The final architecture is based on DenseNet161

Among all the baseline models (MobileNet, Inception, VGG, and ResNet, etc.), the DenseNet model seemed to achieve the best performance. The input image is first passed to the initial convolution with kernel size of 3, stride of 1, and padding of 1. The input image is down-sampled to 16x16 with 96 channels. I followed strictly to the most complicated architecture with the most parameters without drop out layers, which means the skip connection is conducted as concatenation rather than addition.

The loss functions are a bit complicated. During classification, the softmax loss is applied on 2300 people. During fine tuning, I tried lots of loss functions, including center loss, large-margin softmax and angular softmax function. It turned out that the center loss was easier to train (loss decreased faster than the other two). The center loss is companied by softmax, the model gives a feature embedding of 2208 features, which is fed to the center loss. The fc layer is, however, fed into a softmax loss. The final loss is the weighted sum of the two losses. Center loss is supposed to pull the embedding closer to 4300 centers if the final fc layer has no bias. The verification takes two images and uses the 2208 embedding feature to determine the cosine similarity. The angular softmax is harder to train in the beginning, after 10 epochs, the training seems to saturate. The key philosophy of angular softmax is to use a phi approximation (Taylor Series expansion) to speed up the calculation.

The other models are also tested. ResNet50, for example achieved an accuracy of 79% and an accuracy of 84% with drop out modification. However, I found it more important to have the information from multiple scale of gaussian blurs to captures edges of different sizes.

The training uses SGD as optimizer for better convergence. The learning rate starts with 0.1 and momentum of 0.9 as suggested in the paper. Learning rate drops to 10% at epoch 2,3 and 4. The center loss used SGD at first and switched to Adam (lr = 0.5).