차별적인 현지화를 위한 심층 기능 학습

볼레이 저우, 아디트야 코슬라, 아가타 라페드리자, 오드 올리바, 안토니오 토랄바
MIT 컴퓨터 과학 및 인공 지능 연구소 {브저우,코슬라,아가타,올리바,
토랄바}@csail.mit.edu

추상적인

이 작업에서 우리는 [13]에서 제안된 전역 평균 풀링 레이어를 다시 살펴보고, 이것이 이미지 수준 레이블에 대한 훈련에도 불구하고 컨벌루션 신경망이 놀라운 위치 파악 능력을 갖도록 명시적으로 어떻게 활성화하는지 설명합니다. 이 기술은 이전에 훈련을 정규화하기 위한 수단으로 제안되었지만 실제로는 다양한 작업에 적용할 수 있는 일반적인 지역화 가능한 심층 표현을 구축한다는 사실을 발견했습니다. 전역 평균 풀링의 명백한 단순성에도 불구하고 ILSVRC 2014의 객체 위치 파악에 대해 37.1%의 상위 5개 오류를 달성할 수 있습니다. 이는 완전히 감독되는 CNN 접근 방식으로 달성한 34.2%의 상위 5개 오류에 매우 가깝습니다. 우리는 우리의 네트워크가 훈련을 받지 않았음에도 불구하고 다양한 작업에서 식별 가능한 이미지 영역을 위치화할 수 있음을 보여줍니다.

1. 소개

Zhou의 최근 작품*외*[33]은 CNN(컨볼루션 신경망)의 다양한 계층의 컨볼루셔널 유닛이 객체 위치에 대한 감독이 제공되지 않았음에도 불구하고 실제로 객체 감지기처럼 동작한다는 것을 보여주었습니다. 컨볼루셔널 레이어에서 객체를 위치화하는 놀라운 능력이 있음에도 불구하고 완전 연결 레이어를 분류에 사용하면 이 능력이 손실됩니다. 최근 NIN(Network in Network) [13] 및 GoogLeNet [24]과 같은 일부 인기 있는 완전 컨벌루션 신경망은 고성능을 유지하면서 매개변수 수를 최소화하기 위해 완전 연결 레이어 사용을 피하도록 제안되었습니다.

이를 달성하기 위해 [13]은 다음을 사용합니다.*글로벌 평균 풀링*이는 구조적 정규화 역할을 하여 훈련 중 과적합을 방지합니다. 실험에서 우리는 이 전역 평균 풀링 레이어의 장점이 단순히 정규화 역할을 하는 것 이상으로 확장된다는 것을 발견했습니다. 실제로 약간의 조정을 통해 네트워크는 최종 레이어까지 놀라운 위치 파악 능력을 유지할 수 있습니다. 이러한 조정을 통해 네트워크가 원래 훈련되지 않은 작업을 포함하여 다양한 작업에 대한 단일 전달 패스에서 식별 가능한 이미지 영역을 쉽게 식별할 수 있습니다. 그림 1(a)에 도시된 바와 같이,

그림 1. 클래스 활성화 매핑(CAM) 기술과 결합된 전역 평균 풀링 레이어의 간단한 수정을 통해 분류 훈련을 받은 CNN은 단일 순방향 패스(예: 칫솔)에서 이미지를 분류하고 클래스별 이미지 영역을 지역화할 수 있습니다. ~을 위한*양치질*그리고 전기톱은*나무를 자르다*.

객체 분류에 대해 훈련된 CNN은 행동 분류를 위한 식별 영역을 인간 자체가 아닌 인간이 상호 작용하는 객체로 성공적으로 위치화할 수 있습니다.

접근 방식의 명백한 단순성에도 불구하고 ILSVRC 벤치마크[20]의 약하게 감독된 객체 위치 파악에 대해 우리의 최고의 네트워크는 37.1%의 상위 5개 테스트 오류를 달성했습니다. 이는 완전 감독된 개체 위치 파악의 34.2%에 다소 가깝습니다. 알렉스넷[10]. 또한 우리는 접근 방식의 심층 기능의 현지화 가능성을 일반 분류, 현지화 및 개념 발견을 위해 다른 인식 데이터 세트로 쉽게 전송할 수 있음을 보여줍니다.[1].

1.1. 관련된 일

CNN(Convolutional Neural Networks)은 다양한 시각적 인식 작업에서 인상적인 성능을 발휘했습니다[10, 34, 8]. 최근 연구에 따르면 CNN은 이미지 수준 레이블에 대한 교육을 받았음에도 불구하고 객체를 위치화하는 놀라운 능력을 가지고 있는 것으로 나타났습니다[1, 16, 2, 15]. 이 작업에서 우리는 올바른 아키텍처를 사용하면 객체를 지역화하는 것 이상으로 이 기능을 일반화하여 이미지의 어떤 영역이 사용되는지 정확하게 식별할 수 있음을 보여줍니다.

[1]우리 모델은 http://cnnlocalization.csail.mit.edu에서 확인할 수 있습니다.

차별. 여기서는 이 문서와 가장 관련이 있는 두 가지 작업 라인인 약한 지도 객체 위치 파악과 CNN의 내부 표현 시각화에 대해 논의합니다.

약한 감독 객체 현지화:CNN을 사용하여 약한 감독 객체 위치 파악을 탐색하는 최근 작업이 많이 있었습니다[1, 16, 2, 15]. 베르가모*외*[1]은 객체 위치를 파악하기 위해 최대 활성화를 일으키는 영역을 식별하기 위해 이미지 영역을 마스킹하는 것과 관련된 독학 객체 위치 파악 기술을 제안합니다. 신비스*외*[2] 다중 인스턴스 학습과 CNN 기능을 결합하여 객체 위치를 파악합니다. 오크압*외*[15]는 중간 수준 이미지 표현을 전송하는 방법을 제안하고 여러 개의 중첩 패치에서 CNN의 출력을 평가하여 일부 개체 위치 파악이 달성될 수 있음을 보여줍니다. 그러나 저자는 실제로 현지화 능력을 평가하지 않습니다. 반면, 이러한 접근 방식은 유망한 결과를 산출하지만 엔드투엔드(end-to-end) 학습이 이루어지지 않고 객체 위치를 파악하기 위해 네트워크의 여러 전달 전달이 필요하므로 실제 데이터 세트로 확장하기가 어렵습니다. 우리의 접근 방식은 엔드 투 엔드로 훈련되었으며 단일 전달 패스로 객체의 위치를 파악할 수 있습니다.

우리와 가장 유사한 접근 방식은 Oquab의 전역 최대 풀링을 기반으로 한 작업입니다.*외*[16]. 글로벌 대신 *평균*풀링하면 전역적으로 적용됩니다.*최대*물체의 한 지점을 지역화하기 위한 풀링. 그러나 그 위치는 물체의 전체 범위를 결정하기보다는 물체의 경계에 있는 지점으로 제한됩니다. 우리는 그 동안*최대* 그리고*평균*기능이 다소 유사하기 때문에 평균 풀링을 사용하면 네트워크가 개체의 전체 범위를 식별하도록 장려됩니다. 이에 대한 기본 직관은 네트워크가 식별할 때 평균 풀링 이점의 손실이 있다는 것입니다.*모두*최대 풀링과 비교하여 객체의 식별 영역. 이에 대해서는 Sec.에서 더 자세히 설명하고 실험적으로 검증했습니다. 3.2. 또한 [16]과 달리 이 위치 파악 기능은 일반적이며 네트워크가 훈련되지 않은 문제에 대해서도 관찰할 수 있음을 보여줍니다.

