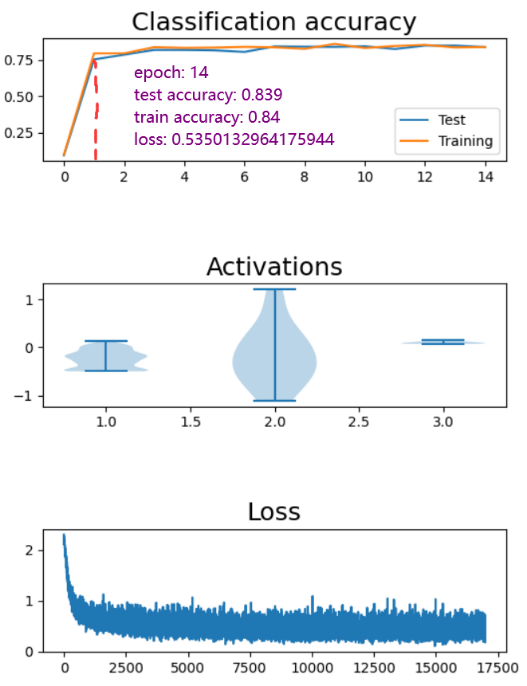
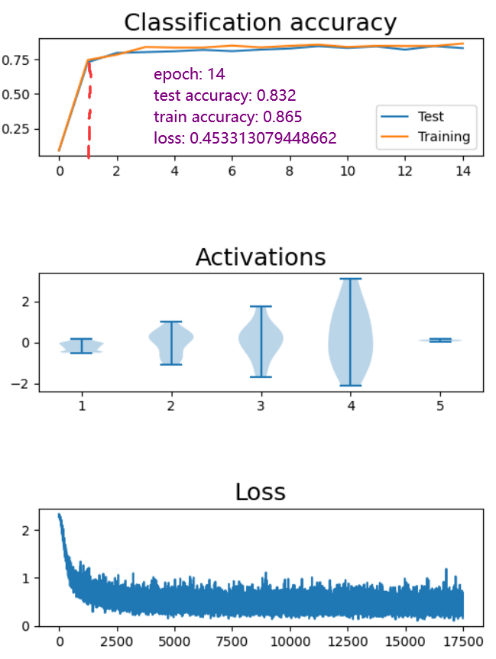
