큰 크기의 이미지 인식을 위한
매우 심층적인 합성곱 네트워크

Karen Simonyan * & Andrew Zisserman

Visual Geometry Group, Department of Engineering Science, University of Oxford

< 초록 >

이 연구에서 우리는 더 깊은 합성곱 신경망 네트워크(CNN)가 대규모 이미지 인식 설정의 정확도에 미치는 영향을 조사했습니다.

우리의 주요 성과는(3 X 3) 합성곱 필터를 사용하는 구조를 이용하고, 깊이를 증가시키는 네트워크에 대한 철저한 평가이며, 이는 깊이를 16 ~ 19 중량 레이어로 밀어냄으로써 이전 기술 구성에 대한 상당한 개선이 달성될 수 있음을 보여주었습니다.
이러한 결과는 ImageNet Challenge 2014의 기초가 되었고 현지화 및 분류 부문에서 우리 팀이 각각 1위와 2위를 차지했습니다.

또한 우리의 방식이 최첨단 결과를 달성하는 다른 데이터 세트에 잘 일반화된다는 것을 보여줍니다. 우리는 컴퓨터 비전에서의 심층 시각적 표현 사용에 대한 추가 연구를 용이하게 하기 위해 두 가지 최고 성능의 ConvNet 모델을 공개적

으로 사용할 수 있게 했습니다다.

1) 도입

합성곱 네트워크(ConvNets)는 최근 ImageNet과 같은 대규모 공공 이미지 저장소와 GPU 또는 대규모 분산 클러스터와 같은 고성능 컴퓨팅 시스템으로 인해 가능해진 대규모 이미지 및 비디오 인식에서 큰 성공을 누리고 있습니다. 특히, 심층 시각 인식 아키텍처의 발전에 중요한 역할은 고차원 얕은 특징 인코딩에서 심층 ConvNets 에 이르기까지 몇 세대에 걸친 대규모 이미지 분류 시스템의 테스트베드 역할을 해온 "ImageNet 대규모 시각 인식 챌린지"(ILSVRC)에 의해 수행되었습니다.

ConvNets이 컴퓨터 비전 분야에서 더 많은 상품이 되면서, 더 나은 정확도를 달성하기 위해 Krizhevsky가 만든 구조를 개선하려는 많은 시도가 있었습니다. 예를 들어, ILSVRC2013에 제출된 결과들 중 최고 성능을 보여준 첫 번째 컨볼루션 레이어의 더 작은 수용 창 크기와 더 작은 보폭을 사용했습니다. 또 다른 개선 사항은 전체 이미지와 여러 척도에 걸쳐 네트워크를 조밀하게 훈련하고 테스트하는 것을 다루었습니다. 본 논문에서는 ConvNet 구조 설계의 또 다른 중요한 측면인 깊이를 다룹니다. 이를 위해, 우리는 아키텍처의 다른 매개 변수를 고정하고, 컨볼루션 레이어를 더 추가하여 네트워크의 깊이를 꾸준히 증가시키는데, 이는 모든 레이어에서 매우 작은 컨볼루션 필터를 사용하기 때문에 실현 가능합니다.

결과적으로 우리는 훨씬 더 정확한 ConvNet 구조를 제시합니다. 이 구조는 ILSVRC 분류 및 지역화 작업에 대한 최고의 정확도를 달성할 뿐만 아니라 다른 이미지 인식 데이터 세트에도 적용할 수 있습니다. 비교적 단순한 파이프라인의 일부로 사용될 때 우리는 추가 연구를 용이하게 하기 위해 두 가지 최고의 성능 모델을 발표했습니다.

논문의 나머지 부분은 다음과 같이 구성되어 있습니다. 2장에서는 ConvNet 구성을 설명합니다. 그런 다음 이미지 분류 훈련 및 평가에 대한 자세한 내용은 3장에서 제시하며, 구성은 4장에서 ILSVRC 분류 작업에 대해 비교합니다. 제5장에서는 논문을 마무리합니다. 또한 완전성을 위해 부록 A의 ILSVRC-2014 객체 현지화 시스템에 대해 설명하고 평가하고 부록 B의 다른 데이터 세트에 대한 매우 심층적인 기능의 일반화에 대해 논의합니다. 마지막으로 부록 C에는 주요 논문 개정 목록이 포함되어 있습니다.

2) ConvNet 구성

공평한 환경에서 ConvNet 깊이 증가로 인한 개선을 측정하기 위해 모든 ConvNet 계층 구성은 Ciresan, Krizhevsky에게서 영감을 받아 동일한 원칙을 사용하여 설계되었습니다. 이 섹션에서는 먼저 ConvNet 구성의 일반적인 레이아웃(섹션 2.1)에 대해 설명한 다음 평가에 사용된 특정 구성(섹션 2.2)에 대해 자세히 설명합니다. 그런 다음 우리의 설계 선택에 대해 논의하고 섹션 2.3의 선행기술과 비교합니다.

