Deep Residual Learning for Image Recognition

Comment

- This paper focuses on *degradation*, which is often occurred using deeper network.

Abstract

- 152 layers - 8x deeper than the state-of-art, still having lower complexity
- learning residual functions with ference to the layer inputs

1. Introduction

- Deeper (naive) network, higher training/test error
- Vanishing/exploding gardients problem can be addressed by normalized init and intermediate normalization
- Using *identity* mapping indicates even depper model should produce no higher training error than its shallower counterpart.
- To handle *degradation* problem, use the nonlinear mapping layer as $\mathcal{F}(\mathbf{x}) := \mathcal{H}(\mathbf{x}) - \mathbf{x}$
- ResNet dominates the most competition such as *ILSVRC 2015* classification, *ImageNet detction*, *ImageNet localization*, *COCO detection*, and *COCO segmentation* in 2015

2. Related Work

Residual Representations

Shortcut Connections

- In the *Inception* paper, an *inception* is composed of a shortcut connections
- "Highway networks" present shortcut connections with gating functions. The gated shortcut is "closed" (approaching zero), it is no longer useful. On the contray, ResNet learns residual functions and the identity shortcuts are never closed.

3. Deep Residual Learning

3.1 Residual Learning

- The reformulation of mapping helps to reasonably **precondition** the problem

3.2 Identity Mapping by Shortcuts

The building block is defined as,

$$y = \mathcal{F}(\mathbf{x}, \{W_i\}) + W_s\mathbf{x}$$

where W_s is used to match the dimension. * $\mathcal{F}$ is flexible, but if it is one layer they could not observe meaningful improvement. * Empricially, you can set W_s as an identity matrix if the dimension is matched.

3.3 Network Architectures

	VGG-19	Plain	Residual
GFLOPS	19.6	3.6	3.6
# of Layer	19	34	34

In case that the dimensions increase,

- (A) Identity Mapping with extra zero entries padded for increasing dimensions
- (B) Projection shortcut

with a stride of 2.

4. Experiments

4.1 ImageNet Classification

Top-1 (%)	plain	ResNet
18 layers	27.94	27.88
34 layers	28.54	25.03

Plain Networks

- Higher training error conjecturally due to low convergence rates

Residual Networks

- Well addressing degradation problem
- Projection shortcuts are not particularly essential and need more time/memory complexity

Deeper Bottleneck Architectures

To reduce the significantly increasing complexity with deeper layers, the following block consisting of 3-layers rather than 2-layer showed better result.

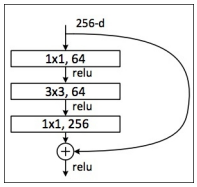

Figure 1: enter image description here

The 1x1 conv layers are responsible for reducing/increasing dimensions. The 3x3 layer is left as a bottleneck with smaller I/O dimensions. This strategy improved both performance and accuracy. (ResNet-152 – 3.57% top-5 error, using less FLOPS than VGG-16/19)

4.2 CIFAR-10 and analysis

Aggressively deep network (1202-layer)

Seems to be overfitting probably because the dataset is relatively too small.

4.3 Object Detection on PASCAL and MS COCO