Deep Residual Learning for Image Recognition

A Review

Abhishek Ranjan

March 10, 2020

Toronto, ON

Table of contents

The Problem

Difficulties in Deep NN Training

1. Vanishing/exploding gradients; claimed to be addressed by:
   - Batch Normalization in forward propagation
   - ReLU + weight initialization
2. Degradation problem (Figure 1):

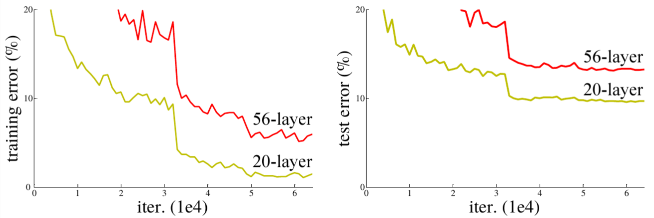

Figure 1: Degradation problem for CIFAR-10

Degradation problem

Degradation problem posed as originating from the solver

- Increasing depth $\not\Longrightarrow$ Low error
- Training error is also high $\Longrightarrow$ Not an overfitting problem
- Possible causes: difficulty in learning Identity functions

Problem statement: **Can we increase depth without degrading accuracy?** [1]

Fundamental Ideas

Proposed solution

Make it easy for the solver to reach identity mapping

- Force identity shortcuts for x to reach the output
- If the original block learns $H(x)$, after shortcut it learns $F(x) = H(x) - x$

Why residuals?

- Residuals are easier to learn (evidenced by other areas)
- Solver can move the weights to zero more easily than to identity mapping
- No extra parameters (for most cases)

Note: No mathematical justification provided in the paper, mostly empirical.

Architecture example

1. Start with a deep, plain network, e.g., a VGG n/w with 34 layers (3x3 Convolutional filters)
2. Add shortcut connections (Figure 2)
    - Identity connections every 2 blocks when dimensions match

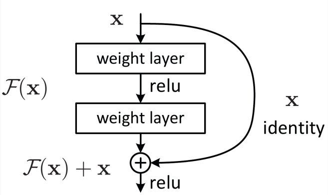

Figure 2: Identity shortcut example

    - When dimensions don't match:
        2.1 Use identity mapping with zero-padding (no extra params)
        2.2 Use projection shortcut (more params!)

Training

- Train as usual with SGD + backpropagation
- Batch normalization after convolution before ReLU
- Learn rate $\eta = 0.1$, weight decay $\lambda = 0.0001$, momentum $\alpha = 0.9$
- Weight initialization [4]: $N(0, \sigma = \sqrt{(}2/fan_{in}))$
- For bottleneck architecture: use identity shortcuts to avoid complexity

Results and Conclusions

Accuracy

Significant improvements in training and validation error across the board

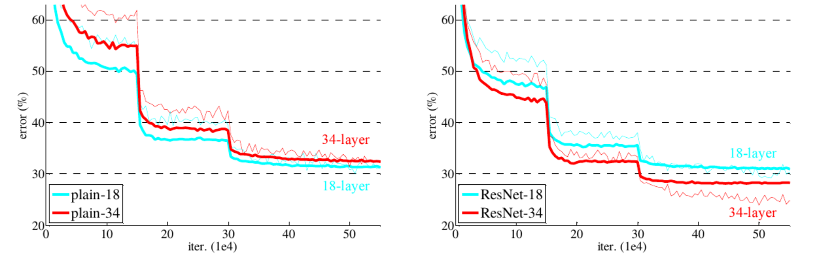

Figure 3: Training (bold line) and validation (thin line) error on ImageNet

Layer responses

ResNet layer responses are smaller $\implies$ Identity shortcuts pass most information

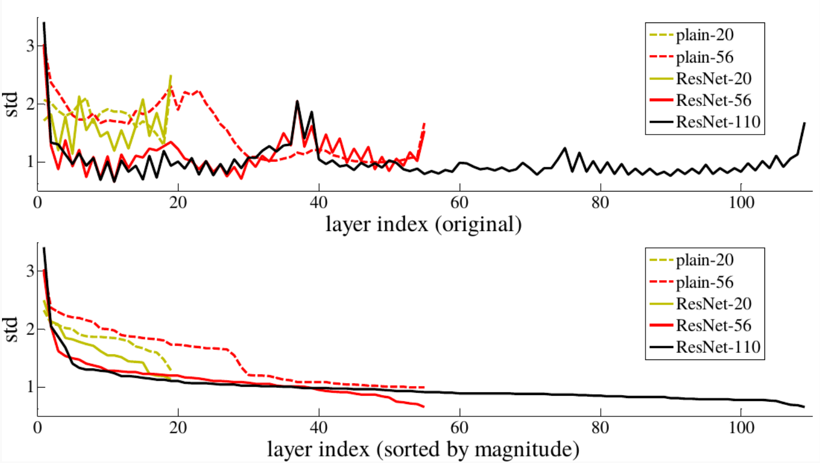

Figure 4: Standard deviation of layer responses on CIFAR-10

Conclusions

- Proposed a new architecture using residuals to address degradation problem
- Identity shortcuts mostly add no extra parameters
- Significant increase in number of layers possible
- Provided strong empirical evidence of its effectiveness across a wide range of tasks
- Opens up many research areas and raises questions:
    - What is the nature of the degradation problem?
    - Is the degradation problem really separate from unstable gradients problem?
    - How do shortcut connections allow deeper networks (mathematically)?

Questions and Discussion

An ensemble of shallow networks [3]

- A collection of many paths of different length (Figure 5)
- Avoids vanishing gradients by leveraging only the short paths
- Layers behave like ensembles

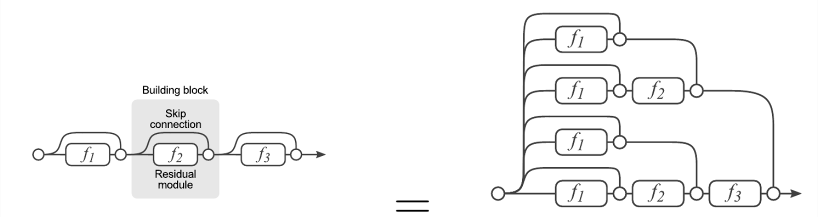

Figure 5: Unraveled paths of residual connections

ResNet as boosting over features [2]

- ResNet output $\sim$ layer-by-layer boosting

$$\begin{aligned} & g_{t+1}(x) = f_t(g_t(x)) + g_t(x) \\ \Longrightarrow\ & g_{t+1}(x) - g_t(x) = f_t(g_t(x)) \\ \Longrightarrow\ & g_{T+1}(x) = \sum_{t=0}^{T} f_t(g_t(x)) \\ & s.t., g_0(x) = 0, f_0(g_0(x)) = x \end{aligned} \tag{1}$$

- ResNet boosts over *representations* and *not estimated labels*

References i

- He, Kaiming, et al. "Deep residual learning for image recognition." *CVPR*, 2016.
- Huang, Furong, et al. "Learning deep resnet blocks sequentially using boosting theory." *ICML*, 2018.
- Veit, Andreas, et al. "Residual networks behave like ensembles of relatively shallow networks." *NIPS*, 2016.
- He, Kaiming, et al. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification." *ICCV*, 2015.