Face2Year : CS 395T Deep Learning Project 1

Haresh Karnan[1] and Manish Reddy[2]

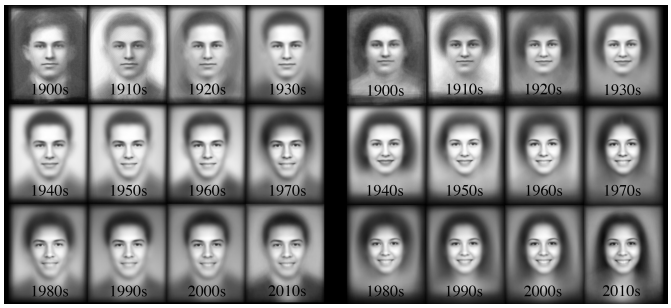

Fig. 1: Images from the yearbook data set containing face photos and the year the images were captured.

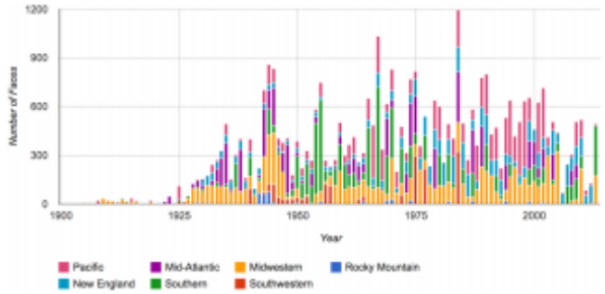

Fig. 2: The distribution of portraits per years and region [1].

***Abstract*— This paper focuses on image level prediction using deep convolutional networks. A data set of frontal-face images of American yearbook photos containing the years (spanning from 1905 to 2013) the pictures were taken as labels is considered in our analysis. Our methods focus on using powerful and time efficient techniques, such as transfer learning and fine tuning, to fit a deep convolutional architecture to predict the year a picture was taken. We compare a pre-trained VGG net and RESnet50 on its L1 accuracy on the validation data to compare how well each architecture performs. We achieved a mean L1 distance of 5.33 years with a validation set accuracy of 4.5% on the VGG model which was the best performer, compared to the RESnet50 which achieved a 9.05 years mean L1 distance on the validation dataset.**

I. INTRODUCTION

Advances in Deep Learning theory and practices have catapulted the research in learning theory forward. Applications in the domain of Computer Vision include Object detection, Feature extraction, feature tracking, etc. Such applications were hard to implement before the era of Deep Learning where progress was through hand tuned features. When the data being dealt with changes slightly, these features had to be handcalculated from the start. Deep Learning models eliminate this need, by introducing a learning architecture that learns features from different samples of data. Generalization is another capability that the Deep Learning architectures have a lot of hope in solving, which is of prime importance in the domain of Computer vision. Through the years, the Deep Learning community has been generating several interesting architectures like the VGG16, VGG19, Inceptionv3, LeNet, RESnet etc. In this project, we pick two architectures VGG16 and RESnet and experiment how well they perform at this task.

One of the most interesting features of Deep architectures that sets it apart from rest of the hand tuned techniques is the concept of transfer learning. Architectures trained towards a particular problem statement can be reused by freezing some of the layers and only training the later layers, hence customizing it for the new problem. This reusability property is beneficial in reducing training time while applying to new datasets. Another interesting property is the concept of fine tuning an architecture where instead of freezing the entire pre-trained model, only first few layers are frozen and the rest of the architecture is open for trainining. Each method has its own advantage and disadvantage and in this project, we employ both transfer learning and fine tuning on the VGG16 and RESnet50 models.

The novelty of our research lies in the fact that we did not use VGG or RESnet pre-trained on imagnet datset that contains random objects other than faces. Since our problem domain is related to faces and features in the faces that may be an unknown function of the year the photo was taken, it is intuitive to use a model that was pre-trained on faces dataset. In our analysis, we used VGGnet and RESnet50 pre-trained on the Oxford Faces dataset. This ensures the base model emphasizes more on learning facial features present in the training dataset.

