Age and Gender Estimation using Optimised Deep Networks

Wade Downton
School of Computer Science and Applied Mathematics,
University of the Witwatersrand
Johannesburg, South Africa
wadedownton@gmail.com

Hima Vadapalli
School of Computer Science and Applied Mathematics,
University of the Witwatersrand
Johannesburg, South Africa
hima.vadapalli@wits.ac.za

ABSTRACT

Age and gender estimation plays a fundamental role in intelligent applications such as access control, marketing intelligence, human-computer interaction etc. The advent of deep architectures have paved a way to improve the performance of estimation models, however, there is still a lack of optimized architectures. This paper focuses on the use of convolutional neural networks, and parameter modeling and optimization, and their effect on accuracy and loss term convergence. This paper first makes use of a generalized deep architecture based on literature and looks at ways of optimizing and reducing complexity without loss of accuracy. Different activation functions such as rectified linear unit (ReLU), linear function, exponential linear unit (ELU), hyperbolic tangent and Googles' proposed Swish function were tested along with the use of additional convolutional and fully-connected layers. Experiments resulted in a less complex architecture for gender classification and results were in line with that of benchmark accuracies found in literature, however, the same couldn't be achieved for age estimation. The inability to find a simpler architecture for age estimation is attributed to the complex nature of features that are associated with age than that of gender and also the multi-class classification nature of the age estimation problem.

CCS CONCEPTS

• **Computing methodologies → Computer vision tasks**; **Neural networks**; *Learning paradigms*; *Cross-validation.*

KEYWORDS

Deep Learning, Computer Vision, Convolutional Neural Network, Age and Gender Estimation, Activation Functions, Adience Dataset, Swish Activation Function

ACM Reference Format:
Wade Downton and Hima Vadapalli. 2019. Age and Gender Estimation using Optimised Deep Networks. In *Conference of the South African Institute of Computer Scientists and Information Technologists 2019 (SAICSIT '19), September 17–18, 2019, Skukuza, South Africa.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3351108.3351123

1 INTRODUCTION

Age and gender estimation from facial images is an innovative idea proposed in the field of computer vision and deep learning. The application for an effective model has been realized for its utility in many fields, such as marketing, entertainment, biometrics, human-computer interaction and so on. A model with improved accuracy, reliability and versatility will notably improve the above fields. On the other hand data representations in the form of features provided by deep learning architectures can improve age and gender estimation.

Deep learning models have an intrinsic ability to mimic the way our own eyes and brains are able to perceive real world objects. Convolutional Neural Networks (CNNs) where the connectivity between neurons in the network resembles that of the organization within the animal visual cortex [12] is one such model. This is the most commonly used model for the task of age and gender estimation due to its proven accuracy [9], [4], [7]. Applications of CNNs have found to be incredibly useful in the task of face detection, where models have been optimized in determining and identifying an individual's facial features responsible for successfully identifying a face in an image or video [11], [18], [13]. A notable contribution made by models developed for facial detection is the use of integral images and integral channel features [5] that drastically improve the accuracy of object detection models in the field of computer vision.

This paper will explore the modifications to a generalized CNN architecture found in the literature for the task of age and gender estimation and explores parameter optimization and reducing the complexity without a loss in accuracy. Activation functions, one of the important parameters, is explored from the available domain [14] and their effect is studied. A common problem faced by researchers is converging to an effective model while avoiding over-complexity [6], [7]. A number of methods have been proposed to overcome this problem, including the use of random neuron dropout [6], Residual Networks (ResNets) [20], [21], utilization of shallower networks [9] etc. However, the effect of altering a network's depth and the effectiveness of different activation functions have not been fully explored for the task on hand. Previous research [7] has noted an improvement in network accuracy with few convolutional network layers, and the reasoning for this was attributed to the idea that the task of age and gender estimation is more prone to that of over-fitting than many other computer vision problems. No deeper reasoning was given to this result and hence more exploration is required.

SAICSIT '19, September 17–18, 2019, Skukuza, South Africa

ACM ISBN 978-1-4503-7265-7/19/09…$15.00
https://doi.org/10.1145/3351108.3351123

2 BACKGROUND AND RELATED WORK

2.1 Overview

There have been a number of approaches proposed to overcome the problem of over-fitting for the task of age and/or gender estimation. Research conducted in [17] found that a CNN with a deeper architecture showed improved accuracy when transitioned from fully connected to sparsely connected. A higher valued dropout rate was also found to be beneficial. In this work, results presented in [2] were also utilized, where it was found that by analyzing the probability distribution of a dataset represented by a large sparsely connected network, an optimal network topology can be constructed iteratively by analyzing the correlation between activations of the last implemented layer and by clustering neurons with the greatest correlated outputs. Not only does this approach reduce the over-fitting of the network but as the network is sparsely connected, the computational complexity does not exponentially increase with the addition of layers.

In [7] and [9], it was shown that the problem of over-fitting and the overall complexity of the network increases as more layers are added to a fully connected network. Use of smaller network architectures and a lower dropout rate was utilized than found in [9], [17] where the network was deeper (6 layers as found in [9] and 22 in [17]). However, the quantifying effect of changing network depth was not thoroughly explored for fully connected networks [7], [9]. Also, there is a lack of research on analyzing the effect of changing network activation functions and if they help to overcome the problem of over-fitting for the task of age and gender estimation where over-fitting is prominent.

2.2 Background Techniques

2.2.1 Existing Models. The most common type of deep learning model used for age and gender estimation is CNN. A CNN is a multi-layered neural network trained using a version of the back-propagation algorithm. The goal of a CNN in the context of computer vision is to recognize patterns in pixel images whilst minimizing preprocessing. This means that a CNN trained on a dataset that contains complex images, such as images with low resolution, poor lighting, and varying facial angles, is still able to perform with a high level of accuracy. The last fully connected layer is fed to a softmax layer where each output class is designated a probability. During the training process, an optimizer uses the output from the network to update the weights using back-propagation. Once trained, predictions are made by determining the class with the greatest probability.

