Queensland University of Technology

IFN680 Assignment 2 - Siamese Network Experiment Report

Doyoo Baek n9544411 | Helen Jeffrey n9416528

---

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 system 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 two different Siamese networks architecture and created by Krizhevsky, Sutskever, & Hinton (2017). and the second based on work by Bertinetto et al (2016). We trained those models with 5 different datasets: an original data of MNIST, slightly warped dataset, larger warped datasets, small and larger warped datasets (40% and 60%), and small and larger warped dataset as a two staged dataset (60% and 40%).

---

2. Methodology

Experiment Methodology

Figure 1 below shows the processes used to setup and conduct the Siamese Network Experiments.

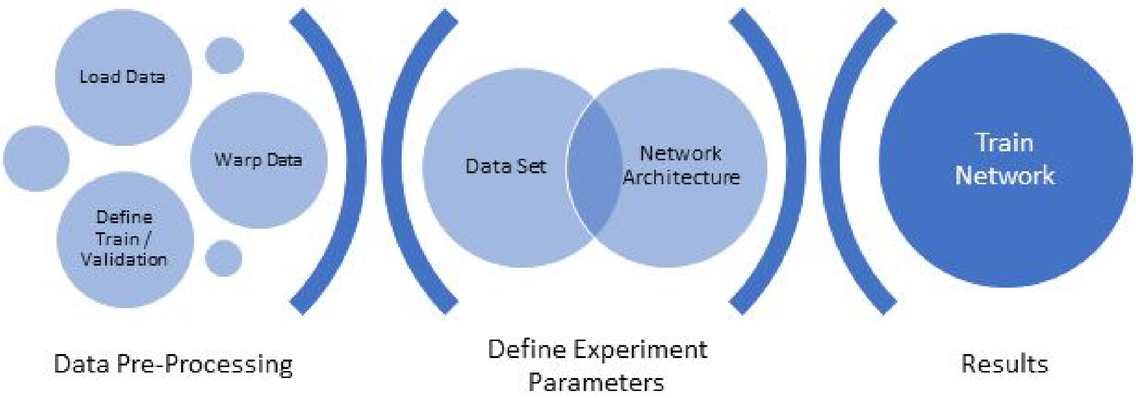

Figure 1 - Experiment Processes

Data Pre-Processing

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 data set is split into a training set of 60,000 examples, and a test set of 10,000 examples. Each image is classified into one of ten possible classes (0 to 9).

Warped Images

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 in section 3.

Training and Testing Dataset

A variable of the experiment is five different training datasets as mentioned above which are used to train the models.

Siamese Network Architectures

Two different Convolutional Neural Network architectures were used in the different experiments to determine how these affected accuracy and loss. The architectures were selected and re-implemented based on descriptions from literature. The first architecture was based on an experiment and article by Krizhevsky, Sutskever, & Hinton (2017). and the second on work by Bertinetto, Valmadre, Henriques, Vedaldi, & Torr (2016).

1. Siamese Architecture based on Krizhevsky et al

The first architecture is based on work by Krizhevsky et al. The architecture they describe was 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 eight layers (five convolutional and three fully connected) of the architecture that was implemented. The architecture was adapted to work on the grey-scale images of the MNIST data set.

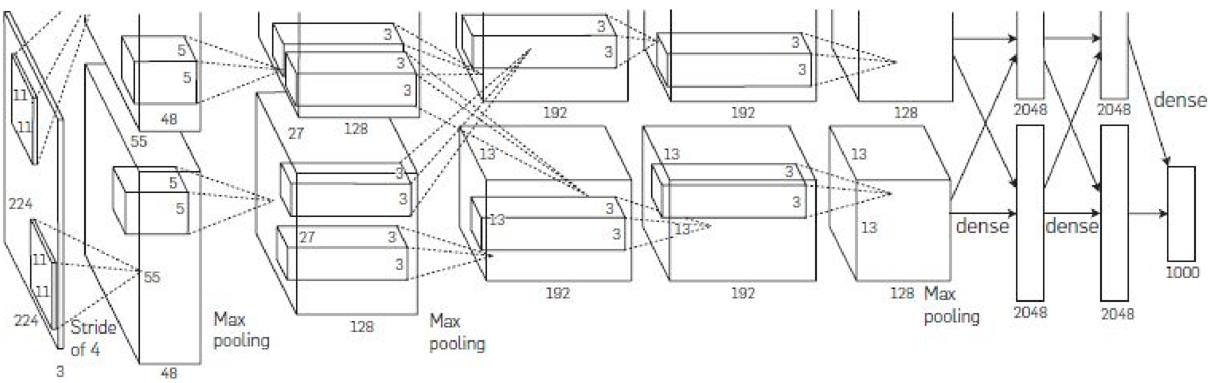

Figure 2 - CNN architecture based on Krizhevsky et al

2. Fully-convolutional Siamese Architecture based on Bertinetto et al

The second architecture implemented is a modified version, which adds a fully connected layer which it improves the accuracy of model for our problem area.

The second architecture is based on work by Bertinetto et al. 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 2* 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
conv5	192	3X3	1
Fully connected layer1			
Fully connected layer2			

Table 2 - CNN architecture based on Krizhevsky et al

As can be seen, Max-pooling is employed after the first two convolutional layers. The ReLU nonlinearity is applied to the output of every convolutional and fully-connected layer except for fifth convolutional layer, the final layer. immediately after every linear layer. The stride of the final representation is eight. During training, batch normalization was inserted immediately after every linear layer.

The Siamese network is composed of two identical subnetworks which share the equal weight followed by a distance calculation layer. The input of the network is a pairs of images pi and pj. If two images are deemed from the same equivalence class, we pair is called a positive pair.

3. Experiments

Once the application was implemented, a series of experiments were conducted. These were to determine the effects of different 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. *Table 1 - Summary of experiments*

Training & Testing datasets:

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

	Data-Set Number	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.

	The number of data set	Average Loss on the initial architecture	Accuracy on the second architecture	Average Loss on the initial architecture	Loss on the second architecture
1 phase	1 original dataset	0.040097	96.01%	0.013	89%
	2. small warped dataset	0.0146283	63.42%	0.0604	93%
	3. Largely warped dataset	2.5736	88.5%	0.0831	92%
2 phase	4. Staged dataset (small warped 60 -> largely warped 40)	TBD			
	5. Staged dataset (small warped 40 -> largely warped 60)	TBD			
	6. Staged dataset (small warped 50 -> larger warped 50)	TBD			

Table 4 - Results from experiments

5. Discussion

After completing the experiments identified in section three, we found that the experiments all performed significantly worse than expected. From our readings and examining other's implementations, results were expected to be around +90%. The best reported result on predicting the class of the MNIST data set is 99.79% (Katariya, 2017). It is our belief that some of this was due to the implementation of the network architecture and supporting code. The complexity of the selected CNN architectures that were implemented also may have contributed to these results.

Convolutional Layers

It was found that the CNN architecture with the smaller number of convolutional layers architecture is working better than the 5 convolutional architecture based on work by Krizhevsky, et al.

Implementation Issues

- Data Transformation speed - In experiments using the warped data, the transformation was being started from scratch each time. This cost a lot of time over the course of the experiments.
- Training speed - each test was taking up to two hours to execute and report accuracy back.

6. Conclusion

The experiments showed better than average prediction of classes of the warped and original images. It is possible that after correcting the implementation issues; these results would have been improved as well.

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/