DEEP LEARNING HW1 REPORT

[redacted]

Question 1: Report your architecture, training time and performance on the test set. Do this for each network:

Answer:

1) The first neural network architecture (suitable for the first and the second network in the homework) includes an input layer with 784(28*28) neurons, a hidden layer(60) with a linear transformation, and an output layer mapping features from the hidden layer to output classes(it can be 3 or 7).
   The output layer map features to output classes, with no activation function applied directly. The forward pass method flattens input and applies ReLU activation to introduce non-linearity.

2) The "NeuralNetwork4weighted" (suitable for the third network) neural network architecture consists of four linear layers with Leaky ReLU activation functions applied after each linear transformation. The input layer has 784 neurons and the hidden layers have a linear transformation from the input layer to a specific number of neurons, introducing non-linearity. The output layer map features from the last hidden layer to output classes, determined by the number of classes in the classification task. The "forward pass" method flattens the input and passes through the layers sequentially, introducing non-linearity after each linear transformation. The final output represents the logits for each class.

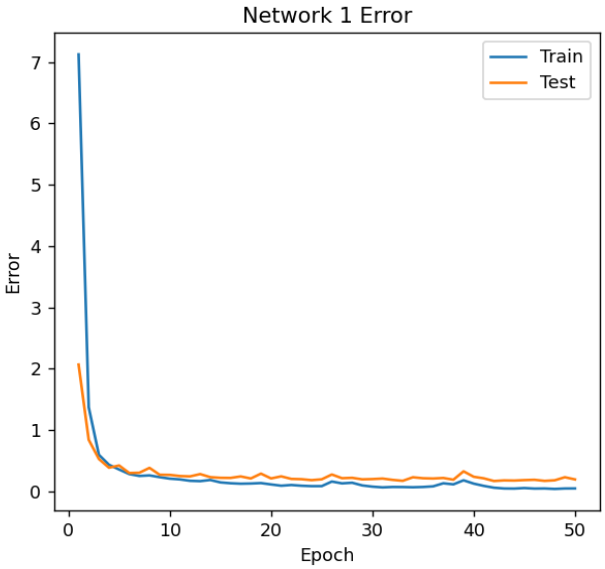
3) The third neural network architecture is a Convolutional Neural Network (CNN) (suitable for the fourth network) designed for image classification tasks. It comprises three layers: "self.conv1", "self.conv2", and "self.conv3". Batch normalization layers are applied after each convolutional layer to stabilize and accelerate training. The pooling layers are "self.pool", a max pooling layer with a kernel size of 2x2, which reduces the spatial dimensions of the input feature maps. Dropout layers are applied after each pooling layer for regularization, randomly zeroing some elements of the input tensor with a probability of 0.3, 0.25, and 0.25, respectively. A fully connected layer is "self.fc1", which maps the flattened output from the last convolutional layer to the output classes number(7). The input size is determined by the spatial dimensions of the output from the last convolutional layer.

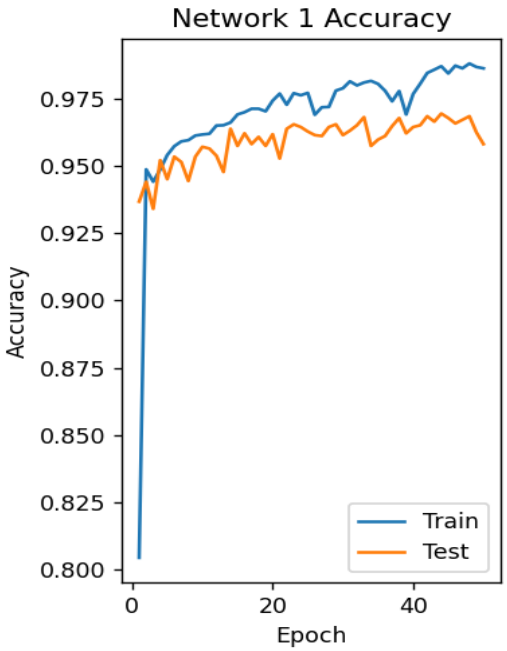
   Activation functions are applied after each convolutional layer to introduce non-linearity and to obtain class probabilities. A helper function is used to compute the number of features in the output of the convolutional layers, which is required for the input size of the fully connected layer.

Question 2: Add a plot showing train error, test error, and accuracy as a function of the number of epochs. Do this for each network.

