대규모 이미지 인식을 위한 매우 심층적인 컨볼루션 네트워크

카렌 시모니안* **& 앤드류 지서먼+**
옥스퍼드 대학교 공학 과학부 시각 기하학 그룹
{karen,az}@robots.ox.ac.uk

초록

이 연구에서는 대규모 이미지 인식 설정에서 컨볼루션 네트워크 깊이가 정확도에 미치는 영향을 조사합니다. 우리의 주요 기여는 매우 작은(3 × 3) 컨볼루션 필터가 있는 아키텍처를 사용하여 깊이를 증가시키는 네트워크에 대한 철저한 평가로, 깊이를 16~19개의 가중치 레이어로 밀어 넣음으로써 이전 구성에서 상당한 개선을 이룰 수 있음을 보여줍니다. 이러한 연구 결과는 이미지넷 챌린지 2014에 출품하여 지역 식별과 분류 트랙에서 각각 1, 2위를 차지한 기반이 되었습니다. 또한 우리의 표현이 다른 데이터 세트에도 잘 일반화되어 최첨단 결과를 얻을 수 있음을 보여주었습니다. 우리는 컴퓨터 비전에서 심층 시각 표현을 사용하는 데 대한 추가 연구를 촉진하기 위해 최고 성능의 두 가지 ConvNet 모델을 공개적으로 사용할 수 있도록 했습니다.

1 소개

컨볼루션 네트워크(ConvNet)는 최근 대규모 이미지 및 비디오 인식 분야에서 큰 성공을 거두었으며(Krizhevsky et al., 2012; Zeiler & Fergus, 2013; Sermanet et al., 2014; Simonyan & Zisserman, 2014), 이는 ImageNet(Deng et al., 2009) 같은 대규모 공개 이미지 저장소와 GPU 또는 대규모 분산 클러스터와 같은 고성능 컴퓨팅 시스템으로 인해 가능해졌습니다(Dean et al., 2012). 특히, 고차원 얕은 특징 부호화(Perronnin et al., 2010)(ILSVRC-2011 우승자)에서부터 심층 ConvNets(Krizhevsky et al., 2012)에 이르기까지 몇 세대에 걸친 대규모 이미지 분류 시스템의 테스트베드 역할을 해온 ImageNet 대규모 시각 인식 챌린지(ILSVRC)가 딥 시각 인식 아키텍처의 발전에 중요한 역할을 했습니다(Russakovsky et al., 2014).

컴퓨터 비전 분야에서 ConvNet이 점점 더 보편화됨에 따라 더 나은 정확도를 달성하기 위해 Krizhevsky 등(2012)의 원래 아키텍처를 개선하려는 시도가 많이 이루어졌습니다. 예를 들어, ILSVRC- 2013에 제출된 논문 중 가장 우수한 성적을 거둔 논문은 더 작은 수용 창 크기와 첫 번째 컨볼루션 레이어의 작은 보폭을 활용했습니다(Zeiler & Fergus, 2013; Sermanet et al., 2014). 또 다른 개선 사항은 전체 이미지와 여러 스케일에 걸쳐 네트워크를 조밀하게 훈련하고 테스트하는 것

이었습니다(Sermanet et al., 2014; Howard, 2014). 이 백서에서는 ConvNet 아키텍처 설계의 또 다른 중요한 측면인 깊이를 다룹니다. 이를 위해 아키텍처의 다른 매개 변수를 수정하고 모든 레이어에 매우 작은(3 × 3) 컨볼루션 필터를 사용하기 때문에 가능한 컨볼루션 레이어를 더 추가하여 네트워크의 깊이를 꾸준히 증가시킵니다.

그 결과, 훨씬 더 정확한 ConvNet 아키텍처를 개발하게 되었으며, 이 아키텍처는 ILSVRC 분류 및 로컬라이제이션 작업에서 최첨단 정확도를 달성할 뿐만 아니라 다른 이미지 인식 데이터 세트에도 적용 가능하며, 비교적 간단한 파이프라인(예: 미세 조정 없이 선형 SVM으로 분류된 심층 특징)의 일부로 사용해도 뛰어난 성능을 달성할 수 있습니다. 추가 연구를 용이하게 하기 위해 가장 성능이 우수한 두 가지 모델을 공개했습니다([1]).

백서의 나머지 부분은 다음과 같이 구성됩니다. 2절에서는 ConvNet 구성에 대해 설명합니다. 그런 다음 이미지 분류 훈련 및 평가에 대한 자세한 내용은 섹션 3에서 설명합니다. 3절, 그리고

[*] 현재 소속: 구글 딥마인드 [+] 현재 소속: 옥스퍼드 대학교 및 구글 딥마인드

[1] http://www.robots.ox.ac.uk/~vgg/research/very_deep/

의 ILSVRC 분류 작업에서 구성을 비교합니다. 4. 5절에서 논문을 마무리합니다. 완전성을 위해 부록 A에서는 ILSVRC-2014 객체 로컬라이제이션 시스템을 설명하고 평가하며, 부록 B에서는 다른 데이터 세트에 대한 매우 심층적인 특징의 일반화에 대해 논의합니다. 마지막으로 부록 C에는 주요 논문 개정 목록이 포함되어 있습니다.

2 CONVNET 구성

공정한 설정에서 ConvNet 깊이가 증가함에 따른 개선 효과를 측정하기 위해, 모든 ConvNet 계층 구성은 Ciresan 외. (2011), Krizhevsky 외. (2012)에서 영감을 받아 동일한 원칙을 사용하여 설계되었습니다. 이 섹션에서는 먼저 ConvNet 구성의 일반적인 레이아웃을 설명한 다음(2.1절), 평가에 사용된 특정 구성을 자세히 설명합니다(2.2절). 그런 다음 2.3절에서 우리의 설계 선택에 대해 논의하고 선행 기술과 비교합니다.

2.1 아키텍처

훈련 중에 ConvNet에 입력되는 것은 고정된 크기의 224×224 RGB 이미지입니다. 우리가 수행하는 유일한 사전 처리는 훈련 세트에서 계산된 평균 RGB 값을 각 픽셀에서 빼는 것입니다. 이미지는 컨볼루션(conv.) 레이어 스택을 통과하며, 여기서 3×3(왼쪽/오른쪽, 위/아래, 중앙의 개념을 포착할 수 있는 가장 작은 크기)의 매우 작은 수용 필드를 가진 필터를 사용합니다. 구성 중 하나에서는 1 × 1 컨볼루션 필터도 사용하는데, 이는 입력 채널의 선형 변환(비선형성 뒤따름)으로 볼 수 있습니다. 컨볼루션 보폭은 1픽셀로 고정되며, 컨볼루션 레이어 입력의 공간 패딩은 컨볼루션 후 공간 해상도가 유지되도록, 즉 3×3 컨볼루션 레이어의 경우 패딩이 1픽셀이 되도록 설정됩니다. 공간 풀링은 5개의 최대 풀링 레이어에 의해 수행되며, 이 레이어는 일부 컨볼루션 레이어를 따릅니다(모든 컨볼루션 레이어에 최대 풀링이 적용되는 것은 아님). 최대 풀링은 2 × 2 픽셀 창에서 보폭 2로 수행됩니다.

컨볼루션 레이어 스택(아키텍처마다 깊이가 다름)에 이어 세 개의 완전 연결(FC) 레이어가 이어집니다. 처음 두 개는 각각 4096개의 채널을 가지고 있고, 세 번째 레이어는 1000방향 ILSVRC 분류를 수행하므로 1000개의 채널(각 클래스당 하나씩)이 포함됩니다. 마지막 레이어는 소프트-맥스 레이어입니다. 완전히 연결된 레이어의 구성은 모든 네트워크에서 동일합니다.

모든 숨겨진 레이어에는 정류(ReLU (Krizhevsky et al., 2012)) 비선형성이 장착되어 있습니다. (하나를 제외한) 모든 네트워크에는 국부 응답 정 규 화 (LRN) 정규화(Krizhevsky et al., 2012)가 포함되어 있지 않다는 점 에 유의하십시오. 4에서 살펴볼 수 있듯이, 이러한 정규화는 ILSVRC 데이터 세트의 성능을 향상시키지는 못하지만 메모리 사용량과 계산 시간을 증가시킵니다. 해당되는 경우, LRN 계층의 파라미터는 (Krizhevsky et al., 2012)의 파 라 미 터 를 사용합니다.

