Age and Gender Classification using Convolutional Neural Networks

Narita S. Pandhe
Department of Computer Science
University of Georgia
Athens, GA, 30605
narita@uga.edu

Abstract

This paper focuses on gender and age classification based on images. Our work builds on the previous work [1, 2] that has developed efficient architecture for such tasks. We intend to extend their approach in order to improve the results. The primary area of experimentation is to tweak previously published architecture[1] in terms of depth of the network, number of parameters in the network, modifications to parameters of the network. The next facet of the project focuses on coupling the architectures for age and gender classification to take advantage of gender-specific age characteristics inherent to images[2]. This stemmed from the observation that gender classification is an inherently easier task than age classification, due to both fewer number of potential classes and more prominent intra-gender facial variations [2]. By training different age classifiers for each gender, an improvement in age classification is observed.

1 Introduction

Recently, the usage of images over the internet has grown at an exponential rate. This afresh wealth of data has now enabled us to tackle computer vision problems that were previously complex to solve. There exists accurate and efficient face detection and recognition frameworks that leverage convolutional neural networks. Its applications range from suggesting "who to tag" to pedestrian detection. The next major step is to extract the characteristics of the subjects in such images. Following the successful example laid down by face detection and recognition systems, similar gains can be obtained with a simple network architecture for age and gender classification, designed by considering the limited availability of accurately labelled datasets.

Data scarcity is mainly because of the nature of the data that is required. The reason behind this is: in order to have age and gender labels for images, access to the personal information like: date of birth, gender of the subjects is needed, which is a rare and private piece of information. Thus, we must take this into consideration and alter the network architectures and algorithmic approaches to cope with these limitations. These reasons are the primary motivations behind [1, 2] choosing to implement a relatively shallow architecture for age and gender classification using convolutional neural networks.

We test our network on the Adience benchmark - collection of unfiltered face images. The data included in this collection is intended to be as true as possible to the challenges of real-world imaging conditions [9]. Although results by [1] provide a remarkable baseline for deep-learning-based age and gender classification approaches, they leave room for improvement for more elaborate system designs, which are presented in this paper.

2 Related Work

Work has been done for decades in the areas of age and gender classification. Early attempts [3] focussed on identification of manually tuned facial features and calculating ratios between different measurements of these features. The features of interest in this approach included the size of the eyes, mouth, ears, and the distance between them [2]. Most of the early experiments have shown to achieve higher accuracies on constrained images (ideal lighting conditions, visibility). Few of them [4] tried to address the challenges that arise in real-world images. An overview of such methods to gender classification can be found in [5]. In 2000s, [6] used support vector machines for gender prediction on thumbnail images of subjects. These images had very low resolution and the paper proposes to apply SVM directly to image intensities.

All of these papers and their methods tackle either gender classification or age classification/regression, but not both. [1] developed one method and architecture to tackle both age and gender. The authors of [1] address the undeniable reality that real-world images are not perfectly aligned, lit or centered. They train their network on images taken from a wide range of angles, lighting conditions, etc. They also oversample the input images by taking multiple crops of the images so that the classifier considers various regions in the images for classification. Their focus on using deep convolutional neural networks (CNNs), follows a pattern in the computer vision community as CNNs are shown to provide unparalleled performance for other types of image classification problems [2]. [7] showed that deep CNN architectures are both feasible and effective, and [8] showed that increasing the depth of such networks shows better performance. The authors of [1] leveraged these advances to build a powerful network that showed state-of-the-art performance. But, they have proposed a relatively shallow network in order to prevent overfitting the small dataset they were operating on.

In this paper, we build off the work of [1], to develop a system that leverages the inherent relationships between age and gender to link these architectures to improve overall performance.

3 Dataset

The dataset used for training and testing for this project is the Adience benchmark - collection of unfiltered face images. This dataset is published by Face Image Project [1] from the Open University of Israel (OUI). It contains total 26,580 images of 2,284 unique subjects that are collected from Flickr [10]. There are 2 possible gender labels: M, F and 8 possible age ranges: 0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, 60+. Each image is labelled with the person's gender and age-range (out of 8 possible ranges mentioned above). The images are subject to occlusion, blur, reflecting real-world circumstances.

The authors of [1] split the images into 5 folds and then perform a subject-exclusive cross-validation protocol. The reason for such bifurcation is because of the nature of the dataset being used. The dataset contains multiple pictures of the same subjects. Therefore, if the images were simply randomly shuffled and divided, there was a possibility of the same subjects appearing in both training and testing [2]. This would have skewed the results and affected correctness. To avoid such an issue, the bifurcation ensures that all the images of a particular subject appear in the same fold.

