Week 4

Abstract

이 비디오는 'Vision Transformer (ViT)' 의 발전과 이미지 인식에서의 효과를 탐구하며, CNN에 대한 의존성을 제거하고 순수한 Transformer 구조를 사용하여 이미지 패치를 처리하는 방법을 소개합니다. ViT가 대규모 데이터셋에서 훈련된 후, 여러 이미지 인식 벤치마크에서 뛰어난 성과를 거두었음을 보여줍니다. Transformer 아키텍처를 이미지 인식에 적용함으로써 얻은 통찰력은 미래의 컴퓨터 비전 연구에서 중요한 변화를 가져올 수 있습니다.

핵심주제

1. 비전 트랜스포머(**ViT**)는 이미지 인식에서 **최첨단** 성과를 이룩하고 있다.

- ViT는 대규모 데이터셋에서 **사전 훈련**한 후 다양한 이미지 인식 벤치마크에서 우수한 결과를 기록했다.
- 비전 트랜스포머는 계산 비용이 적으면서도 다른 **CNN** 구조에 비해 성능이 뛰어난 장점이 있다.
- 특히, ViT는 **자기 주의** 메커니즘을 통해 이미지 정보 통합을 효과적으로 수행한다.

2. 트랜스포머(**Transformer**) 아키텍처는 이미지 처리에 **도전 과제**가 있다.

- 트랜스포머의 나이브 적용은 픽셀 수에 따라 **2차 비용** 문제가 발생해 현실적인 입력 크기에 한계가 있다.
- 이를 해결하기 위해 여러 **근사 방법**이 시도되고 있으며, CNN과의 통합도 연구되고 있다.
- 또한, 최근 연구에서는 **하이브리드 모델**이 자주 등장하여 성능 향상에 기여하고 있다.

3. **자기 감독 학습** 방법이 트랜스포머 성능 향상에 기여한다.

- 자기 지도 사전 훈련을 통해 ViT는 감독된 학습에 비해 **전반적인 성능** 향상을 보여준다.

- 마스크 패치 예측과 같은 방법이 그 예시로, 데이터셋이 적더라도 성능을 개선할 수 있다.
- 하지만 여전히 감독된 학습에 비해 능력이 **제한적**인 부분이 존재한다.

4. 비전 트랜스포머는 **계산 비용**과 성능의 균형에서 유리하다.

- ViT는 다른 CNN 모델들보다 **낮은 비용**으로 최첨단 성능을 발휘할 수 있다.
- 실험 결과, ViT는 이미지넷에서 효과적인 성능을 보여주며, 여러 벤치마크에서 우수한 결과를 도출한다.
- 이로 인해 ViT는 대규모 이미지 데이터셋을 효과적으로 처리할 수 있는 **효율성**을 확보하고 있다.

5. 비전 트랜스포머 아키텍처는 **메모리 효율성**에서 장점을 가진다.

- ViT 모델은 ResNet 모델에 비해 메모리 사용에서 **효율성**을 보이며, 추론 속도 또한 개선된다.
- 이러한 특성은 대규모 데이터셋 처리 시 **확장성**을 높이는 데에도 긍정적인 영향을 준다.
- 모델의 설계와 훈련에서 메모리 관리가 중요하게 고려되고 있음을 알 수 있다.

타임라인

1. 🖼️이미지 인식을 위한 변환기 사용

- 변환기 아키텍처는 자연어 처리 작업의 사실상 표준이 되었지만, 컴퓨터 비전 분야에서는 그 응용이 제한적이다.
- 전통적으로 주의(attention)는 컨볼루션 네트워크와 함께 사용되거나, 컨볼루션 네트워크의 특정 구성 요소를 대체하는 방식으로 적용되지만, 이는 필수가 아님을 보여준다.
- 이미지 패치의 시퀀스에 직접 적용된 순수 변환기는 이미지 분류 작업에서 매우 우수한 성능을 발휘할 수 있다.
- 대량의 데이터로 사전 훈련된 후 여러 중간 혹은 소규모 이미지 인식 벤치마크(ImageNet, CIFAR-100, VTAB 등)에 전이되면, 비전 변환기(Vision Transformer, ViT)는 최첨단 컨볼루션 네트워크와 비교하여 우수한 결과를 얻으면서도 훈련에 필요한 계산 자원을 상당히 적게 소모한다.

