Facial Beauty Prediction with Vision Transformer

Duy Tran[1*], Thang Le [1*], Khoa Tran[1*], Hoang Le[1*] , Cuong Do[1*], Thanh Ha[1*]

* These authors contributed equally to this work

[1] The George Washington Institute of Data Science and Artificial Intelligence, The International Society of Data Scientists, MA, USA

{duy.tran, thang.le, khoa.tran, hoang.le, cuong.do, thanh.ha}@isods.org

Abstract. In addition to its use in the realm of plastic surgery and aesthetics, Facial Beauty Prediction technology also has applications in other areas, such as advertising and social media, where it can be used to optimize marketing strategies and help individuals enhance their online presence. There are applications in other areas as well, such as advertising and social media. This study introduces an effective approach to evaluate human face beauty using a transformer-based architecture. While Convolutional Neural Network (CNN) is a conventional method for this task, our experimental results demonstrate that our Vision Transformer (ViT) based model outperforms the other two effective baselines, VGGNet and ResNet50, in evaluating human face beauty on the widely-used benchmark dataset SCUT-FPB 5500. Our ViT-based model demonstrates superior performance in Mean Absolute Error (MAE) and Mean Squared Error (MSE) compared to VGG16 and ResNet-50, despite employing a simple data pipeline without any data augmentation. Our study suggests that transformer-based architectures offer a more effective means of evaluating human beauty and open new avenues for further research in this field.

Keywords: Artificial Intelligence, Vision Transformer, Facial features extraction, VGG-16, Resnet-50, ViT-based-16-2k

1. Introduction

Facial Beauty Prediction is a Computer Vision task utilizing machine learning methods to process and analyze facial features, attributes, and landmarks. These extracted features are then used as input to machine learning algorithms, such as deep neural networks, to learn the relationship between face beauty and facial patterns. The output of beauty prediction algorithms is usually a numerical score representing facial attractiveness.

In the present study, most state-of-the-art performances for facial beauty prediction are achieved by Convolutional Neural Networks (CNN). Xu et al. (2018) proposed “Transferring Rich Deep Features for Facial Beauty Prediction,” combining both VGG16-feature extraction and a Bayesian Ridge regression to achieve relatively high performance on SCUT-FBP5500 with a Mean Absolute Error score of 0.2595[1]. Following the success of CNN, Bougourzi et al. (2021) experimented with a variety of CNN architectures, including Resnet-50, and proposed a novel architecture integrating all models into one with highly competitive results[2].

The use of Transformers in deep learning models has gained significant popularity recently, being applied to both natural language processing and computer vision tasks. Transformer-based models such as BERT, GPT-3, and the recently released GPT-4 have demonstrated state-of-the-art performance in various language modeling tasks [3]. Likewise, there is a growing trend of employing Vision Transformer (ViT) and its variants in various computer vision tasks, such as image classification [4] and object detection [5]. Recent research by Dosovitskiy et al. (2021)

has demonstrated that ViT exhibits remarkable performance when compared to state-of-the-art CNN while also requiring significantly fewer computational resources during the training process[4]. Realizing the absence of transformer-based models in this task and the potential of ViT, especially when trained on massively large databases such as ImageNet[6], this paper aims at developing ViT, which is pre-trained on a huge dataset, then fine-tuned and evaluated on several benchmark datasets for Facial Beauty Prediction such as SCUT-FBP5500 [1]. Similarly, CNN models such as ResNet-50 and VGG-16, both pre-trained on ImageNet, are also re-implemented to compare their performance with that of Vision Transformers.

This paper walks through some of the related work in the field of Facial Beauty Prediction in Section 2. Then, it will propose the application of ViT architecture for facial beauty prediction in Section 3 before conducting several experiments with ViT along with other benchmarks for performance comparison and analysis. Based on the experimental results, Section 5 will draw conclusions about the architecture of ViT and suggest some future works and progress to improve the performance of our proposed architecture.

2. Related work

In the context of beauty prediction, there have been many research papers suggesting approaches to this task. Iyer et al. (2021) explored machine learning-based facial beauty prediction using facial landmarks and traditional image descriptors, nose symmetry ratio, for instance. These attributes are then input to various traditional machine learning algorithms such as Linear Regression, Random Forest, K-Nearest Neighbour, v.v. to output a score for facial attractiveness [7].

