23.07.27 _ 7주차

딥러닝 논문 요약 및 구현 스터디

발표자

CV

[ViT] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

Related work

2.1. Transformer

2017년도에 나왔으며 NLP task에서 transformer를 활용한 모델들이 SOTA를 찍고 있다. 그러다가 큰 데이터셋에서 pretrain된 다음, task별로 fine-tuning하는 방법을 사용하기 시작했는데, 대표적인 예가 BERT, GPT

- BERT는 denoising self-supervised pre-training task를 사용
- GPT 계열은 pre-training task로서 language modeling을 사용
- 라고 논문에 나오긴 했는데, BERT와 GPT의 차이는 encoder를 쓰는지(BERT)와 decoder를 쓰는지(GPT)인 걸로 알고 있음

Related work

2.2. attention을 이미지에 적용

attention을 단순히 이미지에 활용하는 것은 픽셀이 다른 픽셀에 attention을 수행하는 것을 생각할 수 있다. 그렇게 되면, 픽셀 당 계산해야 하는 수가 너무 커지므로, 실제 입력 크기로 확장할 수 없게 된다. 따라서 이미지와 관련해서 Transformer를 적용하기 위해 과거에 몇가지 비슷한 실험이 시도됨.

2.2.1. 관련 실험들

- Image Transformer: 전역이 아닌 각 쿼리 픽셀에 대해 로컬한 이웃에만 self-attention을 수행.
- local multi head dot-product self-attention block은 convolution block을 완전히 대체한 실험
- Sparse Transformer라는 기술은 이미지에 적용할 수 있도록 global self-attention에 대한 확장 가능한 근사를 사용함.
- Attention을 확장하는 또 다른 방법으로 다양한 크기의 블록에 적용하는 방법을 제안한 실험

이러한 attention architecture를 특수하게 변경하는 방법 중 다수는 컴퓨터 비전 작업에서 유망한 결과를 보여주지만 하드웨어 가속기에서 효율적으로 구현하려면 복잡한 엔지니어링이 필요함

이 논문과 가장 관련이 있는 모델은 Cordonnier가 쓴 모델(On the Relationship between Self-Attention and Convolutional Layers), 입력 이미지에서 크기 2x2의 패치를 추출, 그 위에 full self-attention을 적용함. 하지만, 이 논문은 CNN을 활용한 이미지 task보다 성능이 낫다는 걸 보여주지 못함. ViT 더 나아가 대규모 pretrain을 통한 vanilla transformer를 최신 CNN과 경쟁할 수 있도록(혹은 그 이상) 만든다는 것을 보여줄 수 있다. 또한 위의 논문은 2x2픽셀의 작은 patch size를 사용해 모델을 작은 해상도 이미지에만 적용할 수 있지만, ViT는 중간 해상도 이미지도 처리할 수 있음

Related work

2.3. CNN + self-attention

CNN과 self-attention의 형태를 결합하는데 여러 관심들이 있었음. 이미지 분류를 위해 feature map을 보강하거나(Bello, 2019), self-attention을 사용해 CNN의 출력을 추가로 처리. 객체 탐지(Hu, 2018; Carion, 2020), 비디오 처리(Wang, 2018,...), 이미지 분류(Wu et al., 2020), unsupervised 물체 탐지(Locatello et al., 2020), or 통합된 text-vision tasks(Chen et al., 2020c; Lu et al., 2019; Li et al., 2019).

또 다른 관련 최신 모델은 이미지 해상도와 color space을 줄인 후 이미지 픽셀에 Transformer를 적용하는 **image GPT**(iGPT, 2020)이다. 이 모델은 생성 모델로서 unsupervised로 훈련되며 결과 representation은 분류 성능을 위해 fine-tuning되거나 linear하게 probe되어 **ImageNet에서 최대 72%의 정확도**를 달성.

2.4. larger dataset을 학습해서 성능 향상한 모델들

아래 논문들은 표준 image dataset이 아닌, 더 큰 dataset을 학습해서 성능 향상을 기록한 모델들이다.

추가 데이터 소스를 사용하면 benchmark에서 SOTA를 얻을 수 있는 실험들도 있었음(Mahajan et al., 2018; Touvron et al., 2019; Xie et al., 2020). Moreover, Sun et al. (2017))

또한 Sun(2017)은 CNN 성능이 데이터 셋 크기에 따라 어떻게 확장되는지 연구하고 Djolonga(2020)은 ImageNet-21k 및 JFT-300M과 같은 대규모 데이터 셋에서 CNN transfer learning의 경험적 탐색을 수행.

저자는 이러한 larger 데이터 셋들에도 초점을 맞추지만 위에서 사용된 ResNet 기반 모델 대신 이 논문은 Transformer를 학습한다.