CSC2516 Project Report

Karthik Raja Kalaiselvi Bhaskar
Department of Electrical and Computer Engineering
University of Toronto
karthikraja.kalaiselvibhaskar@mail.utoronto.ca

Manpreet Singh Takkar
Department of Electrical and Computer Engineering
University of Toronto
manpreet.takkar@mail.utoronto.ca

Nishank Thakrar
Department of Electrical and Computer Engineering
University of Toronto
nishank.thakrar@mail.utoronto.ca

April 26, 2019

Abstract

A multi-task learning convolutional neural network for the purpose of performing landmark localization and other correlated tasks is studied and analysed in this project. A different and more challenging task around landmark localization than the one implemented originally is studied using a HyperFace architecture. It is seen that having a multi-task CNN helps improving the accuracy among the correlated tasks. A lot of tasks in Computer Vision are actually correlated and thus architectures like these can prove to be very useful. We present extensive results on the training of multitask network in different scenarios and the complexities in training them. Though our model is modestly trained on a dataset that is new for HyperFace, it is found to give reasonably good results.

1 Introduction:

Face Detection is a problem that has been studied for a long time now. The task however becomes more challenging when the image being fed is no longer an image of frontal face but has face aligned at some angle or when some part of the face is occluded. Another challenging task is to detect landmark points with high local accuracy. Landmark points are basically the corner points on face occurring in a particular sequence and which can be used to describe a face. Some of the computer vision algorithms make use of these landmark points to determine the pose of human face. Thus it would not be wrong to say that these tasks of face detection, landmark localisation and pose estimation are correlated. It would make sense to study these problems as a single correlated problem rather than separate problems. This was the focus point of our project, i.e. to study the training of neural network on these correlated tasks as a single problem.

Among the tasks of face detection, landmark localization and pose estimation, the primary focus is to get robust landmark detections. Using deep CNN allow sharing of weights and between these 3 tasks and thus gives a chance to look at these problems with a completely new viewpoint. The landmark detection is an important component in several applications such as gesture recognition, detection of emotions and face recognition. To perform these tasks efficiently, robust localization of landmark points is necessary. Factors such as occlusion and large pose variation of face hinders the accurate detection of these points.

For this project, the implementation of a CNN which is both deep and wide, and exploits the correlation between the tasks in hand is studied. The network implemented is known by the name of HyperFace. The lower layers of the network implement low level feature detections and are shared between different tasks. The fully connected layers are branch out to learn perform different tasks. The feature detection layers and the classification layers are connected by a bottleneck layer in the middle. The training to learn different tasks is done using multiple loss functions.

2 HyperFace:

2.1 Architecture:

The initial layers which are responsible for feature description, follow the architecture of AlexNet. Layer 1, 3 and 5 of this AlexNet architecture are fused using feature fusion techniques. This is done by adding layers 1a and 3a to the network to bring the features to a common subspace i.e. each to have dimensions of 6 x 6 x 256. These are then concatenated to 6 x 6 x 768-dimension feature map. Feature fusion helps to bring semantically rich features together as we are now convolving the outputs from alternate layers. The concatenated layers are convoluted using a kernel of 1 x 1 to reduce its dimensionality to 6 x 6 x 192. This layer is followed by a fully connected convolutional layer generating a 3072-dimensional feature vector. After this point, the network is split into different branches for performing different tasks. Each branch is a combination of two fully connected layers. The first fully connected layer is followed by a ReLU activation. To learn the weights of the network, task specific loss functions

