Queensland University of Technology

IFN680 Assignment 2 - Siamese Network Experiment Report

Doyoo Baek n9544411 | Helen Jeffrey n9416528

Tutorial:Thu 6-8pm (Ashley Stewart)

---

1. Introduction

This report describes the a series of experiments and their results on an machine learning model based on a Siamese Convolutional Neural Network (CNN) to categorise images which have been geometrically transformed. The problem of most recognition systems based on deep neural network is their performance is subjected to projective transformations that simulate changes in the camera perspective. This is not robust when the perspective of the camera changes dramatically. The category of an object in an image is the category of an object could be various to viewpoint changes.

This causes an issue for an automated system to recognize sea species such as manta rays (Maire, 2017). The objective of the experiments is to to determine whether an image can be correctly classified into it's equivalent class by comparing two images. These equivalence classes are provided in the MNIST dataset. By training the fully convolutional Siamese network with original and transformed datasets, the model can classify equivalence classes which are different images show the same object from various observation point.

Several recent works have aimed to overcome this limitation using a pre-trained deep convolutional network that was learnt for a different but related task (Bertinetto, Valmadre, Henriques, Vedaldi, & Torr, 2016). These experiments aims to reduce to the issue of learning manifolds from a training set by adopting existing architectures. Through these

experiments, we hope to demonstrate that a Siamese convolutional network is a feasible solution to discriminate between patterns subjected to large homographic transformations.

To investigate the suitable architecture based on the Siamese convolutional neural network (CNN), we adopted the Siamese networks architecture and created by Krizhevsky, Sutskever, & Hinton (2017). We modified the structure based on work by Bertinetto et al (2016). Also, we modified the well-known Siamese networks architecture based on Koch, Zemel, & Salakhutdinov, (2015) to compare its performance to Krizhevsky et al. We trained those models with 5 different datasets: an original data of MNIST, slightly warped dataset, larger warped datasets, 3 different small and larger warped datasets (40% and 60%, 60% and 40%) respectively.

---

2. Methodology

Overview.

Architecture approach

Our initial architecture is based on Krizhevsky et al (2017). Their architecture was implemented to identify objects in video which also constantly changes the view of images by Bertinetto et al (2016). It is designed to recognise the object contained in different images from different perspectives. Our second architecture is a modified version of Koch et al's architecture (2015). Koch et al used the MNIST dataset to create the siamese neural network. It is used to compare its performance to the modified architecture based on Krizhevsky et al (2017). The different Convolutional Neural Network architectures were used in the different experiments to determine how these affected accuracy. The architectures were selected and re-implemented based on descriptions from literature.

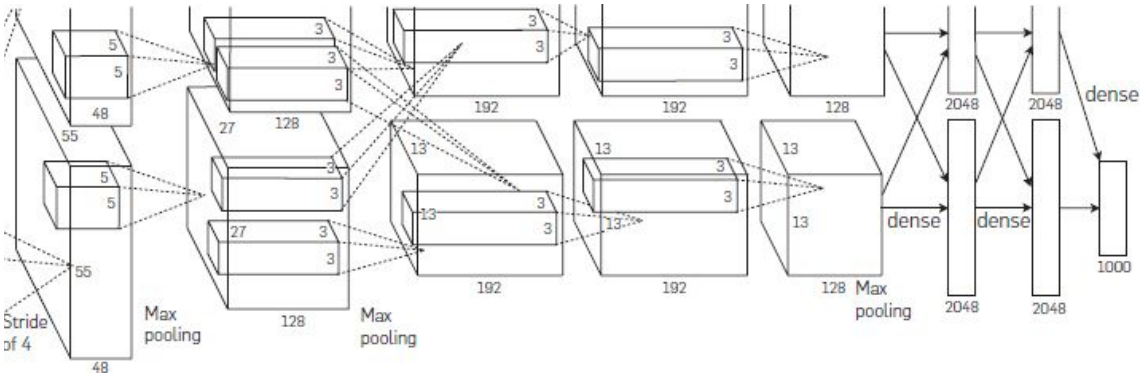

Figure 1 - CNN architecture based on Krizhevsky et al 2017

Siamese Network Architectures

1. Siamese Architecture based on Krizhevsky et al

The initial architecture is based on work by Krizhevsky et al (2017). The architecture they described was initially implemented to classify high resolutions images from the ImageNet data set. The architecture is designed to recognise the object contained in different frames from different perspectives. *Figure 2* below shows the 5 convolutional layers, two fully connected layer, 2 pooling layers of the architecture that was implemented. The ReLU nonlinearity is applied to the output of every convolutional and fully-connected layer except for fifth convolutional layer, the fully-connected layers with drop-outs. Batch normalization were inserted after the first two layers. The architecture implemented below is a modified version, which adds a fully connected layer which it improves the accuracy of model for our problem area. The architecture they describe was implemented to track objects in video where the perspective of the images are constantly changing. The architecture is designed to recognise the object contained in different frames from different perspectives. *Table 1* below shows the fully-convolutional Siamese architecture that was implemented.

Layer	Filters	Kernel/pool_size	Stride
conv1	3	11X11	2
Batch Normalisation			
pool1		3X3	1
conv2	48	5X5	1
Batch Normalisation			
pool2		3X3	1
conv3	256	3X3	1
conv4	192	3X3	1
Conv 5	192	3X3	1
Fully connected layer 1	4096		
Dropout	0.5		
Fully connected layer 2	4096		
Dropout	0.5		

Table 1 - the modified version based on Krizhevsky et al's (2017).

