Deep Residual Learning for Image Recognition

He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR) 2016.

ISL

안재원

CONTENTS

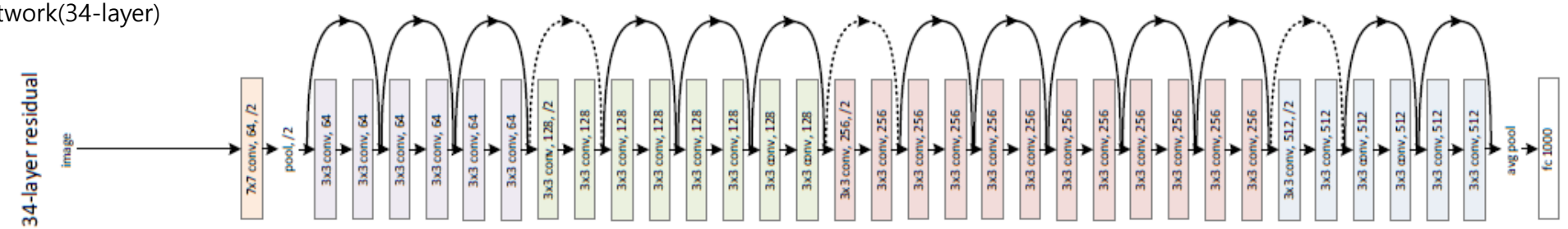
- Introduction
- Deep residual learning
- Experiments

Introduction

Intro

- 신경망이 깊으면 깊을 수록 더 좋지 않을까?
- 어떻게 하면 더 깊은 신경망을 학습할 수 있는가?
- 어떻게 하면 더 깊은 신경망을 빠르게 학습할 수 있는가?

- Residual Network(34-layer)