2.2 구성

이 백서에서 평가한 ConvNet 구성은 표 1에 열당 하나씩 요약되어 있습니다. 아래에서는 넷을 이

름(A-E)으로 지칭합니다. 모든 구성은 2.1절에 제시된 일반적인 설계를 따르며, 네트워크 A의 가중치 레이어는 11개(8개의 변환 레이어와 3개의 FC 레이어)에서 네트워크 E의 가중치 레이어는 19개(16개의 변환 레이어와 3개의 FC 레이어)로 깊이에만 차이가 있습니다. 컨볼루션 레이어의 폭(채널 수)은 첫 번째 레이어에서 64에서 시작하여 최대 풀링 레이어마다 2씩 증가하여 512에 도달할 때까지 다소 작습니다.

표 2에서는 각 구성에 대한 매개변수 수를 보고합니다. 큰 수심에도 불구하고, 우리 그물의 가중치 수는 더 큰 컨버전스 레이어 폭과 수용 필드를 가진 더 얕은 그물의 가중치 수보다 크지 않습니다 (144M 가중치 (Sermanet et al., 2014)).

2.3 토론

우리의 ConvNet 구성은 ILSVRC-2012(Krizhevsky 외., 2012) 및 ILSVRC-2013 대회(Zeiler & Fergus, 2013; Sermanet 외., 2014)의 최고 성능 출품작에 사용된 것과는 상당히 다릅니다. 첫 번째 컨볼루션 레이어에서 비교적 큰 수용 필드(예: 보폭 4인치의 11×11(Krizhevsky et al., 2012) 또는 보폭 2인치의 7×7(Zeiler & Fergus, 2013; Sermanet et al., 2014))를 사용하는 대신, 전체 네트워크에서 매우 작은 3×3 수용 필드를 사용하여 모든 픽셀에서 입력(보폭 1)과 함께 컨볼루션을 수행합니다. 두 개의 3 × 3 컨볼루션 레이어 스택(중간에 공간 풀링이 없음)의 유효 수용 필드는 5 × 5임을 쉽게 알 수 있습니다.

표 1: **ConvNet 구성**(열로 표시). 구성의 깊이는 더 많은 레이어가 추가될수록 왼쪽(A)에서 오른쪽(E)으로 증가합니다(추가된 레이어는 굵게 표시됨). 컨볼루션 레이어 파라미터는 "conv⟨수용 필드 크기⟩-⟨채널 수⟩"로 표시됩니다. 간결성을 위해 ReLU 활성화 함수는 표시되지 않습니다.

ConvNet 구성					
A	A-LRN	B	C	D	E
11 무게 레이어	11 무게 레이어	13 무게 레이어	16 무게 레이어	16 무게 레이어	19 무게 레이어
입력(224 × 224 RGB 이미지)					
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 **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 **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 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
소프트 맥스					

표 2: **매개변수 수**(백만 단위).

네트워크	A,A-LRN	B	C	D	E
매개변수 수	133	133	134	138	144

이러한 레이어는 7 × 7 유효 수신 필드를 갖습니다. 그렇다면 단일 7 × 7 레이어 대신 3 × 3 변환 레이어 3개로 구성된 스택을 사용하면 어떤 이점이 있을까요? 첫째, 단일 레이어 대신 3개의 비선형 정류 레이어를 통합하여 결정 함수의 변별력을 높입니다. 둘째, 매개변수 수를 줄 였습니다. 입력 과 출력 모두 3개 층 3 × 3 컨볼루션 스택에는 C 채널이 있으며, 스택은 $3(3^2C^2) = 27C^2$ 가중치; 동시에 단일 7 × 7 변환 레이어에는 $7^2C^2 = 49C^2$ 파라미터, 즉 81% 더 많은 파라미터가 필요합니다. 이는 7 × 7 변환 필터에 정규화를 적용하여 3 × 3 필터를 통해 분해되도록 강제하는 것으로 볼 수 있습니다(중간에 비선형성이 주입됨).

1 × 1 컨볼루션 레이어(구성 C, 표 1)를 통합하는 것은 컨볼루션 레이어의 수용 필드에 영향을 주

지 않고 결정 함수의 비선형성을 높이는 방법입니다. 우리의 경우 1 × 1 컨볼루션은 본질적으로 동일한 차원의 공간에 대한 선형 투영이지만(입력 및 출력 채널 수가 동일함) 정류 함수에 의해 추가적인 비선형성이 도입됩니다. 1 × 1 컨볼루션 레이어는 최근 Lin 등(2014)의 "네트워크 내 네트워크" 아키텍처에서 활용되고 있다는 점에 유의해야 합니다.

작은 크기의 컨볼루션 필터는 이전에 Ciresan 등(2011)이 사용한 적이 있지만, 그들의 네트워크는 우리보다 훨씬 깊이가 낮고 대규모 ILSVRC 데이터 세트에 대해 평가하지 않았습니다. Goodfellow 등(2014)은 도로 번호 인식 작업에 심층 ConvNet(11개의 가중치 레이어) 을 적용했으며, 그 결과 깊이가 증가할수록 성능이 향상되는 것을 보여주었습니다. ILSVRC-2014 분류 과제에서 가장 우수한 성능을 보인 GoogLeNet(Szegedy et al., 2014)은 본 연구와는 독립적으로 개발되었지만 매우 심층적인 ConvNet을 기반으로 한다는 점에서 유사합니다.

(22개의 가중치 레이어)와 작은 컨볼루션 필터(3×3을 제외하고 1×1 및 5×5 컨볼루션도 사용)를 사용합니다. 그러나 이들의 네트워크 토폴로지는 우리보다 더 복잡하며, 피처 맵의 공간 복원력을 첫 번째 레이어에서 더 적극적으로 줄여 계산량을 줄입니다. 4.5절에서 살펴볼 수 있듯이 4.5에서 살펴볼 수 있듯이, 우리의 모델은 단일 네트워크 분류 정확도 측면에서 Szegedy 등(2014)의 모델보다 성능이 뛰어납니다.

3 분류 프레임워크

이전 섹션에서는 네트워크 구성에 대해 자세히 설명했습니다. 이 섹션에서는 분류 ConvNet 훈련 및 평가에 대한 세부 사항을 설명합니다.

3.1 교육

ConvNet 훈련 절차는 일반적으로 Krizhevsky 등(2012)의 방법을 따릅니다(나중에 설명하는 것처럼 다중 스케일 훈련 이미지에서 입력 작물을 샘플링하는 것은 제외). 즉, 훈련은 운동량과 함께 미니 배치 경사 하강(역전파(LeCun et al., 1989) 기반)을 사용하여 다항 로지스틱 회귀 목표를 최적화하여 수행됩니다. 배치 크기는 256, 모멘텀은 0.9로 설정했습니다. 훈련은 가중치 감쇠에 의해 정규화되었습니다(L_2 페널티 승수는 다음과 같이 설정되었습니다.
5 - 10^{-4})과 완전히 연결된 처음 두 개의 레이어에 대한 드롭아웃 정규화(*드롭아웃* 비율 0.5로 설정
). 학습률은 처음에 10^{-2} 으로 설정한 후 유효성 검사 세트 정확도가 개선되지 않으면 10배로 감소했습니다. 총 3번의 학습률 감소가 이루어졌으며, 학습률은
은 370,000회 반복(74 에포크) 후에 중단되었습니다. (Krizhevsky et al., 2012)에 비해 파라미터 수가 더 많고 그물의 깊이가 더 컸음에도 불구하고, 그물이 수렴하는 데 더 적은 에포크가 필요했던 이유는 (a) 더 큰 깊이와 더 작은 변환 필터 크기로 인한 암시적 정규화, (b) 특정 레이어의 사전 초기화 때문이라고 추측합니다.

네트워크 가중치를 잘못 초기화하면 딥넷의 경사도가 불안정해져 학습이 지연될 수 있으므로 네트워크 가중치를 초기화하는 것이 중요합니다. 이 문제를 피하기 위해 무작위 초기화로 훈련할 수 있을 정도로 얕은 구성 A(표 1)를 훈련하는 것으로 시작했습니다. 그런 다음 더 깊은 아키텍처를 훈련할 때는 처음 4개의 컨볼루션 레이어와 완전히 연결된 마지막 3개의 레이어를 네트워크 A의 레이어로 초기화했습니다(중간 레이어는 무작위로 초기화됨). 미리 초기화된 레이어의 학습 속도를 낮추지 않았기 때문에 학습 중에 레이어가 변경될 수 있었습니다.
무작위 초기화(해당되는 경우)의 경우, 평균이 0이고 분산이 10인 정규 분포에서 가중치를 샘플링했습니다($^{-2}$). 편향은 0으로 초기화했습니다. 논문 제출 후 사전 훈련 없이도 가중치를 초기화할 수 있다는 사실을 알게 되었습니다.
의 무작위 초기화 절차를 사용하여 초기화합니다(Glorot & Bengio, 2010).

