VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION

https://arxiv.org/pdf/1409.1556.pdf

ABSTRACT

CNN의 **깊이**가 대용량 이미지 인식의 정확도에 영향을 미치는 것을 확인

main contribution

- **3 x 3** 컨볼루션 필터의 구조를 사용하면서, **깊이에 변화**(증가)를 주며 평가

  → **16 ~ 19개**의 가중 레이어로 이전의 선행 기술 보다 상당한 향상을 보임
- ImageNet Challenge 2014 → 지역화(localisation)에서 1등, 분류(classification)에서 2등을 달성
- 다른 데이터셋에서 SOTA 달성

1. INTRODUCTION

CNN이 대용량 이미지와 비디오 인식에서 성과를 거두고 있는 이유

- **대용량의 공공 이미지** 저장소(ImageNet)
- 고성능의 컴퓨팅 시스템(**GPU**, 대규모의 distributed clusters)
- ImageNet Large-Scale Visual Recognition Challenge (ILSVRC)의 deep visual recognition architecture

  ILSVRC → high-dimensional shallow feature encoding부터 deep ConvNet까지의 대용량 이미지 분류 시스템의 **테스트 베드**가 되어주고 있음

2012년 Krizhevsky의 구조를 개선해 더 높은 정확도 달성

→ ILSVRC-2013에서 첫번째 합성곱 계층에 **작은** receptive **window size**와 **작은 stride**를 사용

+) 전체 이미지와 여러 스케일의 네트워크를 학습하고 테스트

이 논문에서는 ConvNet 아키텍쳐 중 깊이를 설명

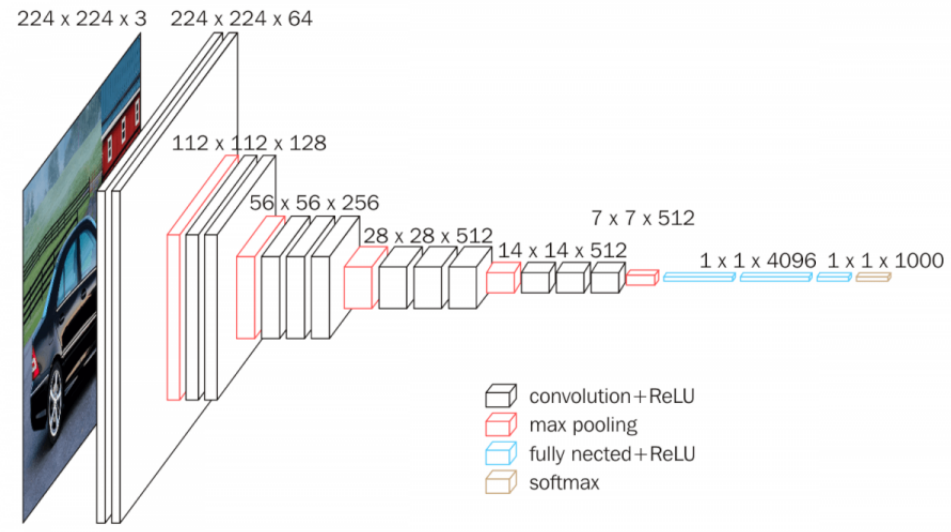
→ **다른 파라미터들은 고정**, **convolutional layer(합성곱 계층)를 추가**해 점진적으로 깊이를 증가시킴

2 CONVNET CONFIGURATIONS

CNN의 깊이 증가에 따른 성능 향상을 측정하기 위해

2011년 Ciresan의 논문, 2012년 Krizhevsky의 논문과 같은 원칙으로 층을 설정

2.1 ARCHITECTURE



입력값 : 224 x 224 RGB 이미지

전처리 : 학습 데이터의 **RGB value의 평균**을 각 **픽셀에서 빼줌**

필터 : 3 x 3, (C모델에서는 1 x 1 도 사용함)

스트라이드 : 1

패딩 : 1

풀링 : 다섯 개의 **max-pooling** 계층 → 몇개의 계층은 conv 계층 뒤에 위치

2×2 픽셀, stride는 2

합성곱 계층 + 3개의 완전 연결 계층

- 처음 2개의 계층 : 각 **4096 채널**
- 3번째 계층 : **1000개의 채널** (1000-way ILSVRC classification를 수행하기 위해)
- 마지막 계층 : **소프트맥스** 계층