VS

- VGG-19 Network

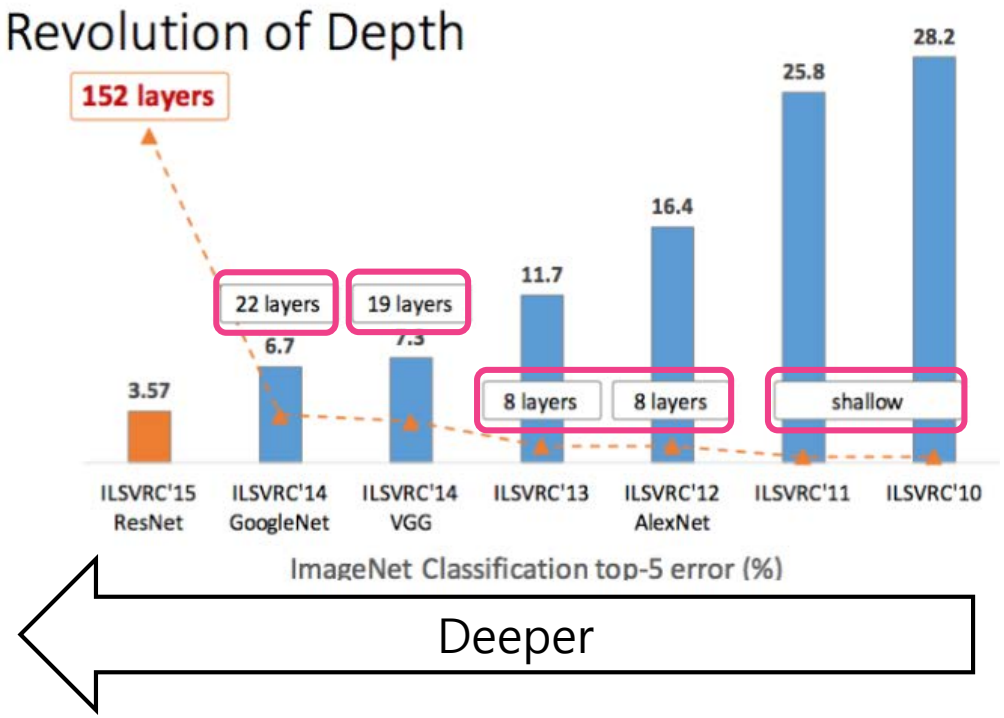
Introduction

Deeper and deeper

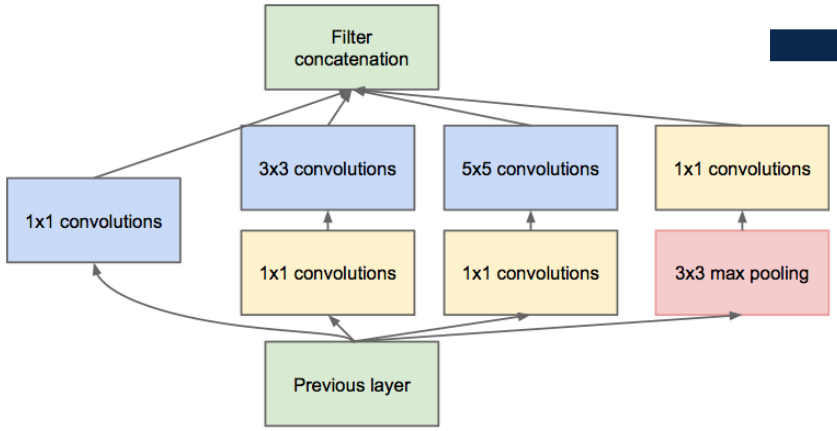
IMAGENET

- 세계 최대의 영상 데이터 베이스
- 약 22000종류, 1500만장의 영상 데이터 보유
- 2010년부터 ILSVRC 개최함

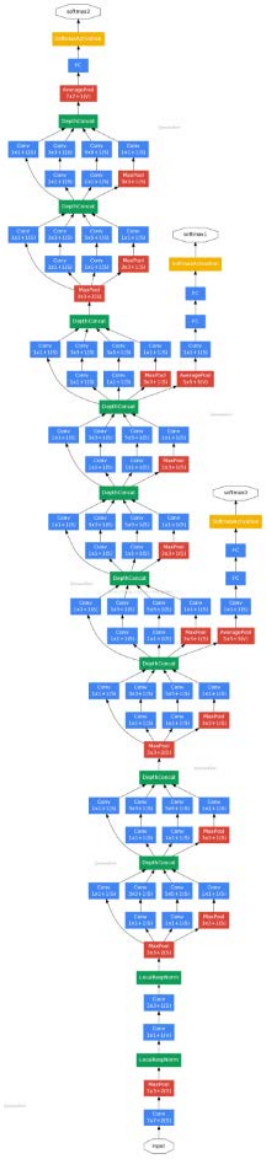
- ImageNet Large Scale Visual Recognition Competition(ILSVRC)
- ImageNet의 영상 데이터를 이용한 영상 인식 대회
- Learning기법을 도입하면서 에러율이 급격하게 감소했다.



