A face recognition framework based on vision transformers

December 19 2022, by Ingrid Fadelli

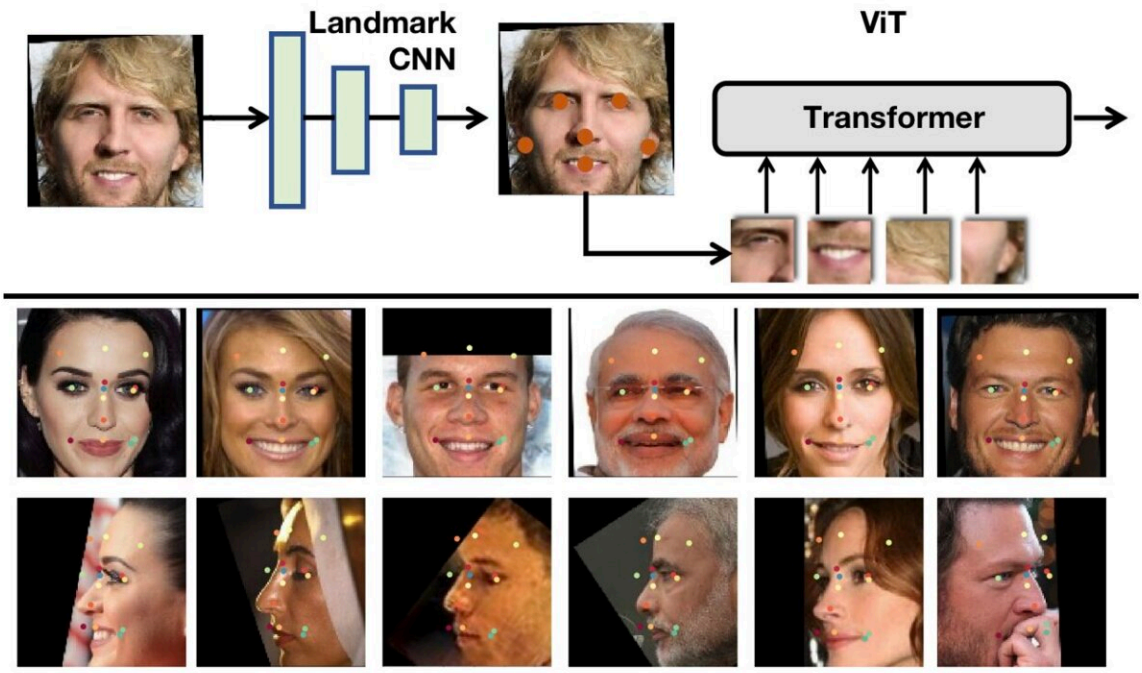

Illustration of the team's part-based ViT for face recognition. A facial image is processed by a lightweight landmark CNN which produces a set of facial landmarks. The landmarks are used to sample facial parts from the input image which are then used as input to a ViT for feature extraction and recognition. The whole system is trained end-to-end without landmark supervision. Examples of landmarks detected by the landmark CNN are shown. Credit: Sun & Tzimiropoulos.

Face recognition tools are computational models that can identify specific people in images, as well as CCTV or video footage. These tools are already being used in a wide range of real-world settings, for instance aiding law enforcement and border control agents in their criminal investigations and surveillance efforts, and for authentication and biometric applications. While most existing models perform remarkably well, there may still be much room for improvement.

Researchers at Queen Mary University of London have recently created a new and promising architecture for face recognition. This architecture, presented in a paper pre-published on *arXiv*, is based on a strategy to extract facial features from images that differs from most of those proposed so far.

"Holistic methods using convolutional neural networks (CNNs) and margin-based losses have dominated research on face recognition," Zhonglin Sun and Georgios Tzimiropoulos, the two researchers who carried out the study, told TechXplore.

"In this work, we depart from this setting in two ways: (a) we employ the Vision Transformer as an architecture for training a very strong baseline for face recognition, simply called fViT, which already surpasses most state-of-the-art face recognition methods. (b) Secondly, we capitalize on the Transformer's inherent property to process information (visual tokens) extracted from irregular grids to devise a pipeline for face recognition which is reminiscent of part-based face recognition methods."

Most widespread face recognition approaches are based on CNNs, a class of artificial neural networks (CNNs) that can autonomously learn to find patterns in images, for instance identifying specific objects or people. While some of these methods achieved very good performances, recent work highlighted the potential of another class of algorithms for

face recognition, known as vision transformers (ViTs).

In contrast with CNNs, which typically analyze images in their entirely, ViTs split an image into patches of a specific size, and then adds embeddings to these patches. The resulting sequence of vectors is then fed to a standard transformer, a deep learning model that differentially weighs different parts of the data it is analyzing.

"The ViT, contrary to CNNs, can actually operate on patches extracted from irregular grids and does not require the uniformly spaced sampling grid used for convolutions," the researchers explained in their paper. "As the human face is a structured object composed of parts (e.g., eyes, nose, lips), and inspired by seminal work on part-based face recognition before deep learning, we propose to apply ViT on patches representing facial parts."

The vision transformer architecture created by Sun and Tzimiropoulos, dubbed part fViT, is made up of a lightweight network and a vision transformer. The network predicts the coordinates of facial landmarks (e.g., nose, mouth, etc.), while the transformer analyzes patches containing the predicted landmarks.

The researchers trained different face transformers using two well-known datasets, namely the MS1MV3, which contains images of 93,431 people and the VGGFace2, featuring 3.1 million images and 8,600 identities. Subsequently, they conducted a series of tests to evaluate their models, also changing some of their features to test how this affected their performance.

Their architecture achieved remarkable accuracies for all the datasets it was tested on, comparable to those of many other state-of-the-art face recognition models. In addition, their models appeared to successfully delineate facial landmarks without being specifically trained for it.

In the future, this recent study could inspire the development of other models for face recognition based on vision transformers. In addition, the researchers' architecture could be implemented in applications or software tools that could benefit from the selective analysis of different face landmarks.

More information: Zhonglin Sun and Georgios Tzimiropoulos, Part-based Face Recognition with Vision Transformers, *arXiv* (2022). DOI: 10.48550/arxiv.2212.00057