DEEP MACHINE LEARNING FOR AGE AND GENDER PREDICTION

George Wanga and Segera Rene Davies
Department of Electrical and Information Engineering, University of Nairobi, Kenya

Abstract
This paper tends to show that by learning feature representations through the employment of convolutional neural networks (CNN), a major increase in performance is obtained on age and gender prediction tasks. An image classifier is built in Matlab. Thousands of facial images are obtained and used to train a convolutional neural network. In this case of deep learning, the CNN essentially constructs abstract features from training image data, which would otherwise have to be handcraft in traditional machine learning model. Feeding in an image in an input, each layer it will perform a series of operations on that data until it outputs a label and classification percentage. Each layer has a different set of abstractions; in the first layers, the network basically teach itself edge detection, then shape detection in the middle layers. They get increasingly more abstract up until the end. The last few layers are the highest-level detectors for the whole object. A lot of computing power and time is spent to train the deep network. The trained network is then used to do predictions of age and gender and can, later after this paper, be integrated with webcam, at home or office to get statistical summary of all guests’ age and gender.

Keywords:
Age, Gender, Prediction, Convolutional Neural Networks, Deep Machine Learning

1. INTRODUCTION

Age and gender, are two key facial qualities which assume a very important role in social collaborations, making age and gender estimation from a face picture a vital task in smart applications [1]. They are the important factors of face analysis. Age and gender recognition have long been recognized as important model for many computer vision applications such as human global interaction, Visual surveillance and passive demographic specific marketing and targeted advertisement and human global interaction. Visual surveillance and passive demographic data collection too. More recently the growing interest in advertising industry for launching specific demographic specific marketing and targeted advertisement and public pages, has attracted the attention of more researchers in the field of computer vision to the field of age and gender recognition.

There are numerous applications for age estimation and here are five of those. First one being access control for example restricting the access of minors to sensible products like alcohol from vending machines go to events with adult content. The second one being human computer interaction (HCI); for example, as much argent estimating the age of a nearby person or advertisement board adopting its offer for young adults or adult people accordingly. The third one being low enforcement for example the automatic scanning video records for suspects for age estimation can help during investigation. Last but obviously not least, the surveillance for example automatic detection of unattended children at usual household places. These can be achieved by performing end-to-end learning and use deep convolutional neural networks to perform image recognition tasks.

The image is taken and faces in the image are detected. The face image is cropped with some padding that is then fed to the neural network. Two lists are produced of which one is used to detect age and the other to detect gender.

2. RELATED WORK

The idea of an algorithm modelled on human neurons has been around at least since 1943, when countries were still fighting world war two. Neurons receive stimuli and fire a signal when those stimuli surpass a given threshold, a process called activation. Today, neural networks have activation functions. Back then neural networks were not deep, limiting what they could do, so people like Marvin Minsky [2] were sceptical of neural networks for a long time, this coincided with an AI winter. Neural networks had a revival in the late 1970s and 1980s, but chips still were not that powerful to let them train well.

Around 2006, a researcher named Geoff Hinton published a paper. He had stacked several shallow neural networks together, to create a deep-belief network. He and his team started to see astonishing results in 2009 [3]. They showed they could train a neural network over three weeks to achieve state of the art voice recognition, equaling the performance of previous algorithms that had taken decades to build.

Adience benchmark [4] is the most recent face image data sets, which was published in 2014. It contains 26,580 photos in 2,284 subjects. All the images contain binary gender tag and a tag from eight different age groups, partitioned into five breaches. The first results of the Adience benchmark realised 45.1% accuracy for age prediction and 77.8% for gender prediction [5]. The same prediction achieves 66.6% accuracy for age prediction and 88.6% for gender prediction on the Ghallagher dataset [6]. Sebastian Lapuschkin, Alexander Binder, Klaus-Robert M and Wojciech Samek, in their paper [7], had to introduce a 3D landmark-based configuration pre-processing phase, to compute versions of the unconstrained face images, this increased gender classification accuracy to 79.3% on the Adience dataset.

