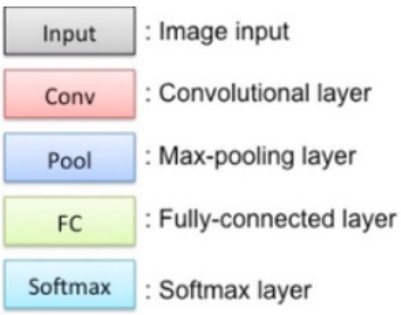
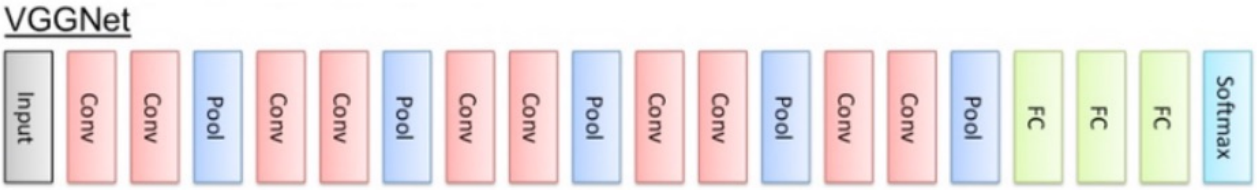
VGG Net





"GooLeNet 에게 물힌 안타까운"

옥스퍼드 대학의 Visual Geometry Group 팀이 개발한 모델로,

2014 년 이미지넷 대회에서 아쉽게 2등!

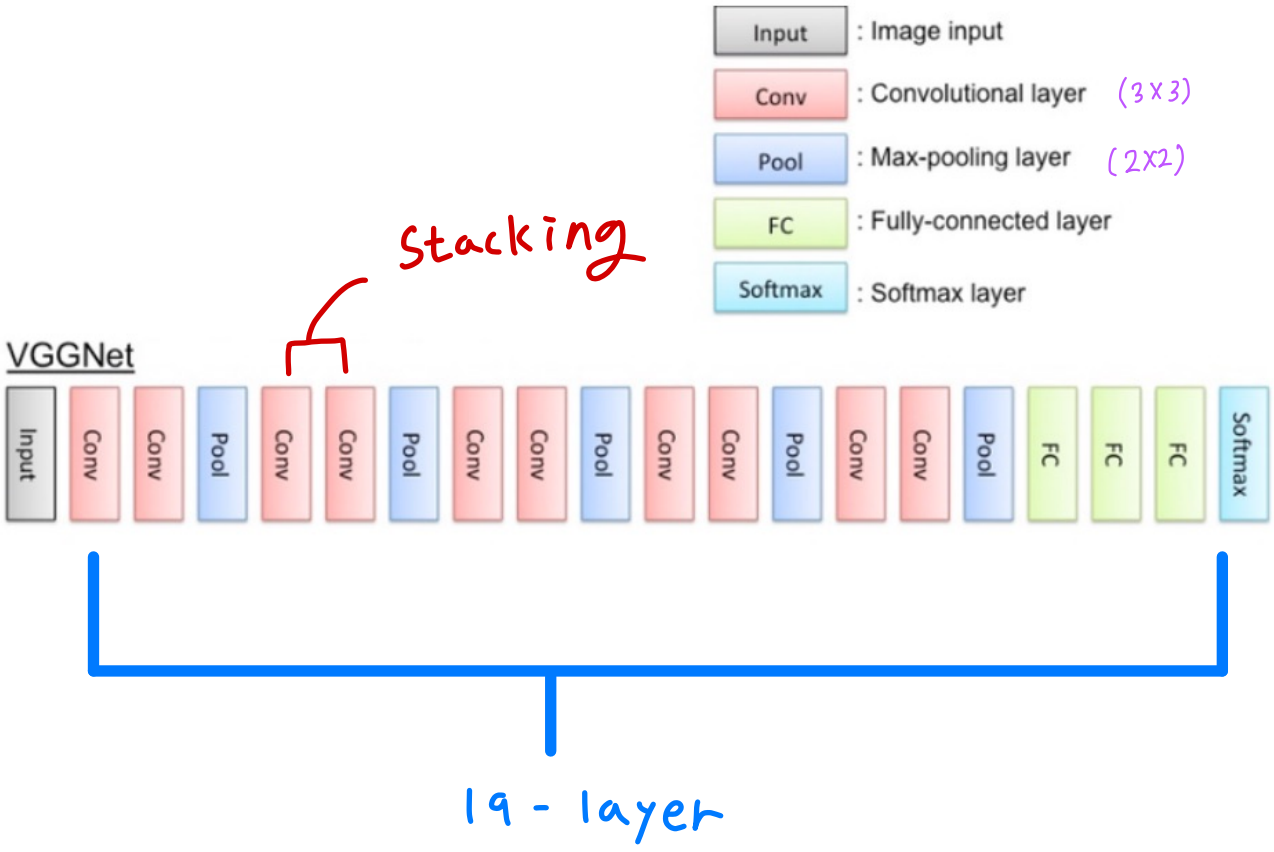
대회에 사용된 모델은 depth - 19 모델로 Inception v1 (22) 과 비슷하게

