Assignment-4

Kshitij Goel **50246430**

1. Assuming the functions used by the auto encoders as following:
   Encoding:
   $$z = \sigma(Wx + b)$$
   And Decoding:
   $$x' = \sigma'(W'z + b')$$
   W, W' = Weight Matrices
   b, b' = Offsets
   z = Hidden Layer
   x = Input
   x' = Output
   σ, σ' = Activation Functions

   Error is calculated by finding the squared difference between the input and the output.
   $$\mathcal{L}(x, x') = \|x - x'\|^2 = \left\|x - \sigma'(W'\big(\sigma(Wx + b)\big) + b')\right\|^2$$
   Thus, error is calculated by the substitution of $\|x - x'\|^2$.

2. The number of nodes are reduced in each layer because the encoding of information takes place in fewer dimensions as compared to the size of original image.

3. The auto encoders are trained one hidden layer at a time because of adoption of greedy approach to calculate the weights of layer. In this process, each layer is trained separately and then combined to the complete model which is then trained again. The process does not require a label thus, making it an unsupervised learning algorithm.

4. The features highlighted in Figure 3 are obtained from the data set digitTrainTestArray which is used as input to the single layered auto encoder. This auto encoder has a hidden layer with 100 nodes and each node is responsible for learning a different feature from the data set, thus providing us with an end result of 100 feature set.

   This is different from the assignment-1 as it made use of k-means classification for finding the dictionary which was interpreted as out feature set. A disadvantage of k-means is that it can only be used for classification purposes whereas the auto encoders can be used for compression of data into smaller dimensions for faster processing while using a soft-max layer.

5. The function 'plotconfusion' is used in MATLAB for returning a confusion plot of the matrix for the output data as compared to target data.

6. The hidden layers in MATLAB use ‘logsig’ (Logistic Sigmoid) as the activation function.

7. ReLU activation function is preferred over sigmoid activation because ReLU is a rectifier and acts as a ramp function.

   For any negative value of x, y maintains the value 0 and then linearly increases with the increase of value of x. It also maintains a slope of 1 i.e. it acts as a straight line whereas sigmoid has a vanishing gradient to 0.

8. It is not a good idea to initialize the network with all values of 0 because if all nodes start with the same weight, same value and properties of gradient will be followed by the nodes during the backpropogation step of the algorithm. This will lead to a symmetry issue as the network would not behave in the way that it is supposed to.

9. Batch Gradient Descent (BGD) computed the gradient after taking into consideration the whole data set. This provides a more accurate result and works smoothly in error manifolds.

   Stochastic Gradient Descent (SGD) computes the gradient after taking into consideration of a single sample from the data set which provides to be advantageous for finding minima and maxima of the data set. It also computes faster than BGD.

   BGD is faster than SGD when the comparison is on the basis of number of epochs because of the way the two function take the data set. Since, BGD uses the complete data set, it will pass through the epoch in one iteration whereas SGD has to consider just a single point of view.

   When compared on the basis of number of iteration, BGD is slower because with each iteration the gradient values of the weights of the entire network are updated while SGD has to update a single value in an iteration.

10. We can observe from the overall functioning of the program that a change in number of hidden layers, drastically changes the outputs. An increase in the number of hidden layer, increases the accuracy but make the program more complex and a decrease in layers, decreases the accuracy.

    The other values which can be altered for testing are MaxEpochs (ME), L2WeightRegularization (L2WR), SparsityRegularization (SR), SparsityProportion (SP), ScaleData (SD). The network fails in many instances while changing the values individually of these parameters. The network failed when the value of SP in 2^{nd} layer was changed to 0. By reading the confusion matrix in this case, we can summarize that the accuracy is at 10%. Changing the number of hidden layer count in 1^{st} layer improves the accuracy but the increase in accuracy is non-advantageous as the computation time increases drastically. Major changes to the value of SR does not change the accuracy of the network by much.