We ran 30 experiments with different parameters such as models, learning rate, batch size, regularization, regularization parameter and error metric for the training phase and evaluate the best L1 distance and training accuracy on the validation dataset. From our experiments, we noticed that the VGG architecture produced the best L1 distance of 5.4 years with a validation dataset accuracy of 4.5%. While training with the imagenet dataset, we achieved the best L1 distance of 14 years on VGGnet with a validation dataset

*This work was not supported by any organization

[1]Haresh Karnan is with Department of Mechanical Engineering, The University of Texas at Austin, TX, USA haresh.miriyala@utexas.edu

[2]Manish Reddy is with the Department of Electrical Engineering, The University of Texas at Austin, TX, USA manishreddy@utexas.edu

accuracy of 0.8%. This shows that clearly, one has to use the VGG architecture pre-trained on the Face dataset instead of imagenet to achieve significant results.

II. TECHNICAL SECTION

This technical section focuses on the different parameters that were used in this experiment. Parameters like the learning rate, regularization parameter, epochs, loss and regularization have been changed between runs and experimented with.

A. Finetuning FaceNet

After thorough visual inspection of the data-set, we noticed that the faces are perfectly aligned at eyes and lips. Some of the images seemed to have been affine-transformed in order to be properly aligned. With this observation, we reasoned that choosing the conventional method of fine tuning network that was trained on CIFAR/ImageNet was an overkill. This is because of how we did not need the capacity of these networks to be scale, color and position invariant in detection/classification. Whatever visual information we might need for classifying these yearbook pictures was present at approximately the same location in every picture with the same scale. Thus, we resorted to using networks that work on understanding faces and faces only and not instead solving the problem of finding faces in the image.

One such application that closely matched our requirements was detecting people given only their cropped, centered faces. FaceNet[**?**] , is the paradigm that does exactly this. Using either VGG/ResNet as a basis, FaceNet aims to detect faces by producing a 128 dimensional embedding for each unique-face. This embedding can then be compared with a L2 metric to identify different people. The embedding was generated by using the Triplet Loss metric [**?**], which is based on distance between 2 images of the same person with respect to a third image of a very different person. Thus, the network learns to produce embedding so as to minimize intra-person image embedding distance while maximizing inter-person image embedding distance. We used these 128 dimensional embeddings as a starting point and constructed various shallow classifiers to convert into years. The results are summarized in the following table.

TABLE I: Embedding Fine Tuning Validation Results

Config	T-10	T-5	T-3	Accuracy	L1
(128x128x104)	54.3%	30.7%	19.3%	4.5%	6.722

We tried varying t he depth of the architecture but noticed severe over-fitting. The above reported result is the best we've achieved considering over-fitting as well as accuracy. The results were surprising since the network was very specific to faces. However, we now reason that maybe the fact that it was specifically trained to learn distinctive features in faces, it might have discarded information about expressions and such, which is a crucial part in discerning between years. Also, this method blatantly ignores all other visual

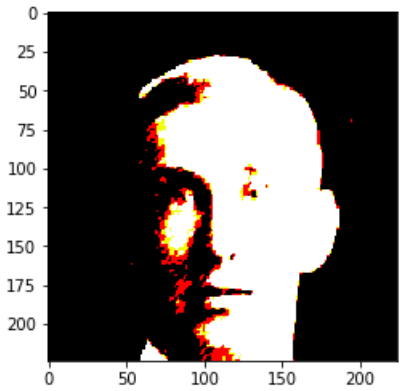

Fig. 3: Test image visualization.

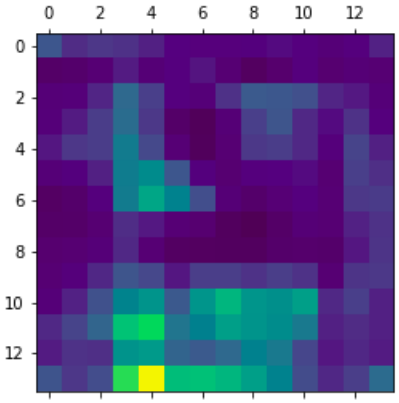

Fig. 4: Heat map.

information like hair styles, image quality, facial expressions and other valuable insights that separate the images between years.

B. Training using VGG16 and RESnet50 Face

We consider the task of predicting the year the photos were taken by posing it as a classification problem involving 105 years between 1905 and 2013. Some of the years did not have any data (For ex. 1907). There were 22840 images of Male and Female students in the training dataset and about 5009 images in the validation dataset. It can be seen from the distribution shown in Fig. 2 that the data is very random and there is no region or year favored in the dataset.

