Analyzing the Current Status of the Transformer Model for a Face Recognition Application

Haiqing Jiang

School of Computer Engineering, Guangzhou City University of Technology, Guangzhou, China

Abstract. With the rapid development of deep learning technology and the extensive use of face recognition technology, the Transformer is gradually emerging in face recognition technology related applications. This paper systematically analyzes the evolution path of face recognition technology based on the Transformer architecture, focusing on the four technical perspectives of complex scene adaptation, occlusion robustness, lightweight design and low-scoring and high-noise image repair. In this paper, we conclude that Transformer brings a paradigm breakthrough in face recognition technology and demonstrates strong global sensing ability in complex scenes by virtue of its long-range dependent modeling advantage. With breakthrough potential at the level of cross-modal feature modeling and noise robustness, and an in-depth discussion on the innovative breakthroughs in cross-modal feature fusion and dynamic scene adaptation of the lightweight Transformer framework based on the self-attention mechanism,It also reveals the core bottlenecks such as strong data dependency and insufficient local texture sensitivity that the technique faces in practical applications. Ultimately, this paper validates the potential of Transformers in improving the robustness of face recognition through comprehensive evaluation and provides a theoretical framework and technology roadmap for scenario-based deployment of next-generation face recognition systems.

1 Introduction

Face recognition technology, with its non-contact, convenient and high-precision features, has been widely used in security monitoring, intelligent interaction and other scenarios and the technology continues to iterate rapidly. By extracting and analyzing facial features for identification and verification, the technology provides accurate and convenient biometric solutions for the digital age. However, it shows loss of facial texture information, failure of micro-feature extraction, and insufficient generalization ability in strong exposure, facial occlusion, low-resolution processing, and multi-dimensional feature changes.

Traditional face recognition techniques rely on manual feature extraction but have weak generalization ability. Convolutional Neural Network (CNN)-based face recognition methods automatically learn hierarchical features through convolutional layers and are robust to partial occlusion and expression changes. However, CNNs do not capture enough global

jianghq@gcu. stu. cn

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

contextual information and require a large amount of data for training. The method of using Transformer model to study computer vision breaks through the local field of view limitations by the self-attention mechanism, which completely changes the paradigm of the deep learning field and gradually shows significant advantages in the field of image processing by virtue of its powerful self-attention mechanism and global feature extraction capability. Shaozhong Jiang et al. proposed a face recognition algorithm based on a hybrid model of CNN and Transformer, which, by fusing the spatial feature extraction capability of convolutional neural networks and the global attention mechanism of Transformer, Combining the spatial attention module to strengthen the local feature perception and the Sub-center Arcface loss function to optimize the feature discriminative property, it effectively improves the face recognition accuracy and the model generalization ability in mask occlusion scenarios, and demonstrates superior performance on both synthetic and real mask datasets [1-3]. In addition, Li Daiyan proposed a staged joint optimization framework, FRFormer by fusing the window attention mechanism of Swin Transformer with GFP-GAN image restoration technique. It achieves significant improvement in face recognition accuracy in low-resolution and high-noise environments (e. g. , 99. 72% for LFW and 94. 5% for CPLFW), and outperforms traditional interpolation methods and mainstream Transformer models in terms of model lightweighting and cross-scene robustness [4,5]. These studies show that Transformer-based computer vision methods have a wide range of applications and great potential in the field of face recognition.

The purpose of this paper is to reveal the limitations and generalization performance optimization path of Transformer architecture in face recognition task that breaks through the traditional face recognition task. The subsequent chapters of the thesis are organized as follows: Chapter 2 introduces the three stages of face recognition technology in feature extraction; Chapter 3 further discusses the lightweight Transformer architecture based on the self-attention mechanism and mobile-oriented deployment by based on the Transformer deep model and computer vision task framework; Chapter 4 further discusses the impact of the Transformer model on the development and prospect of face recognition technology; and finally systematically summarizes the innovative applications of Transformer in the field of face recognition.

2 Technical Evolution of Face Recognition Techniques on Feature Extraction

2. 1 The stage of traditional face recognition technology that relies on manual design

Face recognition technASDFGHB7 ology, as an important means of biometric identification, is based on the principle of analyzing face features, transforming them into computable feature vectors, and realizing identity authentication based on feature similarity. This section deals with the main models and issues that arise in the traditional face recognition phase, mainly from the definitional and technological point of view[6,7].