2013년에 비해 망 깊이가 크게 증가!

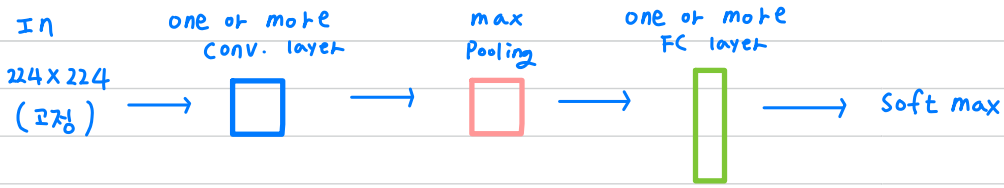
Will Cover

1. 구조적 특성
   - 3x3 kernel stacking
2. 방지 / 개선에 관한 기법들
   - partial pre-trained initializing
   - Scale - jittering

1.1 3x3 Kernel stacking

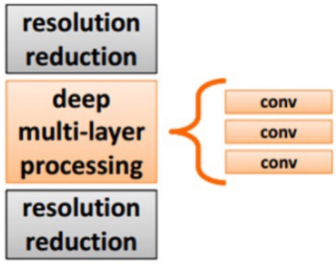


그림에서 확인할 수 있듯이 VGG Net 은 기존의 LeNet , AlexNet 과 구조적으로 큰 차이를 보이지 않음 (GoogLeNet 에 비하면 판박이 수준)



다만, VGG Net 의 출발점은 망의 깊이 가 주는 영향을 알아보기 위해 커널의 사이즈를 모두 3x3 으로 고정 (오직 깊이의 영향만을 잘 알아보기 위해)

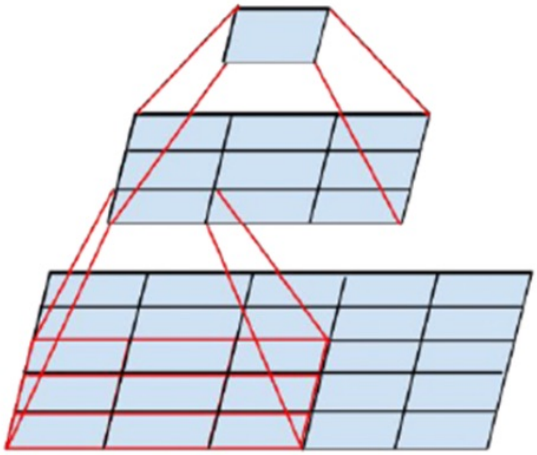
이러한 고정 사이즈 커널로 부터 더 좋은 특성 추출과 적은 연산 비용 효과를 위해 Kernel - stacking 을 고안



커널을 쌓는 다는 점의 효과는 바로 Convolution Factorizing !