2-1) 구조

학습중에 ConvNets에 대한 첫 입력은 224 x 224 RGB 이미지입니다. 우리가 하는 유일한 전처리는 각 픽셀에서 훈련 세트에서 계산된 평균 RGB 값을 빼는 것입니다. 이미지는 컨볼루션(convolutional) 레이어의 스택을 통과하며, 여기서 우리는 매우 작은 수용 필드를 가진 필터를 사용합니다: 3 × 3 (좌/우, 위/아래, 중앙의 개념을 캡처하는 가장 작은 크기). 구성 중 하나에서 우리는 또한 1 × 1 컨볼루션 필터를 사용하며, 이는 입력 채널의 선형 변환으로 볼 수 있습니다(비선형성이 뒤따릅니다). 컨볼루션 스트라이드는 1픽셀로 고정됩니다. 컨볼루션 레이어 입력의 공간 패딩은 컨볼루션 후 공간 해상도가 보존되는 것입니다. 즉, 패딩은 3x3 컨볼루션 레이어에 대해 1픽셀입니다. 공간 풀링은 5개의 최대 풀링 레이어에 의해 수행되며, 이는 일부 컨볼루션 레이어를 따릅니다(모든 컨볼루션 레이어가 최대 풀링을 따르는 것은 아닙니다). 최대 풀링은 2 x 2 픽셀 창에서 스트라이드 2로 수행됩니다.

(다른 아키텍처에서 다른 깊이를 갖는) 컨볼루션 레이어의 스택은 세 개의 완전 연결(FC) 레이어에 이어집니다. 첫 번째 두 레이어는 각각 4096개의 채널을 가지고 있고, 세 번째 레이어는 1000방향 ILSVRC 분류를 수행하며, 따라서 1000개의 채널을 포함합니다(각 클래스에 하나씩). 마지막 레이어는 소프트맥스 레이어입니다. 완전히 연결된 계층의 구성은 모든 네트워크에서 동일합니다.

숨겨진 모든 레이어에는 보정 비선형성이 장착되어 있습니다. 우리는 하나의 네트워크를 제외하고 어떤 네트워크에도 LRN(Local Response Normalization) 정규화를 포함하지 않았다는 것에 주목해 주세요. 4장에서 볼 수 있듯이, 이러한 정규화는 ILSVRC 데이터 세트의 성능을 향상시키지도 않고, 오히려 메모리 소비 및 계산 시간을 증가시킵니다.

2-2) 구성

본 논문에서 실험한 ConvNet 구성은 표 1에 열당 하나씩 요약되어 있습니다. 다음에서 우리는 그것들을 A ~ E로 언급할 것입니다. 모든 구성은 섹션2.1(구조)에 제시된 일반 설계를 따르며, 깊이만 다릅니다. 네트워크 A의 11개의 가중치 계층(8개의 가중치 계층과 3개의 FC 계층)에서 네트워크 E의 19개의 가중치 계층(16개의 가중치 계층과 3개의 FC 계층). 컨볼루션 레이어의 폭(채널 수)은 첫 번째 레이어에서 64개부터 시작하여 최대 풀링 레이어마다 2배씩 증가하여 512개에 도달할 때까지 다소 작습니다.

표 2에서는 각 구성의 파라미터 수를 나타냅니다. 큰 깊이에도 불구하고, 우리 망의 가중치 수는 더 큰 컨볼루션 레이어 폭과 수용 필드를 가진 더 얕은 그물의 가중치 수보다 크지 않습니다.

2-1) 논의

우리의 ConvNet 구성은 첫 번째 변환에서 상대적으로 큰 수용 필드를 사용하는 대신 ILSVRC-2012 및 ILSVRC-2013 대회의 최고 실적 항목에 사용된 것과 상당히 다릅니다. 기존의 레이어(스트라이드가 4인 11×11, 또는 스트라이드가 2인 7×7)에 비해서 우리는 망 전체에 걸쳐 매우 작은 3×3 수용 필드를 사용합니다. 2개의 3x3 conv의 스택이 있다는 것을 쉽게 알 수 있습니다. 레이어(사이에 공간 풀링 없음)는 5×5의 유효 수용 필드를 가집니다.

