Age estimation using Deep Convolutional Neural Networks

Suhas chowdary J
Electrical and Computer Engineering
University of Arizona
Tucson, AZ 85719
suhaschowdaryj@email.arizona.edu

Harshil Pk
Electrical and Computer Engineering
University of Arizona
Tucson, AZ 85719
harshilpk@email.arizona.edu

Abstract

Analysis of facial images has been of growing importance in the recent days. Automatic age estimation is one of the major attributes of facial analysis which can be used in various applications like security control, human-computer interaction and law enforcement. Different people will age at different phase based on the habits, genes and their lifestyle, which makes it difficult to estimate the age of a person accurately. In this paper, we have proposed two network models for the estimation of the age using Convolutional Neural Networks(CNN). First model is built based on [1] and VGG net architecture and second model is based on GoogLeNet architecture. We have implemented the proposed network models on Google Tensor flow library. A comparative study has been conducted on the performance of the two models considering various design constraints like error, computational overhead, memory utilization and design complexity.

1 Introduction

The human face has significant importance in the society. It does reveal various characteristics of a person including gender, identity, ethnicity, age and expressions. Identification of the age of a person plays a key role in various occasions like determining underage drivers, security control and accessing age restricted goods to people. It is very difficult to estimate the age just by his appearance because different people age at different phase depending on their work, habits, location and hygienic conditions. However, with the improvement in the human-computer interaction, the study of the facial analysis has been increased in the recent times. Various aspects of facial image analysis like gender, emotions are being identified accurately but age estimation remains as a challenging problem in the world of machine learning [8]. Some of the reasons are it is not a two to six class problem like in gender classification and emotion detection, the images may be ambiguous without proper timestamp (the date when the picture is taken), poor image quality, multiple faces in the image etc. Estimation of the age used to be implemented by extracting facial features using techniques like SVM which hasn't turned to be attractive due to reasons like outliers and nature of the images. However, with the intervene of Deep Convolutional Neural Networks(DCNN) ([10],[12],[13]) there has been a significant development in age estimation.

Extensive use of Deep Convolutional neural networks has been started since Alexnet [2] won the ImageNet large Scale Visual Recognition Challenge (ILSVRC) by a large margin in 2012. The Alexnet has introduced use of new techniques like Rectified linear units(ReLu) as activation functions and new data augmentation techniques. Whereas It uses large filters with dimensions of 11x11 and 9x9 which are computationally expensive. It also retains more noise in the background which depletes the prediction accuracy. This problem of retaining more noisy background was solved by VGGnet architecture [3] proposed by Simonyan in 2014. The VGGnet architecture introduced small uniform 3x3 filters across all the convolutional layers which decreases the noise in the background significantly. However, the VGG net architecture remained as a computationally expensive model which was solved by GoogLeNet architecture. The GoogLeNet architecture [4] proposed by google has improved the image classification by introducing new inception modules inspired by network in network architecture [5]. GoogLeNet architecture enhances the discriminative ability for local patches within the receptive field.

Figure 1: Sample images of the IMDB-WIKI dataset

In this paper, we have proposed two network models for estimation of the age. First network model is based on [1] and VGG net architecture. We have developed second model based on the inception modules of the GoogLeNet architecture. The designed models are shallow when compared with other deep convolutional architectures which prevents overfitting. The age estimation has been done on IMDB-WIKI dataset. The dataset consists of around 523,000 images which are wild and unconstrained. The dataset contains people of different genders, different ages, regions and of diverse ethnicity. The age labels vary from 1 to 100. Sample images are shown in the Figure 1.

The rest of the paper is organized as follows: Section 2 depicts the network architectures of the two proposed models based on [1] and GoogLeNet. In Section 3, the experiments which includes data preprocessing and testing and training the dataset are included. Results are reported in Section 4. Finally, Section 5 concludes the paper.