↳ = 큰 필터 하나를 여러개의 작은 필터로 인수분해 하여

작은 필터들로부터 큰 필터를 통한 특성 추출의 효과를 만들어 낸다! + α 비용도 적다 !!!



[5x5 커널을 3x3 커널 2개로 인수분해]

- 5x5 커널을 한번 덧대어 하나의 숫자로 계산
  = 25개의 파라미터 필요.

- 3x3 커널로 9번 덧대어 3x3 feature map 을 얻고
  다른 3x3 커널을 1번 덧대어 하나의 숫자로 계산
  = 18개의 파라미터 필요 (9+9)

(-28%)

즉 3x3 커널 2개로 5x5 conv 의 효과를, 3x3 커널 3개로 7x7 Conv 의 효과를 보다 적은 비용으로 얻을 수 있다! ⇒ 파라미터의 개수를 줄일 수 있다

+ PS. 층의 개수가 늘어나면 커널과 커널 사이에 activation function 도 존재해서 실제로 5x5 kernel 1개 보다 3x3 kernel 2개를 통해 추출한 특성이 훨씬 좋다고 한다. (non-linearity 효과적 전달)

+ PS. 다만, 큰 필터를 사용하는 것에 비해 훨씬 파라미터가 적은 거지, VGG Net 자체는 Inception V1 에 비하면 가중치 파티.

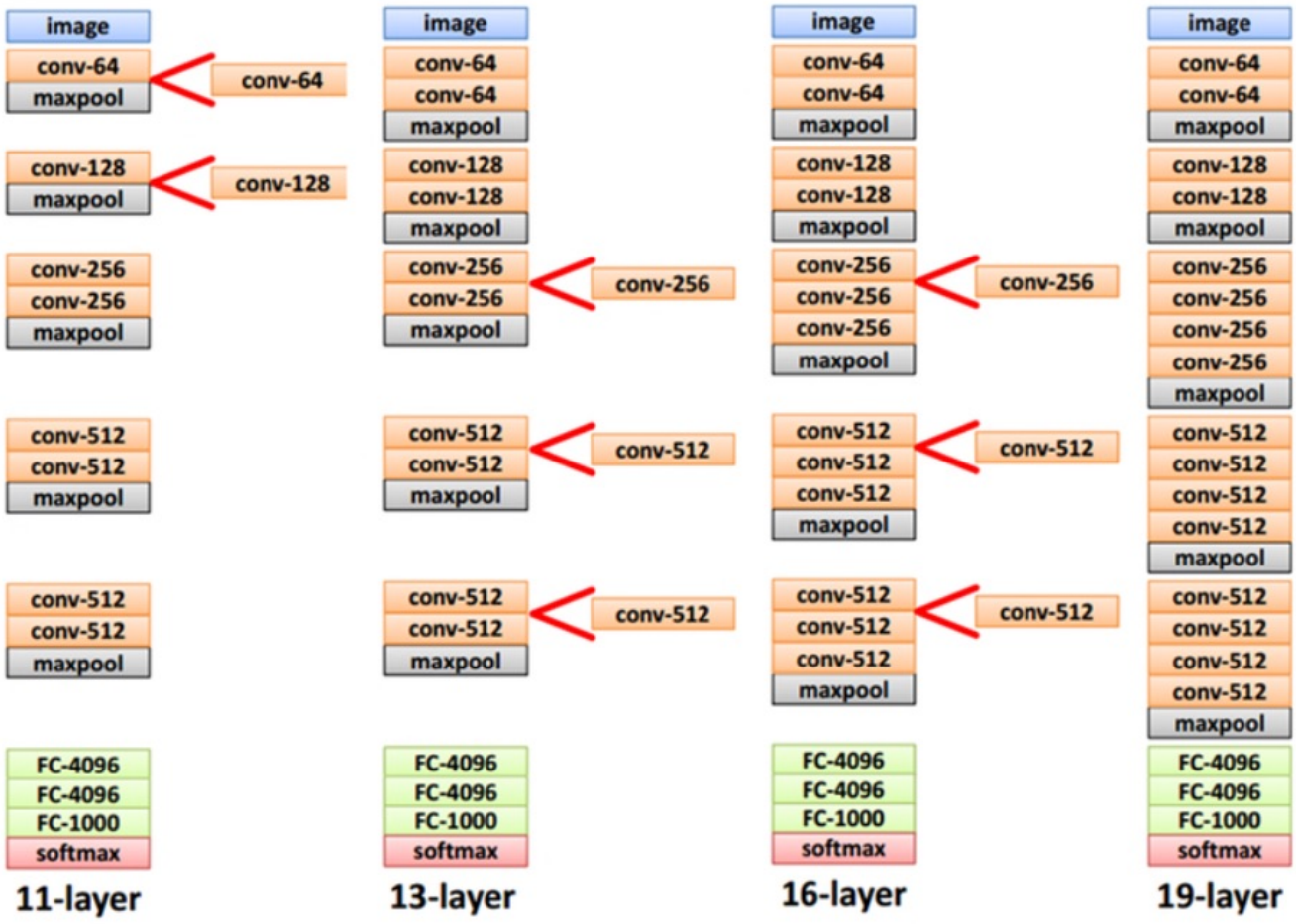
Inception V1 (5 M) <<< VGG A-model (가장간단한 11 Layer) (133 M)