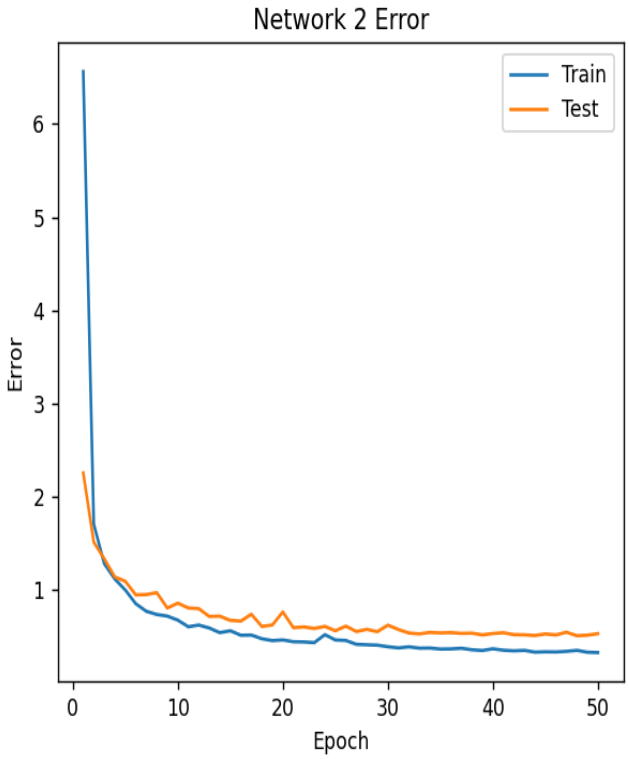
Answer:

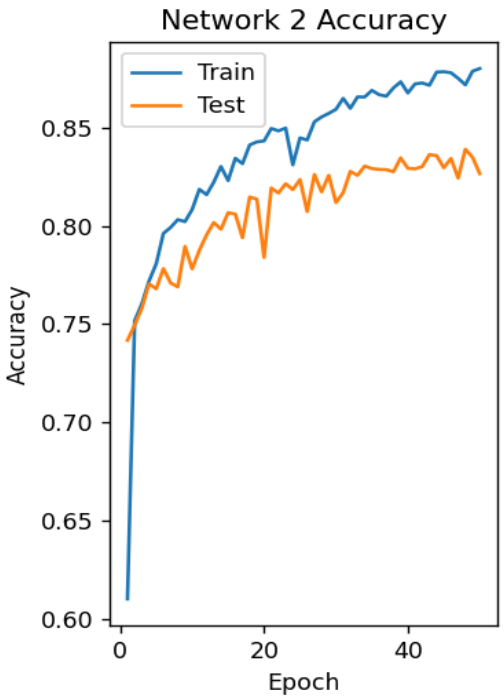
Network 1:



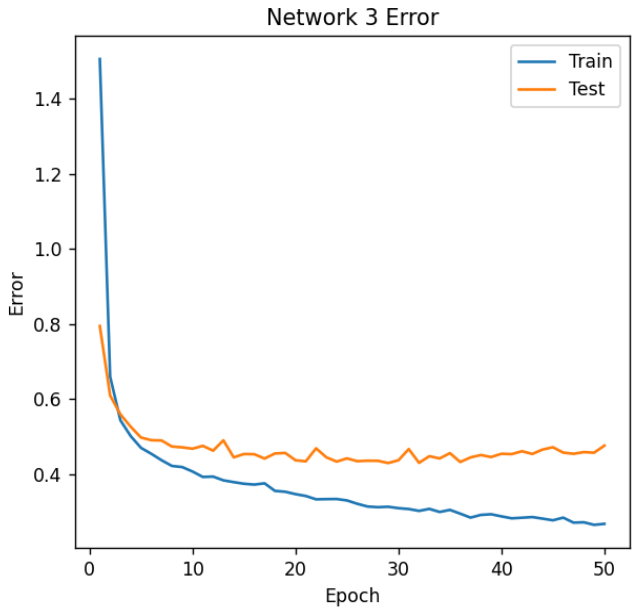


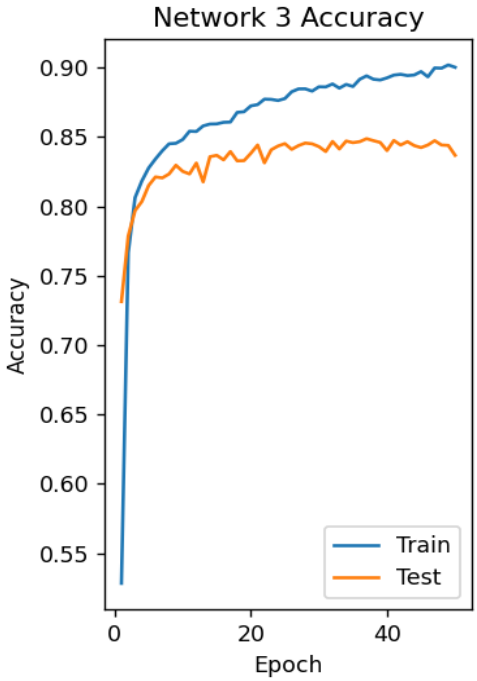
Network 2:



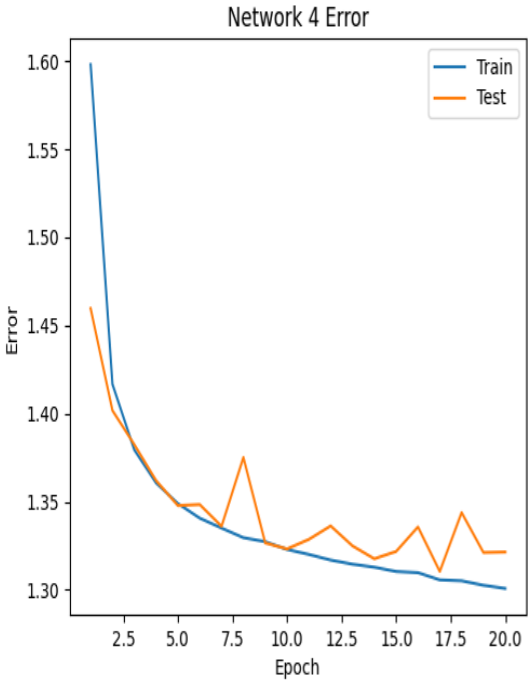


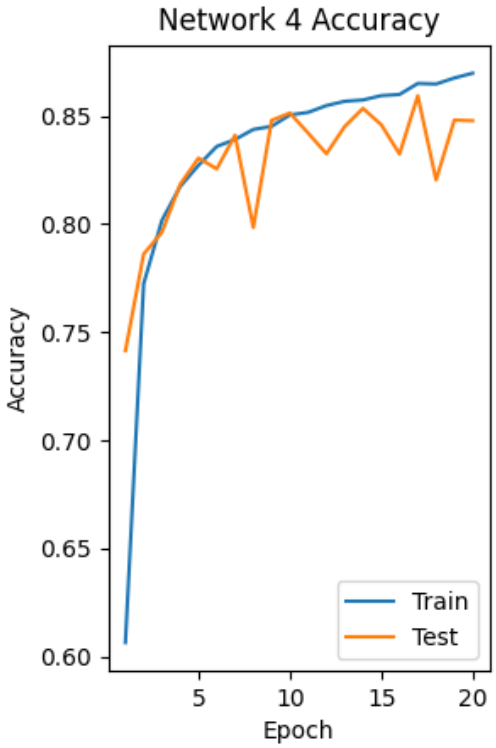
Network 3:





Network 4:





Question 3: Explain why the performance of Network 1 is better than that of Network 2.

Answer:

The performance of Network 1 is better than that of Network 2, because while both Networks have the same number of parameters, in Network 2 we need to classify on 7 classes while Network 1 we classify only on 3 classes, which leads to more complexity into the classification task, as the model needs to differentiate between more potential outcomes. This increased complexity can make it harder for the model to learn discriminative features and boundaries between classes accurately while having the same number of parameters (in Network 2, every class has a smaller number of parameters that learns about him, 50k divided to learn 7 classes, while in Network 1 50k divided to learn on 3 classes).

Question 4: Explain why the performance of Network 3 is better than the performance of Network 2.

Answer:

While both Networks have the same number of parameters, the performance of Network 3 is better than the performance of Network 2, because Network 3 has higher depth than Network 2, higher depth leads the network to learn more complex and hierarchical representations of the data, exploit feature reuse and specialization, benefit from smoother optimization landscapes, and achieve better regularization and accuracy in our case.

Question 5: Explain why the performance of Network 4 is better than the performance of Network 3.

Answer:

While we have the same number of parameters in both networks, but the performance of Network 4 is better than the performance of Network 3, and that is because in Network 4 we use CNN architecture while in Network 3 we use FC.

In data of images, we need to distinguish and learn spatial patterns and features that happens in the same class of the image, and that is what CNN can do by applying convolutional and pooling layers, while FC is focused on each pixel as separate feature, ignoring the spatial pattern and similarity between neighbouring pixels, which are crucial for understanding images effectively, so because of that, a small change (e.g. 10-degree rotation) on the feature vector that completely will go wrong on FC, in CNN this change is handled.

Also, CNN exploits the concept of parameter sharing by applying the same parameters across different regions of the input image, so while both networks have 50k parameter, but Network 4 uses them more efficiently by parameter sharing.

Question 6: Conclude, by questions 3,4,5 above, which architecture is best suited for the FMNIST classification problem.

Answer:

Looking at the results of each network and from 3,4,5 we can conclude that the best architecture for FMNIST classification problem is the one we used in Network 4 which is CNN.