- GoogleNet



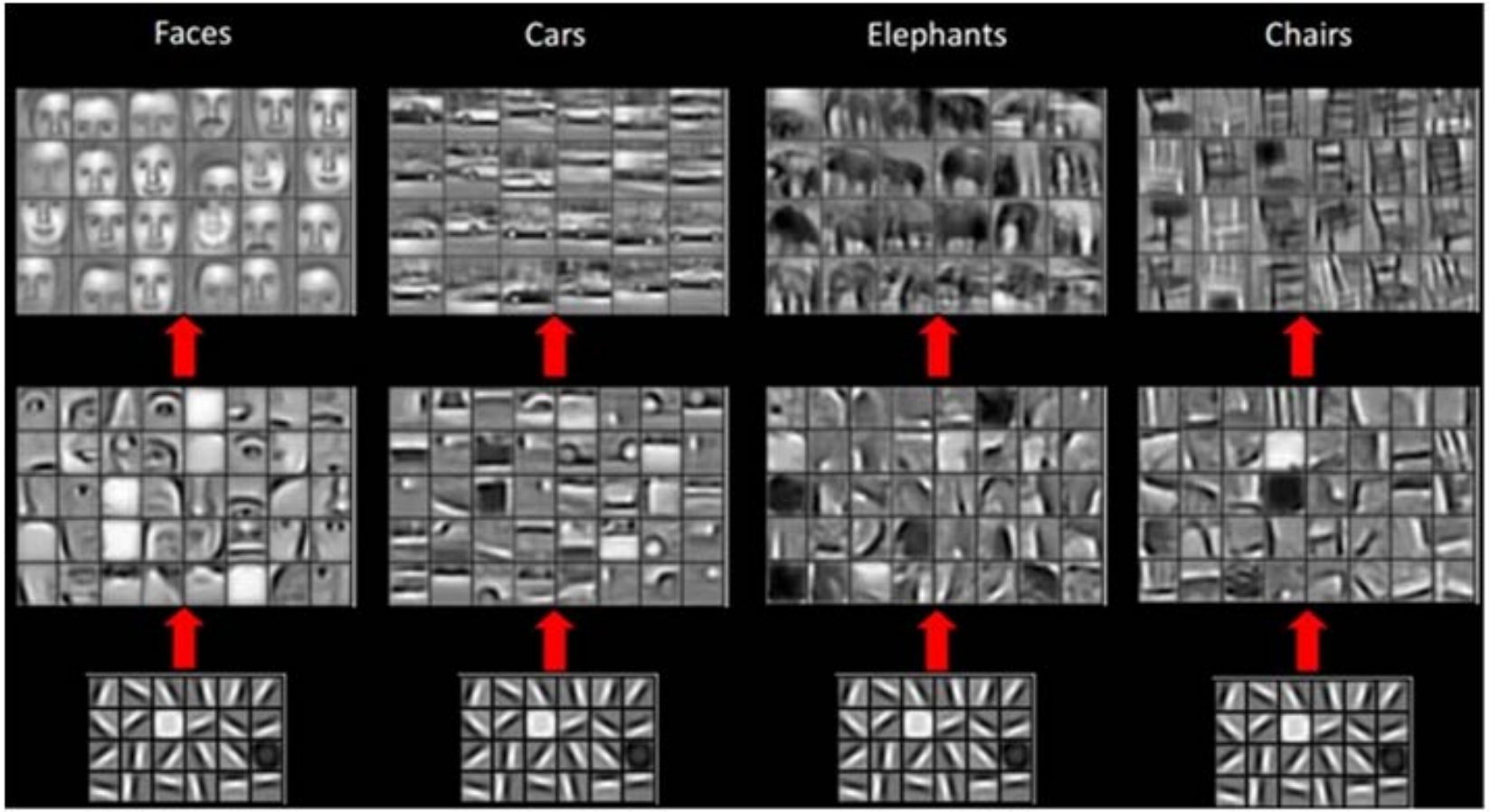
Inception module의 구조



Introduction

Deeper and deeper

- 왜 깊을 수록 좋은가?
- 깊으면 무조건 좋은가?



- High level features
  더 복잡한 특징

- Mid level features
  복잡한 특징

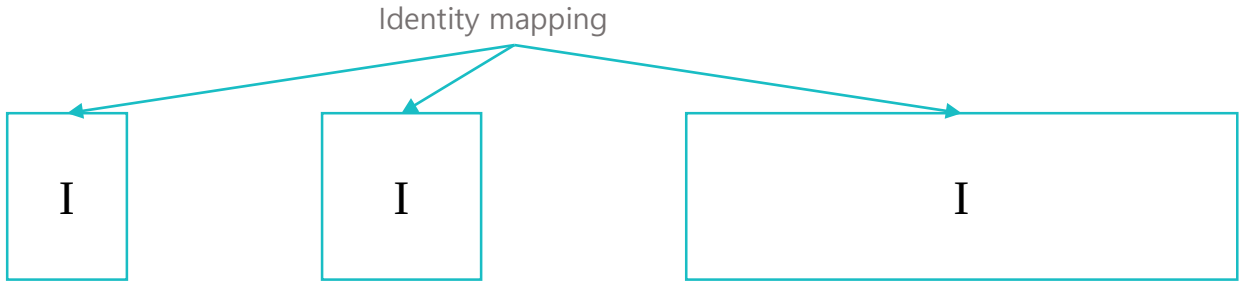
- Low level features
  단순한 특징

Introduction

Deeper network always better?

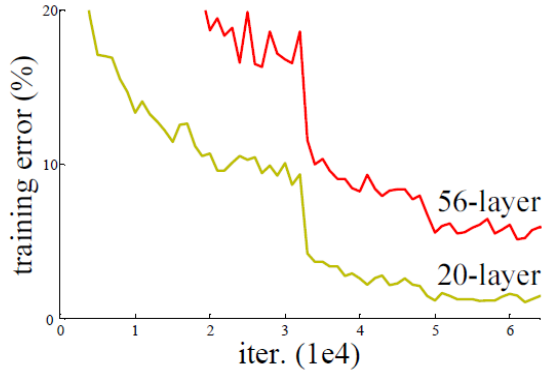
- 왜 깊을 수록 좋은가?
- 깊으면 무조건 좋은가?

