# Facial Expression Recognition with ViT Considering All Tokens towards More Informative Self-attention Outputs

Haorui Song

School of Computing Science Wuhan University Wuhan China

rkant@whu.edu.cn

**Abstract.** Currently few works have been done to apply Vision Transformer (ViT) on facial expression recognition (FER) successfully. In this paper, we put forward a novel idea of processing the outputs from the multi-head attention in ViT by passing through a global average pooling layer, and accordingly design 2 network architectures, namely ViTTL and ViTEH. These 2 models, especially the latter one, show more strength in recognition of local patterns, which normal ideas hold that it is a weakness of ViT. Extensive experiments are conducted on the benchmark dataset FER-2013 with popular state-of-the-art models employed as baselines for comparison. The results demonstrate the superiority of the 2 new models in the realm of FER.

**Keywords:** component; Facial Expression Recognition; ViT.

## 1. Introduction

In order to be more natural in the interplay between human and computer systems, emotions from the human side are significant information that should not be neglected. Also, it is quite plausible that a huge number of industries would spring up providing that emotional data can be collected automatically. Generally, emotion collection can be done in two distinctive methods: bio-based and physical-based. Bio-based emotion recognition is an initiative method, which requires various sensors to collect the biological signals from a body, such as electromyography, electrodermal activity, skin temperature [1], etc. Despite the high recognition rate the bio-based method provides, complex, precision sensors should be worn is the insurmountable drawback of it. Costs would rise quickly if sensors were required in large quantities, and in many circumstances sensors cannot be worn, such as doing heavy sports, letting alone people would refuse to wear them merely because they do not want to.

Physical-based methods are passive ones, which include data of faces, gestures, or speeches. Using these data does not require any complicated sensor to be worn. Among all these physical data, facial emotion recognition (FER) is what has been regarded as the most prominent. According to the research in sociology, it is said that when we communicate face-to-face, 7% of the information is delivered in the linguistic part, like the spoken words, 38% by paralanguage, such as the vocal part or gestures, and most significant 55% is conveyed by the facial expressions [2]. Also, FER is known to have fewer constraints-giving the situation that recognition of emotions through video cameras without any personal consciousness, thus granting more opportunities to get access to greater datasets. Therefore, this study focuses on FER.

FER through images mainly includes classic machine learning methods and deep learning (for the most part CNN-based) learning architecture. Indeed, in sight of the low flexibility and high dependency on the expertise of traditional machine learning methods, CNN-based network has become the de-facto standard for image recognition, and FER is no exempt, with numerous works done before.

Vision Transformer (ViT) [3], is a deep learning model proposed recently, which is famous for taking the attention on the whole picture, rather than only focusing on small, say, $3 \times 3$ pixel squared kernels. This can be great news to FER, for FER usually needs to capture subtle muscle movements, which may hard to be detected in a small area. However, few works have been done on the study of applying it to FER. Existing studies have demonstrated that despite its successes in other applications, original version of ViT has a lethal drawback in FER, and all proposed a much different

network architecture based on transformers. This paper, however, would reveal that the original ViT is not as inept in FER as those studies show, and only a slight change of the structure would give a dramatic leap in accuracy.

In the original version of ViT, we would only make use of the self-attention output of the embedded token called [class]. This, largely, caused the incompetence of ViT. It could be the reason that [class] token is not large enough to summarize all information in the image, such as subtle muscle movements. Thus, we proposed 2 models based on ViT managing to enhance the insufficient [class] token by combining the discarded outputs of the other tokens. Experiments have shown that this action could significantly increase classification accuracy.

Our contributions can be described as follows:

We applied ViT-based models to address the FER problem and empirically evaluate its performance.

To solve the problem the insufficient information in the [class] token, we proposed 2 ViT-based models-ViT normal outputs ToLerance model (ViTTL), ViT normal outputs EnHanced model (ViTEH) by only changing a small part of the original ViT model.

Experimental evaluations on the dataset of FER2013 with currently SOTA models as baselines to validate the effectiveness of the proposed frameworks.

## 2. Related Works

### 2.1.FER Based on CNN

CNN has always been favoured for image recognition since its success on ImageNet for it provides an “end-to-end” learning strategy compared to traditional machine learning methods which rely heavily on physically based models or pre-processing techniques. Tang [4] proposed that using L2-SVM which is differentiable as a layer instead of Softmax on a CNN-model can be of great power and lead to the state-of-the-art (SOTA) on the dataset FER2013. Zhao et al. [5] proposed a network equipped with grid selection feature, which is able to automatically extract and filter facial features by using a feature selection algorithm within AlexNet [6].