모든 은닉 계층 → **ReLU** 사용

Local Response Normalisation**(LRN) 정규화**를 사용하는 네트워크는 (하나 제외하고) 없음

→ 이러한 정규화는 ILSVRC 데이터셋에 대한 **성능을 높이지 않음**(메모리 소비, 계산 시간 증가)

LRN 계층의 파라미터는 2012년 Krizhevsky의 논문과 동일

2.2 CONFIGURATIONS

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					

Table 2: **Number of parameters** (in millions).

Network	A,A-LRN	B	C	D	E
Number of parameters	133	133	134	138	144

Table 1

A에서 E로 갈수록 점진적으로 계층 수 증가(깊이 증가)

A : 11 weight layers = 8 convolutional layers + 3 FC layers

E : 19 weight layers = 16 convolutional layers + 3 FC layers

+) conv layer의 채널은 **64부터 512까지 점차적으로 늘려줌**(사이에 max-pooling layer 추가)

Table 2

깊이가 깊어도, 가중치의 수는 작은 편(깊이가 얕고 더 큰 필터 크기를 갖는 conv 계층의 다른 네트워크에 비해)

2.3 DISCUSSION

기존의 모델과 가장 큰 차이점

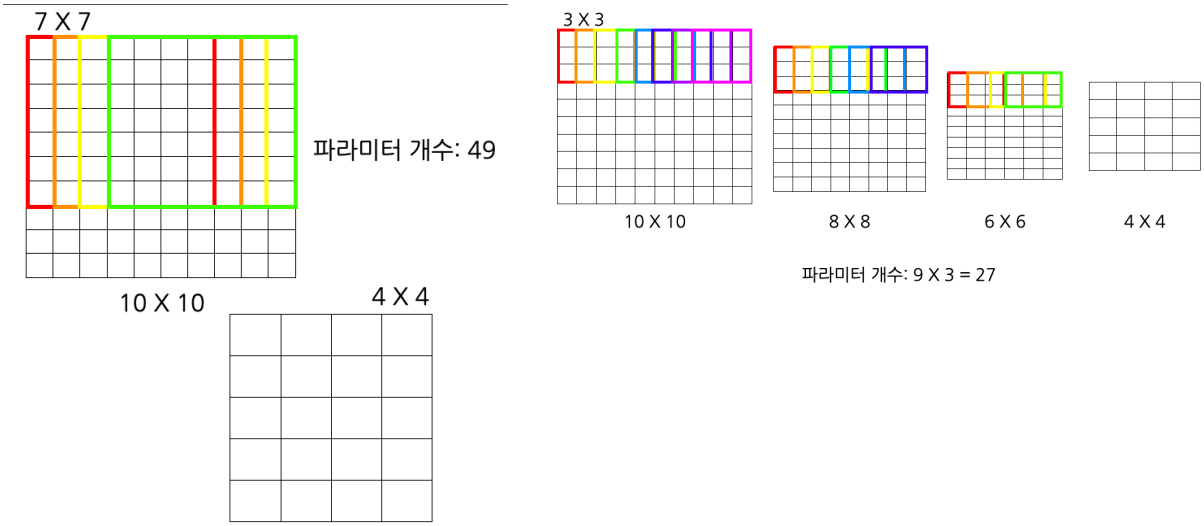
- **필터 크기를 작게** 사용 (3 x 3, 스트라이드 값 1) → 네트워크 전체에 사용해, 입력부터 **모든 픽셀에 합성곱** 해줌

  +) 2012년 Krizhevsky : 11×11, 스트라이드 값 4, 2013년 Zeiler & Fergus / Sermanet : 7×7, 스트라이드 값 2 사용

⇒ 이점?

