Public Figure Recognition Based on Pretrained Fine-Tuning

Zhongyu He[1], Darong Li[2], Linjie Qu[3], Ziming Su[4], Zihan Luo[5],

[1]23020241154398@stu.xmu.edu.cn, Information College Class
[2]23020241154408@stu.xmu.edu.cn, Information College Class
[3]23020241154348@stu.xmu.edu.cn, Information College Class
[4]23020241154437@stu.xmu.edu.cn, Information College Class
[5]23020241154430@stu.xmu.edu.cn, Information College Class

Abstract

With the rapid development of social media and the digital entertainment industry, public figures, especially celebrities, have increasingly become the focus of attention and research in people's lives. Accurately identifying and classifying the facial images of hundreds of celebrities has emerged as a challenging task in the field of computer vision, with significant practical value in applications such as celebrity activity tracking, fan interaction platforms, personalized entertainment content recommendation, and copyright protection. This study explores the application of a combined pretraining and fine-tuning approach to the task of celebrity recognition. Specifically, we utilized convolutional neural networks (CNNs) and Transformers pretrained on large-scale datasets as base models, augmented with classification heads tailored for celebrity recognition, and fine-tuned these models using a dedicated dataset comprising images of hundreds of celebrities. The goal of this study is to assess the feasibility and effectiveness of rapidly adapting pretrained models to the specific application scenario of celebrity recognition. The findings demonstrate that the pretraining and fine-tuning strategy provides a viable pathway for celebrity recognition. Different types of models exhibit distinct advantages in feature extraction, global information capture, and adaptation to complex data distributions. This research not only offers a technical solution for public figure recognition but also provides valuable insights into the application of models in the broader field of image classification. Future research will further explore more efficient pretrained models and optimized fine-tuning techniques to address increasingly complex and dynamic image recognition tasks.

Introduction

With the booming development of social media and the digital entertainment industry, public figures, especially celebrities, are becoming increasingly crucial in people's lives. Their faces not only influence fans' purchasing behaviors and public opinion but also play pivotal roles in film, advertising, and social platforms. Accurately identifying and classifying the facial images of hundreds of celebrities is not only a rich and interesting research topic but also a direction with practical application value in the field of computer vision. By accurately recognizing celebrities, it can aid in building fan interaction platforms, optimizing personalized entertainment content recommendation systems, and play important roles in copyright protection and tracking celebrity activities. Over the past decade, image classification methods based on Convolutional Neural Networks (CNNs) and Transformers have made significant progress. Early CNN models (such as LeNet (LeCun et al. 1998), AlexNet (Krizhevsky, Sutskever, and Hinton 2012), VGG (Simonyan and Zisserman 2014), and ResNet (He et al. 2016)) achieved excellent performance in general image classification tasks by continuously deepening the network structure and introducing residual connections. However, as task complexity and data scale continue to increase, CNNs still face limitations in capturing global dependencies and long-range feature interactions. In recent years, Transformer architectures (such as Vision Transformer (ViT) (Dosovitskiy et al. 2021)) have demonstrated new potentials in high-complexity image classification tasks by effectively modeling global information of images through self-attention mechanisms. However, Transformers typically require high computational resources and large datasets, which limit their direct application in data and resource-constrained environments. Additionally, the computational overhead of Transformers in high-resolution image processing and real-time application scenarios cannot be ignored. To address these issues, the pretraining and fine-tuning strategy has garnered increasing attention. In the fields of natural language processing and computer vision, pretrained models learn universal feature representations on large-scale general datasets and then adapt to specific tasks through fine-tuning, thereby demonstrating excellent training efficiency and adaptability (Radford et al. 2021; Devlin et al. 2019; He et al. 2022). This approach not only accelerates convergence and improves training efficiency but also enhances the model's performance on specific tasks through transfer learning, especially when data is limited (Zoph et al. 2020). We have introduced this paradigm into the task of celebrity recognition. This study explores various pretrained models based on CNNs and Transformers. We selected multiple models pretrained on large-scale datasets such as ImageNet (Deng et al. 2009) as base models and added classification heads suitable for public figure recognition tasks. Subsequently, these models were fine-tuned using a dedicated dataset containing hundreds of celebrities. By systematically evaluating the performance of different models on the public figure recognition task, we analyzed their respective strengths and weaknesses and explored the applicability and performance differences of the pretraining and fine-tuning strategy across different architectures. Additionally, we delved into dataset construction, model optimization, and potential directions for performance improvement, aiming to provide effective technical support and theoretical guidance for public figure recognition. Compared to traditional direct training methods, this strategy is expected to adapt more quickly to the celebrity recognition scenario and achieve satisfactory recognition results under relatively limited data conditions. Research results indicate that the pretraining and fine-tuning strategy is a correct approach for handling large-scale, complex image classification tasks. Different pretrained models exhibit unique characteristics in the task of celebrity recognition: CNN models excel in feature extraction and computational efficiency, while Transformer models have distinct advantages in capturing global dependencies and complex feature interactions. Through fine-tuning, the classification accuracy and generalization ability of all models have improved, validating the effectiveness of the pretraining and fine-tuning method. This study not only provides practical solutions for public figure recognition but also offers valuable references for further optimizing and expanding image classification technologies.