Till now, CNN-based methods have been studied thoroughly in the field of FER, and tend to reach their highest peak in performance.

### 2.2.FER Based on ViT

ViT was introduced in 2021. Ever since its proposal, it has been enjoying increasing popularity not only for its transcendent performance in image classification, but also for its jumping out of the box from the ever-studied CNN-based architecture for feature extraction. Generally speaking, ViT is known for its strength in classification on a global scale, i.e., tasks requiring seeing the whole picture, for it forwards by checking the overall association in different parts of the image parted in a specific fixed rule. Even if numerous computer vision cases have it that ViT is being applied to them such as target detection, target tracking, FER stands still and applications are rare to find.

To study its performance on FER, Xue et al. [7] gave the process ViT-FER which simply parts different faces through different characteristics and then fed to a ViT. Also, Kim et al. [8] squeezed the pure version of ViT to speed up the computation.

It is generally understood that transformers take their strength in a global scope, for each token computes its self-attention with all other tokens, and take its weakness in detecting local patterns, for it may not reflect local features [8]. So it is believed that applications involving small and local patterns are not suitable for ViTs. Experiments have it that pure version of ViT in FER performs worse than CNN-based models. However, this flaw can be conquered by adding the losing information in the deprecated outputs together with the global one. What we need is only to take the average of the outputs and even do not need to change the parameter numbers.

# 3. Method

Classifying emotions has no differences to common image recognition problem. The focus of neural network for this problem is also recognizing patterns or features on images, or specifically, image of human faces. This means researchers could apply majority of common ViT structures to this problem. However, the performance of different structures for the dataset could differ greatly due to features of dataset and environment. In order to construct the structures with the best performance for the dataset, we verify and optimize the ViT-based model on the datasets for multiple times and choose the best one based on its validation accuracy and test accuracy.

## 3.1.Multi-head Attention

Multi-head attention is the core component of ViT, which is an extension version of the normal self-attention. The normal self-attention describes as follows: for any input $\boldsymbol{z} \in \mathbb{R}^{N\times D}$, we start with 3 parts of trainable matrixes $W_q, W_k, W_v \in \mathbb{R}^{D\times h}$, and get the attention $\mathrm{SA}(z)$ as follows:

$$\mathrm{q} = \mathrm{z}W_q \tag{1}$$

$$\boldsymbol{k} = \boldsymbol{z}W_k \tag{2}$$

$$\boldsymbol{v} = \boldsymbol{z}W_v \tag{3}$$

$$\mathbf{A} = \mathbf{softmax}\left(\frac{\boldsymbol{q}\boldsymbol{k}^{\mathrm{T}}}{\sqrt{\boldsymbol{D}}}\right) \tag{4}$$

$$\mathbf{SA}(\boldsymbol{z}) = \boldsymbol{A}\boldsymbol{v} \tag{5}$$

For the multi-head attention (MSA) we have $k$ sets of the common self-attentions (marked as heads) and a trainable transformation matrix $W_0 \in \mathbb{R}^{k\cdot h\times D}$, then we get the $\mathrm{MSA}(z)$ by concatenating multiple $\mathrm{SA}(z)$s multiply the transformation matrix:

$$\mathrm{MSA(z)} = \mathrm{concat}\big(\mathrm{SA_1(z)}, \cdots, \mathrm{SA_k(z)}\big)\mathrm{W_0} \tag{6}$$

Where concat is to concatenate the matrixes in order after their columns.

## 3.2.Encoder Block

Unlike the transformer in NLP, which has both an encoder and a decoder, ViT only make use of an encoder, which a block is structured as bellow:

1) Layer normalization [9]: To overcome the drawbacks batch normalization has, layer normalization do the normalization through the hidden units, where $H$ is the number of hidden units in a layer.