However, one disadvantage of this approach is its heavy dependence on handcrafted features, which requires extensive domain-specific knowledge and subjective design choices about facial beauty. Furthermore, since preferences and opinions about facial beauty may vary among different races and generations, re-implementing feature engineering to align with the interests of a new target audience could be challenging.

Avoiding highly accurate facial characteristics, Xiao et al. proposed Beauty3DFaceNet - a deep CNN predicting attractiveness on 3D faces using both geometry and texture information [8]. The model utilizes a fusion module to combine geometric and texture features and designs a novel sampling strategy based on facial landmarks for improved performance in learning aesthetic features. Nevertheless, this architecture is currently restricted due to difficulties in implementation, and the 3D dataset for facial beauty prediction is limited because of constraints in 3D data collection.

An alternative approach is implementing CNN. CNN is a data-driven model that automatically extracts features and learns the associations between rating scores and facial features and landmarks during the training process. However, in “Facial Expression Recognition using Convolutional Neural Networks: State of the Art,” Christopher Pramerdorfer and Martin Kampel discussed the limitation of using CNN in this sort of task. CNN has a tendency to focus on local regions of an image, which may not capture the global spatial information effectively [9]. This limitation is particularly relevant as facial features often have a complex interplay, influencing the overall face prediction. For instance, in certain East Asian cultures, the culturally desirable traits may include larger eyes with a "double eyelid" and a more "V-shaped" face with a slender jawline.

In contrast, Transformer-based models, as demonstrated by Dosovitskiy et al. (2021), offer a solution to address both the challenges mentioned above. Not only was ViT fully capable of capturing both local and global contextual information in images, but it also demanded fewer computational resources for training compared to state-of-the-art CNN [4], which is why a supervised pre-trained ViT is implemented in this paper to overcome these limitations.

3. Method

3.1 Dataset

The availability of the SCUT-FBP5500 dataset has contributed to the development of more accurate and efficient machine learning models for predicting facial beauty, with potential applications in various industries and fields such as cosmetics, advertising, and social media [10].

The SCUT-FBP5500 dataset contains 5,500 face images, with 2,500 images of males and 3,000 images of females, with participants aged from 15 to 60. The dataset is split into 2000 Asian males, 2000 Asian females, 750 Caucasian males, and 750 Caucasian females.

The SCUT-FBP5500 dataset is evaluated by multiple human evaluators to ensure a diverse and subjective assessment of facial beauty. A total of 60 evaluators individually and independently rate the beauty of the facial images using a 5-point scale with 1 as the smallest score. They consider various facial features such as skin complexion, facial symmetry, attractiveness, and aesthetic appeal. The final beauty score for each image is obtained as the average of the scores given by the evaluators, minimizing subjective biases because different raters may have different perceptions of beauty.

3.2 Model Architecture

3.2.1 Resnet-50 and VGG-16

ResNet-50, as illustrated by Zhang et al. in “Deep Residual Learning for Image Recognition”, [11] consist of 50 layers and include key components such as convolutional layers, batch normalization layers, ReLU activation functions, and skip connections. The convolutional layers extract features from input images, while batch normalization layers normalize inputs to each layer for improved training stability. ReLU activation functions introduce non-linearity for learning complex patterns. The skip connections allow the network to bypass one or more layers, facilitating direct gradient propagation and fluid gradient backpropagation across the network and mitigating the vanishing gradient problem.

VGG16, referenced by Simonyan et al. in “Very Deep Convolutional Networks for Large-Scale Image Recognition”[12], consisting of 16 layers including 13 convolutional layers and 3 fully connected layers, utilizes small 3x3 convolutional filters stacked on top of each other, with occasional max-pooling layers for downsampling.

These two CNN models were initially pre-trained on the ImageNet dataset, and subsequently fine-tuned and trained on the SCUT-FBP5500 dataset for further model evaluation and performance comparison against the proposed Beauty ViT model.

3.2.2 ViT Architecture

The Facial Beauty Prediction Transformer follows the architecture of ViT, depicted in Fig. 1.

