졸업논문
소개

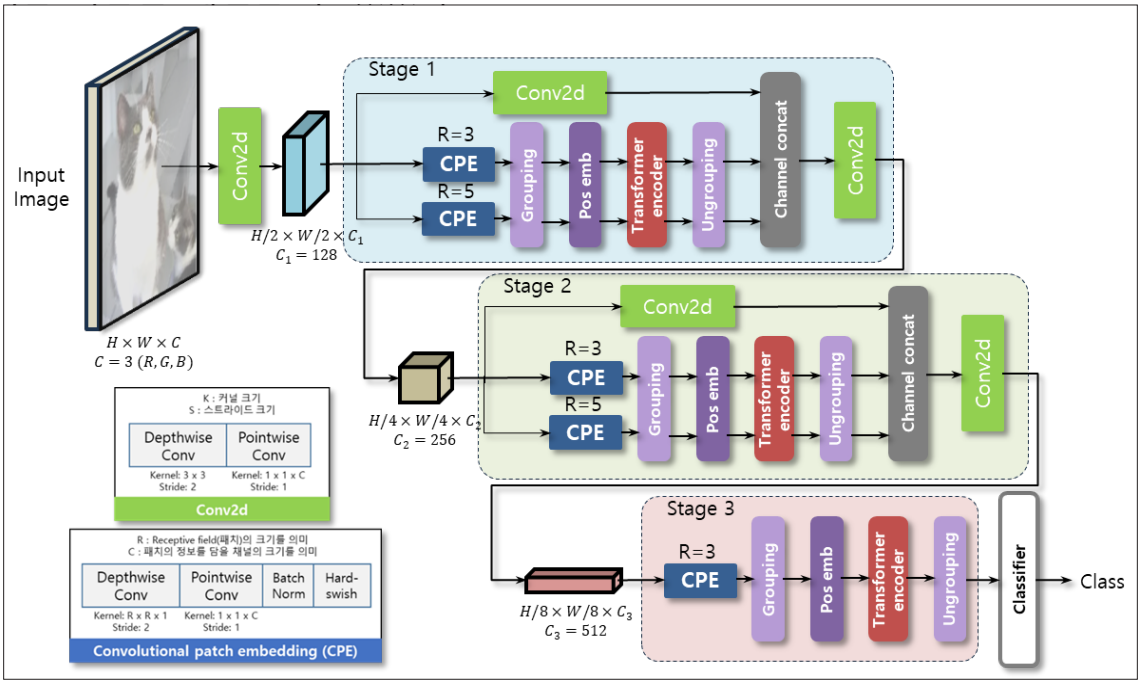
스케일 불변 다중 경로 및 계층 구조 기반 비전 트랜스포머를 활용한 이미지 분류 정확도 개선

이세영 / 광운대학교 영상처리시스템연구실(IPSL)

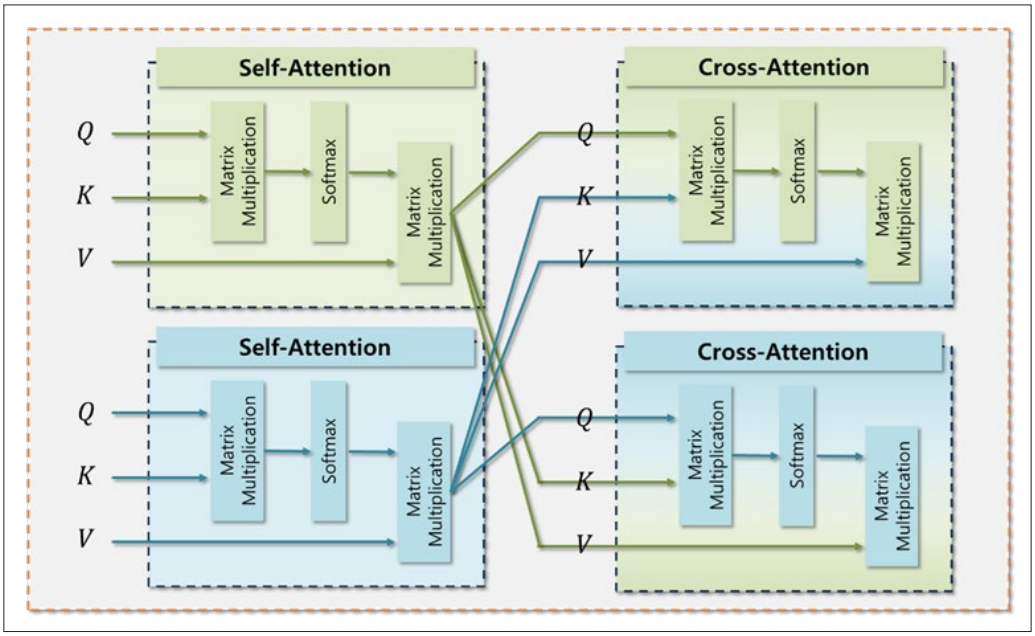
이미지 분류, 객체 인식, 의미적 분할 등 다양한 컴퓨터 비전 분야의 응용에서는 입력 영상에 대한 특징 추출 방법에 따라 성능 차이가 크게 발생한다. 특히 실제 환경에서 취득되는 영상 내 객체의 크기와 종횡비는 매우 다양할 수 있으며, 객체 스케일의 불균일성은 모델 성능 저하의 주요 원인으로 작용한다. 본 논문에서는 객체 스케일이 불균일한 영상에 대해서도 강인한 모델 성능을 보일 수 있는 특징 추출 방법을 제시한다.