Another commonly used model is a Support Vector Machine (SVM) [6]. An SVM views inputs as points in a space in any finite dimension and is used to classify individual points into their corresponding classes under supervised conditions. An SVM can be considered equivalent to a single-layer neural network as noted in [6]. Training and testing of the SVM model in [6] were performed using 5-fold cross-validation with pre-selected splits to eliminate cases where images from the dataset appear in both the training and testing sets in the same fold. The motivation for using a SVM over the more commonly used CNN is its effectiveness in combating over-fitting with the use of a random neuron dropout operation. The model used in [6] is a linear dropout-SVM for gender classification(male and female), and a one-vs-one linear-SVM arrangement is used for the multi-label age classification. In a one-vs-one linear-SVM arrangement, separate classifiers were trained for a different pair of labels.

Lastly, Residual Networks of Residual Networks (RoR) is another popular model used for gender and age classification problem. RoR is an architecture which uses residual mappings on an existing Residual Network (ResNet). A ResNet can be seen as a flow network where a collection of nodes are each connected by differing weights when residual edges are introduced into the flow network between nodes that can admit more flow, then a path flow between nodes is increased and hence the network becomes more optimized. In [20], researchers proposed a CNN model that leverages this RoR architecture. This will allow for a greater exhibition of optimization on the CNN that is leveraging the RoR rather than that of standard CNN models. RoRs are also hypothesized to be more easily optimized than standard ResNets of the same complexity due to the addition of residual mappings and identity shortcuts. These identity shortcuts allow for the preservation of the quality of information as it flows through the network, meaning greater levels of abstraction from the RoR do not affect the quality of the input deeper into the network.

2.2.2 Training Techniques. Training techniques often applied to CNNs are versions of the back-propagation algorithm. Here, the goal is to optimize the accuracy of the model by minimizing the error of the model using the gradient descent algorithm. This is done by calculating the gradient of the loss function used in the model. This training technique is effective because the error generated is fed back into the CNN and this thus is distributed back through the network layers in order for the model parameters to be trained and minimize the error. The loss function used in [10], [6] is a softmax loss layer where the multinomial logistic loss of the softmax layer is generated.

A softmax layer makes use of the softmax function which outputs a probability of the input into the classifiers specified in the model (where 34 are used in [10]), the softmax output is differentiable in order to train by gradient descent algorithm for the training of the CNN. A version of the gradient descent algorithm used to train the outputs of the loss generated by the softmax layer is the Stochastic Gradient Descent method used in [6], [7] which is calculated on single samples in the dataset. This algorithm finds the minimum by randomly selecting a point in the probability distribution of the sample output and finds the minimum using the steepest gradient algorithm.

2.3 Challenges in Prior Research

The most common problem faced by researchers in this field is the problem of over-fitting the model. Many methods were proposed in order to combat over-fitting, such as random neuron dropout in [6], [16], [17] where improvements to accuracy were found. Another method that proved successful was the reduction of the dimensionality of the feature space as demonstrated in [6],[1]. Shallow CNNs utilized in [7],[9] were used as over-fitting was found to be less of an obstacle for shallower networks. For a deeper network proposed in [17], transitioning from the fully connected architecture to a sparsely connected one, and therefore making use of correlations between activations of nodes and therefore clustering nodes with the greatest correlation outputs was found to reduce the effect of over-fitting.

Developing a complex model whereby the increase in complexity outweighed the increase in accuracy was also a challenge faced by researchers. This was demonstrated in [7], [20] where a deeper CNN was found to not only decrease the models' performance but also cause the model to be complex with regard to both time and space. Certain techniques typically used to overcome the problem of over-fitting were found to also be effective in combating this problem. An example of this was given in [17] where the transition from a fully connected to sparsely connected architecture reduced the complexity with the addition of layers.

2.4 Motivation for Current Work

A number of attempts were made to develop models that can accurately estimate age and gender while combating over-fitting. Dimensionality reduction [1] and dropout techniques have proven successful [6], [16], however, the problem of over-fitting still exists in these instances. One might think that complex and deeper neural networks would be the solution to creating the perfect age and gender estimator, however, as seen in [7] and [20], increasing the depth of the network does not only make the model more complex but either maintains or decreases the accuracy of the original model. Image preprocessing with the use of an effective face detector and feature extractor have proven to greatly increase the accuracy of the estimations as seen in [10], [1], [16], [8]. Data augmentation is another key factor in improving the model accuracy, facial alignment of features was proven to improve accuracy in [10], [6]. With regard to quantifying the effect particular alterations have on overfitting the network, little research has been done on CNNs that make use of the fully connected architecture with an alteration in the network depth and activation function used. Therefore further investigation is required on understanding what contributes to over-fitting of models for this particular task in order to overcome the problem.

3 RESEARCH METHODOLOGY

3.1 Research Goals

The primary goal of this paper is to determine the effect different activation functions will have with respect to the model accuracy and convergence, and whether $f(x) = x \cdot sigmoid(\beta x)$ with trained β is the better activation function to use as discovered in [14] within the domain of activation functions. The second goal is to determine the effect a change in the number of convolutional layers or fully-connected layers has on the accuracy and convergence of a model, and whether a deeper model is more greatly affected by over-fitting. The final goal of this study is to determine whether it is possible to benchmark the results as in [6] and [7] with notably fewer network variables.

3.2 Dataset

The Adience dataset proposed in [6] contains age and gender labels for samples captured in the real world. This means that the dataset presents challenging qualities found in real-world images such as low-quality, poor lighting, differing facial angles, and alignment etc. The database contains a total of 19,487 images where the faces are all within the range of a ±45 degree yaw from a frontal face image. However, due to missing data or inaccurate labels, only 16228 images with well-defined age and gender labels are used. Out of 16228 samples, 7657 are from males and 8571 are from females. The 8 age groupings are classified as follows: 0-2, 4-6, 8-12, 15-20, 25-32, 38-43, 48-53, and 60+. The number of samples in each age grouping differs greatly, with the 25-32 age group containing the most number of samples (4,951).