고정 크기 224×224 ConvNet 입력 이미지를 얻기 위해 리스케일링된 훈련 이미지에서 무작위로 크롭했습니다(SGD 반복당 이미지당 한 번씩 크롭). 훈련 집합을 더욱 보강하기 위해 자른 이미지에 무작위 수평 뒤집기와 무작위 RGB 색상 이동을 적용했습니다(Krizhevsky et al., 2012). 훈련 이미지 리스케일링은 아래에 설명되어 있습니다.

훈련 이미지 크기. *S*는 등방성으로 크기 조정된 훈련 이미지의 가장 작은 면으로, 이로부터 ConvNet 입력이 잘립니다(*S*를 훈련 스케일이라고도 합니다). 잘라내기 크기는 224 × 224로 고정되어 있지만 원칙적으로 *S*는 224 이상의 값을 취할 수 있습니다. $S = 224$인 경우 잘라내기는 훈련 이미지의 가장 작은 면에 완전히 걸쳐 전체 이미지 통계를 캡처하고, $S \gg 224$인 경우 잘라내기는 작은 물체 또는 물체 부분이 포함된 이미지의 작은 부분에 해당합니다.

첫 번째는 단일 스케일 훈련에 해당하는 S를 고정하는 것입니다(샘플링된 작물 내의 이미지 콘텐츠는 여전히 다중 스케일 이미지 통계를 나타낼 수 있음). 실험에서는 두 가지 고정 스케일로 훈련된 모델을 평가했습니다: $S = 256$(선행 기술에서 널리 사용되어 왔습니다(Krizhevsky 외., 2012; Zeiler & Fergus, 2013; Sermanet 외., 2014)) 및 $S = 384$. ConvNet 구성이 주어지면 먼저 $S = 256$을 사용하여 네트워크를 훈련했습니다. $S = 384$ 네트워크의 훈련 속도를 높이기 위해 다음과 같은 가중치로 초기화했습니다.
$S = 256$으로 사전 학습되었으며, 초기 학습률은 10^{-3} 으로 더 작게 사용했습니다.
*S*를 설정하는 두 번째 접근 방식은 다중 스케일 훈련으로, 각 훈련 이미지가 특정 범위[S_{min} , S_{max}]($S_{min} = 256$ 및 $S_{max} = 512$ 사용)에서 무작위로 샘플링하여 개별적으로 스케일을 재조정합니다. 이미지 속 물체의 크기가 다를 수 있으므로 훈련 시 이를 고려하는 것이 좋습니다. 이는 스케일 지터링에 의한 훈련 세트 증강으로 볼 수도 있습니다.

모델은 다양한 스케일의 물체를 인식하도록 훈련됩니다. 속도상의 이유로, 고정 $S = 384$로 사전 학습된 단일 스케일 모델의 모든 레이어를 동일한 구성으로 미세 조정하여 다중 스케일 모델을 학습했습니다.

3.2 테스트

테스트 시 훈련된 ConvNet과 입력 이미지가 주어지면 다음과 같은 방식으로 분류됩니다. 먼저, 미리 정의된 가장 작은 이미지 측면으로 등방성 리스케일링하여 Q(테스트 스케일이라고도 함)로 표시합니다. Q가 반드시 훈련 스케일 S와 같을 필요는 없습니다(4절에서 설명할 것입니다. 4절에서 설명하겠지만, 각 S에 대해 여러 값의 Q를 사용하면 성능이 향상됩니다.) 그런 다음 네트워크는 (Sermanet et al., 2014)와 유사한 방식으로 재조정된 테스트 이미지에 조밀하게 적용됩니다. 즉, 완전히 연결된 레이어는 먼저 컨볼루션 레이어로 변환됩니다(첫 번째 FC 레이어는 7×7 컨볼루션 레이어로, 마지막 두 FC 레이어는 1×1 컨볼루션 레이어로 변환). 그런 다음 결과물인 완전 컨볼루션 네트워크를 전체(자르지 않은) 이미지에 적용합니다. 그 결과 채널 수가 클래스 수와 동일한 클래스 점수 맵과 입력 이미지 크기에 따라 가변적인 공간 해상도가 생성됩니다. 마지막으로 이미지에 대한 클래스 점수의 고정 크기 벡터를 얻기 위해 클래스 점수 맵을 공간적으로 평균화(합산 풀링)합니다. 또한 이미지를 수평으로 뒤집어 테스트 세트를 보강하고, 원본 이미지와 뒤집힌 이미지의 소프트 최대 클래스 후위를 평균하여 이미지의 최종 점수를 얻습니다.

완전 컨볼루션 네트워크는 전체 이미지에 적용되기 때문에 테스트 시 여러 작물을 샘플링할 필요가 없지만(Krizhevsky et al., 2012), 각 작물에 대해 네트워크를 다시 계산해야 하므로 효율성이 떨어집니다. 동시에 Szegedy 등(2014)의 연구처럼 대규모 작물 세트를 사용하면 완전 컨볼루션 네트워크에 비해 입력 이미지를 더 세밀하게 샘플링할 수 있기 때문에 정확도가 향상될 수 있습니다. 또한 다중 작물 평가는 서로 다른 컨볼루션 경계 조건으로 인해 밀도 평가와 상호 보완적입니다. 작물에 ConvNet을 적용할 때 컨볼루션 특징 맵에 0이 채워지는 반면, 밀도 평가의 경우 동일한 작물에 대한 패딩이 이미지의 인접한 부분에서 자연스럽게 제공되므로(컨볼루션과 공간 풀링 모두로 인해) 전체 네트워크 수용 필드가 크게 증가하여 더 많은 컨텍스트를 포착할 수 있습니다. 실제로는 여러 크롭의 계산 시간 증가가 정확도의 잠재적 향상을 정당화하지 못한다고 생각하지만, 참고로 저희는 스케일당 50개의 크롭(5×5 일반 격자, 2번 뒤집기)을 사용하여 네트워크를 평가했으며, 이는 3스케일에 걸쳐 총 150개의 크롭으로, Szegedy 등(2014)이 사용한 4스케일에 걸쳐 144개의 크롭과 비슷한 수준입니다.

3.3 구현 세부 정보

우리의 구현은 공개적으로 사용 가능한 C++ Caffe 툴박스(Jia, 2013)에서 파생되었지만(2013년 12월에 분기됨), 여러 가지 중요한 수정 사항이 포함되어 있어 단일 시스템에 설치된 여러 GPU에서 훈련 및 평가를 수행할 수 있을 뿐만 아니라 위에서 설명한 대로 여러 스케일의 풀 사이즈(자르지 않은) 이미지에서 훈련 및 평가할 수 있습니다. 멀티 GPU 트레이닝은 데이터 병렬 처리를 활용하

며, 트레이닝 이미지의 각 배치를 여러 개의 GPU 배치로 분할하여 각 GPU에서 병렬로 처리하는 방식으로 수행됩니다. GPU 배치 그라데이션이 계산된 후에는 평균을 내어 전체 배치의 그라데이션을 얻습니다. 그라디언트 계산은 GPU 전체에서 동기식으로 이루어지므로 단일 GPU에서 훈련할 때와 결과가 완전히 동일합니다.

최근 네트워크의 여러 계층에 모델 및 데이터 병렬 처리를 사용하는 보다 정교한 ConvNet 훈련 속도 *향상* 방법이 제안되었지만(Krizhevsky, 2014), 개념적으로 훨씬 더 간단한 방식이 이미 상용 4-GPU 시스템에서 단일 GPU를 사용하는 것보다 3.75배의 속도 향상을 제공한다는 사실을 발견했습니다. 4개의 NVIDIA 타이탄 블랙 GPU가 장착된 시스템에서 단일 네트워크를 트레이닝하는 데는 아키텍처에 따라 2~3주가 걸렸습니다.

4 분류 실험

데이터 세트. 이 섹션에서는 ILSVRC-2012 데이터 세트(ILSVRC 2012-2014 대회에 사용됨)에 대해 설명한 ConvNet 아키텍처를 통해 얻은 이미지 분류 결과를 제시합니다. 이 데이터 세트에는 1000개의 클래스 이미지가 포함되어 있으며, 훈련(1.3M 이미지), 검증(50만 이미지), 테스트(클래스 레이블이 보류된 10만 이미지)의 세 가지 세트로 나뉩니다. 분류 성능은 상위 1%와 상위 5% 오류의 두 가지 측정값을 사용하여 평가됩니다. 전자는 다중 클래스 분류 오류, 즉 잘못 분류된 이미지의 비율이며 후자는

ILSVRC에서 사용되는 주요 평가 기준이며, 실사 범주가 예측된 상위 5개 범주에 속하지 않는 이미지의 비율로 계산됩니다.

