ECCV, 2022

[11]D.Kollias: "ABAW: Valence-Arousal Estimation, Expression Recognition, Action Unit Detection & Multi-Task Learning Challenges", IEEE CVPR, 2022

[12]D.Kollias, et. al.: "Distribution Matching for Heterogeneous Multi-Task Learning: a Large-scale Face Study", 2021

[13]D.Kollias, et. al.: "Analysing Affective Behavior in the second ABAW2 Competition". ICCV, 2021

[14]D.Kollias,S. Zafeiriou: "Affect Analysis in-the-wild: Valence-Arousal, Expressions, Action Units and a Unified Framework, 2021

[15]D.Kollias, et. al.: "Analysing Affective Behavior in the First ABAW 2020 Competition". IEEE FG, 2020

[16]D.Kollias, S. Zafeiriou: "Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and ArcFace". BMVC, 2019

[17]D.Kollias, et at.: "Face Behavior a la carte: Expressions, Affect and Action Units in a Single Network", 2019

[18]D.Kollias, et. al.: "Deep Affect Prediction in-the-wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond". International Journal of Computer Vision (IJCV), 2019

[19]S.Zafeiriou, et. al. "Aff-Wild: Valence and Arousal in-the-wild Challenge". IEEE CVPR, 2017

[20]Ekman, Paul, and Wallace V. Friesen. "Facial action coding system." Environmental Psychology & Nonverbal Behavior (1978).

[21]Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.