Related Work

Person recognition is a key research topic in computer vision and pattern recognition, aiming to identify, verify, or classify individuals by analyzing their features in images or videos. In the field of visual recognition, CNN-based and Transformer-based models have played an essential role, not only advancing the technology but also providing valuable insights and guidance for subsequent research.

CNN-based Models

EfficientNet (Tan and Le 2019) scales the network's depth, width, and resolution within a unified framework, achieving high accuracy while reducing computational resource usage, thus advancing research into network scaling methods. RegNet (Radosavovic et al. 2020) searches for optimal network architectures by regularizing the design space, offering a fresh perspective on network structure design, particularly in network regularization and design space exploration. ResNet (He et al. 2016) addresses the degradation problem in deep network training by introducing residual learning, which alleviates the vanishing gradient issue. Its skip connection design has had a profound impact on deep learning architectures. ResNeXt (Xie et al. 2017) divides the network into multiple parallel branches and uses group convolutions within each branch, increasing the number of parameters while maintaining the same computational cost. This design has further advanced research on deep network structures.

Transformer-based Models

Swin Transformer (Liu et al. 2021) adopts a Transformer encoder structure and introduces a ”Swin Transformer Block” that includes Multi-Head Self-Attention (MHSA) and a Multi-Layer Perceptron (MLP), demonstrating excellent performance in image data processing and inspiring subsequent research on Transformer-based visual models. Vision Transformer (ViT) (Dosovitskiy et al. 2021) divides an image into patches and uses a self-attention mechanism to capture global dependencies, achieving a breakthrough in computer vision and sparking extensive research on Transformer-based models. CaiT (Touvron et al. 2021) builds upon the Transformer architecture by introducing a class-attention mechanism, enabling the model to dynamically focus on features from different categories during self-attention computation. This has further promoted the application of Transformer architectures in visual recognition tasks and showcased the potential of class-attention mechanisms to enhance model performance.

Method

In this section, we will introduce our approach. We adopt a method of fine-tuning pre-trained weights from different models on the same dataset (as shown in Figure 1), aiming to compare the performance of various methods on the same dataset. The pre-trained weights are obtained from large-scale datasets such as ImageNet. Our approach is divided into two parts: one based on CNN architectures and the other based on Transformer architectures. Among the CNN-based models, the best-performing one is ResNeXt-50, while the best-performing Transformer-based model is the CaiT model.

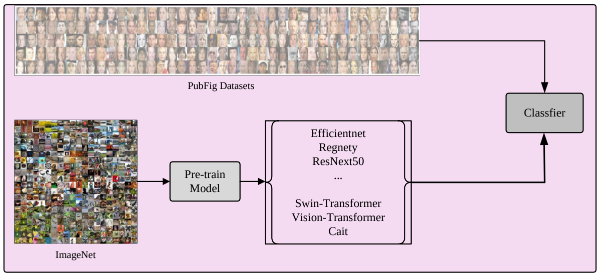

Figure 1: Overview of the proposed method framework.We fine-tune multiple different ImageNet pre-trained models on the PubFig dataset to enable person recognition of public figures.

ResNeXt

ResNeXt is an enhanced Convolutional Neural Network (CNN) architecture initially proposed by the Facebook AI Research team. It builds upon the traditional Residual Networks (ResNet) by introducing the concept of Cardinality, which refers to the number of parallel paths in each residual block. This allows the network to achieve stronger representational power while maintaining the same computational budget. The core idea of ResNeXt is to improve model performance by increasing the number of paths in the network, rather than simply increasing depth or width. ResNeXt uses larger convolution kernels (e.g., 7x7) for initial feature extraction from the input image, typically followed by a pooling layer to reduce spatial dimensions. The network is built by stacking multiple residual blocks, each containing several parallel paths, with each path performing computations using grouped convolutions. At each stage, multiple paths are employed to enhance the representational capacity. After passing through multiple layers of convolution and residual blocks, the feature maps are downsampled using a global average pooling layer to generate a fixed-dimensional feature vector. Finally, the feature vector is passed through a fully connected layer for classification tasks. In the multi-path structure of ResNeXt, each residual block contains G parallel paths (i.e., Cardinality), which are implemented through grouped convolutions. Let the number of input channels of the feature map be denoted by C_{in}, and the number of output channels be C_{out}. Each path processes $\frac{C_{in}}{G}$ input channels. For each path $g \in \{1, 2, \ldots, G\}$, the computation of grouped convolution is given by:

$$\mathbf{y}_g = \mathbf{W}_g * \mathbf{x}_g + \mathbf{b}_g \tag{1}$$

$\mathbf{x}_g$ is the input feature subset for path $g.\mathbf{W}_g$ is the convolution kernel weight for path $g.\mathbf{b}_g$ is the bias term for path $g.*$ represents the convolution operation.$\mathbf{y}_g$ is the output feature subset for path g. The outputs from each path are combined into the final output feature map by either concatenation or summation. The merging operation is given by:

$$\mathbf{y} = \sum_{g=1}^{G} \mathbf{y}_g \tag{2}$$

ResNeXt's residual block combines the multi-path structure and the residual connection, and the output is formulated as:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{\mathbf{W}_g\}) + \mathbf{x} \tag{3}$$

$\mathbf{x}$ is the input feature map. $\mathcal{F}(\mathbf{x}, \{\mathbf{W}_g\})$ is the output computed by the G parallel paths using grouped convolutions.The output channels of $\mathbf{x}$ and $\mathcal{F}$ must be consistent. If they are not, a 1×1 convolution is used to adjust the dimensions.

CaiT

Based on previous experiences, increasing the depth of a model allows the network to learn more complex representations. For example, ResNet shows an improvement in accuracy as the depth increases from 18 layers to 152 layers. However, as the depth increases, the model becomes increasingly difficult to converge. To address the depth issue, CaiT (Class-Attention in Transformer) introduces two key improvements: LayerScale and Class-Attention. The CaiT model is a Transformer architecture that combines causal reasoning and visual perception, aiming to enhance the performance of visual tasks through improved causal relationship modeling. It introduces a causal reasoning mechanism into the visual model by adjusting the attention calculation, thereby improving the model's comprehension and generative abilities. The core idea of the CaiT model is to construct the information flow via causal self-attention, enabling the model to better capture long-range dependencies and more effectively perform image understanding. One of the key innovations of CaiT is placing the class token later in the model. This approach allows the earlier layers to focus solely on learning self-attention, eliminating the optimization conflicts encountered in earlier layers. The last two layers are replaced with Class-Attention, while the overall structure remains based on self-attention. However, in the Class-Attention mechanism, only the class token is updated, while the patch embeddings are kept fixed. The model structure begins with a patch embedding operation, where the input image is divided into various patches. Next, the Self-Attention operation is applied. Unlike ViT, CaiT introduces LayerScale, which multiplies the output of each residual block by a learnable diagonal matrix. The addition of LayerScale does not alter the representational power of the structure, yet it allows deeper models to converge more effectively. Finally, the Class-Attention (CA) module is employed. The purpose of the CA layer is to extract information from the processed patch embeddings using the class token (CLS). LayerScale adds a learnable diagonal matrix to the output of each residual block, which is initialized close to zero. Adding this simple layer after each residual block improves the training dynamics, enabling the training of deeper and higher-capacity Transformers.

$$x_l^{'} = x_l + diag(\lambda_1, ..., \lambda_d) \times SA(Norm(x_l)) \tag{4}$$

$$x_{l+1} = x_l^{'} + diag(\lambda_1^{'}, ..., \lambda_d^{'}) \times FFN(Norm(x_l^{'})) \tag{5}$$

Class-Attention is a structure similar to the Encode/Decode mechanism. Unlike Self-Attention, which focuses on self-representations, Class-Attention emphasizes extracting information from processed patch tokens. In contrast to Self-Attention (SA), where the query variable z is typically computed as the query vector Q, in Class-Attention, the query is replaced by the class token x_{class}, while the keys K and values V remain unchanged. The formula is as follows, where:

$$\mathbf{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} + \mathbf{M}\right)\mathbf{V} \tag{6}$$

$$Q = W_q x_{class} + b_q \tag{7}$$

$$K = W_k z + b_k \tag{8}$$

$$V = W_v z + b_v \tag{9}$$

$z = [x_{class}, x_{patches}]$.Here, x_{class} is the class token, and $x_{patches}$ are the processed patch embeddings.

