ECSE 4965: Final Project

Name: Li Dong
RIN: 661243168

May 8, 2017

Learning strategy

In this project, the network architecture is replicated from [1]. Although the data is downsampled from 224×224 to 64×64, the same architecture with the same hyperparameters still achieves comparable precision. Visualization of the architecture can be seen from the following figures. It's possible that the testing data is not big and diverse enough and too similar to the training data. One improvement over [1] in this project is that the learning rate is adaptive. Specifically, with a initial learning rate 1e-3, it is bisected if the average testing precision over the next five epochs is larger than that of the previous five epochs or decrease less than 5e-3. The loop terminates when learning rate is less than 1e-9 or the max epoch number is reached. All the data are loaded into the RAM at the beginning of the code and a batch of 250 is used in each learning iteration, which means one training epoch contains 192 iterations considering 48,000 data points. Loss function value, train and testing precision are evaluated in each epoch. One thing to note in current model is that the input image data are **NOT normalized** as in most cases and the **raw image data** are fed into the model, since substantial overfitting was observed starting from the very first few epochs when using the normalized data. One explanation to this is that the network is too large for this amount of data and normalizing the data makes the model too easy to learn, which leads to fast overfitting.

Overfitting

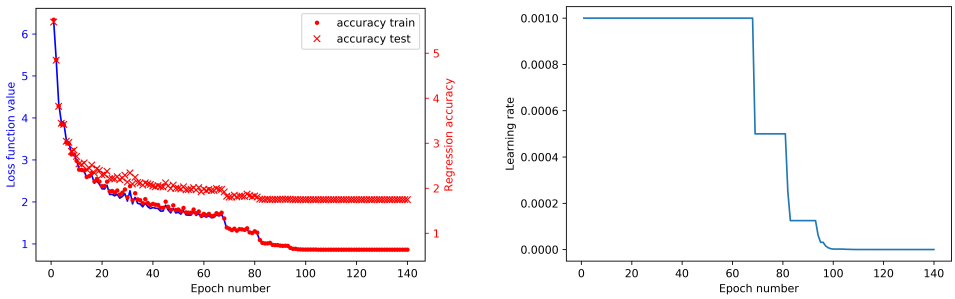
In order to deal with overfitting, a few treatments are explored. As in [1], weight decay is applied to all training variables and the penalization parameter is fixed to 5e-4. In addition to that, data augmentation by altering RGB and dropout with 50% keep probability for fully connected layers [2] were explored but suboptimal results were observed somehow and thus deserted.

Results

Different pathway combinations are explored, including both eyes + face + face grid (E+E+F+FG), both eyes + face grid (E+E+FG) and face + face grid (F+FG). Convergence plots (loss function value, training precision and testing precision), learning rate decay and model architecture visualization for E+E+F+FG can be seen in Figures 1 and 2, for E+E+FG Figures 3 and 4, for F+FG Figures 5 and 6. The best testing errors for the three combinations are E+E+F+FG: 1.747 cm, E+E+FG: 1.764 cm and F+FG: 1.703 cm. Since F+FG performs the best, the attached my-model files are saved from it. This is an interesting scenario that even the downsampled images without information from eyes can generate very descent results. This may indicate the head position plays a larger role than people initially thought in gaze capture process.

References

[1] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.



(a) (b)

Figure 1: Convergence plot and learning rate decay for four-path-way learning: E+E+F+FG.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