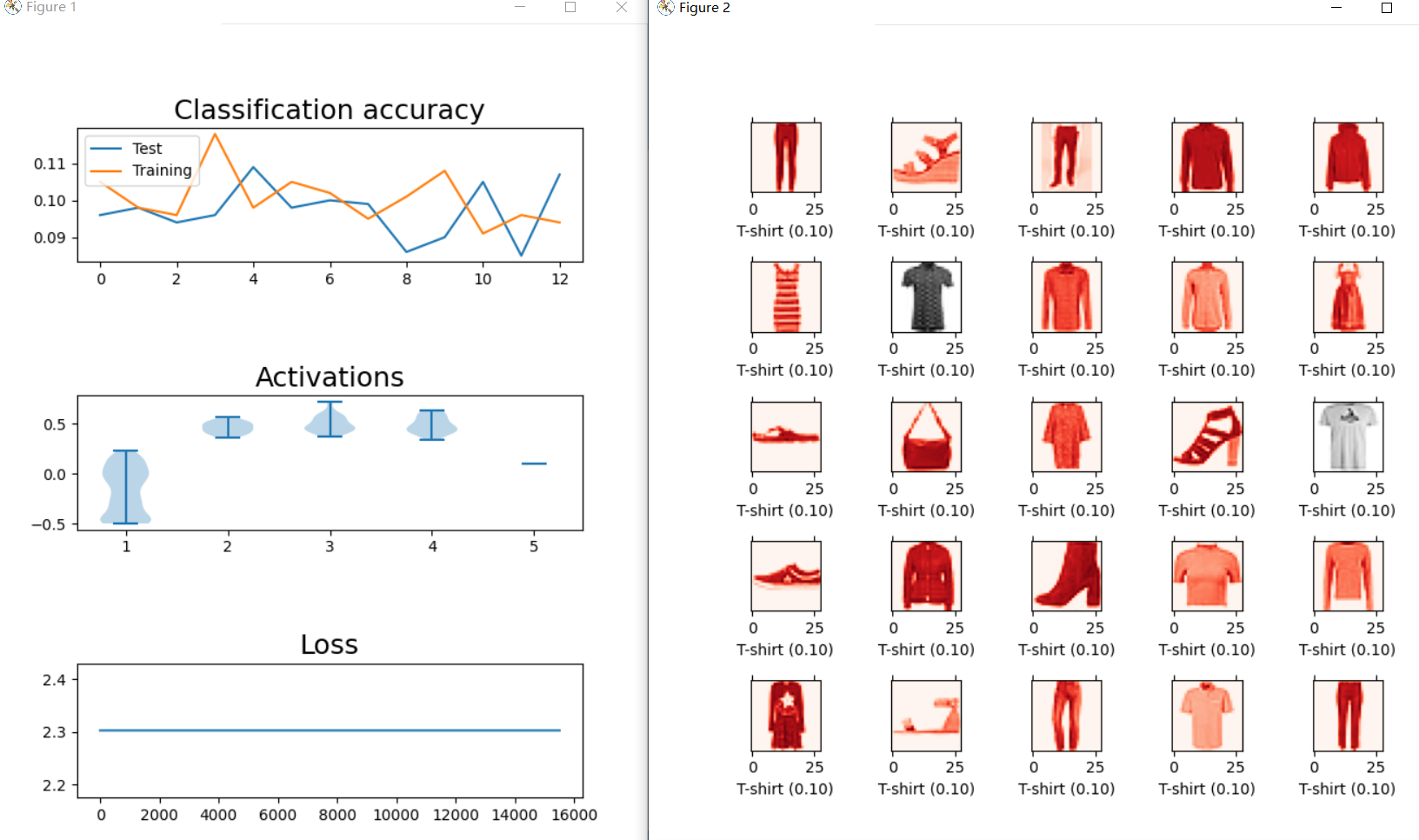
Activation Function