2 Network architecture

2.1 Proposed approach

We have proposed two network models for the age estimation. The network of the first model has been built by using three convolutional layers with filter size of 5x5 and 3x3 followed by two FC (fully connected) layers. A detailed implementation has been described in section 2.1. The schematic diagram of the architecture is shown in figure 2. Second model is built based on the inception module of GoogLeNet architecture. It does consist of 3x3 and 5x5 filters pre-convolved with 1x1 filters to decrease the computational overhead. The network design is explained in the section 2.3.

2.2 Model 1: Network based on [1] and VGGnet architecture

The model has been built as follows:

1. First convolution layer is built by using 32 filters each of size 5x5x3 convolved with the input layer. Rectified Linear Unit (ReLu) and max pooling are applied to the convolutional layer followed by local response normalization layer.

2. Second convolutional layer has been built by passing the output of the first convolution layer to 64 kernels each of size 3x3x32 followed by ReLu layer, max pooling layer and a local normalization layer.

3. The resultant of the second convolution layer has been fed to 128 filters each having dimensions of 3x3x64 and then treated with ReLu layer and max pooling layer. Local normal response layer is applied to the third convolution layer.

4. The First fully connected layer consists of 1024 neurons which is connected with the output of the third convolution layer. It is followed by a ReLu layer and finally a drop out layer with dropout ratio 0.75 has been applied to the resultant.

5. The output of the first fully connected layer has been fed to the second fully connected layer which contains 100 labels helps in estimating the age of the face image.

The input images of size 256x256x3 are fed into the above proposed model. The output of the second fully connected layer is given to softmax layer to find the probabilities. The weights are updated using stochastic gradient method. However, the fully connected layers use many parameters which is shown in Table 1. This enlarged network makes it more prone to overfitting and computations sum up which becomes a major bottleneck in the model 1. To overcome this, we have proposed model 2 which moves from fully connected layer to sparsely connected architecture.

