Question1

The best output was selected by doing a grid search of hyperparameters. Therefore, the best output was found at batch size 64, learning rate 0.001. The number of epochs used was 5 and early stopping was used with a patience of 2. This setup was used for all three optimizes and SGD was found to provide the best accuracy. The highest accuracy was obtained with model1 and SGD.

The considered batch sizes are: 16, 32, 64, 128.
The considered learning rates are: 0.01, 0.001, 0.0001.
The selected batch size and learning rate are 64, 0.001 with 5 epochs.

Model	SGD	Adam	RMSprop
Model1	99.18	98.93	98.86
Model2	99.10	98.87	98.49
Model3	99.11	98.24	97.95

Here all accuracy are in percentage of test accuracy.
So, the model1 where filters increased in each layer obtained the highest accuracy with SGD.
My observation is taking a small batch size takes too much time for convergence and a larger batch size reduces the accuracy. In all cases SGD performs better than Adam but it takes a slightly more time than Adam.

Question2

1. For a lower learning rate it does not converge to minima for the number of epochs. For a higher learning rate it keeps jumping over minima. The best learning rate was found to be 0.0001 for my setup.
2. Too large of a batch size leads to poor generalization. Using a batch equal to the entire dataset guarantees convergence to the global optima of the objective function. However, this is at the cost of slower, empirical convergence to that optima. The optimal batch size selected in the problem is 64.
3. The hyperparameters were Adam optimizer with learning rate 0.0001, batch size 64, a 20% train-validation split was used to observe the performance. Categorical crossentropy was used as the loss function. The best test accuracy obtained was 54.57%.
4. a) The feed forward neural network fails to perform with the given size of dense layer. It obtained only 11.10% validation accuracy. However, using a different layer size it could obtain highest 27.46% test accuracy for 25 epochs.
b) The total parameters for the FFNN is 31604. It is not worth it as the test accuracy is pretty lower than a CNN.

Question3

1. The filter size is 3x3, image size is 6,6,1. Both image and kernel dimension is 2 and the depth of the Image and kernel if 1.
2. The activation map is calculated with stride convolution. The function can be found in the code.
3. 2x2 Max pooling is applied. The output can be found in the code.