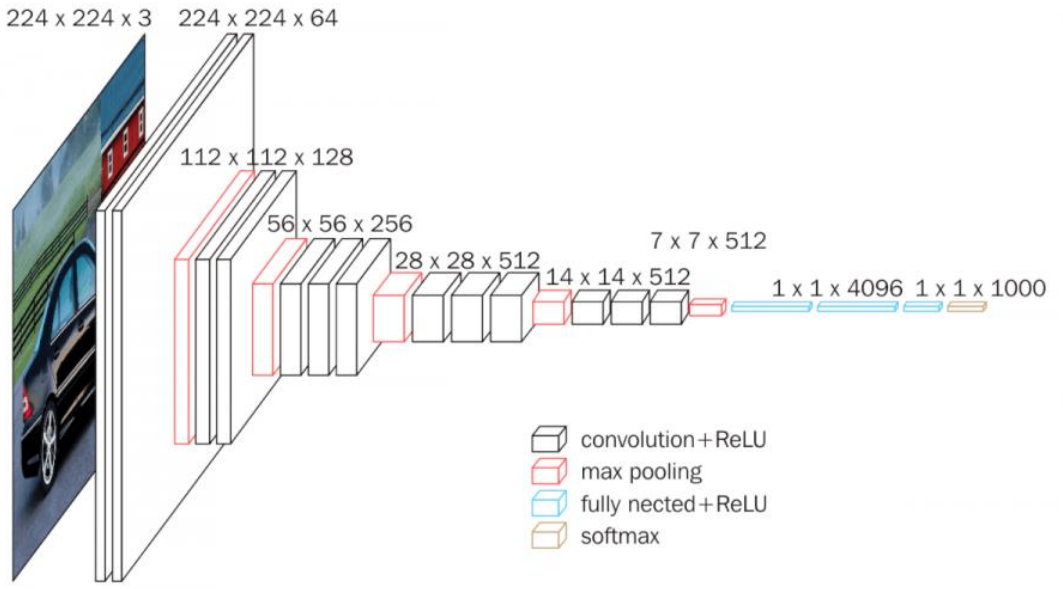
표 1: ConvNet 구성(열에 표시). 구성의 깊이는 더 많은 레이어가 추가됨에 따라 왼쪽(A)에서 오른쪽(E)으로 증가합니다(추가된 레이어는 굵게 표시됨). 컨볼루션 레이어 매개변수는 "convreceptive field sizei-hnumber of channelsi"로 표시됩니다. ReLU 활성화 함수는 간결함을 위해 표시되지 않습니다.

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

[표1, 표2]



[그림1]

기존의 레이어는 7 x 7 유효 수용 필드를 가집니다. 만약, 단일 7x7 레이어가 아닌 3x3 컨볼루션 레이어 3개를 스택으로 사용함으로써 얻을 수 있는 것은 무엇일까요?

첫째, 우리는 단일 수정 레이어 대신 세 개의 비선형 수정 레이어를 통합하여 결정 기능을 더 차별적으로 만듭니다. 둘째, 우리는 매개 변수의 수를 줄입니다. 3계층 3 × 3 컨볼루션 스택의 입력과 출력이 모두 C 채널을 가지고 있다고 가정하고, 스택은 $3(3^2C^2) = 27C^2$ 가중치로 매개 변수화됩니다. 그리고 단일 7 × 7 컨볼루션 레이어는 $7^2C^2 = 49C^2$ 매개 변수, 즉 81%가 더 필요합니다. 이는 7 × 7 컨볼루션 필터에 정규화를 부과하여 3 × 3 필터를 통해 분해를 강요하는 것으로 볼 수 있습니다(비선형성이 주입됨).

1 × 1 컨볼루션 레이어의 통합(구성 C, 표 1)은 컨볼루션 레이어의 수용 필드에

영향을 미치지 않고 결정 함수의 비선형성을 높이는 방법입니다. 우리의 경우 1 × 1 컨볼루션(convolution)은 본질적으로 동일한 차원(입력 및 출력 채널의 수가 동일)의 공간에 대한 선형 투영이지만, 수정 기능에 의해 추가적인 비선형성이 도입됩니다. 최근 Line et al. (2014)의 "Network in Network" 아키텍처에서 1×1 컨볼루션 레이어가 활용되고 있다는 점에 유의해야 합니다.

