복소수 위치 임베딩을 적용한 비전 트랜스포머 활용 이미지 분류

김한영[1], 조영준[2]
[1]전남대학교 인공지능융합학과 석사과정
[2]전남대학교 인공지능융합학과 부교수
codebyhy@gmail.com, yj.cho@jnu.ac.kr

Image classification using vision transformers with complex positional embeddings

Han-Young Kim[1], Yeong-Jun Cho[1]
[1]Dept. of Artificial Intelligence Convergence, Chonnam National University

요 약

본 연구에서는 Complex Order Position Embedding (COPE)을 Vision Transformer (ViT)에 적용하여 컴퓨터 비전 태스크에서의 효과성을 검증하였다. COPE는 복소수 연산을 활용하여 위치 정보를 인코딩하는 방법으로, 기존에 자연어 처리 분야에서 성공적으로 적용된 바 있다. ImageNet-Tiny 데이터셋을 사용한 실험에서, COPE를 적용한 ViT-Tiny 모델은 기존 모델 대비 1.8%p 높은 34.0%의 정확도를 달성하였다. 이는 파라미터 수의 미미한 증가(약 37,000개)만으로 이루어진 성능 향상이다. 본 연구 결과는 COPE가 컴퓨터 비전 분야에서도 효과적임을 입증하며, 특히 객체 검출이나 의미론적 분할과 같이 위치 정보가 중요한 고난도 비전 태스크에서의 잠재적 성능 향상 가능성을 제시한다. 이는 복소수 위치 임베딩의 응용 범위를 확장하고, 트랜스포머 기반 비전 모델의 성능 개선을 위한 새로운 방향을 제시한다는 점에서 의의가 있다.

.

1. 서론

자연어 처리(NLP) 분야에서 트랜스포머[1] 계열 모델의 등장은 큰 변화를 가져왔다. 트랜스포머 등장 이전에는 문장 단위의 입력을 처리하기 위해 각 단어를 순차적으로 인코딩하는 순환 신경망(RNN)[2] 계열의 모델이 사용되었으나, attention 메커니즘을 기반으로 하는 트랜스포머의 등장으로 기존 모델보다 월등한 성능을 보였다. Attention 메커니즘은 모든 입력을 한 번에 처리하면서 순열 불변성을 가진다. 이 문제를 해결하기 위해 위치 임베딩(PE)이 도입되었다.

비전 분야에서도 트랜스포머 모델이 큰 주목을 받고 있다. 비전 트랜스포머(ViT)[3]의 등장 이전에는 주로 합성곱 신경망(CNN) 계열의 모델이 이미지 처리에 사용되었다. CNN[4]은 지역화된 필터를 통해 이미지의 공간적 계층 구조를 효과적으로 포착한다. 반면, ViT는 self-attention 메커니즘을 활용하여 이미지 전체에 걸친 장거리 의존성을 모델링 한다. 이러한 특성으로 인해 ViT는 지역적 특성에 국한되지 않고 이미지의 전역적 맥락을 더 효과적으로 파악할 수 있다. 결과적으로 ViT는 기존 CNN 모델들과는 차별화된 방식으로 이미지를 처리하며, 다양한 비전 태스크에서 우수한 성능을 보여주고 있다.

최근에는 기존의 위치 임베딩 방식의 한계를 극복하기 위한 더 정교한 접근 방식이 탐구되고 있다. 복소수 위치 임베딩(Complex Order Position Embedding, COPE)[5]이 그 예이다. 복소수 연산을 활용한 COPE는 자연어 처리 분야에서 처음 사용되었는데, 단어 임베딩(WE)과 위치 임베딩(PE)를 각각 실수부 허수부로 분할하여 모델링 하는 방식이다.