2. 📊셀프 어텐션 기반 구조의 이미지 인식 연구

- 셀프 어텐션 구조, 특히 **트랜스포머**는 자연어 처리(NLP)에서 선택된 모델이 되었다.
- 이 모델은 큰 텍스트 코퍼스에서 사전 훈련한 다음, 더 작은 작업 특화 데이터셋에서 미세 조정하는 주도가 명확하다.
- 하지만 대규모 이미지 인식에서는 고전적인 ResNet과 같은 구조가 여전히 **최첨단**을 유지하고 있다.
- 이 연구에서는 이미지를 패치로 나누고 이 패치의 선형 임베딩 시퀀스를 트랜스포머에 입력하여 직접적으로 이미지에 트랜스포머를 적용함으로써, 대규모 데이터셋에서 훈련한 경우 **탁월한 성과**를 달성할 수 있음을 발견하였다.
- 특히 Vision Transformer(ViT)는 ImageNet 및 JFT-300M 데이터셋에서 사전 훈련한 후 여러 이미지 인식 벤치마크에서 최상의 결과를 기록하였다.

3. 트랜스포머와 이미지 처리 연구

- 트랜스포머는 **기계 번역**을 위해 Vaswani et al. (2017)에 의해 제안되어, 많은 NLP 작업에서 **최첨단 방법**이 되었다.
- 이 모델들은 대규모 코퍼스에서 사전 학습된 뒤, 특정 작업에 맞게 조정된다; 예를 들어, BERT는 **노이즈 제거** 자가 감독 사전 학습 작업을 사용한다.
- 그러나 이미지에 대한 나이브 자가 주의 적용은 픽셀 수에 따라 **2차 비용**이 드는 문제로 인해 현실적인 입력 크기에는 적용할 수 없다.
- 따라서 이미지 처리에 트랜스포머를 적용하기 위해 여러 근사 방법이 시도되었으며, 그 중 일부는 **국소적인** 이웃에서만 자가 주의를 적용하였다.
- 또한, 최신 연구에서는 CNN과 자가 주의를 결합하거나, 이미지 픽셀에 트랜스포머를 적용한 모델이 등장하여 이미지 인식 성능을 높이고 있다.

4. 비전 트랜스포머(ViT)의 설계 및 조정

- 모델 설계에서 우리는 원래의 **트랜스포머**를 최대한 따르며, 이는 스케일 가능한 NLP 트랜스포머 아키텍처를 거의 즉시 사용할 수 있는 장점이 있다.
- 비전 트랜스포머(ViT)는 2D 이미지를 처리하기 위해 이미지를 2D 패치로 변형하고 이를 1D 시퀀스로 변환하여 입력으로 사용한다.

- MLP 블록과 다중 헤드 자기 주의(MSA) 계층이 교차하는 트랜스포머 인코더를 통과하면서 항상 상수의 잠재 벡터 크기를 유지한다.
- 비전 트랜스포머는 CNN보다 이미지에 대한 **유도 편향**이 적으며, 혼합 아키텍처를 통해 CNN의 기능 맵을 입력으로 사용할 수도 있다.
- Fine-tuning 과정에서 고해상도 이미지를 처리할 때, 미리 훈련된 위치 임베딩을 2D 보간을 통해 조정하여 더욱 효율적인 성능을 낸다.