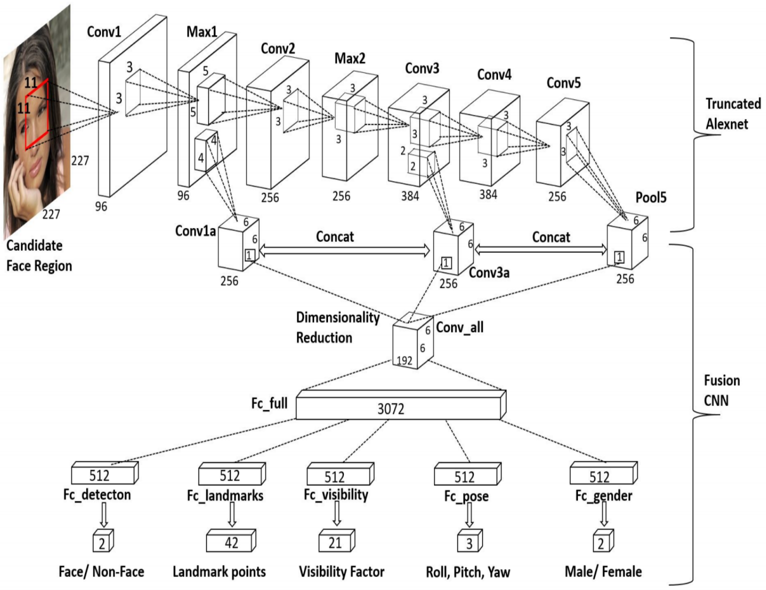

Figure 1: Network architecture for HyperFace [12]

are used. The network architecture for original HyperFace implementation is shown in Figure 1.

2.2 Training:

The training is performed using weighted sum of loss for different tasks. For face detection, softmax loss function used is described by the equation below:

$$log_D = -(1-l).log(1-p) - l.log(p)$$

For landmark localization, the loss function used is mean squared error. The equation describing the loss is given below:

$$loss_L = 1\frac{1}{2N}\sum_{i=1}^{N}(\hat{x_i} - a_i)^2 + (\hat{y_i} - b_i)^2$$

the landmark points are mapped to the range [0,1] using the formulae below:

$$(a_i, b_i) = \left(\frac{x_i - x}{w}, \frac{y_i - y}{h}\right)$$

For pose estimation, the loss used is again mean squared error:

$$loss_P = \frac{(\hat{p_1} - p_1)^2 + (\hat{p_2} - p_2)^2 + (\hat{p_3} - p_3)^2}{3}$$

The way pose labels are generated give approximate results as described later in the report. The algorithm to generate labels for pose is described in [3]

For overall loss, we use a weighted sum of these losses and the weights for face loss, landmarks loss and pose loss is in ratio 1:5:5.

3 Related Work:

The idea to perform multitask learning for the problems of face detection, landmark localization and pose estimation was given in [9], [10]. Their method is based on using shared pool of parts such that every landmark point is modeled as a part and uses global mixtures to capture topological changes. Multi-task using neural networks is performed in [12], [4]. Problems such as pose estimation and action detection have been studied in the scope of multi-task leaning in [5] which is done using CNN and was claimed to be giving state-of-the-art results. Face detection using CNN based architecture has been proposed in [11]. These CNN based architectures overcome the limitation of variation in illumination and pose as faced by a traditional method such as HOG. Landmark Localization using CNN based architectures has been used in [6] and these have been found to give accuracy comparable to previous regression based methods studied in [18]. A survey of pose estimation methods for human face has been provided in [8].

Multi-task learning for problems related to human face has been studied in [12]. It does not mention anything about the impact of training for individual task such as landmark localization on the other tasks other than the fact the other tasks help improve the accuracy. The converging rate of loss on the other hand is an important feature to be compared between the single-task networks and multi-task networks. One of the focus points of this project is to study the correlation of these tasks and complexities that arise due to multi-task learning. Also, an attempt has been made for prediction of 68 landmark points instead of 21 points as done in [12].

4 Comparisons and Demonstrations:

For the first attempt, the aim was to create a dataset of images with annotations of 68 landmark points, and values for roll, pitch and yaw for human faces. The dataset was prepared by combining various datasets from [7][13][14][15][1][2][16][17]. All of these have images and annotations of 68 landmark points on face. However, the value of pose for these faces is not available. This was added to the dataset by making use of Posit algorithm which tries to determine the pose of human face based on six landmark points (left corner of left eye, right corner of right eye, two corner points on lips, chin and tip of nose) and their approximate position in 3-D coordinate system. The pose estimates generated using this approach is not the most accurate one but suffices for the purpose of this project.

To generate data for face detection part, selective search algorithm was used to give different regions from an image. If the overlap of this region with a rectangular box tightly fit around face is more than 50%, it is classified as face. If the overlap is less than 10%, it is classified as not-face. In the HyperFace paper however, 30% overlap is selected to classify not-face images. This however was found to give highly biased results as number of non-face images turned out to be much larger than number of face images. From here on in this report, this dataset is referred to as 'GenFace' dataset.