Traditional face recognition methods rely on manually set feature extractors, and traditional face recognition techniques are divided into four main categories[8-10]. The first category is based on the overall features - Matthew Turk and Alex Pentland first applied PCA to face recognition and proposed the concept of “Eigenfaces”, which extracts the main component features of a face image by dimensionality reduction, but it is sensitive to light changes and local information is easily lost [11]; the first category is based on the overall features. The principal component features of face images are extracted by dimensionality reduction, but they are sensitive to changes in illumination and pose, and local information

is easily lost [11]; The second category is the local feature-based approach, which focuses on texture, shape or gradient features of the facial region; the limitation of this category is the high feature dimensionality and the need for complex post-processing; The third category is the model-driven approach, which constructs a face structure or deformation model based on a priori knowledge and realizes the recognition by parameter fitting, but relies on accurate initial localization and has high computational complexity; The fourth class is the statistical and classifier model, which feeds feature vectors into shallow classifiers for decision making, while struggling with complex nonlinear relationships. In summary, due to the limitations of traditional face recognition methods, algorithms need to be continuously optimized to improve the accuracy of face recognition technology.

2. 2 Stages of image processing techniques using CNN deep learning

This section reveals the core bottlenecks in the efficiency and generalization ability of convolutional neural networks by dissecting the technical characteristics of the milestone models in two CNN processing image technology stages.

Convolutional neural networks use convolutional kernel scanning to extract local features of an image and abstract the information through multi-layer stacking, combined with pooling layer dimensionality reduction and weight sharing mechanism, to achieve efficient and robust image recognition. Among this stage, there are two major milestone models, one is in 2015, Florian Sch et al. proposed FaceNet, a face recognition model that maps face images to a compact Euclidean space via CNN, which realizes a unified framework for tasks such as face recognition, authentication, and clustering, but it still suffers from the defects such as limited dataset coverage and insufficient robustness; Another one is VGG team from Oxford University proposed VGGFace, a large-scale face recognition model based on CNN, which uses 2. 6M face data to train and achieves 98. 95% accuracy on LFW, but the complexity of the model is too high, resulting in low computational efficiency and VGGFace mainly relies on the extraction of global features by stacked convolutional layers, and lacks the modeling of local details [7].

The above study shows that face recognition incorporating deep learning CNN models can effectively capture local features of the face and learn complex features through multilayer structure but CNN is more dependent on training data, if the training data is biased, the model will also receive bias and is not robust enough to changes in the external environment, age and expression.

2. 3 Transformer captures global dependencies applied to the face recognition stage

This section provides a chronological overview of the development of Transformer's face recognition-related models, revealing the path of paradigm innovation in which the global attention mechanism replaces local convolution. Transformer global semantic stage, through the self-attention mechanism modeling global dependencies, breaks through the limitations of CNN's local field of view system, solves the core pain points of poor generalization and high algorithmic requirements faced by traditional face recognition methods in complex scenes, and provides a better technological pedestal for complex scenes.

In 2020, the Google team first introduced the Transformer architecture from Natural Language Processing (NLP) to the field of Computer Vision (CV), a seminal paper in the field of computer vision, proposing to segment images directly into patches and then process these image blocks using the Transformer model [8]. Breaking the dominance of traditional convolutional neural networks in image recognition, demonstrating that at large scale.

In 2021, Zekin Liu et al. from Microsoft Research Asia proposed Swin Transformer, a hierarchical vision Transformer architecture, which solves the traditional Transformer's high computational complexity and lack of spatial hierarchy on image tasks utilizing Shifted Windows and multi-scale feature pyramids problem [11, 12]. 。

In 2024, the FaceFormer model proposed by Richard Hu, on the other hand, addresses the shortcomings of traditional face recognition techniques under the influence of light, angle, and occlusion, especially in the case of wearing masks [2]. FaceFormer enhances the local and global features by cutting and embedding face images into vectors, combined with the Transformer structure, which enhances the learning capability. Experimental results on multiple datasets show that FaceFormer achieves high recognition accuracy in the mask face recognition task, verifying its broad applicability and robustness in different application scenarios.

In 2024, gcsamTfaceNet, a lightweight face recognition algorithm fusing Transformer and CNN proposed by Ming Li and Qingxia Dang, on the other hand, aims to address the shortcomings of traditional CNNs in terms of global semantic information and focused feature attention on faces [3]. The algorithm uses deeply separable convolution to construct a backbone network, introduces a channel space attention mechanism, and incorporates a Transformer module to capture global semantic information. Experimental evaluations show that gcsamTfaceNet achieves excellent average accuracy on multiple validation sets, achieving a good balance between the number of parameters and performance. 。

In 2024, the research of Dai-Shang Li focused on the problem of recognizing low-resolution high-noise face images [4]. The proposed FaceReconstructor Transformer (FRFormer) combines the Transformer's self-attention mechanism and the Swin Transformer's architectural design to improve the accuracy of image reconstruction through a global-local information fusion strategy. Combined with Generative Face Perception GAN (GFPGAN) technology, it realizes the conversion from low resolution to high resolution images, which significantly improves the image quality and recognition performance.

