CSEye: A Proposed Solution for Accurate and Accessible One-to-Many Face Verification

Ameer Dharamshi*, Rosie Yuyan Zou*

University of Waterloo, Waterloo, Ontario, Canada
{adharams, y53zou}@edu.uwaterloo.ca

Abstract

Facial verification is a core problem studied by researchers in computer vision. Recently published one-to-one comparison models have successfully achieved accuracy results that surpass the abilities of humans. A natural extension to the one-to-one facial verification problem is a one-to-many classification. In this abstract, we present our exploration of different methods of performing one-to-many facial verification using low-resolution images. The CSEye model introduces a direct comparison between the features extracted from each of the candidate images and the suspect before performing the classification task. Initial experiments using 10-to-1 comparisons of faces from the Labelled Faces of the Wild dataset yield promising results.

Introduction

The CSEye face verification model was conceptualized based on three eventual objectives: the model is easily trainable with limited computing power and data, it retains a threshold value of accuracy for it to be functional, and it is small enough to be stored on mobile devices. These objectives were set because while recent developments in one-to-one facial verification models have demonstrated stunning performance, model portability and practicality are still hurdles preventing broader usage. For instance, the Transferred Deep Feature Fusion model produced a 97.9% true accept rate and 0.001 false positive rate on the IJB-A dataset (Xiong et al. 2018). However, a model of that caliber is unlikely to be practical on mobile devices, or for applications such as 1-to-N face verification using blurry security footage, due to the computing power required, and the large quantity of quality training images needed. Hence CSEye was designed to address the tradeoff between accuracy and practicality. We can describe the face verification problem as follows: given a suspect facial image and N candidate images, determine which image is the match.

* These authors contribute equally to this abstract

Model Architecture

To answer the one-to-many facial verification problem, we propose a three-stage model: feature extraction, comparison and classification. To maintain the integrity of the comparisons, the model employs a weight-sharing paradigm. This ensures that each candidate image receives identical mapping and has the added benefit of drastically reducing the number of trainable parameters and facilitates scaling the model to a larger number of candidates.

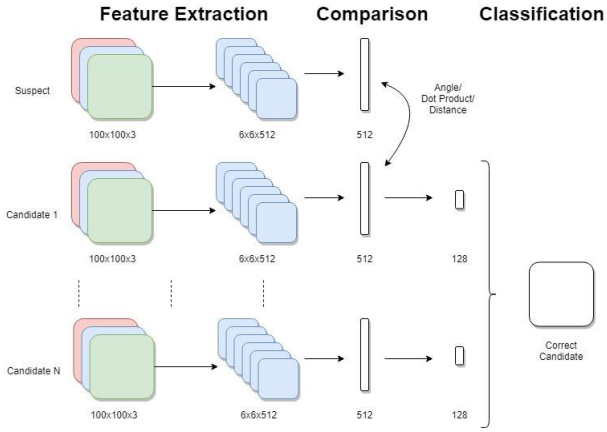

Figure 1: CSEye Model Architecture

The initial stage, feature extraction, is based off a modified VGG19 architecture (Simonyan and Zisserman 2015). To accommodate the lower resolution images, the initial input to the neural network is reduced from the standard (224,224,3) to (100,100,3). Each of the following layers is reduced in dimensions by the appropriate quantity. The second major modification is that the network is abruptly halted after the final convolution layer. The purpose of this stage is simply to extract the same features from each image. In an N-to-1 comparison, each of the N+1 images has its features extracted by the modified VGG19 network, generating N+1 sets of 512 6x6 matrices.

The comparison step is inspired by the paper "Image Question Answering using Convolutional Neural Network with Dynamic Parameter Prediction" which uses values generated by an RNN based on a question to influence an answer (Noh, Seo and Han 2015). In the CSEye model, we intend to use features extracted from the suspect image to directly influence the selection of the candidate. To accomplish this, we compare the results of each specific feature extraction of the suspect to each candidate and then use the results to perform the classification task. We consider three comparison measures for the 6x6 feature map matrices: angle, dot product and distance. If we vectorize the matrices, we consider the angle between the two vectors, the dot product between the two vectors and the sum of the element-wise differences between each vector.

Using the values computed in the comparison step, the model passes each set of comparisons through three dense layers. The output is then merged into a single vector and passed through a softmax layer for the final classification.

Model Testing

To test the validity of the model, we consider a 10-to-1 face verification problem. Datasets are generated from the Labeled Faces in the Wild (LFW). An individual with more than one image in LFW is randomly selected and one image is chosen as a suspect and one as the correct candidate. These images are different to establish ground truth and ensure that the data does not cheat the model. The remaining nine candidates are randomly selected from LFW to complete the sample. Next, the image sets are passed through the truncated VGG in order to obtain the encoding at the 4th Convolution Layer of Block 5, which is used as the argument for the comparisons. Each feature map, represented as a 6x6 matrix, is vectorized before computing the angle, dot product, and distance between the suspect and each of the 10 candidates. A dense network then selects the candidate that is most similar to the suspect. This candidate is identified as the suspect's match.

In our model testing, we have used a training set consisting of 5000 randomly generated samples created using the method discussed above. Of this sample, 10% is set aside as a validation set. We also construct a test set of 1000 randomly generated samples. With kernels initialized using the Glorot Normal function and L2-regularized at 0.01, the model is trained using categorical cross entropy and back-propagation. On the validation set, dot product resulted in a 78% accuracy, whereas angle and distance produced 84% accuracy. On the test set, the angle metric produced the highest accuracy of 86% after being trained for 50 epochs. When considering the overarching objectives to create a more accessible model, we note that each epoch takes less than 2 seconds to train without the use of a GPU.

Following the initial tests, we performed further testing using training and test sets of 10,000 samples each. We proceed with the angle and distance models given their better performance. In this test, we observed improved performance with the angle and distance models achieving 98.74% and 90.35% true positive rates respectively. This improvement is expected as the increased training sample leads to a more generalized final model. With the second-stage test results and overall training time, the model architecture proves that it can be scaled to larger datasets.

Conclusion

In our initial tests, we observed that angle and distance-based models achieve high predictive accuracy. The model architecture also maintains a low number of parameters and low training time, which improves model portability. This does not guarantee that strong performance will extend to 1-to-N comparisons for large N. Further testing using larger datasets beyond LFW with more sophisticated sampling techniques will be required to prove this. Additionally, while the dot product model did not perform well in testing, the viability of the concept is not eliminated. The current model uses a generalized feature extraction network. A natural next step is to train this component specifically to detect subtle features embedded in face images. This would result in more relevant features extracted and ideally would improve performance in all three models.

Some additional aspects to be further investigated are related to model practicality, specifically factors such as image resolution, training time, and response time. These key performance indicators will allow us to better assess the trade-off between practicality and performance.

Acknowledgements

We would like to thank fellow students Kye Xinkai Wei and Glen Chalatov for their contributions. We would also like to thank Professor Ali Ghodsi for his continued advice.

References

Noh, H., Seo, P. and Han, B. 2015. Image Question Answering using Convolutional Neural Network with Dynamic Parameter Prediction. arXiv preprint arXiv:1511.05756.

Simonyan, K. and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In ICLR 2015.

Taigman, Y., Yang, M., Ranzato, M. and Wolf, L. 2014. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In CVPR 2014.

Xiong, L., Karlekar, J., Zhao, J., Feng, J., Pranata, S. and Shen, S. 2018. A Good Practice Towards Top Performance of Face Recognition: Transferred Deep Feature Fusion. arXiv preprint arXiv:1704.00438.