Deep Learning Project

Convolution Neural Network: BEYOND HUMAN EYES

Case Study: "*Transgender and Gender Classification*"

Authors:
Allan Nhapulo (m20201091)
Geraldo Timbe (m20200603)
Lyinder Swale (m20201009)
Manuel Carreiras (m20200500)
Venâncio Munhangane (m20200579)

Table of Content

Abstract:

Face is the most dominant part of human body and a lot of information can be extracted from facial features. In this study deep learning models are used for classifying gender using facial images of male and female and applying the same model to classify the born gender of people who have gone through medical procedures to change their appearances and sometimes gender. The images were trained by several models. The performance of the models when tested on the images of subjects without alteration scored above 92%. The same models when tested on the transgender dataset were not able to reach the same success. The best score was 25% which indicates that the model is not able to identify transgenders born gender.

Keywords: *Gender Classification*, *Image Recognition, Transgender, Deep Learning, Convolutional neural Networks, Transfer Learning*

1. Introduction

Face is the most dominant part of human body and a lot of information can be extracted from facial features. Human gender classification is one of the fundamental fields which has recently gained a lot of traction in research communities as well as industries due to its substantial role in notable number of real-world applications such as social security, borders control, passport control and others [1].

With advancement in the medical field, people are able to alter their facial appearances and typical gender features using medical procedures such as cosmetic surgeries and hormonal therapies.

In recent times gender recognition and classification tasks have been employed using face features and gender attributes by applying deep learning models enabling various techniques in computer vision to increase accuracy and efficiency. One of these techniques is Convolutional Neural Network (CNN) which utilizes convolutional, poling and fully connected layers on image classification problems.

This study focuses on classifying gender using facial images of male and female and applying the same model to classify the born gender of people who have gone through medical procedures to change their appearances and sometimes gender.

2. Materials and Methodology

This section describes all the procedures and the materials that were used to conduct the study in order to build the best model for our problem.

2.1. Dataset Description and Partition

The data (images) used in this work were collected from different sources. The first and main source of data is IMDB which contains the total of 58691 images 29,607 being male and the rest (29,084) female images who haven't done any facial transformation medical or cosmetic procedures. From the dataset, 43009 images containing both genders were used for training and 11,682 for validation of the model while 4000 images were used for testing.

The second dataset was extracted from several social media channels i.e., YouTube, Instagram etc. The dataset contains the total of 1320 images of transgender males (660) and females (660) after hormonal and surgical transformations. The images are labelled based on their born gender, for example a transgender who was born female and transitioned to male is labelled as female. This dataset is used for further testing of the model.

Figure 1: Male and Female

Figure 2: Transgender Male and Female

2.2. Image Pre-Processing

All the images were first pre-processed before being passed to the model. Different pre-processing techniques were performed on the images; All the images were **RESCALED** by **1/255** and then augmented. The augmentation was performed by using **FLOW_FROM_DIRECTORY** package from the **KERAS** library where images were **ROTATED BY 40 DEGREES**, **WIDTH SHIFT**, **HEIGHT SHIFT**, **SHEAR**, **ZOOM** and **HORIZONTAL SHIFT** all by **0.2.**

2.3. The Model

A sequential base model tailored to our problem was created with 7 layers. The first layer in the model is a 2-dimensional convolutional layer with 32 output filters, each with a kernel size of 3x3, and relu as the activation function.

For the second layer, a 2-dimensional max pooling layer with a pooling size of 2x2 was added in order to pull and reduce the dimensionality of the data.

The second convolutional layer (third layer of the model) was added with the same output filters and kernel size as the first (32 and 3x3 respectively). The layer is then followed by another MaxPool2D layer with same dimensions as the previous one.

We then Flatten the output from the convolutional layer and pass it to a Dense layer with 128 nodes and relu as an activation function. This dense layer is then followed by the output layer of the network which contains 2 nodes, one for male and one for female. We then use SoftMax activation function on our output so that the output for each sample is a probability distribution over the outputs of male and female.

2.4. Model Optimization

Adam and MRSProp were used as optimizers with their default learning rate. In order to improve the model and reduce overfitting, several optimization techniques were applied namely, *Dropout*, *Early Stopping* and *Image Augmentation*.

- ***Dropout***

Dropout was applied to reduce overfitting whereby randomly selected neurons were omitted during the training process. This process was applied using two different dropout rates in order to find the best percentage of neurons to omit. On the first dropout, half of the neurons were dropped by passing 0.5 as the rate and on the second one 0.3.

- ***Early Stopping***

The model was initially set to 20 epochs and the early stopping technique was applied in order stop whenever the model runs 3 epochs in a row without improving its accuracy.