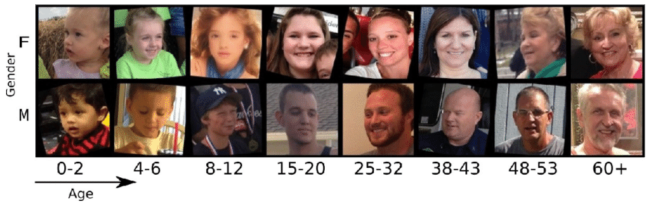


Figure 1: Samples from Adience Dataset

3.3 Network Architecture

The CNN that is utilized in this paper is adapted from [7], [9] and is a fully connected architecture. Different network architectures are used for gender and age classification. The base model comprises of 3 convolution layers followed by 3 fully connected layers. As in [9], [7], all three color channels are processed directly by the network itself. Input images are rescaled to 227 × 227 pixels from the original 256 × 256. The parameter values for network layers are adapted from both [7] and [9]. The 3 subsequent convolutional layers are as follows:

(1) **Convolutional Layer 1**

Gender model uses 24 convolutional filters and age model uses 16 convolutional filters. If there are additional convolutional or fully connected layers, then the number of filters is 16 for both age and gender. All filters are of size 7x7 and convolved with stride 4 and zero padding. This is followed by the chosen activation function and max-pooling with size 3x3 and stride 2. Batch normalization is then performed. Zero padding is then applied with size 2.

(2) **Convolutional Layer 2**

Gender model uses 32 convolutional filters and age model uses 20 convolutional filters. If there are additional convolutional or fully connected layers, then the number of filters are 20 for both age and gender. All filters are of size 5x5 and convolved with stride 1 and no padding as it was applied in the previous layer. This is also followed by the chosen activation function and max-pooling with size 3x3 and stride 2. Batch Normalization is then performed. Zero padding is then applied with size 1.

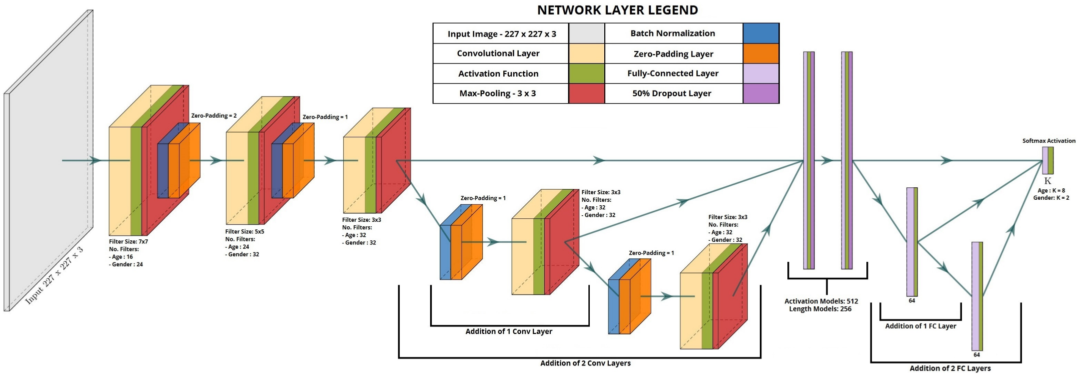


Figure 2: Full Network Architecture

(3) **Convolutional Layer 3**

The filter sizes for the third convolutional layer are determined based on whether additional layers have been used in the particular model. The reasoning for this is to allow for the preservation of information for deeper layers by filtering fewer times in the case of additional layers in the network. If there are additional convolution or fully connected layers, then the 24 filters are used for both age and gender. If not, 32 convolution filters for both age and gender models of size 3x3 are convolved with stride 1 and zero padding as it was already performed in the previous layer, followed by the chosen activation function and max-pooling with size 3x3 and stride 2.

(4) **Convolutional Layer N+3 (Only for N>0)**

This is where additional convolutional layers are added to test the variations of the network depth. Here N corresponds to the number of additional convolutional layers. 24 filters of size 3x3 are convolved with stride 1 and padding 1. This is followed by the chosen activation function and max-pooling with size 3x3 and stride 2.

(5) **Fully-Connected Layer 1**

First, the outputs of the last convolutional layer are flattened in order to obtain the correct input shape for the fully connected layers. 512 neurons fully connected to the output of the previous convolutional layer, followed by the chosen activation function layer and dropout layer with a dropout rate of 0.5 are used. If there are additional convolution or fully connected layers, then the layer size is 256 for both age and gender.

(6) **Fully-Connected Layer 2**

512 neurons fully connected to the output of the previous layer, followed by the chosen activation function layer and dropout layer with a dropout rate of 0.5 is used. If there are additional convolution or fully connected layers, then a layer size of 256 is used for both age and gender estimation.

(7) **Fully-Connected Layer 3 to 3+M (Only for M>0)**

This is where additional fully-connected layers shall be added to the network. This layer(s) shall consist of 64 neurons fully connected to the output of previous layer followed by the chosen activation function layer. Here no dropout layer is used as there is a relatively small number of neurons.

(8) **Fully-Connected Layer K where K is the Final Fully-Connected Layer**

This is the final fully-connected layer in the network. The output of this network must be either 2 or 8 neurons to accommodate the number of classes for gender and age. These neurons are fully-connected to the outputs of the previous layer. Softmax layer is then used to calculate the loss term that is optimized during training and it produces the class probabilities for either gender or age.

