3.4 Convolutional networks

Table (1) and (2) presents the classification errors and number of training epochs for each of the two networks.

Table 1: Classification errors for each network

Data set	Network 1	Network 2
Training set	0.3271	0.2537
Validation set	0.3892	0.2953
Test set	0.3855	0.2862

Table 2: Number of training epochs for each networks

Networks	Number of epochs
Network 1	120
Network 2	120

From table (1) it is clear that network 2 performs best, due to having lower classification error than network 1. Since network 1 has one more fully connected hidden layer with 50 neurons than network 2, network 1 might experience problems with overfitting due to having more neurons. Overfitting the data will in turn decrease the learning of the network and in turn result in a higher classification error. Network 2 also contains two more convolutional layers than network 1, which is beneficial since convolutional networks has fewer neurons than fully connected networks. This decreases the risk of overfitting the data.

Another difference between the two networks is that network 2 has one more Max-pooling layer than network 1. The Max-pooling layer simplifies the outputs of the convolution layers by summarising the outputs of close feature maps, by taking the maximum over the feature maps outputs. The Max-pooling layers thereby reduce the dimensionality and the of amount of parameters and computational work in the network, which increases the performance.

Network 2 also contains batch-normalisation layers, in contrast to network 1. In fully connected networks the batch-normalisation layers have a role in reducing the unstable gradient problem, by preventing local fields of hidden neurons to grow. However, it is unclear to which extent the batch-normalisation layers reduce the unstable gradient in the the convolutional networks with the ReLU function. From the lectures in the Artificial Neural Network course, it has been taught that the batch-normalisation layers may have a reducing effect to the unstable gradient problem, even though it is not known exactly how. Thereby this results in yet another reason to why network 2 performs better than network 1.

Finally it has been noted from table (2) that early stopping did not occur for any of networks, since the maximal number of epochs were set to 120.