Unveiling Facial Attributes with Convolutional Neural Networks

Lee and Sike

---

01 Introduction

Facial attribute recognition is an important problem in the field of deep learning and more broadly, machine intelligence. It has far reaching applications in human computer interaction, security and research.
In this project, we build and train neural networks on the classic CelebA dataset to perform the task of facial attribute recognition. First, we build from scratch, train and tune a convolutional neural network to solve the task. Then, we built a second model which uses transfer learning to leverage Chollet's Xception model. Our aim is to gain the pedagogical benefits of hands-on experience and achieve a fine culmination to this semester's foray into introductory deep learning.

02 Dataset

We use the classic CelebA dataset **[1]** for this project. It was collected by the Multimedia Laboratory at the Chinese University of Hong Kong. The dataset is well-documented, well-maintained and widely used in research, being featured in over 2_000 papers over the last five years **[1]**. It holds 202_599 unique images where each image is labeled as positive or negative for each of 40 facial attributes.
Performing exploratory data analysis on the dataset revealed several and severe class imbalances. In particular, there are a lot more images labeled as negative than positive for most of the facial attributes. This will be important to keep in mind during evaluation.

03 Previous Solutions

Previous solutions in image-to-image translation and face attribute prediction have utilized advanced deep learning techniques to address inherent challenges. Hao Tang et al. proposed Attention-Guided Generative Adversarial Networks (AGGAN) to improve image quality by focusing on discriminative image parts and reducing unwanted changes, leveraging attention mechanisms within the network. Ziwei Liu et al. developed a deep learning framework with cascaded CNNs (LNet and ANet) for predicting face attributes in varied conditions, achieving superior performance by pre-training on different datasets and demonstrating effective face localization with image-level annotations. Louay Hazami et al. introduced efficient modifications to Very Deep VAEs (VDVAE) to enhance convergence speed, reduce memory usage, and improve training stability, achieving competitive performance on image datasets. These advancements highlight the importance of specialized network architectures and training strategies in overcoming the limitations of previous models.

04 Proposed Methods

4.0 Preprocessing

We start our project by loading in our images, and converting them to TFRecords (a binary file format which takes up less memory and can be read in faster). Then we visualize these images, searching the

labels for null values of which we found none, and run some exploratory data analysis to investigate class distribution on the training and test sets. We adhere to the 80-10-10 split recommended by the authors of the dataset, with images 1-162770 for training, 162771-182637 for validation, and 182638-202599 for testing.

4.1 Convolutional Neural Network

Our first model was a convolutional neural network– which uses layers of convolutional filters to automatically learn spatial hierarchies of features. From the author's recommended splits, we take out 40_000 files for training, 10_000 files for validation and 10_000 files for testing. Then we normalize them to the 0:1 range.

Our CNN architecture consisted of 1_711_624 (6.53 MB) trainable parameters, with ten conv2D layers, five maxpool2D layers, and three dense layers (including the output layer). It included no drop-out layers because training on a massive dataset let us relax regularization. Configured with early stopping set at min_delta 0.01 and a patience number of 5, it trained for 40 epochs.

4.2 Transfer Learning

Our next attempt was transfer learning– which involves taking a model pre-trained on a large dataset and fine-tuning it to a specific, perhaps smaller, dataset. This approach allows us to take advantage of learned features from the original dataset, improving the efficiency and effectiveness of the model on the new task.

We utilized the Xception model, pre-trained on the ImageNet dataset, as our base model. This model is particularly suitable for image-based tasks due to its deep architecture and efficient use of model parameters. We started by loading the Xception model without its top layer and freezing its weights to retain learned features. The input shape was set to (150, 150, 3). Then a new top layer was added to adapt the model to our specific task, consisting of a global average pooling layer followed by a dropout layer for regularization, and a dense layer with 40 outputs using the sigmoid activation function for multi-label classification.

The initial training process involved 20 epochs, where we monitored the performance and adjusted the model parameters as needed. Subsequently, fine-tuning was conducted by unfreezing the base model and continuing training with a very low learning rate to make subtle adjustments without overriding the pre-learned features.

Both models were compiled with binary cross entropy loss, and we utilized both custom metrics for a comprehensive evaluation during training. To improve data handling efficiency during training, we employed batching and prefetching techniques.