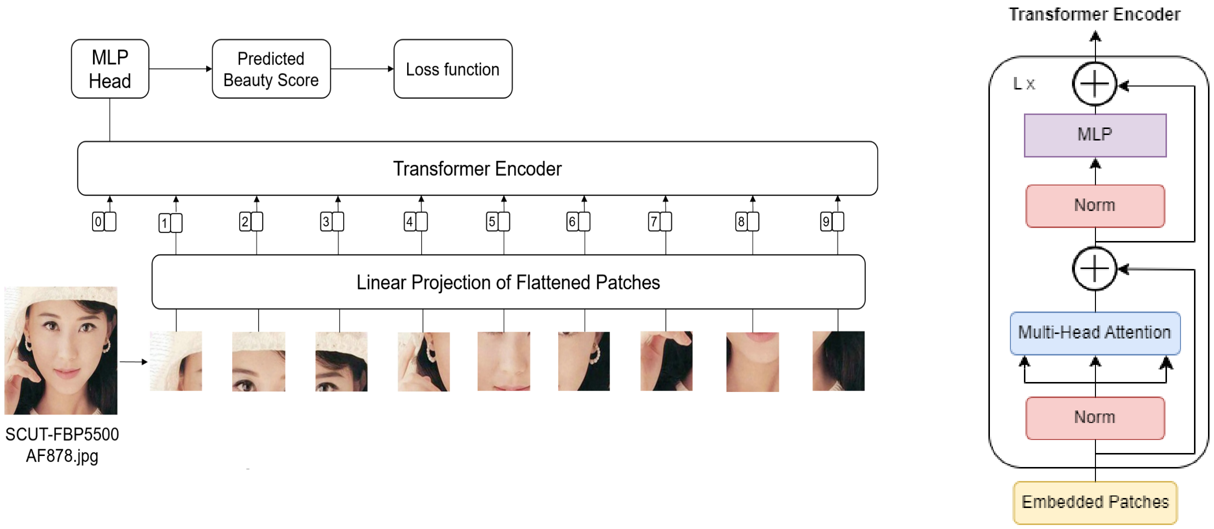

Fig. 1. Facial Beauty Prediction ViT architecture and Transformer Encoder component

The ViT model for Face Beauty Prediction is an effective transformer-based model tailored for facial recognition tasks. The architecture comprises a patch embedding layer, a transformer encoder, and an MLP (Multilayer Perceptron) head for output. The patch embedding layer dissects the input image into a multitude of patches, which are subsequently flattened and fed to the transformer encoder. The transformer encoder, stacked with self-attention layers, enables the model to capture long-range dependencies among various aspects of the face image. The output feature vector from the transformer encoder, representing the holistic face image, feeds into the MLP head, which produces the predicted beauty score for evaluation and backpropagation.

3.2.3 Loss function

The loss function was employed to assess the performance of the models on the SCUT-FBP5500 dataset integrates two critical metrics: Mean Squared Error (MSE) and Mean Absolute Error (MAE). These metrics offer a quantitative measure of the disparity between the predicted facial beauty scores (y_{l}) from the MLP and the corresponding ground truth scores (y) for each image (i) in the dataset.

The MSE is computed by averaging the squared differences between the predicted and ground truth beauty scores.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2$$

The symbol 'n' is the total number of instances in the dataset. The squared term in MSE amplifies larger prediction errors, making it more sensitive to significant discrepancies between predicted and ground truth scores.

Conversely, the MAE is the average of the absolute differences between the predicted and ground truth beauty scores.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{y_i}|$$

The MAE calculates the average of the absolute differences between the predicted and actual beauty scores. Unlike MSE, MAE is equally sensitive to all differences between the predicted and actual scores, irrespective of their size.

Employing both MSE and MAE in the loss function offers a comprehensive evaluation of the model's performance from unique perspectives. Smaller MSE and MAE values indicate superior model performance, denoting lesser deviations between the predicted and actual beauty scores, thus indicating higher model precision and accuracy. Additionally, the Pearson correlation coefficient is employed to offer additional insight into the model's performance.

4. Experiments and Result

Although the ViT does require more computational resources to train, the superior performance and faster convergence time make it a highly effective and efficient choice. Table 1 illustrates the results of our experiments using the 3 models.

Table 1. The number of parameters, Pearson-correlation (PC) scores, MSE, and MAE of each model

Models	Parameters	PC	Mean Squared Error	Mean Absolute Error
VGG16	134M	0.78	0.4675	0.5680
ResNet-50	23M	0.81	0.1609	0.3113
Vision Transformer	86M	**0.9127**	**0.0758**	**0.2087**

Our findings revealed that ViT significantly outperforms state-of-the-art Convolutional Neural Networks (CNNs), achieving a Mean Absolute Error (MAE) score of 0.2087, with 86 million parameters. Despite possessing a higher number of parameters than ResNet-50, ViT demonstrated superior performance across all three evaluation metrics. Furthermore, ViT showed remarkable efficiency during training, converging to the final optimal result after around ten epochs, as opposed to the approximate 100 epochs required for the CNN models. This efficiency may be attributed to the successful application of transfer learning from a pre-trained version of ViT on ImageNet.