5. 🧪실험: ResNet, 비전 트랜스포머의 학습 능력 평가

- 우리는 ResNet, Vision Transformer(ViT), 그리고 하이브리드 모델의 **표현 학습 능력**을 평가한다.
- 각 모델의 데이터 요구 사항을 이해하기 위해 다양한 크기의 데이터셋으로 사전 훈련을 수행하고 몇 가지 벤치마크 작업을 평가한다.
- 사전 훈련의 계산 비용을 고려할 때, ViT는 상당히 유리하며 대부분의 인식 벤치마크에서 낮은 비용으로 최첨단 성능을 달성한다.
- 마지막으로, **자기 감독** 방법을 사용하는 소규모 실험을 실시하여, 자기 감독된 ViT가 미래에 가능성을 가지는 것을 보여준다.
- 모델 확장을 탐구하기 위해, ILSVRC-2012 ImageNet 데이터셋(1k 클래스, 1.3M 이미지)을 사용하며, 이를 포함한 더 큰 데이터셋인 ImageNet-21k와 JFT도 활용한다.

6. 비전 트랜스포머 모델 변형 및 평가

- 비전 트랜스포머(ViT) 모델의 변형에 대한 세부사항은 표 1에 제시되어 있으며, ViT-Base, ViT-Large, ViT-Huge 모델들이 포함된다.
- VTAB 분류 스위트를 사용해 19개의 작업을 평가하며, 각 작업에 1,000개의 훈련 예제를 사용한다.
- 모델 변형은 BERT(BERT Devlin et al., 2019) 구성을 기반으로 하며, “Base” 및 “Large” 모델은 BERT에서 직접 채택되었고, “Huge” 모델이 추가되었다.
- 모델 크기와 입력 패치 크기를 간략하게 표시할 때 ViT-L/16은 “Large” 변형을 나타낸다.
- 전이 학습을 개선하기 위해 ResNet을 기본 CNN으로 사용하되, 배치 정규화 계층을 그룹 정규화로 대체하고, 표준화된 합성곱을 사용하였다.

7. 모델 훈련 및 미세 조정 방법

- 모든 모델, 특히 ResNet을 훈련하기 위해 Adam 옵티마이저를 사용하며, 배치 크기는 4096이고 높은 **가중치 감소** 0.1를 적용했다.
- 저희 연구에서는 Adam이 ResNet의 경우 SGD보다 약간 더 나은 성능을 보임을 발견했다.
- 미세 조정을 위해 SGD와 모멘텀을 사용하고, 배치 크기는 512로 설정했다.
- 이미지넷 결과에 대한 미세 조정은 해상도가 512인 ViT-L/16과 518인 ViT-H/14를 사용했으며, Polyak & Juditsky의 평균화를 적용했다.
- 결과는 다운스트림 데이터셋에서 몇 샷 또는 미세 조정 정확도를 통해 보고되며, 각 모델의 성능을 캡처한다.

8. 📊최신 기술과의 비교 분석

- 우리는 큰 모델인 ViT-H/14 및 ViT-L/16을 기존 CNN들과 비교하였다.
- 비교 지점 중 하나는 Big Transfer(BiT)이며, 다른 하나는 라벨이 제거된 대규모 Semi-supervised 학습을 통해 훈련된 Noisy Student이다.
- 모델들은 TPUv3 하드웨어에서 훈련되었으며, 각 모델의 프리트레인에 소요된 TPUv3-core-days를 보고하였다.
- 결과적으로, ViT-L/16 모델은 JFT-300M에서 프리트레인 되었으며 BiT-L을 모든 태스크에서 능가하였고, 더 적은 컴퓨팅 리소스를 사용하였다.
- 마지막으로, ViT-L/16 모델은 공개된 ImageNet-21k 데이터셋에서 훈련된 결과, 대부분의 데이터셋에서도 우수한 성과를 보였다.

9. 비전 트랜스포머의 사전 훈련 데이터 요구사항

- 비전 트랜스포머(ViT)는 JFT-300M 데이터셋에서 사전 훈련 시 성능이 우수하다.
- 우리는 먼저 ImageNet, ImageNet-21k, JFT-300M의 크기를 증가시키며 ViT 모델을 사전 훈련하였다.
- 작은 데이터셋인 ImageNet에서 사전 훈련된 ViT-Large 모델은 ViT-Base 모델에 비해 성능이 떨어지지만, JFT-300M에서 사전 훈련되었을 때 큰 모델의 이점을 볼 수 있다.

