Facial Landmarks Detection with ResNet18

Matjaž Zupančič
CS231n 22/23, Stanford
mm1706@student.uni-lj.si

***Abstract*—In this work, we develop a facial landmarks detection model using a ResNet18 neural network that has been previously trained for image classification on the ImageNet dataset. Our approach involves framing the landmark detection task as a regression problem and minimizing the Mean Squared Error (MSE) loss across 194 landmarks. The ResNet18 model serves as the basis for our facial landmarks detection model, and its performance on the ImageNet classification task serves as a starting point for fine-tuning on the landmark detection task. Our based model achieved a MSE loss of 27.84 on the test set.**

I. Introduction

Facial landmark detection is a key problem in computer vision that involves identifying unique features of a person's face, such as the eyes, nose, mouth and yaw line. This technology has a wide range of applications, including face recognition, augmented reality, animation, face detection, emotion detection, and medical imaging. In this paper, we will train a CNN based facial landmarks detector.

II. Related Work

Facial landmark detection has been an active research area in computer vision for many years. Early work in this field focused on developing hand-crafted features and classifiers, such as the Active Appearance Model (AAM) proposed by Cootes et al. [1]. In recent years, there has been a shift towards the use of deep learning techniques for facial landmark detection.

One of the first deep learning approaches for facial landmark detection was the Multi-Task Convolutional Neural Network (MT-CNN) proposed by Zhang et al. [2]. This model uses a multi-task learning framework to jointly predict facial landmarks and classify facial attributes. Another popular approach is the Stacked Hourglass Network (SHN) proposed by Newell et al. [3], which uses a stacked hourglass architecture to learn a high-resolution representation of the face.

Other notable approaches for facial landmark detection include the DenseReg network proposed by Güçlütürk et al. [4], which uses dense regression to predict the locations of facial landmarks, and the Face Alignment by Deep Regression (FAN) model proposed by Bulat et al. [5], which uses a combination of convolutional and recurrent layers to learn a mapping from image pixels to facial landmarks.

III. Methodology

To train our network we used the Helen facial landmarks dataset ([6]). The dataset consists of 2000 training images and 330 test images. Images are of varying resolution and are annotated with 196 landmarks. We start by cropping the annotated faces from the original images. We resize the images and corresponding landmarks to a resolution of 300×300. Figure 1 shows a sample from the Helen dataset. The starting point to our optimization is a ResNet18 ([7]) neural network that has been previously trained for image classification on the ImageNet dataset ([8]). We swap the final fully connected layer and randomly reinitialize it. We frame the landmark detection task as a regression problem and minimize the Mean Squared Error (MSE) loss across 194 landmarks. To be more concrete we adopt the DeepPose architecture ([9]) shown on figure 2. Before training the model, images are re-scaled to a resolution of 256×256 and normalized by the mean and standard deviation of the ImageNet dataset. We perform hyper-parameter optimization to find optimal values for learning rate and weight decay parameter of the Adam optimizer. We randomly search for optimal values in the log space by performing 10 epochs and selecting the values for learning rate and weight decay that yield the minimal validation loss. We generate additional training samples, by introducing the following transformations: randomly rotate image and landmarks, randomly crop image and randomly adjust pixel intensities. After data augmentation we are left with 8000 samples used for training and validation. Finally we train the network for 100 epochs and back-propagate through the entire ResNet18 model. We use the Adam optimizer with previously estimated best learning rate and weight decay.

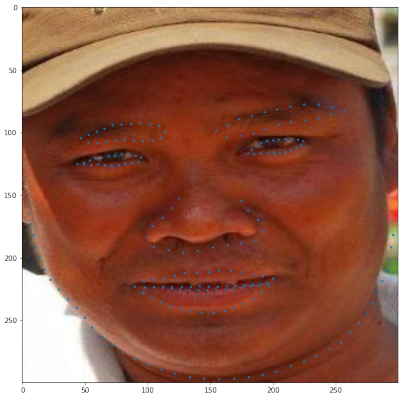

Fig. 1. Example image and corresponding landmarks from the Helen Database

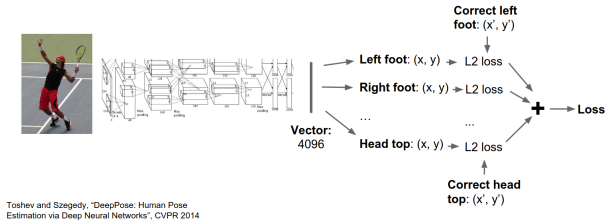

Fig. 2. We adopt the DeepPose architecture ([9]). The network on the image is in our case ResNet18 with a fully connected layer of shape 88×1