2. 관련 연구들

2.1 Transformer, ViT

트랜스포머는 기존의 RNN 계열 모델들과 달리

모든 입력을 동시에 처리할 수 있는 self-attention 메커니즘을 도입하여 큰 성과를 거두었다. Vision Transformer(ViT)은 컴퓨터 비전 분야에서 트랜스포머 모델을 적용하였다. ViT는 self-attention을 사용하여 이미지를 처리하며, 전통적인 CNN과 달리 전역적인 문맥을 효과적으로 캡처할 수 있다. CNN은 국소적 특징을 학습하는 데 유리한 inductive bias를 가지며, 이는 이미지와 같은 데이터에서 중요한 패턴을 학습하는 데 도움이 된다. 반면, ViT는 이러한 국소적 제한 없이 모든 토큰 간의 상호작용을 학습할 수 있어 더 유연하고 강력한 전역 패턴 학습이 가능하다.

2.2. Absolute/Relative Position Embedding

Absolute Positional Embedding(APE)은 입력 시퀀스의 각 위치에 고유한 embedding vector를 할당한다. 이는 sin과 cos 함수를 이용해 위치를 인코딩하는 방법이 일반적이다. 트랜스포머은 이러한 방식으로 입력 토큰의 순서를 모델에 전달한다.

Relative Positional Embedding(RPE)[6]은 입력 시퀀스의 상대적 위치 정보를 인코딩한다. RPE는 각 quary와 key 간의 상대적 위치를 고려하는 임베딩을 제안하여, 보다 유연하고 일반화된 위치 정보를 제공할 수 있었다. 이는 특히 긴 시퀀스나 다양한 길이의 입력에서 효과적이다.

APE의 장점은 구현이 간단하고, 모델이 절대 위치 정보를 명확하게 인식할 수 있도록 돕는다는 것이다. 그러나 단점으로는 입력 시퀀스의 길이가 달라지면 재학습이 필요하고, 고정된 위치 정보가 모델의 유연성을 제한할 수 있다는 점이 있다. 반면, RPE의 장점은 상대적인 위치 정보를 사용하여 다양한 길이의 시퀀스에서 더 유연하게 작동할 수 있으며, 긴 문맥을 더 효과적으로 처리할 수 있다는 것이다. 단점으로는 구현이 복잡하고, 절대 위치 정보를 명확하게 제공하지 않기 때문에 일부 응용에서 성능이 떨어질 수 있다는 점이 있다.

2.3. Complex Order Position Embedding(COPE)

Complex Order Position Embedding(COPE)은 NLP 분야에서 단어의 순서를 복소수(complex number) 임베딩으로 인코딩하는 방법으로, 앞서 언급한 방식 중 APE에 일종이다. COPE는 위치 정보의 표현력을 향상시키며, 이를 통해 더 정확한 의미 파악이 가능하다. 복소수 임베딩은 단어의 위치를 복소수 함수로 표현함으로써 단어의 순서와 그 관계를 모델링 한다. 이는 단어의 위치가 증가함에 따라 단어 표현이 부드럽게 이동할 수 있도록 하여, 서로 다른 위치에 있는 단어 표현이 연속적인 함수로 상호 연관될 수 있게 한다. 이 방법은 CNN, RNN 및 트랜스포머 모델에 적용할 수 있으며, 이를 통해 텍스트 분류, 기계 번역 및 언어 모델링에서 성능 향상을 보여주었다.

수식은 다음과 같다. r(진폭)과 w(주파수, 주기의 역수), θ(초기 위상)를 학습 파라미터로 설정하여 z를

$$z = r \circ exp(i\theta) = r(\cos\theta + \sin\theta) \quad (1)$$

이와 같이 모델링한다. 이를 통해 위치 임베딩 함수 $g(pos)$를 아래와 같이 정의한다.

$$g(pos) = z_2 z_1^{pos} = r_2 e^{i\theta_2}(r_1 e^{i\theta_1})^{pos} = r_2 r_1^{pos} e^{i(\theta_2 + \theta_1 pos)}. \quad (2)$$

D차원의 벡터의 각 성분의 인덱스를 j라고 하고, 이를 오일러 공식을 통해

$$g_{pe(j,pos)} = [e^{i(w_{j,1}pos+\theta_{j,1})}, \cdots, e^{i(w_{D,1}pos+\theta_{D,1})}] \quad (3)$$

으로 정의할 수 있다. 이를 통해 복소수 위치 임베딩을 구현하여 사용한다.