특히 FC layer 에서 파라미터가 너무 많음

⇒ GoogleNet의 Avg pooling layer 가 해결!!

이러한 kernel-stacking 기법을 통해 중간중간 쌓는 커널 수를 변경해가며 총 6개 모델에 대한 학습과 비교를 진행하였다고 함.

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					



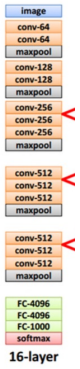
2.1 Partial pre-trained initializing

Google Net 에서는 Auxiliary Classifier 를 통해

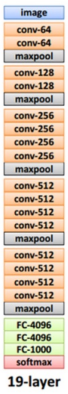
Gradient V/E 문제를 해결하였는데 VGG Net 은 어떻게 ?

⇓

간단한 구조
먼저 학습



학습결과를
복잡한 모델의
일부 레이어
가중치 초기화에 활용!!



VGG Net 에서는 기울기 소실/폭주 문제 방지를 위해서

가장 간단한 A 모델로 (위의 표 A~E 모델중) 먼저 학습 후,

학습된 가중치로 C, D, E 모델같은 복잡한 모델들의

첫 4개의 레이어 와 FC 레이어 초기화 함으로서 이를 방지 했다.

※ Transfer Learning, Fine-tuning 과 혼동하면 X

애초에 전이학습의 경우 FC 레이어를 제거하고 (만약에 FC를 사용하는 모델이라면)
목적에 맞는 커스텀 FC를 쌓아 학습한다.

GAP 를 사용하는 모델이라도 Conv. 레이어 일부를 freeze 하는 경우는 있어도
"일부만" 가중치 초기화 하진 않음.

2.2 Scale-jittering

VGG Net 은 3x3 kernel 로만 이루어진 단순한 구조로부터 최대한 성능을 끌어올리기 위해 train/test 과정에서도 여러 기법을 활용함.

