Deep Residual Learning for Image Recognition

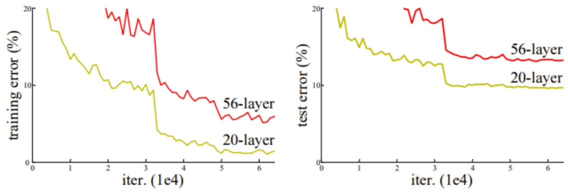
Abstract

Deeper neural networks are more difficult to train
reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions.
easier to optimize, accuracy from depth.
8x VGG but lower complexity
ensemble
The depth is importance for many visual recognition tasks

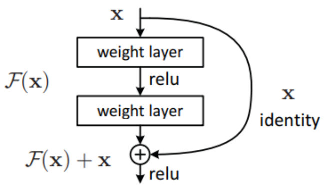
Introduction

Deep networks - integrate low/mid/highlevel features / classifiers in an end-to-end multilayer fashion, “levels” = depth
depth is of crucial importance
vanishing/exploding gradients problem
normalized initialization / intermediate normalization layers for gradient descent (SGD)

degradation problem = accuracy gets saturated / degrades rapidly.
but not overfitting higher training error, as reported in [11, 42] and thoroughly verified by our experiments.
Fig. 1 shows a typical example.



The original mapping is recast into F(x)+x.
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.



Identity shortcut connections = extra parameter X / computational complexity.X
still be trained end-to-end by SGD with backpropagation
easily implemented using common libraries (e.g., Caffe [19]) without modifying the solvers.

1) easy to optimize
2) easily enjoy accuracy gains
ImageNet, CIFAR-10

Our ensemble has 3.

excellent generalization performance on other recognition tasks
residual learning principle is generic - applicable in other vision and non-vision problems.

Related Work

Residual Representations.
VLAD and Fisher Vector
shallow representations for image retrieval and classification
For vector quantization, encoding residual vectors >> encoding original vectors.

Partial Differential Equations
Multigrid method / hierarchical basis preconditioning
these solvers converge much faster than standard solvers

Shortcut Connections
highway networks - shortcut connections w/ gating functions
never closes

Deep Residual Learning

Residual Learning
Residual Learning은 H(x)가 아닌 출력과 입력의 차인 H(x) - x를 얻도록 목표를 수정
F(x) = H(x) - x를 최소화시켜야 하고 이는 즉, 출력과 입력의 차을 줄인다는 의미
H(x)를 x로 mapping 하는 것이 학습의 목표
pre-conditioning으로 인해 Optimal function이 zero mapping보다 identity mapping에 더 가깝다면, solver가 identity mapping을 참조하여 작은 변화를 학습하는 것이 새로운 function을 학습하는 것보다 더 쉬울 것

Identity Mapping by Shortcuts
neither extra parameter nor computation complexity
identity mapping is sufficient for addressing the degradation problem and is economical, and thus Ws is only used when matching dimensions

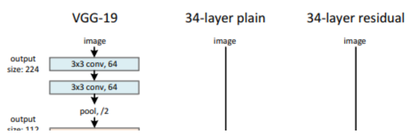
Network Architectures

consistent phenomena

Plain Network
VGG
3×3 filters
stride of 2
global average pooling layer and a 1000-way fully-connected layer with softmax
The total number of weighted layers is 34

two design rules
for the same output feature map size, the layers have the same number of filters;
feature map size is halved, the number of filters is doubled so as to preserve the time complexity per layer.

Residual Network
added shortcut connections



When the dimensions increase 2 options
extra zero entries padded
Ws

Implementation

ImageNet
randomly sampled in [256, 480]
horizontal flip, with the per-pixel mean subtracted [21
standard color augmentation
batch normalization
SGD with a mini-batch size of 256
The learning rate starts from 0.1 and is divided by 10 when the error plateaus,
to 60 × 104 iterations
weight decay of 0.0001 and a momentum of 0.9
do not use dropout

Experiments **ImageNet Classification**