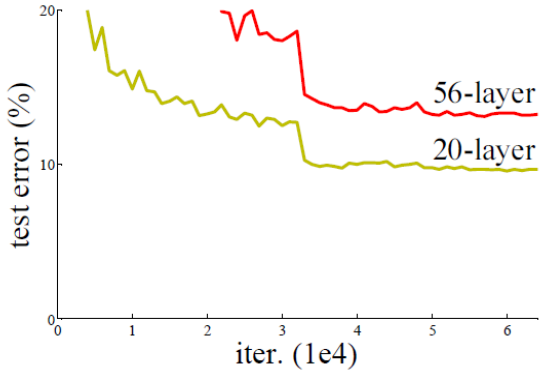
- 잘 학습된 Shallower Network



- Identity mapping에 의해 잘 학습된 특징이 그대로 전달 된다.
- 그렇기 때문에 적어도 Training error가 유지되거나 줄어들 것이다.

But





- 성능 저항의 원인이 Overfitting만의 문제는 아니다.

  - K. He and J. Sun. Convolutional neural networks at constrained time cost. In CVPR, 2015.
  - R. K. Srivastava, K. Greff, and J. Schmidhuber. Highway networks. arXiv:1505.00387, 2015.

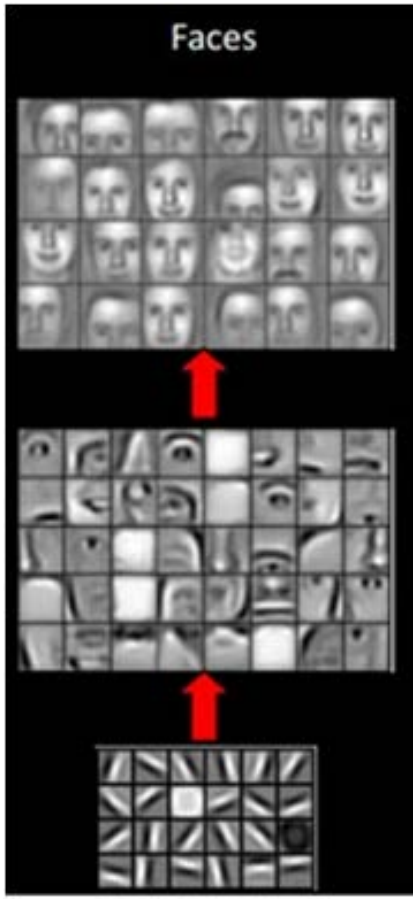
- 현재의 네트워크 구조와 학습 형태가 필요한 정보를 온전히 전달하도록 구성되어 있지 않기 때문에 발생하는 문제다.

Introduction

Paradigm shift

- 현재의 네트워크 구조와 학습 형태가 필요한 정보를 온전히 전달하도록 구성되어 있지 않기 때문에 발생하는 문제다.

- 얼굴을 학습할 수 있는 깊이의 네트워크에서 눈, 코, 입을 학습하고 싶다면?



- High level features
  더 복잡한 특징

- Mid level features
  복잡한 특징

- Low level features
  단순한 특징

- 기존 네트워크의 학습 방향

층을 거듭할 수록, Low level의 특징(feature)를 이용해 얼마나 더 복잡한 특징을 학습 시킬 것인가.

- High level features
  사람1, 사람2, 사람3

- Mid level features
  눈, 코, 입

- Low level features
  점, 선, 면

- Residual network의 학습 방향

층을 거듭할 수록, Low level의 특징과 얼마나 더 (복잡한 방향으로)다른 형태의 특징을 학습시킬 것인가.

- Mid level features
  눈, 코, 입

- Low level features
  점, 선, 면

Deep Residual Learning

Residual?

- A Residual is generally a quantity left over at the end of process.
- Error.

- Residual function.

$$\mathcal{F}(x) := \mathrm{H}(x) - x$$

Hypothesis(End of process)

Input

- 각 층(Layer)의 Hypothesis.

