FACE RECOGNITION WITH ONE-SHOT LEARNING USING A SIAMESE CNN

Ross Erskine (ppxre1)

University of Nottingham
School of Computer Science,
Jubilee Campus,
Nottingham NG8 1BB, UK

ABSTRACT

Face recognition with one-shot learning can be an important task such as border-control, where a certain individual needs to be found and there is little known identity or images of the person, or a employee confirming his/her attendance via a register at their place of work. This task creates many challenges such as illumination, pose, expression and occlusion. In this paper we aim to make a Siamese Convolutional neural network (SCNN) which can recognise faces with only one image of each task *One-shot learning*. Although the results of the experiments were not fully finished to give us a full description of our model, there is perhaps a base from which to progress with further experiments to develop this work.

Index Terms— Computer vision, one-shot learning, Face recognition, Siamese CNN

1. INTRODUCTION

Face recognition is the task of identifying or matching two or more faces from an image or database; in a way that a company may have a digital registration, application security, border-control, which may be looking for certain individuals or, identifying people on group photos via social media. Face recognition is a challenging and sophisticated problem to solve, with many concerns such as lighting, occlusion, aging, expression, and backgrounds. All of these challenges could be on one single image from a vast database of images. Add to this problem a limited number of classes such as, only having one image of each individual you are trying to recognise. also, it mat be of the one images you are trying to match may not be aligned, there may be occlusion such as, sunglasses or a hat, or, the image you have, could be of them being younger or they may have different expressions.

There have been many methods which have address similar problems for example *one-shot learning*[1]; having only one image of each class, and tested on multiple images in different scenarios see figure(1). One of the major issues of *one-shot learning* is, with only one image that image takes up a very small amount of feature space and so finding similarities can be difficult. The idea is to push the class partitions further away from each other whilst also enlarging each partition.

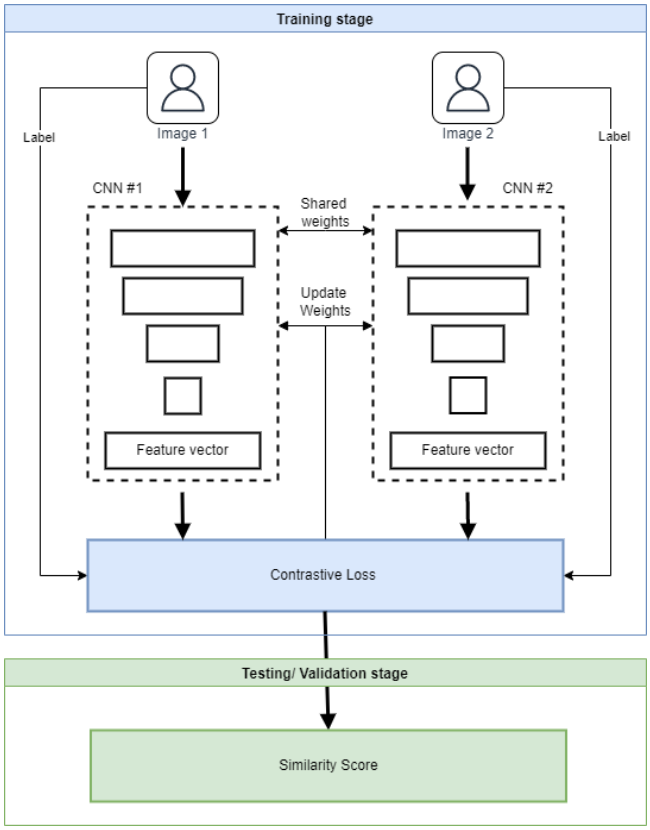

Figure 1. Siamese CNN architecture

The aim of this paper is to create a model that can use a small data-set of single images *one-shot learning*, one class each; such as employees of a company that can be added to the model,which will recognise faces from other images potentially from a registration or boarder-check which have different scenarios for instance, lighting change, occlusions, age, pose or expression and still be able to recognise faces with high accuracy. The model created will be tested against a set

of 100 images(one class each) and tested on a test set of 1300 images of the same class with different scenarios.

2. METHODOLOGY

It can be relatively easy to distinguish if an image is the exact image you are comparing. However, at what point do we say that there are enough similar features to be able to say that the two images you are comparing are the same person, but in a different pose or facial expression etc. To give each face a uniqueness, so we are able to work out if they are the same person or not. A final fully-connected layer is used to gain an embedded feature vector which encodes where in high or low dimensional space the vector of the face resides; for instance a unique address of a certain class of images. This allows us to provide two images that we are matching a similarity score and whether they are similar or dissimilar.