Model Training

The loss function used in this experiment is the cross-entropy loss (Bishop 2006),, which is commonly used in classification tasks, especially in multi-class problems. It computes the loss by measuring the difference between the predicted probability distribution and the true label distribution. For binary classification problems, the definition of the cross-entropy loss is as follows:

$$\mathcal{L}_{binary} = -\left[y \log(p) + (1-y)\log(1-p)\right] \quad (10)$$

In this experiment, the optimizer used is Adam (Kingma and Ba 2015).

Experiment

Dataset

The PubFig dataset, meticulously curated by a research team at the Massachusetts Institute of Technology (MIT), is designed to provide a rich and diverse facial image resource for researchers in the fields of computer vision and machine learning (Kumar et al. 2009). The dataset primarily consists of facial photographs of public figures (e.g., actors, politicians, singers) collected from the internet, ensuring a wide range of sources and high levels of authenticity and diversity. The PubFig dataset contains thousands of high-quality images representing hundreds of different public figures. Images for each individual were captured under various conditions, including different facial expressions, poses, lighting, and backgrounds, enhancing the robustness and generalization capabilities of facial recognition algorithms in real-world applications. In this study, we extended the PubFig dataset by incorporating 20 carefully selected images from the KunKun dataset, forming a new category—Xukun Cai. This addition aims to simulate a realistic scenario of quickly and accurately identifying a specific individual, Xukun Cai, among multiple celebrities. By mixing Xukun Cai's images with those of other public figures from the PubFig dataset, we constructed a more challenging and realistic dataset to test and evaluate the performance of various computer vision models in fine-grained facial recognition tasks.

Experimental Setup

For the experimental setup, we utilized the Timm (PyTorch Image Models) library provided by Hugging Face (Wightman 2019), which facilitates access to various computer vision models pretrained on the ImageNet dataset (Deng et al. 2009). In the experiments, we selected four CNN-based models and three Transformer-based models for comparative analysis. The models used are as follows: The models used in this study include CNN-based architectures such as EfficientNet-B3 (Tan and Le 2019), RegNetY-160 (Radosavovic et al. 2020), ResNet50 (He et al. 2016), and ResNeXt50_32x4d (Xie et al. 2017), as well as Transformer-based architectures like Swin Transformer Base Patch4 Window7 224 (Liu et al. 2021), ViT Base Patch16 224 (Dosovitskiy et al. 2021), and CaiT-S24-224 (Touvron et al. 2021). These models were selected to cover a range of widely used and high-performing methods in computer vision.

Result

From the table 1, it can be observed that CNN-based models outperform Transformer-based models overall. Notably, ResNeXt50_32x4d achieves a classification accuracy of 0.956, significantly higher than the other models. In contrast, Transformer-based computer vision models demonstrate comparatively lower performance on this experimental dataset. Even the best-performing Transformer-based model, CaiT-S24-224, achieves a classification accuracy of only 0.867. Further analysis of the same models under different pretraining conditions reveals that base models initialized with ImageNet pretrained parameters exhibit a clear performance advantage on this experimental dataset. Specifically, all CNN-based models show higher classification accuracy when pretrained parameters are used compared to their counterparts without pretraining, confirming the effectiveness of pretrained parameters in improving model performance. Based on the above experimental results, we selected the best-performing model, ResNeXt50_32x4d, and saved its parameters for subsequent generalization experiments. These experiments aim to investigate whether the model overly relies on specific features of Xukun Cai, such as the signature middle-part hairstyle, for identity recognition. To this end, we collected two additional celebrity photos from the internet: one featuring a middle-part hairstyle but not Xukun Cai, and the other featuring Xukun Cai but without the middle-part hairstyle. The generalization experiment framework is illustrated in the figure 2 .

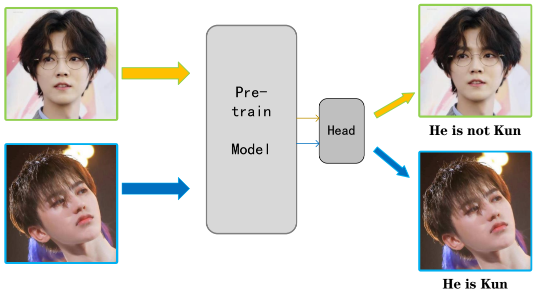

Figure 2: Overview of the proposed method framework. We use trained models and classification heads to identify Xukun Cai's face images and evaluate the model's prediction performance through relevant indicators.

We input the two selected images into the saved model for classification and evaluated its ability to effectively recognize the specific public figure using the Top-N Accuracy metric.