We fine-tuned in 2 variations : First, in the transfer learning step we freeze the entire pre-trained base model and then add a fully connected layer followed by Dropout of 0.5 and Soft-max that we train. Second, we unfreeze last 4 layers in the base model followed by a fully connected layer, add a Dropout of 0.5 and a Soft-max layer, with weights from the transfer learning step. We initially did not add Dropout and then noticed that the generalization was very poor - The training accuracy reaches 90% but the validation accuracy is less than 1%. This led us to add the dropout layer and we did not increase the Dropout value beyond 0.5 as doing so did not improve the validation accuracy.

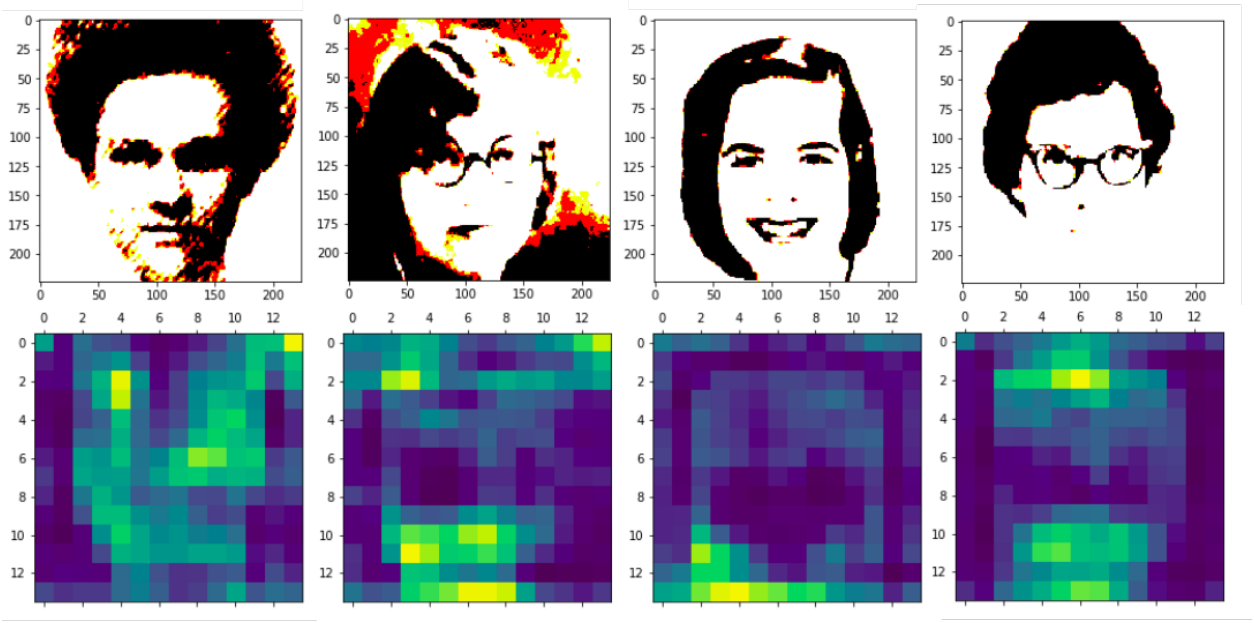

Fig. 5: Heat map.

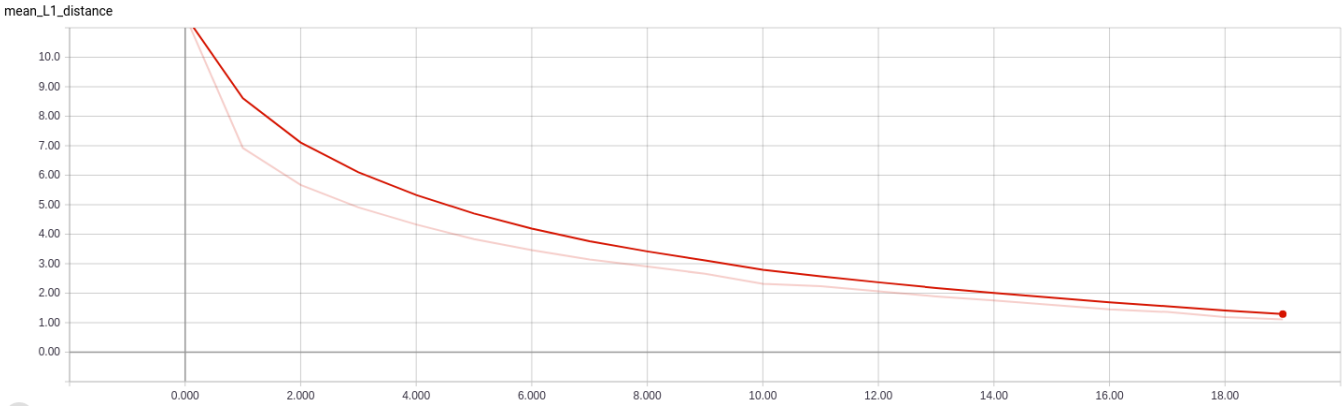

Fig. 6: Mean L1 distance for the training set.

