ImageNet Classification with Deep Convolutional Neural Networks (2012)

태그 AlexNet

논문 리뷰: ImageNet Classification with Deep Convolutional Neural Networks (AlexNet)

1. 서론 (Introduction)

객체 인식 성능을 향상시키기 위해서는 더 큰 데이터셋, 더 강력한 모델, 그리고 과적합 방지 기술이 필수적입니다. 기존의 작은 데이터셋(수만 장 규모)으로는 한계가 있었으나, ImageNet과 같이 수백만 장의 이미지를 포함하는 대규모 데이터셋의 등장은 새로운 가능성을 열었습니다. 이 논문은 이러한 대규모 데이터셋에서 효과적으로 학습할 수 있는 CNN을 제안합니다. CNN은 이미지의 통계적 정상성 및 픽셀 의존성의 지역성과 같은 특성을 잘 활용하며, 기존의 완전 연결 신경망보다 훨씬 적은 연결과 파라미터로 효율적인 학습이 가능합니다.

비록 CNN이 매력적이지만, 고해상도 이미지에 대규모로 적용하기에는 계산 비용이 매우 컸습니다. 다행히 GPU 기술의 발전과 최적화된 2D 합성곱 구현 덕분에 대형 CNN의 학습이 가능해졌습니다. 본 논문은 ILSVRC-2010 및 ILSVRC-2012 대회에서 사용된 ImageNet 부분집합에 대해 당시 가장 큰 CNN 중 하나를 학습시켜 최고의 결과를 달성한 내용을 다룹니다. 또한 성능 향상 및 학습 시간 단축을 위한 새롭고 독특한 특징들, 그리고 과적합 방지 기법들을 소개합니다. 모델의 깊이가 성능에 중요하며, 최종 네트워크 크기는 주로 GPU 메모리와 학습 시간에 의해 제한된다고 언급합니다.

2. 데이터셋 (The Dataset)

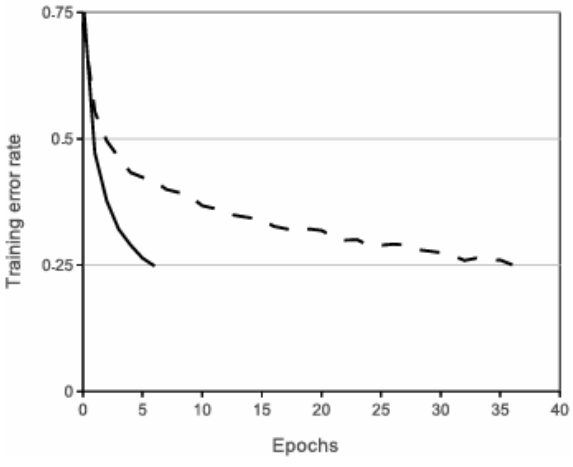
ImageNet은 약 22,000개 카테고리에 속하는 1,500만 개 이상의 레이블된 고해상도 이미지 데이터셋입니다. 본 연구에서는 주로 ILSVRC-2010 버전을 사용했으며, 이는 1,000개 카테고리 각각에 약 1,000개의 이미지가 포함되어 총 120만 개의 훈련 이미지, 50,000개의 검증 이미지, 150,000개의 테스트 이미지로 구성됩니다. ImageNet에서는 top-1 및 top-5 오류율을 보고하는 것이 일반적입니다.

ImageNet 이미지는 다양한 해상도를 가지므로, 256x256으로 다운샘플링한 후 중앙에서 224x224 패치를 잘라내어 일정한 입력 크기를 사용했습니다. 각 픽셀에서 훈련 세트의 평균 활성도를 빼는 것 외에는 별도의 전처리를 수행하지 않고 원시 RGB 값으로 네트워크를 훈련했습니다.

3. 아키텍처 (The Architecture)

네트워크는 5개의 합성곱 계층과 3개의 완전 연결 계층, 총 8개의 학습 가능한 계층으로 구성됩니다. 주요 특징은 다음과 같습니다.

3.1 ReLU 비선형성 (ReLU Nonlinearity)



(그림 1 - ReLU와 tanh의 학습 속도 비교 그래프)

기존의 tanh나 sigmoid 같은 포화(saturating) 비선형 함수 대신, ReLU(Rectified Linear Unit, f(x)=max(0,x))를 사용했습니다. ReLU는 경사 하강법으로 훈련 시 포화 비선형 함수보다 몇 배 더 빠르게 학습되며, 이는 대형 신경망 실험에 필수적이었습니다.