From the original dataset we used mostly frontal face images reducing the dataset size to 17,523 images. Table 1 displays the total number of images comprising every fold.

Table 1: Total number of images comprising the dataset

Fold	Total #images
0	3996
1	3619
2	3127
3	3322
4	3468

Fold 0 - Fold 3 were used as train, validation datasets and Fold 4 was used as test data set. Further, each of the Table 1 folds were divided into males, females. As described in Section 5.2, this was necessary for the types of classifiers we were aiming to build. All of the sub-folds coming from the Fold 4 were separated as test data, never to be trained on or validated against. Then the remaining 4 folds(Fold 0 - Fold 3), and their sub-folds, were used for training, and cross-validation. Table 2 shows the gender wise distribution of images in every fold.

Table 2: Total number of images based on gender

Fold	#Males	#Females	Total
0	2047	1949	3996
1	1611	1999	3619
2	1363	1764	3127
3	1502	1820	3322
4	1597	1871	3468

4 CNN for Age and Gender Classification

Collecting labelled training images is a tedious task because it requires access to personal information like age and gender of the subjects in the images. Real world dataset for such problem is relatively small and incomparable with the large size of image classification datasets like ImageNet [11]. Such small collection size can lead to overfitting. Problem worsens when employing deep convolutional neural networks due to their huge numbers of model parameters. Efforts must therefore be taken to safeguard against overfitting.

4.1 Network Architecture

Figure 1: Network architecture proposed by [1, 2]

The network architecture is based off of the work in [1, 2]. As stated, this network is intended to be relatively shallow so as to reduce the risk of overfitting the data as well as the nature of the problem: Age and Gender classification on the Adience dataset, which requires differentiating between 8 age classes and only 2 gender classes. The input to the network is an RGB 3x256x256 image. It is center cropped to 3x227x227. The details of the network:

1. 96 filters of size 3x7x7 pixels are applied to the input with a stride of 4 and 0 padding.

This is followed by a rectified linear operator (ReLU), a Max-Pooling layer taking the maximal value of 3x3 regions with two-pixel strides and a local response normalization layer.

2. Second convolutional layer then processes the 96x28x28 output of the previous layer with 256 filters of size 96x5x5. Followed by ReLU, Max-Pooling layer and LRN

3. In the third layer, 384 filters of size 256x3x3 are convolved with stride 1 and padding 1, followed by a ReLU and Max-Pool.

4. The fully connected layers:

    1. 512 neurons are fully connected to the 256x7x7 output of the third convolutional layer, followed by a ReLU layer and dropout of 0.5, layer .

    2. 512 neurons are fully connected to the 1x512 output of previous fully connected layer followed by a ReLU layer and dropout 0.5, layer.

    3. Third, fully connected layer maps to the final 8 classes for age or 2 classes for gender

The output of the last fully connected layer is input to softmax. Prediction is made by considering the class with the maximal probability for the given test image.

4.2 Training and Testing

Data Augmentation. We take a random crop of 227 × 227 pixels from the 256 × 256 input image and randomly mirror it in each forward-backward training pass.

Cross Validation. As described in Section 3, the dataset was divided into 5 subject exclusive folds, we further divided each of those folds into males, females. This was necessary for the types of classifiers we were aiming to build. All of the sub-folds coming from the 5th original fold were separated as test data, never to be trained on or validated against. Then the remaining 4 folds, and their sub-folds, were used for training, and cross-validation. For example, when training male-age classifier, we would use 3 of the male sub-folds as training dataset and the 4th male sub-fold as validation dataset. This arrangement would rotate within the first 3 sub-folds for every possible assignment of the validation fold. The classifier would be tested against the 5th (previously separated) male sub-fold.

Initialization. The modified fold separation prevented from using the pre-trained weights provided by [1]. The network is trained from scratch. The weights in all layers are initialized with random values from a zero mean Gaussian with standard deviation of 0.01. Target values for training are represented as one-hot vectors corresponding to the ground truth classes.

Network Training. Training is performed using Stochastic Gradient Descent having a batch size of 50. The initial learning rate is 1e−3, reduced to 5e-4 after every 10,000 iterations.

Prediction. Center Crop: We fed the network with the face image, cropped to 227 × 227 around the face center.

5 Experiments

Our method is implemented using the open-source framework: Tensorflow [12]. Training was performed on a laptop with 1070 GPU having 1920 CUDA cores and 8GB of video memory. Training each network required about 5 hours.

