Deep Residual Learning for Image Recognition

Abstract

> Deeper neural networks are more difficult to train. We present a residual learning framework to ease the training of networks that are substantially deeper than those used previously.

> Deep residual nets are foundations of our submissions to ILSVRC & COCO 2015 competitions1, where we also won the 1st places on the tasks of ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation.

1.Introduction

> Recent evidence [41, 44] reveals that network depth is of crucial importance,

> Driven by the significance of depth, a question arises: Is learning better networks as easy as stacking more layers?

> When deeper networks are able to start converging, a degradation problem has been exposed: with the network depth increasing, accuracy gets saturated (which might be unsurprising) and then degrades rapidly. Unexpectedly, such degradation is not caused by overfitting, and adding more layers to a suitably deep model leads to higher training error, as reported in [11, 42] and thoroughly verified by our experiments. Fig. 1 shows a typical example.

> In this paper, we address the degradation problem by introducing a deep residual learning framework. Instead of hoping each few stacked layers directly fit a desired underlying mapping, we explicitly let these layers fit a residual mapping.

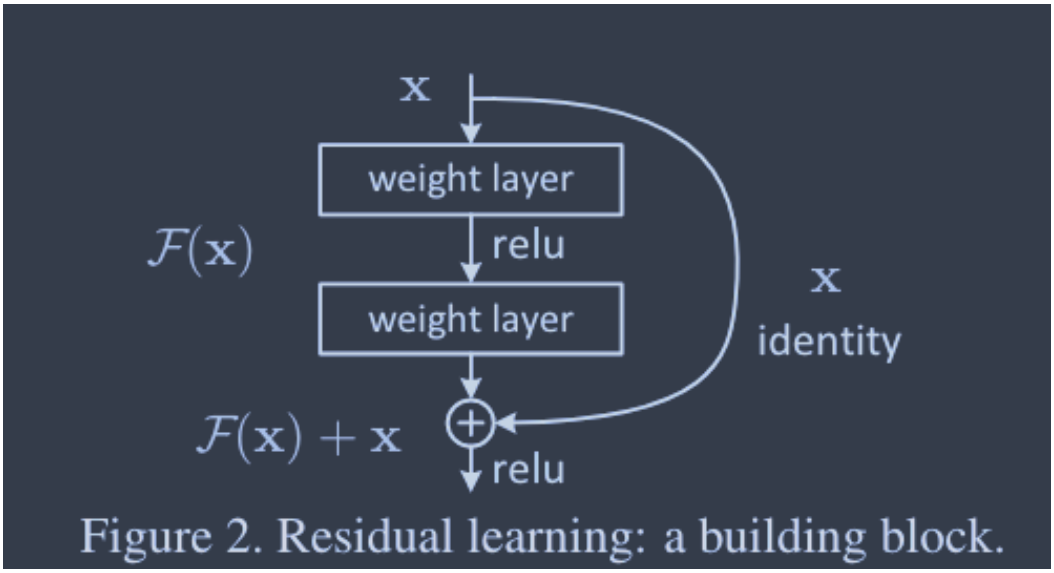

Figure 2. Residual learning: a building block.

> Formally, denoting the desired underlying mapping as H(x), we let the stacked nonlinear layers fit another mapping of F(x) := H(x) − x. The original mapping is recast into F(x)+x. We hypothesize that it is easier to optimize the residual mapping than to optimize the original, unreferenced mapping. To the extreme, if an identity mapping were optimal, it would be easier to push the residual to zero than to fit an identity mapping by a stack of nonlinear layers.

> The formulation of F(x) + x can be realized by feedforward neural networks with “shortcut connections” (Fig. 2). Shortcut connections [2, 34, 49] are those skipping one or more layers. In our case, the shortcut connections simply perform identity mapping, and their outputs are added to the outputs of the stacked layers (Fig. 2). Identity shortcut connections add neither extra parameter nor computational complexity. The entire network can still be trained end-to-end by SGD with backpropagation, and can be easily implemented using common libraries (e.g., Caffe [19]) without modifying the solvers.

3.Deep Residual Learning

3.1 Residual Learning

> Let us consider H(x) as an underlying mapping to be fit by a few stacked layers (not necessarily the entire net), with x denoting the inputs to the first of these layers. If one hypothesizes that multiple nonlinear layers can asymptotically approximate complicated functions2, then it is equivalent to hypothesize that they can asymptotically approximate the residual functions, i.e., H(x) − x (assuming that the input and output are of the same dimensions). So rather than expect stacked layers to approximate H(x), we explicitly let these layers approximate a residual function F(x) := H(x) − x. The original function thus becomes F(x)+x. Although both forms should be able to asymptotically approximate the desired functions (as hypothesized), the ease of learning might be different.

Summary

This paper points out that they found a method using residual network to solve the degradation problem that occurs in the deep layers network.

To elaborate, they construct a framework that use F(x)=H(x)-x instead of just underlying it as H(x)=x.In this way, they can easily learn the features of the parameter than learn the hole mapping of H(x), at the same time, it ensure that the performance of deeper layer won't worse than the former ones.

Besides, they also design a special framework that will make the VAE fully learn the features of input parameters and reduce the amount of cauculating.

Problem Thiking

Pooling

> （1）保留主要特征的同时减少参数和计算量，防止过拟合。
>
> （2）invariance(不变性)，这种不变性包括translation(平移)，rotation(旋转)，scale(尺度)。

Translation Invariance and Translation Equivariance

> https://zhuanlan.zhihu.com/p/382569419

> **平移不变性（Translation Invariance）**：在图像分类任务中，不变性意味着，当所需要识别的目标出现在图像的不同位置时，模型对其识别所得到的标签应该相同。即当输出进行变换后，还能得到相同的输出。
>
> $$F(x) = F[transform(x)]$$
>
> **平移相等性（Translation Equivariance）**：指在目标检测任务中，如果输入的图像中，对应的目标发生了平移，那么最终检测出的候选框也应发生相应的变化。即对输入进行变换后，输出也会发生相应的变换。
>
> $$transform[F(x)] = F[transform(x)]$$

Global Average Pooling(GAP)

> https://zhuanlan.zhihu.com/p/345183296

Class Activation Mapping(CAM)

> https://zhuanlan.zhihu.com/p/51631163

Interpolation

> https://zhuanlan.zhihu.com/p/428523385
>
> https://www.jianshu.com/p/055706fd32ee

Sampling

> https://zhuanlan.zhihu.com/p/579702765
>
> https://blog.csdn.net/zhibing_ding/article/details/125254670

Channel

> https://blog.csdn.net/sscc_learning/article/details/79814146

1x1 Convolutional Kernel

> https://zhuanlan.zhihu.com/p/40050371