2. The Siamese Architecture version based on Koch et al (2015)

The original architecture is simplified without a regulator and one extra fully-connected layer is added to the original architecture to meet the requirement of this experiment. *Table 2* below shows the fully-convolutional Siamese architecture that was implemented.

Layer	Filters	Kernel/pool_size	Stride
conv1	64	10X10	N/A
pool1		2X2	N/A
conv2	128	7X7	N/A
pool2		2X2	N/A
conv3	128	1x1	N/A
conv4	256	1x1	N/A
Fully connected layer 1	4096		
Fully connected layer 2	4096		

Table 2 - the modified version based on Koch et al (2015)

Data set

The data used in these experiments is based on the MNIST data set. It is a collection of classified images of handwritten digits that have been taken from many scanned documents. Each image is a centered and normalised to be 28 by 28 pixels square.

The images will be paired to used to create input data to train the models. The five different data set will be used. Original dataset is created by pairing the original MNIST dataset. The dataset is warped to simulate potential changes in camera perspective. The original images are warped by different degrees and strength as defined in the experiments. Small warped dataset is transformed with a rotation of 15 degrees and a projective transformation with a strength of 0.1. Larger warped dataset is transformed with a rotation of at 45 degrees and a projective transformation with a strength of 0.3. Staged dataset is 100000 datasets of which are transformed in the different ratio.

3. Experiments

A series of experiments were conducted on two different architecture mentioned above with the 5 different datasets below. These were to determine the effects of different siamese CNN architectures and training approaches on the effectiveness of the siamese network. The experiments are described in this section, with the results in Section 4.

Training & Testing datasets

The following table shows the datasets used in training and testing for the different experiments:

	Data-Set	Training Phase1	Training Phase2	Testing on the training dataset	Testing on validation dataset
1 phase	1 original dataset	100,000	N/A	100,000	1,000
	2. small warped dataset	100,000	N/A	100,000	1,000
	3. Largely warped dataset	100,000	N/A	100,000	1,000
2 phase	4. Staged dataset (small warped 60 -> largely warped 40)	60,000	40,000	100000 (combined training set)	
	5. Staged dataset (small warped 40 -> largely warped 60)	40,000	60,000	100000 (combined testing set)	
	6. Staged dataset (small warped 50 -> larger warped 50)	*50,000*	*50,000*	*100000 (combined testing set)*	

Table 3 - Training & Testing datasets

4. Results

After executing all of the identified experiments, the accuracy and average loss results are shown in Table 4 below. This table shows the experiments completed, the average loss reported and the final accuracy.

First and Second Architecture

	Data-Set	Average Loss on the Krizhevsky Architecture	Accuracy on The Krizhevsky Architecture	Average Loss on the Koch Architecture	Loss on the Koch Architecture
1 phase	1 original dataset	0.30097	50%	0.2283	77.88%
	2. small warped dataset	0.446283	50%	0.2933	72.05%
	3. Largely warped dataset	0.95736	50%	0.2601	73.06%
2 phase	4. Staged dataset (small warped 60 -> largely warped 40)	0.80 0.13	50%	0.21 0.101	73.14%
	5. Staged dataset (small warped 40 -> largely warped 60)	0.343 0.0130	50%	0.14 0.301.	74.14%

Table 4 - Results from experiments

5. Discussion

After completing the experiments identified in section three, we found that the experiments on the Krizhevsky architecture all performed significantly worse than expected. From our readings and examining other implementations, the results were expected to be around +90% based on Koch's architecture. The reported best result on predicting the class of the MNIST data set is 99.79% (Katariya, 2017). We successfully trained them. However, the result of the model is not accurate and they all indicates 50% regardless of what loss is reported.

We discussed several reasons behind this. First the dataset is different since the initial structure is designed for the video. There are more strides in the architecture than the second architecture since image in video is a bigger image than MNIST dataset. Also, they used batch size of 8 with 50 epochs on the half size of samples than ours. It would be too time consuming to conduct the experiment in as described so it was decided to test the architecture with 128 batch size and 10 epochs.

To discuss an effect of two staged training, we refer to the accuracy result based on Koch's architecture. The model trained by the staged dataset performed slightly better than the models trained by the slightly warped and larger warped dataset. However, it is hard to determine that the two staged training improves its accuracy significantly.

6. Conclusion

The Krizhevsky architecture is required to be adjusted its structure for MNIST dataset to produce a better result such as kernel size, strides, and the number of convolutionally layers. It implies that ,depending on what dataset is analysed, an architecture of simonse neural network should be modified and adapted accordingly. Also, this experiment showed two staged trainings increases its accuracy slightly. However, it is difficult to confirm its effectiveness. When it comes to the Koch's architecture, the architecture can be useful to adjust it in the marine object detection once it improves its performance by modifying its layers and structure considering image dataset is color-scale and bigger object than MNIST dataset.

7. References

Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A., & Torr, P. H. S. (2016). Fully-convolutional siamese networks for object tracking. Paper presented at the , 9914 850-865. doi:10.1007/978-3-319-48881-3_56

Kartariya A, (2017). Applying Convolutional Neural Network on the MNIST dataset. https://yashk2810.github.io/Applying-Convolutional-Neural-Network-on-the-MNIST-dataset/

Koch, G., Zemel, R., & Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition. In ICML Deep Learning Workshop (Vol. 2).

Krizhevsky, A., Sutskever, I., & Hinton, G. (2017). ImageNet classification with deep convolutional neural networks. NEW YORK: ACM. doi:10.1145/3065386

Maire, F. (2017, September 27). 2017 IFN680 - Assignment Two (Siamese Network). Retrieved from Advanced Topics in Artificial Intelligence IFN680_17se2 (Blackboard QUT): https://blackboard.qut.edu.au/