However, the training time remained a challenge; on an NVIDIA Tesla T4 15GB graphic card, training ViT for ten epochs took nearly the same amount of time as training VGG16 or ResNet-50 for 100 epochs—approximately 30 minutes. Training ViT for 50 epochs took two hours, confirming that this model requires more computational resources and time. Future work will focus on optimizing our models and algorithms to enhance performance and computational efficiency

5. Conclusion and Future Work

Overall, our paper emphasizes the outstanding performance of ViT architecture for facial beauty prediction and the potential of using ViT variants for this task. The success of ViT can be attributed to its complex architecture and transferability, which allows it to be pre-trained on highly large-scale datasets such as ImageNet and then fine-tuned on SCUT-FBP5500 for performance optimization, which demonstrated that the ViT model is superior to other models such as VGG and ResNet for human beauty prediction.

This finding opens new opportunities for further research in using transformers and other advanced techniques for feature detection in computer vision tasks, particularly in applications such as human beauty evaluation, where accurate feature extraction is critical. Future research could explore the application of ViT and other transformer-based models to related tasks such as facial expression recognition, age estimation, and gender classification. These tasks also require accurate feature extraction and could benefit from using transformer-based models.

One potential avenue for future research is to use better pre-trained models for ViT. Currently, ViT is pre-trained on ImageNet, which consists of only a small number of instances for humans in general and facials specifically. Therefore, in the future, if the team can have access to models pre-trained on facial datasets, the performance is likely to be better.

Another potential avenue for future work is to explore other transformer variants, such as Swin-Transformer, which has shown promising results in other computer vision tasks. It would be interesting to investigate whether DeiT (Data-efficient Image Transformers) could achieve even higher accuracy for face beauty evaluation than ViT [13].

Another area of future research could be to explore ensembling methods for combining multiple models. Ensembling has been shown to be an effective way to improve the accuracy of deep learning models by combining the strengths of multiple models. It would be interesting to investigate whether ensembling ViT with other models, such as ResNet or VGG, could achieve even higher accuracy for face beauty evaluation.

Overall, there are several promising directions for future research in this area, and we hope that our work will inspire further investigation into the use of transformer-based models for evaluating human beauty.

References

1. Xu, L., Xiang, J., Yuan, X.: Transferring Rich Deep Features for Facial Beauty Prediction (Version 1) (2018).
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision, LNCS, pp. 213-229. Springer, Cham (2020).
3. OpenAI. GPT-4 Technical Report. ArXiv, (2023). Available: https://arxiv.org/abs/2303.08774
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv, (2020). Available: https://arxiv.org/abs/2010.11929
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision, LNCS, pp. 213-229. Springer, Cham (2020).
6. Russakovsky, O., Deng, J., Su, H., Krause, J., Sathe, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, AC., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge (Version 3) (2014).

7. Iyer, J., K T, R., Nersisson, R., Zhuang, Z., Joseph Raj, AN., Refayee, I.: Machine Learning-Based Facial Beauty Prediction and Analysis of Frontal Facial Images Using Facial Landmarks and Traditional Image Descriptors. In: López Rubio, E. Computational Intelligence and Neuroscience, vol. 2021, p. 1-14. Hindawi Limited (2021).
8. Xiao, Q., Wu, Y., Wang, D., Yang, Y-L., Jin, X.: Beauty3DFaceNet: Deep geometry and texture fusion for 3D facial attractiveness prediction. In: Computers & Graphics, vol. 98, pp. 11-18. Elsevier BV (2021).
9. Pramerdorfer, C., & Kampel, M.: Facial Expression Recognition using Convolutional Neural Networks: State of the Art (Version 1) (2016).
10. Liang L, Lin L, Jin L, Xie D, Li M.: SCUT-FBP5500: A Diverse Benchmark Dataset for Multi-Paradigm Facial Beauty Prediction (Version 1) (2018).
11. He, K., Zhang, X., Ren S., Sun J.: Deep Residual Learning for Image Recognition (Version 1) (2015).
12. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition (Version 6) (2014).
13. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles A., Jégou, H.: Training data-efficient image transformers & distillation through attention (Version 2) (2020).