Alex Net 은 모든 입력 이미지를 256x256 으로 고정하고 그중 랜덤하게 224x224 영역을 골라 학습하는 방식으로 2048 배의 image-augmentation 을 진행하였는데,

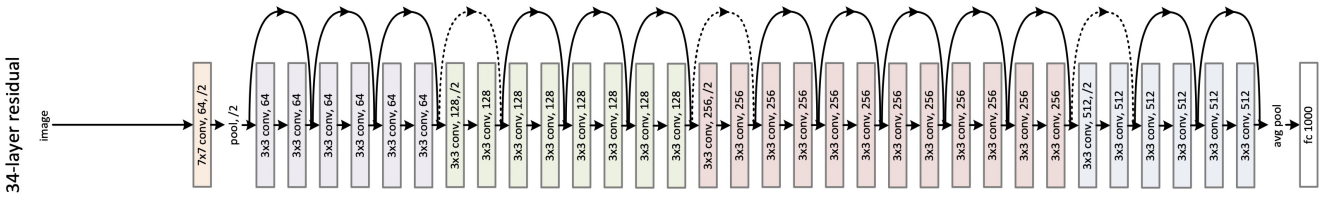
VGG Net 의 경우 한 단계 더 나아가 입력 이미지를 384x384 로 고정시킨 영역에서 무작위로 추출한 224x224 영역으로 먼저 학습을 진행하고 이후 입력 이미지를 256x256 ~ 512x512 사이의 랜덤 스케일링 하여 그중 추출되는 224x224 영역의 이미지로 fine-tuning 을 진행하였다.

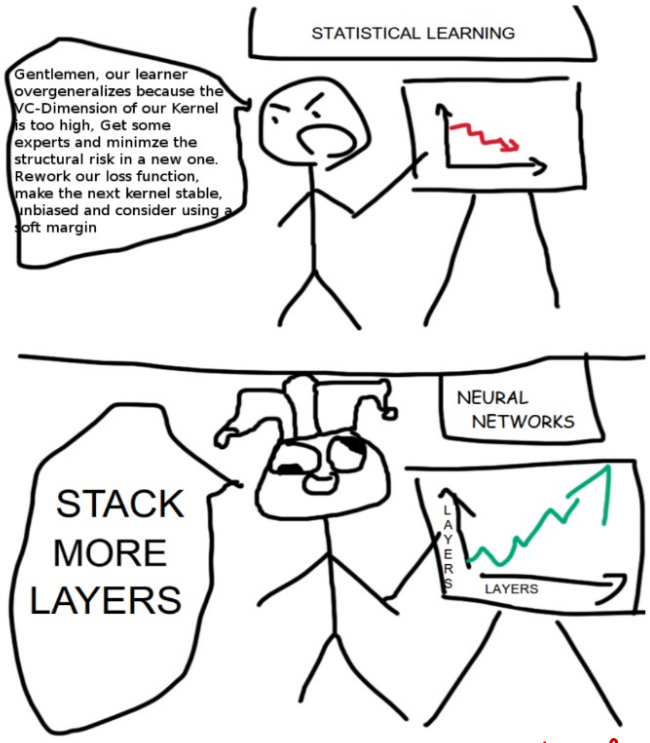
Alex Net 과 같이 영역추출 전 입력 이미지를 한 사이즈로 통일시키는 것을 Single-Scaling, 여러 사이즈로 조정하는 것을 Multi-scaling 이라고 하는데, VGG Net 과 같이 정해진 사이즈 종류가 아닌 무작위 방식을 통해 multi-scaling 을 진행하는 것을 "Scale-jittering" 이라고 한다.

+ test 과정에서도 새로운 기법인

Dense-evaluation 개념에 대해 언급되어 있지만 시간 상 생략하고 다음 시간에 가능하면 Over-feat 주제로 공부할 것.

Res Net





"152 Layers"

"Ultra-deep net"

2015년 이미지넷 모든 부문에서 우승을 거머쥔 MS의 괴물 모델!

classification, detection, localization, segmentation

Resnet 설계자인 Kaiming He 는 classification 뿐만 아니라 detection 분야에도 크게 영향을 미침 (R-CNN → Fast R-CNN → ···)