5.1 Baseline

The first step was to reproduce the results of [1] and set the baseline. Similar to [1], we trained the network using SGD with learning rate of 1e-3 which decays by a factor of 10 every 10,000 iterations. The batch size was of 50 images. This was successful and the network achieved accuracies that were close to [1] for both age and gender classification. Henceforth, the baseline model is referred as Model #1.

Table 3: Baseline Accuracies

	Gender	**Age**
Benchmark[1]	85.9	45.9
Observed Results	85.89	43.5

5.2 Objectives

Our initial objective was to verify if the proposed network architecture[1] was truly optimal. The authors of [1] claim that any deeper network would lead to overfitting. In an attempt to validate this, we experimented by adding convolutional layers, using Adam optimizer instead of SGD, replacing the fully connected layers. The goal was to improve existing architecture and achieve better results than the observed baseline. These modifications made the system perform the same or, sometimes, worse.

As no clear benefit was achieved after modifying the existing architecture, we tried to experiment with a different approach for constructing these classifiers. According to the observations in [2] gender classification is an easier task as compared to age classification, due to the fewer number of classes to distinguish between and more marked differences exist between genders than between many age groups. As [2] states, there is a possibility of using one of the attributes to better inform the prediction of the other, like gender can be used to predict age better. For example: Receding hairline can be a useful indicator of male's age whereas this isn't the case for females.

Based on the above mentioned observations, we then tried: chaining the architectures for gender and age classification, separating the age classifiers based on genders.

5.3 Chaining Gender and Age Classifiers

As stated, gender can be used to predict the age better, we trained a chained gender - age network that would first classify gender (using the same network architecture as [1]), and then based on the gender classification feed the image into an age classifier trained solely on male or solely on females. Figure 2 depicts the variance in the architecture from [1].

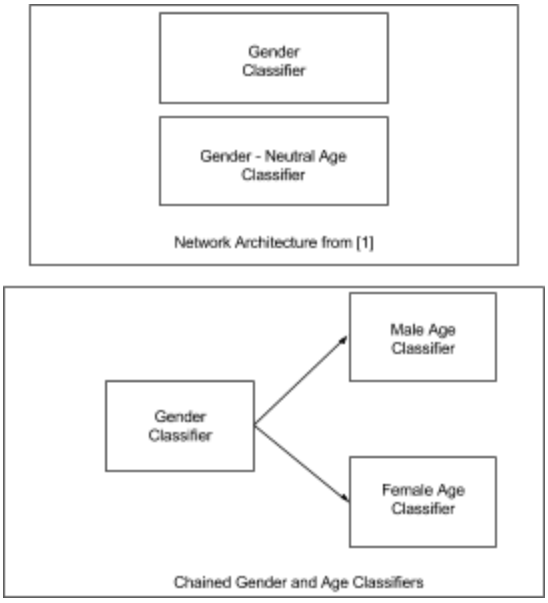

Figure 2: Differences in architecture

5.4 Gender based Age Classifiers

Separating the age classifiers based on genders frees the network from having to learn a gender - neutral concept of age. It was observed that when training a classifier only on male images, accuracy of predicting the age of men increased as compared to training a gender-neutral age classifier and then predicting men's age. Henceforth, in the paper chained gender - age network with separate networks for male age and female age classifier are referred as Model #2.

Table 4: Age-Classification accuracies achieved by gender-neutral age classifier, Model #1

Model #1 : Gender Neutral Age Classifier	**Exact Match**		**One - Off Accuracy**	
Males Total: 1343	440	32.76%	721	53.68%
Females Total: 1636	562	34.35%	955	58.37%
Total: 2979	1002	33.63%	1676	56.26%

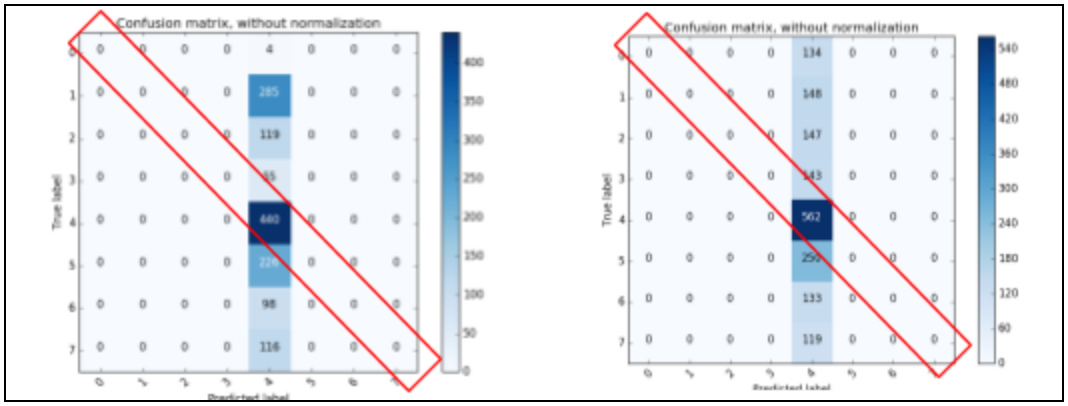

Figure 3: Age-Classification confusion matrix of gender-neutral age classifier, Model #1