IV. Experiments

We evaluate the model performance by computing the MSE loss on the Helen dataset ([6]), using the 330 provided test samples. We display the top 3 detections and the worst 3 detections. We display the distribution of MSE Losses over all test images. In order to better understand which pixels network finds the most important we compute the activation map $\frac{\partial \mathcal{L}}{\partial I_{ij}}$ for each pixel I_{ij}. Activation maps are shown for the 3 best and 3 worst detections.

V. Results and Discussion

A. Results

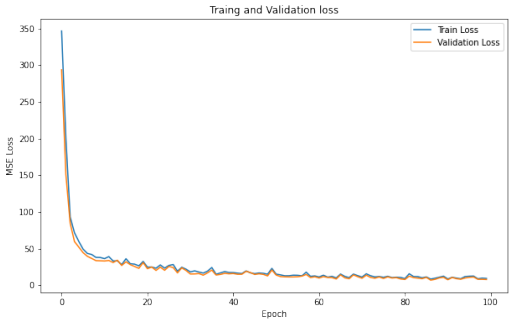

Fig. 3. Training and Validation MSE Loss as a function of epochs

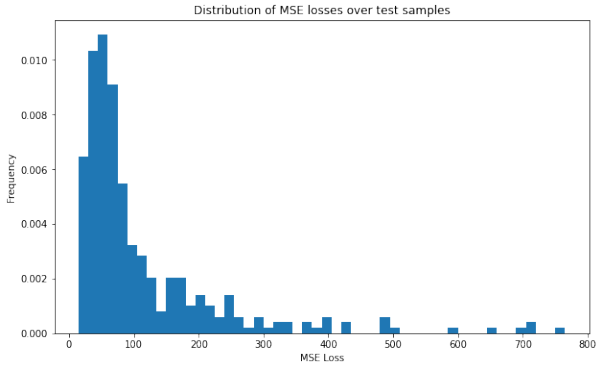

Fig. 4. MSE distribution over all test images

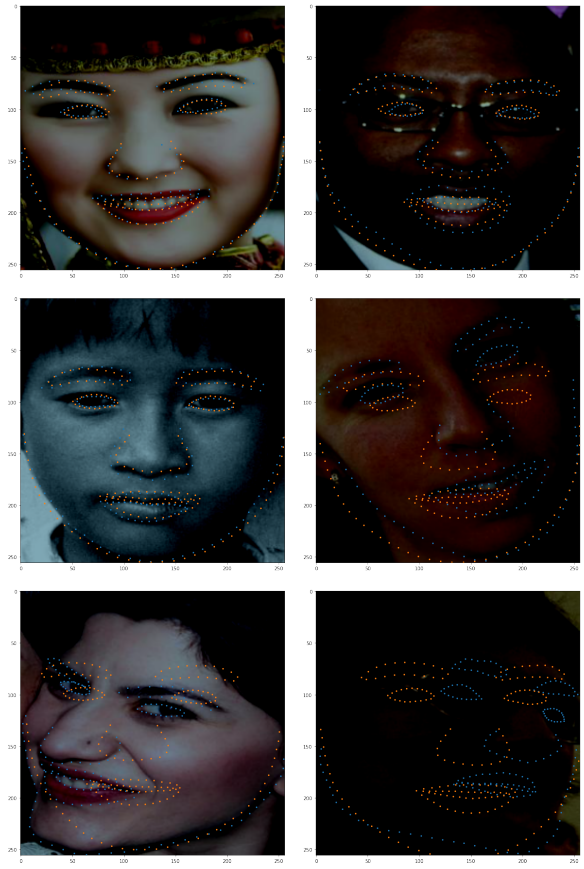

Fig. 5. A collection of 3 best and 3 worst detections

B. Discussion

Figure 3 shows training and validation MSE Loss as a function of epochs. The results showed that the model learned quickly during the first 20 epochs, but the loss reduction slowed down afterwards. The lowest validation loss of 8.00 was achieved in the 100th epoch. When tested on 330 samples, the mean MSE was 27.84. Figure 4 shows the distribution of losses on the test set. We found that there were some outliers with MSE values above 200. Figure 5 displays a selection of the 3 best and 3 worst detections, illustrating that even the best detections were not perfect. The worst detections showed that the network simply placed an average face in the center of the image. Figure 6 shows the activation maps corresponding to the detections in Figure 5. Overall, the activations were not strong, with the exception of the first image in the grid, which clearly highlighted the face. The activations for the worst detections were more diffuse.

VI. Conclusion