Table 1. Age and Gender Prediction results on the Audience benchmark in recent years

<table>
<thead>
<tr>
<th>Year</th>
<th>Age</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td>77.80%</td>
<td>45.10%</td>
</tr>
<tr>
<td>2015</td>
<td>79.30%</td>
<td>-</td>
</tr>
<tr>
<td>2015</td>
<td>86.80%</td>
<td>50.70%</td>
</tr>
<tr>
<td>2016</td>
<td>-</td>
<td>64.00%</td>
</tr>
<tr>
<td>2016</td>
<td>91.00%</td>
<td>61.30%</td>
</tr>
</tbody>
</table>

The current best outcomes for age and gender predictions are 64% and 91% accuracy respectively, from the winner of the ‘ChaLearn Looking at People 2015’ challenge [8]. The model
uses the VGG-16-layer architecture, which has been pertained on the IMDB-WIKI [9] face dataset. The authors point their achievement to large amounts of pre-training data and suitable choice of network architecture. The 91% accuracy achieved by the commercial system unfortunately has no details given about the model architecture in use.

In summary, major factors to improve performance are identified among models as shown in Table.1.

3. CNN MODEL

The CNN is built to copy human brain. This model divides age and gender into two problems with inputs given through an image. The CNN outputs labels around the faces showing the age and gender output. The model was constructed according to Fig.1 [10].

![Fig.1. Flowchart showing how the classification is done](image1.png)

**Input layer**

Takes an input of dimensions

\[
W \times H \times D
\]

**Convolution Layers**

Requires four hyper parameters:

- Number of filters, \( K \)
- Height and width of filters, \( F \)
- Step size for traversing input, \( S \)
- Size of padding, \( P \)

\[
P = (F-1)/2
\]

Its input is from the previous layer as:

\[
W_1 \times H_1 \times D_1
\]

Produces an output of size

\[
W_2 = (W_1 - F + 2P)/(S+1)
\]

\[
H_2 = (H_1 - F + 2P)/(S+1)
\]

\[
D_2 = K
\]

Output = \( W_2 \times H_2 \times D_2 \) = number of neurons

The neurons in this output are connected to the receptive field

\[
\text{Receptive field} = F \times F \times D_1 = \text{Weights per field}
\]

\[
\text{biases} = K
\]

\[
\text{Parameters} = \text{Output} \times \text{Weights per field} \times K
\]

\[
\text{Memory in Bytes} = \text{Parameters} \times 4
\]

**Pooling Layer**

Input is similar to Eq.(1)

Produces an output of size

\[
W_2 = (W_1 - F)/(S+1)
\]

\[
H_2 = (H_1 - F)/(S+1)
\]

where \( D \) - number of neurons, weights, biases, number of parameters and memory size remains as in Eq.(6) - Eq.(11)

Each Age and Gender models are constructed with the flow in Fig.2 below.

![Fig.2. Layers flowchart](image2.png)

It is then designed to use the model in Fig.3 for training purposes.

![Fig.3. Constructed CNN Model Layout](image3.png)

After training and testing, the CNN is ready for use and the designed model is as shown in Fig.5.

---

2000
4. EXPERIMENTAL RESULTS

4.1 TRAINING

The IMDB-WIKI dataset [9], [12] and APPA-REAL Dataset [13], [14] is used to train the model. The model is created to train with resume capability. When inspecting the data, most of the photos don't make sense for the age to which they are set, some are not readable and some do not have faces. Therefore, learning something reliable from such data seems impossible for the CNN. It becomes necessary to clean the data before using it [15]. Matlab load function is used to read label files which points to specific images in the dataset. The loaded images are seen unbalanced as in these plots making it necessary to attempt fixing almost equal widths to avoid skewed training. This is achieved by dropping some images, and the result is as seen in the Fig.5-Fig.9.

The above data has to be trimmed so as to ensure that the network learns features across all ages uniformly, otherwise it would make most age predictions towards the age with the larger number of images, 26 years in this case. The result of trimming is as below

Only 1500 images for every age was considered so as to achieve uniformity and large training data as well. In Fig.6, the target number of images was only achieved for ages between 10 years to 70 years. This means that the network may work satisfactorily within these age limits and this, together with the 1500 which is considered a small number, will have an impact when testing accuracy of the model as will be seen later.