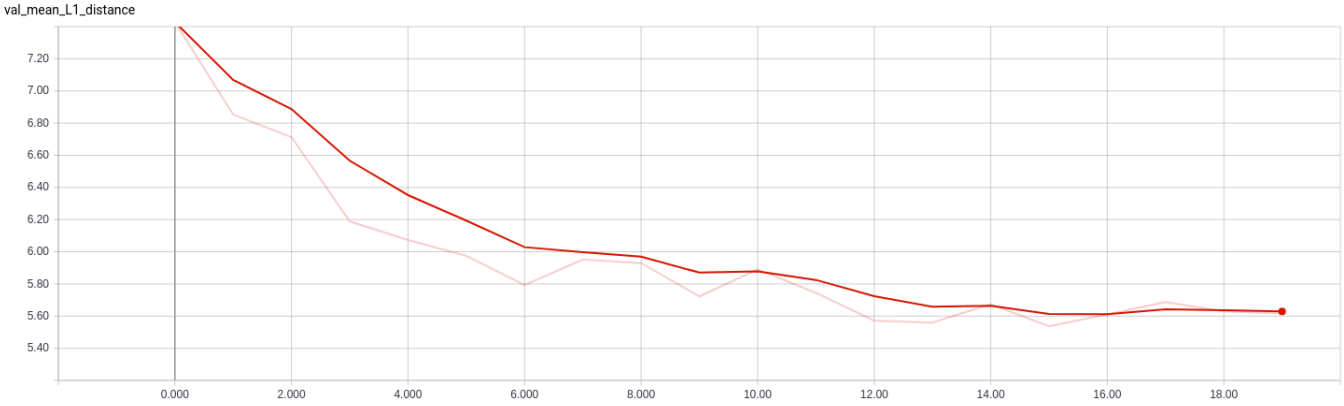

Fig. 7: Mean L1 distance for the validation set.

For the python implementation of our architecture, we used the VGG_face model from [3] that is pre-trained on the Oxford Face dataset [2]. We used Keras to implement our models architectures and to perform the training using TensorFlow on Tesla K80 GPUs. We experiment between two different optimizers ADAM (Learning Rate = e-5) and SGD (Learning Rate = e-4). For each model VGG and RESnet50 pre-trained on the face dataset and VGG and Resnet50 pre-trained on imagenet dataset, we run different experiments changing the optimization, regularization on and off, regularization parameter and batch size (64 and 128). All 30 experiments were run for 30 epochs and were programmed for early stoppage if there were no improvements in the L1 validation accuracy over continuous 5 epochs. As expected, the VGG and RESnet50 pre-trained on the imagenet dataset fail miserably on the yearbook dataset. VGG16 achieved a mean L1 distance of 15 years and the RESnet50 achieved 28 years, which is nowhere near the best achieved by VGG16 pre-trained on the face dataset. This shows that using architectures pre-trained on specific domains related to the dataset achieve better performance over any other well trained architecture unrelated to the dataset.

Fig. 7 shows that the mean L1 validation was around 5.333 which is the best accuracy achieved by the VGG16 architecture pre-trained on the Oxford Face dataset. Fig. 8 shows the training loss decreases over subsequent epochs. Fig. 9 shows the top 5 validation categorical accuracy increases over the epochs. We also plotted intermediate activations to to visualize the network, as shown in Fig. 3. We also generate heat map of class activations for the particular test example shown below, which gives us more inference on what the Deep Net has learnt from the test sets, as shown in Fig. 4 and 5.

A Dropout layer of value 0.5 is added between the last layer and the fc layer to help in overfitting. During the training phase, we perform batch normalization that makes the training robust to hyperparameter tuning. The normalization process is not just applied to the first layer, but also to all the deep layers within the neural network. During the hyperparameter tuning phase, we tried a bunch of learning rates $\alpha = 0.1, 0.001, 0.0001$ and found that 0.001 yielded the best mean L1 distance for the validation dataset.