본 논문에서 제안하는 〈그림 1〉의 스케일 불변 비전 트랜스포머 구조는 합성곱 기반 패치 임베딩 방법을 활용한 다중 경로의 특징 추출 방식을 사용하며, 다중 경로 간 크로스 어텐션 메커니즘을 통해 저수준의 특징 추출 단계에서 강인한 특성을 갖춘다. 그러나 비전 트랜스포머가 활용하는 어텐션 연산은 많은 연산량을 수반한다는 단점이 있다. 따라서 본 논문에서는 사전 정의된 윈도우 단위로 어텐션 연산을 수행하는 계층 구조를 결합함으로써 네트워크 복잡도를 줄였다. 또한, 〈그림 1〉의 제안하는 방법에서 다중 경로 간 특징 표현의 효과적인 통합을 위해 〈그림 2〉의 크로스 어텐션 메커니즘을 적용하였다. 셀프 어텐션 연산을 수행한 각 경로의 특징 맵은 같은 경로에 위치한 크로스 어텐션의 Query를 생성하는데 사용되며, 다른 경로의 특징 맵은 크로스 어텐션의 Key와 Value를 생성하는데 활용된다. 크로스 어텐션 구조를 통해, 서로 다른 수용 영역에서 추출된 다양한 스케일의 문맥 특징의 관계성을 각 경로에서 유기적으로 학습하여 복잡한 시각 패턴에 대한 학습을 개선할 수 있었다.

본 논문에서는 컴퓨터 비전 작업의 하나의 응용 예시인 이미지 분류 벤치마크 ImageNet과 가공 데이터셋을 사용하여, 제안하는 방법의 정확도 개선 효과를 평가하였다. ImageNet 데이터셋 중 'sliding door'와 'pole' 클래스의 객체들과 같이 종횡비가 크거나 객체의 경계가 모호한 이미지에서 모델 성능이 하락하는 것을 고려하여 ImageNet 데이터셋으로부터 가공 데이터셋을 생성하였다. ImageNet 데이터셋의 이미지 내 객체의 비율과 그 크기를 다양화하기 위해 원본 이미지에 대해 공간 영역 다운샘플링을 수행하고, 보간 과정을 통해 네트워크의 입력 텐서 크기로 조정하였다. 또한, 다운샘플링 과정에서 객체가 제외되는 경우를 막기 위해, 다운샘플링 후의 이미지 최소 사이즈는 원본 이미지의 4분의 3 크기로 제한하고, 안티 에일리어싱 필터를 적용하였으며, 보간 필터로



<그림 1> 제안하는 스케일 불변 비전 트랜스포머 구조



<그림 2> 제안하는 방법에서 사용한 크로스 어텐션 메커니즘

는 DCT 보간 필터를 사용하였다. 즉, 직접 가공한 데이터셋은 원본 ImageNet 데이터셋보다 객체의 스케일 다양성이 증가하고, 객체의 경계가 상대적으로 불명확해진 이미지들로 구성되었다.

계층 구조 기반 비전 트랜스포머 중 하나인 Nest-B 모델을 참조 네트워크로 활용하여, 다중 경로 구조로 확장 적용한 제안 방법 1과 크로스 어텐션을 추가한 제안 방법 2에 대해 실험한 결과, 제안 방법 2, 제안 방법 1, Nest-B, Swin-B, CVT-13, PVT, DeiT-B, ViT-B/16, ResNet-50 순으로 분류 정확도가 높게 나타났다. 제안 방법 1은 Nest-B 대비 분류 정확도가 0.2%p만큼 향상되었으며, 제안 방법 2는 약 0.5%p만큼 향상되었다. 제안 방법 1, 2는 참조 네트워크인 Nest-B보다 네트워크 파라미터 수가 각각 약 8M, 약 14M만큼 증가하였으나, Swin-B 또는 비계층 구조 기반 비전 트랜스포머인 ViT-B/16과 DeiT-B보다 파라미터 수가 약 4M~6M 이상 적었다. 또한, 선행 연구와 제안 방법을 ImageNet-1k 데이터셋에 대하여 사전 학습한 뒤, 가공 데이터셋에 대하여 전이 학습을 수행하고 실험한 결과, 가공 데이터셋에 대해서도 제안 방법 2는 참조 네트워크인 Nest-B 대비 여전히 약 0.4%p만큼 더 높은 분류 정확도를 보였다.

본 논문에서는 스케일 적응성을 개선한 다중 경로 및 계층 구조 기반 비전 트랜스포머를 제안하여, 이미지 분류, 객체 탐지 등을 비롯한 다양한 컴퓨터 비전 작업에서 활용될 수 있을 것으로 기대된다.

이 세 영

- 2022년 8월 : 광운대학교 컴퓨터공학부 학사
- 2025년 2월 : 광운대학교 컴퓨터공학과 석사
- 2025년 3월 ~ 현재 : (주)LG전자 CTO부문 SoC센터
- 주관심분야 : 영상처리, 비디오코덱, 디지털IP설계