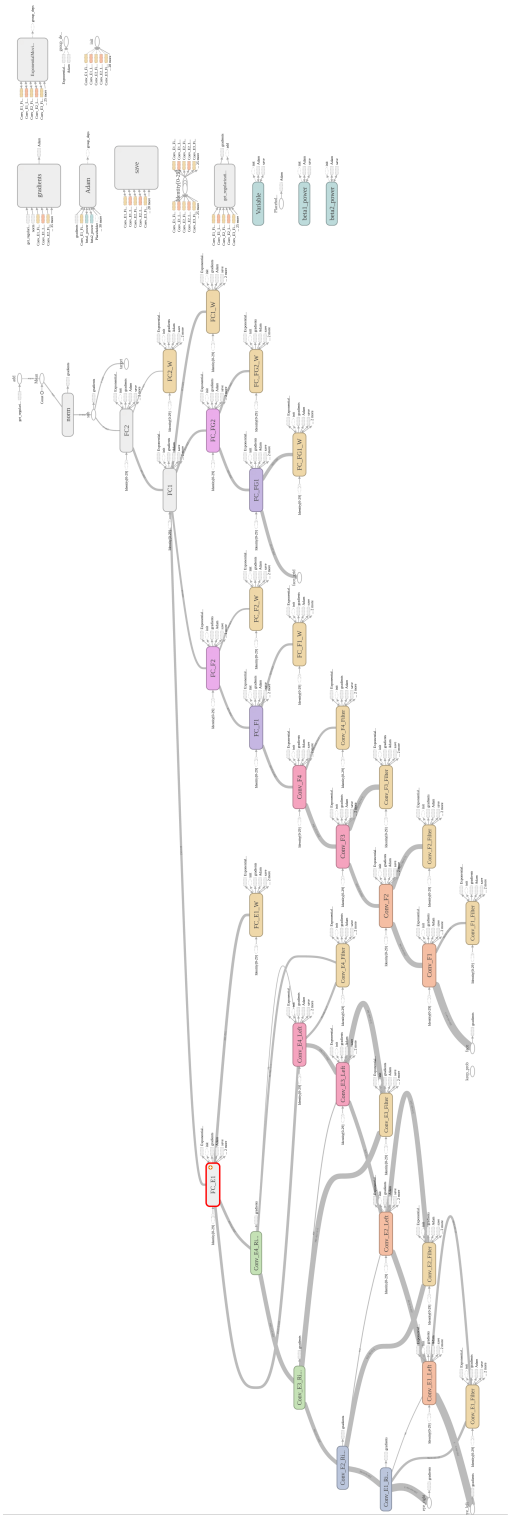
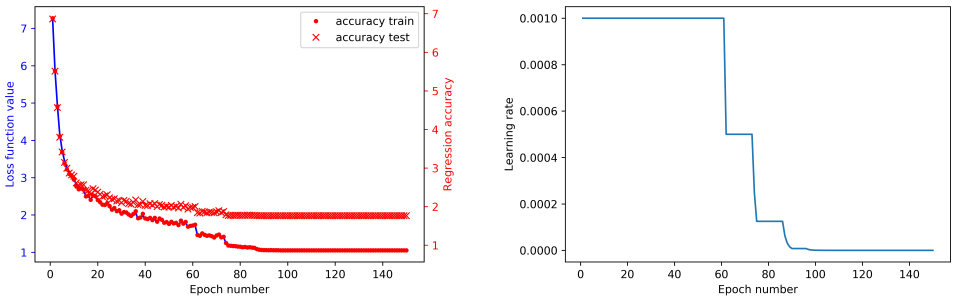


Figure 2: Graph visualization for four-path-way learning: E+E+F+FG.



(a) (b)

Figure 3: Convergence plot and learning rate decay for three-path-way learning: E+E+FG.