In this work, a ResNet18 model was trained to perform facial landmarks detection using the Helen dataset. The model was then evaluated on 330 test images, yielding an MSE of 27.84. It was found that the best validation loss was 8.00. The gap between test and validation set suggest that with more data

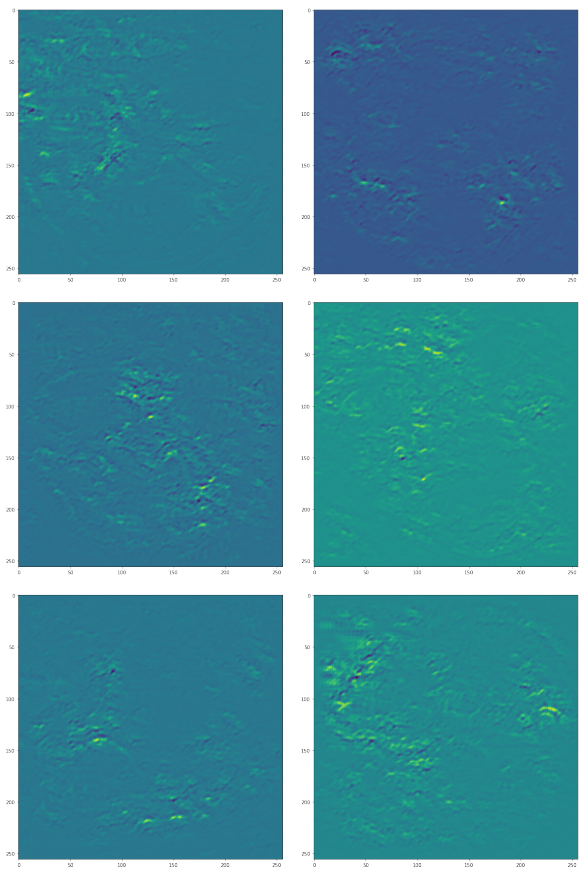

Fig. 6. A collection of activation maps corresponding to the 3 best and 3 worst detections

augmentation and more training time the model performance could be further improved. The distribution of losses on the test images was examined, and a selection of good and poor detections was shown. The activation maps for the best and worst detections were also analyzed. As potential future work, we could consider generating more training data using similar approaches, or even using generative adversarial networks (GANs).

REFERENCES

[1] T. F. Cootes, G. J. Edwards, C. J. Taylor, and D. H. Cooper, "Active appearance models," *International journal of computer vision*, vol. 61, no. 2, pp. 137–154, 2005.

[2] Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[3] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*. Springer, 2016, pp. 483–499.

[4] R. Alp Guler, G. Trigeorgis, E. Antonakos, P. Snape, S. Zafeiriou, and I. Kokkinos, "Densereg: Fully convolutional dense shape regression in-the-wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6799–6808.

[5] A. Bulat, G. Tzimiropoulos, and S. Zafeiriou, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *International Conference on Computer Vision*. Springer, 2017, pp. 1049–1068.

[6] N. Le, J. Susskind, and C. Fowlkes, "Interactive facial feature localization," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3069–3076.

[7] X. Yu and S.-H. Wang, "Abnormality diagnosis in mammograms by transfer learning based on resnet18," *Fundamenta Informaticae*, vol. 168, no. 2-4, pp. 219–230, 2019.

[8]

[9] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1653–1660.