Swin Transformer

☑ 복습 ☐

swin.pdf

❓ **1) Swin Transformer: Hierarchical Vision Transformer using Shifted Windows**

주요 내용: 컴퓨터 비전을 위한 범용 백본(backbone)으로서 기능하는 새로운 Vision Transformer, Swin Transformer. 해당 모델은 계층적 구조를 특징으로 하며, 여러 크기의 비주얼 엔티티와 이미지의 고해상도라는 시각과 언어 사이의 도메인 차이를 극복하기 위함

2) Swin Transformer V2: Scaling Up Capacity and Resolution

주요 내용: Swin Transformer의 성능을 확장하기 위한 연구로, 더 큰 모델 용량과 높은 해상도에서의 성능 증가가 목표이다. 해당 논문은 Vision Task에서 대규모 모델을 탐구하며 훈련 과정에서의 세 가지 주요 문제를 해결한다.

3) Performance Evaluation of Swin Vision Transformer Model using Gradient Accumulation Optimization Technique

주요 내용: Swin Transformer 모델의 성능을 평가하고, 그래디언트 누적 최적화 기술을 사용하여 모델 성능을 향상시키는 방안을 제시하는 연구

주요 방법론

- 계층적 구조: 이미지의 다양한 스케일을 효과적으로 처리하기 위해 계층적 구조를 도입
- modified self-attention computation (Swin Transformer blocks)을 적용
  → local 윈도우 내에서 self-attention computation을 진행하고, 이미지를 겹치지 않는 패치로 분할하여 계층적인 특징 맵을 구축
  이를 통해 효율성을 높이며, Window 간 상호 작용을 가능하게 하여 전체적인 관계를 학습할 수 있게 한다.

논문에서 풀고자 하는 문제

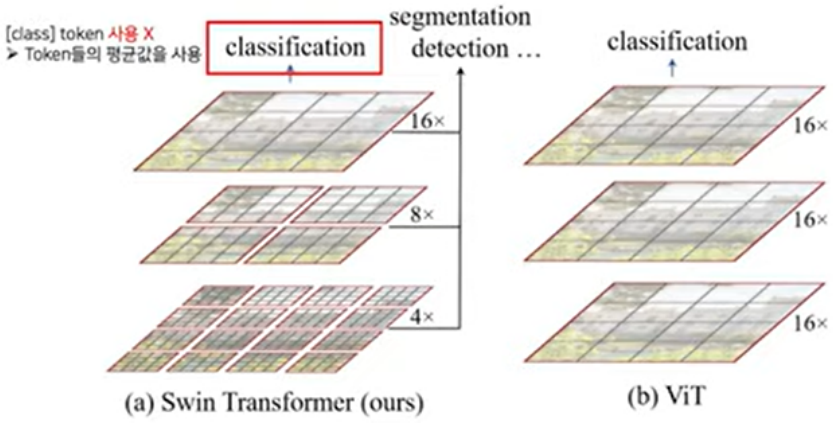
언어에서 시작된 트랜스포머를 비전 분야로 전환하는 과정에서 발생하는 문제 해결

- 이미지 내에서 다양한 크기와 스케일을 가지는 시각적 객체에 의한 큰 변화 문제
- 이미지의 픽셀 해상도가 텍스트의 단어와 비교했을 때 높은 차이가 발생하는 문제

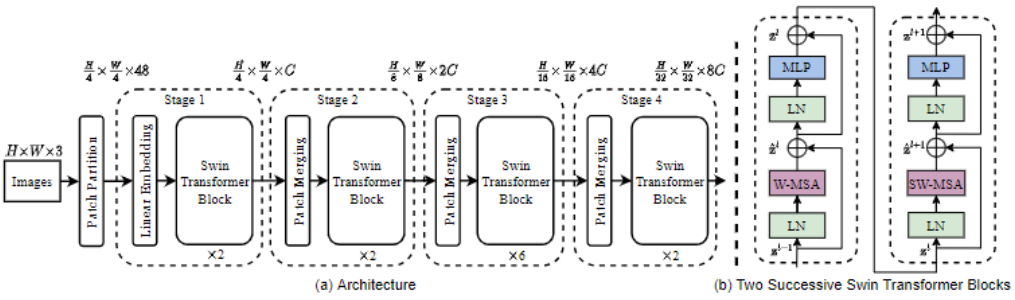
기존 연구들의 접근 방식

기존 연구들은 주로 CNN(Convolutional Neural Network) 기반의 아키텍처를 사용하여 시각적 특징을 추출하였으나, 이는 이미지의 전체적인 context를 파악하는데 한계가 있다.