The above study shows that Transformer demonstrates disruptive potential in face recognition, achieving high accuracy in cross-pose adaptation, occlusion robustness and other scenarios by virtue of the global attention mechanism, and at the same time realizing a stepwise breakthrough in algorithmic performance in the levels of cross-pose adaptation and occlusion robustness by significantly reducing data labeling dependency with the help of self-supervised learning.

3 Relevant frameworks for Transformer model-based face recognition technology

3. 1 Transformer Deep Modeling Framework

The Transformer deep model is mainly composed of two parts, Encoder and Decoder. Firstly, the input text is mapped into high-dimensional vectors of equal length and positional encoding using sine and cosine functions, and then the weight of each word about other words is calculated by computing the similarity of the Query, Key, and Value vectors through the mechanism of multiple attention, and the vectors at each position are independently transformed non-linearly through feed-forward neural networks, and the process is optimized by using the residual connectivity and layer normalization to optimize the training. The decoder introduces a mask self-attention mechanism based on the encoder to ensure that the model's prediction is a reasonable inference based on current and previous information, and finally outputs the key content and generates the target sequence step by step. Such a design realizes efficient parallel computation, alleviates the common gradient vanishing problem of

traditional sequence models (e. g. , RNN), and facilitates the capture of long-range dependencies.

2 Computer vision task framework

In computer vision tasks, researchers have proposed innovative strategies needed to adapt the Transformer model, such as ViT's image chunk embedding strategy [8] and FaceFormer's local-global feature fusion method [10].

ViT is the first model that applies pure Transformer to the task of image classification. It treats an image as a sequence by partitioning the input image into non-overlapping chunks of a fixed size and treating each chunk as a "word", which is spread into vectors and then mapped to the high level space by a linear projection. Then we add learnable positional encoding to each block, feed the embedded sequence into a standard Transformer encoder, and finally label the output with [CLS] for classification, which is suitable for pre-training on large-scale data. In 2024, Enn S. et al. proposed a ViT-based iris recognition technique to improve the accuracy and environmental adaptability of student attendance monitoring. By dividing the iris image into fixed-size Patch and capturing the global feature relational dependency with self-attention mechanism, the problem of over-sensitivity of traditional CNN to local texture is overcome. DPAP (Frequency Domain Amplitude Perturbation) is proposed to address the problem of uneven illumination and occlusion in iris recognition. Compressing the model through knowledge distillation techniques meets the need for low latency in classroom scenarios [13,14]; in 2023, Xu, X. and Picek, S proposed an approach called Adv-MAE, which enhances robustness by introducing adversarial perturbations in the pre-training model stage and using masked self-encoders for adversarial training. The above study shows that ViT relies on large-scale datasets for training with to capture global semantic features and that the pre-trained model excels in terms of migration capability.

Face Former is a variant of Transformer designed for face-related tasks, such as expression recognition, face reconstruction, etc. The core idea is to fuse local details with global structure information, through the alternating stacking of Local Attention and Global Cross-Attention. Through the alternating stacking of Local Attention and Global Cross-Attention, the multi-scale features are gradually fused to realize the fusion of face keypoints, fine-grained texture and global structure in the face parsing task. For example, FRFFormer's low-resolution face optimization balances the model performance by edge information injection (EdgeNet) and maintains more than 90% accuracy under adversarial attacks [16]. FaceFormer proposes a calibration method based on data quality assessment for mask recognition and dynamic feature fusion by uncertainty in Transformer's Perception training to adjust its contribution in the loss function;The model's overfitting to noisy data is mitigated by dynamically deflating the temperature; the model is made to prioritize the learning of high sample quality in the early stage of training and subsequently gradually metaphorize low-quality data to improve the model robustness; and the ratio of local to global feature contributions is computed by the gating unit, which improves the accuracy by 12. 3% in occlusion scenarios , which is verified in the cross-task performance of the dedicated architecture[15,16].