4.1 Experiment 1:

For the first attempt, the weights of the network are initialized randomly. A network initialized on these weights was found to be overfitting very quickly, which was expected. The reason for this is the limited amount of dataset available (12,000 images), as opposed to the huge amount of data used for training AlexNet.

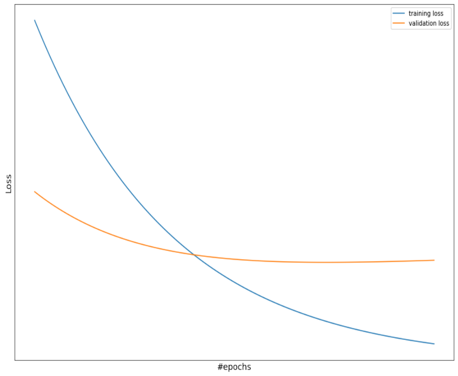

Figure 2: Comparison of validation loss and training loss for model trained on random initialisation of weights. The loss plotted here is a log normalized. The actual validation loss is much higher than the training loss. The figure just demonstrates how the validation loss stops reducing quite early.

4.2 Experiment 2(a):

For the next attempt, we decided to divide the training process into two steps. First, a Fast-RCNN model is trained to make prediction for face detection only, i.e. a binary classification of whether the image is a face or not. The dataset used for this purpose was GenFace. The initialization of weights for the first five convolutional layers was done using pre-trained AlexNet weights as with random initialization of weights, it overfits quickly. Then, the weights from this model were used to initialize weights for the full HyperFace architecture. The training was done using AFLW dataset (21 landmark points). The accuracy for landmark localization of this experiment is shown by the graph in Figure 3.

4.3 Experiment 2(b):

Another model was trained in which Fast-RCNN training was done with another custom dataset in which non-face images were chosen from the CIFAR-10 dataset. The rest of the training was similar to the previous section with the full HyperFace architecture being used. The predictions performed for face detection part turn out to be qualitatively less accurate. Also, the accuracy for landmark localization reduces by around 2%. This can be observed from Figure 3. The reason for this is that to perform other tasks with robustness, it must learn to classify an image is a non face image that has a very small portion of face (say less than 30%).

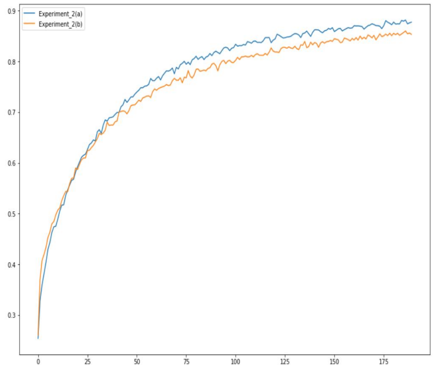

Figure 3: Accuracy of Landmark Localization for Exp 2(a) and Exp 2(b).

4.4 Experiment 3:

Next, we decided to drop the task of detecting pose and gender and train the network following the same procedure as of experiment 2(a).

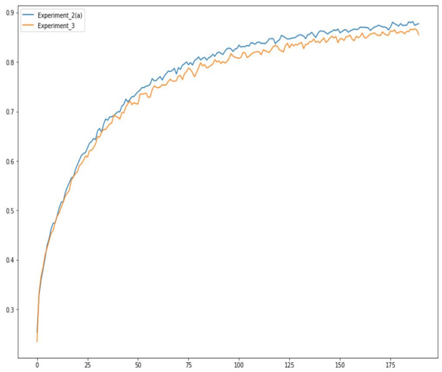

Figure 4: Accuracy of Landmark Localization for Exp 3 and Exp 2(a).

As we can see from the comparison of accuracy of landmark localization in the graph from Figure 4 that the accuracy drops by a small fraction i.e. 1%. It is then reasonable to assume that training this multi-task network by dropping some of the tasks is not a bad idea. However, training it this way doesn't reduce the number of epochs it takes for the weights to converge as the loss becomes constant after almost the same number of epochs in both the cases.