- ***Image Augmentation***

Deep learning models perform best with more data which allows the models to learn patterns and/or functions of the data. To take advantage of this feature, we artificially expanded the training dataset by using image augmentation technique. Image augmentation was done by rotating the images by 40%, width and height shifting both by 0.2, zooming the images, shearing them and horizontally flipping them all by 0.2.

2.5. Grid Search

In order to tune the model, different combinations of dense layers, convolution layers, dropout values and optimizers were tested. Using tensorboard it was possible to view which combination provides better results and the right epoch to stop training the model to avoid overfitting.
The Grid of combinations: dense_layers = [0, 1, 2], layers_sizes= [32, 64, 128], conv_layers = [1,2,3], dropout_rate = [0.3, 0.5], optimizer_values = ["adam", 'rmsprop'] resulted on 108 models.

2.6. Transfer Learning

The developed model was then compared against the top four models in the field of image recognition namely Xception, VGG16, VGG19 and ResNet50. Although the models were not built for face recognition, they can produce good results on gender classification [2]. In addition to that, the model was also compared with the best model in the face recognition field, VGGFace2 (Sonet50) [3]. This comparison was performed by transfer learning technique whereby the models' original architecture, parameters and weights are kept the same [4] while the last layer is adapted to our problem. Due to high computational cost, all the transferred learning models were trained up to 5 epochs.

2.7. Training Process in CNN with Keras

On this chapter will be visualized how our layer structure processes the data in terms of visualization of each intermediate activation, which consists of displaying the feature maps that are output by the convolution and pooling layers.
Figure 5 shows the process above and is possible to verify that, the first layer is maybe retaining the full shape of the face, although there are some filters that are not activated and are left blank, but it is retained almost all the information presented in the initial picture.
As we go deeper in the layers, the activations become increasingly abstract and less visually interpretable. They begin to encode higher-level features like eyes, lips, and noise.

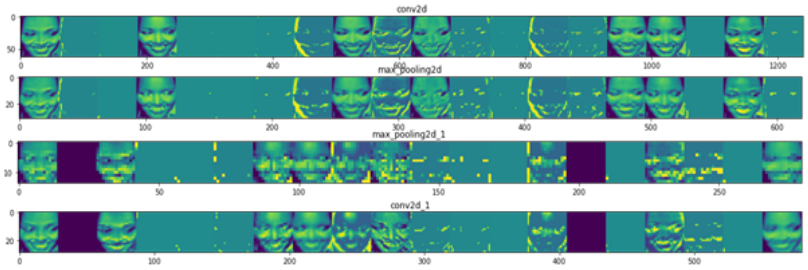

Figure 3: Visualizing Every Channel in Every Intermediate Activation

3. Results and Discussion

As documented in the prior chapter, the images were trained by four different models namely, the base model, the second model built on top of the base model with implementation of dropout (0.5), the third model built on top of the second model with implementation of data augmentation. On the fourth model several combinations of parameters were tested using grid search and the best parameters were used. The results of these models are discussed below.

3.1. Base Model

On the base model, the accuracy tends to increase on training set but remained stable on the validation. On the training, the accuracy value starts at 0.89 and ends at 0.99. On other hand the validation started at 0.94 and ended at 0.95.

The loss value tends to decrease on training but for validation starts increasing after the 5th epochs. On the training, the loss value starts at 0.26 and ends at 0.01. On other hand the validation starts at 0.17 and ends at 0.29. In general, the model had good results although, there is overfitting on the training set.

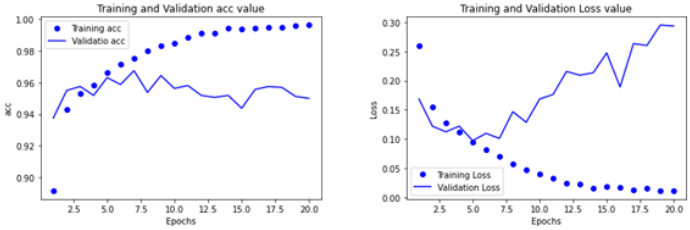

Figure 4: Base Model Performance

The model was tested on the two test sets and the results on the first test dataset gave the scores of 0.96 for both precision and recall. This indicates that male and female are being well classified. In addition, the model was applied on the transgenders test dataset and the results were deplorable for both precision (0.28) and recall (0.32) which indicates that the model is not able to identify transgenders born gender.

The results of the base model when applied on the two test sets are documented on Table 1.