우리는 사용*클래스 활성화 맵*섹션 2에 설명된 대로 각 이미지에 대해 생성된 가중 활성화 맵을 참조합니다. 전역 평균 풀링은 여기서 제안하는 새로운 기술은 아니지만 정확한 식별 위치화에 적용될 수 있다는 관찰은 다음과 같다는 점을 강조하고 싶습니다. , 우리가 아는 한, 우리 업무에 고유한 것입니다. 우리는 이 기술의 단순성으로 인해 휴대성이 뛰어나고 빠르고 정확한 위치 파악을 위해 다양한 컴퓨터 비전 작업에 적용될 수 있다고 믿습니다.

CNN 시각화:CNN의 속성을 더 잘 이해하기 위해 CNN이 학습한 내부 표현을 시각화하는 최근 연구[29, 14, 4, 33]가 많이 있습니다. 자일러*외*[29] 각 유닛을 활성화하는 패턴을 시각화하기 위해 역합성곱 네트워크를 사용합니다. 저우*외.*[33] CNN은 장면을 인식하도록 훈련하는 동안 객체 감지기를 학습하고 동일한 결과를 보여줍니다.

네트워크는 단일 순방향 패스로 장면 인식과 객체 위치 파악을 모두 수행할 수 있습니다. 이 두 작품 모두 컨벌루션 레이어만 분석하고 완전히 연결된 레이어는 무시하여 전체 스토리의 불완전한 그림을 그립니다. 완전히 연결된 레이어를 제거하고 대부분의 성능을 유지함으로써 네트워크를 처음부터 끝까지 이해할 수 있습니다.

마헨드란*외*[14] 그리고 도소비츠키*외*[4] 다양한 레이어의 심층 기능을 반전시켜 CNN의 시각적 인코딩을 분석합니다. 이러한 접근 방식은 완전 연결 레이어를 반전시킬 수 있지만 이 정보의 상대적 중요성을 강조하지 않고 심층 기능에 어떤 정보가 보존되어 있는지만 보여줍니다. [14] 및 [4]와 달리 우리의 접근 방식은 이미지의 어느 영역이 식별에 중요한지 정확하게 강조할 수 있습니다. 전반적으로 우리의 접근 방식은 CNN의 영혼을 다시 한 번 엿볼 수 있는 기회를 제공합니다.

2. 클래스 활성화 매핑

이 섹션에서는 생성 절차를 설명합니다. *클래스 활성화 맵*(CAM)은 CNN에서 GAP(Global Average Pooling)를 사용합니다. 특정 카테고리에 대한 클래스 활성화 맵은 해당 카테고리를 식별하기 위해 CNN이 사용하는 식별 이미지 영역을 나타냅니다(예: 그림 3). 이러한 맵을 생성하는 절차는 그림 2에 설명되어 있습니다.

우리는 Network in Network [13] 및 GoogLeNet [24]와 유사한 네트워크 아키텍처를 사용합니다. 네트워크는 주로 컨볼루셔널 레이어로 구성되며 최종 출력 레이어(분류의 경우 소프트맥스) 직전에 전역 평균 풀링을 수행합니다. 컨벌루션 기능 맵을 생성하고 이를 원하는 출력(범주형 또는 기타)을 생성하는 완전 연결 레이어의 기능으로 사용합니다. 이 간단한 연결 구조가 주어지면 출력 레이어의 가중치를 클래스 활성화 매핑이라고 부르는 기술인 컨벌루션 기능 맵에 다시 투영하여 이미지 영역의 중요성을 식별할 수 있습니다.

그림 2에서 볼 수 있듯이 전역 평균 풀링은 마지막 컨볼루셔널 레이어에서 각 유닛의 특징 맵의 공간 평균을 출력합니다. 이러한 값의 가중치 합계가 최종 출력을 생성하는 데 사용됩니다. 마찬가지로, 클래스 활성화 맵을 얻기 위해 마지막 컨볼루셔널 레이어의 특징 맵에 대한 가중치 합을 계산합니다. 우리는 이것을 소프트맥스의 경우에 대해 아래에서 더 공식적으로 설명합니다. 회귀 및 기타 손실에도 동일한 기술을 적용할 수 있습니다.

주어진 이미지에 대해*에프케이*(*엑스, 와이*)유닛의 활성화를 나타냅니다.*케이*공간 위치의 마지막 컨벌루션 레이어에서 (*엑스, 와이*). 그러면 단위에 대해서는*케이*Σ , 글로벌 수행의 결과 평균 풀링,*에프케이*~이다 $_{x,y}$에프케이(엑스, 와이),Σ따라서 주어진 수업*씨*, 소프트맥스에 대한 입력,*에스씨*, 이다 케이승베이에프케이어디승씨 *케이* 클래스에 해당하는 가중치입니다.*씨*단위용*케이*. 본질적으로, *승쌔이*나타냅니다*중요성*~의*에프케이*반을 위해서*씨*. 마지막으로 클래스에 대한 소프트맥스 출력*씨*,*피씨*에 의해 주어진다Σ특급(*에스씨*)/Σ특급(에스씨). 여기서는 편향항을 무시합니다. 입력을 명시적으로 설정합니다.

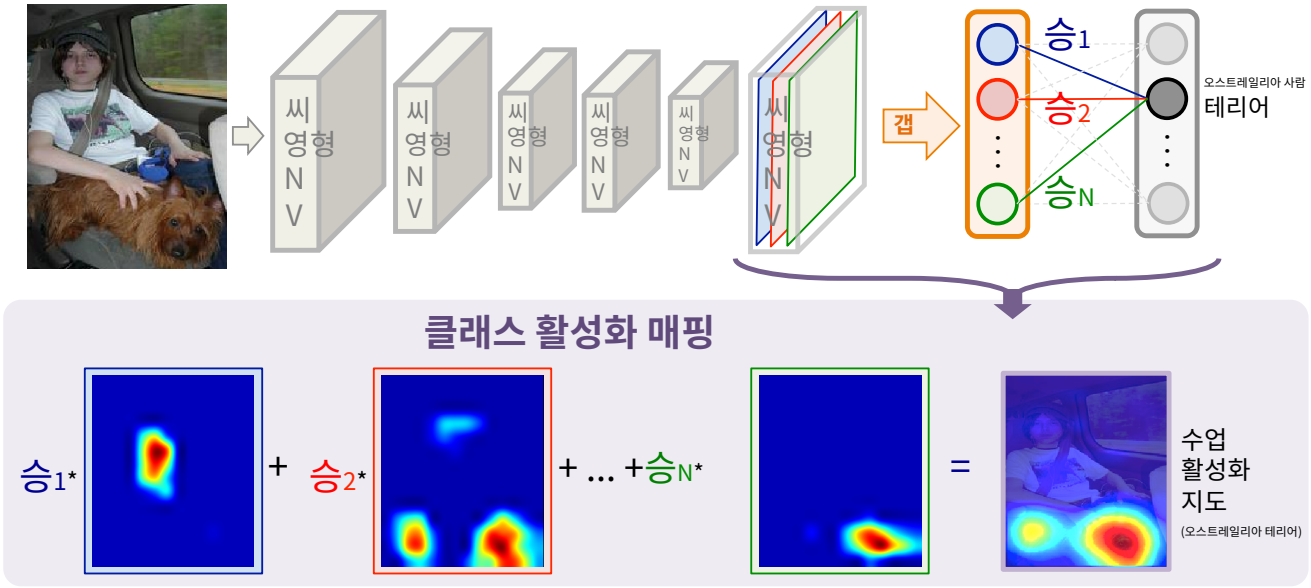

그림 2. 클래스 활성화 매핑: 예측된 클래스 점수는 클래스 활성화 맵(CAM)을 생성하기 위해 이전 컨볼루셔널 레이어에 다시 매핑됩니다. CAM은 클래스별 차별 영역을 강조합니다.