3.4 Training and Validation

The Adience dataset is split into 5 folds where it is ensured that no two folds contain the same sample simultaneously. The first 4 folds are used as training and validation (20% out of the data from 4 folds). The final training set contains a total of 10286 samples and the validation set will contain 2572 samples. A validation error will thus be calculated for each epoch and is recorded for every model. As in [9], the weights in all layers are initialized with random values from a zero-mean Gaussian with a standard deviation of 0.01. The network is trained from scratch. The target values are represented by binary vectors which are either of length 2 or 8. The target vectors will contain a 1 where the index of the vector matches the classification and a 0 elsewhere.

In order to avoid over-fitting, dropout layers are utilized. Here the dropout rate is fixed at 0.5. Also, the training set is shuffled for every epoch that is run to ensure no patterns in the data order are detected. Training is performed using the Stochastic Gradient Descent (SGD) optimizer. As there are 2 gender classes, gender models will calculate the loss term using binary cross-entropy, and categorical cross-entropy is used to calculate the loss term on the age models.

The activation function is chosen from a domain of usable activation functions and the effect of each on the network accuracy and loss term is analysed. The base model will use **Rectified Linear Unit (ReLU)**:

$$f(x) = x^{+} = max(0, x)$$

The other activation functions to be analyzed are:

Linear Activation Function:

$$f(x) = x$$

Exponential Linear Unit Activation Function (ELU):

$$x < 0 : f(x) = exp(x) - 1$$

$$x > 0 : f(x) = x$$

Hyperbolic Tangent Activation Function:

$$f(x) = \tanh(x)$$

Swish [15], [14] Activation Function:

$$f(x) = x \cdot sigmoid(\beta x)^{*}$$

This domain is partially derived from the list used in [14] and where the Swish activation function was discovered to be the best performing empirically proven function from their domain and this claim is tested in this paper for CNNs. β is a trainable parameter in this paper as to maximize its potential. In order to account for the sample count imbalance in the age and gender classes, class weights are calculated for the training set specifically and this is used to optimize the network output as to not over-fit a particular class. Class weights are calculated by inverting each class count and multiplying this value for each class by the maximum sample count found between all classes. Therefore the minimum weight possible is 1.0 and classes that are more under-represented will have weights greater than others.

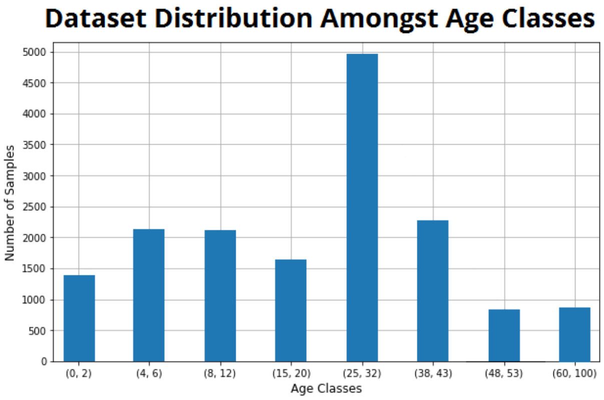


Figure 3: Sample Age Distribution In New Dataset Indicating Notable Class Imbalance

3.5 Testing Benchmark

In order to determine the effect of variations introduced into the models, exact classification accuracy is calculated. The Adience dataset is originally split into 5 folds. The last fold is excluded from the training and validation process and is used as the test set. This test set contains 3370 images, 1555 males and 1815 females. The age distribution for the test set is checked to ensure it contains enough samples from each age class to provide accurate testing results for the creation and analysis of confusion matrices.

Figure 4: Sample of the Test Set

Conclusions are drawn from results of the training, validation and testing accuracies and changes in loss terms for each network. Predictions made on the test set will provide insight into how well the validation accuracy represents the model output. The change in validation loss for each network of different sizes will give an accurate representation into how well particular networks trained. Correlation between validation accuracy and training accuracy is an effective measure for the detection of overfitting, as in the instance training accuracy increases and validation accuracy decreases, we know that we have trained the model to portray the training data in excess. The convergence of the loss term is evaluated from networks of varying length by calculating the relative change in the loss term between the initial and final training epoch.

4 EXPERIMENTS

4.1 Experimental Setup

Firstly, current work will explore the effect of the addition of convolutional and fully-connected layers has on the networks accuracy and loss term. Therefore a total of 5 pairs of models were created. With each pair consisting of a model for gender estimation and the other age estimation. Each pair was set with the exact same network parameters. The only change between pairs was the addition of either the number of convolutional layers (Conv) or fully-connected layers (FC). For each model the training and validation history is recorded for each epoch trained. Note that class weighting was not used in these tests as to determine whether the addition of layers would be able to account for the imbalance in the classes.

In order to form a base model as to compare the modified models with, a pairing of models were trained with 3 convolutional and 3 fully-connected layers. Then consideration is made for the addition of 1 and 2 convolutional layers added to the base model structure. The same is done with the addition of 1 and 2 fully-connected layers to the base structure. For training, the batch size is set to 50, and 5 epochs are run. The learning rate is set to 0.01 with a decay rate of 0.0001 and momentum value of 0.9.

In order to determine the effect different activation functions have on the network, a total of 5 pairs of models are created. With each pair consisting of a model for gender estimation and age estimation. The domain of activation functions are explored one-by-one and for every new model pair a different activation function is used. This activation function is applied to every layer excluding the output layer which will always use softmax as to allow training with the optimizer. Based on the analysis of the training and validation history of these models, those that tend to converge with a consistently decreasing loss term are tested on the test data and these results are further analysed.

All 5 pairs of models will have exactly the same values set for their network parameters expect for the activation function. In order to determine whether these networks can perform accurately at a level consistent with that of prior research such as [7], [6], class weights are used in order to account for the imbalance in the number of samples between classes. The epoch count for these models is 10. The learning rate is set to 0.01 with a decay rate of 0.001 and momentum value of of 0.9.