- 모델의 내재적 속성을 평가하기 위해 9M, 30M, 90M의 무작위 샘플 하위 집합 및 전체 JFT-300M 데이터셋에 대해 모델을 학습했다.
- 결과적으로 작은 데이터셋에서는 컨볼루션 유도 편향이 유용하나, 큰 데이터셋에서는 직접 데이터를 통해 필요한 패턴을 학습하는 것이 효율적이다.

10. 🔍비전 트랜스포머와 모델 성능 분석

- 4.4 스케일링 연구에서는 JFT-300M에서 **전이 성능**을 평가하여 다양한 모델의 제어된 스케일링 연구를 수행하였다.
- 모델 집합에는 7개의 ResNet과 6개의 비전 트랜스포머, 5개의 하이브리드 모델이 포함되며, 각 모델의 성능과 사전 훈련 비용을 평가하였다.
- 비전 트랜스포머는 ResNet에 비해 성능과 계산 비용의 균형에서 우세하며, 하이브리드 모델은 소규모 계산 예산에서 약간 더 나은 성능을 보인다.
- 또한 Vision Transformer는 시도된 범위 내에서 **포화되지 않는** 경향이 있으며, 향후 스케일링 작업을 위한 동기를 부여한다.
- 4.5 비전 트랜스포머의 입력 주의를 조사하는 과정에서 자기 주의 기법이 이미지 정보 통합을 가능하게 하여, 네트워크가 이 기능을 얼마나 활용하는지 분석하였다.

11. 📊 자기 지도 학습을 통한 Transformer 성능 향상

- Transformer 모델은 NLP 작업에서 **인상적인 성능**을 보여주지만, 그 성공은 우수한 확장성 외에도 대규모 자기 지도 사전 훈련에서 기인한다.
- 이 연구에서는 BERT에서 사용되는 마스킹 언어 모델링 작업을 모방한 마스크 패치 예측에 대한 초기 탐색을 수행한다.
- 자기 지도 사전 훈련을 통해 더 작은 ViT-B/16 모델이 이미지넷에서 79.9%의 정확도를 달성했으며, 이는 처음부터 훈련했을 때보다 2% 향상된 수치이다.
- 하지만 여전히 감독된 사전 훈련에 비해 4% 뒤처진 상황이다.

Conclusion

- Transformers을 이미지 인식에 직접 적용하는 방법을 탐구했다.

- 이전 연구와 달리, 우리는 초기 패치 추출 단계를 제외하고는 아키텍처에 **이미지 특화된 유도 편향**을 도입하지 않았다.
- 대신 이미지를 패치 시퀀스로 해석하고 자연어 처리에 사용되는 표준 Transformer 인코더로 처리했다.
- 이 단순하면서도 확장 가능한 전략은 대규모 데이터셋에서의 사전 훈련과 결합했을 때 놀라운 성능을 발휘했다.
- VIT는 많은 이미지 분류 데이터셋에서 최신 기술과 비교하여 동등하거나 초과하는 결과를 보였지만 여전히 많은 도전 과제가 남아 있다.

Additional Analysis

14. 하이퍼파라미터 및 멀티헤드 셀프 어텐션 개요

- 모델 훈련을 위한 하이퍼파라미터는 배치 크기를 4096으로 설정하고, 학습률 워밍업을 10,000 스텝으로 설정하였다.
- ImageNet에 대한 훈련에서는 전역 정규화 1에서 그래디언트 클리핑을 추가 적용하는 것이 유익하다는 것을 발견했다.
- 입력 시퀀스의 각 요소에 대해 가중합을 계산하며, 어텐션 가중치는 쿼리와 키 표현 간의 쌍별 유사성에 기반한다.
- 멀티헤드 셀프 어텐션(MSA)은 여러 개의 셀프 어텐션 연산을 병렬로 실행하고 이들의 출력을 연결하여 프로젝션하는 확장된 형태이다.