소프트맥스의 편향0분류 성능에 거의 영향을 미치지 않기 때문입니다. Σ CE.

연결하여 *에프케이= 에스씨*, *x,y에프케이(엑스, 와이)*수업 점수에, 우리는 얻습니다

$$\text{에스씨} = \sum_{\text{케이}} \text{승케이} \sum_{x,y} \text{에프케이(엑스, 와이)}$$

$$= \sum_{x,y} \sum_{k} \text{승씨} \text{에프케이(엑스, 와이).} \quad (1)$$

우리는 정의합니다*중씨*클래스의 클래스 활성화 맵으로*씨*, 어디 각 공간 요소는 다음과 같이 주어진다.

$$\text{중씨(엑스, 와이)} = \sum_{\text{케이}} \text{승씨} \text{에프케이(엑스, 와이).} \quad (2)$$

따라서,*에스씨= x,y중씨(엑스, 와이)*, 따라서*중씨(엑스, 와이)*곧장 공간 격자에서의 활성화의 중요성을 나타냅니다. (*엑스, 와이*) 이미지를 클래스로 분류하는 방법*씨*.

직관적으로 이전 연구[33, 29]를 기반으로 각 단위가 수용 필드 내의 일부 시각적 패턴에 의해 활성화될 것으로 예상합니다. 따라서*에프케이*는 이 시각적 패턴의 존재에 대한 지도입니다. 클라 SS AC자극 엄마 pl네 가중 린트들 p개의 공간 위치 다시일번째 ese시각적 ~에따르다NT 의 귀 합계 . 비 y 단순히 위 플리ng c 라활성 ㅏ- s개의 이미지 영 화하 놓다영상, 승 이데 그 역 m에 대한 지도ost r중요한 ~에그 특별한 고양이 오리.

그림 3에서 우리는 쇼 w CAM 출력의 몇 가지 예 위의 접근 방식을 사용합니다. 우리는 차별이 있음을 알 수 있습니다. 다양한 클래스에 대한 이미지의 기본 영역이 강조 표시됩니다. 그림 4에서는 CAM의 차이점을 강조합니다. 다른 클래스를 사용할 때 단일 이미지의 경우*씨*일반- 지도를 먹었다. 우리는 관찰한다 그 차별은 원주민 지역

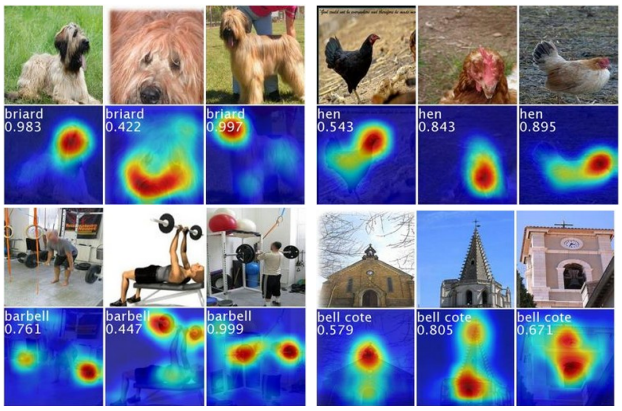

그림 3. ILSVRC의 4개 클래스 CAM [20]. 지도는 이미지 분류에 사용되는 식별 이미지 영역(예: 동물의 머리)을 강조 표시합니다.*브리아드* 그리고*암탉*, 접시 *바벨*, 그리고 벨이*벨 코트*.

특정 이미지에 대해서도 카테고리마다 다릅니다. 이는 우리의 접근 방식이 예상대로 작동함을 의미합니다.

우리앞으로의 섹션에서 이를 정량적으로 설명합니다.

G전역 평균 풀링(GAP)과 전역 최대 풀-GMP):감독된 개체 ing (위치 파악을 위해 GMP를 사용하는 것에 대한 이전 작업 와 [16]을 고려할 때 GAP GMP 간의 직관적인 차이점을 강조 포트하는 것이 중요하다고 생각합니다. 우리는 GAP 손실이 네 그리뵤워크가 개체의 범위를 식별하도록 장려한다고 믿습니 다.

단 하나의 차별적인 부분을 식별하도록 권장하는 GMP입니다. 왜냐하면 지도의 평균을 낼 때 다음을 구함으로써 그 값을 극대화할 수 있기 때문이다.*모두*차별적인 물체의 일부 모두 그렇듯이 w 활성화 빨간색 의 출력을 uce

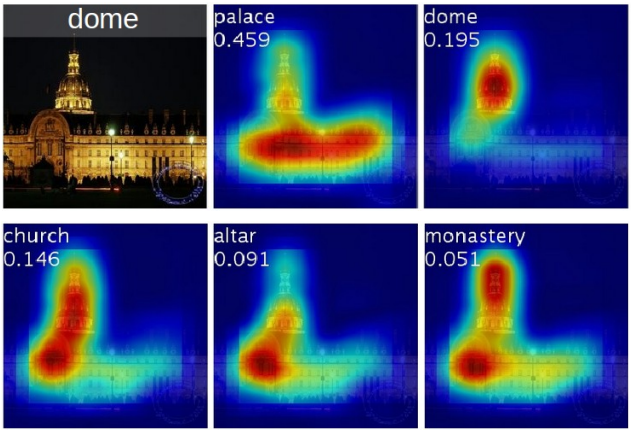

그림 4. 실측을 돔으로 사용하여 주어진 이미지에 대해 상위 5개 예측 범주에서 생성된 CAM의 예. 예측된 클래스와 해당 점수는 각 클래스 활성화 맵 위에 표시됩니다. 강조 표시된 영역이 예측 클래스에 따라 달라지는 것을 관찰합니다.*둥근 천장*위쪽 둥근 부분을 활성화하는 동시에 *궁전*화합물의 아래쪽 평평한 부분을 활성화합니다.

특별한 지도. 반면 GMP의 경우 가장 차별적인 영역을 제외한 모든 이미지 영역에 대한 낮은 점수는 최대값을 수행하므로 점수에 영향을 주지 않습니다. 우리는 Sec.의 ILSVRC 데이터세트에서 이를 실험적으로 검증합니다. 3: GMP는 GAP와 유사한 분류 성능을 달성하는 반면 GAP는 현지화 측면에서 GMP보다 성능이 뛰어납니다.

3. 약하게 감독되는 객체 위치 파악

이 섹션에서는 ILSVRC 2014 벤치마크 데이터세트[20]에 대해 훈련할 때 CAM의 위치 파악 능력을 평가합니다. 먼저 Sec.에서 사용된 실험 설정과 다양한 CNN을 설명합니다. 3.1. 그런 다음 Sec. 3.2 우리는 약한 감독 객체 위치 파악에 대한 자세한 결과를 지역화하고 제공하는 방법을 학습할 때 우리 기술이 분류 성능에 부정적인 영향을 미치지 않는다는 것을 확인합니다.

3.1. 설정

실험을 위해 AlexNet [10], VG-Gnet [23] 및 GoogLeNet [24]과 같은 인기 있는 CNN에서 CAM을 사용하는 효과를 평가합니다. 일반적으로 이러한 각 네트워크에 대해 최종 출력 전에 완전 연결 레이어를 제거하고 이를 GAP와 완전 연결 소프트맥스 레이어로 대체합니다.