The *Siamese network* is a network that works in parallel, usually two, that share the same weights see figure(1). They have been used in many recognition tasks such as: image recognition [1, 2, 3], Chart classification [4], and signature verification[5].The advantages of a Siamese CNN are: it can generalise well to never before seen data with an infinite amount of classes, can improve with more data, and less overfitting. The disadvantages of a Siamese CNN can be very computationally expensive. A *Convolutional neural network* (CNN) will be used in the *Siamese network* to break-down the images into our *feature vector*. There will be an input layer of 105x105x3 with 4 *convolutional layers* all with *Relu* activation function, followed by a *max-pooling* layer apart from the last layer, which is directly connected to a *fully connected layer* with an output of 4096 see figure (2). Some of the issues with CNN can be that they are hard to interpret and verify, noisy images can cause miscalculations, and require a large amount of labeled data. The idea of the *Siamese CNN* (SCNN) is to create *Siamese batches* of similar and dissimilar pairs, so the model can learn using a loss function usually *cross entropy*, *contrastive loss* or *triplet loss*.

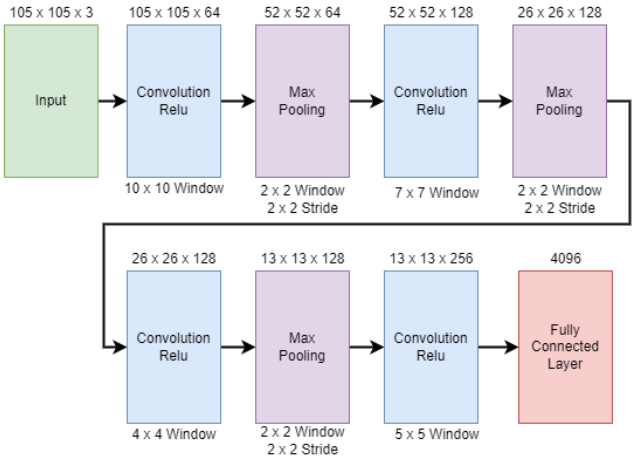

Figure 2. CNN architecture

The loss function *contrastive loss* created by [6] maps each image into embedded vectors which are similar to neighboring points, and dissimilar vectors to distant points. This loss function acts as if it pushes dissimilar items away, whilst attracting similar items towards its local neighborhood in high and low dimensional spaces. The loss function is represented here.

$$L(W, Y, \mathbf{X_1}\mathbf{X_2}) = (1-Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}\max(0, m - D_W)^2 \quad (1)$$

Where D_w is the Euclidean distance between vectors $(\mathbf{X_1}, \mathbf{X_2})$, while m is the margin that defines whether a vector is dissimilar or not, the loss penalises similar vectors that are to far apart and penalise dissimilar vectors that are too close. Training will be split into two stages: **Stage one** will be training on a larger data-set of *Labelled faces in the wild* (LFW) deep funneled [7, 8] which compromises of 13000 images of celebrity faces and 1680 of those with more than one per class; these have been aligned for enhanced results and will be trained using *Siamese mini-batches* and *K-fold cross validation* of similar or dissimilar pairs.

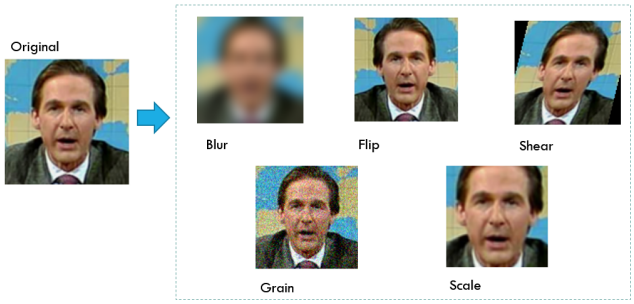

Figure 3. Data augmentation

Stage two is a smaller set of 100 images which have similarities of the set that the model will be tested on see figure(3); these images will be augmented to enlarge the data-set. The problems which augmentation in the data-sets are trying to solve are: illumination, pose,occlusion, aging, expression

and background. Pose, occlusion, aging and expression is the challenge of finding enough uniqueness in an image that gives a distinct encoding, which will standout when there is less information to compare. This challenge will be addressed by transformation of images such as: shear, blur, flipping and scale. Background could addressed by scale where each image can focus more on the face rather than the background. [3] Discuses the conventional pipeline of detect $\rightarrow$ align $\rightarrow$ represent $\rightarrow$ classify; with this in mind, we will detect the image and align them, so we will be able to match. The experiments will be using *Viola Jones Cascade object detector*[9], detects face in an image with coordinates that allows the image to be cropped under similar conditions to help comparison. figure(4).