05 Evaluation Methods

In the evaluation of our multi-label classification models, classic metrics we calculated were the hamming loss, hamming score, recall, precision, F1 measure.

5.1 Custom Metrics for Evaluation

Given the multi-label nature of our classification task, traditional accuracy metrics were insufficient. Instead, we implemented two custom metrics:

- CustomAccuracy: Tracks the accuracy across all labels by considering true positives, true negatives, false positives, and false negatives, offering insights into both correct identifications and rejections.
- MultiLabelAccuracy: Computes the Jaccard index (IoU) for each sample, providing a measure of overlap between predicted and actual labels, acknowledging partial correctness in predictions. By evaluating the intersection and union, it inherently normalizes the impact of label frequency, making for a more robust measure of performance across all labels.

We also explored different thresholds for each label to maximize the F1 score, a critical step given the imbalance across labels.

6.0 Results and Discussion

- Hamming Loss: The model achieved a hamming loss of 0.11176, indicating that approximately 11.18% of the total labels per instance were incorrect. In the context of multi-label classification, a lower hamming loss suggests fewer misclassifications and thus a better model performance.
- Recall: The recall score reached 0.80475, highlighting the model's strong capability to capture relevant labels. A higher recall indicates fewer false negatives, meaning the model was able to identify a large majority of the relevant labels across the dataset.
- Precision: The precision achieved was 0.67229, which suggests that approximately 67.23% of the labels predicted by the model as positive were correct. While this is a strong score, especially in a complex multi-label setting, the disparity between precision and recall suggests that the model may still be predicting more false positives than ideal.
- F1 Measure: With an F1 score of 0.71656, the model shows a balanced capacity in terms of precision and recall. This score is critical in scenarios where a trade-off between recall and precision is essential, indicating robustness in the model's predictive accuracy and reliability.
- Hamming Score: Finally, the hamming score (IoU) of 0.57828 demonstrates the model's effectiveness in overlapping its predictions with the true labels. This score is significant for evaluating how well the predictions align with the actual labels on an average basis, emphasizing the model's precision in handling multi-label data.

Interpretation and Implications

The evaluation metrics suggest that the model performs robustly across multiple fronts, particularly in terms of recall and the overall F1 score. However, the precision score indicates potential overfitting to

certain labels, leading to a higher number of false positives. This is a common challenge in multi-label classification, where the balance between capturing as many relevant labels as possible (high recall) and ensuring that the labels predicted are correct (high precision) can be delicate.

The hamming loss and score provide additional context to the model's performance, with a relatively low hamming loss pointing to fewer misclassifications and a moderate hamming score reflecting a reasonable overlap between the predicted and actual labels.

Conclusion

Overall, the metrics demonstrate the model's capacity to effectively handle a multi-label classification task, offering a detailed insight into its strengths and areas for improvement. Future work could focus on optimizing the model further to reduce false positives, potentially through more targeted data preprocessing, enhanced feature engineering, or experimenting with different thresholds for classification to find an optimal balance between recall and precision.

References

1. Project Github Repository
2. Large-scale CelebFaces Attributes (CelebA) Dataset | The Chinese University of Hong-Kong
3. CelebA Dataset Usage | Papers With Code
4. Transfer learning and Fine-tuning | Keras
5. Xception: Deep Learning with Depthwise Separable Convolutions | Francois Chollet
6. Evaluating Multi-label Classifiers Towards Data Science | Aniruddha Karajgi
7. Metrics for Multilabel Classification | Mustafa Murat ARAT
8. Hao Tang, Dan Xu, Nicu Sebe, and Yan Yan. "Attention-Guided Generative Adversarial Networks for Unsupervised Image-to-Image Translation." Department of Information Engineering and Computer Science, University of Trento, Italy; Department of Engineering Science, University of Oxford, UK; Department of Computer Science, Texas State University, USA. Available at: https://github.com/Ha0Tang/AttentionGAN.
9. Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. "Deep Learning Face Attributes in the Wild." Multimedia Laboratory, The Chinese University of Hong Kong. International Conference on Computer Vision (ICCV) 2015.
10. Louay Hazami, Rayhane Mama, and Ragavan Thurairatnam. "Efficient-VDVAE: Less is More." 25 March 2022. Available at: https://github.com/Rayhane-mamah/Efficient-VDVAE.
11.