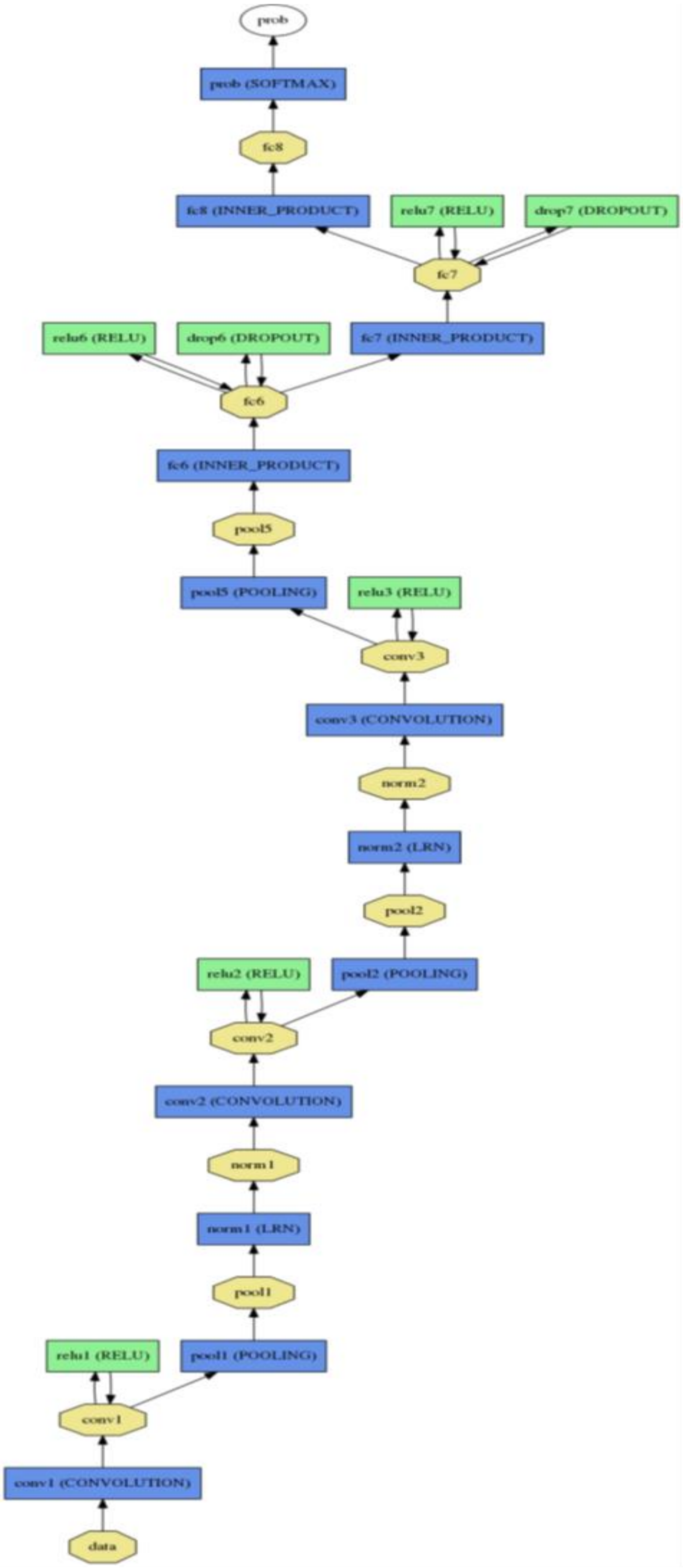

Figure 2: Schematic diagram of proposed model 1

2.2 Model 2: Network based GoogLeNet architecture

The proposed design has been built on the GoogLeNet Inception v3 architecture. The design uses 1x1 convolution filters before convolving with 3x3 and 5x5 filters to decrease the computational overhead. The implementation is depicted in the figure 3.

The built model is as follows:

1. Conv_1x1: 32 filters each of size 1x1x3 are convolved with the input layer to build the first convolution layer. Rectified Linear Unit (ReLu) and local response normalization layer are applied to the convolutional layer.

2. Conv_3x3: The resultant of the first convolutional layer has been fed to 64 filters each having dimensions of 3x3x32 and then treated with ReLu layer, max pooling layer and Local normalization response layer.

3. Conv_5x5: The output of the first convolutional layer is also applied to 64 kernels each of size 5x5x32 and then followed by ReLu layer, max pooling layer and a local normalization layer.

4. Maxpool_3x3: The max pooling layer has been applied to the input layer which is followed by a ReLu layer.

5. Conv_1x1_maxpool: The output of the Maxpool_3x3 has been treated with 64 filters with the dimensions of 1x1x3. The obtained layer is followed by a ReLu layer and a normalization layer.

6. The final layer is obtained by filter concatenation of the convolved layers Conv_1x1, Conv_3x3, Conv_5x5 and Conv_1x1_maxpool which is fed into softmax layer.

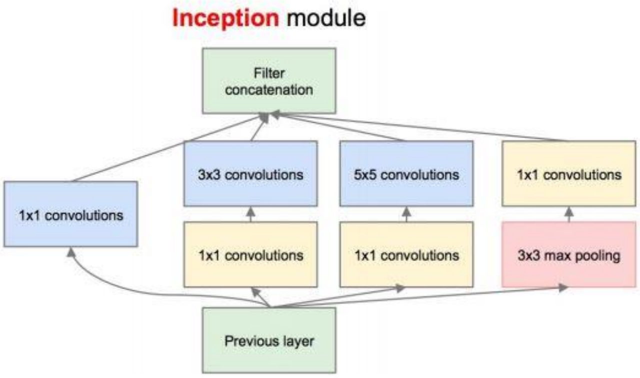

Figure 3: Schematic diagram of proposed model 2

3 Experiments

We have trained our model on Wiki-IMDB dataset which contains 523,051 images. The images of the dataset are collected only if timestamp (the date when photo is taken) is available on the image so that the age labels are authentic. We have used 15,000 images for training and 1000 images for testing due to the resource constraints like GPU and memory expenses. The data has been trained and tested on CPU with 16 GB memory.