대부분의 실험에서는 검증 세트를 테스트 세트로 사용했습니다. 일부 실험은 테스트 세트에서 수행되어 ILSVRC-2014 대회에 "VGG" 팀 출품작으로 공식 ILSVRC 서버에 제출되기도 했습니다 (Russakovsky et al., 2014).

4.1 단일 척도 평가

2.2절에서 설명한 레이어 구성으로 단일 규모에서 개별 ConvNet 모델의 성능을 평가하는 것부터 시작합니다. 테스트 이미지 크기는 다음과 같이 설정했습니다: 고정 S의 경우 $Q = S$, 지터링된 $S \in [S_{min}, S_{max}]$의 경우 $Q = 0.5(S_{min} + S_{max})$. 결과는 표 3에 나와 있습니다.

먼저, 로컬 응답 정규화(A-LRN 네트워크)를 사용하면 정규화 레이어가 없는 모델 A에서 개선되지 않는다는 점에 주목합니다. 따라서 더 심층적인 아키텍처(B-E)에서는 정규화를 사용하지 않습니다.

둘째, ConvNet 깊이가 증가함에 따라 분류 오류가 감소하는 것을 관찰할 수 있습니다(A의 경우 11개 레이어에서 E의 경우 19개 레이어). 특히, 동일한 깊이에도 불구하고 1 × 1 컨볼루션 레이어 3개가 포함된 구성 C가 네트워크 전체에 3 × 3 컨볼루션 레이어를 사용하는 구성 D보다 성능이 더 나쁩니다. 이는 추가적인 비선형성이 도움이 되기는 하지만(C가 B보다 낫다), 사소한 수신 필드가 아닌 컨볼루션 필터를 사용하여 공간 컨텍스트를 캡처하는 것도 중요하다는 것을 나타냅니다(D가 C보다 낫다). 이 아키텍처의 오류율은 깊이가 19층에 도달하면 포화 상태에 이르지만, 더 깊은 모델이 더 큰 데이터 세트에 유리할 수 있습니다. 또한 3 × 3 컨볼루션 레이어 쌍을 5 × 5 컨볼루션 레이어 하나로 대체하여 B에서 파생된 5 × 5 컨볼루션 레이어 5개가 있는 얕은 넷 B와 비교했습니다(2.3절에서 설명한 것과 동일한 수용 필드를 가짐). 얕은 그물의 상위 1 오차는 (중앙 작물에서) B보다 7% 더 높은 것으로 측정되었으며, 이는 작은 필터가 있는 깊은 그물이 더 큰 필터가 있는 얕은 그물보다 성능이 우수하다는 것을 확인시켜 줍니다.

마지막으로, 훈련 시 스케일 지터링($S \in [256; 512]$)은 테스트 시 단일 스케일을 사용하더라도 가장 작은 면이 고정된 이미지($S = 256$ 또는 $S = 384$)에서 훈련하는 것보다 훨씬 더 나은 결과를 가져옵니다. 이는 스케일 지터링에 의한 훈련 세트 증강이 다중 스케일 이미지 통계를 캡처하는 데 실제로 도움이 된다는 것을 확인시켜 줍니다.

표 3: **단일 테스트 규모에서의 ConvNet 성능.**

ConvNet 구성 (표 1)	가장 작은 이미지 측면		TOP-1 VAL. ERROR (%)	TOP-5 VAL. 오류 (%)
	기차 (S)	테스트 (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**

4.2 다중 규모 평가

단일 스케일에서 ConvNet 모델을 평가했으므로 이제 테스트 시 스케일 지터링의 영향을 평가합니다. 이 평가는 테스트 이미지의 여러 재조정된 버전(서로 다른 Q 값에 해당)에 대해 모델을 실행한 다음 결과 클래스 포스트어에 대한 평균을 구하는 방식으로 이루어집니다. 훈련과 테스트 스케일 간의 불일치가 크면 성능이 저하된다는 점을 고려하여 고정 S로 훈련된 모델은 훈련 이미지에 가까운 세 가지 테스트 이미지 크기에 대해 평가되었습니다: $Q =$ $\{S - 32, S, S + 32\}$. 동시에 훈련 시 스케일 지터를 사용하면 테스트 시 네트워크를 더 넓은 범위의 스케일에 적용할 수 있으므로 변수 $S \in [S_{min} ; S_{max}]$로 훈련된 모델을 더 넓은 범위의 크기 $Q = \{S_{min}, 0.5(S_{min} + S_{max}), S_{max}\}$에 대해 평가했습니다.

표 4에 제시된 결과는 테스트 시 스케일 지터링이 더 나은 성능으로 이어진다는 것을 나타냅니다(표 3에 표시된 단일 스케일로 동일한 모델을 평가하는 것과 비교). 이전과 마찬가지로 가장 심층적인 구성(D 및 E)의 성능이 가장 우수하며, 스케일 지터링은 고정된 최소 측면 *S로* 훈련하는 것보다 낫습니다. 검증 세트에서 가장 우수한 단일 네트워크 성능은 24.8%/7.5% 상위 1 / 상위 5 오류입니다(표 4에서 굵은 글씨로 강조 표시됨). 테스트 세트에서 구성 E는 7.3%의 상위 5위 오류를 달성합니다.

표 4: **여러 테스트 규모에서의 ConvNet 성능.**

ConvNet 구성 (표 1)	가장 작은 이미지 측면		TOP-1 VAL. ERROR (%)	TOP-5 VAL. 오류 (%)
	기차 (*S*)	테스트 (*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 작물 평가

표 5에서는 밀도 높은 ConvNet 평가와 다중 크롭 평가를 비교합니다(꼬리 제거에 대해서는 3.2절 참조). 또한 두 평가 기법의 소프트 최대 출력의 평균을 구하여 두 평가 기법의 상호 보완성을 평가합니다. 보시다시피, 여러 작물을 사용하는 것이 밀도 평가보다 약간 더 나은 성능을 보이며, 두 접근법의 조합이 각각보다 더 나은 성능을 보이기 때문에 두 접근법은 실제로 상호보완적입니다. 위에서 언급했듯이 이는 컨볼루션 경계 조건의 처리 방식이 다르기 때문이라는 가설을 세웠습니다.

표 5: **ConvNet 평가 기법 비교.** 모든 실험에서 훈련 척도 *S는* [256; 512]에서 샘플링되었으며, 세 가지 테스트 척도 *Q가* 고려되었습니다: {256, 384, 512}.

ConvNet 구성 (표 1)	평가 방법	TOP-1 VAL. ERROR (%)	TOP-5 VAL. 오류(%)
D	밀도	24.8	7.5
	다중 자르기	24.6	7.5
	멀티 크롭 및 고밀도	**24.4**	**7.2**
E	밀도	24.8	7.5
	다중 자르기	24.6	7.4
	멀티 크롭 및 고밀도	**24.4**	**7.1**

4.4 CONVNET 퓨전

지금까지는 개별 ConvNet 모델의 성능을 평가했습니다. 이 실험에서는 여러 모델의 소프트맥스 클래스 후방 평균을 구하여 여러 모델의 결과를 결합합니다. 이 방법은 모델의 상호보완성으로 인해 성능을 향상시키며, 2012년(Krizhevsky 외., 2012)과 2013년(Zeiler & Fergus, 2013; Sermanet 외., 2014)에 제출된 상위 ILSVRC 논문에서 사용되었습니다.

결과는 표 6에 나와 있습니다. ILSVRC 제출 당시에는 단일 규모 네트워크와 다중 규모 모델 D(모

든 레이어가 아닌 완전히 연결된 레이어만 미세 조정)만 훈련했습니다. 그 결과 7개 네트워크의 앙상블은 7.3%의 ILSVRC 테스트 오류를 보였습니다. 제출 후, 가장 성능이 좋은 두 가지 멀티스케일 모델(구성 D와 E)로만 구성된 앙상블을 고려한 결과, 밀도 평가에서는 테스트 오류가 7.0%, 밀도 평가와 멀티크롭 평가를 결합한 경우 6.8%로 감소했습니다. 참고로, 가장 성능이 좋은 단일 모델의 오류는 7.1%입니다(모델 E, 표 5).

4.5 최첨단 기술과의 비교

마지막으로, 우리의 결과를 표 7의 최신 기술과 비교해 보았습니다. ILSVRC-2014 챌린지 (Russakovsky et al., 2014)의 분류 과제에서 "VGG" 팀은 다음과 같이 2위를 차지했습니다.