Table 1: Precision, Recall and F1 Score for the Base Model on Test Sets

Dataset	Label	Precision	Recall	F1
Male/Female	Female	0.95	0.96	0.96
	Male	0.96	0.95	0.96
	Average	0.96	0.96	0.96
Transgender	Born Gender Female	0.21	0.18	0.20
	Born Gender Male	0.28	0.32	0.30
	Average	0.25	0.25	0.25

3.2. Base Model with Dropout

To reduce overfitting and increase the score of the base model, dropout was applied with a rate of 0.5. The graphs below on fig 6 show that overfitting has been reduced.

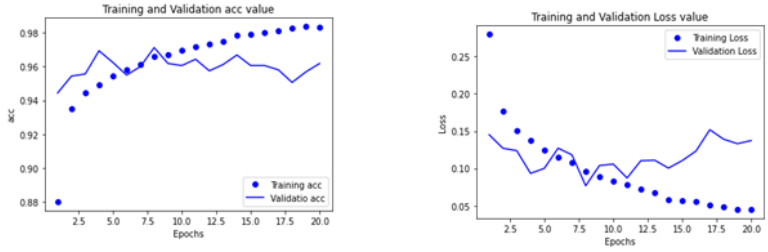

Figure 5: Model Performance with Dropout

The model with dropout had slightly better results from the base model on precision and recall. The precision for female and recall for male improved from 0.95 to 0.96. The model performance on the transgender dataset decreased from 0.25 to 0.23.
The results of the model with dropout when applied on the two test sets are documented on the table below.

Table 2: Precision, Recall and F1 Score for the Base Model with Dropout on Test Sets

Dataset	Label	precision	recall	F1
Test (male/female)	Female	0.96	0.96	0.96
	Male	0.96	0.96	0.96
	Average	0.96	0.96	0.96
Transgender	Born Gender Female	0.18	0.15	0.16
	Born Gender Male	0.28	0.33	0.30
	Average	0.23	0.24	0.23

3.3. Base Model with Dropout and Data Augmentation

When augmentation was applied on top of dropout, overfitting decreased slightly more although the performance decreased.

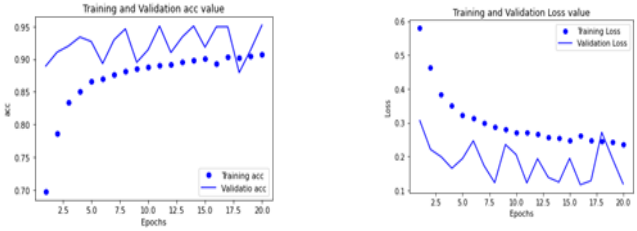

Figure 6: Performance on Model with Dropout and Augmentation

The average precision and recall decreased from 0.96 to 0.94 on the test dataset and 0.23 to 0.19 on transgender dataset. Because of that, it was decided to not implement data augmentation on the model.
The results of the model with dropout and augmentation when applied on the two test sets are documented on the table below.

Table 3: Precision, Recall and F1 Score for the Base Model with Dropout and Augmentation on Test Sets

Dataset	Label	precision	recall	F1
Test (male/female)	Female	0.95	0.93	0.94
	Male	0.93	0.95	0.94
	Average	0.94	0.94	0.94
Transgender	Born Gender Female	0.22	0.24	0.23
	Born Gender Male	0.16	0.15	0.16
	Average	0.19	0.19	0.19

3.4. Grid Search (layers, dropout and optimizer)

The grid of combination resulted to 108 models where the following was concluded:

- The optimizer Adam provides better results than *rmsprop*
- Models with few convolution layers and layers performance better with more nodes.
- The more nodes and convolution layers you add more chance you have to overfitting.
- Although, the best model was with dropout rate of 0.3, most of the model with rate 0.5 had better results.

From grid search, the best model had to be trained with 3 convolution layers, 1 dense layer, 64 nodes on layers sizes, dropout rate of 0.3 and Adam as optimizer on until the 13 epochs.

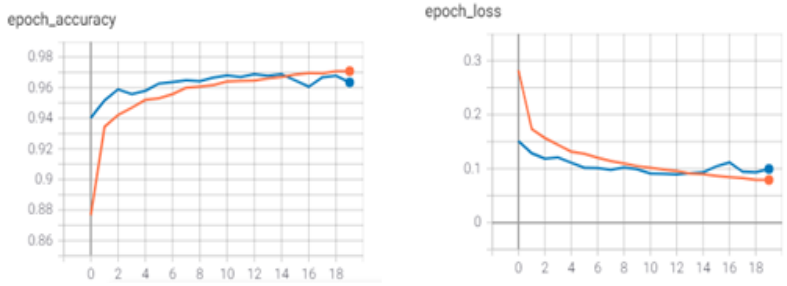

Figure 7: Performance of the Best Grid Search Model

With this model the male misclassified on the test dataset decreased from 71 to 62 (see fig 7 and 8). Also, the female sightly decreased from 88 to 87. On the transgender dataset the behaviour was different. The F1 score decreased from 0.23 to 0.21.