- 한개가 아닌 **3개의 비선형 활성화 함수를 통합** 가능 → **차이를 더 정확히 구분**해냄
- **파라미터의 수를 줄일 수 있음**

ex)



- 작은 크기의 필터는 이전에도 사용되어 옴 → 2011년 Ciresan, 하지만 VGG보다 깊지 않음
- GoogLeNet : 22개의 가중치 계층과 작은 필터를 사용, 하지만 VGG보다 복잡

3 CLASSIFICATION FRAMEWORK

3.1 TRAINING

학습 과정의 대부분은 2012년 Krizhevsky 논문의 모델을 따름

- batch size : 256
- momentum : 0.9
- L2 regularization : 5 x 10^-4
- dropout ratio : 0.5
- learning rate : 10^-2 → validation set accuracy의 상승이 없을 때마다 10배 감소시킴

  → 총 3번 감소, 74 epochs(370K iteration)에서 학습을 멈춤

⇒ 2012년 Krizhevsky 논문의 모델보다 **모델이 깊고 파라미터의 수가 많지만**,

- 필터 사이즈가 더 작고

- 특정 계층의 **사전 초기화**

  로 인해 **더 낮은 에폭수**로 네트워크를 수렴시킬 수 있었음

가중치 초기화의 중요성 → 초기화를 잘못 시켜주었을 때, 깊은 네트워크에서 기울기의 불안정성 때문에 **학습을 지연**시킴

문제 해결 위해

- **랜덤 초기화**로 학습 시키기에 충분히 **얕은 A모델**(11개의 계층)로 학습을 시작
- 더 깊은 구조를 학습시킬 때는, A모델의 처음 4개의 합성곱 계층과 마지막 3개의 FC 계층을 초기화(중간 계층들은 랜덤하게 초기화 됨)

  → 랜덤 초기화를 위해, **가중치를 평균 0, 분산 10^-2의 정규분포**를 따르는 값 설정

  ⇒ 사전 초기화된 계층에는 learning rate를 줄이지 않음(학습하는 동안 변할 수 있게 함)

⇒ 논문 제출 후 **Glorot & Bengio의 무작위 초기화 절차**를 사용해, 사전 훈련없이 가중치를 초기화 할 수 있음을 발견

- biases : 0
- input image : 224 x 224 x 3 → 224 x 224를 만들기 위해 스케일링된 학습 이미지를 **랜덤하게 크롭**

  훈련 세트를 추가로 늘리기 위해, 크롭된 이미지를 랜덤하게 **수평 뒤집기**(horizontal flipping), **RGB 색상 전환**을 거침

Training image size

입력값에서 크롭된 학습 이미지에서 등방성 재조정된(isotropically-rescaled) 이미지 중 **가장 작은면을 S**라고 하면,

크롭 사이즈가 224 x 224이기 때문에, 원칙적으로 **S는 224보다 작지 않은 값**을 가짐

S >>224인 경우, 크롭된 이미지는 작은 사물이나 사물의 부분을 포함하는 이미지의 작은 부분에 해당할 것

S를 세팅하기 위한 2가지의 접근법

- **single-scale training** : S를 고정

  → S=256, S=384 : **두 가지의 고정**된 S값 사용

  → S=384 네트워크의 학습 속도를 위해, **S=256에서 사전 학습된 가중치로 초기화** + 더 작은 learning rate(10^-3) 사용

- **multi-scale training** : 각각의 **학습 개별적으로 랜덤하게 샘플링된 S**를 사용

  → **256와 512 사이**에서 **무작위로 S** 선정

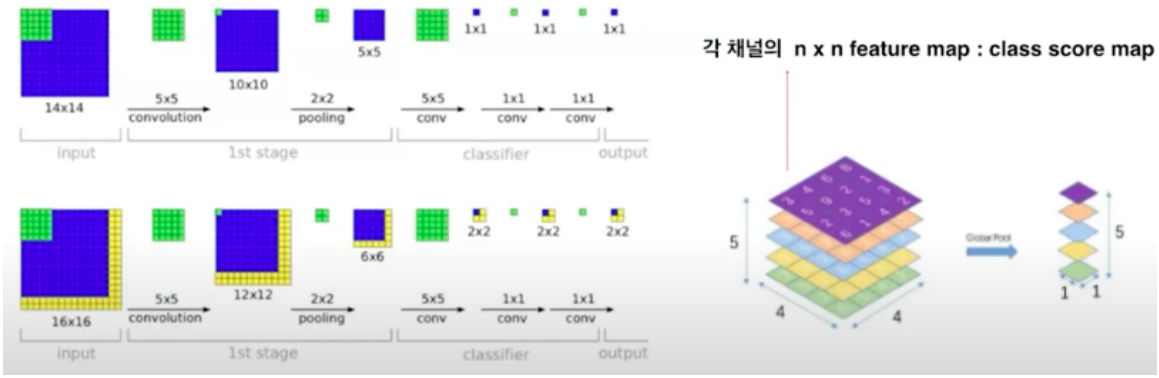
  → 이미지 속 **사물은 다른 사이즈**이기 때문에, 학습 중에 이것을 고려하는 것이 유리