작은 크기의 컨볼루션 필터는 이전에 Ciresan이 사용했지만 그 네트워크는 우리보다 훨씬 깊지 않으며 대규모 ILSVRC 데이터로 평가하지 않았습니다. Goodfellow는 도로 번호 인식 작업에 Deep ConvNet(11개 가중치 레이어)을 적용했으며 깊이가 증가하면 성능이 향상된다는 것을 보여주었습니다. ILSVRC-2014 분류 작업의 최고 성능을 보여준 GoogLeNet은 우리 작업과 별도로 개발되었지만 매우 깊은 ConvNet(22개의 가중치 레이어)과 작은 컨볼루션 필터(3 × 3 제외, 그들은 또한 1 × 1 및 5 × 5 컨볼루션을 사용합니다. 그러나 그들의 네트워크 토폴로지는 우리보다 복잡하고 첫 번째 계층에서 피쳐 맵의 공간 해상도를 더 적극적으로 줄여 계산량을 줄입니다. 섹션4.5에서 보여주겠지만, 우리 모델은 단일 네트워크 분류 정확도 측면에서 Szegedy보다 성능이 뛰어납니다.

3) 분류 프레임워크

이전 장에서는 네트워크 구성에 대한 세부 정보를 제공했습니다. 이 섹션에서는 ConvNet의 학습 및 평가 분류에 대해 자세히 설명합니다.

3-1) 학습

ConvNet 학습 절차는 일반적으로 Krizhevsky의 방식을 따릅니다(다중 스케일 훈련 이미지에서 입력 작물을 샘플링하는 것은 제외). 즉, 학습은 추진력을 가진 미니 배치 경사 하강(역전파 기반)을 사용하여 다항 로지스틱 회귀 목표를 최적화하여 수행됩니다. 배치 크기는 256으로, 모멘텀은 0.9로 설정되었습니다. 학습은 weigt 감쇠(5 · 10-4로 설정된 L2 페널티 승수)와 처음 두 완전 연결 레이어에 대한 중도 하차 정규화에 의해 정규화되었습니다(중퇴 비율은 0.5로 설정됨). 학습 속도는 처음에 10-2로 설정되었다가 유효성 검사 세트 정확도가 향상되지 않을 때 10배 감소했습니다. 총 3회 학습률이 감소했고, 370K 반복(74에폭) 후에는 학습이 중단되었습니다. 우리는 (a) 더 깊은 깊이와 더 작은 컨볼루션 필터 크기에 의해 부과된 암시적 정규화; (b) 특정 레이어의 사전 초기화 때문에 더 많은 매개변수와 더 큰 그물망 깊이에도 불구하고, 그물망이 수렴하는 데 덜 필요한 기간이라고 추측합니다.

초기화가 잘못되면 심층 망의 기울기 불안정성으로 인해 학습이 지연될 수 있기 때문에 네트워크 가중치의 초기화가 중요합니다. 이 문제를 피하기 위해, 우리는 무작위 초기화로 학습할 수 있을 만큼 얕은 구성인 A를 훈련하는 것으로 시작했습니다. 그런 다음, 더 깊은 아키텍처를 훈련할 때, 우리는 망A의 레이어로 처음 4개의 컨볼루션 레이어와 마지막 3개의 완전히 연결된 레이어를 초기화했습니다(중간 레이어는 무작위로 초기화되었습니다). 우리는 미리 초기화된 계층에 대한 학습 속도를 줄이지 않아 학습 중에 변경될 수 있습니다. 랜덤 초기화(해당되는 경우)의 경우 평균이 0이고 분산이 10^{-2}인 정규 분포의 가중치를 샘플링했습니다. 편향은 0으로 초기화되었습니다. 논문 제출 후 Gloot & Bengio(2010)의 무작위 초기화 절차를 사용하여 사전 교육 없이 가중치를 초기화할 수 있다는 것을 알게 되었습니다.

고정 크기의 224×224 ConvNet 입력 이미지를 얻기 위해 크기가 조정된 훈련 이미지에서 무작위로 잘라냈습니다(SGD 반복당 이미지당 하나의 자르기). 훈련 세트를 더욱 강화하기 위해 결과물은 무작위 수평 뒤집기와 무작위 RGB 색상 이동을 거쳤습니다. 훈련 이미지 리스케일링은 아래에 설명되어 있습니다.

학습용 이미지 크기. S를 ConvNet 입력이 잘린 등방성으로 다시 조정된 훈련 이미지의 가장 작은 면이라고 가정합니다(S를 훈련 척도라고도 함). 자르기 크기는 224 × 224로 고정되어 있지만 원칙적으로 S는 224 이상의 값을 가질 수 있습니다. S ≫ 224의 경우 자르기는 이미지의 작은 부분에 해당하며 작은 개체 또는 개체 부분을 포함합니다.