4.2 Experimental Results

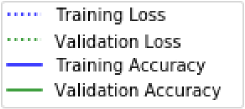
The pairs of models each with varying layer lengths were trained and their training and validation accuracies and losses are plotted below. The following table describes the total change in the loss term from epoch 1 to epoch 5 during training and validation:

Table 1: Total Change In Loss Term After 5 Epochs For Gender And Age Models

	Total Change in Loss Term For Gender Model	Total Change in Loss Term For Age Model
Base Model	-35.39%	-26.74%
Conv Increase by 1	-17.61%	24.69%
Conv Increase by 2	-0.18%	21.57%
FC Increase by 1	0.19%	0.75%
FC Increase by 2	-0.05%	-0.05%

From this table and the following diagrams, the base models far outperform models with additional layers. There is a notable convergence rate for the base model after 5 epochs. When we examine the addition of convolutional layers, the convergence rate of the loss term for gender decreases notably for each layer added. An important observation is an increase in the divergence rate for the age model with additional convolutional layers, thus the estimation in age gets worse for every epoch. A possible reason for this is that with the addition of convolutional layers, the model begins to learn features of the image that no longer correspond to age or gender, and therefore the model drops the features we are interested in.

When we examine the addition of fully-connected layers, we note an almost immediate halt in the change of the loss term. Therefore for more fully-connected layers we add, the models ability to achieve convergence on the loss term diminishes almost completely. These findings are consistent with [7] as it was found that for a model with this architecture and size, the addition of layers to the network would only lead to models of the same or worse performance. The legend for the graphs is as follows:



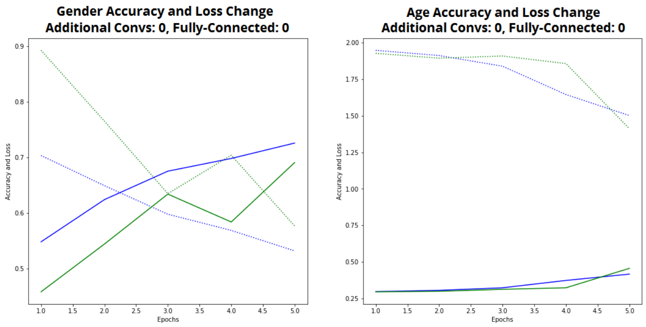


Figure 5: Results for gender (left) and age (right) of the base model with no additional layers added

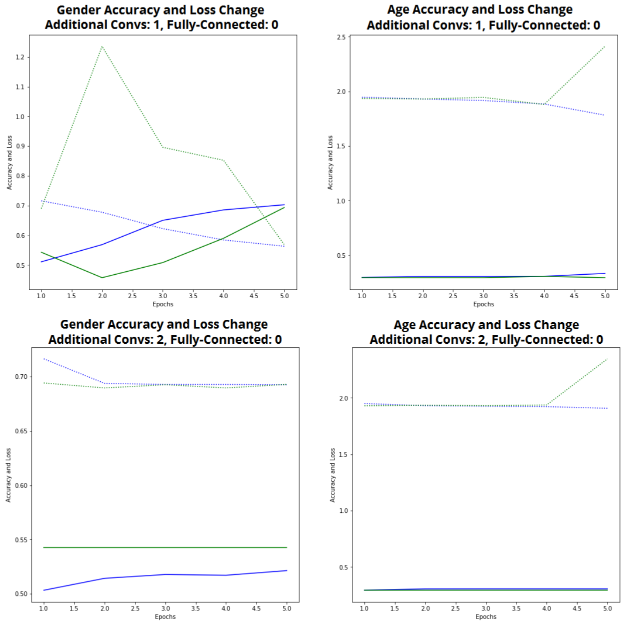


Figure 6: Results for gender (left) and age (right) with the addition of convolutional layers

Next we will examine the results from the use of different activation functions in the network as well as the use of class weights to account for the imbalance in class samples.

From the results in Fig. 8 we can observe a steady convergence for the models that use the ReLU, Linear and Swish activation functions. ELU offers no convergence over the 10 epochs trained and that may be the reason why, as ELU is typically used in networks that are trained for a greater number of epochs. Also ELU is found to be useful in much deeper networks where the layer count usually exceeds 40 as noted in [3]. The hyperbolic tangent function also performs poorly as an activation function here. A reason for this

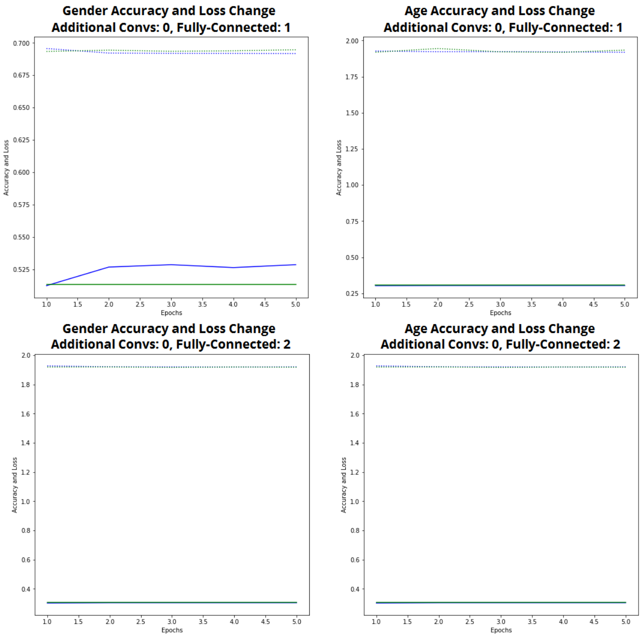


Figure 7: Results for gender (left) and age (right) with the addition of fully-connected layers

is that tanh is known to saturate for particular networks as it is both bounded above and below [19]. This is in contrast with ReLU, Linear and Swish, which are all unbounded in at least one direction [15], [14]. It is unlikely that ELU has become saturated although it is possible to saturate in the negative direction [3]. However the effect of the small number of epochs is more likely.