From the confusion matrix of Model #1 it can be inferred that the network classifies age correctly for males and females only for age group 4 and fails to classify other age groups. It also highlights that the number of images for age group 4 must be very high and thus the network is biased.

Considering the above observation, we slightly modified the dropouts and included weighted losses in Model #2. Table 5 and Figure 4 depicts the results.

Table 5: Age-Classification accuracies achieved by gender-based age classifiers, Model #2

Model #2 : Gender Based Age Classifier	**Exact Match**		**One - Off Accuracy**	
Males Total: 1343	679	50.55%	1106	82.35%
Females Total: 1636	651	39.79%	1140	69.68%
Total: 2979	1330	44.64%	2246	75.39

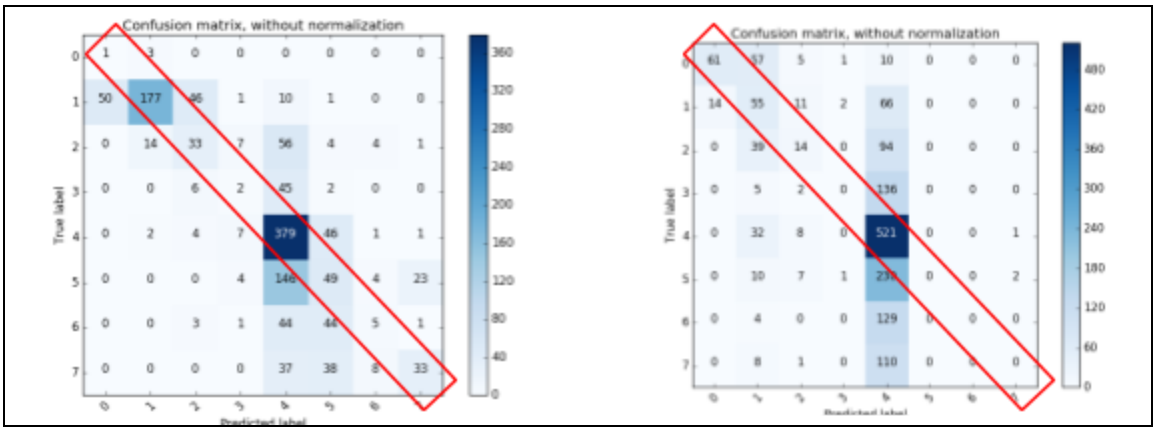

Figure 4: Age-Classification confusion matrix of gender-based age classifiers, Model 2

From Figure 4, it can be concluded, that Model #2 performs certainly better than Model #1 in classifying age, both for males and females respectively. But, the model predicts mostly for age group 4. The bias still seems to exists. On observing the training data closely, we realized that the age distribution is skewed towards younger users for both males and females. Table 6, Table 7 highlights the distribution of images for every age group.

Table 6: Dataset distribution for every age group for males

Male Folds	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	287	180	60	41	929	337	144	69
1	2	93	265	387	343	520	41	128
2	364	224	74	74	229	159	41	128
3	72	35	367	203	413	306	57	49

Table 7: Dataset distribution for every age group for females

Female Folds	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	254	313	156	112	743	219	83	69
1	0	387	497	318	407	240	72	78
2	236	134	401	212	499	120	88	74
3	4	203	130	314	819	221	60	69

According to Table 6, Table 7, it is observed that age group 4 is usually one of the 3 age groups with higher number of images for most of the folds, whereas, for some folds, there are very few images for some age groups. For example: Female Fold 1 has 0 images for age-group 0. To reduce this inconsistency and the dominance of age group 4, we balanced out the the dataset across every age group ensuring maximum subject exclusivity along with weighted loss. We refer to this as Model #3. We added weighted loss, because, in spite of the balanced distribution, the number of images in age group 4 is high. That can be because, these images are downloaded from Flickr and younger users tend to upload considerably more images on social media websites as compared to other age groups. It can also be due to the fact, that age group 4 has the largest interval.