3. 제안 방법론

COPE을 ViT에 적용한다. 이를 위해, WE과 Attention block 사이의 PE를 복소수 위치 임베딩으로 변경한다. (3)의 수식에서 pos 값은 트랜스포머에서 사용하는 삼각함수 테이블을 그대로 사용하고, ViT에서 사용하는 클레스 토큰도 그대로 적용한다.

4. 실험 및 결과

실험은 COPE가 비전 과업에서도 좋은 영향을 주는지 검증하는 방향으로 진행하였다. 특히 가장 기본적인 과업인 이미지 분류에서의 성능을 검증하기 위해, ImageNet-Tiny[7] 데이터셋을 사용하였다. 또한 사용 모델은 ViT 중에서 가장 가벼운 모델인 ViT_Tiny를 사용하였다.

학습 가능한 파라미터를 사용한 기존 모델과 COPE를 적용한 모델을 비교한다.

<표 1> ImageNet-Tiny 검증 결과

	정확도	파라미터 수
기본 PE	0.322	5,563,016
COPE	0.340	5,600,840

학습 가능한 파라미터를 사용한 기본 ViT-Tiny는 0.322의 정확도, COPE를 적용한 모델은 0.340의 정확도를 보인다. 또한 각 모델의 파라미터 수는 약 37,000개의 차이를 보인다(<표 1> 참조).

5. 결론

본 연구에서는 Complex Order Position Embedding(COPE)을 Vision Transformer(ViT)에 적용하여 이미지 분류 태스크에서의 성능을 평가하였다. 실험 결과, COPE를 적용한 ViT 모델이 기존의 위치 임베딩을 사용한 모델보다 향상된 성능을 보여주었다. COPE를 적용한 ViT-Tiny 모델은 ImageNet-Tiny 데이터셋에서 기존 방식 대비 높은 정확도를 달성하였다. 이는 COPE가 비전 과업에서도 효과적임을 입증한다. 또한, COPE 적용으로 인한 파라미터 증가는 약 37,000개로, 전체 모델 크기에 비해 미미한 수준이다. 이는 COPE가 적은 추가 비용으로 성능 향상을 이끌어낼 수 있음을 시사한다.

본 연구의 결과는 원래 자연어 처리를 위해 개발된 COPE가 컴퓨터 비전 태스크에서도 효과적임을 보여줌으로써, 복소수 위치 임베딩의 응용 범위가 확장될 수 있음을 확인하였다. 이는 COPE가 다양한 분야에서 활용될 수 있는 가능성을 제시한다.

마지막으로, 위치 정보가 더욱 중요한 역할을 하는 객체 검출이나 의미론적 분할과 같은 고난도 비전 태스크에서 COPE의 적용은 더욱 현저한 성능 향상을 가져올 것으로 기대된다. 이는 COPE가 제공하는 정교한 위치 정보 인코딩이 이미지 내 객체의 공간적 관계와 세밀한 경계를 포착하는 데 특히 유용할 것으로 예상되기 때문이다.

감사의 글

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 인공지능융합혁신인재양성사업(IITP-2023-RS-2023-00256629) 및 대학ICT연구센터사업(IITP-2024-RS-2024-00437718)의 연구결과로 수행되었음

참고문헌

[1] Vaswani, A. "Attention is all you need." Advances in Neural Information Processing Systems (2017).

[2] Sherstinsky, Alex. "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network." Physica D: Nonlinear Phenomena 404 (2020): 132306.

[3] Dosovitskiy, Alexey. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

[4] O'Shea, K. "An introduction to convolutional neural networks." arXiv preprint arXiv:1511.08458 (2015).

[5] Wang, Benyou, et al. "Encoding word order in complex embeddings." arXiv preprint arXiv:1912.12333 (2019).

[6] Shaw, Peter, Jakob Uszkoreit, and Ashish Vaswani. "Self-attention with relative position representations." arXiv preprint arXiv:1803.02155 (2018).

[7] Le, Ya, and Xuan Yang. "Tiny imagenet visual recognition challenge." CS 231N 7.7 (2015): 3.