⇒ 속도 문제로, S = 384(**single-scale model**)로 **사전 훈련**된 모든 계층에 **multi-scale model**로 **파인튜닝**

3.2 TESTING

1. 사전 정의된 가장 작은 이미지 면을 등방성 재조정 = Q(Q는 S와 달라도 됨)
2. FC 계층을 **합성곱 계층으로 변환**(첫 번째 FC 계층은 7 × 7으로, 마지막 두 FC 계층은 1 × 1으로)
3. 결정된 FC 네트워크는 **크롭되지 않은 이미지 전체에 적용**됨 ⇒ FC 네트워크가 전체 이미지에 적용되기 때문에, **샘플을 여러번 크롭할 필요 없음**
4. 그 결과는 클래스의 수와 같은 채널의 수를 가지는 class score map
5. 이미지의 class score의 고정된 크기의 벡터를 얻기 위해, **class score map을 공간적으로 평균화**

Testing



+) 이미지를 수평으로 뒤집어 테스트 세트를 보강

→ **원본 이미지와 뒤집힌 이미지**의 **소프트맥스 클래스** 사후를 **평균**해 **이미지의 최종 점수**를 얻음

4 CLASSIFICATION EXPERIMENTS

4.1 SINGLE SCALE EVALUATION

테스트 이미지 사이즈는 **Q = 0.5(Smin + Smax)** for fixed **S ∈ [Smin, Smax]**

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**
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**
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**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Table 2: **Number of parameters** (in millions).

Network	A,A-LRN	B	C	D	E
Number of parameters	133	133	134	138	144

Table 3: **ConvNet performance at a single test scale.**

ConvNet config. (Table 1)	smallest image side		top-1 val. error (%)	top-5 val. error (%)
	train (S)	test (Q)		
A	256	256	29.6	10.4
A-LRN	256	256	29.7	10.5
B	256	256	28.7	9.9
C	256	256	28.1	9.4
	384	384	28.1	9.3
	[256;512]	384	27.3	8.8
D	256	256	27.0	8.8
	384	384	26.8	8.7
	[256;512]	384	25.6	8.1
E	256	256	27.3	9.0
	384	384	26.9	8.7
	[256;512]	384	**25.5**	**8.0**

- LRN이 성능 향상에 그리 영향을 주지 않음 → B모델 부터 정규화를 수행하지 않음
- ConvNet 깊이가 증가할수록 분류 에러가 감소함을 확인

→ B보다 C가 성능이 더 좋음 : 추가된 비선형이 도움이 됨

→ C보다 D가 성능이 더 좋음 : **합성곱을 활용**해 공간 문맥을 캐치하는 것도 중요

4.5 COMPARISON WITH THE STATE OF THE ART

Table 7: **Comparison with the state of the art in ILSVRC classification**. Our method is denoted as "VGG". Only the results obtained without outside training data are reported.

