# Supervised Nets and Knowledge Distillation

**Paper 1:** Deeply Supervised Nets

**Paper 2:** Network in Network

**Paper 3:** Distilling the Knowledge in a Neural Network

1. How deep supervision in a network help it learn highly discriminative feature maps?

2. What are the three aspects in convolutional neural network style architectures are being looked at in paper 1?

3. How deep supervision is different from normal supervised training? How deep su pervision affects (a) Training, (b) Backpropagation

4. What differences you observe in the feature visualization shown in Figure 3 of pa per 1. Explain briefly

5. Label smoothing is a technique to train a model to be less confident about its prediction by smoothing out the one-hot vector, e.g., from [0, 1, 0] to [0.1, 0.8, 0.1]. How the soft labels generated using a teacher model is different from it? Explain.

6. How the SoftMax activation is modified to control the soft labels generated by the teacher model? What is the relation between the temperature T and soft labels produced?

7. What is the implicit assumption made by a conventional CNN model? How paper 2 address this problem? Explain

8. The MLP is shared among all local receptive fields” Explain this sentence from paper 2.

9. Define Global Average Pooling (GAP) and state its advantage(s). Also, explain how the application of GAP is more robust to spatial translations?

10. Describe each term briefly: (a) Companion Objective, (b) deep supervision, (c) Teacher and student networks, (d) Knowledge Distillation