3.1 Data preprocessing

The images in the dataset are wild and unconstrained which are collected from various instances without any conditioning. We have used MATLAB to extract the respective age labels and image paths. Each image has been resized to 256x256. Images have been subjected to mean subtraction across three channels(RGB) such that the cloud of the data will be centered around the origin.

3.2 Training and testing

Weights are initialized with random normal variables close to zeros such that each weight vector is unique. If all the weights are initialized with zero, then during back propagation, the neurons may be saturated due to the same gradient for all weights. So, the random weight initialization ensures that the neurons will not be saturated while back propagation due to different weights.

Each convolution layer is treated with a ReLu layer, max pooling layer and a local normal response layer. We have used Rectified linear unit(ReLu) as activation function which helps in increasing the speed of convergence of the stochastic gradient descent. The resultant is treated with a Max pooling layer. It is used to reduce the spatial dimensions of the model which results in decrease of number of parameters and computation involved in the model. Local response normalization layer(LRN) has been used which helps in contrasting high frequency features with that of low frequency features. This results in eliminating the noise in the background. Drop out layers are used to prevent overfitting. We have trained the data using stochastic gradient descent method with learning rate of 0.001 and batch size of 30.

4 Results and analysis

The first network is designed based on [1] and VGG net architecture, whereas the second network utilizes GoogLeNet architecture. The model based on VGG net architecture has design simplicity. The three convolutional layers of the model uses one 5x5 filter and two 3x3 filters produces an effective receptive field of 7x7 filters. This model emulates receptive field of a larger filter while using small filters, hence decreasing the computation. Also, the designed model increases in depth while the spatial dimensions shrinks which result in improving minute activations. We have achieved an error rate of 0.51 with network depth of 5 layers.

However, the first model based on VGG net architecture still have high computational overhead. We proposed a second model based on the inception module which uses 1x1 filters before using 3x3 and 5x5 filters. Initial convolution with 1x1 kernels results in dimensionality reduction. The 1x1 and 3x3 filters present extracts fine grain details about the image whereas, the 5x5 filters helps in creating large receptive fields. Also, these layers are treated with ReLu layer and max pooling layer resulting in higher activations and spatial dimension reduction. Apart from performing all these operations, the computational overhead is decreased.

Table 1: Error rate

Network model	**Error**
Model(1) based on [1] and VGG net architecture	0.51
Model(2) based on GoogLeNet architecture	0.46

Fully connected layers are not used in the second model which consumes major computational resources in the first model. The total number of parameters involved in the two models are illustrated in tables 2 and 3 respectively. The two Fully connected layers in the model 1 requires 134M parameters which is a major expense in the computation and memory. The model 2 uses filter concatenation and max pooling which removes the necessity of the fully connected layers. So, the number of parameters involved decreases by 1M in the second model. Despite all this, the error rate has been brought down to 0.46. The errors of both the models are depicted in Table 1. However, the design complexity of the second model increases as new layers like filter concatenation are involved.

Table 2: Model 1 layer dimensions and parameters

Layer	**Dimensions**	**Weights**	**Parameters**
Input layer	[256x256x3]	0	0
Convolutional layer 1	[256x256x32]	32-(5*5*3)	2400
Max pool 1	[128x128x32]	0	0
Convolutional layer 2	[128x128x64]	64-(3*3*3)	1728
Max pool 2	[64x64x64]	0	0
Convolutional layer 3	[64x64x128]	128-(3*3*3)	3456
Max pool 3	[32x32x128]	0	0
Fullyconnected layer1	[1x1x1024]	1024-(32*32*128)	134M
Fullyconnected layer2	[1x1x100]	100-(1024*1*1)	102K
Total parameters			134.15M