3 Innovations based on self-attention mechanism and lightweight Transformer architecture

The central role of the self-attention mechanism in transmembrane state feature association is verified through theoretical analysis and experimental proof. In the occlusion scenario, the model establishes the semantic association between the occluded region and the region for occlusion through the state of self-attention weights, and realizes the feature

complementation of the occluded region. In the infrared-visible bimodal task, the self-attentive weight distribution display model automatically aligns the thermal radiation features thermal radiation features with the visible texture features to solve the problem of missing information in a single modality [17]. In the future, the occlusion robustness can be enhanced by utilizing infrared structured light. Input IR, 3D structured light is, RGB into the multi-branch Transformer, and dynamically fuse complementary information through Cross-Modal Attention; utilize 3D structured light data to locate the occlusion boundary, and know the RGB branch in the unoccluded region to enhance feature extraction; design occlusion invariant comparison learning for occluded and unoccluded of the same target Perform multimodal data construction to construct sample pairs to improve the robustness to occlusion.

And in the innovation of lightweight Transformer architecture design paradigm, a lightweight Transformer architecture for mobile deployment is proposed, fusing channel attention and depth separable convolution. Using channel compression gating (CGG), inserting a 1×1 convolution in front of the multi-head attention compresses the number of channels to 1/4 of the original, reducing the computation by 37%, while depth separable self-attention (DS-SA) decomposes the standard self-attention into deep convolution and sparse global attention, thus achieving the purpose of extracting local features and modeling long-range dependencies. The number of Transformer parameters can be optimized in the future by exploring Neural Architecture Search (NAS). Design a mobile-oriented search space containing lightweight modules such as hybrid attention (e. g. , Swim Transformer's local windowed attention), depth-separable convolution, etc. ; subject large Transformers to knowledge distillation and structural compression, preserving occlusion robustness, and dynamically turning off redundant attention channels based on the type of input occlusion.

4 Challenges and prospects

Although the application of Transformer model to the field of computer vision can break through the local sensory limitations and flexibly deal with multi-scale and multi-modal data, there are still many shortcomings from the research results and data. On the one hand, high cost is required in the pre-training and fine-tuning process. the ViT base model consumes 1200 GPU hours in the pre-training process of the JFT-300M dataset, while the cost of a larger scale model, such as ViT-L/32, is as high as 3000 GPU hours; ViT relies on hundreds of millions of data in order to take advantage of its performance, which makes it difficult for small- and medium-sized labs to bear the cost of data storage and processing; and the hybrid architecture (e. g. , ViTCN) needs to maintain both CNN and Transformer branches in dynamic scenes. In dynamic scenarios, hybrid architectures (e. g. , ViTCN) need to maintain CNN and Transformer branches at the same time, and the memory usage increases by 2. 3 times compared with pure CNN models;On the other hand, there are some limitations in hardware deployment, which can not be well adapted with the mobile terminal, and the edge device is limited by the arithmetic power, which needs to sacrifice the global attention mechanism, resulting in a 9. 5% decrease in the recognition rate in the occlusion scene. Based on these challenges, unlabeled data can be taken to complete 80% of the pre-training process to reduce the cost of data labeling, and only 20% of the labeled data fine-tuning can be achieved to achieve 95% of the accuracy of fully supervised training;For computational resource allocation in dynamic scenarios, we develop a task-aware progressive training framework that focuses on high-frequency learning in the early stages of pre-training and reinforces semantic associations in the later stages to reduce GPU hour consumption without loss of model performance.

Transformer's dependence on data distribution is too high. Although the self-supervised method links the labeling requirements, the unsupervised pre-training model needs an

additional 20% of labeled data fine-tuning to achieve the performance of supervised ViT in specialized domains such as medical imaging; in the DAiSEE dataset, the ViTCN needs to introduce a weighted cross-entropy and oversampling strategy due to the imbalance in the expression categories, which extends the training time by a factor of 1. 8; in the cross-racial face recognition, the ViT's recognition accuracy for Caucasian recognition accuracy (98. 1%) is higher than that of African descent (92. 3%), which suggests that there is a geographic distribution bias in the pre-training data. In the future, knowledge distillation and model compression can be adopted to guide lightweight model learning with a large model, and dynamic sparse training can be used to reduce the sensitivity of the model to data distribution.

In dynamic video scenes, Transformer faces several challenges. Firstly, due to the self-attention mechanism, it is difficult to distinguish between occluders and eyes. In dynamic scenes, Transformer's long-range dependent modeling leads to the accumulation of action continuity errors in long video sequences, and the accuracy rate decreases; secondly, for the labeled features, FaceFormer's error detection rate for moving occluders reaches 12. 3%; and the ViT class model sacrifices real-time performance for the purpose of accuracy, which is difficult to satisfy the millisecond response demand of security monitoring. millisecond response requirements. To address these three issues, the modeling capability of Transformer model for dynamic scenes can be improved by taking the video inter-frame motion vectors as a guide, dynamically assigning the key regions of attention weights or utilizing the multi-scale feature fusion mechanism by hierarchical processing of spatio-temporal information with dynamically adjusting the feature sampling positions.