훈련 척도 S를 설정하기 위한 두 가지 접근 방식을 고려합니다. 첫 번째는 단일 척도 훈련에 해당하는 S를 수정하는 것입니다.(샘플링된 결과물 내의 이미지 콘텐츠는 여전히 다중 척도 이미지 통계를 나타낼 수 있음). 실험에서 우리는 S = 256(이는 선행 기술에서 널리 사용됨) 및 S = 384라는 두 가지 고정 척도로 훈련된 모델을 평가했습니다. ConvNet 구성이 주어지면 먼저 S = 256을 사용하여 네트워크를 훈련했습니다. S = 384 네트워크의 up training은 S = 256으로 pre-train된 weight로 초기화되었고 우리는 10^{-3}의 더 작은 초기 학습 주기를 사용했습니다.

S 설정에 대한 두 번째 접근 방식은 다중 스케일 훈련입니다. 여기서 각 훈련 이미지는 특정 범위 $[S_{min}, S_{max}]$에서 S를 무작위로 샘플링하여 개별적으로 크기가 조정됩니다(S_{min} = 256 및 S_{max} = 512 사용). 이미지의 개체는 크기가 다를 수 있으

므로 훈련 중에 이를 고려하는 것이 좋습니다. 이것은 또한 단일 모델이 광범위한 스케일에 걸쳐 객체를 인식하도록 훈련되는 스케일 지터링에 의한 훈련 세트 증대로 볼 수 있습니다. 속도상의 이유로 고정 S = 384로 사전 훈련된 동일한 구성으로 단일 규모 모델의 모든 레이어를 미세 조정하여 다중 규모 모델을 훈련했습니다.

3-2) 시험

테스트 시 훈련된 ConvNet과 입력 이미지가 주어지면 다음과 같이 분류됩니다. 첫째, Q로 표시되는 미리 정의된 가장 작은 이미지 측면으로 등방성으로 다시 조정됩니다(테스트 스케일이라고도 함). 우리는 Q가 훈련 척도 S와 반드시 같지는 않다는 점에 주목합니다(4장에서 보겠지만, 각 S에 대해 여러 Q 값을 사용하면 성능이 향상됩니다). 그런 다음, 네트워크는 Sermanet과 유사한 방식으로 다시 스케일된 테스트 이미지에 조밀하게 적용됩니다. 즉, 완전 연결 레이어는 먼저 컨볼루션 레이어로 변환됩니다(첫 번째 FC 레이어는 7 × 7 전환 레이어로, 마지막 두 FC 레이어는 1 × 1 전환 레이어로). 그런 다음 결과로 생성된 완전 컨볼루션 네트가 전체(자르지 않은) 이미지에 적용됩니다. 결과는 입력 이미지 크기에 따라 채널 수가 클래스 수와 가변 공간 해상도를 갖는 클래스 점수 맵입니다. 마지막으로, 이미지에 대한 클래스 점수의 고정 크기 벡터를 얻기 위해 클래스 점수 맵은 공간적으로 평균화됩니다(합계 풀링). 또한 이미지를 수평으로 뒤집음으로써 테스트 세트를 보강합니다. 원본 이미지와 뒤집힌 이미지의 soft-max 클래스 사후값을 평균하여 이미지의 최종 점수를 얻습니다.

전체 이미지에 완전 컨볼루션 네트워크가 적용되기 때문에 테스트 시 여러 개의 결과물을 샘플링할 필요가 없으며, 각 결과물에 대해 네트워크 재계산이 필요하기 때문에 효율성이 떨어집니다. 동시에 Szegedy가 수행한 것처럼 대량의 결과

물 세트를 사용하면 완전 컨볼루션 네트에 비해 입력 이미지의 더 미세한 샘플링이 이루어지기 때문에 정확도가 향상될 수 있습니다. 또한, 다중 결과물 평가는 다른 컨볼루션 경계 조건으로 인해 밀도 평가를 보완합니다. 결과물에 ConvNet을 적용할 때 컨볼루션 피쳐 맵은 0으로 패딩되는 반면, 밀도 평가의 경우 동일한 결과물에 대한 패딩은 자연스럽게 이미지의 인접 부분(두 가지 모두 컨볼루션으로 인해)에서 나옵니다.이온 및 공간 풀링)을 통해 전체 네트워크 수용 필드를 크게 증가시켜 더 많은 컨텍스트를 캡처할 수 있습니다. 실제로 여러 결과물의 계산 시간이 증가했다고 해서 정확도의 잠재적 이득을 정당화할 수는 없다고 믿지만, 참고로 우리는 또한 3척에 걸쳐 총 150개의 결과물에 대해 스케일당 50개의 작물(5 x 5 일반 그리드(2플립)을 사용하여 네트워크를 평가하는데, 이는 Szegedy가 4척에 걸쳐 사용한 144개의 결과물과 맞먹습니다.