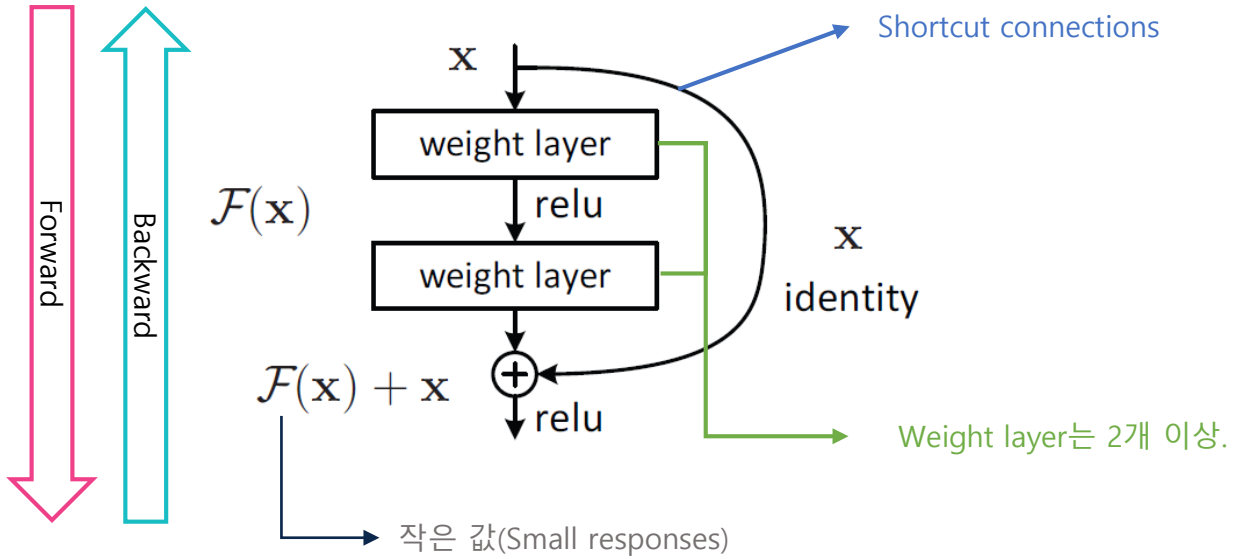
$$\mathrm{H}(x) = \mathcal{F}(x) + x$$

Learing을 통해 학습하고자 하는 것.

아래층(Low layer)의 특징(Feature)과 얼마나 달라지는가를 학습한다.

Residual

- Residual learning framework.

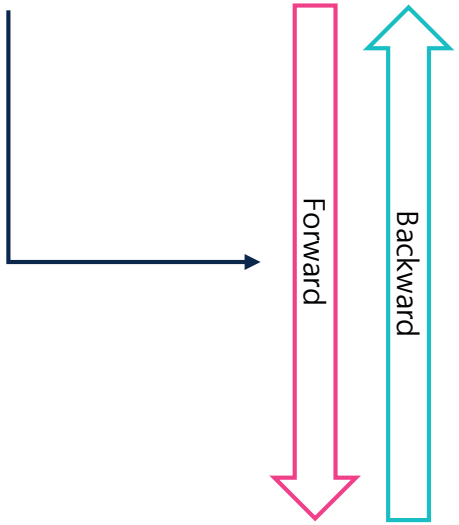


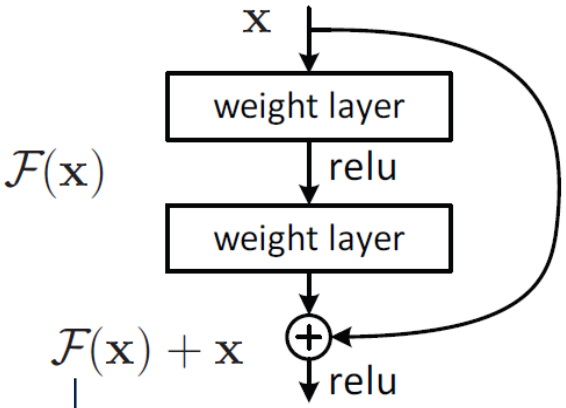
Deep Residual Learning

Residual learning framework

- Residual learning

- 일반적인 네트워크처럼 동작한다.





- 학습에 유리한 구조를 갖는다.
- Precondition

$$\mathrm{H}(x) = \mathcal{F}(x) + x$$

x
identity

학습되고자 하는 형태가 사전에 정해져 있다.

작은 변화(Perturbations)

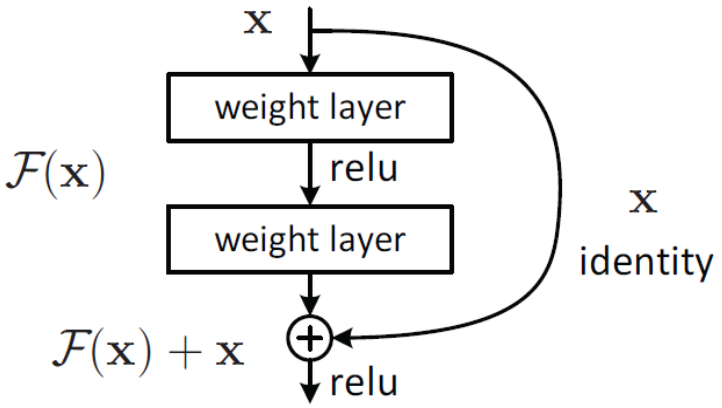
작은 값(Small responses)