이 논문의 강점



- 효율적인 계산: Swin Transformer는 self-attention computation을 local로 제한하고, 그 효율을 증대시킨다.
- 확장성: 이 아키텍처는 다양한 크기와 해상도의 이미지에 적용 가능하여 모델의 범용성을 높인다.
  → 선형 계산 복잡성을 유지하면서도 밀도 있는 예측을 가능하게 하며,
  다양한 비전 작업에 적용 할 수 있는 일반적인 백본 모델
- 성능 향상: 계층적 구조와 Shifted Windows 기법을 통해 기존 방법보다 상당한 성능 향상 달성



모델 핵심 개념

1. Patch Merging
   각 Stage의 첫 번째 레이어에서, 인접한 Patch를 합쳐 feature dimension을 늘리는 방식으로 작동한다. 동시에, 이 과정은 spatial dimension을 줄여 계산 효율성을 높인다.

- Patch Partition: 입력 이미지를 4x4의 Patch로 나눈다. 해당 과정은 이미지를 고정된 크기의 Patch로 분할하여 Transformer에 적합한 입력 형태로 변환하는 역할을 한다.
- Linear Embedding: 각 Patch를 1D 벡터로 평탄화하고, Linear Layer를 통해 차원을 변환한다. 해당 단계에서, 인접한 Patch를 결합하여 하나의 큰 Patch로 만들거나(즉, Merging), 차원을 증가시키는 역할을 하여 더 높은 수준의 특징을 포착할 수 있게 한다.

→ 이미지 내의 다양한 크기의 객체를 효과적으로 처리하고, 계산 효율성을 높이며, 서로 다른 스케일에서의 특징을 학습할 수 있는 계층적 구조를 구축

2. Swin Transformer Block

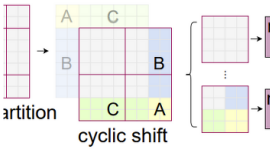
- Shifted Window 기반의 Self-Attention: 이는 고정 크기의 Window 내에서만 Self-Attention을 계산함으로써, 더 효율적인 계산을 가능하게 합니다.
- Swin Transformer는 기본 Window와 Shifted Window 방식을 번갈아가며 사용하여, Window 간의 정보 교환을 촉진하고 더 넓은 범위의 컨텍스트를 커버합니다.
- Layer Normalization: 각 Sub-layer의 입력 앞에 적용되며, 모델의 안정성과 학습 속도를 향상
- W-MSA(Window based Multihead Self Attention): 이미지를 고정된 크기의 여러 개의 window로 나누고, 각 window 내에서 Multihead Self Attention을 계산
  → local pattern과 특징을 효과적으로 학습

- SW-MSA(Shifted Window based Multihead Self Attention): 전체 이미지를 window로 나눈 뒤, 각 window의 위치를 일정량 shift(이동)시킨 후 Multihead Self Attention을 계산
  → 경계에서 정보가 격리되는 것을 방지하고, 인접 window 간의 통합된 context 정보를 활용 가능
- MLP (Multi-Layer Perceptron): Non-linear 변환을 수행하는 데 사용되며, 레이어 간의 비선형 결합을 가능하게 한다.
- Skip Connection: 입력을 출력에 직접 더하는 구조로, 모델의 깊이가 깊어질수록 발생할 수 있는 학습 문제를 완화

[Swin Transformer 논문 리뷰] - Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

*Swin Transformer 논문 리뷰를 위한 글이고, 질문이 있으시다면 언제든지 댓글로 남겨주세요! Swin Transformer 논문: [2103.14030] Swin Transformer: Hierarchical Vision Transformer using Shifted Windows (arxiv.org) Swin Transformer: Hierarchical Vision Transformer using Shifted Windows This paper presents a new vision Transformer, called Swin

https://kyujinpy.tistory.com/14



Future Work

- 모델 구조와 학습 방법의 확장: 대용량 데이터셋과 더 큰 모델을 활용하여 성능을 더욱 향상시키기
- 새로운 domain 및 task로의 적용: Swin Transformer 아키텍처를 vision task 외에 다른 domain에 적용하여 활용 범위를 넓혀보기
- 결합 모델: CNN과 Transformer를 결합한 모델로, 각각의 장점을 활용하여 더 효과적인 이미지 인식 모델을 개발하는 연구가 진행 가능