Gender data was also cleaned to a distribution as below:

Just like the age data, this was also trimmed to 15000 images for each gender as shown in Fig.8.
The graph below, Fig. 9, shows the training curves for the first training of the model. More training can be done to achieve better results.

Gender training goes to a validation accuracy of 67% with a rapidly dropping validation loss as seen in Fig. 9. This is due to a sufficiently large training data as mentioned before and also only two possible and very distinct outcome; male and female. The network is seen to learn the difference in features too fast from the start.

Age model had its training characteristics as shown in Fig. 10. The age model performs poorly but as expected. It goes to a validation accuracy of 1.02% with almost constant validation loss. This is on the first training which has only 1500 images for every age between 10 years to 70 years, and less outside this range but between 1 and 100. This is considered a very small data for the deep learning model having 101 possible outputs.

The test accuracy goes to 82% for randomly picked 50 test images, only 10 displayed on Table 2. The first column shows the actual age while the second column shows what our model predicted. Digit 1 here represents male while 2, female. In few cases the prediction is wrong. This indicates that the network requires further training and maybe further fine tuning with technological advancement to have the accuracy as close to unity as possible. The other reason for a good test accuracy is that there was no overfitting, which is usually characterized by the loss starting to rise after falling for a few epochs. The recall rate [16] is 100% in this case since there was cleaning done before images were used.

The trained Age model gave test results in Table 2. Age test accuracy is 7% for 100 test images with only the first 10 displayed in Table 3. This can be accounted for the training data of 1500 is too small as mentioned earlier. Ages from 0-9 years and 71-100 years are not properly covered by the training data, yet the range appear in the test data. The CNN imitates human neurons and thus brain, the same way it would be difficult for us humans to sport age difference between two individuals who are 25 and 26 years old.
old or say 80 and 81 years old, this CNN finds this even more hard too, therefore may miss the exact figure of a person’s age. These percentages are determined on precision thus the closeness to a given age is ignored. This makes the age prediction to even be much lower as predicting an age of 45 instead of actual 46 is considered inaccurate.

<table>
<thead>
<tr>
<th>Table 3. Age model test result</th>
</tr>
</thead>
<tbody>
<tr>
<td>net_Age_accuracy = 0.0700</td>
</tr>
<tr>
<td>net_Age_recall = 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Actual age</th>
<th>Predicted age</th>
</tr>
</thead>
<tbody>
<tr>
<td>ans = 10×2 int8 matrix</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>25</td>
</tr>
<tr>
<td>85</td>
<td>58</td>
</tr>
<tr>
<td>99</td>
<td>64</td>
</tr>
<tr>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>88</td>
<td>29</td>
</tr>
<tr>
<td>87</td>
<td>38</td>
</tr>
<tr>
<td>57</td>
<td>53</td>
</tr>
<tr>
<td>47</td>
<td>47</td>
</tr>
</tbody>
</table>

A lot of training using large datasets is to be done to improve on the model accuracy. This requires longer durations and high machine specifications to achieve high accuracy especially with age training. It is trivial to achieve recall [16] of 100% which appear in this model too. Therefore, recall alone is not enough but one needs to measure accuracy.

The models work together to predict age and gender. From the tests, a computation speed could be calculated as shown in Fig. 11.

\[
\text{computation_speed} = \frac{\text{age\_comp\_time}/\text{nume1}(Y\text{Test\_Age})}{100} + \frac{\text{gender\_comp\_time}/\text{nume1}(Y\text{Test\_Gender})}{100} \\
\text{computation_speed} = 3.8028
\]

Fig. 11. CNN computation speed

This is a speed of 3.8028 seconds per image which looks a little slow. This can be improved by use of machine with higher hardware capabilities than the one in Fig. 12.