Table 3: Model 2 layer dimensions and parameters

Layer	**Dimensions**	**Weights**	**Parameters**
Input layer	[256x256x3]	0	0
Conv_1x1	[256x256x32]	32-(1*1*3)	96
Maxpool_3x3	[128x128x3]	0	0
Conv_3x3	[256x256x64]	64-(3*3*3)	1728
Max pool 2(a)	[128x128x64]	0	0
Conv_5x5	[256x256x64]	64-(5*5*3)	4800
Max pool 2(b)	[128x128x64]	0	0
Conv_1x1_maxpool	[128x128x64]	64-(1*1*3)	192
Filter concatenation	[128x128x192]	0	0
Total parameters			6K

5 Conclusion

We proposed two models for the estimation of the age. We have shown the internal structure and working of the two models. First model is based on the [1] and VGGnet architecture. The design is relatively simple which consists of three convolution layers and two fully connected layers. This model uses filters of size 3x3 and 5x5 which produces large receptive field. This model also performs well in extracting discriminant information. However, the fully connected layers became bottleneck in the model as they consume majority of the resources. Second model uses 1x1 filters to decrease the spatial dimensionality. This model has advantage of computational overhead and memory utilization over first model due to replacement of fully connected layers with sparsely connected architecture using average pooling and filter concatenation. Also, second model outperforms the first model in terms of performance accuracy by a considerable margin as it captures both fine grain details and create large receptive fields with filters of different dimensions like 1x1, 3x3 and 5x5. Future work can be done in fine tuning both the architectures to achieve high accuracy while meeting the design constraints like number of hyperparameters, memory and computational demands more effectively.

References

[1] Levi, Gil, and Tal Hassner. "Age and gender classification using convolutional neural networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2015.

[2]. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.

[3] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

[4] Szegedy, Christian, et al. "Going deeper with convolutions." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015

[5].Lin, Min, Qiang Chen, and Shuicheng Yan. "Network in network." *arXiv preprint arXiv:1312.4400* (2013).

[6] Geng, Xin, Zhi-Hua Zhou, and Kate Smith-Miles. "Automatic age estimation based on facial aging patterns." *IEEE Transactions on pattern analysis and machine intelligence* 29.12 (2007): 2234-2240.

[7] Geng, Xin, Chao Yin, and Zhi-Hua Zhou. "Facial age estimation by learning from label distributions." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.10 (2013): 2401-2412.

[8] Hewahi, Nabil, et al. "Age estimation based on neural networks using face features." *Journal of Emerging Trends in Computing and Information Sciences* 1.2 (2010): 61-67.

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

[10] Lawrence, Steve, et al. "Face recognition: A convolutional neural-network approach." *IEEE transactions on neural networks* 8.1 (1997): 98-113.

[11] Can Malli, Refik, Mehmet Aygun, and Hazim Kemal Ekenel. "Apparent Age Estimation Using Ensemble of Deep Learning Models." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2016

[12] Giacinto, Giorgio, and Fabio Roli. "Design of effective neural network ensembles for image classification purposes." *Image and Vision Computing* 19.9 (2001): 699-707.

[13] Kanellopoulos, I., and G. G. Wilkinson. "Strategies and best practice for neural network image classification." *International Journal of Remote Sensing* 18.4 (1997): 711-725.

[14] http://cs231n.github.io/convolutional-networks

[15] Park, Soo Beom, Jae Won Lee, and Sang Kyoon Kim. "Content-based image classification using a neural network." *Pattern Recognition Letters* 25.3 (2004): 287-300.

[16] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[17] Yu, Wei, et al. "Visualizing and Comparing AlexNet and VGG using Deconvolutional Layers." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2015

[18] Erhan, Dumitru, et al. "Scalable object detection using deep neural networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014.