

Facial Expression Recognition using Vanilla ViT backbones with MAE Pretraining

Jia Li

School of Computer Science and Information Engineering, Hefei University of Technology
jiali@hfut.edu.cn

Abstract

Humans usually convey emotions voluntarily or involuntarily by facial expressions. Automatically recognizing the basic expression (such as happiness, sadness, and neutral) from a facial image, i.e., facial expression recognition (FER), is extremely challenging and attracts much research interests. Large scale datasets and powerful inference models have been proposed to address the problem. Though considerable progress has been made, most of the state-of-the-art employing convolutional neural networks (CNNs) or elaborately modified Vision Transformers (ViTs) depend heavily on upstream supervised pretraining. Transformers are taking place the domination of CNNs in more and more computer vision tasks. But they usually need much more data to train, since they use less inductive biases compared with CNNs. To explore whether a vanilla ViT without extra training samples from upstream tasks is able to achieve competitive accuracy, we use a plain ViT with MAE pretraining to perform the FER task. Specifically, we first pretrain the original ViT as a Masked Autoencoder (MAE) on a large facial expression dataset without expression labels. Then, we fine-tune the ViT on popular facial expression datasets with expression labels. The presented method is quite competitive with 91.07% on RAF-DB, 62.42% on AffectNet, and 90.18% on FER-Plus, which can serve as a simple yet strong ViT-based baseline for FER studies. The code and pretrained models will be publicly available.

Introduction

Facial expression analysis plays an important role in understanding one's feelings, attitudes and intensions (Gong, An, and Elfiky 2022; Zhao, Liu, and Zhou 2021), which facilitate many other techniques such as new-type human-computer interaction and mental health assessment. The task of facial expression recognition (FER) aims to infer the basic emotion class, such as happiness, sadness, and anger, of the person in a given facial image. Researches have made substantial progress in the past few years, benefiting from the development of more powerful deep neural networks, such as ResNet (He et al. 2016) and Vision Transformer (Dosovitskiy et al. 2020), and more representative benchmarks, such as AffencNet (Mollahosseini, Hasani, and Mahoor 2017) and RAF-DB (Li, Deng, and Du 2017).

Predicting the discrete emotion class of a facial image is difficult due to the expression ambiguity, imbalanced and noisy emotion labels. Therefore, most of the state-of-the-art (SOTA) methods (Xue, Wang, and Guo 2021; Li et al. 2021; She et al. 2021; Hwang et al. 2022; Wang et al. 2020) commonly pretrain deep models on large scale and less noisy datasets, for instance, Ms-Celeb-1M (Guo et al. 2016) and even ImageNet, to gain better feature representations first and fine-tune deep models on specific FER datasets. Most recently, CNNs equipped with attention modules (Ma, Sun, and Li 2021; Xue, Wang, and Guo 2021) and ViT variants (Li et al. 2021) have been introduced into FER and obtained promising performance. However, whether a vanilla ViT without extra training samples from upstream tasks can achieve competitive performance has not been explored thus far. In this paper, we demonstrate that a vanilla ViT-base backbone pretrained by MAE method (He et al. 2022) without extra data is comparable to the SOTA work.

Our Approach

We initialize the original ViT as a Masked Autoencoder (MAE) (He et al. 2022) on AffectNet (Mollahosseini, Hasani, and Mahoor 2017) dataset such that it can extract a stronger representation of expressive semantic information. MAE pretraining force the ViT to learn the facial local and global structures (Chen et al. 2022). Different facial emotions often share one or more facial muscle movements. Thus, facial expressions with low occurrence frequencies, for example, fear, could benefit from the structure learning of frequent facial expressions, for example, happiness, during the pretraining.

MAE. The masked autoencoder (He et al. 2022) employs an asymmetric encoder-decoder architecture to implement the self-supervised task by randomly masking the patches of input images and reconstructing missing pixels. It is experimentally demonstrated that MAE can learn a large number of visual semantics and has excellent performance in several downstream tasks. Inspired by this, we adopt MAE to pretrain a vanilla ViT with powerful expression semantic representation and narrow the gap between self-supervised task and downstream FER task.

Dataset. The upstream pretraining task requires large-

scale data, which enables the model to have strong representational capabilities. For FER, AffectNet is the largest available in-the-wild dataset that contains more than one million face images collected from the Internet by keyword searching. Among them, about 400K images are manually labeled. We clean the dataset based on the label file and use OpenFace (Baltrušaitis et al. 2018) for face alignment, leaving 269,316 images for self-supervised training.

Experiment

Implementation

We choose the ViT-base and do self-supervised pretraining on the AffectNet training set. The pretraining length is 300 epochs and the other settings of pretraining and fine-tuning are same as MAE (He et al. 2022). Due to the training set of AffectNet is imbalanced whereas the validation set is balanced, oversampling strategy is used while fine-tuning. We train the ViT model on each FER training subset and then evaluate on each FER testing subset.