Deep Residual Learning

Residual learning framework

- Shortcut connections

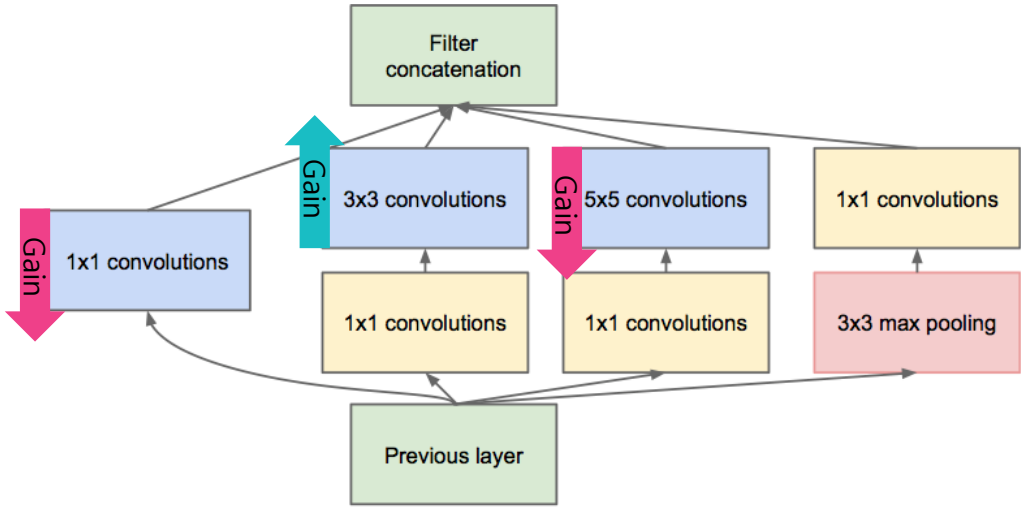
  Residual learning framework에서 처음 등장한 개념은 아니다.



$$H(x) = \mathcal{F}(x) + x$$

- Short connection이 언제나 활성화되어 있다.
- 즉, Residual한 특성이 언제나 유지된다.

- Inception module(v3) in GoogleNet



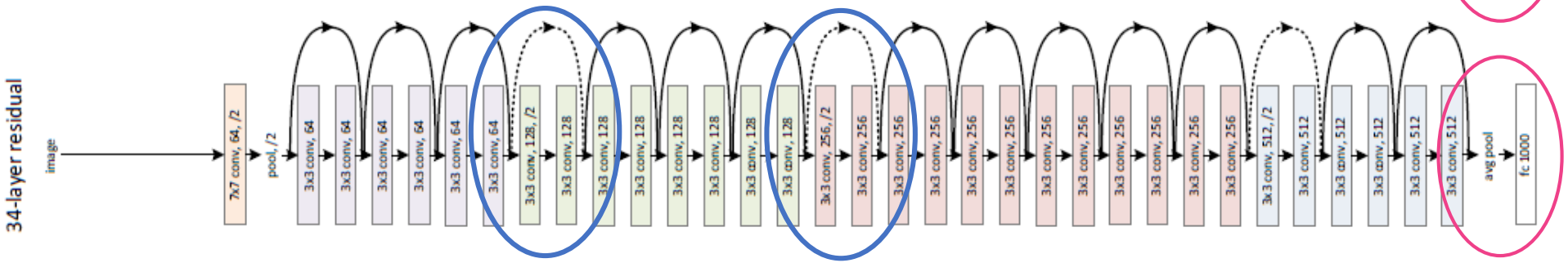
- 획득하고자 하는 특징에 적합한 Convolutions의 크기를 알 수 없다.
- 그렇기 때문에 여러 크기의 Convolutions으로 학습을 진행한다.
- 특징 검출에 따라 선택적으로 Short connection이 활성화 된다.

Deep Residual Learning

Residual Network

- Residual Network(34-layer)
  - VGG-19 network를 기반으로 만들었다.
    - 각 층(Layer)에서 출력되는 특징(feature map)의 수가 같으면, 같은 수의 filter를 사용한다.
    - 출력되는 특징의 수가 반으로 줄면, 사용하는 filter의 수는 2배가 된다.
    - 위의 특징에 의해 각 층에서 출력되는 특징의 수에 상관 없이 각 층의 복잡도는 보존된다.
  - VGG-19 network보다 덜 복잡하다.
    - VGG-19 network : 19.6 billion FLOPs
    - Residual network : 3.6 billion FLOPs
  - VGG-19 network보다 filter의 수도 더 적다.

