EE-559 Deep Learning: Course Project 1

Futong Liu **Shilin Wang** **Sicheng Xie**
futong.liu@epfl.ch shi-lin.wang@epfl.ch sicheng.xie@epfl.ch
EPFL, Lausanne, Switzerland

Abstract

Image classification is a significant topic in the field of computer vision. In this project, we implemented four deep learning classifiers to compute the relationship between two MNIST digit images. Trained with weight-sharing and auxiliary loss, our best Siamese model achieves a test error of $2.39\% \pm 0.43\%$.

1 Introduction

Image classification is a significant topic in the field of computer vision. Given as input a pair of two MNIST images (14×14) of hand-written digits, the project aims at implementing a classifier with deep learning that can tell if the digit in the first image is less or equal than the digit in the second one.

Four different architectures are implemented and compared in the project:

- CNN with input as a two-channle image pair (Baseline)
- CNNs without weight sharing
- Siamese CNN with MLP for decision
- Siamese CNN with logical decision: Siamese CNN digit classifier

Particularly, the influences of weight sharing and auxiliary loss are assessed on how they help with the performance of the classifier.

2 Model Description

2.1 Model 1: Baseline

As shown in 1, the baseline model treats the input image pair as a single two-channel image and predicts the relationship between the two digits the images with a CNN.

A CNN with a LeNet-like architecture is implemented to recognize and classify the input images. 2D-convolutions are used since there is no spatial or time implication by the order of the channels. The details of the network are listed as below in 1:

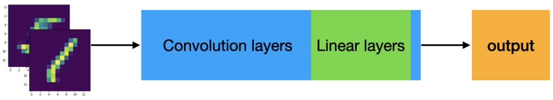

Figure 1: Architecture of baseline

Layer	Structure		Nb. Params
1	Convolution Layer Max Pooling Relu	(2, 32, 3) 2	608
2	Convolution Layer Max Pooling Relu	(32, 64, 3) 2	18,496
3	Linear Layer Relu	(256, 200)	51,400
4	Linear Layer	(200, 2)	402

Table 1: Details of the CNN

The loss of model 1 is simply the cross entropy loss between the output 2×1 vector and the ground truth target.

To treat the input image pair as a single two-channel image does not fully exploit the advantages of CNNs. Usually a channel of an image represents the colour and the kernel of the convolution layers extracts the information from different channels of an image. Therefore, it is better to train a classifier with input as two single channel images.

2.2 Model 2: Parallel CNNs

Since the exact class information of digit images is also provided, we believe that to include this information can improve the performance of the network. According to literature [2], auxiliary loss can push down the information and promote discrimination to lower stages in deep networks.

Namely, to achieve our main objective of relationship classification, we also accomplish an auxiliary task of digit classification.

As shown in 2, each image in the input image pair undergoes a CNN and produces a vector of size 10 × 1, which represents the probability of each class. An auxiliary loss can be computed, that is the cross entropy loss between the vector generated and the class information provided. At the output of the fully connected layers, the final decision is made and the cross entropy loss between the decision and the target is computed as well. Hence, the loss function consists of three terms: two auxiliary losses and one output loss.

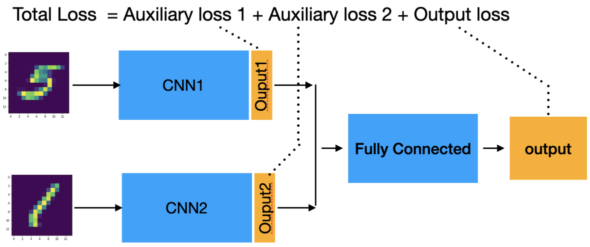
Total Loss = Auxiliary loss 1 + Auxiliary loss 2 + Output loss
CNN1
Ouput1
CNN2
Ouput2
Fully Connected
output

Figure 2: Architecture of model 2

2.3 Model 3: Siamese Network

A Siamese neural network shares the weights when deal with different inputs and produces comparable outputs [1]. As shown in 3, model 3 has similar architecture as model 2, except that the two CNNs share the same weights, which forms a Siamese network. Each image in the image pair goes through the same CNN and hence produces two comparable output vectors. Same as model 2, the loss function of model 3 has three terms.

