Problem 1

part a:
Padding generates image of a size 12×12. Kernel is 3×3 and stride is 3. Hence, 4 kernels can fit columns and 4 kernels can fit rows. **n=4, m=4**. In this case, k=m*n. **k=16**.

part b:
$\mathbf{W}_{conv}$: kernel size $\times$ channels, $\mathbf{W}_{conv} : 3 \times 3$,
For the section that k $\times$ 1 is input and 10 $\times$ 1 (after fully connection) is output, b must have the same dimension as the output, $\mathbf{b} : 10 \times 1$. System needs to turn a 10 $\times$ 1 matrix after multiplying the input via $\mathbf{W}_{fc}$ so, $\mathbf{W}_{fc} \cdot (\mathrm{k} \times 1) : 10 \times 1$. Where k has been found 16, $\mathbf{W}_{fc} : 10 \times 16$.

Problem 2

List of η's used for this problem is $[1e-5, 1e-4, 1e-3, 0.01, 0.1]$.

Test Set Loss/Error Rate for SGD					
η	1e-5	1e-4	1e-3	0.01	0.1
Loss	2.07339	0.49175	0.11177	0.00675	0.00001
Error Rate	0.5388	0.1403	0.0732	0.0280	0.0222

Test Set Loss/Error Rate for Adagrad					
η	1e-5	1e-4	1e-3	0.01	0.1
Loss	1.82669	0.46747	0.10285	0.00343	0.00114
Error Rate	0.4467	0.1258	0.0653	0.0227	0.0309

Test Set Loss/Error Rate for RMSprop					
η	1e-5	1e-4	1e-3	0.01	0.1
Loss	0.12837	0.00451	0.02352	0.11376	1.83383
Error Rate	0.0747	0.0254	0.0279	0.0775	0.7350

Test Set Loss/Error Rate for Adam					
η	1e-5	1e-4	1e-3	0.01	0.1
Loss	0.12498	0.00469	0.00004	0.00835	2.29933
Error Rate	0.0737	0.0254	0.0265	0.0706	0.8972

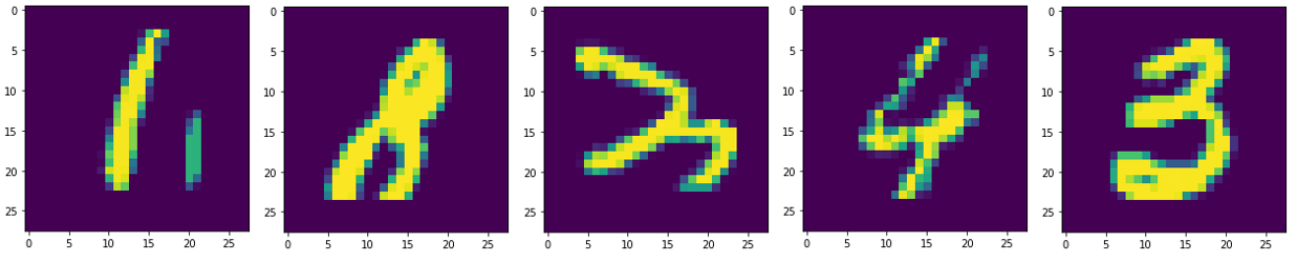
I applied 10 epochs for each optimizer & learning rate pair. This is very low in general but they yielded reasonable results and I didn't want to wait hours to see test accuracy results. As expected, larger learning rates yielded better results in the same number of epochs until over-fitting starts. However, if we had the opportunity to apply high numbers of epochs, using smaller learning rates is safer to avoid any sudden divergence. Using larger learning rates may cause over-fitting and therefore divergence. According to information on tables, SGD with $\eta = 0.1$ was the best. However, Adagrad is better in general, especially while using smaller learning rates and using smaller learning rates is safer.

Problem 3

List of η's used for this problem is $[1e-5, 1e-4, 1e-3, 0.01, 0.1]$.

Test Set Loss/Error Rate for SGD					
η	1e-5	1e-4	1e-3	0.01	0.1
Loss	2.03939	0.57878	0.24927	0.01622	0.00468
Error Rate	0.6997	0.2200	0.0831	0.0388	0.0280

According to information in the above table, best learning rate is $\eta = 0.1$ for SGD using CNN. Loss = 0.00468, and error rate = 0.0280.



I randomly selected 5 images which the model incorrectly predicted. You can see the correct labels and the incorrect labels that model predicted below the images.
With increased number of epochs, we could see that the model would be successful in correctly predicting labels of more images. I tried different numbers of epochs. With less epochs, there were more incorrectly predicted labels and high number of epochs gives better results.

For the first image, the separate greenish part on the right probably caused the model to be mistaken. For the second image, it looks like a bold 1 and the incomplete line at the bottom caused the model not to understand that the image was 8. For the third image, that is a very weird 3 and it reminds me of 7 in shape. For the fourth image, it looks like a 9 in shape and the model probably couldn't distinguish the incomplete line above or there weren't enough 4s in the data set to train. For the fifth image, the model was probably confused by the bump at the bottom of the image because it looks very similar to the bump in 5.