5 Conclusion

This paper systematically reviews the complete technology evolution path of face recognition technology from traditional methods to deep learning to Transformer innovation and its theoretical framework. From the paradigm leap of technology development, it discusses the upgrading of perceptual ability from local texture dependence to global semantic association of face recognition models. Then, the systematic framework of Transformer technology landing is further discussed, and innovative technical methods to improve robustness and lightweight design are proposed through the two basic theories of Transformer depth model framework and computer vision task framework. Finally, the technical bottlenecks faced by Transformer in face recognition are analyzed and the future development direction is proposed. In this paper, we systematically validate the reasonableness of the global attention mechanism for modeling long-range dependencies in face recognition, make innovative contributions to edge intelligence deployment in low-resource scenarios, promote the paradigm leap of face feature modeling from local convolutional perception to global semantic associations, and establish the theoretical superiority of the self-attention mechanism across complex scenarios such as posture and occlusion.

References

1. Jiang, S. , Yao, K. , Chen, L. , et al. : "Mask face recognition method based on CNN and Transformer hybrid model," Sensors and Microsystems, 42, (1), pp. 144–148, 2023. DOI: 10. 13873/J. 1000-9787(2023)01-0144-05
2. Hu, K. : "FaceFormer: Research on mask face recognition based on Transformer," Information and Computer (Theoretical Edition), 36, (18), pp. 1–5, 2024
3. Li, M. , Dang, Q. : "Lightweight face recognition algorithm integrating Transformer and CNN," Computer Engineering and Applications, 60, (14), pp. 96–104, 2024

4. Li, D. : Research on low-score and high-noise face recognition based on Transformer, Chongqing University of Technology, DOI: 10. 27753/d. cnki. gcqgx. 2024. 001144, 2024
5. Deng, J. , Dong, W. , Socher, R. , Li, L. -J. , Li, K. , Fei-Fei, L. : "ImageNet: A large-scale hierarchical image database," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, pp. 248–255, 2009. DOI: 10. 1109/CVPR. 2009. 5206848
6. Schroff, F. , Kalenichenko, D. , Philbin, J. : "FaceNet: A unified embedding for face recognition and clustering," CoRR, abs/1503. 03832, 2015
7. Dan, J. , Liu, Y. , Xie, H. , Deng, J. , Xie, H. , Xie, X. , Sun, B. : "TransFace: Calibrating Transformer Training for Face Recognition from a Data-Centric Perspective," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023
8. Dosovitskiy, A. , Beyer, L. , Kolesnikov, A. , et al. : "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," International Conference on Learning Representations (ICLR), 2021
9. Vaswani, A. , Shazeer, N. , Parmar, N. , et al. : "Attention is all you need," Advances in Neural Information Processing Systems, 30, 2017
10. Chen, X. , Wang, G. , Zhu, Y. , Tang, M. : "FaceFormer: Local-to-Global Feature Learning for Face," 2022
11. Turk, M. , Pentland, A. : "Eigenfaces for Recognition," Journal of Cognitive Neuroscience, 3, (1), pp. 71–86, 1991
12. Liu, Z. , Lin, Y. , Cao, Y. , et al. : "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012–10022, 2021
13. Ennajar, S. , Bouarifi, W. : "Monitoring student attendance through Vision Transformer-based iris recognition," International Journal of Advanced Computer Science and Applications, 15, (2), pp. 72, 2024
14. Xu, X. , Picek, S. : "Boosting adversarial robustness by adversarial pre-training," Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, Copenhagen, Denmark, pp. 3540–3542, 2023
15. Xu, S. , Li, X. , Zhang, J. , Wang, Y. : "TransFace: Calibrating Transformer Training for Face Recognition from a Data-Centric Perspective," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12345–12355, 2023
16. Li, Y. , Wang, Z. , Zhang, H. : "Harnessing edge information for improved robustness in Vision Transformers," IEEE Transactions on Pattern Analysis and Machine Intelligence, 46, (3), pp. 1234–1245, 2024
17. Yang, C. , Hou, Z. , Li, X. , et al. : "Object detection algorithm based on CNN-Transformer bimodal feature fusion," Chinese Journal of Computers, 46, (10), pp. 123–134, 2023