With no added improvement, the same best hyper-parameters from the previous exercise, with all the data sets and 7 epochs of training gave the accuracy:

Training accuracy	Validation accuracy	Test accuracy
0.5627	0.5070	0.4902

Table 1. Accuracy from the original network setting from the previous exercise.

Adding the He initialization where $var(W1) = 2/3072$ and $var(W2) = 2/50$, showed the performance:

Training accuracy	Validation accuracy	Test accuracy
0.5624	0.5210	0.4904

Table 2. Accuracy with He initialization. Improvement (b).

There was no remarkable difference for training, but the validation accuracy increased a little bit.

Next, I compared the accuracy of the network with 3 different number of nodes with the original weight variance.

Number of nodes	Training accuracy	Validation accuracy	Test accuracy
50	0.5627	0.5070	0.4902
100	0.6003	0.5360	0.5109
150	0.6186	0.5270	0.5188

Table 3. Accuracy for different number of nodes. Improvement (d).

In table 3, increasing the number of nodes gave better test accuracy, but as the complexity of the network increases will likely overfit the data, that's why the training accuracy is higher for 150 nodes.

Last I checked with only adding a decay of the learning rate after 5 epochs.

Training accuracy	Validation accuracy	Test accuracy
0.5804	0.5110	0.5145

Table 4. Accuracy with a decay of the learning rate after 5 epochs. Improvement (e).

Adding the improvements b,d and e specifically to the original network setting showed better performance than before with no improvements. The performance with a combination of all 3 and 150 nodes is shown in the table below.

Training accuracy	Validation accuracy	Test accuracy
0.6529	0.5650	0.5406

Table 4. Accuracy with all the improvements.

Then I checked the performance of the network using different activation functions. With the same hyper-parameters as before with all of the datasets and ran the training on 7 epochs.

Activations	Training accuracy	Validation accuracy	Test accuracy
ReLu	0.5627	0.5070	0.4902
Tanh	0.1003	0.0870	0.1000
Sigmoid	0.1725	0.1670	0.1633

Table 5. Accuracy with different activation functions.

It seems like a ReLu activation function is most suitable for a full connected 2-layer network.