우리는 GAP 이전의 마지막 합성곱 계층이 더 높은 공간 해상도를 가질 때 네트워크의 위치 파악 능력이 향상된다는 것을 발견했습니다.*매핑 해상도*. 이를 위해 일부 네트워크에서 여러 컨볼루셔널 레이어를 제거했습니다. 구체적으로 다음과 같이 수정했습니다. AlexNet의 경우 다음 레이어를 제거했습니다. 전환5(즉,풀5에게문제)결과적으로 매핑 해상도는 다음과 같습니다.13 ×13.VGGnet의 경우 다음 레이어를 제거했습니다.전환5-3(즉,풀5에게문제),결과-

매핑 해상도는 다음과 같습니다.14 ×14.GoogLeNet의 경우 다음 레이어를 제거했습니다.인셉션4e(즉,풀4 에게문제),결과적으로 매핑 해상도는 다음과 같습니다.14 ×14. 위의 각 네트워크에 크기의 컨벌루션 레이어를 추가했습니다.삼 ×삼,보폭 1,인주11024개의 단위가 있고 그 뒤에 GAP 레이어와 소프트맥스 레이어가 있습니다. 그런 다음 각 네트워크를 미세 조정했습니다.[2]1000방향 객체 분류를 위한 ILSVRC [20]의 130만 훈련 이미지를 사용하여 최종 네트워크 AlexNet-GAP, VGGnet-GAP 및 GoogLeNet-GAP를 각각 생성했습니다.

분류를 위해 우리의 접근 방식을 원래 AlexNet [10], VGGnet [23] 및 GoogLeNet [24]과 비교하고 NIN(Network in Network) [13]에 대한 결과도 제공합니다. 현지화를 위해 원본 GoogLeNet과 비교합니다.[삼], NIN 및 CAM 대신 역전파 [22] 사용. 또한 평균 풀링과 최대 풀링을 비교하기 위해 전역 최대 풀링(GoogLeNet-GMP)을 사용하여 훈련된 GoogLeNet에 대한 결과도 제공합니다.

분류 및 현지화 모두에 대해 ILSVRC와 동일한 오류 측정항목(상위 1, 상위 5)을 사용하여 네트워크를 평가합니다. 분류의 경우 ILSVRC 검증 세트를 평가하고 현지화의 경우 검증 및 테스트 세트를 모두 평가합니다.

3.2. 결과

우리는 우리의 접근 방식이 분류 성능을 크게 저하시키지 않는다는 것을 보여주기 위해 먼저 객체 분류에 대한 결과를 보고합니다. 그런 다음 우리의 접근 방식이 약한 감독 객체 위치 파악에 효과적이라는 것을 보여줍니다.

분류:Tbl. 1은 원본 네트워크와 GAP 네트워크의 분류 성능을 요약합니다. 우리는 대부분의 경우 약간의 성능 저하가 있음을 발견했습니다.1-2%다양한 네트워크에서 추가 레이어를 제거할 때. 우리는 AlexNet이 완전 연결 레이어 제거로 인해 가장 큰 영향을 받는 것을 관찰했습니다. 이를 보완하기 위해 GAP 바로 앞에 두 개의 컨벌루션 레이어를 추가하여 AlexNet*-GAP 네트워크를 생성합니다. 우리는 AlexNet*-GAP가 AlexNet과 비슷한 성능을 보인다는 것을 발견했습니다. 따라서 전반적으로 GAP 네트워크에 대한 분류 성능이 크게 유지된다는 것을 알 수 있습니다. 또한 GoogLeNet-GAP와 GoogLeNet-GMP는 예상대로 분류 성능이 유사하다는 것을 확인했습니다. 위치 파악에서 높은 성능을 얻으려면 네트워크가 분류에서 잘 수행되는 것이 중요합니다. 이는 객체 카테고리와 경계 상자 위치를 모두 정확하게 식별하는 것과 관련되기 때문입니다.

현지화:지역화를 수행하려면 경계 상자와 관련 개체 범주를 생성해야 합니다. CAM에서 경계 상자를 생성하기 위해 간단한 임계값 기술을 사용하여 히트맵을 분할합니다. 먼저 값이 20%를 초과하는 영역을 분할합니다.

[2]처음부터 훈련도 비슷한 성과를 거두었습니다.
[삼]이는 GoogLeNet-GAP보다 매핑 해상도가 낮습니다.

탭le 1. ILSVRC 검증의 분류 오류 세트.

네트워크	상위 1개 값 오류	상위 5위 가치. 오류
VGGnet-GAP	33.4	12.2
GoogLeNet-GAP	35.0	13.2
알렉스넷*-갭	44.9	20.9
AlexNet-GAP	51.1	26.3
구글르넷	31.9	11.3
VGGnet	31.2	11.4
알렉스넷	42.6	19.5
닌	41.9	19.6
GoogLeNet-GMP	35.6	13.9

CAM의 최대값. 그런 다음 분할 맵에서 가장 큰 연결된 구성 요소를 덮는 경계 상자를 사용합니다. 상위 5개 지역화 평가 지표에 대한 상위 5개 예측 클래스 각각에 대해 이 작업을 수행합니다. 그림 6(a)는 이 기술을 사용하여 생성된 경계 상자의 몇 가지 예를 보여줍니다. ILSVRC 검증 세트의 현지화 성능은 Tbl에 표시됩니다. 2, 그림 5의 출력 예.

우리는 GAP 네트워크가 GoogLeNet-GAP를 사용하여 모든 기본 접근 방식을 능가하여 가장 낮은 위치 오류를 달성한다는 것을 관찰했습니다.43%상위 5위. 이 네트워크가 주석이 달린 단일 경계 상자에 대해 훈련되지 않았다는 점을 고려하면 이는 놀라운 일입니다. 우리는 우리의 CAM 접근법이 [22]의 역전파 접근법보다 훨씬 뛰어난 성능을 보인다는 것을 관찰했습니다(출력 비교를 위해 그림 6(b) 참조). 또한 우리는 GoogLeNet-GAP가 분류를 위해 역전되었음에도 불구하고 현지화 측면에서 GoogLeNet보다 훨씬 뛰어난 성능을 보인다는 것을 관찰했습니다. 우리는 GoogLeNet의 낮은 매핑 해상도(7 ×7)정확한 위치 파악을 방해합니다. 마지막으로, 우리는 GoogLeNet-GAP가 객체의 범위를 식별하기 위해 최대 풀링보다 평균 풀링의 중요성을 보여주는 합리적인 마진으로 GoogLeNet-GMP보다 성능이 우수하다는 것을 관찰했습니다.

우리의 접근 방식을 기존의 약한 감독 [22] 및 완전 감독 [24, 21, 24] CNN 방법과 추가로 비교하기 위해 ILSVRC 테스트 세트에서 GoogLeNet-GAP의 성능을 평가합니다. 여기서는 약간 다른 경계 상자 선택 전략을 따릅니다. 상위 1차 및 2차 예측 클래스의 클래스 활성화 맵에서 두 개의 경계 상자(하나는 촘촘하고 하나는 느슨함)를 선택하고 상위 3차 예측 클래스에서 하나의 느슨한 경계 상자를 선택합니다. 우리는 이 경험적 방법이 검증 세트의 성능을 향상시키는 데 도움이 된다는 것을 발견했습니다. 성능은 Tbl에 요약되어 있습니다. 3. 휴리스틱을 사용하는 GoogLeNet-GAP는 약한 감독 설정에서 상위 5개 오류율 37.1%를 달성합니다. 이는 완전 감독 설정에서 AlexNet의 상위 5개 오류율(34.2%)에 놀라울 정도로 가깝습니다. 인상적이기는 하지만, 지역화를 위해 동일한 아키텍처(즉, 약한 감독 GoogLeNet-GAP 대 완전 감독 GoogLeNet)를 사용하는 완전 감독 네트워크를 비교할 때 아직 갈 길이 멀습니다.

표 2. ILSVRC 검증 세트의 현지화 오류. *역전파*CAM 대신 위치 파악을 위해 [22]를 사용하는 것을 의미합니다.