<table>
<thead>
<tr>
<th>Processor:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel(R) Core(TM) i3-3120M</td>
</tr>
<tr>
<td>CPU @ 2.50GHz</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Installed Memory (RAM):</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.00 GB (7.89 GB usable)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>System type:</th>
</tr>
</thead>
<tbody>
<tr>
<td>64-bit Operating System, ×64-based processor</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pen and Touch:</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Pen or Touch input is available for this display</td>
</tr>
</tbody>
</table>

Fig. 12. Hardware specifications of the testing machine

Face detection was employed to enhance the prediction accuracy, it can also be improved further by employment of face alignment and introduction of some reasonable padding around the face before passing it for prediction, both of which are not covered in this paper, and any other technological development that may appear after this paper.

4.2.2 Comparison with Other Models:
Some prior accuracy results for apparent age and gender prediction. Amongst them is Eidinger et al. [4] in their paper uses age and gender estimation of unfiltered faces [4]. They get accuracies of 76.1±0.9 on gender and 45.1±2.6 on age. Another one is Levi and Hassner [5] finds 86.8±1.4 on gender and 50.7±5.1 on age. Both use classification models but have larger age group classes like (0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53 and 60-100). These two are improved on here by reducing the classes to unity classes, and further training and adjustments will enable us get better results.

Ideally, Age prediction should be approached as a regression problem, this is because the expected output is a real number as the output. However, regression is challenging with currently available technology here, and even humans cannot accurately predict the exact age based on looking at a person. However, classification is used with unit classes in this case. Because of this reason, a low accuracy on age prediction is expected.

4.3 PREDICTION USING THE NETWORK

The Fig. 13-Fig. 16 below shows the results of age and gender prediction from an image labelled Original Input Image. We analysed the employment of face detection then cropping the face before doing predictions.

Fig. 13. Trained CNN in action, predicting my Age and Gender

This is an image showing the network in action. From left to right, the model is given the first image ‘Original Input Image’. Using MATLAB image processing library, the face is detected and a square drawn on the detected face as shown in the second image ‘Detected face’. The tool is further applied to crop the face along the drawn line to give the third image ‘Cropped Face’. All these pre-processing is done so as to have image almost similar to the ones on which the network was trained on. The cropped image is then passed through the CNN to do the age and gender prediction which give the output with labels as in the fourth image, in this case ‘Age: 28 and Gender: M’.

Fig. 14. Side by side comparison of features extracted by first convolution layer of the Age CNN
The Fig.14 shows feature descriptor which is a representation of the image that simplifies it by extracting useful information and throwing away extraneous information [17]. They are obtained from the output of the first and second convolution layers, during the prediction in Fig.12. These are the features which would otherwise have to be handcraft for a purely machine learning model. The number of these abstract features keeps growing as more convolution layers are added [11]. The CNN is able to extract a variety of features as the image is processed down the network before it finally does the prediction based on these abstract features.

Fig.14. Trained CNN in action, predicting Age and Gender for a generic image from the internet

To test gender diversity, the Fig.15 was picked randomly from the internet. It was passed through the same model as the image of Fig.13 and as a result, a female with a much lower age could be predicted in this case. This shows range in age prediction as well.

Another image showing Will Smith face was used too with results shown in Fig.16. The model still works to produce a satisfactory result as output. More images were tested on which all may not be displayed on this paper.

In this model, the predictions improved for a few samples while becoming worse for some other images. Proper detection and alignment may be necessary if you are mainly operating with non-frontal faces.

Fig.15. Trained CNN in action, predicting Age and Gender for a generic image from the internet

Fig.16. Trained CNN in action, predicting Age and Gender for another generic image from the internet

5. CONCLUSIONS

In this paper, a CNN has been built and trained to predicting age and gender from an input image. Using two combined architecture, both were trained and prediction was done on various test images. Overly, the accuracy of the models is satisfactory, however, it can be improved more by exploitation additional data, defining more data augmentation and improving to higher network architectures. One may also attempt to use a regression model rather than classification for age prediction if enough dataset is obtainable.

REFERENCES


on Automatic Face and Gesture Recognition, pp. 30-36, 2017.


[16] Precision and Recall, Available at: https://en.wikipedia.org/wiki/Precision_and_recall.