Method	top-1 val. error (%)	top-5 val. error (%)	top-5 test error (%)
VGG (2 nets, multi-crop & dense eval.)	**23.7**	**6.8**	**6.8**
VGG (1 net, multi-crop & dense eval.)	24.4	7.1	7.0
VGG (ILSVRC submission, 7 nets, dense eval.)	24.7	7.5	7.3
GoogLeNet (Szegedy et al., 2014) (1 net)	-	7.9	
GoogLeNet (Szegedy et al., 2014) (7 nets)	-	**6.7**	
MSRA (He et al., 2014) (11 nets)	-	-	8.1
MSRA (He et al., 2014) (1 net)	27.9	9.1	9.1
Clarifai (Russakovsky et al., 2014) (multiple nets)	-	-	11.7
Clarifai (Russakovsky et al., 2014) (1 net)	-	-	12.5
Zeiler & Fergus (Zeiler & Fergus, 2013) (6 nets)	36.0	14.7	14.8
Zeiler & Fergus (Zeiler & Fergus, 2013) (1 net)	37.5	16.0	16.1
OverFeat (Sermanet et al., 2014) (7 nets)	34.0	13.2	13.6
OverFeat (Sermanet et al., 2014) (1 net)	35.7	14.2	-
Krizhevsky et al. (Krizhevsky et al., 2012) (5 nets)	38.1	16.4	16.4
Krizhevsky et al. (Krizhevsky et al., 2012) (1 net)	40.7	18.2	-

→ 전통적인 ConvNet 구조를 사용했지만, 깊이를 늘려 성능에서 상당한 효과를 보임

5 CONCLUSION

이번 작업으로 대용량의 이미지 분류를 위해 DNN(최대 19개의 가중치 계층)을 평가함

전통적인 ConvNet 구조를 깊게함으로써

- 분류의 정확도를 높임
- ImageNet challenge dataset에서 SOTA를 달성함을 입증

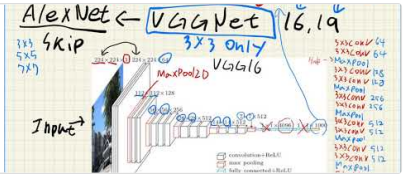
VGGNet | AI 인공지능 기초 CNN 아키텍쳐 VERY DEEP CONVOLUTIONAL NETWORKSFOR LARGE-SCALE IMAGE RECOGNITION 논문 리뷰

안녕하세요 PIEW9입니다.PIEW9 은 비전공자 전공자 상관없이 모두 모여서 인공지능 AI 논문을 읽고 리뷰하는 모임입니다.VGGNet이라고 널리 알려진 'VERY DEEP CONVOLUTIONAL NETWORKSFOR LARGE-SCALE IMAGE RECOGNITION' 논...

https://www.youtube.com/watch?v=bweAXJgZGyE

4일차-01 VGGNet소개

https://www.youtube.com/watch?v=MaDakbMDBrI



위키독스

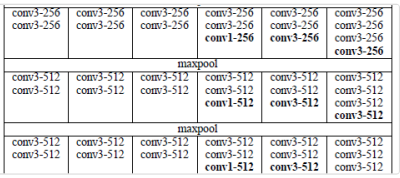
온라인 책을 제작 공유하는 플랫폼 서비스

https://wikidocs.net/118514

Very Deep Convolutional Networks For Large-Scale Image Recognition

논문의 목적은 아주 작은 3x3 필터들을 사용해 깊이를 늘리는 것이 정확도에 어떤 영향을 주는지 밝혀내는 것이다. 논문의 모델은 ImageNet Challenge 2014에서 localization과 classification에서 각각 첫 번째, 두 번째로 뛰어난 성능을 보였고, 다른 dataset에서도 통했다. 거대한 이미지 dataset, 고성능의 계산 능력(GPU), large-scale distributed

https://creamnuts.github.io/paper/VGGnet/

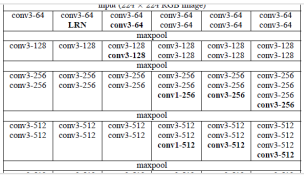


conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	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

[논문 요약5] Very Deep Convolutional Networks For Large-Scale Image Recognition

업데이트 2018.04.12 16:54] 다섯번째 요약할 논문은 "Very Deep Convolutional Networks For Large-Scale Image Recognition"(https://arxiv.org/pdf/1409.1556.pdf) 입니다. VGG Net이라고 불리우는 심층 신경망 모델로, 2014 ILSVRC (ImageNet Large-Scale Visual Recognition Challenge)에서 우승하진 못했지만 top-5 test error

https://arclab.tistory.com/160



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					