방법	상위 1개 Val.error	상위 5위 가치. 오류
GoogLeNet-GAP	56.40	43.00
VGGnet-GAP	57.20	45.14
구글르넷	60.09	49.34
알렉스넷*-갭	63.75	49.53
AlexNet-GAP	67.19	52.16
닌	65.47	54.19
GoogLeNet의 역전파	61.31	50.55
VGGnet의 역전파	61.12	51.46
AlexNet의 역전파	65.17	52.64
GoogLeNet-GMP	57.78	45.26

표 3. 다양한 약한 감독 및 완전 감독 방법에 대한 ILSVRC 테스트 세트의 위치 파악 오류.

방법	감독	상위 5개 테스트 오류
GoogLeNet-GAP(휴리스틱)	약하게	37.1
GoogLeNet-GAP	약하게	42.9
역전파 [22]	약하게	46.4
구글리넷 [24]	가득한	26.7
오버핏 [21]	가득한	29.9
알렉스넷 [24]	가득한	34.2

4. 일반 지역화를 위한 심층적인 기능

CNN의 상위 계층의 응답(예: FC6, FC7AlexNet의)은 다양한 이미지 데이터세트에서 최첨단 성능을 갖춘 매우 효과적인 일반 기능인 것으로 나타났습니다[3, 19, 34]. 여기에서는 GAP CNN이 학습한 기능이 일반 기능만큼 잘 수행되고 보너스로 특정 작업에 대해 훈련을 받지 않았음에도 불구하고 분류에 사용되는 식별 이미지 영역을 식별한다는 것을 보여줍니다. 원래 소프트맥스 레이어와 유사한 가중치를 얻으려면 GAP 레이어의 출력에 대해 선형 SVM [5]을 훈련하면 됩니다.

먼저, SUN397 [27], MIT Indoor67 [18], Scene15 [11], SUN Attribute [17], Caltech101 [6], Caltech256 [9], Stanford Action40 [28], UIUC Event8 [12]. 실험 설정은 [34]와 동일하다. Tbl에서. 5에서는 최고의 네트워크인 GoogLeNet-GAP의 기능 성능을 다음과 비교합니다.FC7 AlexNet의 기능 및애비뉴 풀GoogLeNet에서.

예상대로 GoogLeNet-GAP와 GoogLeNet은 AlexNet보다 훨씬 뛰어난 성능을 보입니다. 또한 우리는 GoogLeNet-GAP와 GoogLeNet이 전자의 컨볼루션 레이어 수가 적음에도 불구하고 유사한 성능을 보이는 것을 관찰했습니다. 전반적으로 우리는 GoogLeNet-GAP 기능이 일반적인 시각적 기능으로서 최첨단 기능과 경쟁적이라는 것을 발견했습니다.

더 중요한 것은 GoogLeNet-GAP와 함께 CAM 기술을 사용하여 생성된 위치 파악 맵이 이 시나리오에서도 유익한지 여부를 조사하고 싶습니다. 그림 8은 다양한 데이터 세트에 대한 몇 가지 예시 맵을 보여줍니다. 우리는 가장 차별적인 영역이 모든 데이터 세트에서 강조 표시되는 경향이 있음을 관찰합니다. 전반적으로 우리의 접근 방식은 효과적입니다.

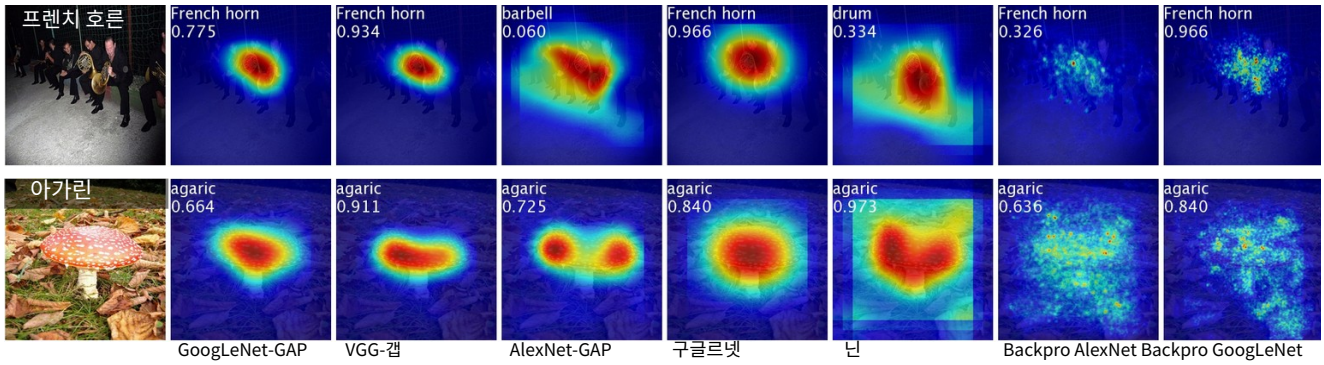

그림 5. CNN-GAP의 클래스 활성화 맵과 역전파 방법의 클래스별 돌출 맵.

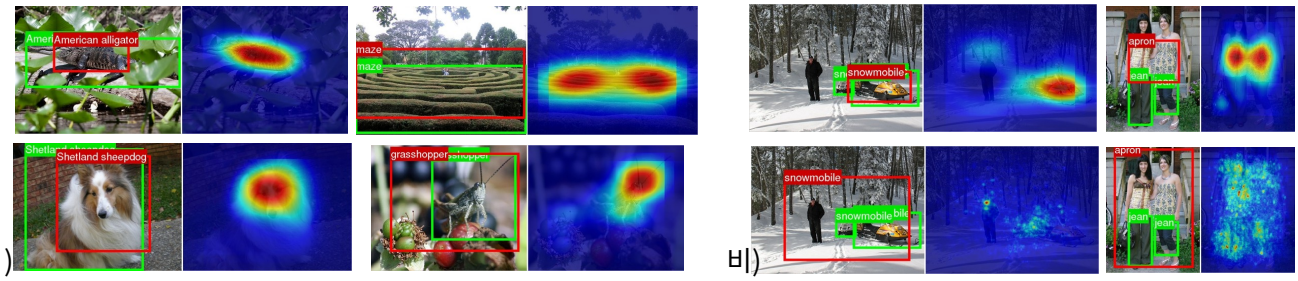

그림 6. a) GoogleNet-GAP의 현지화 예. b) GooleNet-GAP의 지역화(상위 2개)와 AlexNet을 사용한 역전파(하위 2개) 비교. 실측 상자는 녹색으로 표시되고 클래스 활성화 맵에서 예측된 경계 상자는 빨간색으로 표시됩니다.

일반적인 작업을 위해 지역화 가능한 심층 기능을 생성합니다.

초에서. 4.1에서는 새에 대한 세밀한 인식을 탐구하고 일반적인 위치 파악 능력을 평가하고 이를 사용하여 성능을 더욱 향상시키는 방법을 보여줍니다. 초에서. 4.2에서는 GoogLeNet-GAP를 사용하여 이미지에서 일반적인 시각적 패턴을 식별하는 방법을 보여줍니다.

4.1. 세밀한 인식

이 섹션에서는 CUB-200-2011[26] 데이터 세트에서 200종의 조류 종을 식별하는 데 일반적인 지역화 가능한 심층 기능을 적용합니다. 데이터세트에는 훈련용 이미지 5,994개, 테스트용 이미지 5,794개 등 11,788개의 이미지가 포함되어 있습니다. 이 데이터 세트에는 현지화 능력을 평가할 수 있는 경계 상자 주석도 포함되어 있으므로 선택합니다. Tbl. 4는 결과를 요약한다.