3.2 다중 GPU에서의 훈련 (Training on Multiple GPUs)

단일 GPU(GTX 580, 3GB 메모리)의 메모리 제약으로 인해 네트워크를 두 개의 GPU에 분산시켜 학습했습니다. 각 GPU에 커널(뉴런)의 절반을 할당하고, 특정 계층에서만 GPU 간 통신이 이루어지도록 하여 효율성을 높였습니다. 이 방식은 단일 GPU로 절반 크기의 네트워크를 학습시킨 경우보다 top-1 및 top-5 오류율을 각각 1.7% 및 1.2% 감소시켰습니다.

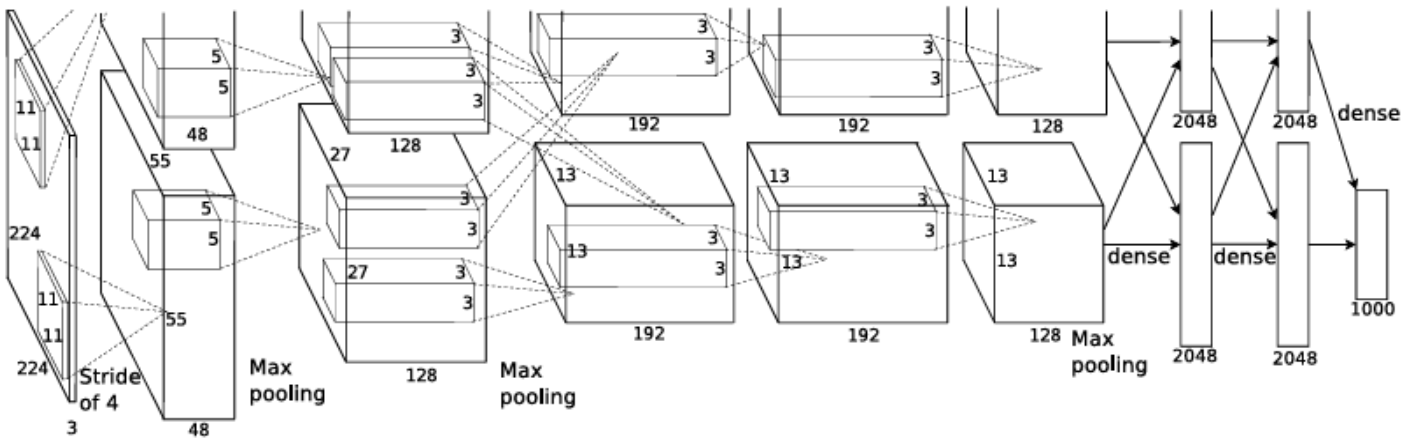
3.3 지역적 반응 정규화 (Local Response Normalization, LRN)

ReLU는 입력 정규화가 필요 없는 장점이 있지만, LRN을 적용하여 일반화 성능을 향상시켰습니다. LRN은 실제 뉴런의 측면 억제(lateral inhibition) 현상에서 영감을 받아, 서로 다른 커널로 계산된 뉴런 출력 간의 경쟁을 유도합니다. 이 방식은 top-1 및 top-5 오류율을 각각 1.4%와 1.2% 감소시켰습니다.

3.4 오버래핑 풀링 (Overlapping Pooling)

전통적인 비오버래핑 풀링 대신, 풀링 유닛의 이웃 영역이 겹치도록 하는 오버래핑 풀링(스트라이드 s=2, 풀링 윈도우 z=3)을 사용했습니다. 이는 비오버래핑 방식에 비해 top-1 및 top-5 오류율을 각각 0.4%와 0.3% 감소시켰으며, 과적합을 약간 더 어렵게 만드는 경향이 관찰되었습니다.

3.5 전체 아키텍처 (Overall Architecture)



(그림 2 - AlexNet 전체 아키텍처 다이어그램)

네트워크는 5개의 합성곱 계층과 3개의 완전 연결 계층으로 구성되며, 최종적으로 1000-way 소프트맥스 출력을 가집니다.

GPU 병렬 처리: 두 번째, 네 번째, 다섯 번째 합성곱 계층의 커널은 동일 GPU 내의 이전 계층 커널 맵에만 연결되고, 세 번째 합성곱 계층은 이전 계층의 모든 커널 맵에 연결됩니다. 완전 연결 계층은 모든 이전 뉴런에 연결됩니다.

LRN은 첫 번째와 두 번째 합성곱 계층 뒤에, 최대 풀링은 LRN 계층 뒤와 다섯 번째 합성곱 계층 뒤에 적용됩니다. ReLU는 모든 합성곱 및 완전 연결 계층의 출력에 적용됩니다.

