1. After each stride-2 conv, why do we double the number of filters?

2. Why do we use a larger kernel with MNIST (with simple cnn) in the first conv?

3. What data is saved by ActivationStats for each layer?

4. How do we get a learner's callback after they've completed training?

5. What are the drawbacks of activations above zero?

6.Draw up the benefits and drawbacks of practicing in larger batches?

7. Why should we avoid starting training with a high learning rate?

8. What are the pros of studying with a high rate of learning?

9. Why do we want to end the training with a low learning rate?