Deep Learning Course 2 - Week 1 (Jan 12, 2018 11:00 am)
Practical aspects of Deep Learning

Definitions/Notes

- Splitting data
  - The train set is used to fit the model
  - The dev set is used to tune hyperparameters
  - The test set is used to evaluate the model
- Rule of thumb
  - Old: 70/30/. or 60/20/20 split
  - New (with big data): 98/1/1 or 99.5/0.25/0.25 split
- Basic recipe
  - If high bias then train a bigger network or train for longer
  - If high variance then get more data or apply regularization
- Bias Variance Tradeoff
  - Small bias and variance is better
  - With deep learning there are methods with very little tradeoff between minimizing bias/variance; i.e., we can decrease bias without increasing variance (much), and we can decrease variance without increasing bias (much).
- Regularization
  - Helps to prevent overfitting
  - L1 regularization (also called lasso regression) produces sparse models
  - L2 regularization (also called ridge regression or weight decay) is more common than l1 regularization
  - The regularization hyperparameter lambda is chosen using cross-validation.
- Dropout
  - A regularization technique that randomly drop neurons and set their connections equal to zero
  - Inverted dropout: drop units with a probability p, then divide the activations by 1-p.
  - Dropout forces you the model to learn a “smaller” neural network
  - With dropout, the neural network can rely on any one feature too much.
- Data augmentation
  - Intuition: create synthetic data that are realistic to train on (more data helps training)
  - For example, we can mirror flip images
- Early stopping
  - Monitor dev set error and stop training when dev set error starts to increase
- Normalize inputs
  - Subtract the mean
  - Divide by the standard deviation
  - This puts the inputs on relatively the same scale which helps gradient descent converge
- Weight Initialization
  - Careful initialization of the network’s weights can help with training
  - If using a relu activation then then initialize as a normal with std=sqrt(2/n)
  - If using tanh activation then initialize as a normal with std=sqrt(1/n) or use xavier initialization

Questions

- When is it okay to not have a test set (and only a dev set)?
  If we only care about tuning the model and not getting an unbiased estimate of the generalization error then we only need a dev set and not a test set.

- Why is it important for the dev set comes from the same distribution as the test set.
  If we have a test set, then test performance is what we care about and so it makes sense to have a dev set similar to the test set.
- Why does regularization prevent overfitting?
  See Statistical Learning Ch6 Notes.
- Does the size of the dev set always need to be the same as the test set?