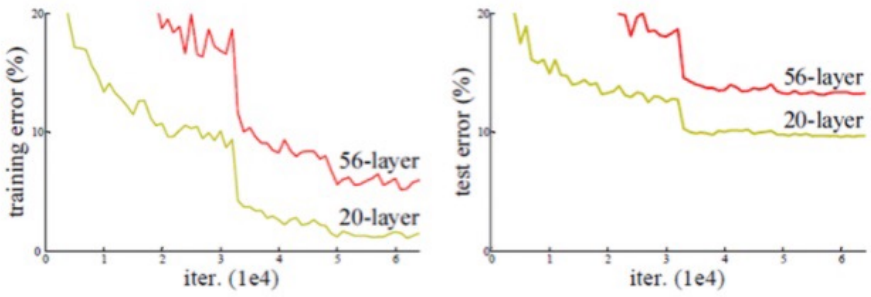
구글팀은 VGG 모델을 참고하여 Inception V_2, V_3 를 만들었으며
Res Net 모델을 참고하여 Inception V_4 를 만들었다.

+ 2016 년에는 Alex Net, VGG Net, Inception, Res Net 이 모두 합쳐진 "Res Next" 라는 끔찍한 혼종이 완성된다.

Will Cover

1. 구조적 특성
   - deep residual-learning
   - comparison w VGG
2. 기법 및 부가적 내용
   - bottle neck application
   - other field performance, other dataset

1.1 Deep residual learning



Res Net 연구팀이 초기에 VGG 와 같은 일반적 Conv. Net 을 각각 20-layer, 56-Layer 로 구성하여 CIFAR-10 데이터를 통해 학습성능을 비교해 보았는데 이상하게 train 할때, test 할때 모두 56-Layer 가 더 낮게 나왔다.

Over fitting 문제라면 train 할때는 56-Layer 가 더 좋게 나와야 하는거 아닌가 ??

↳ 망의 깊이가 더욱 깊어져 파라미터가 곱으로 증가함에 따라 그 많은 파라미터가 다 학습할 환경이 제한 (ex. 데이터 부족)

↳ 나는 개인적으로 그냥 under fitting 말하는거 아닌가 이해하고 있는데 다른 모든 정리글에서 "degradation" 이란 표현을 씀. 둘이 엄연히 다른건지는 모르겠음 (질문남김)
확실한건 degradation != over-fitting

이때 Res Net 연구팀은 이런 생각을 했다. "아니 그럼 처음 20-Layer 는 원래대로 설계하고 아무일도 안하고 그냥 전달만 하는 36-Conv. Layer 를 붙여서 56-Layer 를 만들면 아무리 학습이 안되도 20-Layer 랑 같은 성능이 지 않겠어 ?"

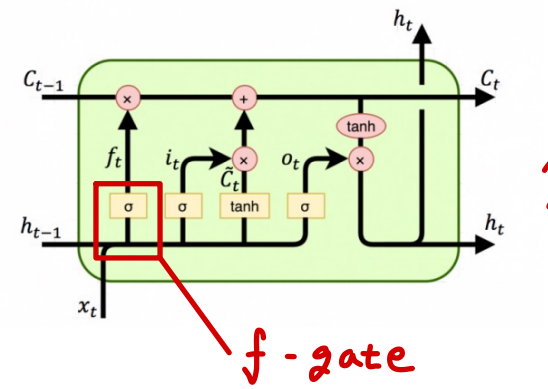
이것이 바로!! residual-learning 의 시작이었다.

"이전층의 정보를 전달한다." 자체는 사실 간단하다. 필터를 $\begin{bmatrix}0&0&0\\0&1&0\\0&0&0\end{bmatrix}$ 이런 꼴로 설계하고 각 파라미터가 학습이 되지 않도록 Non-trainable parameter 취급하면 된다.

그러나 Res Net 팀이 원한 것은 "현재 층에서 이전층의 정보를 전달만 할지, 추가 정보를 붙여 전달할지 결정하는 방향의 학습" 이었다.

