Cyclical Learning Rates for Training Neural Networks

Gobind Puniani

MATH 190: Fall 2019

09/19/19

Learning Rate

- Learning rate is one of the factors involved in updating the parameters of a neural network
    - Stochastic gradient descent
    - Too small = slow convergence
    - Too large = divergence
- Learning rate is usually adjusted experimentally to find relative optimal value

Gradient Descent Equation

- Stochastic Gradient Descent: $\theta^t = \theta^{t-1} - \epsilon_t \frac{\partial L}{\partial \theta}$
  - θ represents the weight parameter, L is a loss function, and ϵ_t is the learning rate

Cyclical Learning Rates

- Traditional approach: pick a learning rate value that decreases during training
- New approach: let learning rate cycle through a set of values
- Benefits of Cyclical Learning Rates (CLR):
    - More accurate
    - Removes the guesswork in changing the learning rate
    - Does not require additional computation (unlike adaptive learning rate methods, which set local values instead of a global one)

Experiment

- CIFAR-10 dataset
  - Tuning method: 81.4% accuracy after 70,000 iterations
  - CLR method: Same accuracy after 25,000 iterations
- CLR outperformed tuning method in accuracy for CIFAR-10 and CIFAR-100

CLR Optimization

- Original observation: increasing learning rate has short-term negative effect but long-term positive effect
- Different functional forms yielded same results:
  - triangular window (linear)
  - Welch window (parabolic)
  - Hann window (sinusoidal)
- Triangular learning rate policy used because it is simplest
- Stepsize = number of iterations in half-cycle
- Boundaries determined by “LR Range Test”
  - Run through several epochs, set stepsize to maximum number of iterations, and note where the accuracy increases and peaks

Future Areas of Research

- Training other architectures, such as recurrent neural networks
- Combining CLR with adaptive learning rate methods
- Theoretical analysis needed for full understanding