Summary for Lecture 7

1. Optimization
2. SGD: Very slow progress along shallow dimension, local minima and saddle point problem
3. SSD + Momentum: build up ‘velocity’ as a running mean of gradients
4. Nesterov Momentum: different way of taking actual step
5. AdaGrad: added element-wise scaling of the gradient based on the historical sum of squares
6. RMSProp: reflect the sum of square by updating itself
7. Adam: uses both momentum and AdaGrad/RMSProp method
8. First Order Optimization & Second Order Optimization
9. Regularization
10. Model Ensembles: train multiple independent model and average their results at test time
11. Add term to loss: R(W)
12. Dropout: randomly set some neurons to zero in forward passs and prevents co-adaptation of features
13. Data Augmentation: transform the images in various way
14. Transfer Learning
15. We don’t need too much data to use CNN because already learned weights will be shared in frozen part