- VGG-19 network



  - 입출력의 크기가 다른 구간은 두 옵션 중 하나를 선택한다.
    - A : Zero padding(입력에 대한 Identity를 보장할 수 없다.)
    - B : Projection shortcut(1x1 convolutions : dimension 조절(증가, 감소)을 위한 convolutions.)

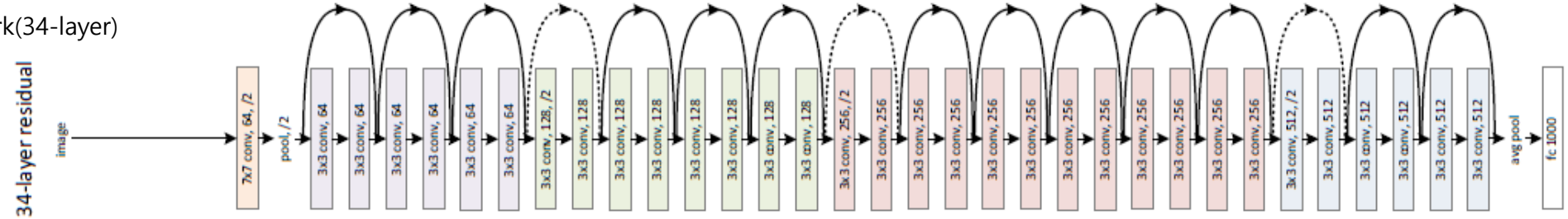
$$H(x) = \mathcal{F}(x) + W_s x$$

Experiments

Training on ImageNet

- ImageNet을 이용한 학습 결과 비교.
- Batch normalization
- SGD
- Don't use dropout
- 600,000 iterations

- Residual Network(34-layer)