15. 모델 훈련 및 미세 조정 세부사항

- 모델 훈련 설정을 요약하면, 강력한 정규화가 **ImageNet**에서 모델을 처음부터 훈련할 때 중요하다는 것을 발견했다.
- 드롭아웃은 모든 밀집 레이어 후에 적용되며, qkv-프로젝션과 패치 임베딩에 positional 정보를 추가한 직후에는 제외된다.
- 미세 조정을 위해 SGD를 사용하며, 학습률에 대한 작은 그리드 검색을 수행하고, 학습율 범위는 테이블 4에서 확인할 수 있다.

- VTAB 및 기타 데이터셋으로 모델을 전이할 때, 전체 헤드를 제거하고 타겟 데이터셋에 필요한 클래스 수를 출력하는 단일 **제로 초기화** 선형 레이어로 교체하는 것이 더 견고한 결과를 얻을 수 있었다.

16. 자기 감독을 위한 마스킹 패치 예측 목표

- **자기 감독 실험**을 위해 50%의 패치 임베딩을 손상시키고, 이때 80%는 학습 가능한 [마스크] 임베딩으로 대체하며, 10%는 다른 랜덤 패치 임베딩으로, 나머지 10%는 그대로 유지한다.
- 우리는 각 손상된 패치의 3비트 평균 색상을 예측하기 위해, 이들의 패치 표현을 사용하여 1M 단계(약 14 에폭) 동안 모델을 훈련한다.
- 예측 목표로는, 평균 3비트 색상(512색 중 1예측) 예측, 16×16 패치를 3비트 색상으로 병렬 예측, 전 패치에 대한 L2 회귀(각 RGB 채널에 대해 256회귀) 등을 시도했고, 모두 **상당히 좋은 결과**를 보였다.
- 특히, 옵션 1에서 최고 성능을 보였으며, 15%의 손상률을 적용했으나 결과는 다소 낮았다.
- 마지막으로, **우리가 사용한 마스킹 패치 예측**은 대규모 데이터셋이나 사전 훈련 없이도 ImageNet 분류 성능 향상을 가져오는 것을 확인했다.

17. 📊비전 트랜스포머 성능 및 분석 결과

- 본 논문에서는 다양한 ViT 모델의 사전 훈련 데이터셋 크기에 따른 변환 성능 결과를 제시한다.
- 특히, 이미지넷 및 JFT-300M에서 사전 훈련된 모델들의 performance가 날로 증가함을 확인하였다.
- SGD와 Adam 옵티마이저의 차이점을 실험을 통해 보여주었으며, Adam이 더 나은 성능을 보임을 알 수 있었다.
- 또한, 트랜스포머 아키텍처의 다양한 매개변수를 조정해보며 각 매개변수가 성능에 미치는 영향을 관찰하였다.
- 마지막으로, 위치 인코딩 방식을 다양하게 시도하여 각 방식의 성능 차이를 비교하였다.

18. 위치 임베딩의 모델 성능 분석

- 다양한 하이퍼파라미터로 훈련된 모델의 위치 임베딩에 대한 연구 결과를 요약하였다.

- 정렬 임베딩이 없는 모델과 위치 임베딩이 있는 모델 간의 성능 차이는 크지만, 공간 정보를 인코딩하는 방법 간에는 **상당한 차이**가 없음을 알 수 있다.
- 패치 수준 입력으로 작동하는 트랜스포머 인코더 때문에, 이번 실험에서 두 가지 방식 간의 **차이점**이 덜 중요하다고 추측된다.
- 특히, 패치 수준 입력에서 공간 차원은 원래 픽셀 수준 입력보다 훨씬 작으며, 이러한 해상도에서 공간 관계를 표현하는 것은 **서로 다른** 위치 인코딩 전략들에 대해 쉽게 학습될 수 있다.
- 그러나 네트워크가 학습한 위치 임베딩 유사성의 특정 패턴은 훈련 하이퍼파라미터에 따라 달라진다.