Table 4: Precision, Recall and F1 Score for the TensorBoard Model on Test Sets

Dataset	Label	precision	recall	F1
Test (male/female)	Female	0.97	0.96	0.96
	Male	0.96	0.97	0.96
	Average	0.96	0.96	0.96
Transgender	Born Gender Female	0.13	0.09	0.11
	Born Gender Male	0.28	0.36	0.32
	Average	0.21	0.23	0.21

3.5. Best Model vs Transferred Learning Models

The best performance overall for the first dataset was by VGGFace2 with 0.97 score for both precision, recall and F1. VGGFace2 was trained for with a large-scale face dataset containing 3.31 million images of 9131 subjects, with an average of 362.6 images for each subject [4]. Since facial recognition neural networks like VGGFace2 have already been trained to distinguish human features, the features that they extract may be more useful for determining gender from a photo than the features extracted by a more general neural network like VGG16, VGG19, Xception and ResNet trained to recognize more that 1000 different objects.

On the transgender dataset, the best performance was by the base model without any optimization with the score of 0.25 on precision, recall and F1.

The table below displays the results on the average F1 score, precision and recall for the transferred learning models on the two test datasets.

Table 5: Average Precision, Recall and F1 Score for the Built Model and Transfer Learning Models on Transgender Test Set

Model	Label	precision	recall	F1
Base Model	Test (male/female)	0.96	0.96	0.96
	Transgender (born gender)	0.25	0.25	0.25
TensorBoard Model	Test (male/female)	0.96	0.96	0.96
	Transgender (born gender)	0.21	0.23	0.21
Xception	Test (male/female)	0.94	0.94	0.94
	Transgender (born gender)	0.15	0.15	0.15
VGG19	Test (male/female)	0.92	0.92	0.92
	Transgender (born gender)	0.18	0.19	0.18
VGGFace	Test (male/female)	0.97	0.97	0.97
	Transgender (born gender)	0.14	0.15	0.14
VGG16	Test (male/female)	0.92	0.92	0.92
	Transgender (born gender)	0.20	0.20	0.20
ResNet50	Test (male/female)	0.95	0.95	0.95
	Transgender (born gender)	0.14	0.15	0.14

4. Conclusion

In this study, we trained a deep learning network model to recognize gender based on facial features of male and female subjects without facial alterations and then test it twice. The first test was performed on the subjects without alterations and then on those who had surgical procedures and hormonal therapy to alter born gender features. All the models tested on the images of subjects without alteration performed at the score above 92%. In contrast, the same models when tested on the transgender dataset were not able to reach the same success. The best scores were 25% which indicates that the model is not able to identify transgenders born gender. The reason for the poor performance on the transgender dataset may be attributed to different factors. Based on the visualization of the intermediate activation in CNN (see Figure 5), as the model goes deeper in the layers, the activations begin to encode higher level features such as eyes, lips, beard, hairline and nose. These features are usually the focus during the facial transformation procedures and the results go to show that medical procedures.

The task of classifying the born gender after medical procedures for transformation is still a challenge for deep learning models. For this study, the main limitations where computational costs as deep learning models require a great deal of computational power, lack of transgender images availability and insufficient time. Using other architecture models, pre-processing techniques or fine-tuning the parameters may achieve better results.

5. References

[1] A. Mustafa and K. Meehan, "Gender Classification and Age Prediction Using CNN and ResNet in Real-Time," in *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, Bahrain, 2020.

[2] O. M. Parkhi, A. Vedaldi and A. Zisserman, "Deep Face Recognition," in *British Machine Vision Conference*, 2015.

[3] G. Antipov, M. Baccouche, S. A. Berrani and J. L. Dugelay, "Effective Training of Convolutional Neural Networks fo Face-based Genderand Age Prediction," *Pattern Recognition,* vol. 72, pp. 15-26, 2017.

[4] J. Yosinski, J. Clune, Y. Bengio and H. Lipson, "How Transferable are Features in Deep Neural Networks?," *Advances in Information Processing Systems,* vol. 27, no. Online, pp. 3320-3328, 2014.

[5] Q. Cao, L. Shen, W. Xie, O. M. Parkhi and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *IEEE Conference on Automatic Face and Gesture Recognition (F&G)*, 2018.