4.5 Experiment 4:

The observations from experiment 3 suggest that we can drop a few tasks from the original HyperFace network and it doesn't affect the accuracy of landmark localization much. Thus it gives us green signal for us to train HyperFace network on GenFace dataset. This dataset doesn't have the labels for gender. Also, pose is a more informative feature for human interpretation as well, so we decided to remove visibility of landmark points instead of the pose for this experiment. Also, for this experiment we trained the network using only the 'face' images from GenFace dataset.

An interesting thing to note from these transitions, is that through a lot of training epochs, the position of the landmark points relative to each other remains rather constant. Even if it changes, the changes are similar in proportion.

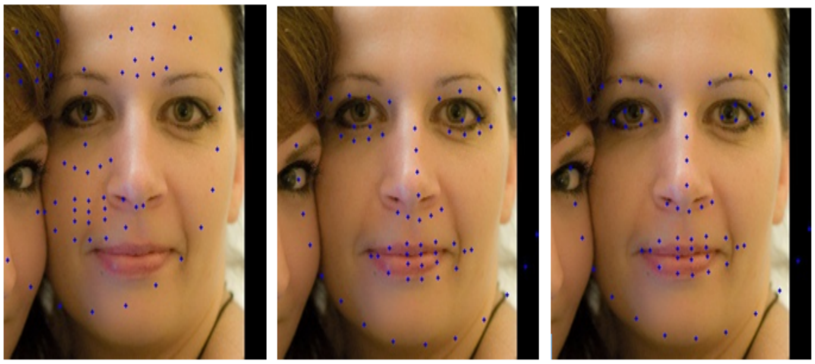

Figure 5: Transition of prediction of landmark points over the training process.

4.6 Experiment 5:

This is more of an additional experiment, to understand how learning for one task impacts the output for other tasks. In this, the network was trained initially to give reasonable results. Later we put the network to train only for landmark points. However, the branches for other auxiliary tasks were still kept connected. They did not contribute to the weight updates, but their losses were still logged. To our surprise, the loss for pose estimation was found to be dropping (at much less rate however) along with the landmark localization loss. This is shown in the graphs from of Figure 6.

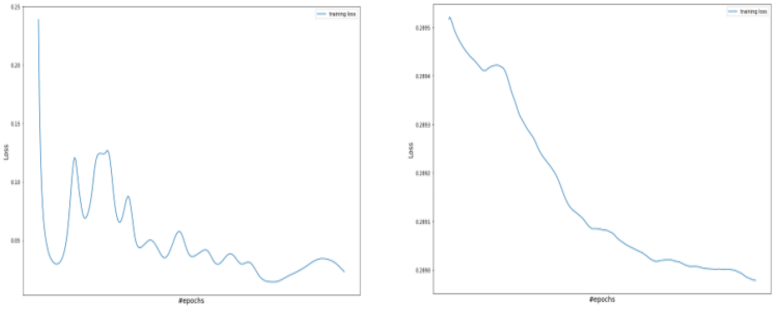

Figure 6: Loss of Landmark Localization for Exp 5. The graph on left shows the reduction in loss of landmark localization and the graph on the right shows reduction in the loss of pose estimation. The loss in second graph falls at a lower pace (Please note the scale in the graphs).

5 Limitations:

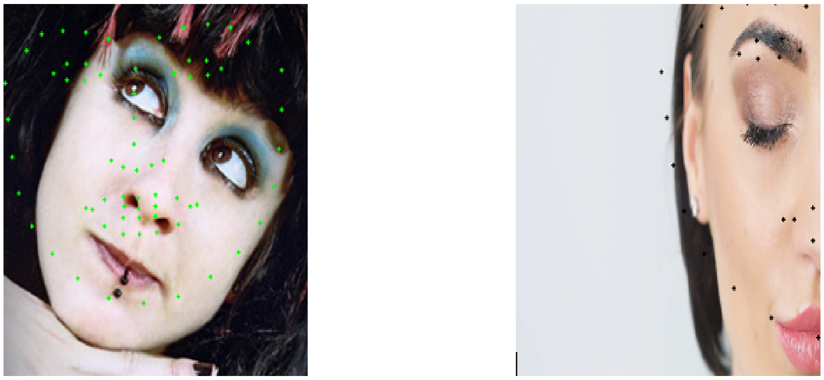

Figure 7: Model Predictions on challenging images.

