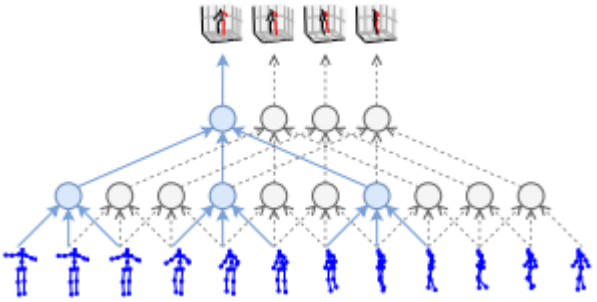
3D human pose estimation in video with temporal convolutions and semi-supervised training

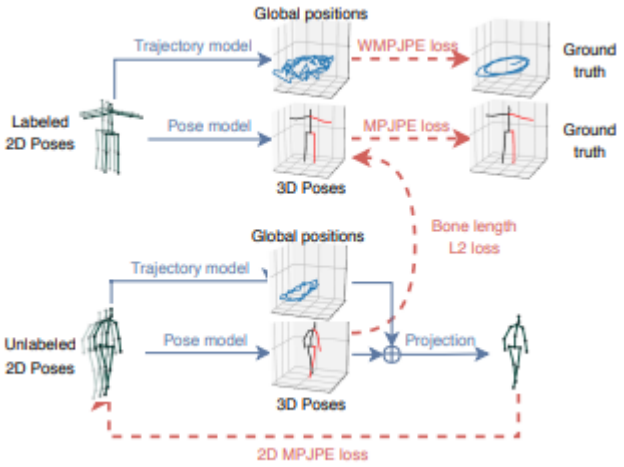


Abstract

Unlabeled video data를 이용한 semi-supervised learning을 이용. unlabeled에서 예측한 2d keypoints를 이용하여 3D posed를 예측하고 이를 다시 2d keypoint로 back-project

Introduction

기존의 방법들은 2D keypoint가 가지고 있는 애매모호함 때문에 3D pose를 정확히 예측하지 못했으므로 이를 해결하기 위해 RNN 같은 모델 이용. CNN 모델도 temporal 정보를 모델링하는 것에 성공적인 모습을 보여주고 있음. Fully convolutional 모델 구조를 이용하여 2d keypoint로부터 3D pose를 예측하는 모델 제안.
라벨링된 데이터가 부족하기 때문에 semi-supervised learning 이용. 이를 위해 cycle consistency 이용



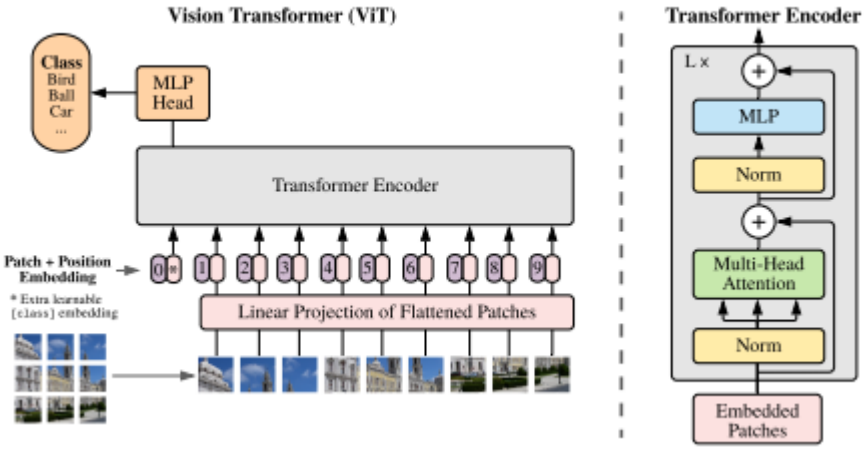
AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Transfomer만을 사용하여 이미지 분류 task 수행

- 이미지는 이미지 조각의 시퀀스로 처리
- 대량의 데이터에 대해 사전 학습한 후 작은 이미지 인식 벤치마크(이미지넷, CIFAR-100, VTAB)에 적용

- 그 결과 Vision Transformer(ViT)는 여타의 SOTA CNN 기반의 모델과 비교했을 때 훌륭한 성능을 얻음
- 동시에 학습 과정에서의 계산 자원은 훨씬 적게 소모

표준 Transformer을 최대한 변형 없이 직접적으로 이미지에 적용. 이를 위해 이미지를 패치 조각으로 쪼갬. 각각의 패치에 대해 선형적인 임베딩의 시퀀스를 제공하여 Transformer에 입력. 이때 이미지 패치는 NLP에서의 토큰과 같이 다룸. 모델은 이미지 분류 task에 대해 지도 학습 방식으로 학습.
ImageNet과 같은 중간 크기의 데이터셋에 이 모델을 적용해본 결과, 같은 크기의 ResNet 모델의 정확도보다 몇 퍼센트 정도 낮은 정확도 얻음. 이 결과는 기대에 미치치 못한 결과처럼 보이지만 모델을 더 많은 데이터셋에 대햏 학습했을 때 CNN이 학습하는 bias의 힘보다 강함. 충분히 큰 스케일에서 ViT를 사전 학습한 결과, 더 적은 데이터셋을 가진 하위 task에 전이 학습하여 좋은 성능을 얻을 수 있음.



모델 : 이미지는 고정된 크기의 패치로 쪼개고, 각각을 선형적으로 임베딩한 후 위치 임베딩을 더하여 결과 벡터를 일반적인 transformer 인코더의 input으로 입력.

Vision Transformer

이미지 인풋
- 일반적인 transformer은 토큰 임베딩에 대한 1차원의 시퀀스를 입력으로 받음
- 2차원의 이미지를 다루기 위해 논문에서는 이미지를 flatten된 2차원의 패치의 시퀀스로 변환

위치 임베딩
- 각각의 패치 임베딩에 위치 임베딩을 더하여 위치 정보를 활용할 수 있도록 함
- 학습 가능한 1차원 임베딩 사용

하이브리드 아키텍처
- 이미지 패치를 그대로 사용하는 대신, CNN 결과 나온 feature map을 인풋 시퀀스로 사용할 수 있음
- 하이브리드 모델에서는 패치 임베딩 프로젝션을 CNN feature map에서 결과로 나온 패치에 대해 적용함

임베딩 프로젝션

- ViT는 펼쳐진 패치를 더 낮은 차원의 공간으로 매핑

위치 임베딩
- 선형 프로젝션 이후, 각각의 패치 representation에는 위치 임베딩이 더해지게 됨
- 시각화 결과를 보면, 모델은 이미지 내의 거리 개념을 인코딩하여 위치 임베딩에서 유사성이 나타난다는 것을 알 수 있음

self-attention
- 가장 밑단에 있는 레이어에서부터 ViT가 전체 이미지에 있는 정보를 통합하도록 도움
- 일관적으로 작은 거리의 패치에 집중
- CNN 밑단에서 일어나는 것과 비슷한 작용