**Summary for lecture 6, Training Neural Networks**

1. Activation Functions
2. Sigmoid: Gradient Descenting problem, compute expensive due to exponential calculation, outputs are not zero centered
3. Tanh: can solve two problems of sigmoid function but still gradient descenting problem
4. ReLU: treat the gradient problem and actually more biologically plausible than sigmoid but not zero centered output. Should initialize with slightly positive biased since there can be dead ReLU problem
5. Leaky ReLU
6. ELU
7. Maxout ‘Neuron’
8. Data Preprocessing: to get zero mean data and get hypothetically optimal weights
9. Weight Initialization
10. Small random numbers: problems in deeper networks
11. Xavier initialization:
12. He et el
13. Batch Normalization: get gaussian normalized outputs and improve gradient flow through the networks
14. Babysitting the learning Process
15. Hyper-parameter Opttimization
16. Cross-validation
17. Deal with network architecture, learning rate, decay, regularazation