Given above are two cases where the model fails to generalize. The first one being when faces in the image appear at tilted angles or awkward positions. The model expects the faces to be at the at a straighter angle. Second case is when the face in the image is occluded. Here the model does better than the first case, but the qualitative predictions are less accurate. This can be assumed to be a general trend, wherein the model predicts the landmark points better for images which have faces at the center, and the pose is relatively straight.

We believe both these limitations are owing to the fact that the data used to train the model lacks the latent variability to capture the inherent distribution of landmark points relative to the pixel values.

6 Conclusion and Future Scope:

In this project, we study and analyse the implementation of a multi-task CNN over the problems of detection, landmark localization and pose estimation of human face. The network architecture used is HyperFace. The training is done for various publicly available dataset having annotations for 68 landmark points on human face. The experiments show that having a multi-task network help improve the accuracy for landmark localization but the improvement is not too high and that is expected as well. The network with random initialization of weights overfits early. The network with five tasks takes almost the same number of epochs as the network trained for three tasks which supports the use of multi-task learning wherever possible. From the observations of the output of landmark points on more challenging images such as the ones in which face is highly tilted or in which just only a portion of face is visible, the model seems to fail. Training with a dataset that is suited to fight these limitations would surely help.

Training the model with a dataset that has higher variation in the pose of these images and different portions and different percent of face occluded, is what we propose for future task. Another possible extension of this project could be to train convolutional neural networks to improve the accuracy of landmark points locally. What we suggest is to train CNN that would take a small patch (say 7 x 7) of image around the predicted point as input and try to find a more accurate position for that landmark point. Once we can train for highly accurate landmark points, these could be used to perform tasks such as emotion analysis and gesture recognition.

7 List of contributions:

As far as the contributions of people for this project are concerned, almost all the work has been done together in group meetings and sessions. Since our project structure is sequential in nature, building on steps one after the other, it was not really possible to assign individual tasks for people to work on. It has mostly been about a number of brainstorming sessions for coding and debugging, and making decisions on how to proceed with the project. Essentially the individual contributions for the project is divided equally among the three authors.

References

[1] Peter N Belhumeur, David W Jacobs, David J Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2930–2940, 2013.

[2] Grigoris G Chrysos, Epameinondas Antonakos, Stefanos Zafeiriou, and Patrick Snape. Offline deformable face tracking in arbitrary videos. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 1–9, 2015.

[3] Daniel F Dementhon and Larry S Davis. Model-based object pose in 25 lines of code. *International journal of computer vision*, 15(1-2):123–141, 1995.

[4] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015.

[5] Georgia Gkioxari, Bharath Hariharan, Ross Girshick, and Jitendra Malik. R-cnns for pose estimation and action detection. *arXiv preprint arXiv:1406.5212*, 2014.

[6] Amit Kumar, Rajeev Ranjan, Vishal Patel, and Rama Chellappa. Face alignment by local deep descriptor regression. *arXiv preprint arXiv:1601.07950*, 2016.

[7] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *European conference on computer vision*, pages 679–692. Springer, 2012.

[8] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head pose estimation in computer vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):607–626, 2009.

[9] Deva Ramanan and Xiangxin Zhu. Face detection, pose estimation, and landmark localization in the wild. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2879–2886. Citeseer, 2012.

[10] Deva Ramanan and Xiangxin Zhu. Face dpl face detection, pose estimation, and landmark localization in the wild. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2879–2886. Citeseer, 2015.

[11] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. A deep pyramid deformable part model for face detection. In *2015 IEEE 7th international conference on biometrics theory, applications and systems (BTAS)*, pages 1–8. IEEE, 2015.

[12] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):121–135, 2019.

[13] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: Database and results. *Image and vision computing*, 47:3–18, 2016.

[14] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013.

[15] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. A semi-automatic methodology for facial landmark annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 896–903, 2013.

[16] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of the IEEE*

International Conference on Computer Vision Workshops, pages 50–58, 2015.

[17] Georgios Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3659–3667, 2015.

[18] Xuehan Xiong and Fernando De la Torre. Global supervised descent method. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2664–2673, 2015.