- 계층별 세부 구성:
  - 1번 합성곱: 11x11×3 커널 96개 (스트라이드 4)
  - 2번 합성곱: 5x5×48 커널 256개
  - 3번 합성곱: 3x3×256 커널 384개
  - 4번 합성곱: 3x3×192 커널 384개
  - 5번 합성곱: 3x3×192 커널 256개
  - 완전 연결 계층: 각 4096개 뉴런

4. 과적합 줄이기 (Reducing Overfitting)

6,000만 개의 매개변수를 가진 이 네트워크는 과적합 위험이 컸습니다. 이를 해결하기 위해 다음 두 가지 주요 방법을 사용했습니다.

4.1 데이터 증강 (Data Augmentation)

두 가지 형태의 데이터 증강을 사용했습니다:

1. **이미지 이동 및 수평 반전**: 256x256 이미지에서 무작위로 224x224 패치(및 수평 반전)를 추출하여 훈련 세트를 2048배 증가시켰습니다. 테스트 시에는 5개 패치(네 모서리, 중앙)와 그 반전, 총 10개 패치의 예측을 평균 냈습니다.
2. **RGB 채널 강도 변경**: ImageNet 훈련 세트의 RGB 픽셀 값에 PCA를 수행하고, 주성분에 고유값과 가우시안 확률 변수를 곱한 값을 더하여 조명의 강도 및 색상 변화에 대한 불변성을 학습하도록 했습니다. 이는 top-1 오류율을 1% 이상 줄였습니다.

4.2 드롭아웃 (Dropout)

처음 두 개의 완전 연결 계층에 드롭아웃을 사용했습니다. 각 은닉 뉴런의 출력을 0.5의 확률로 0으로 만들어, 뉴런 간의 복잡한 상호 적응을 줄이고 더 강건한 특징을 학습하도록 했습니다. 테스트 시에는 모든 뉴런을 사용하되 출력을 0.5 곱하여 여러 드롭아웃 네트워크의 예측을 평균 내는 효과를 얻었습니다. 드롭아웃은 수렴에 필요한 반복 횟수를 대략 두 배로 늘렸습니다.

5. 결과 (Results)

- **ILSVRC-2010**: top-1 오류율 37.5%, top-5 오류율 17.0%를 달성하여 당시 최고 성능(각각 47.1%, 28.2%)을 크게 앞섰습니다.
- **ILSVRC-2012**:
  - 단일 모델: top-5 오류율 18.2%
  - 5개 유사 CNN 앙상블: top-5 오류율 16.4%
  - ImageNet Fall 2011 전체 데이터로 사전 훈련 후 미세 조정한 2개 모델 + 위 5개 모델 앙상블: **top-5 오류율 15.3%로 우승** (2위는 26.2%)
- **ImageNet Fall 2009 버전 (10,184 범주, 890만 이미지)**: 6번째 합성곱 계층을 추가한 모델로 top-1 67.4%, top-5 40.9%를 달성하여 이전 최고 결과(78.1%, 60.9%)를 능가했습니다.

5.1 정성적 평가 (Qualitative Evaluations)

- (그림 3 위치 - 학습된 커널 시각화)

  학습된 커널은 다양한 주파수 및 방향 선택적 특징, 색상 얼룩 등을 보여줍니다. 두 GPU 간의 제한된 연결로 인해 GPU 1은 주로 색상 무관 커널을, GPU 2는 주로 색상 특화 커널을 학습하는 특수화 현상이 나타났습니다.

- (그림 4 위치 - 테스트 이미지 예측 및 유사 이미지 검색 결과)

  테스트 이미지에 대한 상위 5개 예측은 합리적이었으며, 중심에서 벗어난 객체도 인식했습니다. 마지막 은닉 계층의 특징 벡터 간 유클리드 거리를 사용하여 유사 이미지를 검색한 결과, 픽셀 수준에서는 다르지만 의미론적으로 유사한 이미지를 잘 찾아냈습니다. 이는 원시 픽셀 기반 검색보다 우수함을 시사합니다.

6. 논의 (Discussion)

이 연구는 크고 깊은 CNN이 순수 지도 학습만으로도 매우 어려운 데이터셋에서 기록적인 결과를 달성할 수 있음을 보여줍니다. 네트워크의 깊이는 성능에 매우 중요하며, 중간 계층 하나를 제거해도 성능이 약 2% 저하되었습니다. 비지도 사전 훈련을 사용하지 않았지만, 향후 더 큰 네트워크와 데이터셋이 가능해지면 도움이 될 것으로 예상합니다. 인간 시각 시스템 수준에 도달하려면 아직 갈 길이 멀며, 궁극적으로 시간적 정보를 활용할 수 있는 비디오 시퀀스에 매우 크고 깊은 CNN을 적용하고자 합니다.