Tricks, and going farther

Sunday, August 25, 2024 9:57 AM

Tricks of the trade

- Self-supervised learning (image masking)
- data augmentation (eg, add shifts, rotations, reflections ...)
- "ablation" studies to determine which aspects of a successful architecture are needed
- batch normalization (helps w/ vanishing gradients)
- residual networks (resnets) i.e. skip-connections
  learn perturbations of identity. Helps w/ vanishing gradients
- gradient clipping (for exploding gradients)
- dropout (to regularize)
- proper initialization (i.e., variance depends on layer)
  Ex: "Glorot (Xavier) initialization" or, for ReLU, "He initialization"
- fancier optimization
  - momentum, acceleration, adaptive stepsizes
    AdaGrad, RMSprop, Adam
  - approx. 2nd order methods: KFac
  - Don't do pure Newton (even if computationally feasible) due to nonconvexity
    nor do nonlinear CG (instead, L-BFGS way more stable)

Field is changing rapidly! New techniques all the time, old ones fall out of favor

- check internet, blogs, ...
- "The Deep Learning Book"
  (reliable though not up-to-date), ch. 8 especially