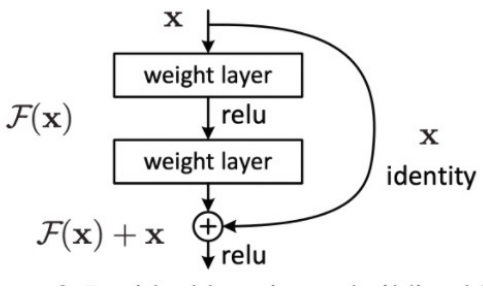
이에 연구팀은 LSTM의 Forget Gate 에서 아이디어를 얻어 다음과 같은 residual-block 을 설계하게 된다.

LSTM block



f-gate

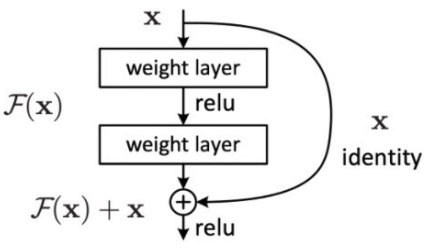
Residual block



다른 점은, f-gate 처럼 '얼마나' 전달할지에 대한 파라미터를 두지 않고 이전층의 정보를 100% 전달하는 식으로 설계하였다. 이러한 설계 방식을 Shortcut-Connection 혹은 Skip-Connection 이라 부르며 Res Net 의 핵심이다.

이 residual-block 을 통한 학습의 성능은 뛰어났는데, 이유를 들여다보면 다음과 같다.

1. 레이어에게 옵션이 생긴다.

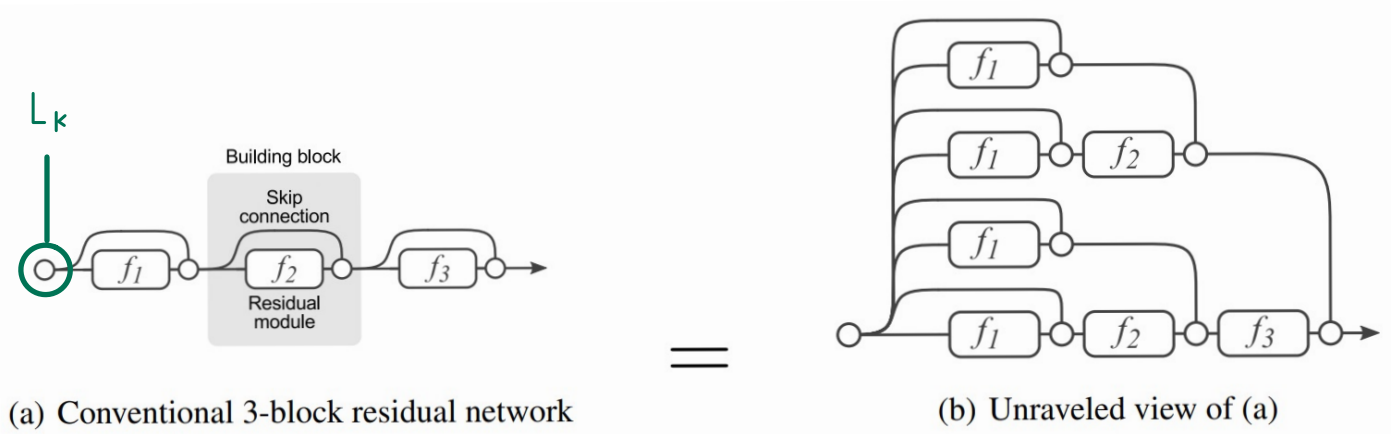


- Skip-connection 이 없는 구조에서 Weight layer 들은 별다른 옵션이 없이 최적의 어떤 숫자를 출력해야한다. (몇인지 아무도 모름)
- Skip-connection 이 있는 구조에서 ★★ Weight layer 들은 옵션이 생긴다.
  - 정보를 추가하는게 이득이다 → 기존정보에 추가할 정보만을 출력하도록 학습한다. (F(x) != 0)
  - 그냥 전달만 하는게 이득이다 → 출력이 0이 되도록 학습한다. (F(x) = 0)

★ 출력이 0이 되게 끔 학습하는 것은 쉽다. (목표 고정)

차이!