Table 8: Age-Classification accuracies achieved by gender-based age classifiers, Model #3

Model #3 : Gender Based Age Classifier	**Exact Match**		**One - Off Accuracy**	
Males Total: 1343	644	47.95%	1146	85.33%
Females Total: 1636	651	39.79%	1292	78.97
Total: 2979	1295	43.47%	2438	81.83%

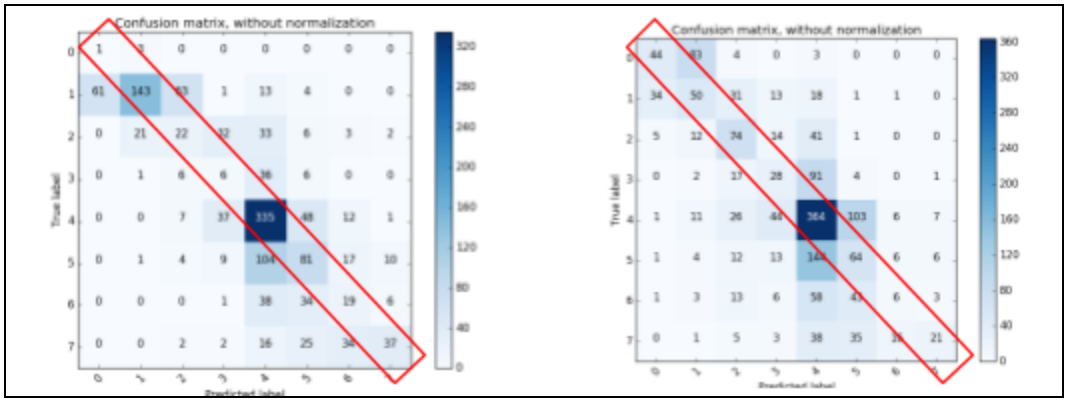

Figure 5: Age-Classification confusion matrix of gender-based age classifiers, Model #3

It can be concluded that, because of distribution of images across all folds and weighted losses, the network is trained with better features of every age group for both males and females. Thus, Model #3, achieves considerable improvement over Model #1 and Model #2. As suggested by [2], we tried modifying the network architecture of Model #3, by replacing one of the fully connected layers with convolutional layer along with the Adam optimizer. With these modifications, we thought, we could maybe use Adam to decrease training time and increase (or at least match) performance levels of Model #3. But, we couldn't find a setting that matched the performance of Model #3.

5 Conclusion

Work has been done to address the problem of age and gender classification, but, most of it has been focussed on constrained images. Such images do not adequately reflect appearance variations common to the real-world images. Authors of [1] build a powerful network that showed state-of-the-art performance on real-world images. Unlike CNN-based image classification problems, [1] advocate the use of a relatively shallow network to prevent overfitting considering the availability of limited labeled data. [2] improves on these results by modifying the approach of constructing the classifiers. Based on observations, [2] states, that one of the attributes can be used to better inform the prediction of the other, like gender can be used to predict age better. In line with this observation, the authors of [2] have chained the architectures for gender and age classification and separated the age classifiers based on genders. This paper further improves by modifying the dropouts used in the network of the gender-based age classifiers along with weighted losses.

The most difficult part of this project was setting up the training data to properly divide it into folds such that they don't overlap and cross-validation. Currently, this approach works with only single face images. We foresee future directions building off of this work to be able to classify gender and age of multiple faces in images, aid face recognition and much more. We plan to apply the proposed method on the IMDB-WIKI – 500k+ face images with age and gender labels. We hope that additional training data will become available for this task which will enable successful techniques from other types of classification with huge datasets to be applied to this area as well.

References

[1] Gil Levi & Tal Hassner Alexander, Age and Gender Classification using Convolutional Neural Networks. 2015

[2] Ari Ekmekji. Convolutional Neural Networks for Age and Gender Classification, Stanford University. 2016

[3] Y. H. Kwon and N. da Vitoria Lobo. Age classification from facial images. In Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on, pages 762–767, Jun 1994.

[4] E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. IEEE Transactions on Information Forensics and Security, 9(12):2170–2179, Dec 2014.

[5] E. Makinen and R. Raisamo. Evaluation of gender classification methods with automatically detected and aligned faces. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(3):541–547, March 2008.

[6] B. Moghaddam and M.-H. Yang. Learning gender with support faces. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(5):707–711, May 2002.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, Advances in Neural Information Processing Systems 25, pages 1106–1114. 2012

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. CoRR, abs/1409.4842, 2014.

[9] The OUI-Adience Face Image Project, http://www.openu.ac.il/home/hassner/Adience/data.html

[10] Flickr, https://www.flickr.com/

[11] ImageNet, http://image-net.org/

[12] Tensorflow, https://www.tensorflow.org/