$$\mu^l = \frac{1}{H}\sum_{i=1}^{H} a_i^l \tag{7}$$

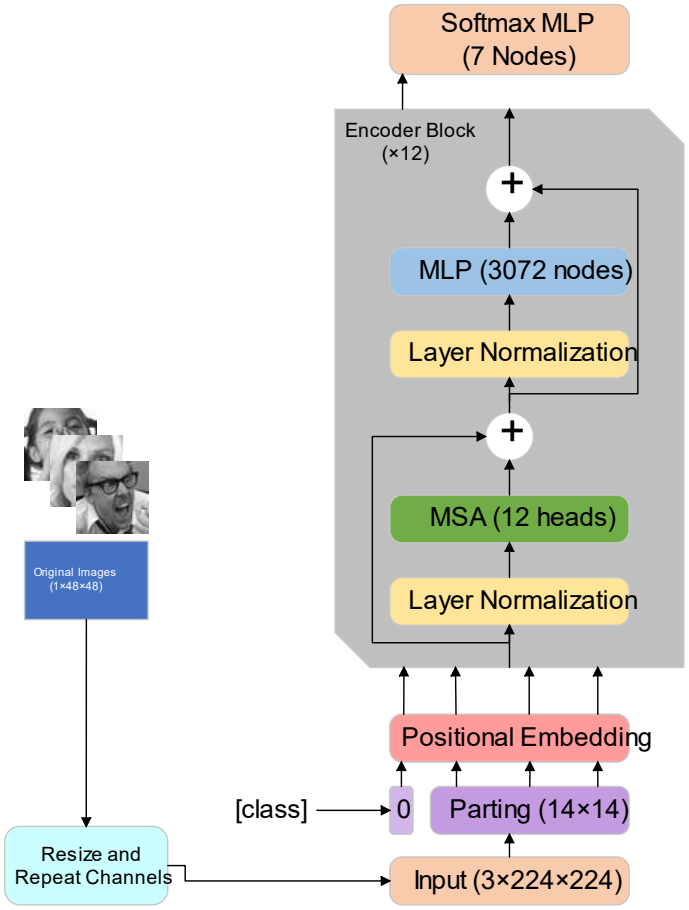


**Figure 1.** The origin version of ViT

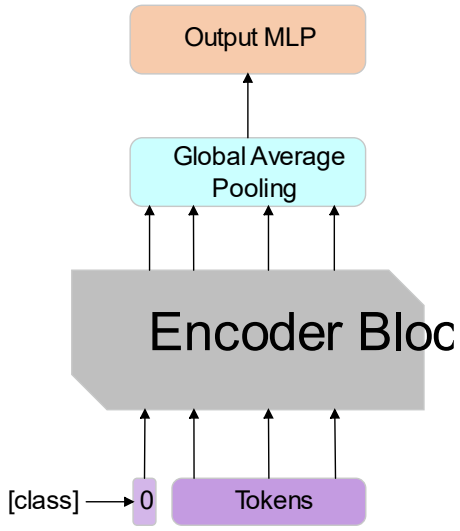


**Figure 2.** ViTTL: Making use of every output and treat [class] as a normal output to go through the Global average pooling layer

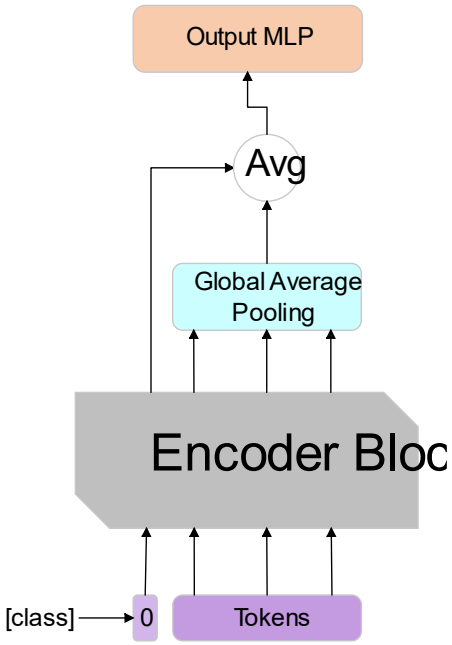


**Figure 3.** ViTEH: Average the output of [class] token and other pooled attention outputs

$$\sigma^l = \sqrt{\frac{1}{H}\sum_{i=1}^{H}\left(a_i^l - \mu^l\right)^2} \quad (8)$$

2) Multi-head attention: As mentioned before, we have 12 heads in each layer.
3) Residual adding: We get the output from layer 1) and add it with the last layer output (layer 2)).
4) Layer normalizaion: Another layer norm after residual adding.
5) MLP: A fully connected layer with 3072 nodes.
6) Residual adding: Add the output from layer 3) and 5).
This is an encoder block, which can be piled up to form bigger network.

### 3.3.Overall archetecture

After describing previous network blocks, we can now build the overall architecture of ViT. We proposed and trained 3 ViT models, one the same as the original version [3], while the other two uses a pooling layer for further process. Apart from the layers above, the model has some additional ones:

*Input* Layer*:* Passing the photo into the model.