표 6: **여러 ConvNet 융합 결과.**

결합된 ConvNet 모델	오류		
	TOP-1 VAL	TOP-5 VAL	TOP 5 테스트
ILSVRC 제출			
(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
제출 후			
(D/[256;512]/256,384,512), (E/[256;512]/256,384,512), 고밀도 평가.	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), 멀티 크롭 및 고밀도 평가.	**23.7**	**6.8**	**6.8**

7개의 모델로 구성된 앙상블을 사용한 테스트 오류율 7.3%. 제출 후 오류율을 다음과 같이 낮췄습니다. 6.8% 2가지 모델의 앙상블 사용.

표 7에서 볼 수 있듯이, 우리의 매우 심층적인 ConvNet은 ILSVRC-2012 및 ILSVRC-2013 대회에서 최고의 결과를 얻은 이전 세대의 모델보다 훨씬 뛰어난 성능을 보입니다. 또한 분류 과제 우승자(6.7%의 오류를 기록한 GoogLeNet)와 비교해도 경쟁력이 있으며, 외부 훈련 데이터를 사용했을 때 11.2%, 사용하지 않았을 때 11.7%를 기록한 ILSVRC-2013 우승작 Clarifai보다 훨씬 뛰어난 성능을 보였습니다. 대부분의 ILSVRC 출품작에서 사용되는 것보다 훨씬 적은 두 개의 모델만 결합하여 최고의 결과를 얻었다는 점을 고려하면 이는 놀라운 결과입니다. 단일 넷 성능 측면에서도, 우리의 아키텍처는 단일 *GoogLeNet*을 0.9% 능가하는 최고의 결과(테스트 오류 7.0%)를 달성했습니다. 주목할 만한 점은 LeCun 등(1989)의 고전적인 ConvNet 아키텍처에서 벗어나지 않고 심도를 크게 높여 개선했다는 점입니다.

표 7: **ILSVRC 분류의 최신 기술과의 비교**. 우리의 방법은 "VGG"로 표시됩니다. 외부 훈련 데이터 없이 얻은 결과만 보고됩니다.

방법	TOP-1 VAL. ERROR (%)	TOP-5 VAL. 오류 (%)	상위 5위 테스트 오류(%)
VGG(2망, 다중 작물 및 밀도 평가)	**23.7**	**6.8**	**6.8**
VGG(1망, 다중 작물 및 밀도 평가)	24.4	7.1	7.0
VGG(ILSVRC 제출, 7망, 밀도 평가)	24.7	7.5	7.3
GoogLeNet(Szegedy 외, 2014) (1망)	-	7.9	
GoogLeNet(Szegedy 외, 2014)(7망)	-	**6.7**	
MSRA (He et al., 2014) (11 그물)	-	-	8.1
MSRA(He et al., 2014) (1 순)	27.9	9.1	9.1
클라리파이(러시아코프스키 외, 2014) (다중 그물)	-	-	11.7
클라리파이(러사코브스키 외, 2014) (1망)	-	-	12.5
자일러 & 퍼거스 (자일러 & 퍼거스, 2013) (6망)	36.0	14.7	14.8
자일러 & 퍼거스 (자일러 & 퍼거스, 2013) (1망)	37.5	16.0	16.1
오버팻(Sermanet et al., 2014) (7망)	34.0	13.2	13.6
오버팻(Sermanet 외, 2014) (1 순)	35.7	14.2	-
(크리제프스키 외, 2012) (5 그물) (5 그물)	38.1	16.4	16.4
(크리제프스키 외, 2012) (1망)	40.7	18.2	-

5 결론

이 연구에서는 대규모 이미지 분류를 위해 매우 심층적인 컨볼루션 네트워크(최대 19개의 가중치 레이어)를 평가했습니다. 표현 깊이가 분류 정확도에 도움이 되며, 깊이가 상당히 증가된 기존의 ConvNet 아키텍처(LeCun 외., 1989; Krizhevsky 외., 2012)를 사용하여 ImageNet 챌린지 데이터

세트에서 최첨단 성능을 달성할 수 있음을 입증했습니다. 또한 부록에서는 딥러닝 모델이 다양한 작업과 데이터 세트에 잘 일반화되어 덜 심도 있는 이미지 표현을 중심으로 구축된 더 복잡한 인식 파이프라인과 일치하거나 그 성능을 능가한다는 것을 보여줍니다. 이번 연구 결과는 시각적 표현에서 깊이가 얼마나 중요한지 다시 한 번 확인시켜 줍니다.

감사

이 작업은 ERC 보조금 VisRec 번호. 228180. 이 연구에 사용된 GPU를 기증해 주신 NVIDIA Corporation의 지원에 감사드립니다.

참조

Bell, S., Upchurch, P., Snavely, N. 및 Bala, K. 컨텍스트 데이터베이스의 자료를 사용한 야생에서의 자료 인식. *CoRR*, abs/1412.0623, 2014.

채트필드, K., 시모니안, K., 베달디, A., 지서만, A. 세부적인 악마의 귀환: 컨볼루션 그물 깊숙이 파고들기. In *Proc. BMVC.*, 2014.

심포이, M., 마지, S., 및 베달디, A. 텍스처 인식 및 분할을 위한 심층 컨볼루션 필터 뱅크. *CoRR*, abs/1411.6836, 2014.

Ciresan, D. C., Meier, U., Masci, J., Gambardella, L. M. 및 Schmidhuber, J. 이미지 분류를 위한 유연한 고성능 컨볼루션 신경망. In *IJCAI*, pp. 1237-1242, 2011.

딘, J., 코라도, G., 몽가, R., 첸, K., 데빈, M., 마오, M., 란자토, M., 시니어, A., 터커, P., 양, K., 레, Q. V., 응, A. Y. 대규모 분산 심층 네트워크. In *NIPS*, pp. 1232-1240, 2012.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L. Imagenet: 대규모 계층적 이미지 데이터베이스. In *Proc. CVPR*, 2009.

도나휴, J., 지아, Y., 빈얄스, O., 호프만, J., 장, N., 쳉, E., 대럴, T. 디카프: 일반적인 시각 인식을 위한 심층 컨볼루션 활성화 기능. *CoRR*, abs/1310.1531, 2013.

에버링햄, M., 에슬라미, S. M. A., 반 굴, L., 윌리엄스, C., 윈, J., 지서만, A. 파스칼 시각 객체 클래스의 도전: 회고. *IJCV*, 111(1):98-136, 2015.

Fei-Fei, L., Fergus, R. 및 Perona, P. 몇 가지 훈련 예제에서 생성적 시각 모델 학습: 101개의 객체 범주에 대해 테스트한 증분적 베이지안 접근법. *생성 모델 기반 비전의 IEEE CVPR 워크샵*, 2004.

Girshick, R. B., Donahue, J., Darrell, T., Malik, J. 정확한 객체 감지 및 의미적 세분화를 위한 풍부한 기능 계층 구조. *CoRR*, abs/1311.2524v5, 2014. Proc. CVPR, 2014.

Gkioxari, G., Girshick, R. 및 Malik, J. 전체와 부분의 작용 및 속성. *CoRR*, abs/1412.2604, 2014.

글로롯, X. 및 벤지오, Y. 심층 피드포워드 신경망 훈련의 어려움에 대한 이해. In *Proc. AISTATS*, 9권, 249-256쪽, 2010.

굿펠로우, I. J., 불라토프, Y., 이바르즈, J., 아누드, S., 셰트, V. 심층 컨볼루션 신경망을 이용한 스트리트 뷰 이미지에서 여러 자리 숫자 인식. In *Proc. ICLR*, 2014.

Griffin, G., Holub, A. 및 Perona, P. Caltech-256 객체 범주 데이터 세트. 기술 보고서 7694, 캘리포니아 공과대학, 2007.

시각 인식을 위한 심층 컨볼루션 네트워크의 공간 피라미드 풀링(He, K., Zhang, X., Ren, S. 및 Sun, J.). *CoRR*, abs/1406.4729v2, 2014.

Hoai, M. 이미지 분류를 위한 정규화된 최대 풀링. In *Proc. BMVC.*, 2014.

Howard, A. G. 심층 컨볼루션 신경망 기반 이미지 분류에 대한 몇 가지 개선 사항. In *Proc. ICLR*, 2014.

Jia, Y. Caffe: 빠른 기능 임베딩을 위한 오픈 소스 컨볼루션 아키텍처 . http://caffe.berkeleyvision.org/, 2013.

Karpathy, A. 및 Fei-Fei, L. 이미지 설명 생성을 위한 심층 시각적 의미론적 정렬. *CoRR*, abs/1412.2306, 2014.