Result

We compare our method with two pretraining methods using supervised learning and MAE on ImageNet-1K in Table 1. Although AffectNet is almost $4\times$ smaller than ImageNet1K, pretraining with MAE on AffectNet yields better fine-tuning results than both supervised learning and MAE pretraining methods on ImageNet-1K. Our method based on **ViT-base** achieves the best accuracy of 90.22%, 61.73% and 89.00% on RAF-DB, AffectNet and FER-Plus respectively without any data augmentation, which gives us a powerful pretraining model for the subsequent work. And the performance reaches **91.07%** on RAF-DB, **62.42%** on AfectNet, and **90.18%** on FER-Plus if regular augmentation is used (random horizontal flipping, color jitter and affine transformation).

Table 1: MAE pretraining vs. supervised pretraining, numbers evaluated by fine-tuning the ViT-base weights from ImageNet-1K and AffectNet on downstream FER datasets without data augmentation.

method	pretrain data	RAF-DB	AffectNet
Supervised	IN1K	85.14%	57.10%
MAE	IN1K	88.49%	60.87%
MAE	AffectNet	90.22%	61.73%

Table 2: More ViT-based results on RAF-DB, AffectNet and FER-Plus. Here, data augmentation is employed. The variants, ViT-small and ViT-tiny are proposed in DeiT (Touvron et al. 2021).

method	pretrain data	ViT variants	RAF-DB	AffectNet	FER-Plus
Supervised	IN1K	ViT-samll	87.19%	57.91%	88.46%
Supervised	IN1K	ViT-tiny	87.03%	58.28%	88.56%
Supervised	IN1K	ViT-base	87.22%	57.99%	88.91%
MAE	AffectNet	Vit-base	91.07%	62.42%	90.18%

In addition, to demonstrate the effectiveness of our method, we compare it with the existing state-of-the-art methods. In Table2, SCN (Wang et al. 2020) and RUL (Zhang, Wang, and Deng 2021) adopt ResNet as the backbone and introduce different uncertainty learning methods to tackle the uncertainty problem in FER. MVT (Li et al. 2021) proposes the combination of grid-wise local attention and token-based global attention to overcome the weakness of CNN-based in long-range relation modeling. Compared with these well-designed SOTA methods, our results on the three FER datasets demonstrate the effectiveness and scalability of our method.

Table 3: Comparsion with the state of the art on RAF-DB, AffectNet and FER-Plus.

method	year	RAF-DB	AffectNet	FER-Plus
SCN (Wang et al. 2020)	2020	88.14%	60.23%	89.35%
RUL (Zhang, Wang, and Deng 2021)	2021	88.98%	61.43%	
MVT (Li et al. 2021)	2021	88.26%	61.40%	89.22%
Ours	2022	91.07%	62.42%	90.18%

Conclusions

In this paper, we demonstrate that a vanilla ViT-based backbone pretrained by MAE method (He et al. 2022) without extra data is comparable to the SOTA work. The presented method achieves 91.07% on RAF-DB, 62.42% on AfectNet, and 90.18% on FER-Plus, which can serve as a transparent ViT-based baseline for FER studies.

References

- Baltrušaitis, T.; Zadeh, A.; Lim, Y. C.; and Morency, L.-P. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 59–66. IEEE.
- Chen, X.; Ding, M.; Wang, X.; Xin, Y.; Mo, S.; Wang, Y.; Han, S.; Luo, P.; Zeng, G.; and Wang, J. 2022. Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Gong, W.; An, Z.; and Elfiky, N. M. 2022. Deep learning-based microexpression recognition: a survey. *Neural Computing and Applications* 1–24.
- Guo, Y.; Zhang, L.; Hu, Y.; He, X.; and Gao, J. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, 87–102. Springer.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009.
- Hwang, H.; Kim, S.; Park, W.-J.; Seo, J.; Ko, K.; and Yeo, H. 2022. Vision transformer equipped with neural resizer on facial expression recognition task. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing*.

Conference on Acoustics, Speech and Signal Processing (ICASSP),
2614–2618. IEEE.

Li, H.; Sui, M.; Zhao, F.; Zha, Z.; and Wu, F. 2021. Mvt: mask vision transformer for facial expression recognition in the wild. *arXiv preprint arXiv:2106.04520*.

Li, S.; Deng, W.; and Du, J. 2017. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2852–2861.

Ma, F.; Sun, B.; and Li, S. 2021. Facial expression recognition with visual transformers and attentional selective fusion. *IEEE Transactions on Affective Computing*.

Mollahosseini, A.; Hasani, B.; and Mahoor, M. H. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* 10(1):18–31.

She, J.; Hu, Y.; Shi, H.; Wang, J.; Shen, Q.; and Mei, T. 2021. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6248–6257.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357. PMLR.

Wang, K.; Peng, X.; Yang, J.; Lu, S.; and Qiao, Y. 2020. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6897–6906.

Xue, F.; Wang, Q.; and Guo, G. 2021. Transfer: Learning relation-aware facial expression representations with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3601–3610.

Zhang, Y.; Wang, C.; and Deng, W. 2021. Relative uncertainty learning for facial expression recognition. *Advances in Neural Information Processing Systems* 34.

Zhao, Z.; Liu, Q.; and Zhou, F. 2021. Robust lightweight facial expression recognition network with label distribution training. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 3510–3519.