2. 역전파 시 앙상블 효과를 얻을 수 있다.

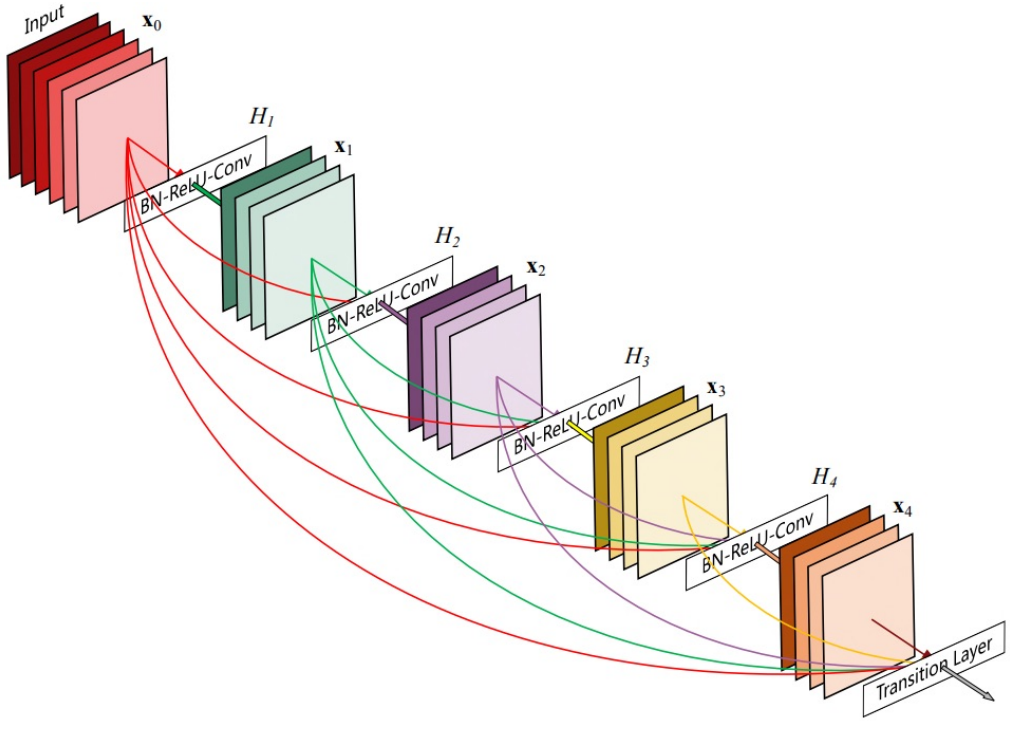


(a) Conventional 3-block residual network (b) Unraveled view of (a)

L_k 라고 표시한 위치에서는 마치 서로 다른 8개의 모델로 부터 각각 역전파되어 온 그래디언트를 종합하여 현재 위치의 그래디언트를 계산하는 듯한 모습을 띄게 된다.

↳ Gradient V/E solve, stable gradient

+ FYI Res Net 은 인접한 Conv. layer ⟷ Conv. layer 사이 정보를 전달하는 정도였으나 네트워크 전체에 통로를 만들어 정보를 전달하는 시도에 의해 2016 년 Dense Net 이 등장함



1.2 Comparison with VGG

	VGG	ResNet
kernel	3x3	3x3
classifier	FC	GAP
reducing map size	max pool	conv. with 2-stride

2.1 Bottle Neck application → 학습에 걸리는 시간을 고려하여 나온 구조

Inception V1 리뷰에서도 다뤘듯 ResNet 에서도 실험을 진행한 여러 모델 중, 50-Layer 이상의 복잡한 모델들은 1x1 Conv. 층을 활용하여 차원을 감소시켜 계산을 함으로서 연산 시간을 절약했다