Stem: the image of size 28x28 pixels is divided into 196 non-overlapping patches of size 2x2, the weight is of size 4xd and biases are applied to each vectorised patch (where d=4). This is then stored in matrix of size 196x4 (where 196 is number of patches and 4=d). Figure 1 shows a visual representation of the stem structure:

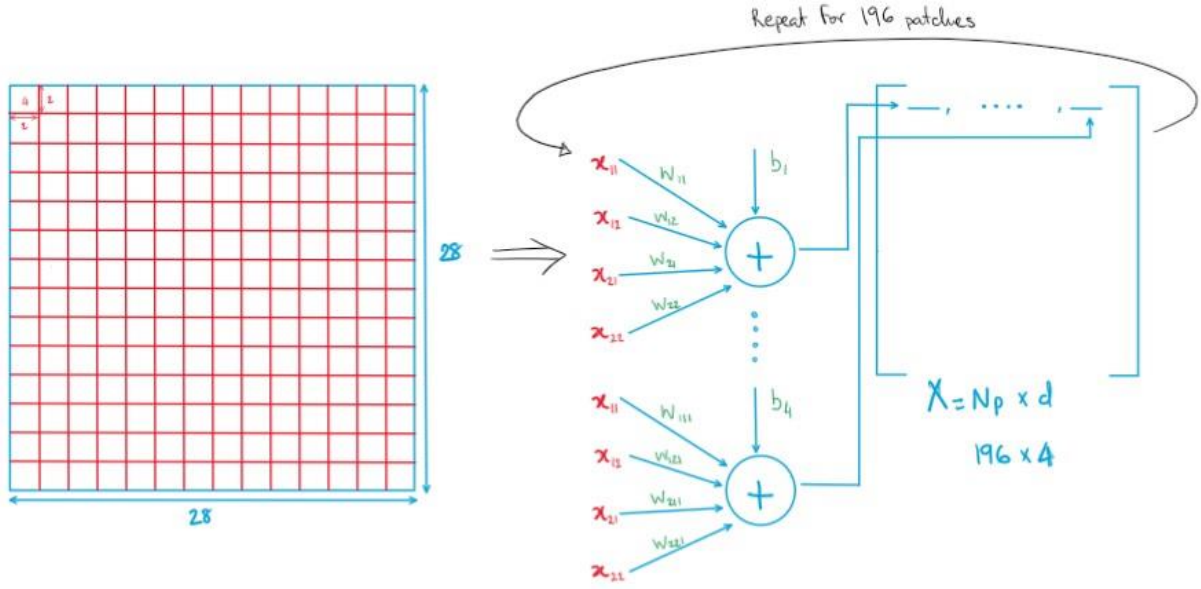


Figure 1

Backbone and classifier: the backbone consists of 4 blocks; the first block uses the sigmoid activation function, and the rest of the blocks use the tanh activation function. Additionally, the mean feature is calculated and fed into a SoftMax regression classifier to output 1x10 values. Figure 2 shows a visual representation of the backbone and classifier structure:

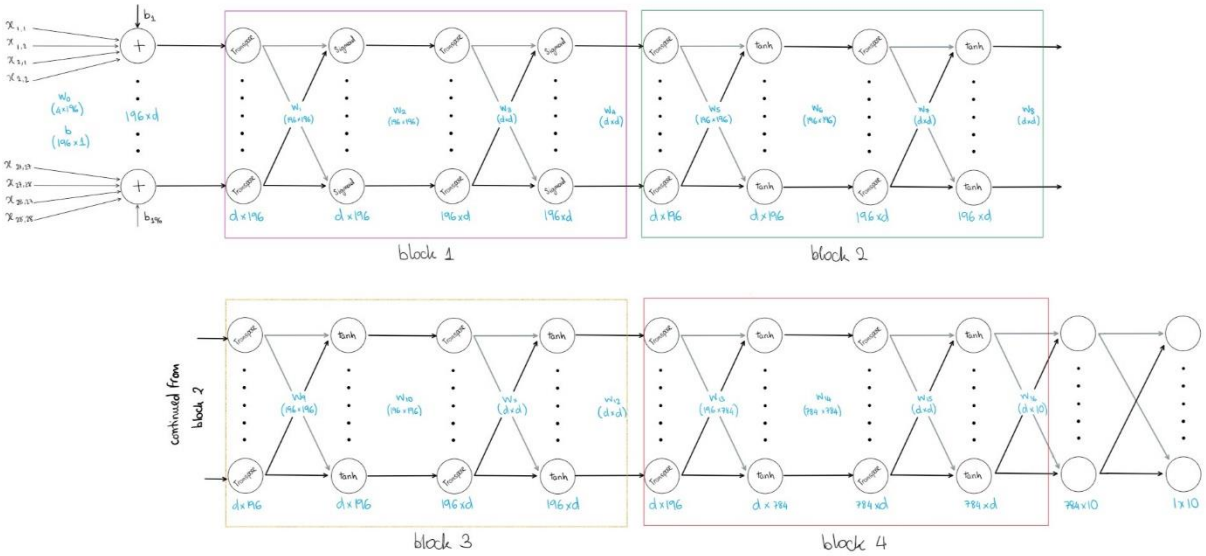


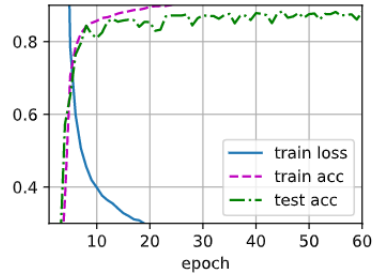
Figure 2

Training the network:

```
start_time = time.time() #initiate the timer

loss = nn.CrossEntropyLoss() #loss fuction
wd,lr = 0.0, 0.9 #weight decay and learning rate
#optimization algorithm
optimizer = torch.optim.SGD(model.parameters(),weight_decay=wd, lr=lr) #stochastic gradient descent
```

```
num_epochs = 60 #number of epochs
mu.train_ch3(model, train_iter, test_iter, loss, num_epochs, optimizer) #SoftMax regression classifier
elapsed_time = time.time() - start_time #calculate the time taken to train the model in seconds
```




```
accuracy= mu.evaluate_accuracy(model, test_iter)
print(accuracy) #print the test accuracy
```

```
0.8816
```

```
print(elapsed_time/60) #output the time taken to train the model in minutes
```

```
16.859029126167297
```

Figure 3

Figure 3 shows a graph illustrating the curve for the evolution of loss (train loss), the curve for the evolution of training (train acc) and the curve for the validation accuracies (test acc).

In this instance it can be observed the test accuracy is 88.16%, to train the network for 60 epochs the time required was around 17 minutes. The learning rate is 0.9, the weight decay is 0 as no significant overfitting issues can be observed, the batch size is 256. Overall, the network accuracy is consistently above 87% and the time required is below 20 minutes.