Kiros, R., Salakhutdinov, R. 및 Zemel, R. S. 시각적 의미 임베딩과 다중 모드 신경 언어 모델의 통합. *CoRR*, abs/1411.2539, 2014.

크리제프스키, A. 컨볼루션 신경망을 병렬화하는 한 가지 이상한 트릭. *CoRR*, abs/1404.5997, 2014. Krizhevsky, A., Sutskever, I., Hinton, G. E. 심층 컨볼루션 신경망을 사용한 이미지넷 분류-.

작동합니다. *NIPS*, 1106-1114쪽, 2012.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. 및 Jackel, L. D. 필기 우편 번호 인식에 적용된 백그라운드. *신경 계산*, 1(4):541-551, 1989.

Lin, M., Chen, Q., and Yan, S. 네트워크 내 네트워크. In *Proc. ICLR*, 2014.

Long, J., Shelhamer, E. 및 Darrell, T. 의미론적 세분화를 위한 완전 컨볼루션 네트워크. *CoRR*, abs/1411.4038, 2014.

Oquab, M., Bottou, L., Laptev, I. 및 Sivic, J. 컨볼루션 신경망을 사용한 중간 수준의 이미지 표현 학습 및 전송. In *Proc. CVPR*, 2014.

Perronnin, F., Sa´nchez, J. 및 Mensink, T. 대규모 이미지 분류를 위한 피셔 커널 개선. In *Proc. ECCV*, 2010.

Razavian, A., Azizpour, H., 설리반, J., 칼슨, S. CNN 기성품 기능: 인식을 위한 놀라운 기준선. *CoRR*, abs/1403.6382, 2014.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. 및 Fei-Fei, L. ImageNet 대규모 시각 인식 챌린지. *CoRR*, abs/1409.0575, 2014.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y. OverFeat: 컨볼루션 네트워크를 사용한 통합 인식, 로컬라이제이션 및 탐지. In *Proc. ICLR*, 2014.

Simonyan, K. 및 Zisserman, A. 비디오에서 동작 인식을 위한 2스트림 컨볼루션 네트워크. *CoRR*, abs/1406.2199, 2014. Proc. NIPS, 2014.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. 컨볼루션으로 더 깊이 들어가기. *CoRR*, abs/1409.4842, 2014.

웨이, Y., 샤, W., 황, J., 니, B., 동, J., 자오, Y., 얀, S. CNN: 단일 레이블에서 다중 레이블로. *CoRR*, abs/1406.5726, 2014.

Zeiler, M. D. and Fergus, R. 컨볼루션 네트워크 시각화 및 이해. *CoRR*, abs/1311.2901, 2013. Proc. ECCV, 2014.

A 현지화

이 백서의 본론에서는 ILSVRC 챌린지의 분류 과제를 고려하고 다양한 깊이의 ConvNet 아키텍처에 대한 철저한 평가를 수행했습니다. 이 섹션에서는 2014년에 25.3%의 오차로 우승한 바 있는 이 챌린지의 로컬라이제이션 과제에 대해 살펴봅니다. 이 과제는 클래스의 실제 객체 수와 관계없이 상위 5개 클래스 각각에 대해 하나의 객체 경계 상자를 예측해야 하는 객체 감지의 특수한 경우로 볼 수 있습니다. 이를 위해 ILSVRC-2013 지역화 챌린지 우승자인 Sermanet 외(2014)의 접근 방식을 약간의 수정을 가하여 채택했습니다. 우리의 방법은 Sect. A.1에 설명되어 있으며 Sect. A.2.

A.1 로컬라이제이션 컨버넷

오브젝트 로컬라이제이션을 수행하려면 클래스 점수 대신 마지막으로 완전히 연결된 레이어가 바운딩 박스 위치를 예측하는 매우 심층적인 ConvNet을 사용합니다. 바운딩 박스는 중심 좌표, 너비, 높이가 저장된 4D 벡터로 표현됩니다. 바운딩 박스 예측을 모든 클래스에서 공유할지(단일 클래스 회귀, SCR(Sermanet et al., 2014)) 아니면 클래스별로 수행할지(클래스별 회귀, PCR)를 선택할 수 있습니다. 전자의 경우 마지막 계층은 4-D이고, 후자의 경우 데이터 집합에 1000개의 클래스가 있기 때문에 4000-D입니다. 마지막 바운딩 박스 예측 레이어와는 별도로, 16개의 가중치 레이어를 포함하고 분류 작업에서 가장 우수한 성능을 보이는 것으로 밝혀진 ConvNet 아키텍처 D(표 1)를 사용합니다(섹션 4).

훈련. 로컬라이제이션 ConvNet의 훈련은 분류 ConvNet의 훈련과 유사합니다(3.1절). 가장 큰 차이점은 로지스틱 회귀 목표를 유클리드 손실로 대체하여 예측된 바운딩 박스 매개변수의 실측값과의 편차에 불이익을 준다는 것입니다. 각각 단일 스케일로 두 개의 로컬라이제이션 모델을 훈련했습니다: $S = 256$ 및 $S = 384$(시간 제약으로 인해 ILSVRC-2014 제출에는 훈련 스케일 지터링을 사용하지 않았습니다). 훈련은 다음과 같이 진행되었습니다.

해당 분류 모델(동일한 척도로 훈련)로 초기화했으며, 초기 학습률은 10^{-3}으로 설정했습니다. 모든 레이어를 미세 조정하는 방법과 (Sermanet et al., 2014)에서 수행한 것처럼 완전히 연결된 처음

두 개의 레이어만 미세 조정하는 방법을 모두 탐색했습니다. 마지막으로 완전히 연결된 레이어는 무작위로 초기화되고 처음부터 학습됩니다.

테스트. 두 가지 테스트 프로토콜을 고려합니다. 첫 번째는 유효성 검사 집합에서 서로 다른 네트워크 수정 사항을 비교하는 데 사용되며, 분류 오류를 고려하기 위해 기준값 클래스에 대한 경계 상자 예측만 고려합니다. 바운딩 박스는 이미지의 중앙 크롭에만 네트워크를 적용하여 얻습니다.

두 번째 본격적인 테스트 절차는 분류 작업(3.2절)과 유사하게 전체 이미지에 로컬라이제이션 ConvNet을 밀도 있게 적용하는 것을 기반으로 합니다. 차이점은 클래스 점수 맵 대신 마지막으로 완전히 연결된 레이어의 출력은 경계 상자 예측 집합이라는 점입니다. 최종 예측을 도출하기 위해 먼저 공간적으로 가까운 예측을 병합한 다음(이들의 평균을 구하여) 분류 ConvNet에서 얻은 클래스 점수를 기반으로 등급을 매기는 Sermanet 등(2014)의 욕심 병합 절차를 활용합니다. 여러 로컬라이제이션 ConvNet을 사용하는 경우 먼저 바운딩 박스 예측 집합의 합을 취한 다음 합에 대해 병합 절차를 실행합니다. 다중 풀링은 사용하지 않았습니다.

오프셋 기법을 사용하여 바운딩 박스 예측의 공간 해상도를 높이고 결과를 더욱 개선할 수 있습니다(Sermanet et al., 2014).

A.2 현지화 실험

이 섹션에서는 먼저 가장 성능이 좋은 현지화 설정을 결정한 다음(첫 번째 테스트 프로토콜 사용), 본격적인 시나리오(두 번째 프로토콜)에서 이를 평가합니다. 로컬라이제이션 오차는 ILSVRC 기준(Russakovsky et al., 2014)에 따라 측정됩니다. 즉, 실측 바운딩 박스와 결합 *비율에* 대한 교차가 0.5 이상이면 바운딩 박스 예측이 올바른 것으로 간주됩니다.

설정 비교. 표 8에서 볼 수 있듯이 클래스별 회귀(PCR)가 클래스에 구애받지 않는 단일 클래스 회귀(SCR)보다 성능이 뛰어나며, 이는 PCR이 SCR보다 성능이 뛰어났다는 Sermanet 등(2014)의 연구 결과와는 다릅니다. 또한 로칼라이제이션 작업을 위해 모든 레이어를 미세 조정하는 것이 완전히 연결된 레이어만 미세 조정하는 것보다 눈에 띄게 더 나은 결과를 가져온다는 사실에 주목합니다(Sermanet et al., 2014). 이 실험에서는 가장 작은 이미지 측면을 $S = 384$로 설정했으며, $S = 256$의 결과도 동일한 동작을 나타내므로 간결성을 위해 표시하지 않았습니다.

표 8: 간소화된 테스트 프로토콜을 사용한 **다양한 수정에 대한 로컬라이제이션 오류**: 바운딩 박스는 단일 중앙 이미지 크롭에서 예측되고 실측 기준 클래스가 사용됩니다. 마지막 레이어를 제외한 모든 ConvNet 레이어는 구성 D(표 1)이며, 마지막 레이어는 단일 클래스 회귀(SCR) 또는 클래스별 회귀(PCR) 중 하나를 수행합니다.

