Part III (Bonus question) - Exploring Out-of-Distribution (OOD) generalization on Colab (15pts) [15min]

3.5 Translation and rotation (15pts) We modify the original test datapoints, by moving up the images of the test set by 4 pixels. By doing this, we create a new translated test set. Run the code to see the original and modified images. You will notice that to a human eye, the new test set is not any more difficult than the original test set.

(i). But can our MLP model still do well on the new test set? What's the test accuracy on the two-layer MLP? (1pt)

Our MLP model does worse on the new test set with a test accuracy of 0.45559999346733093.

(ii). What if we try a different model architecture? For example, in class we saw that convolutional neural networks are good at dealing with image translation. We replace the first linear layer with a convolutional layer of 64 kernels with size 7 × 7, followed by max-pooling. This can be done with just one or two lines of code in TensorFlow. Calculate the number of parameters of the CNN model and the original 2-layer MLP model (show your calculation), and verify that the CNN has fewer parameters than the MLP model. (3pts)

Parameters of CNN: (7*7*1+1 [bias])*64 + 0 [no params for pooling] + 0 [no params for flattening to 7744 elements] + (7744+1)*10 [dense layer with 10 nodes] = 80,650 parameters
Parameters of MLP: (784+1)*128 [layer 1 with 128 nodes] + (128+1)*10 [layer 2 with 10 nodes] = 100480+1290 = 101,770 parameters
As expected, the CNN parameters of 80,650 are less than the 101,770 parameters of the original 2-layer MLP model, so the CNN model is more efficient in terms of parameters.

(iii). Train the 2-layer CNN on the original training data and test it on both the original and the translated test sets. What is the in-domain test accuracy and the translated test accuracy of the CNN model? Can you provide some intuition behind these numbers? (2pts) [6min]

in-domain test accuracy: 0.8708999752998352
translated test accuracy: 0.5270000100135803
The in-domain test accuracy is better than the translated test accuracy as expected, since the original training data equips the CNN model to better classify on the non-altered test set, while translating the data by 4 may cause a huge difference in whether the model can classify correctly, since the pixels in the translated set may be cut off and not show the full image, unlike the original test set, resulting in a worse test accuracy.

(iv). Going one step further, we can make the CNN deeper by adding one more convolutional layer of 64 × 128 (64 input channels and 128 output channels) kernels with size 2 × 2 between the two existing layers, followed by max-pooling. Verify that the deeper 3-layer CNN model has fewer parameters than the 2-layer CNN model (show your calculation). (3pts)

Parameters of CNN: (7*7*1+1 [bias])*64 + 0 [no params for pooling] + (2*2*64+1 [bias])*128 + 0 [no params for pooling] + 0 [no params for flattening to 3200 elements] + (3200+1)*10 [dense layer with 10 nodes] = 68,106 parameters
As expected, the deeper 3-layer CNN model has less parameters than that of the 2-layer CNN model.

(v). What is the in-domain and the translated test accuracy of the deeper 3-layer CNN model? Provide some intuition on why it is better than the 2-layer CNN on the translated set. (3pts) [9min] You will notice that the deeper model takes some time to train. If we trained on GPUs instead of CPUs, training would be much faster. This is because GPUs are much more efficient at matrix computations, which is exactly what neural network training demands.

In-domain test accuracy: 0.8718000054359436
Out-of-domain test accuracy: 0.5652999877929688

It may be better than the 2-layer CNN due to having more layers to extract more features of the translated test data and capture more complexity, so the 3-layers can identify more details of the images and thus have a higher chance of classifying the image correctly and a higher test accuracy.

Next, we create 3 more OOD test sets by rotation. We rotate the images in the original test set by 90, 80 and 270 degrees.

(vi). Provide the test accuracy of the 2-layer MLP model, the 2-layer CNN model and the 3-layer CNN model on the three rotation test sets. Are the 2-layer CNN and the 3-layer CNN still doing well? (3pts)

Rotation 90 degrees:
MLP model accuracy: 0.020800000056624413
2-layer CNN model accuracy: 0.05009999871253967
Deeper 3-layer CNN model accuracy: 0.052799999713897705

Rotation 180 degrees:
MLP model accuracy: 0.18690000474452972
2-layer CNN model accuracy: 0.21649999916553497
Deeper 3-layer CNN model accuracy: 0.04650000110268593

Rotation 270 degrees:
MLP model accuracy: 0.05530000105500221
2-layer CNN model accuracy: 0.04740000143647194
Deeper 3-layer CNN model accuracy: 0.04650000110268593

Both forms of CNN are not doing well on the rotated test tests. The 2-layer CNN model seems to do better than the 3-layer by a good amount for the 180-rotated test set and slightly better for the 270-rotated test set. However, overall all test accuracies, whether MLP or 2-layer or 3-layer CNN models, have massively decreased due to the rotation in the test sets.

Deliverables for Problem 3: Code for 3.1 as a separate Python file `neural_networks.py`. Screenshot for 3.1. Plots for part 3.1, 3.2, 3.3 and 3.4. Number of epochs and gradient updates for part 3.2. Analysis and explanation for parts 3.2, 3.3, 3.4. For the optional bonus question 3.5, submit the test accuracy on in-domain and OOD test sets of the three models, the number of parameters of the three models, and the explanations we ask for.