Figure 4. Cascade object detection

3. METHOD EVALUATION

Experiments on **stage two** with augmented images were assessed first, to identify the robustness of the model; this experiment gave us a target to pass of 25% accuracy, after that we experimented with the larger **stage one** on a clean network, with fresh weights initiated with narrow-normal initialisation that is a normal distribution with zero mean and standard deviation 0.01. However, our first results from this experiment were extremely poor, between 1 - 4%; this indicated that there was a problem with feature vector encoding as the model did not seem to be learning. This was the point were we added Cascade object detection by [9]. The next experiments gave us the accuracy of 14%, indicating that something was not right with **stage one** see figure(4). The k-fold cross-validation indicating that the model was overfitting. So the experiment was run again with lower batch size from 64-32, and we lowered the learning rate from 0.0001 to 0.000001; this appeared to help in training as the cross-validation seemed to curve with the training data. However, with time constraints and technical issues, this experiment did not finish and an accuracy was unable to be attained.

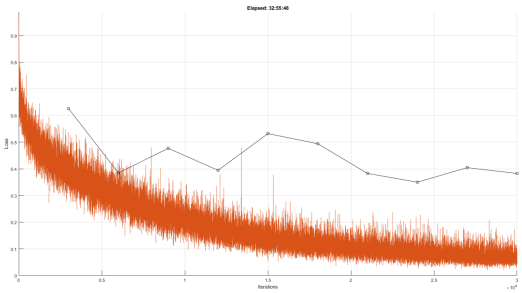

Figure 5. Stage one training

Table 1:Experimental results.

Results			
Accuracy	Precision	Recall	F-Score
14%	?	?	?

4. FURTHER WORK

Further experiments will need to be completed with attention on the lower batch and lower learning rate, and to run the training longer than 6000 X 10 batches to make sure the model converges. Without the results of these experiments it is difficult to gauge what would be needed to improve the model. However, some initial thoughts are to add more augmented images to the **stage one** training, this would add more variation in images such as jitter filters, and to align with augmentation of **stage two** training. **Stage two** training images would be more efficient if they were in an image data store like **stage one**; as stage two had memory restrictions and more augmentation can be achieved, or even **stage one** and **stage two** concatenated together into one training stage. It may be useful to look into triplet loss as a possible loss function used in[2]. With many obstacles such as occlusions, lighting, pose, aging and background, one would have to tackle each obstacle one at a time. This paper has only touched the surface with cascade object detection; it has helped eliminate the background problem with focusing on the face. To help with occlusions of images random images, could have partial images blocked out with patches such as, with the research of [10]. And lastly to speed up prediction if the *one-shot learning* set of 100 images was passed through the model and saved, so that each image does not need to be passed through the network each test would save massive computation.

5. CONCLUSION

The conclusion of this paper is that face recognition using a SCNN is a difficult task with many obstacles, especially with the amount of possible variations of an image such as occlusions, lighting, pose, aging and background. This paper has only scratched the surface in exploring the use of *Cascade object detection* to eliminate the background of each image. Al-

though the results of the experiments were not fully finished there is perhaps a base from which to progress with further experiments to enhance this work.

6. REFERENCES

[1] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov, "Siamese Neural Networks for One-shot Image Recognition," p. 8.

[2] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman, "Deep Face Recognition," in *Procedings of the British Machine Vision Conference 2015*, Swansea, 2015, pp. 41.1–41.12, British Machine Vision Association.

[3] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 2014, pp. 1701–1708, IEEE.

[4] Filip Bajić and Josip Job, "Chart Classification Using Siamese CNN," *Journal of Imaging*, vol. 7, no. 11, pp. 220, Oct. 2021.

[5] Sounak Dey, Anjan Dutta, J. Ignacio Toledo, Suman K. Ghosh, Josep Llados, and Umapada Pal, "SigNet: Convolutional Siamese Network for Writer Independent Offline Signature Verification," *arXiv:1707.02131 [cs]*, Sept. 2017, arXiv: 1707.02131.

[6] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality Reduction by Learning an Invariant Mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, June 2006, vol. 2, pp. 1735–1742, ISSN: 1063-6919.

[7] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," p. 11.

[8] Gary B Huang, Marwan A Mattar, Honglak Lee, and Erik Learned-Miller, "Learning to Align from Scratch," p. 9.

[9] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, Dec. 2001, vol. 1, pp. I–I, ISSN: 1063-6919.

[10] Lingxue Song, Dihong Gong, Zhifeng Li, Changsong Liu, and Wei Liu, "Occlusion Robust Face Recognition Based on Mask Learning with PairwiseDifferential Siamese Network," *arXiv:1908.06290 [cs]*, Aug. 2019, arXiv: 1908.06290.