Machine Learning
Team Pi

Daniel Moorhead
201860527
daniel.moorhead.2018@uni.strath.ac.uk

Alex Russell
201882769
alexandra.russell.2018@uni.strath.ac.uk

Garry McBride
201867279
garry.mcbride.2018@uni.strath.ac.uk

April 9, 2019

1 MNIST

1.1 Combination 1

The best combination found was a convolutional neural network with one convolutional layer, followed by a pooling layer, flattening layer and then two fully connected layers. This network took in hand drawn digits (0-9) and classified them into the digits.

1.1.1 Description of the combination

The convolution layer is a 2D convolution with a 5x5 kernel and 32 feature maps. This is followed by the pooling layer. This layer is a 2D max pooling layer with a 3x3 kernel. The two fully connected layers are 100 sigmoid neurons followed by 10 softmax neurons.

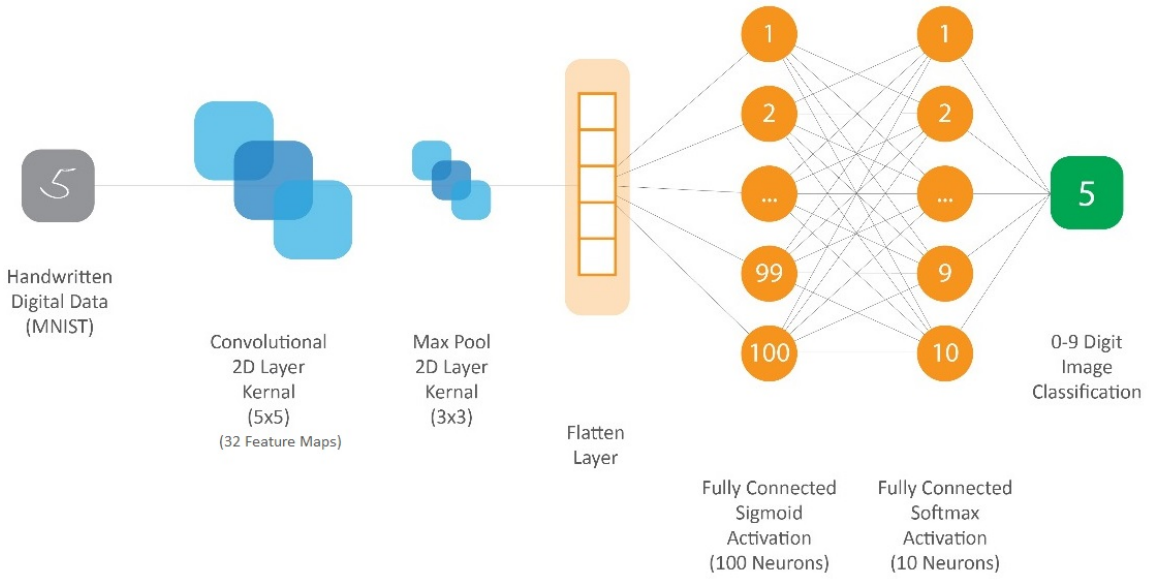

Figure 1: Network diagram of MNIST combination 1

1.1.2 Parameters Explored

When exploring the parameters the goal was to find the simplest model that achieved the best accuracy. Increasing the number of epochs or using a more complex structure didn't improve accuracy, merely increased the runtime, as we can see with Combination 3. Various batch numbers and learning rates were investigated (with 1200 and 0.01 respectively being the best combination). Finally, different node numbers and activation functions were tested. The functions included relu, exponential and linear, but sigmoid and softmax proved the most effective. The seed used was 12345.

1.2 Combination 2

Combination two is a simple multilayer perceptron. It has one fully connected layer with 100 neurons and a sigmoid activation function. This simple architecture produces respectable accuracy rates, if

slightly overfitted. Adding more layers was not found to improve the model, as we can see from Combination 4. It was found that the number of batches must be kept low or the model overfits.

1.3 Other Combinations

1.3.1 Combinations 3, 5 and 6

These combinations were all variations on combination 1. They change the structure by adding extra convolutional layers, changing the activation functions and changing the dropout rate.

1.3.2 Combination 4

Combination 4 adds two additional hidden layers to combination 2. We can see that the results are very similar, but the additional complexity doesn't improve results.

1.4 Justification

Since convolutional neural networks tend to handle images better (by preserving the structure), it is unsurprising that the highest accuracy generated was achieved with a convolutional neural network. Combination 1 gave the most accurate results and avoids over and under fitting. The number of batches, epochs and layers were chosen as they achieve the best accuracy without being unnecessarily complex or compromising runtime. As such, Combination 1 gives the best results in terms of both accuracy and efficiency.

Example	Combination	Parameters and Configuration	Training Accuracy	Testing Accuracy
a	**1**	**LR:0.01, Epochs:5, Batches: 1200, Network as described above**	**0.9849167**	**0.9823**
b	2	LR:0.3, Epochs:100, Batches: 20, Network describes above	0.995	0.9029
c	3	LR:0.01, Epochs:5, Batches: 1200 , Network the same as Combination 1 but with an additional convolutional layer	0.9684	0.9705
d	4	LR:0.3, Epochs:100, Batches: 20, Network is the same as combination 2 but with 2 additional hidden layers with 100 neurons on each layer	0.989	0.8937
e	5	LR:0.01, Epochs:5, Batches: 1200, Network the same as Combination 1 but without the softmax activation function on the output layer	1	0.4787
f	6	LR:0.01, Epochs:5, Batches: 1200, Network the same as Combination 1 but with a higher dropout value of 0.9	0.95968336	0.962

2 IMDB

2.1 Combination 1

The best combination found was a basic neural network. This network was many-to-one that took in text data of IMDB reviews and classified it to either positive or negative. The model was built as a sequential model.