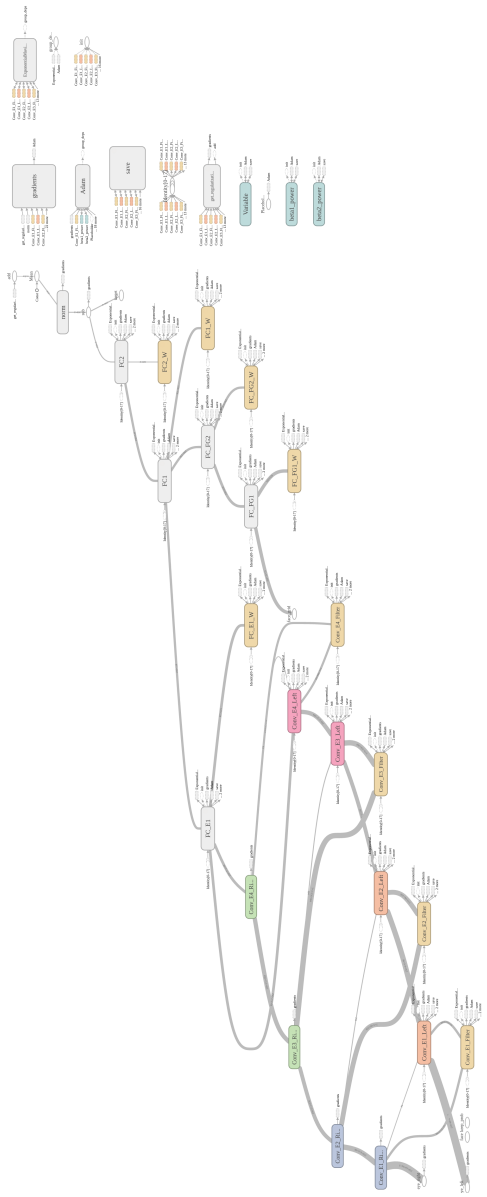
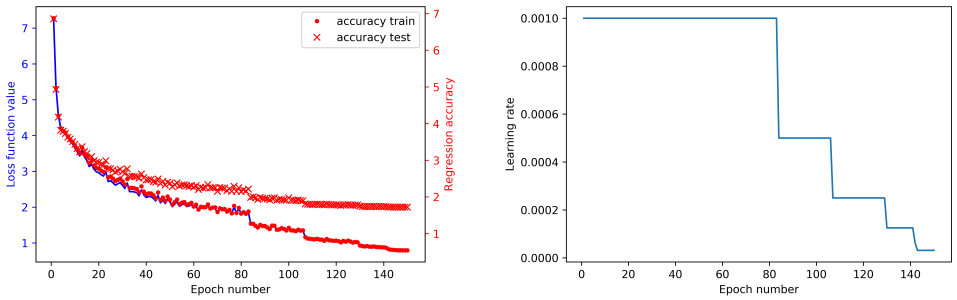


Figure 4: Graph visualization for three-path-way learning: E+E+FG.



(a) (b)

Figure 5: Convergence plot and learning rate decay for two-path-way learning: F+FG.

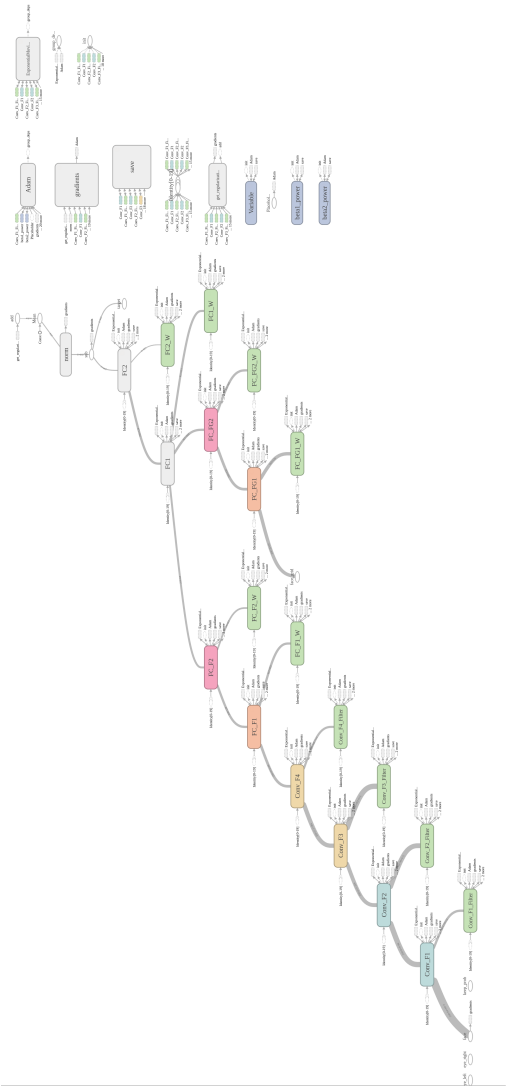


Figure 6: Graph visualization for two-path-way learning: F+FG.