미세 조정된 레이어	회귀 유형	GT 클래스 현지화 오류
1, 2차 FC	SCR	36.4
	PCR	34.3
모두	PCR	**33.1**

본격적인 평가. 최상의 로컬라이제이션 설정(PCR, 모든 레이어의 미세 조정)을 결정한 후, 이제 이를 본격적인 시나리오에 적용하여 최고 성능의 분류 시스템(4.5절)을 사용하여 상위 5개의 클래스 라벨을 예측하고, 밀도 높게 계산된 여러 바운딩 박스 예측을 Sermanet 등(2014)의 방법을 사용하여 병합합니다. 표 9에서 볼 수 있듯이, 지상 실측 대신 상위 5개의 예측 클래스 레이블을 사용했음에도 불구하고 전체 이미지에 로컬라이제이션 ConvNet을 적용하면 중앙 자르기(표 8)를 사용할 때보다 결과가 크게 향상됩니다. 분류 작업(섹션 4)과 마찬가지로, 여러 스케일에서 테스트하고 여러 네트워크의 예측을 결합하면 성능이 더욱 향상됩니다.

표 9: **현지화 오류**

가장 작은 이미지 측면		상위 5위 현지화 오류(%)	
기차 (*S*)	테스트 (*Q*)	val.	테스트.
256	256	29.5	-
384	384	28.2	26.7
384	352,384	27.5	-
퓨전: 256/256 및 384/352,384		**26.9**	**25.3**

최신 기술과의 비교. 표 10에서 최고의 로컬라이제이션 결과와 최신 기술을 비교했습니다. 25.3%의 테스트 오류를 기록한 "VGG" 팀은 ILSVRC-2014의 로컬라이제이션 챌린지에서 우승했습니다(Russakovsky et al., 2014). 주목할 만한 점은 더 적은 스케일을 사용하고 해상도 향상 기법을 사용하지 않았음에도 불구하고 ILSVRC-2013 우승팀인 오버피트(Sermanet et al., 2014)의 결과보다 훨씬 나은 결과를 얻었다는 점입니다. 우리는 이 기법을 우리의 방법에 통합하면 더 나은 로컬라이제이션 성능을 달성할 수 있을 것으로 예상합니다. 이는 매우 심층적인 ConvNet이 가져온 성능 향상을 나타냅니다. 더 간단한 로컬라이제이션 방법으로 더 나은 결과를 얻었지만 더 강력한 표현을 얻을 수 있었습니다.

B 매우 심층적인 기능의 일반화

이전 섹션에서는 ILSVRC 데이터 세트에 대한 매우 심층적인 ConvNet의 훈련과 평가에 대해 설명했습니다. 이 섹션에서는 ILSVRC에서 사전 훈련된 ConvNet을 다음과 같은 특징으로 평가합니다.

표 10: **ILSVRC 로컬라이제이션의 최신 기술과의 비교**. 우리의 방법은 "VGG"로 표시됩니다.

방법	TOP-5 VAL. 오류 (%)	상위 5위 테스트 오류(%)
VGG	**26.9**	**25.3**
GoogLeNet(Szegedy 외., 2014)	-	26.7
과제중(Sermanet et al., 2014)	30.0	29.9
(크리제프스키 외, 2012)	-	34.2

추출기를 다른 소규모 데이터 세트에 사용할 수 있으며, 과적합으로 인해 처음부터 대규모 모델을 훈련할 수 없습니다. 최근 이러한 사용 사례에 대한 관심이 높아졌는데, ILSVRC에서 학습된 심층 이미지 표현이 다른 데이터 세트에도 잘 일반화되어 수작업으로 만든 표현을 큰 차이로 능가하는 것으로 밝혀졌기 때문입니다(Zeiler & Fergus, 2013; Donahue 외., 2013; Razavian 외., 2014; Chatfield 외., 2014). 이러한 작업의 연장선상에서, 저희의 모델이 최신 방법에서 사용되는 얕은 모델보다 더 나은 성능을 제공하는지 조사했습니다. 이 평가에서는 ILSVRC(섹션 4)에서 분류 성능이 가장 우수한 두 가지 모델, 즉 "Net-D" 및 "Net-E" 구성(공개적으로 제공됨)을 고려합니다.

다른 데이터 세트의 이미지 분류를 위해 ILSVRC에 대해 사전 훈련된 ConvNets를 활용하기 위해, 마지막으로 완전히 연결된 레이어(1000방향 ILSVRC 분류를 수행)를 제거하고 두 번째 레이어의 4096-D 활성화를 여러 위치와 규모에 걸쳐 집계된 이미지 피처로 사용합니다. 결과 이미지 설명자는 정규화되고 대상 데이터 세트에 대해 학습된 선형 SVM 분류기와 결합된 L_2 입니다. 간소화를 위해 사전 학습된 ConvNet 가중치는 고정된 상태로 유지됩니다(미세 조정은 수행되지 않음).

특징의 집계는 ILSVRC 평가 절차와 유사한 방식으로 수행됩니다(3.2절). 즉, 먼저 이미지의 가장 작은 면이 Q와 같도록 이미지의 크기를 조정한 다음 이미지 평면에 네트워크 작업을 조밀하게 적용합니다(모든 가중치 레이어가 컨볼루션으로 처리될 때 가능). 그런 다음 결과 피처 맵에 대해 글로벌 평균 풀링을 수행하여 4096-D 이미지 설명자를 생성합니다. 그런 다음 이 디스크립터를 가로-세로로 반전된 이미지의 디스크립터와 평균을 냅니다. 4.2절에서 설명한 바와 같이 4.2에서 살펴본 바와 같이 여러 스케일에 걸쳐 평가하는 것이 유리하므로 여러 스케일에 걸쳐 특징을 추출합니다. 결과물인 다중 스케일 특징은 스케일 간에 스택하거나 풀링할 수 있습니다. 스태킹을 사용하면 후속 분류기가 다양한 스케일에 걸쳐 이미지 통계를 최적으로 결합하는 방법을 학습할 수 있지만, 설명자 차원이 증가한다는 단점이 있습니다. 아래 실험에서 이 설계 선택에 대한 논의를 다시 이어가겠습니다. 또한 두 개의 네트워크를 사용하여 계산된 특징의 후기 융합을 평가하는데, 이는 각각의 이미지 설명자를 쌓아서 수행됩니다.

표 11: **VOC-2007, VOC-2012, Caltech-101 및 Caltech-256의 최신 이미지 분류 기술 비교**. 당사의 모델은 "VGG"로 표시됩니다. 로 표시된 결과는 *확장된* ILSVRC 데이터 세트(2000개 클래스)에 대해 사전 훈련된 ConvNets를 사용하여 얻은 결과입니다.

방법	VOC-2007 (평균 AP)	VOC-2012 (평균 AP)	Caltech-101 (평균 클래스 리콜)	Caltech-256 (평균 클래스 리콜)
자일러 & 퍼거스(Zeiler & Fergus, 2013)	-	79.0	86.5 ± 0.5	74.2 ± 0.3
(채트필드 외, 2014)	82.4	83.2	88.4 ± 0.6	77.6 ± 0.1
(He et al., 2014)	82.4	-	**93.4 ± 0.5**	-
웨이 등(Wei et al., 2014)	81.5 (85.2)*	81.7 (**90.3)***	-	-
VGG Net-D(16 레이어)	89.3	89.0	91.8 ± 1.0	85.0 ± 0.2

VGG Net-E(19 레이어)	89.3	89.0	92.3 ± 0.5	85.1 ± 0.3
VGG Net-D & Net-E	**89.7**	**89.3**	92.7 ± 0.5	**86.2 ± 0.3**

VOC-2007 및 VOC-2012에 대한 이미지 분류. 먼저 PASCAL VOC-2007 및 VOC-2012 벤치마크의 이미지 분류 과제에 대한 평가부터 시작합니다(Everingham et al., 2015). 이 데이터 세트에는 각각 10K 및 22.5K 이미지가 포함되어 있으며, 각 이미지에는 20개의 객체 범주에 해당하는 하나 또는 여러 개의 레이블이 주석으로 지정되어 있습니다. VOC 주최자는 훈련, 검증 및 테스트 데이터로 사전 정의된 분할을 제공합니다(VOC-2012의 테스트 데이터는 공개되지 않고 대신 공식 평가 서버가 제공됨). 인식 성능은 클래스 간 평균 정밀도(mAP)를 사용하여 측정됩니다.