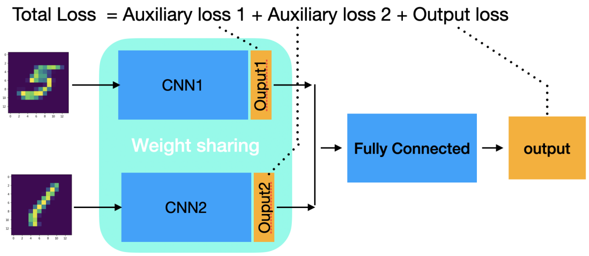
Total Loss = Auxiliary loss 1 + Auxiliary loss 2 + Output loss
CNN1
Ouput1
Weight sharing
CNN2
Ouput2
Fully Connected
output

Figure 3: Architecture of model 3

2.4 Model 4: Siamese Network

Since we know that the ultimate objective of the classifier is to predict for each pair of images whether the first digit is less or equal to the second one, we can hard-code this comparison logic immediately after the digits are classified.

The architecture of model 4 is shown in 4. From the output vector of the CNN, the model can determine which number the digit image is most likely to be. By comparing the two numbers, the decision is finally made.

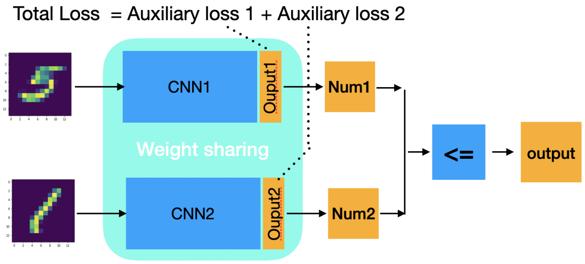

Figure 4: Architecture of model 4

As the fully connected layer is replaced by a deterministic procedure, where there is no parameter to be learned, the loss function simply consists of two auxiliary losses. Essentially, the model only learns a CNN that classifies the number.

3 Model Training

All the four models are trained over 25 epochs with a batch size of 10. The parameters are optimised by SGD. Since the loss functions are different, the learning rate are roughly tuned for each model with grid search.

4 Results and Discussions

In order to test the performance of the models, each model has been tested 15 times to get the average test error rate and the standard deviation. Below in 2 are the results of the performance of each model,

Model	Test Error	Nb. Params
1	18.22% ± 2.20%	70,906
2	5.96% ± 0.87%	150,614
3	5.35% ± 0.84%	79,708
4	2.39% ± 0.43%	70,906

Table 2: Results of Models

We see that the baseline method performs the worst. It treats the input image pair as a two-channel image pair, which doesn't fit the physical meaning of channel and hence the CNN can't extract advantageous information by performing convolution over the over-lapped two-channel image.

Model 2 uses two separate CNNs to classify the input images first and then use the fully connected layers to make the decision. The improvement is largely due to the usage of auxiliary loss, which includes the information of digit class into the model as well and hence reduces over-fitting and produces more satisfactory performance.

On top of model 2, model 3 uses a Siamese network architecture that has a twin CNN with exactly the same parameters such that each image in the input image pair undergoes the same procedure before digit classification. It can be concluded that weight sharing improves the performance: in model 3, the CNN is trained with the whole training dataset, while in model 2 each CNN can only see half of the input images. Therefore, there are fewer parameters in model 3 than in model 2, which reduces the complexity of the model and alleviates over-fitting.

Model 4 has the best performance. Model 4 does not have to learn the decision logic since the decision is simply a numeric operation. The only parameters to be learned are the ones in the CNN of the Siamese network. This reduces the number of parameters significantly and hence reduces the effect of over-fitting as well.

5 Conclusion

By comparing the four models implemented for the classification problem, it is shown that appropriate architecture of a network play a significant role of its performance. Additionally, weight sharing and auxiliary loss can reduce the complexity of the model and include more information to the training process and hence improve the performance of the network.

References

[1] S. Chopra, R. Hadsell, and Y. LeCun. “Learning a similarity metric discriminatively, with application to face verification”. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*. Vol. 1. 2005, 539–546 vol. 1.

[2] C. Szegedy et al. “Going deeper with convolutions”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1–9.