The following experiments will explore the 3 activation functions that were shown to have a degree of convergence on the loss term after the 10 epochs, those activation functions being: ReLU, Linear and Swish. The following results depict the accuracy of the 3 model pairs on the testing set.

Test accuracies of the gender model were 79% for ReLU and Swish, and 77% for Linear functions. These results are therefore consistent with the results in [7] where a benchmark accuracy for gender was found to be 80.8%. An important note is that the network used in this paper was much larger, with filter sizes of 96, 256 and 384 for each convolution layer as opposed to the 24, 32, 32 for the gender models.

The results from the confusion matrices in Fig.9 show a pattern in age classification amongst all 3 models. Most notably having a very high exact accuracy for the age group (0, 2) and the worst for the age group (48, 53). A possible reason for this is the networks ability to learn features indicative with the age group (0, 2). Also, as seen in Fig. 3, the age group (48, 53) had the fewest number of samples in the entire dataset. A possible reason for the low accuracy can be attributed to the insufficient number of samples for that particular class. It should be noted that even with the use of class weights, the poor accuracy for the age group (48, 53) could not be alleviated. This is evident from the confusion matrices, as we see a large number of misclassifications being attributed to

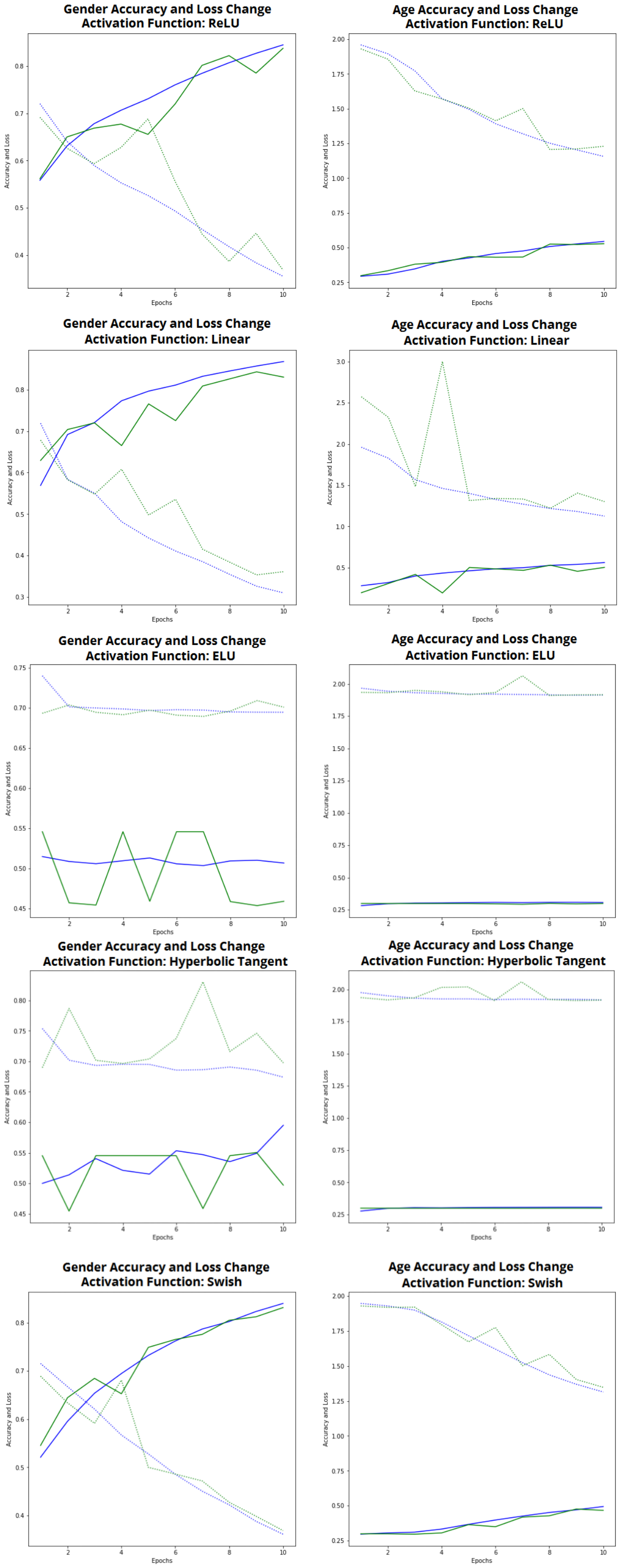


Figure 8: Results for gender (left) and age (right) with models using ReLU, Linear, ELU, tanh **and Swish activation functions**