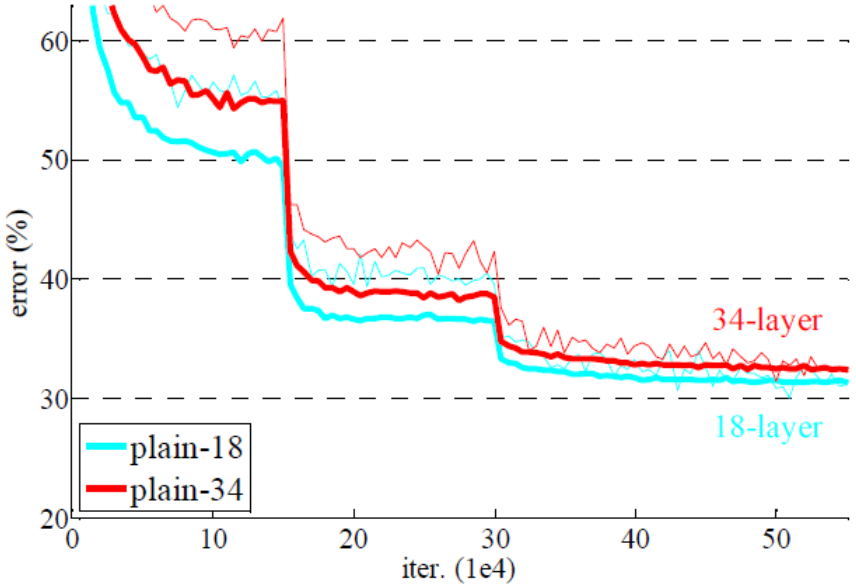


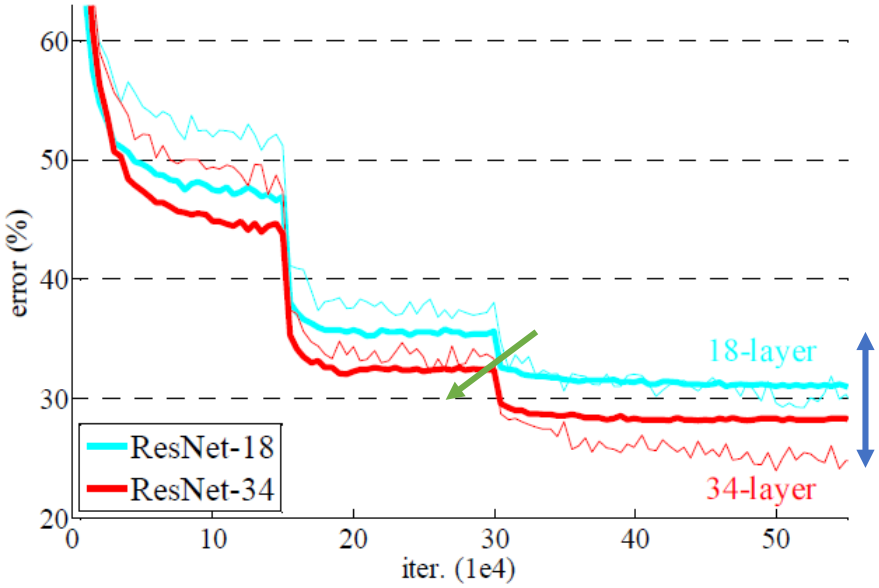
- Plain Network(34-layer)

Experiments

Training on ImageNet

- Error 변화 비교





- 굵은 선 : Training error
- 얇은 선 : Validation error

- 더 깊은 네트워크가 더 좋은 성능을 보인다. →
- 학습 속도가 더 빠르다. →
- 깊을 수록 성능차이가 더 두드러지는 것으로 보인다.

- Top-1 error

	plain	ResNet
18 layers	27.94	27.88
34 layers	28.54	**25.03**

Experiments

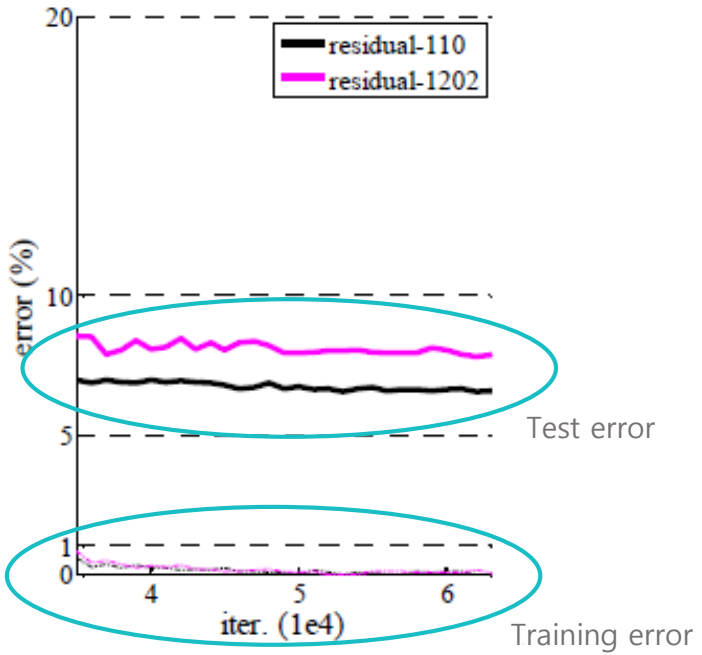
Exploring over 1000 Layers

- Training data set : CIFAR-10

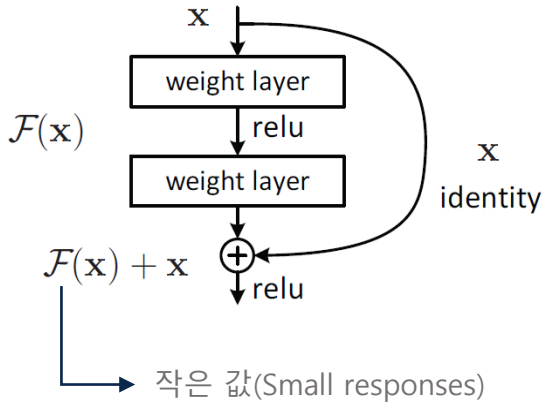
- Error 비교.

method			error (%)
Maxout [10]			9.38
NIN [25]			8.81
DSN [24]			8.22
	# layers	# params	
FitNet [35]	19	2.5M	8.39
Highway [42, 43]	19	2.3M	7.54 (7.72±0.16)
Highway [42, 43]	32	1.25M	8.80
ResNet	20	0.27M	8.75
ResNet	32	0.46M	7.51
ResNet	44	0.66M	7.17
ResNet	56	0.85M	6.97
ResNet	110	1.7M	**6.43** (6.61±0.16)
ResNet	1202	19.4M	7.93

- Error 비교.



- Over-fitting에 의한 결과라고 판단 되지만..



작은 값(Small responses)

Identity mapping이 보장되는가?

Q & A