Deep Residual Learning for Image Recognition

Gobind Puniani

MATH 190: Fall 2019

09/26/19

Network Depth

- The accuracy of deep learning networks depends on number of layers (depth)
- Hypothesis: improve networks by stacking more layers
  - Wrong!
- “Degradation”: accuracy peaks and then declines with deeper networks
  - NOT caused by overfitting
  - Higher training error

Performance of Deeper Networks

- If a given network has layers added, there should be a solution by construction for which the training error does not increase
  - Added layers are “identity mapping,” while other layers are the same
  - Experiments unable to find such a solution
- Proposal: deep residual learning framework

Deep Residual Learning

- Layers fit to a residual mapping
  - Easier to optimize
  - Can be made to resemble identity mappings if such mappings are optimal
- “Shortcut connections” are used in the forward direction of the network
  - One or more layers skipped
  - Identity mapping with outputs added to outputs of stacked layers
  - No extra parameter or computational complexity

Findings

- Proposed model has fewer filters and lower complexity (3.6bn FLOPS vs. 19.6bn FLOPS)
- Extremely deep residual networks are easy to optimize
    - 18-layer ResNet converges faster than 18-layer plain network (similar accuracy)
- “Plain” networks (stacked layers) have higher training error with increased depth
    - Higher training error with 34-layer plain network than 18-layer plain network
- Deep residual networks experience higher accuracy with increased depth, unlike previous networks
    - 34-layer ResNet is better than 18-layer ResNet (much lower training error)

Questions

- What is the difference between identity mapping and residual mapping?
- What are top-1 and top-5 error rates?
- What is a bottleneck architecture?