GoogLeNet-GAP는 기차와 테스트 모두에 대해 경계 상자 주석 없이 전체 이미지를 사용할 때 63.0%의 정확도를 달성하여 기존 접근 방식과 비슷한 성능을 발휘한다는 것을 확인했습니다. 경계 상자 주석을 사용하면 이 정확도가 70.5%로 증가합니다. 이제 네트워크의 현지화 기능을 고려하여 Sec와 유사한 접근 방식을 사용할 수 있습니다. 3.2(즉, 임계값 지정)를 사용하여 기차와 테스트 세트 모두에서 새 경계 상자를 먼저 식별합니다. 그런 다음 훈련 및 테스트를 위해 GoogLeNet-GAP를 사용하여 경계 상자 내부의 작물에서 기능을 다시 추출합니다. 이를 통해 성능이 67.8%로 상당히 향상되었음을 알 수 있습니다.

표 4. CUB200 데이터세트의 세분화된 분류 성능 GoogLeNet-GAP는 중요한 이미지 자르기를 성공적으로 현지화하여 분류 성능을 향상시킬 수 있습니다.

행동 양식	훈련/테스트 Anno.	정확성
전체 이미지의 GoogLeNet-GAP 자르기의 GoogLeNet-GAP	해당사항 없음 해당사항 없음	63.0% 67.8%
BBox의 GoogLeNet-GAP	비박스	70.5%
정렬 [7]	해당사항 없음	53.6%
정렬 [7]	비박스	67.0%
DPD [31]	B박스+부품	51.0%
디CAF+DPD [3]	B박스+부품	65.0%
팬더R-CNN [30]	B박스+부품	76.4%

이러한 현지화 능력은 범주 간의 구별이 미묘하고 더 집중된 이미지 자르기를 사용하면 더 나은 식별이 가능하므로 세밀한 인식에 특히 중요합니다.

또한 GoogLeNet-GAP는 5.5%의 우연 성능과 비교하여 0.5 IoU(Intersection Over Union) 기준에 따라 이미지의 41.0%에서 새의 위치를 정확하게 파악할 수 있음을 발견했습니다. 우리는 그림 7에 몇 가지 예를 시각화했습니다. 이는 우리 접근 방식의 위치 파악 능력을 더욱 검증합니다.

4.2. 패턴 발견

이 섹션에서는 우리 기술이 이미지를 넘어 이미지의 공통 요소나 패턴을 식별할 수 있는지 여부를 살펴봅니다.

표 5. 다양한 심층 기능에 대한 대표적인 장면 및 객체 데이터세트의 분류 정확도.

	SUN397	MIT 실내67	장면15	태양 속성	칼텍101	칼텍256	액션40	이벤트8
FC7AlexNet에서	42.61	56.79	84.23	84.23	87.22	67.23	54.92	94.42
애비뉴 풀GoogLeNet에서 갭	51.68	66.63	88.02	92.85	92.05	78.99	72.03	95.42
GoogLeNet-GAP에서	51.31	66.61	88.30	92.21	91.98	78.07	70.62	95.00

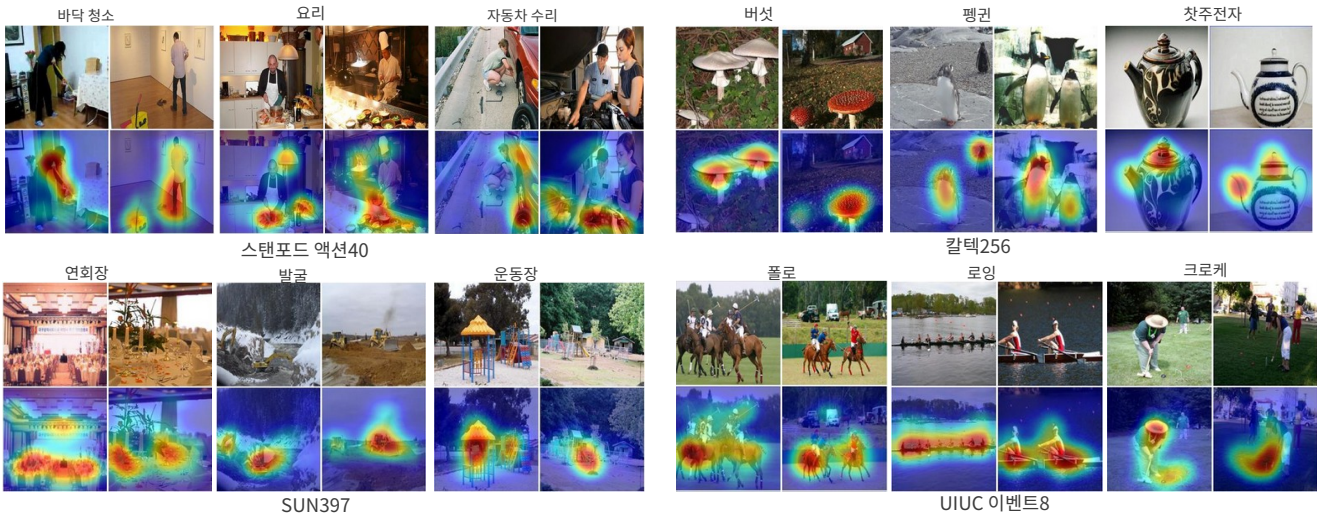

그림 8. GoogLeNet-GAP 심층 기능(객체를 인식하도록 훈련됨)을 사용한 일반적인 식별적 위치 파악. 4개 데이터 세트에 대한 3개 클래스의 각각 2개의 이미지와 그 아래에 해당 클래스 활성화 맵을 표시합니다. 우리는 이미지의 식별 가능한 영역이 종종 강조 표시되는 것을 관찰합니다. 예를 들어 Stanford Action40에서는 걸레가*바닥 청소*, 동안*요리*팬과 그릇은 현지화되어 있으며 다른 데이터세트에서도 유사한 관찰이 이루어질 수 있습니다. 이는 우리의 심층 기능의 일반적인 현지화 능력을 보여줍니다.

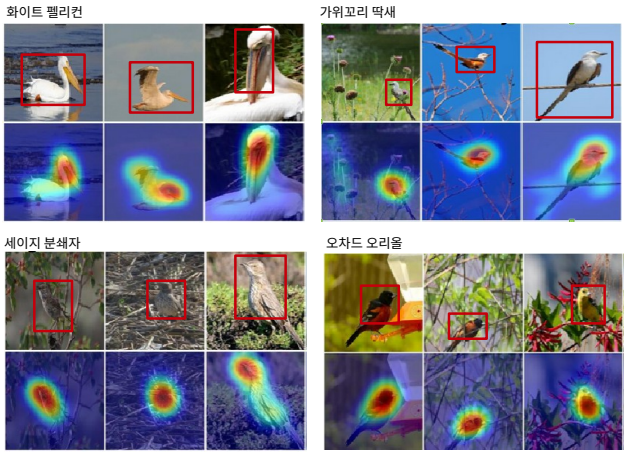

그림 7. CUB200의 4가지 새 카테고리에서 선택한 이미지에 대한 CAM 및 추론된 경계 상자(빨간색). 초에서. 4.1 경계 상자의 품질을 정량적으로 평가합니다(0.5 IoU에 대해 정확도 41.0%). 우리는 이러한 CAM 경계 상자에서 GoogLeNet-GAP 기능을 추출하고 SVM을 다시 훈련하면 새 분류 정확도가 약 5% 향상된다는 것을 발견했습니다(표 4).