*Parting:* The image is parted with $16 \times 16$ pixels to form $\left(\frac{224}{16}\right)^2 = 196$ patterns. Each patten is flattened to be one of the tokens. Then we prepend an extra learnable token [class] in the position 0 to serve as an image representative.

*Positional Embedding:* Tokens are positional embedded with a 1D learnable positional encoding parameters.

*Output Layer:* A 7-node MLP with Softmax function to output the classification.

The structure of the first model is depicted exactly in Figure 1, where we *only* use the attention output of the [class] token for classification, i.e., *discarding* the attention output of every other token *except* the [class] token and connecting it to the classification MLP.

In order to make use of the self-attention outputs of ordinary tokens, in the second model ViTTL, we treat all the attention outputs as equal, keeping and connecting them to a global average pooling layer [10] for summarization, and then connect to the MLP, as Figure 3 shows.

**Figure 5.** Images considered to be "abnormal"

ViTTL does not put the speciality of [class] token into consideration, in the third model ViTEH, as depicted in Figure 2, we emphasize the speciality of the [class] token by letting that token out when going through the global average pooling layer. The "Avg" here simply means adding the two tensors together then divided by 2.

## 4. Experiment

This section evaluates the proposed models on the FER-2013 dataset. We first provide an overview of the specific dataset. Then, we display the comparison over the state-of-the-art (SOTA) networks (including the original version of ViT) and our models. At last, we discuss the possible reasons why original networks perform worse than our variant.

### 4.1.Dataset

The 2013 Facial Expression Recognition dataset (FER-2013) is a dataset provided by Kaggle, introduced at the International Conference on Machine Learning (ICML) in 2013 [11] by Pierre-Luc Carrier and Aaron Courvill.

This dataset gets its source from the internet, not being an academic lab-base dataset. It classifies each face accordingly, by the emotion they show, to their categories they belong to. All the images in it chop their channels and turned into grayscale images, then stretch to $48 \times 48$ pixel squared. The total data number is 35,887, which is made up of 7 different types of expression, which is described in TABLE I.

**Table 1.** Number of Data in FER-2013 Dataset

| **Expression (Classification)** | **Validation Data** | | **Training Data** | **Total** |
|---|---|---|---|---|
| | ***Public*** | ***Private*** | | |
| Angry | 467 | 491 | 3995 | 4953 |
| Disgust | 56 | 55 | 436 | 547 |
| Fear | 496 | 528 | 4097 | 5121 |
| Happy | 895 | 879 | 7215 | 8989 |
| Sadness | 653 | 594 | 4830 | 6077 |
| Surprise | 415 | 416 | 3171 | 4002 |
| Contempt | 607 | 626 | 4965 | 6198 |
| | 3589 | 3589 | 28709 | 35887 |

### 4.2.Experiment Setup

As we are fine-tuning a model, the size of the training data does not matter a lot. Nevertheless, data augmentation still can be helpful. Emotion classification only contains images of faces, and data augmentation can be done in a wider range. Specifically, training data are randomly rotated to a maximum of 10 degrees, followed by a random horizontal flipping and a width and height shift of 0.1 for data augmentation. Figure 2 shows the example of the original data and the processed data. Also, for equality reasons, all the input images are stretched into $224 \times 224$ pixel squared and expend their channel from grayscale to "RGB" (repeat them 3 times) to $3 \times 224 \times 224$ shaped tensor, which is shown in Figure 4.

**Figure 4.** Origin image (left) and its augmentation varient (right)

Our models are implemented with the popular TensorFlow framework. For a fair comparison, all the models are using the optimizer AdamW [12] with the default hyperparameters in TensorFlow (with learning rate equals to $3 \times 10^{-5}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay rate to $0.05$). The ViT base model is implemented with Hugging Face 🤗's TFViTModel with the original weights of "google/vit-base-patch16-224-in21k" However, when compared to the classical networks, these parameters cannot fit in and might cause somewhere falling into the trap of infinity. In this case, notes will be available.

The whole training process took place on Google Colaboratory of a free licence, whose GPU is NVIDIA Tesla T4 at hand when we do the training. We divide the dataset into an 85% training set and a 15% validation set. All the models got the training for a maximum of 10 epochs with a mini-batch size of 32, with an Earlystopping callback monitoring the validation loss, bearing the tolerance of 1 step and restoring to their best weights.

### 4.3.Experiment results

We compared our ViT-based models with the CNN base models like ResNet and VGG in their original forms. Results are summarized in TABLE II. , where the "#Param" refers to the number of parameters of the model.

