BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2024
http://excel.fit.vutbr.cz

Vision Transformers for Facial Recognition

Šimon Strýček

Abstract

This paper focuses on applying vision transformer based neural networks to face recognition related tasks. It focuses on exploring modern ViT architectures, experimenting with alternative data, and finding the right parameters to train vision transformers to compete with the already established dominance of convolutional neural networks in face recognition.

The goal was not to create a state-of-the-art solution, but it was to show the potential in this kind of neural networks for this specific field. The output of this work contains results of various experiments, demonstrations of benefits and drawbacks of some of the modern and popular ViTs, definition of an optimal setup when wanting to employ vision transformers for facial recognition, and interesting observations from working with vision transformers.

*xstryc06@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

For the past several years, convolutional neural networks have been the number one when considering models for image processing and pattern recognition. This fact changed shortly after transformers were introduced to solve vision-related problems in the paper named *An Image is Worth 16x16 Words* [1]. Since then, various forms of ViTs have been dominant in tasks like image classification or segmentation. With the rise of such architectures, a question of suitability to other vision tasks arises, which is the main motivation of this work.

The problem this paper tries to solve is the dilemma of whether it is worth switching to this new type of architecture or whether it is better to stick to conventional solutions with CNNs. This question is tried to be answered by performing various experiments and finding the best conditions for training such a neural network. The goal of each experiment performed in this work was to maximize prediction accuracy. The best results achieved are all compared with the current state-of-the-art methods.

Existing solutions for face recognition already use some aspects of vision transformers (at least few self-attention layers) but are in most cases still primarily focused around convolutional layers. The vast majority of today's state-of-the-art models, like the one described in the last years *AdaFace* [2], achieve very good performance by exploiting a large number of parameters using deep convolutional backbone. (In case of *ArcFace* paper it is ResNet-100.) On the other hand, from the performance perspective, they are able to achieve astonishing results above 99% binary classification Accuracy on LFW dataset.

The approaches described in this paper include using different variations of transformers as backbones in different scenarios. These backbones were not modified, but were presented with different types of data, tasks, and parameters. The results show that even though experimenting with these models had some potential, primarily due to unavailability of clean data, it is best to stick to more traditional setups and rather exploiting pretrained variants.

Although competing with high levels of accuracy described in SoTA methods is difficult, it is achievable even with pure ViTs. This paper demonstrates that by using simple vision transformers, we can get respectably close to this state without making significant changes to the architecture. The best performance achieved so far in this paper was peaking at 98.98% Accuracy on LFW with only a "tiny" version of such a transformer.

2. Performed experiments

This paper summarizes the most important takeaways from a series of experiments performed with multiple ViT architectures. These experiments can be categorized into the following themes:

- **Experimenting with modern architectures**, which consisted of selecting a set of such models and comparing their ability to perform on face recognition task. The testing included training all selected models under similar conditions and evaluating their performance during and after training. Good-performing architectures were then used for further experimentation.
- **Combining different types of training data** focused primarily on realizing multitask learning and exploiting CLIP architecture implementation. Experiments of this type led to better results but were very data-dependent.
- **Testing multiple public datasets** played a crucial role in the resulting performance of the trained models. Although some datasets contained better quality images, they lacked alternative annotations that could potentially help to get the best from tested architectures.
- **Finding ideal training conditions** was one of the most challenging sets of experiments, including tuning the training hyperparameters and choosing the most suitable loss function. Training vision transformers for such a task proved to be quite challenging.

3. Exploring modern ViTs

For the first experiment, multiple architecture types based on vision transformers were selected. The source for those architectures were past two ICCV and CVPR conferences (2022 and 2023). All selected architectures were evaluated by training them on a face recognition task, and all of them were tested afterwards. On the basis of the evaluation results, the best candidates for further experimentation were picked. Among the ViTs used were:

- Swin transformer [3]
- FLatten transformer [4]
- CMT (Convolutional Vision Transformer) [5]
- BiFormer [6]
- Scale aware modulation transformer [7]
- CLIP [8]
- Transformer with deformable attention [9]

Among the best performing models were the Swin transformer, CLIP and FLatten transformer.

4. CLIP model

The best performing model among all the tested ones was the pretrained CLIP model. Using it's pretrained variant made by Laion, I was able to achieve results close to current state-of-the-art facial recognition systems.

Due to this success, several other experiments were performed with this model. One of them being the exploitation of its ability to combine textual and visual data. With a combination of face datasets containing alternative annotations, there is potential to guide learning using text prompts. Using this approach together with multitask learning focused on predicting the alternative annotations, there was a noticable performance increase. Unfortunately, since the quality of the dataset containing alternative annotations was not ideal, the best results achieved with CLIP was with a different approach - using a good quality large dataset with just the visual transformer part of the architecture.

5. Optimal training conditions

Using both *ArcFace* and *CosFace* losses in combination with ViTs showed to be very sensitive to selecting the right parameters and therefore further tuning was crucial. Nonoptimal hyperparameter selection usually led to poor convergence and plateau at low levels of accuracy.

Using empirical testing, it was found that through *ArcFace* loss being currently a state-of-the-art method for training face recognition tasks, it resulted in much worse performance with ViTs than training using the *CosFace* loss function. This can be either because ViTs have different distribution of the extracted features coming from their architectural difference or due to *CosFace* function being less sensitive to selecting the margin and other hyperparameters. In either way, using the *CosFace* function proved to be able to push the model to respectable performance.

6. Achieved results

As mentioned above, the CLIP model was the most successful ViT architecture used in the scope of this paper with the best accuracy peaking at 98.98% on the LFW benchmark. In this experiment, only the visual transformer was extracted from the smallest variant of the CLIP model. Current TOP 10 implementations are reaching above 99% with the SoTA method having 99.87% on this specific benchmark. Other experiments trained using the loss function *CosFace* reached between 95 and 98% Accuracy.

References

[1] An image is worth 16x16 words: Transformers for image recognition at scale. ICLR 2021.

[2] Adaface: Quality adaptive margin for face recognition, 2022.

[3] OpenAI. Swin transformer: Hierarchical vision transformer using shifted windows, 2021.

[4] Flatten transformer: Vision transformer using focused linear attention. ICCV 2023.

[5] Cmt: Convolutional neural networks meet vision transformers, 2021.

[6] Biformer: Vision transformer with bi-level routing attention. CVPR 2023.

[7] Scale-aware modulation meet transformer. ICCV 2023.

[8] Learning transferable visual models from natural language supervision, 2021.

[9] Vision transformer with deformable attention. CVPR 2022.