3-3) 구현 세부 정보

우리의 구현은 공개적으로 사용할 수 있는 C++ Caffe toolbox 에서 파생되었지만, 여러 가지 중요한 수정 사항을 포함하고 있어 단일 시스템에 설치된 여러 GPU에 대한 훈련 및 평가를 수행할 수 있을 뿐만 아니라 여러 스케일(a)에서 전체 크기(잘리지 않은) 이미지에 대한 훈련 및 평가를 수행할 수 있습니다. 위에서 설명한 바와 같습니다). 다중 GPU 훈련은 데이터 병렬화를 활용하며, 각 GPU에서 병렬로 처리되는 여러 GPU 배치로 각 훈련 이미지 배치를 분할하는 방식으로 수행됩니다. GPU 배치 그레이디언트를 계산한 후, 전체 배치의 그레이디언트를 얻기 위해 평균을 구합니다. 그레이디언트 계산은 GPU 전체에서 동기화되므로 결과는 단일 GPU에서 훈련할 때와 정확히 동일합니다.

ConvNet 훈련 속도를 높이는 보다 정교한 방법이 최근에 제안되었지만, 네트워크의 다른 계층에 대해 모델 및 데이터 병렬 처리를 사용하지만 개념적으로 훨

씬 더 간단한 계획은 이미 3.75배의 속도 향상을 제공한다는 것을 발견했습니다. 단일 GPU를 사용하는 것과 비교하여 기성품 4-GPU 시스템. 4개의 NVIDIA Titan Black GPU가 장착된 시스템에서 단일 네트를 교육하는 데 아키텍처에 따라 2~3주가 소요되었습니다.

4) 분류 실험

데이터세트. 이 장에서는 ILSVRC-2012 데이터 세트(ILSVRC 2012-2014 과제에 사용됨)에서 설명된 ConvNet 아키텍처에 의해 달성된 이미지 분류 결과를 제시합니다. 데이터 세트에는 1000개 클래스의 이미지가 포함되어 있으며, 교육(130만 개 이미지), 검증(50K 이미지), 테스트(보류된 클래스 레이블이 있는 100K 이미지)의 세 가지 세트로 분할됩니다. 분류 성능은 top-1 및 top-5 오차라는 두 가지 측도를 사용하여 평가됩니다. 전자는 다중 클래스 분류 오류, 즉 잘못 분류된 이미지의 비율입니다. 후자는 ILSVRC에서 사용되는 주요 평가 기준이며, 실측 실측 범주가 상위 5개 예측 범주를 벗어나도록 이미지의 비율로 계산됩니다. 대부분의 실험에서 검증 세트를 테스트 세트로 사용했습니다. 테스트 세트에 대한 특정 실험도 수행되었으며 ILSVRC-2014 대회의 "VGG" 팀 엔트리로 공식 ILSVRC 서버에 제출되었습니다.

4-1) 단일 규모 평가

2장에 설명된 레이어 구성대로 단일 규모에서 개별 ConvNet 모델의 성능을 평가하는 것으로 시작합니다. 테스트 이미지 크기는 고정 S의 경우 Q = S, 지터링

된 S ∈ [S_{min}, S_{max}]의 경우 Q = 0.5(S_{min} + S_{max})로 설정되었습니다. 결과는 표 3에 나와 있습니다. 먼저 로컬 응답 정규화(A-LRN 네트워크)를 사용했을 때 정규화 계층이 없는 모델 A에 비해서 개선된 점이 없어 보입니다다. 따라서 우리는 더 깊은 구조(B ~ E)에서 정규화를 사용하지 않습니다.

둘째, 우리는 분류 오차가 ConvNet 깊이가 증가함에 따라 감소한다는 것을 관찰합니다: A의 11개 레이어에서 E의 19개 레이어로. 특히, 동일한 깊이에도 불구하고, 구성 C(3개의 1 × 1 컨볼루션 레이어를 포함)는 네트워크 전체에서 3 × 3 컨볼루션 레이어를 사용하는 구성 D보다 성능이 떨어집니다. 이것은 추가적인 비선형성이 도움이 되는 반면(C가 B보다 낫지만), 사소한 수용 필드가 없는 컨볼루션 필터를 사용하여 공간 컨텍스트를 포착하는 것도 중요하다는 것을 나타냅니다(D가 C보다 낫습니다). 우리 아키텍처의 오류율은 깊이가 19개 계층에 도달하면 포화되지만, 더 깊은 모델도 더 큰 데이터 세트에 유용할 수 있습니다. 우리는 또한 B와 5개의 5 × 5 컨볼루션 레이어를 가진 얕은 B를 비교했는데, 이는 3 × 3 컨볼루션 레이어의 각 쌍을 단일 5 × 5 컨볼루션 레이어로 교체함으로써 B에서 파생되었습니다(섹션 2.3에 설명된 것과 동일한 수용 필드를 갖는). 얕은 네트의 상위 1개 오차는 B보다 7% 높은 것으로 측정되었으며, 이는 작은 필터를 가진 깊은 네트가 큰 필터를 가진 얕은 네트보다 성능이 우수함을 증명하는 것입니다.