**Table 2.** Comparison to SOTA Methods on the FER-2013 Test Set

| Arch. | Model | #Param | Top-1 Acc. |
|---|---|---|---|
| ResNet | ResNet18[a] | 11.2M | 56.67 |
| | ResNet34[a] | 21.3M | 44.09 |
| | ResNet50 | 23.6M | 64.99 |
| | ResNet101 | 42.7M | 64.32 |
| | ResNet152 | 58.4M | 67.15 |
| VGG | VGG16 | 14.7M | 66.91 |
| | VGG19 | 20.0M | 67.21 |
| Trans | ViT | 86.4M | 67.90 |
| | ViTTL | 86.4M | 69.14 |
| | ViTEH | 86.4M | **70.37** |

a. These models do not have pre-trained weights to fine-tune.

### 4.4.Result Explanation

As can be inferred from the experiment results, ViT based model really does a better job in FER than CNN-based architectures. The reasons are supposed to be as the following:

Error tolerance: ViT uses transformers which are initially used in the language model; thus they are born to the high noise in language training materials. Also, the dataset FER-2013 consists of a number of "abnormal" images, which are shown in Figure 5. As a result, it is quite plausible that transformer-based models outweigh the CNN-based ones in the field of FER.

Channel changing: All the CNN-based models with their pre-trained weights are originally trained on colourful images, where their kernels have learnt to detect patterns in different channels. FER-2013, however, even repeats the channels, are still grayscale images, so as to CNN kernels failing to fit into this situation. ViT, on the other hand, does not require kernels to make predictions, then they might suit new conditions more easily.

The result of Model 2 and Model 3 may extirpate the conclusion that in the field of FER, we have to consider subtle muscle movements, so that a simple [class] token does not include all these subtle information, and we must consider all other self-attention outputs to see a bigger picture. Nonetheless, since the [class] token is trainable, it still does consist of a large proportion of information, and only treating it as a normal token is not the best idea.

## 5. Conclusion

In this study, we propose 2 new types of ViT-based networks by slightly changing the original. To solve the problem that ViT performs not as good as CNN-based models, we propose a structure making use of the self-attention outputs from the other tokens other than [class] token by calculating the average. Based on the experimental results on the dataset FER-2013, we draw the conclusion that performance could be notably improved when using the self-attention outputs of [class] token and others at the same time, and the latter acts as a complement to the former on local feature detections.

## References

[1] A. Haag, S. Goronzy, P. Schaich, and J. Williams, 'Emotion Recognition Using Bio-sensors: First Steps towards an Automatic System', in Affective Dialogue Systems, vol. 3068, E. André, L. Dybkjær, W. Minker, and P. Heisterkamp, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 36–48.

[2] COMMUNICATION THEORY. LONDON: ROUTLEDGE, 2017. Accessed: Jul. 11, 2022.

[3] A. Dosovitskiy et al., 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale'. arXiv, Jun. 03, 2021.

[4] Y. Tang, 'Deep Learning using Linear Support Vector Machines'. arXiv, Feb. 21, 2015. Accessed: Jun. 26, 2022.

[5] S. Zhao, 'Feature Selection Mechanism in CNNs for Facial Expression Recognition', p. 12.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, 'ImageNet classification with deep convolutional neural networks', Commun. ACM, vol. 60, no. 6, pp. 84–90, May 2017.

[7] F. Xue, Q. Wang, and G. Guo, 'TransFER: Learning Relation-aware Facial Expression Representations with Transformers', in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, Oct. 2021, pp. 3581–3590.

[8] S. Kim, J. Nam, and B. C. Ko, 'Facial Expression Recognition Based on Squeeze Vision Transformer', Sensors, vol. 22, no. 10, p. 3729, May 2022, doi: 10.3390/s22103729.

[9] J. L. Ba, J. R. Kiros, and G. E. Hinton, 'Layer Normalization'. arXiv, Jul. 21, 2016.

[10] M. Lin, Q. Chen, and S. Yan, 'Network In Network'. arXiv, Mar. 04, 2014.

[11] I. J. Goodfellow et al., 'Challenges in Representation Learning: A Report on Three Machine Learning Contests', in Neural Information Processing, vol. 8228, M. Lee, A. Hirose, Z.-G. Hou, and R. M. Kil, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 117–124.

[12] I. Loshchilov and F. Hutter, 'Decoupled Weight Decay Regularization'. arXiv, Jan. 04, 2019.