19. 🏎️아키텍처의 실제 속도 및 메모리 효율성

- 우리는 아키텍처의 **실제 속도**에 관심이 있으며, 이는 이론적인 FLOPs로 잘 예측되지 않는 경우가 많다.
- 주요 모델의 추론 속도를 TPUv3 가속기에서 측정한 결과, 추론 속도와 역전파 속도 간의 차이는 상수 모델 독립적 요인으로 나타났다.
- 그래프는 각 입력 크기에서 하나의 코어가 초당 처리할 수 있는 이미지 수를 보여주며, 이론적인 비차원 스케일링은 최대 해상도에서만 시작된다.
- 모델이 코어에 적합할 수 있는 최대 배치 크기도 중요한데, 큰 배치 크기가 대규모 데이터셋에 대한 확장성에 더 유리하다.
- 결과적으로 큰 ViT 모델은 ResNet 모델에 비해 메모리 효율성에서 명확한 장점을 보인다.

20. 📊축 방향 주의(Axial Attention)의 개요

- 축 방향 주의(Axial Attention)는 다차원 텐서로 구성된 큰 입력에서 **자기 주의**를 적용하기 위한 간단하면서도 효과적인 기술이다.
- 이 방법은 입력 텐서의 각 축을 따라 여러 가지 주의 작업을 수행하여 정보를 혼합하며, 다른 축에 있는 정보는 독립적으로 유지된다.
- Wang et al.은 AxialResNet 모델을 제안하여 ResNet50의 모든 3x3 커널 크기 합성곱을 축 방향 자기 주의로 대체하고, 상대적 위치 인코딩을 추가하였다.
- 우리는 AxialResNet를 기준 모델로 구현하였으며, ViT를 수정하여 입력을 1차원 패치 시퀀스가 아닌 2차원 형태로 처리하고, 축 방향 변환기 블록을 통합하였다.

- Axial-ViT 모델은 전반적으로 성능이 ViT-B 모델보다 우수하지만, 더 많은 계산 비용이 수반된다.

21. 📊ViT의 자기 주의 메커니즘 분석

- ViT가 이미지 전반에 걸쳐 정보를 통합하기 위해 자기 주의를 사용하는 방법을 이해하기 위해, 우리는 각 레이어에서 주의 가중치의 평균 거리를 분석했다.
- 이 '주의 거리'는 CNN의 수용 영역 크기에 해당하며, 낮은 레이어에서 평균 주의 거리는 매우 변동성이 크고 일부 헤드는 이미지의 대부분을 처리하는 반면, 다른 헤드는 쿼리 위치 근처의 작은 영역에만 주목한다.
- 깊이가 증가함에 따라 모든 헤드의 주의 거리가 늘어나고, 네트워크의 후반부에서는 대부분의 헤드가 토큰 전반에 걸쳐 넓게 주목한다.
- 주의 맵을 계산하기 위해 우리는 Attention Rollout 방법을 사용하며, ViT-L/16의 주의 가중치를 모든 헤드에 걸쳐 평균내고, 모든 레이어의 가중치 행렬을 재귀적으로 곱하여 토큰 간 주의 혼합을 고려한다.

22. 객체 네트워크 결과 및 VTAB 성과 분석

- D.9 객체 네트워크 결과: 우리의 주력 모델 ViT-H/14는 Kolesnikov et al. (2020)의 평가 설정에 따라 ObjectNet 벤치마크에서 상위 5개 정확도 82.1%와 상위 1개 정확도 61.7%를 기록했다.
- D.10 VTAB 분석: 표 9는 각 VTAB-1k 과제에서 달성한 점수를 보여준다.
- ViT-H/14 (JFT) 모델의 성과는 Caltech101에서 95.3, CIFAR-100에서 85.5, 그리고 Flowers102에서 99.7로 나타났으며, 평균적으로 77.6의 점수를 기록했다.