텍스트 또는 상위 수준 개념과 같은 개체. 공통 개념이 포함된 이미지 세트가 주어지면 네트워크가 중요하다고 인식하는 영역이 무엇인지, 그리고 이것이 입력 패턴과 일치하는지 식별하고 싶습니다. 우리는-

이전과 유사한 접근 방식으로 GoogLeNet-GAP 네트워크의 GAP 계층에서 선형 SVM을 교육하고 CAM 기술을 적용하여 중요한 영역을 식별합니다. 우리는 심층 기능을 사용하여 세 가지 패턴 발견 실험을 수행했습니다. 결과는 아래에 요약되어 있습니다. 이 경우 학습 및 테스트 분할이 없습니다.-우리는 시각적 패턴 발견을 위해 CNN을 사용합니다.

장면에서 유익한 객체 발견:우리는 SUN 데이터세트[27]에서 최소한 다음을 포함하는 10개의 장면 카테고리를 선택합니다.200완전히 주석이 달린 이미지로 총 4675개의 완전히 주석이 달린 이미지가 생성됩니다. 각 장면 카테고리에 대해 일대다 선형 SVM을 훈련하고 선형 SVM의 가중치를 사용하여 CAM을 계산합니다. 그림 9에서는 예측된 장면 카테고리에 대한 CAM을 플롯하고 두 장면 카테고리에 대해 높은 CAM 활성화 영역과 가장 자주 겹치는 상위 6개 객체를 나열합니다. 우리는 높은 활성화 영역이 특정 장면 범주를 나타내는 개체에 자주 해당하는 것을 관찰합니다.

약한 라벨이 붙은 이미지의 개념 현지화:[32]의 하드 네거티브 마이닝 알고리즘을 사용하여 개념 탐지기를 학습하고 CAM 기술을 적용하여 이미지의 개념을 위치화합니다. 짧은 문구에 대한 개념 탐지기를 훈련하기 위해 포지티브 세트는 텍스트 캡션에 짧은 문구가 포함된 이미지로 구성되고, 네거티브 세트는 텍스트 캡션에 관련 단어가 없는 무작위로 선택된 이미지로 구성됩니다. 그림 10에서 우리는 시각화합니다.

그림 9. 두 장면 카테고리에 대한 정보 개체. 식당 및 욕실 카테고리의 경우 원본 이미지 예시(상단)와 해당 카테고리의 목록을 보여줍니다.6 해당 장면 카테고리에서 가장 빈번한 객체를 해당 출현 빈도와 함께 표시합니다. 하단: CAM과 높은 활성화 영역과 가장 자주 겹치는 6개 개체 목록입니다.

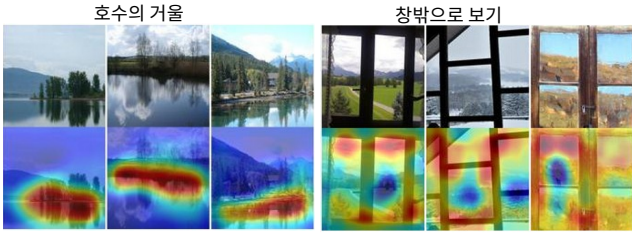

그림 10. 약하게 레이블이 지정된 이미지에서 학습된 개념에 대한 정보 영역. 상당히 추상적임에도 불구하고 개념은 GoogLeNet-GAP 네트워크에 의해 적절하게 현지화되었습니다.

그림 11. 약하게 지도되는 텍스트 감지기 학습. 네트워크가 텍스트나 경계 상자 주석으로 훈련되지 않았더라도 이미지에서 텍스트가 정확하게 감지됩니다.

두 개의 개념 감지기에 대한 최고 순위의 이미지와 CAM. 문구가 일반적인 개체 이름보다 훨씬 추상적임에도 불구하고 CAM은 개념에 대한 정보 영역을 지역화합니다.

약한 감독 텍스트 감지기:우리는 SVT 데이터 세트 [25]의 텍스트를 포함하는 350개의 Google StreetView 이미지를 포지티브 세트로 사용하고 SUN 데이터 세트 [27]의 야외 장면 이미지에서 무작위로 샘플링한 이미지를 네거티브 세트로 사용하여 약한 감독 텍스트 감지기를 훈련합니다. 그림 11에서 볼 수 있듯이 우리의 접근 방식은 경계 상자 주석을 사용하지 않고 텍스트를 정확하게 강조 표시합니다.

시각적 질문 응답 해석:우리는 시각적 질문 답변을 위해 [35]에서 제안된 기준선에 우리의 접근 방식과 지역화 가능한 심층 기능을 사용합니다. 개방형 트랙의 테스트 표준에서 전체 정확도는 55.89%입니다. 그림 12에서 볼 수 있듯이 우리의 접근 방식은 예측된 답변과 관련된 이미지 영역을 강조 표시합니다.

그림 12. 시각적 질문 답변에서 예상 답변 클래스에 대해 강조 표시된 이미지 영역의 예.

5. 클래스별 단위 시각화

저우*외*[33]은 CNN의 다양한 계층의 컨볼루션 단위가 질감이나 재료와 같은 낮은 수준의 개념부터 객체나 장면과 같은 높은 수준의 개념을 식별하는 시각적 개념 탐지기 역할을 한다는 것을 보여주었습니다. 네트워크 속으로 깊이 들어갈수록 단위는 점점 더 차별적이 됩니다. 그러나 많은 네트워크에서 완전히 연결된 레이어를 고려할 때 다양한 카테고리를 식별하기 위한 다양한 단위의 중요성을 식별하는 것은 어려울 수 있습니다. 여기에서는 GAP와 순위가 매겨진 소프트맥스 가중치를 사용하여 특정 클래스에 대해 가장 차별적인 단위를 직접 시각화할 수 있습니다. 여기서 우리는 그들을*클래스별 유닛*CNN의.

그림 13은 AlexNet의 클래스별 단위를 보여줍니다. *-객체 인식을 위한 ILSVRC 데이터 세트(상단) 및 장면 인식을 위한 장소 데이터베이스(하단)에 대해 훈련된 GAP. 수용 필드를 추정하고 최종 컨볼루셔널 레이어에서 각 유닛의 상위 활성화 이미지를 분할하기 위해 [33]과 유사한 절차를 따릅니다. 그런 다음 소프트맥스 가중치를 사용하여 특정 클래스에 대한 단위의 순위를 매깁니다. 그림을 통해 분류 시 가장 구별되는 물체 부분과 이러한 부분을 감지하는 장치가 정확히 무엇인지 식별할 수 있습니다. 예를 들어, 강아지 얼굴과 몸의 털을 감지하는 유닛은*레이크랜드 테리어*; 소파, 테이블, 벽난로를 감지하는 장치는*거실*. 따라서 우리는 CNN이 실제로 각 단어가 차별적인 클래스별 단위인 단어 모음을 학습한다고 추론할 수 있습니다. 이러한 클래스별 단위의 조합은 CNN이 각 이미지를 분류하는 데 도움이 됩니다.

6. 결론

이 연구에서 우리는 전역 평균 풀링을 사용하는 CNN을 위한 CAM(Class Activation Mapping)이라는 일반적인 기술을 제안합니다. 이를 통해 분류 훈련을 받은 CNN은 경계 상자 주석을 사용하지 않고도 객체 위치 파악을 수행하는 방법을 학습할 수 있습니다. 클래스 활성화 맵을 사용하면 주어진 이미지에서 예측된 클래스 점수를 시각화하여 CNN이 감지한 식별 가능한 객체 부분을 강조할 수 있습니다. 우리는 ILSVRC 벤치마크에서 약하게 감독되는 객체 위치 파악에 대한 접근 방식을 평가하여 전역 평균 풀링 CNN이 정확한 객체를 수행할 수 있음을 보여줍니다.