1. If we use the Identity activation function (i.e. a[i] = z[i]) in hidden layer and softmax function as output layer, the performance (Image 1.1) is good. It indicates that this might be a potentially linear problem. To verify, we use one hidden layer instead of three, because the last layer is always the linear function of the first layer, no matter how many hidden layers there are in the neural network. As expected, the performance (Image 1.2) is almost the same as 3 hidden layers.

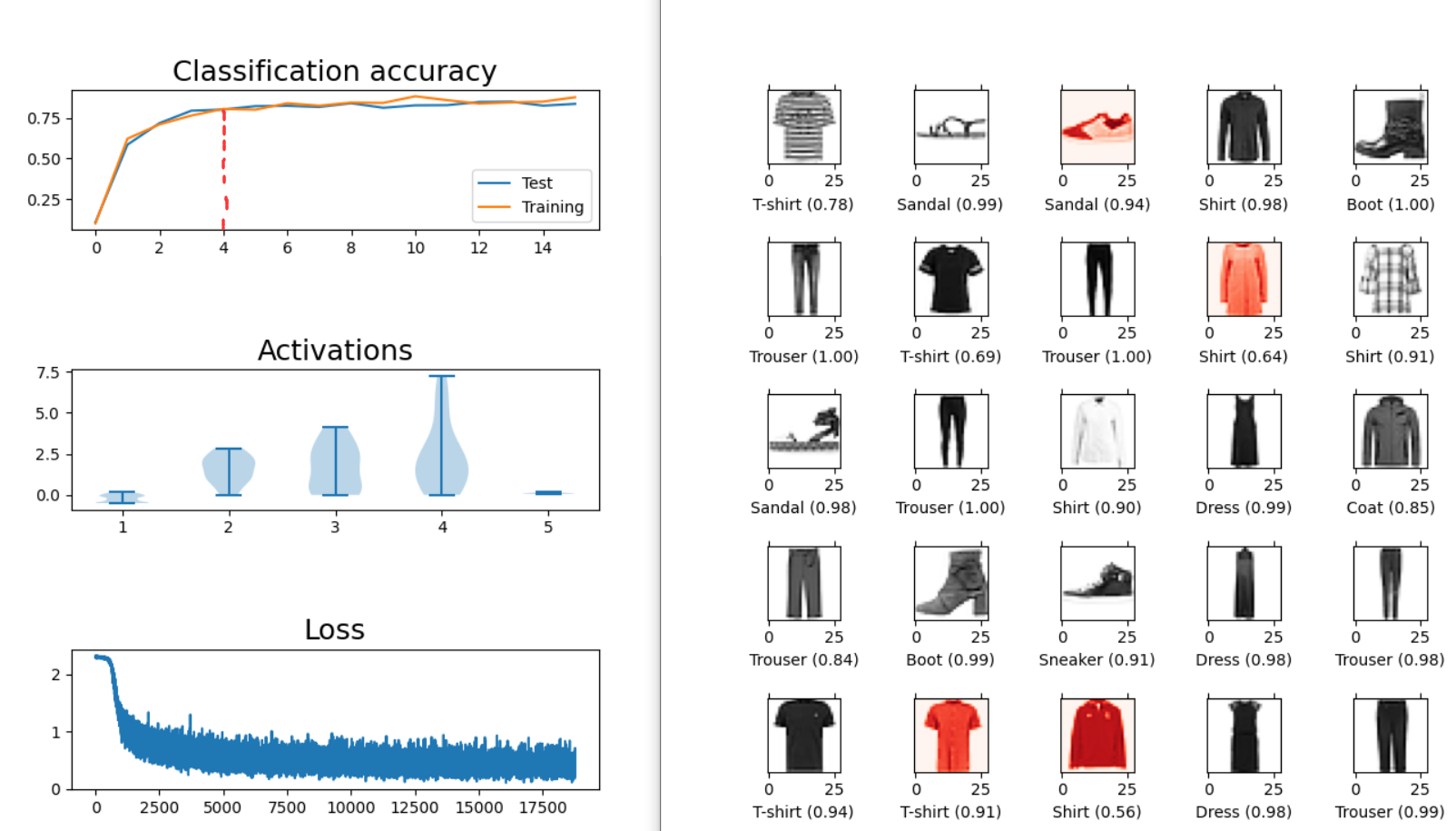




1. If we use sigmoid as activation function in hidden layer and SoftMax function for output layer, the accuracy fluctuates around 0.1 and the loss fluctuates around 2.3. This is because when backpropagation through the derivative of the sigmoid, the point which far from the 0, the gradient becomes very small or even disappears. It also called vanishing gradient problem. This cause the weight has almost no effect to the loss function. Thus, here the loss almost not change. In addition, the sigmoid function is exponent arithmetic, which has high computational cost.
2. If we use sigmoid as activation function in hidden layer and SoftMax function for output layer, the accuracy growing very slowly and the loss is still high. The image below is about 20 minutes runtime but only with 0.25 accuracy. The accuracy can keep growing but will cost a lot of time. This is because when backpropagation through the derivative of the sigmoid, the point which far from the 0, the gradient becomes very small or even disappears. It also called vanishing gradient problem. This cause the weight has low or almost no effect to the loss function. In addition, the sigmoid function is exponent arithmetic, which has high computational cost.



1. If we use Relu as activation function in hidden layer, and SoftMax function for output layer, we can obtain a high-performance model. It can make prediction accurately. ReLU function learning fast as the accuracy becomes ideal with x-axis only 4.



1. If we use ReLU as activation function in hidden layer, and Sigmoid function for output layer, it has a good performance in this dataset. In this case Sigmoid function has the similar action as SoftMax. They both convert a value to range (0,1), the difference is all outputs of SoftMax are sum to 1. Thus, SoftMax is often used to solve the multi categories problem and Sigmoid often used to solve the problem with only 2 categories.