2.1.1 Description of the combination

The first layer was an embedding layer with an input dimension the same as the incoming data, and six outputs. After this there is a dropout with a rate of 0.4.

After this the data is pooled using global max pooling. This connects to two fully connected hidden layers. The first had 16 relu activation nodes, and the second had 1 sigmoid activation node as the output. The network was optimised with the Adam optimiser.

Below is a diagram of the architecture:

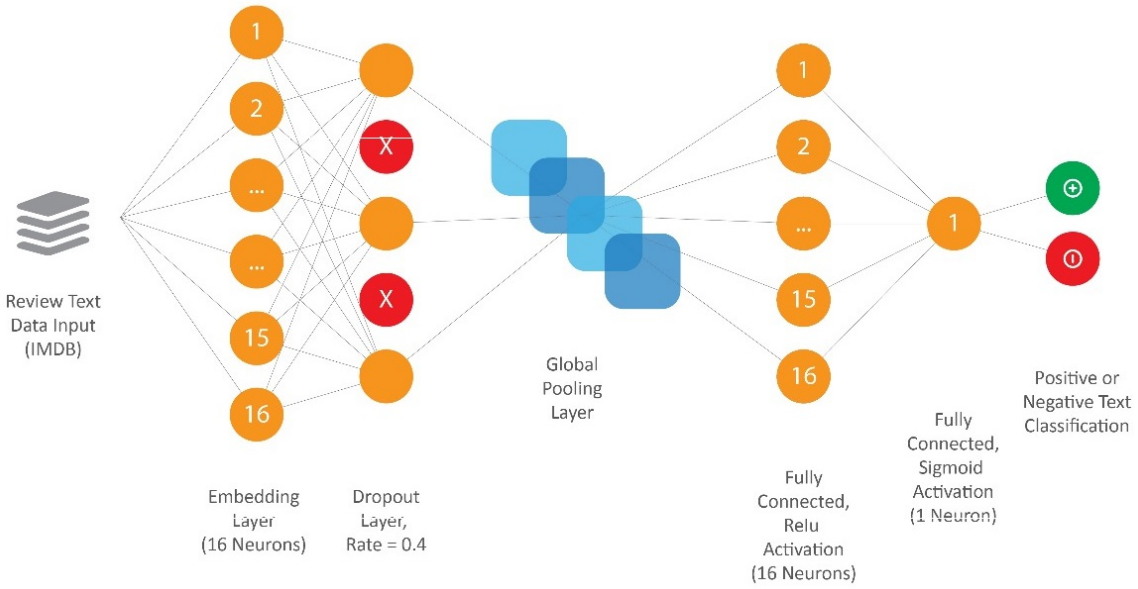

Figure 2: Network diagram of IMDB combination 1

2.1.2 Parameters Explored

One of the main parameters explored was the number of hidden layers and the number of nodes. The activation functions for each layer was also investigated. The functions investigated were linear, sigmoid, hard sigmoid, exponential, softmax, relu, elu, selu, softplus, softsign and tanh.

The optimisation functions were also investigated, these were chosen from SGD, RMSprop, Adam, Adamax, Nadam, Adagrad and Adadelta. Diffrent types of pooling were also investigated. These were; max pooling, average pooling and global max pooling. While no parameters

were needed to be explored with global poolings, the other options were explored with different stride lengths and pool sizes.

The dropout rate, epochs and and number of batches were also investigated. For all investigations the seed was 12345

2.2 Combination 2

Combination two was a recurrent neural network based around the LSTM layer. It had a an embedding layer with 16 points, a dropout layer with rate 0.2. This was followed by an LSTM layer of one node and a tanh activation function. This was followed by a layer of 32 sigmoid nodes and a final layer of one sigmoid node. This uses the Adamax optimiser

2.3 Other Combinations

2.3.1 Combination 3

Another combination was much the same as combination 2 however, after the LSTM layer there was a time distributed layer of 16 softmax nodes. The optimiser was a Nadam optimiser.

2.3.2 Combination 4

A different combination was a bidirectional wrapped LSTM layer of 16 softmax nodes followed by 32 relu layers, 16 relu layers and one sigmoid layer. This was done with the SGD optimiser

2.4 Justification

Although theory suggests that recurrent neural networks are most suited to text classification, this proved to be too computationally demanding. And so the basic neural network was decided.

The parameters chosen provided the best combination of accuracy and efficiency. Parameters such as epochs and batch results were increased to the point where no difference was made. Other parameters were tested for a variety of different values.

2.5 Combination results

Example	Combination	Parameters and Configuration	Training Accuracy	Testing Accuracy
a	**1**	**learning rate 0.001; epochs 20; batches 100; as described above**	**0.95560**	**0.8385040**
b	2	learning rate 0.05; epochs 10; batches 50; as described above	0.53247	0.510520
c	1	learning rate 0.4; epochs 12; batches 10; plus Nadam optimiser and an extra layer of 50 tanh neurons	0.50353	0.50000
d	2	learning rate 0.04; epochs 5; batches 25 plus no dropout and an extra layer of 1000 sigmoid neurons	51034	0.48974
e	3	learning rate 0.07; epochs 2; batches 20; as described above	0.501000	0.49647
f	4	learning rate 0.5; epochs 1; batches 20; as described above	0.50353	0.499888