특히 VOC-2007 및 VOC-2012의 검증 세트에 대한 성능을 검토한 결과, 여러 스케일로 계산된 이미지 설명자를 평균하여 집계하는 것이 시뮬레이션 성능을 향상시키는 것으로 나타났습니다.

스태킹에 의한 집계와 유사합니다. 이는 VOC 데이터세트에서 객체가 다양한 척도에 걸쳐 나타나기 때문에 분류자가 활용할 수 있는 특정 척도별 의미가 없기 때문이라는 가설을 세웠습니다. 평균을 내면 설명 차원이 부풀려지지 않는다는 이점이 있기 때문에 광범위한 스케일에 걸쳐 이미지 설명자를 집계할 수 있었습니다: $Q \in \{256, 384, 512, 640, 768\}$. 하지만 $\{256, 384, 512\}$의 더 작은 범위에서의 *개선은* 0.3%로 다소 미미하다는 점에 주목할 필요가 있습니다.

테스트 세트 성능은 표 11에 다른 접근 방식과 비교하여 보고되어 있습니다. 당사의 네트워크 "Net-D"와 "Net-E"는 VOC 데이터 세트에서 동일한 성능을 보여주며, 이 두 가지를 결합하면 결과가 약간 개선됩니다. 우리의 방법은 ILSVRC 데이터 세트에 대해 사전 학습된 이미지 표현 전반에 걸쳐 새로운 첨단 기술을 설정했으며, 이전 최고 결과인 Chatfield 등(2014)의 결과를 6% 이상 능가합니다. VOC-2012에서 1% 더 나은 mAP를 달성한 Wei 등(2014)의 방법은 VOC 데이터세트에 의미적으로 가까운 1000개의 카테고리를 추가로 포함하는 확장된 2000클래스 ILSVRC 데이터세트에서 사전 학습되었다는 점에 유의해야 합니다. 또한 객체 감지 지원 분류 파이프라인과의 융합을 통해 이점을 얻을 수 있습니다.

Caltech-101 및 Caltech-256의 이미지 분류. 이 섹션에서는 이미지 분류 벤치마크인 Caltech-101(Fei-Fei 외., 2004) 및 Caltech-256(Griffin 외., 2007)에 대한 매우 심층적인 기능을 평가합니다. Caltech-101은 102개의 클래스(101개의 객체 카테고리와 배경 클래스)로 분류된 9K 이미지를 포함하며, Caltech-256은 31K 이미지와 257개의 클래스로 더 큽니다. 이러한 데이터 세트에 대한 표준 평가 프로토콜은 훈련 데이터와 테스트 데이터로 여러 개의 무작위 분할을 생성하고 평균 클래스 리콜(클래스당 테스트 이미지 수가 다른 것을 보정)로 측정되는 분할 전체의 평균 인식 성능을 보고하는 것입니다. Chatfield 외(2014), Zeiler & Fergus(2013), He 외(2014)에 따라 Caltech-101에서는 훈련 및 테스트 데이터로 3개의 무작위 분할을 생성하여 각 분할에 클래스당 30개의 훈련 이미지와 클래스당 최대 50개의 테스트 이미지가 포함되도록 했습니다. Caltech-256에서도 3개의 분할을 생성했으며, 각 분할에는 클래스당 60개의 훈련 이미지가 포함되어 있습니다(나머지는 테스트에 사용됨). 각 분할에서 훈련 이미지의 20%는 하이퍼파라미터 선택을 위한 검증 세트로 사용되었습니다.

VOC와 달리, Caltech 데이터 세트에서는 다중 스케일로 계산된 디스크립터 스태킹이 평균 또는 최대 풀링보다 성능이 더 우수하다는 것을 발견했습니다. 이는 Caltech 이미지에서 객체가 일반적으로 이미지 전체를 차지하기 때문에 다중 스케일 이미지 특징이 의미적으로 다르며(전체 *객체와* 객체 부분 캡처), 스태킹을 통해 분류기가 이러한 스케일별 표현을 활용할 수 있기 때문으로 설명할 수 있습니다. 여기서는 $Q \in \{256, 384, 512\}$의 세 가지 스케일을 사용했습니다.

표 11에서 두 모델을 서로 비교하고 최신 기술과 비교했습니다. 표에서 볼 수 있듯이, 더 깊은 19층 Net-E가 16층 Net-D보다 성능이 더 우수하며, 두 모델을 조합하면 성능이 더욱 향상됩니다. Caltech-101에서 우리의 표현은 He 등(2014)의 접근 방식과 경쟁력이 있지만, VOC-2007에서 우리의 그물보다 성능이 현저히 떨어집니다. Caltech-256에서 우리의 기능은 최신 기술(Chatfield 외., 2014)을 큰 차이(8.6%)로 능가합니다.

VOC-2012의 액션 분류. 또한 동작을 수행하는 사람의 경계 상자가 주어지면 단일 이미지에서 동작 클래스를 예측하는 PASCAL VOC-2012 동작 분류 작업(에버링햄 외., 2015)에서 가장 성능이 우수한 이미지 표현(Net-D 및 Net-E 특징의 스택)을 평가했습니다. 이 데이터 세트에는 11개의 클래스로 분류된 4.6K 훈련 이미지가 포함되어 있습니다. VOC-2012 객체 분류 작업과 유사하게 mAP를 사용하여 성능을 측정합니다. 두 가지 훈련 설정을 고려했습니다. (i) 전체 이미지에 대해 ConvNet 특징을 계산하고 제공된 경계 상자를 무시하는 것, (ii) 전체 이미지와 제공된 경계 상자에 대해 특징을 계산하고 이를 스택하여 최종 표현을 얻는 것입니다. 결과는 표 12에서 다른 접근 방식과 비교됩니다.

제공된 바운딩 박스를 사용하지 않고도 VOC 동작 분류 작업에서 최고 수준의 표현을 달성했으며, 이미지와 바운딩 박스를 모두 사용할 경우 결과가 더욱 향상됩니다. 다른 접근 방식과 달리 작업별 휴리스틱을 통합하지 않고 매우 심층적인 컨볼루션 특징의 표현력에 의존했습니다.

기타 인식 작업. 이 모델은 공개 이후 연구 커뮤니티에서 다양한 이미지 인식 작업에 활발히 사용되어 왔으며, 얕은 표현보다 지속적으로 더 나은 성능을 보였습니다. 예를 들어, Girshick 등(2014)은 Krizhevsky 등(2012)의 ConvNet을 16층 모델로 대체하여 물체 감지 결과와 같은 상태를 달성했습니다. 더 얕은 아키텍처의 Krizhevsky 등(2012)에 비해 유사한 이득을 얻었습니다.

표 12: **VOC- 2012의 단일 이미지 액션 분류에 대한 최신 기술과의 비교**. 우리의 모델은 "VGG"로 표시되어 있습니다. 로 표시된 결과는 *확장된* ILSVRC 데이터 세트(1512개 클래스)에 대해 사전 학습된 ConvNets를 사용하여 얻은 결과입니다.

방법	VOC-2012(평균 AP)
(오캅 외, 2014)	70.2*
(Gkioxari 외, 2014)	73.6
(Hoai, 2014)	76.3
VGG Net-D 및 Net-E, 이미지 전용	**79.2**
VGG Net-D 및 Net-E, 이미지 및 바운딩 박스	**84.0**

는 의미론적 세분화(Long et al., 2014), 이미지 캡션 생성(Kiros et al., 2014; Karpathy & Fei-Fei, 2014), 텍스처 및 재질 인식(Cimpoi et al., 2014; Bell et al., 2014)에 사용되었습니다.

C 논문 개정

여기에서는 독자의 이해를 돕기 위해 주요 논문 개정 목록을 제시하고 실질적인 변경 사항을 간략하게 설명합니다.

v1 초기 버전. ILSVRC 제출 전에 수행한 실험을 제시합니다.

v2에서는 스케일 지터링을 사용하여 훈련 세트 증강이 포함된 제출 후 ILSVRC 실험을 추가하여 성능을 개선합니다.

v3에서는 PASCAL VOC 및 Caltech 이미지 분류 데이터 세트에 대한 일반화 실험(부록 B)을 추가합니다. 이 실험에 사용된 모델은 공개적으로 사용 가능합니다.

v4 논문이 ICLR-2015 제출 형식으로 변환되었습니다. 또한 분류를 위해 여러 작물을 사용한 실험을 추가합니다.

v6 카메라 지원 ICLR-2015 컨퍼런스 논문. 얕은 그물망 B와 얕은 그물망의 비교와 PASCAL VOC 액션 분류 벤치마크의 결과를 추가합니다.