REAL-TIME FACIAL ATTRIBUTE CLASSIFICATION USING THE *Transformer* ARCHITECTURE

A PREPRINT

Zsombor Szenyán
Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics
zsomborszenyan@edu.bme.hu

March 7, 2024

Keywords Real-time · Facial Attribute Recognition · Transformer · Deep Learning

1 Related Work

Facial attribute recognition, a vital task in computer vision, has undergone significant advancements with the advent of machine learning techniques. In recent years, transformer models have gained prominence in various natural language processing (NLP) tasks and have been increasingly applied to computer vision tasks, including facial attribute recognition. In this section, we review the related work on facial attribute recognition with a specific focus on the application of transformer models.

1.1 Earlier Approaches to Facial Attribute Recognition using Deep Learning

The advent of deep learning revolutionized facial attribute recognition by enabling end-to-end learning of feature representations directly from raw data. Convolutional Neural Networks (CNNs) emerged as the backbone of many state-of-the-art systems, leveraging their ability to automatically learn hierarchical features.

Hand and Chellappa [2016] have developed a multi-task deep convolutional neural network (MCNN) for attribute classification. The proposed architecture, along with an auxiliary network (AUX), significantly improves attribute classification accuracy compared to traditional methods. Their method achieved state-of-the-art performance on various attributes from the CelebA and LFWA datasets, with some attributes showing up to a 15% improvement over other methods. The MCNN architecture significantly reduces the number of parameters and training time required for attribute classification compared to independent CNNs, making it more efficient. The learned relationships among attributes in the auxiliary network provide insights into the correlations between different attributes, contributing to a better understanding of the underlying data.

Günther et al. [2016] have demonstrated that the application of data augmentation techniques, including random scaling, rotation, shifting, blurring, and horizontal flipping, not only does not compromise performance but also yields significant benefits. Their findings underscore the importance of leveraging data augmentation as a powerful strategy to enhance model robustness and performance in various tasks. By introducing variations in the training data through augmentation, models can learn more generalized features and exhibit improved performance across different scenarios.

Han et al. [2017] present a novel approach to heterogeneous face attribute estimation using Deep Multi-Task Learning (DMTL) with convolutional neural networks (CNNs). Unlike previous methods that either focused on estimating a single attribute or used separate models for each attribute without considering attribute correlation and heterogeneity, the proposed DMTL approach addresses these issues explicitly. The DMTL framework consists of shared feature learning for all attributes followed by category-specific feature learning for heterogeneous attribute categories. To handle attribute heterogeneity, the paper categorizes attributes into nominal vs. ordinal and holistic vs. local. Nominal attributes, such as race, are handled using classification schemes with cross-entropy loss, while ordinal attributes, such as age, are handled using regression schemes with Euclidean loss. Additionally, attributes are categorized as holistic or

local based on whether they describe characteristics of the whole face or local facial components, respectively. The proposed DMTL approach outperforms state-of-the-art methods in face attribute estimation, as demonstrated through experiments on various benchmark datasets. The approach not only achieves high accuracy but also demonstrates excellent generalization ability, particularly in cross-database testing scenarios.

1.2 Transforming Image Recognition: A Comparative Review of Transformer Networks Versus CNNs

The transformer architecture, initially proposed for natural language processing (NLP) tasks, has recently been applied to computer vision tasks, including image recognition and object detection, however the application of transformer models to facial attribute recognition has been relatively limited. In this section we review the recent work on the application of transformer models to computer vision tasks, with one example of their application in the domain.

One notable approach is the Vision Transformer (ViT) by Dosovitskiy et al. [2020], which treats image patches as tokens (similar to words in NLP) and processes them using a standard Transformer architecture. ViT achieves impressive results when pre-trained on large datasets and transferred to various image recognition benchmarks, surpassing state-of-the-art CNN-based models while requiring fewer computational resources. In contrast to CNNs, ViT exhibits less image-specific inductive bias, with only the MLP layers being local, while the self-attention layers are global. Positional information is preserved through the addition of learnable position embeddings.

The Swin Transformer by Liu et al. [2021] addresses the challenges of the ViTs by proposing a hierarchical Transformer architecture with shifted windows. This design enables modeling at various scales while maintaining linear computational complexity with respect to image size, while enhancing modeling power without sacrificing computational efficiency.

It has been demonstrated by Liu et al. [2022] that the transformer architecture can be effectively applied to facial attribute recognition tasks. Inspired by the visualization of feature attention map of different attributes, they naturally group attributes with similar attention regions into the same category. The proposed TransFa model utilizing the Swin Transformer architecture achieves state-of-the-art performance on the CelebA dataset, outperforming previous methods.

One drawback of the transformer architecture is its computational complexity, which is higher than that of CNNs. However, Li et al. [2022] have proposed the EfficientFormer, a lightweight transformer architecture that achieves competitive performance with state-of-the-art CNNs while being more efficient in terms of computational resources. With an inference speed lower than the frametime of most displays, makes this model suitable for real-time applications. Li et al. [2023] later introduced the second version of the EfficientFormer, which further improves the performance of the model while maintaining its efficiency. They achieved this by giving the multi-head self attention mechanism (MHSA) an input computed by several local convolutional layers, which allows the model to capture local features more effectively and reducing the number of parameters given to the MHSA. Furthermore they improve on the MHSA by downsampling the input and interpolating the output, which allows the model to capture global features more efficiently.

References

Emily M. Hand and Rama Chellappa. Attributes for improved attributes: A multi-task network for attribute classification. *CoRR*, abs/1604.07360, 2016. URL `http://arxiv.org/abs/1604.07360`.

Manuel Günther, Andras Rozsa, and Terrance E. Boult. AFFACT - alignment free facial attribute classification technique. *CoRR*, abs/1611.06158, 2016. URL `http://arxiv.org/abs/1611.06158`.

Hu Han, Anil K. Jain, Shiguang Shan, and Xilin Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *CoRR*, abs/1706.00906, 2017. URL `http://arxiv.org/abs/1706.00906`.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. URL `https://arxiv.org/abs/2010.11929`.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *CoRR*, abs/2103.14030, 2021. URL `https://arxiv.org/abs/2103.14030`.

Decheng Liu, Weijie He, Chunlei Peng, Nannan Wang, Jie Li, and Xinbo Gao. Transfa: Transformer-based representation for face attribute evaluation, 2022.

Yanyu Li, Geng Yuan, Yang Wen, Ju Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang, and Jian Ren. Efficientformer: Vision transformers at mobilenet speed. *Advances in Neural Information Processing Systems*, 35: 12934–12949, 2022.

Yanyu Li, Ju Hu, Yang Wen, Georgios Evangelidis, Kamyar Salahi, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Rethinking vision transformers for mobilenet size and speed. In *Proceedings of the IEEE international conference on computer vision*, 2023.