As shown in Table 2, the results indicate that the model did not misclassify Xukun Cai as another celebrity, nor did it misidentify other celebrities as Xukun Cai. This demonstrates that the model does not solely rely on superficial features, such as hairstyle, but is capable of accurately capturing more distinctive facial features for identification.

Discussion

This experiment revealed that Transformer-based models underperformed on the dataset used in this study. Despite their strong performance on large-scale datasets, such as Vision Transformer (ViT), Transformer models pretrained on ImageNet failed to meet expectations in this facial recognition task.

We hypothesize that ViT's patch-based processing struggles with facial recognition, as celebrity face datasets often

Model	Acc(Pre-trained)	Acc(Non)
EfficientNet-B3	0.9515	0.7183
RegNetY-160	0.8010	0.8023
ResNet50	0.9219	0.7517
ResNeXt50_32x4d	0.9563	0.8178
Swin Transformer Base Patch4 Window7 224	0.0369	0.0352
ViT Base Patch16 224	0.0926	0.0403
CaiT-S24-224	0.5931	0.5426

Table 1: Performance comparison of models under pre-trained and non-pre-trained conditions.

Model	Top-1 Acc	Top-3 Acc	Top-5 Acc
EfficientNet-B3	0.9213	0.9814	0.9903
RegNetY-160	0.8231	0.8722	0.9010
ResNet50	0.9042	0.9531	0.9754
ResNeXt50_32x4d	0.9446	0.9811	0.9952
Swin Transformer Base Patch4 Window7 224	0.0571	0.1031	0.1544
ViT Base Patch16 224	0.0822	0.2104	0.3147
CaiT-S24-224	0.5387	0.6933	0.8023

Table 2: Comparison of model performance in recognizing specific public figures.

exhibit high similarity between patches, making it difficult to capture subtle distinguishing features. Additionally, facial recognition heavily relies on fine-grained details, such as the eyes, nose, and mouth. CNN-based models, with their strength in local feature extraction, are better suited for this task compared to Transformers, which focus on global features.

In conclusion, while Transformers hold great promise for various vision tasks, CNN-based models remain superior for celebrity facial recognition. Future work could explore hybrid CNN-Transformer architectures or optimize Transformers for facial feature extraction to improve performance in this domain.

Conclusion

This study focuses on the task of celebrity face recognition and systematically explores the application of pretraining and fine-tuning strategies in different computer vision models. The experimental results show that CNN-based models outperform Transformer-based models in this task, with ResNeXt50_32x4d achieving the best classification accuracy of 95.6%. This validates the superiority of CNNs in extracting local features and adapting to high-similarity image classification tasks. However, Transformer-based models did not perform as well in celebrity face recognition. We speculate that this is related to the limitations of Transformers in capturing fine-grained features. Although Transformers have demonstrated performance comparable to or even surpassing traditional CNNs on large-scale datasets, their global feature extraction capabilities fail to effectively distinguish subtle differences between individuals in datasets with highly similar facial images. The experimental analysis also highlights the importance of pretraining strategies. For both CNN and Transformer architectures, pretraining significantly improved classification performance, indicating that leveraging large-scale pretrained models and fine-tuning them is an effective approach under limited data conditions. This study not only provides a feasible technical solution for celebrity recognition tasks but also offers valuable references for further optimization of image classification technologies. Future research could explore hybrid architectures combining CNNs and Transformers or design Transformer models tailored for face recognition tasks to further enhance their performance in fine-grained face recognition scenarios.

References

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. Springer.

Deng, J.; et al. 2009. ImageNet: A large-scale hierarchical image database. *CVPR*.

Devlin, J.; et al. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*.

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.; et al. 2022. Masked autoencoders are scalable vision learners. *CVPR*.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 1–13.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 1097–1105.

Kumar, N.; Berg, A. C.; Belhumeur, P. N.; and Nayar, S. K. 2009. Attribute and Simile Classifiers for Face Verification. *ICCV*.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Liu, Z.; et al. 2021. Swin Transformer: Hierarchical vision transformer using shifted windows. In *CVPR*.

Radford, A.; et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Radosavovic, I.; et al. 2020. Designing network design spaces. *CVPR*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tan, M.; and Le, Q. V. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, 6105–6114.

Touvron, H.; et al. 2021. Going deeper with image transformers. *Proceedings of NeurIPS*.

Wightman, R. 2019. timm: PyTorch Image Models. *GitHub repository*.

Xie, S.; et al. 2017. Aggregated residual transformations for deep neural networks. *CVPR*.

Zoph, B.; et al. 2020. Rethinking pre-training and self-training. *NeurIPS*.