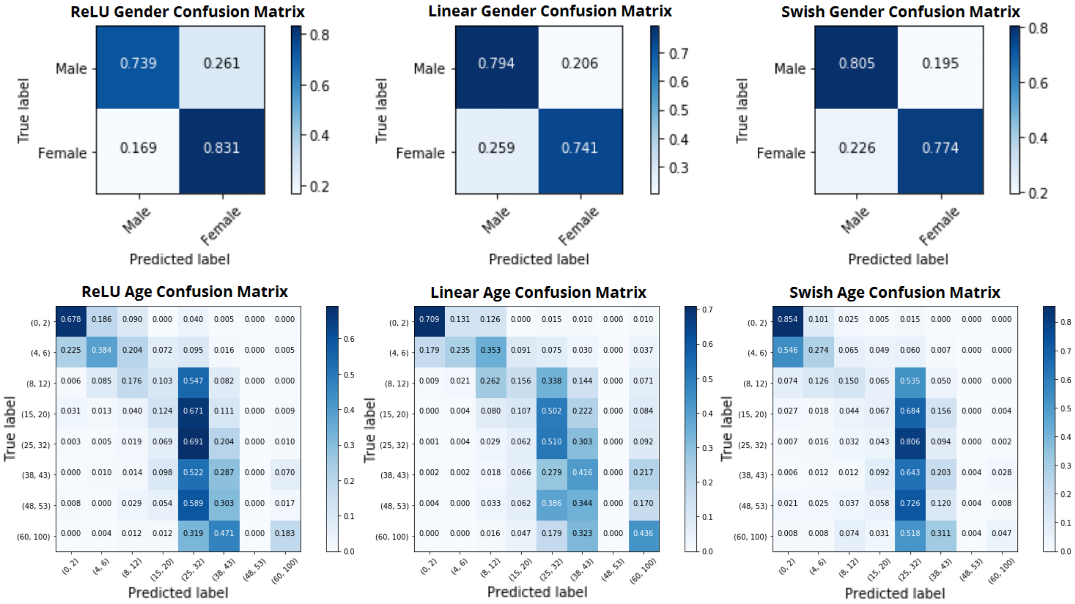


Figure 9: Confusion Matrices For Age Models

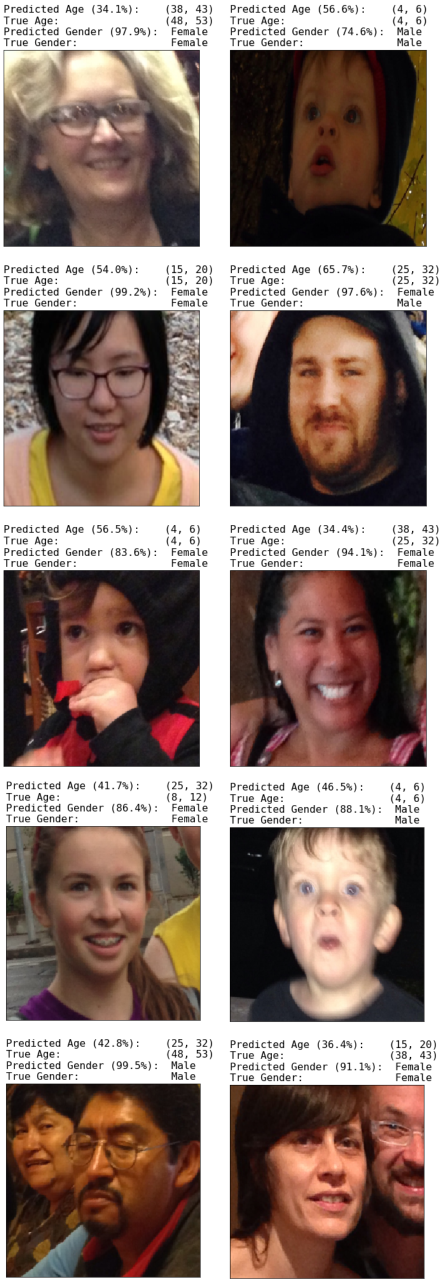
the (25, 32) age class, as this class contained the greatest number of samples in the dataset. This distribution amongst accuracies is also discussed in [7], Therefore this pattern may be inherent in the network architecture or dataset samples, and not necessarily the parameters of the network. With a benchmark accuracy of 50.2% for age classification as given in [7] and 49.5% as given in [6], our models accuracy fall short with an accuracy of 40% from both ReLU and Swish activations. Therefore more training as well as an increase in the number of convolutional filters is required in order to obtain similar results to these benchmarks.

5 CONCLUSION

The primary goal of this paper was to determine the effect different activation functions will have with respect to the model accuracy and convergence of the loss term, and whether $f(x) = x \cdot sigmoid(\beta x)$ with trained β is the better activation function to use as discovered in [14] for their activation function domain. From the experimental results, only 3 of the 5 activation functions tested proved to show convergence for the loss term for 10 training epochs. A notable property for these functions is that they are all unbounded in at least one direction and therefore are far less likely to suffer from saturation. Lack of convergence for the exponential linear unit (ELU) activation function can be attributed to the small number of epochs and shallow network architecture, therefore further testing with a deeper network and more epochs is required to truly determine the effectiveness of ELU in the task of age and gender estimation. With regard to performance of Swish, both ReLU and Swish achieved the same average accuracy results for both age and gender classification. Although ReLU had greater accuracy on higher age classes than that of Swish, this was negated by the performance of Swish on specific age classes (0, 2) and (25, 32). Therefore it should be noted that ReLU was able to generalize the age classification better than that of Swish. A possible reason ReLU and Swish performed almost identically is for a great enough β value in the Swish function, the function converges to the form of the ReLU activation function.

The second goal was to determine the effect a change in the number of convolutional layers or fully-connected layers has on the classification accuracy and convergence of the loss term of a model, and whether a larger model is more affected by over-fitting. Based on the results, both an increase in the number of convolutional layers and fully-connected layers saw a notable decrease in the models accuracy and loss term convergence rate. A possible reason for this decrease with added convolutional layers was the models preference to over-fit on redundant features that are not relevant to the task of age or gender classification. Additional fully-connected layers saw a complete inability to achieve convergence on the loss term. This may be due to vital information being lost as it travels through the deeper network. However, further testing should be done with addition fully-connected layers of different sizes as to confirm whether this is the case, as it may be possible that the size of 32 neurons as used in this paper was not enough to maintain the integrity of the convolved information for softmax based classification.