마지막으로, 훈련 시간에 스케일 지터링(S ∈ [256; 512])은 테스트 시간에 단일 스케일이 사용되더라도 고정된 가장 작은 변(S = 256 또는 S = 384)이 있는 이미지에서 훈련하는 것보다 훨씬 더 나은 결과를 가져옵니다. 이것은 스케일 지터링에 의한 훈련 세트 증대가 다중 스케일 이미지 통계를 캡처하는 데 실제로 도움이 된다는 것을 확인합니다.

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**

[표3]

4-2) 다중 규모 평가

단일 스케일에서 ConvNet 모델을 평가한 후 이제 테스트 시간에 스케일 지터링의 영향을 평가합니다. 테스트 이미지의 여러 가지 크기 조정된 버전(Q의 다른 값에 해당)에 대해 모델을 실행한 다음, 결과 클래스 사후의 평균을 구하는 것으로 구성됩니다. 훈련 스케일과 테스트 스케일 간의 큰 불일치가 성능 저하로 이어진다는 점을 고려하여 고정 S로 훈련된 모델은 훈련에 가까운 세 가지 테스트 이미지 크기에 대해 평가되었습니다. Q = {S − 32, S, S + 32}을 동시에 훈련 시 스케일 지터링을 통해 테스트 시 네트워크를 더 넓은 범위의 스케일에 적용할 수 있으므로 모델은 변수 S ∈ [S_{min}; S_{max}]는 더 넓은 범위의 크기 Q = {S_{min}, 0.5(S_{min} + S_{max}), S_{max}}에 대해 평가되었습니다.

표 4에 제시된 결과는 테스트 시 스케일 지터가 (표 3에 나와 있는 것과 같이 단일 스케일에서 동일한 모델을 평가하는 것과 비교하여) 더 나은 성능을 제공한다는 것을 나타냅니다. 이전과 같이 가장 깊은 구성(D 및 E)이 가장 잘 수행되며, 가장 작은 변 S를 고정한 훈련보다 스케일 지터가 더 좋습니다. 검증 세트에서

NAT의 최고의 단일 네트워크 성능은 24.8%/7.5% 입니다(표 4에서 굵은 글씨로 강조). 테스트 세트에서 구성 E는 상위 5개 오류 중 7.3%를 달성했습니다.

Table 4: **ConvNet performance at multiple test scales.**

ConvNet config. (Table 1)	smallest image side		top-1 val. error (%)	top-5 val. error (%)
	train (S)	test (Q)		
B	256	224,256,288	28.2	9.6
C	256	224,256,288	27.7	9.2
	384	352,384,416	27.8	9.2
	[256; 512]	256,384,512	26.3	8.2
D	256	224,256,288	26.6	8.6
	384	352,384,416	26.5	8.6
	[256; 512]	256,384,512	**24.8**	**7.5**
E	256	224,256,288	26.9	8.7
	384	352,384,416	26.7	8.6
	[256; 512]	256,384,512	**24.8**	**7.5**

4-3) 다중 결과물 평가

표 5에서는 고밀도 ConvNet 평가와 다중 작물 평가를 비교합니다(자세한 내용은 섹션 3-2 참조). 또한 소프트맥스 출력을 평균화하여 두 평가 기법의 상호 보완성을 평가합니다. 보시다시피, 다중 결과물을 사용하는 것이 밀도 평가보다 약간 더 나은 성능을 발휘하며, 두 가지 접근 방식은 조합이 각각을 능가하기 때문에 실제로 상호 보완적입니다. 위에서 언급한 바와 같이, 우리는 이것이 컨볼루션 경계 조건의 다른 처리 때문이라고 가정합니다.

Table 5: **ConvNet evaluation techniques comparison.** In all experiments the training scale S was sampled from $[256; 512]$, and three test scales Q were considered: $\{256, 384, 512\}$.

ConvNet config. (Table 1)	Evaluation method	top-1 val. error (%)	top-5 val. error (%)
D	dense	24.8	7.5
	multi-crop	24.6	7.5
	multi-crop & dense	**24.4**	**7.2**
E	dense	24.8	7.5
	multi-crop	24.6	7.4
	multi-crop & dense	**24.4**	**7.1**

4-4) 앙상블

지금까지 개별 ConvNet 모델의 성능을 평가했습니다. 실험의 이 부분에서는 soft-max 클래스 후보를 평균화하여 여러 모델의 출력을 결합합니다. 이는 모델의 상호보완성으로 인해 성능이 향상되었으며 2012년과 2013년에 ILSVRC가 제출한 상위 항목에 사용되었습니다.