C. Finetuning Emotion Recognition CNNs

In the paper accompanying the release of Years to Age dataset, Ginosar et.al [4] present an analysis on how expressions of the subjects changed over the course of years in the photographs, noting that smile was an important factor to have varied significantly across years. We hoped to exploit this observation by fine-tuning a 7-way emotion detection CNN developed by Albanie et.al [5]. This network was trained over a data-set of annotated *Deal or No Deal* video frames. The network classifies each image into representing one of 7 emotions: *Anger, Disgust, Fear, Sadness, Happiness, Neutral, Surprise*.

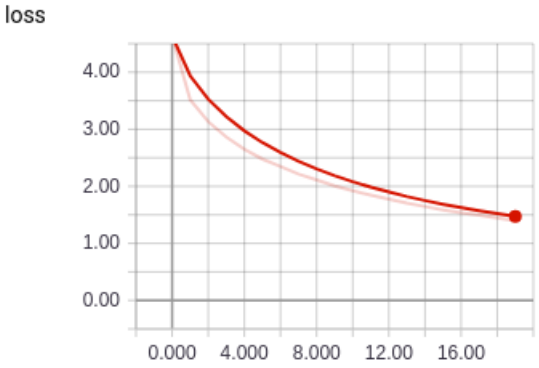

Fig. 8: Training loss for the best VGG16 model.

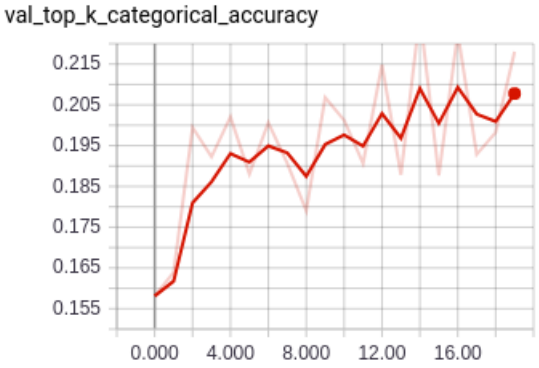

Fig. 9: Top 5 validation accuracy for the validation dataset.

We fine-tuned the network in 2 ways. In the first, simpler way, we replaced the final fully connected layer of dimensions $(512, 7)$ by our 104 way year-classifying $(512, 104)$ layer. In the second approach, we fine-tuned the entire network. Both networks results are summarized in the following table.

In both cases, we added a dropout of 0.5 in the last *fc (fully connected)* to prevent the network from over-fitting, which was predominant in the beginning.

TABLE II: Emotion Net Fine Tuning Validation Results

Config	T-10	T-5	T-3	Accuracy	L1
Full Network	54.3%	30.7%	19.3%	4.5%	6.722
FC-Only	49.4%	27.7%	16.4%	4.2%	**5.095**

We noticed that retraining the full network overfit the data, sometimes resulting in a 100% accuracy on the training set. We use early stopping, L2 regularization and dropout to avoid this over fitting. Also, one can notice that despite the higher accuracy of the fully-fine tuned network, the L1 distance is lesser for the FC-Layer fine-tuned network. This may be again be attributed to over-fitting, while the second network generalized well enough to all years the entire set in-order actually reduce the L1 distance without significant drop in accuracy.

III. CONCLUSIONS

From our repeated experiments, we conclude that the VGG16 architecture pre-trained on the Oxford face dataset is better than the RESnet50 trained on Oxford face or imagenet dataset. We also noticed in the training phase that increasing the learning rate improves the mean L1 accuracy of the validation samples. This could be because when the learning rate is too less, the model learns too much of the data and is not able to generalize well on the validation dataset.

ACKNOWLEDGMENT

We would like to acknowledge the professor Philipp Krahenbuhl for his time and teaching in the seminars. Special thanks to our peers in the class for creating a condusive atmosphere for deep learning study.

REFERENCES

[1] A century of portraits: A visual historical record of american high school yearbooks.

[2] http://www.robots.ox.ac.uk/~vgg/data/vgg_face/

[3] https://github.com/rcmalli/keras-vggface

[4] http://people.eecs.berkeley.edu/~shiry/projects/yearbooks/yearbooks.html

[5] https://arxiv.org/pdf/1706.01509.pdf