The final goal of this study was to determine whether it is possible to recreate the results in [6] and [7] with fewer network variables in the form of convolution filters, as well as notably fewer training epochs. The model used for gender classification with ReLU and Swish activation functions was within a 2% accuracy range with the initial benchmark set in [7]. With regard to age classification, more training and a more complex model is required in order to achieve similar results to that in the literature, and from this it can be noted that the features representing age in image data are far more complex than that of gender.



```
Predicted Age (34.1%):    (38, 43)
True Age:                 (48, 53)
Predicted Gender (97.9%):  Female
True Gender:               Female
```

```
Predicted Age (56.6%):     (4, 6)
True Age:                  (4, 6)
Predicted Gender (74.6%):  Male
True Gender:               Male
```

```
Predicted Age (54.0%):    (15, 20)
True Age:                 (15, 20)
Predicted Gender (99.2%):  Female
True Gender:               Female
```

```
Predicted Age (65.7%):    (25, 32)
True Age:                 (25, 32)
Predicted Gender (97.6%):  Female
True Gender:               Male
```

```
Predicted Age (56.5%):    (4, 6)
True Age:                 (4, 6)
Predicted Gender (83.6%):  Female
True Gender:               Female
```

```
Predicted Age (34.4%):    (38, 43)
True Age:                 (25, 32)
Predicted Gender (94.1%):  Female
True Gender:               Female
```

```
Predicted Age (41.7%):    (25, 32)
True Age:                 (8, 12)
Predicted Gender (86.4%):  Female
True Gender:               Female
```

```
Predicted Age (46.5%):     (4, 6)
True Age:                  (4, 6)
Predicted Gender (88.1%):  Male
True Gender:               Male
```

```
Predicted Age (42.8%):    (25, 32)
True Age:                 (48, 53)
Predicted Gender (99.5%):  Male
True Gender:               Male
```

```
Predicted Age (36.4%):    (15, 20)
True Age:                 (38, 43)
Predicted Gender (91.1%):  Female
True Gender:               Female
```

Figure 10: Prediction output including prediction certainties using the ReLU activation function

REFERENCES

[1] A. Anand, R. D. Labati, A. Genovese, E. Muñoz, V. Piuri, and F. Scotti. 2017. Age estimation based on face images and pre-trained convolutional neural networks. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. 1–7. https://doi.org/10.1109/SSCI.2017.8285381

[2] Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. 2013. Provable Bounds for Learning Some Deep Representations. *CoRR* abs/1310.6343 (2013). arXiv:1310.6343 http://arxiv.org/abs/1310.6343

[3] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2015. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *CoRR* abs/1511.07289 (2015). arXiv:1511.07289 http://arxiv.org/abs/1511.07289

[4] Afshin Dehghan, Enrique G. Ortiz, Guang Shu, and Syed Zain Masood. 2017. DAGER: Deep Age, Gender and Emotion Recognition Using Convolutional Neural Network. *CoRR* abs/1702.04280 (2017). arXiv:1702.04280 http://arxiv.org/abs/1702.04280

[5] Piotr Dollar, Zhuowen Tu, Pietro Perona, and Serge Belongie. 2009. Integral Channel Features. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 91.1–91.11. doi:10.5244/C.23.91.

[6] E. Eidinger, R. Enbar, and T. Hassner. 2014. Age and Gender Estimation of Unfiltered Faces. *IEEE Transactions on Information Forensics and Security* 9, 12 (Dec 2014), 2170–2179. https://doi.org/10.1109/TIFS.2014.2359646

[7] Ari Ekmekji. 2016. Convolutional Neural Networks for Age and Gender Classification.

[8] Z. Kuang, C. Huang, and W. Zhang. 2015. Deeply Learned Rich Coding for Cross-Dataset Facial Age Estimation. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*. 338–343. https://doi.org/10.1109/ICCVW.2015.52

[9] G. Levi and T. Hassncer. 2015. Age and gender classification using convolutional neural networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 34–42. https://doi.org/10.1109/CVPRW.2015.7301352

[10] Refik Can Malli, Mehmet Aygun, and Hazim Kemal Ekenel. 2016. Apparent Age Estimation Using Ensemble of Deep Learning Models. *CoRR* abs/1606.02909 (2016). arXiv:1606.02909 http://arxiv.org/abs/1606.02909

[11] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. 2014. Face detection without bells and whistles. In *ECCV*.

[12] Masakazu Matsugu, Katsuhiko Mori, Yusuke Mitari, and Yuji Kaneda. 2003. Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural Networks* 16, 5 (2003), 555 – 559. https://doi.org/10.1016/S0893-6080(03)00115-1 Advances in Neural Networks Research: IJCNN '03.

[13] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep Face Recognition. In *BMVC*, Vol. 1. 41.1–41.12.

[14] Prajit Ramachandran, Barret Zoph, and Quoc V. Le. 2017. Searching for Activation Functions. *CoRR* abs/1710.05941 (2017). arXiv:1710.05941 http://arxiv.org/abs/1710.05941

[15] Prajit Ramachandran, Barret Zoph, and Quoc V. Le. 2017. Swish: a Self-Gated Activation Function. (10 2017).

[16] R. Ranjan, S. Zhou, J. C. Chen, A. Kumar, A. Alavi, V. M. Patel, and R. Chellappa. 2015. Unconstrained Age Estimation with Deep Convolutional Neural Networks. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*. 351–359. https://doi.org/10.1109/ICCVW.2015.54

[17] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2014. Going Deeper with Convolutions. *CoRR* abs/1409.4842 (2014). arXiv:1409.4842 http://arxiv.org/abs/1409.4842

[18] Paul Viola and Michael Jones. 2001. Robust Real-time Object Detection. In *International Journal of Computer Vision*.

[19] Bing Xu, Ruitong Huang, and Mu Li. 2016. Revise Saturated Activation Functions. *CoRR* abs/1602.05980 (2016). arXiv:1602.05980 http://arxiv.org/abs/1602.05980

[20] K. Zhang, C. Gao, L. Guo, M. Sun, X. Yuan, T. X. Han, Z. Zhao, and B. Li. 2017. Age Group and Gender Estimation in the Wild With Deep RoR Architecture. *IEEE Access* 5 (2017), 22492–22503. https://doi.org/10.1109/ACCESS.2017.2761849

[21] Ke Zhang, Miao Sun, Tony X. Han, Xingfang Yuan, Liru Guo, and Tao Liu. 2016. Residual Networks of Residual Networks: Multilevel Residual Networks. *CoRR* abs/1608.02908 (2016). arXiv:1608.02908 http://arxiv.org/abs/1608.02908