결과는 표6에 나와 있습니다. ILSVRC 제출 당시 우리는 단일 스케일 네트워크뿐만 아니라 다중 스케일 모델 D만 훈련했습니다(모든 레이어가 아닌 완전히 연결된 레이어만 미세 조정함). 7개 네트워크의 결과 앙상블은 7.3% ILSVRC 테스트 오류를 가집니다. 제출 후, 우리는 오직 두 개의 최고 성능의 다중 스케일 모델(구성 D 및 E)의 앙상블을 고려했는데, 이는 밀도 평가를 사용하여 테스트 오류를 7.0%로, 그리고 밀도 및 다중 작물 복합 평가를 사용하여 6.8%로 감소시켰습니다. 참고로, 우리의 최고 성능의 단일 모델은 7.1%의 오차를 달성합니다(모델 E, 표 5).

4-5) SOTA와 비교

마지막으로, 우리는 우리의 결과를 표 7의 최첨단 기술과 비교합니다. ILSVRC-2014 챌린지의 분류 과제에서 우리의 "VGG" 팀은 2위를 차지했습니다.

Table 6: **Multiple ConvNet fusion results.**

Combined ConvNet models	Error		
	top-1 val	top-5 val	top-5 test
ILSVRC submission			
(D/256/224,256,288), (D/384/352,384,416), (D/[256;512]/256,384,512) (C/256/224,256,288), (C/384/352,384,416) (E/256/224,256,288), (E/384/352,384,416)	24.7	7.5	7.3
post-submission			
(D/[256;512]/256,384,512), (E/[256;512]/256,384,512), dense eval.	24.0	7.1	7.0
(D/[256;512]/256,384,512), (E/[256;512]/256,384,512), multi-crop	23.9	7.2	-
(D/[256;512]/256,384,512), (E/[256;512]/256,384,512), multi-crop & dense eval.	**23.7**	**6.8**	**6.8**

7개 모델의 앙상블을 사용하여 7.3%의 테스트 오류가 발생했습니다. 제출 후, 우리는 두 모델의 앙상블을 사용하여 오류율을 6.8%로 줄였습니다.

표 7에서 볼 수 있듯이, 매우 심층적인 ConvNets는 ILSVRC-2012 및 ILSVRC-2013 대회에서 최고의 결과를 달성한 이전 세대의 모델을 크게 능가합니다. 우리의 결과는 1등(오류 6.7%의 GoogleLeNet)에 대해서도 경쟁력이 있으며 외부 교육 데이터로 11.2%, 없이 11.7%를 달성한 ILSVRC-2013 수상 제출자 Clarifai를 크게 능가합니다. 이는 대부분의 ILSVRC 제출에서 사용된 것보다 훨씬 적은 두 가지 모델만 결합해도 최상의 결과를 얻을 수 있다는 점을 고려하면 놀라운 일입니다. 단일 네트워크 성능 측면에서, 우리의 아키텍처는 단일 GoogleLeNet을 0.9% 능가하는 최고의 결과(7.0% 테스트 오류)를 달성합니다. 특히, 우리는 LeCun et al.(1989)의 고전적인 ConvNet 아키텍처에서 벗어나지 않고, 깊이를 상당히 증가시켜 개선했습니다.

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	-

5) 결론

이 작업에서 우리는 대규모 이미지 분류를 위해 매우 심층적인 컨볼루션 네트워크(최대 19개의 가중치 계층)를 평가했습니다. 표현 깊이가 분류 정확도에 유용하며, 깊이가 상당히 향상된 기존 ConvNet 아키텍처를 사용하여 ImageNet 챌린지 데이터 세트에서 최첨단 성능을 달성할 수 있음을 입증했습니다. 부록에서, 우리는 또한 우리의 모델이 덜 깊은 이미지 표현을 중심으로 구축된 더 복잡한 인식 파이프라인과 일치하거나 능가하는 광범위한 작업 및 데이터 세트에 잘 일반화된다는 것을 보여줍니다. 우리의 결과는 시각적 표현에서 깊이의 중요성을 다시 한번 확인합니다.

감사의 말

이 작업은 ERC VisRec 번호 228180에 의해 지원되었습니다. 이 연구에 사용된

GPU를 기부한 NVIDIA Corporation의 지원에 감사드립니다.

참조

(생략)