**Deep Residual Learning for Image Recognition**

Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun

Deep networks are generally harder to train and show degradation problems when getting extremely deeper. Deep residual networks show generally better performance than plain networks and easily perform better when getting deeper.

1. Introduction

It is widely known that deep networks integrate low/mid/high level features and the levels of features are enriched by deeper layers. However, deeper networks show degradation problems in which deeper layers show more errors. This indicates that deeper layers are difficult to optimize. In this study, we address the degradation problem by a deep residual learning framework. We hypothesize that residual mapping is easier to optimize than a unreferenced mapping and will show better error results.

We use formulations of F(x) + x, which shortcut connections skipping one or more layers and only perform identity mapping. We present experiments on ImageNet and prove that (1) very deep residual networks are easy to optimize, but plain networks show higher error rates when getting deeper; (2) our deep residual nets show better error rates with increased depth.

2. Related Work

Some early studies implement shortcut connections which add a linear layer which directly connects the network input to the output. “Highway networks” present shortcut connections with gating functions, unlike identity shortcuts which are parameter-free. While a gated shortcut can be closed and act like non-residual functions, identity shortcuts are always open. Moreover, highway networks do not show increased accuracy in very deep networks.

3. Deep Residual Learning

텍스트, 도표, 폰트, 라인이(가) 표시된 사진

자동 생성된 설명

Figure 1. a building block for residual learning

Residual learning is adopted to every few stacked layers following this building block. This building block is defined as:

y = F(x, {Wi}) + x.

the Function F(x, {Wi})

If the dimensions of x and F do not match, we can perform a linear projection simply to match dimensions:

y = F(x, {Wi}) + Wsx.

The networks used in this experiment is mainly inspired by VGG nets, however having fewer filters and lower complexity than VGG nets.

4. Experiments

텍스트, 라인, 그래프, 도표이(가) 표시된 사진

자동 생성된 설명

Figure 2. Training on ImageNet

If we compare plain-18 and plain-34 networks, the plain-34 network has higher training error throughout the whole training procedure

But if we compare ResNet-18 and ResNet-34 networks, the ResNet-34 network has significantly lowr training errors than the 18-layer one and is generalizable to the validation data.

Moreover, ResNet-34 network shows a significant decrease in error rates- for reducing the top- 1 error by 3.5%

Also, ResNet using residual mapping eases the optimization process by providing faster convergence at an early stage.

To compare identify shortcuts and projection shortcuts, we compare three options (A) all shortcuts are zero-padded and parameter free (B) projection is only used for increasing dimensions (C) all shortcuts are projections. All options are considerably better than the plain counterpart. B is slightly better than A, and C in marginally better than B- however the differences are not significant and therefore projection shortcuts are not essential for the degradation problem.

To figure out the effects of residual learning in deeper networks, we modify the building block as a bottleneck design and use a 50-layer, 101-layers, and 152-layers ResNets. The 152-layer ResNet has lower complexity compared to VGG-16 or 19 nets. These 50, 101, 152- layer ResNets are considerably more accurate than the 34-layer ResNet and degradation problem is not observed.

The same experinents were conducted on the CIFAR-19 datasets. The results are shown by the following graph.

텍스트, 도표, 스크린샷, 그래프이(가) 표시된 사진

자동 생성된 설명

Figure 3. Training on CIFAR-10

Deep plain nets show degradation problems and show higher error rates when getting deeper. It acts similar with the ImageNet dataset. ResNets, in contrast, overcome the optimization difficulty and gains accuracy when getting deeper. Also, ResNets have generally smaller responses than plain networks, which may indicate that residual functions may be closer to zero than non-residual functions.

In an extremely deep residual model of 1202 layers, it shows no optimization difficulty but shows worse error rates than 110-laeyr networks, which is considered to be induced by overfitting.

ResNets have also shown significant performances in PASCAL and MS COCO also.