Systems for Machine Learning (CMSC828G)

Optimizing DL Kernels

Abhinav Bhatele, Daniel Nichols

UNIVERSITY OF
MARYLAND

Announcements

- Interim report for the project is due on April 17

Machine learning modifications for systems optimizations

- Optimizers
- Mixture of experts and grouped GEMMs
- Offloading data to CPU

Optimizers in deep learning

- Used to adjust parameters to minimize loss
- Critical for effective model convergence
- Types of optimizers:
  - First-order: rely only on first-order gradients
  - Second-order: use both first-order gradients and second-order derivatives (Hessian matrix)

First-order optimizers

- Advantages:
  - Computationally efficient
  - Scale well for large models
- Examples:
  - SGD: Stochastic Gradient Descent
  - AdamW: Adaptive Moment Estimation with weight decay
- Why is AdamW popular:
  - Effective balance of speed and stability
  - Robust to different hyper parameters such as batch sizes, learning rate, weight decay

Second-order optimizers

- Advantages:
    - Faster convergence
    - Better suited for complex loss landscapes
- Examples:
    - Newton’s method
    - K-FAC: Kronecker-factored Approximate Curvature
    - Shampoo
- Challenges: computationally expensive

UNIVERSITY OF MARYLAND