그림 13. ImageNet(상단)과 Places(하단)에서 각각 훈련된 AlexNet*-GAP에 대한 클래스별 단위 시각화. 선택한 3개 클래스의 상위 3개 단위가 각 데이터세트에 대해 표시됩니다. 각 행에는 해당 장치의 수용 필드로 분할된 가장 자신감 있는 이미지가 표시됩니다. 예를 들어 칠판, 의자, 테이블을 감지하는 단위는 분류에 중요합니다. *교실*장면 인식을 위해 훈련된 네트워크용.

현지화. 또한 우리는 CAM 현지화 기술이 다른 시각적 인식 작업에 일반화된다는 것을 보여줍니다. 즉, 우리 기술은 다른 연구자가 CNN이 작업에 사용하는 차별의 기초를 이해하는 데 도움이 될 수 있는 일반적인 현지화 가능한 심층 기능을 생성합니다.

참고자료

[1] A. Bergamo, L. Bazzani, D. Anguelov 및 L. Torresani. 심층 네트워크를 통한 독학 객체 위치 파악. *arXiv 사전 인쇄본 arXiv:1409.3964*, 2014.

[2] RG Cinbis, J. Verbeek 및 C. Schmid. 다중 다중 인스턴스 학습을 통한 약한 지도 객체 위치 파악. *IEEE 트랜스. 패턴 분석 및 기계 지능에 관한*, 2015.

[3] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell. Decaf: 일반적인 시각적 인식을 위한 심층 컨벌루션 활성화 기능입니다. *머신러닝에 관한 국제 컨퍼런스*, 2014.

[4] A. Dosovitskiy 및 T. Brox. 컨벌루션 네트워크를 이용한 컨벌루션 네트워크 반전. *arXiv 사전 인쇄 arXiv:1506.02753*, 2015.

[5] R.-E. 팬, K.-W. 장, C.-J. Hsieh, X.-R. 왕(Wang), C.-J. 린. Lilinear: 대규모 선형 분류를 위한 라이브러리입니다. *기계 학습 연구 저널*, 2008.

[6] L. Fei-Fei, R. Fergus 및 P. Perona. 몇 가지 교육 예제를 통해 생성적 시각적 모델 학습: 101개 개체 범주에 대해 테스트된 증분 베이지안 접근 방식입니다. *컴퓨터 비전 및 이미지 이해*, 2007.

[7] E. Gavves, B. Fernando, CG Snoek, AW Smeulders 및 T. Tuytelaars. 세분화된 분류를 위한 로컬 정렬. *국제컴퓨터비전저널*, 2014.

[8] R. Girshick, J. Donahue, T. Darrell 및 J. Malik. 정확한 객체 감지 및 의미론적 분할을 위한 풍부한 기능 계층. *진행 CVPR*, 2014.

[9] G. Griffin, A. Holub 및 P. Perona. Caltech-256 객체 카테고리 데이터세트. 2007.

[10] A. Krizhevsky, I. Sutskever 및 GE Hinton. 심층 컨벌루션 신경망을 사용한 Imagenet 분류. *신경 정보 처리 시스템의 발전*, 2012.

[11] S. Lazebnik, C. Schmid 및 J. Ponce. 다양한 기능 제공: 자연 장면 카테고리를 인식하기 위한 공간 피라미드 매칭. *진행 CVPR*, 2006.

[12] L.-J. Li와 L. Fei-Fei. 무엇, 어디서, 누구? 장면 및 객체 인식별로 이벤트를 분류합니다. *진행 ICCV*, 2007.

[13] M. Lin, Q. Chen, S. Yan. 네트워크 속의 네트워크. *학습 표현에 관한 국제 컨퍼런스*, 2014.

[14] A. Mahendran 및 A. Vedaldi. 이미지를 반전하여 깊은 이미지 표현을 이해합니다. *진행 CVPR*, 2015.

[15] M. Oquab, L. Bottou, I. Laptev 및 J. Sivic. 컨벌루션 신경망을 사용하여 중간 수준 이미지 표현을 학습하고 전송합니다. *진행 CVPR*, 2014.

[16] M. Oquab, L. Bottou, I. Laptev 및 J. Sivic. 객체 현지화는 무료인가요? 컨볼루셔널 신경망을 이용한 약한 지도 학습. *진행 CVPR*, 2015.

[17] G. 패터슨과 J. 헤이스. Sun 속성 데이터베이스: 장면 속성을 발견하고 주석을 달고 인식합니다. *진행 CVPR*, 2012.

[18] A. Quatoni 및 A. Torralba. 실내 장면을 인식합니다. *진행 CVPR*, 2009.

[19] AS Razavian, H. Azizpour, J. Sullivan 및 S. Carlsson. Cnn의 기성품 기능: 인식을 위한 놀라운 기준선입니다. *arXiv 사전 인쇄본 arXiv:1403.6382*, 2014.

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, AC Berg, L. Fei-Fei. 이미지넷 대규모 시각인식 챌린지. ~ 안에 *국제컴퓨터비전저널*, 2015.

[21] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus 및 Y. LeCun. Overfeat: 컨볼루셔널 네트워크를 사용한 통합 인식, 위치 파악 및 감지. *arXiv 사전 인쇄본 arXiv:1312.6229*, 2013.

[22] K. Simonyan, A. Vedaldi 및 A. Zisserman. 컨벌루션 네트워크 심층 분석: 이미지 분류 모델 및 돌출 맵 시각화. *학습 표현 워크숍에 관한 국제 회의*, 2014.

[23] K. Simonyan 및 A. Zisserman. 대규모 이미지 인식을 위한 매우 깊은 컨벌루션 네트워크. *학습 표현에 관한 국제 컨퍼런스*, 2015.

[24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke 및 A. Rabinovich. 컨볼루션을 통해 더 깊이 들어가 보세요. *arXiv 사전 인쇄본 arXiv:1409.4842*, 2014.

[25] K. Wang, B. Babenko 및 S. Belongie. 엔드투엔드 장면 텍스트 인식. *진행 ICCV*, 2011.

[26] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie 및 P. Perona. Caltech-UCSD Birds 200. 기술 보고서, 캘리포니아 공과대학, 2010.

[27] J. Xiao, J. Hays, KA Ehinger, A. Oliva 및 A. Torralba. Sun 데이터베이스: 수도원부터 동물원까지 대규모 장면 인식. *진행 CVPR*, 2010.

[28] B. Yao, X. Jiang, A. Khosla, AL Lin, L. Guibas 및 L. Fei-Fei. 행동 속성과 부분을 기반으로 학습하여 인간의 행동을 인식합니다. *진행 ICCV*, 2011.

[29] MD Zeiler 및 R. Fergus. 컨벌루션 네트워크를 시각화하고 이해합니다. *진행 ECCV*, 2014.

[30] N. Zhang, J. Donahue, R. Girshick 및 T. Darrell. 세분화된 카테고리 감지를 위한 r-cnn 기반. 부분-*진행 ECCV*, 2014.

[31] N. Zhang, R. Farrell, F. Iandola 및 T. Darrell. 세분화된 인식 및 속성 예측을 위한 변형 가능한 부품 설명자입니다. *진행 ICCV*, 2013.

[32] B. Zhou, V. Jagadeesh 및 R. Piramuthu. 개념학습자: 약하게 레이블이 지정된 이미지 컬렉션에서 시각적 개념을 발견합니다. *진행 CVPR*, 2015.

[33] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva 및 A. Torralba. 객체 탐지기는 심층 장면 CNN에 나타납니다. *학습 표현에 관한 국제 컨퍼런스*, 2015.

[34] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba 및 A. Oliva. 장소 데이터베이스를 활용하여 장면 인식을 위한 심층 기능을 학습합니다. *신경 정보 처리 시스템의 발전*, 2014.

[35] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam 및 R. Fergus. 시각적 질문 답변을 위한 간단한 기준선입니다. *arXiv 사전 인쇄 arXiv:1512.02167*, 2015.