STAT 526

Topic X
Optimization and Training of Deep Learning II

Dr. Qifan Song

SGD convergence for DNN

In the previous lecture https://proceedings.mlr.press/v202/zhu23h/zhu2

- Failed conditions
- Overparamerization
- Landscape
- Lazy training for shallow network

Implicit Bias

- Reduce model size by randomly dropping out hidden/input nodes. Avoid overfitting.
- During each iteration of training, hidden (input) nodes are discarded from network (i.e., output from them are fixed to be 0) with probability $p \approx 0.5$ (or smaller). Hence model updates will not use these node.
- During the testing, all nodes are retained, but their output is multiplied by p. This can be viewed as a surrogate of model average of sparse networks.
- Originally for MLP; Some modification is need for CNN layer (pooling dropout/cutout) and RNN (variational dropout: drop the same network units at each time step)

SGD and generalization

- Towards Theoretically Understanding Why SGD Generalizes Better Than ADAM in Deep Learning

NTK

A sufficiently over-parameterized random network contains a sub-network that achieves high accuracy without any training.

- Compare to UAT, (S)LTH claims the existence of an approximation within an random dense network.
- In practice, usually, 90% or more reduction in parameters without significant loss in accuracy
- Finding winning tickets typically requires iterative pruning and retraining

Neural Collapse

- High-dimensional data (such as images, text, or audio) actually resides on or around a lower-dimensional, structured manifold (or the union of multiple low dimensional manifold) embedded within the high-dimensional ambient space.
- It justifies the usefulness of nonlinear dimensional reduction technique
- Manifold learning algorithm