# 이산 푸리에 변환을 이용한 패치 단위의 위치 임베딩 적용

송승헌, 이재구*
국민대학교

*jaekoo@kookmin.ac.kr

# Application of Patch-based Position Embedding Using Discrete Fourier Transform

Seungheon Song, Jaekoo Lee*

College of Computer Science, Kookmin University

요 약

최근 컴퓨터 비전에서 유도편향을 완화하고 패치 단위의 입력을 사용한 트랜스포머와 같은 모델들은 다양한 과업에서 주목할만한 성능을 보인다. 하지만 영상에서 패치의 사용은 원본 데이터를 분할함으로써 모델이 패치 간의 위치 관계 정보 손실이 있고, 해당 문제를 해결하기 위해 위치 인코딩 기법이 적용되었다. 우리 방법은 기존 방법과 달리 패치 단위에서 푸리에 임베딩을 이용하여 위치 관계를 주입하는 방법을 제안한다. 패치 단위 입력에서 푸리에 임베딩에서 주파수 신호는 위치정보를 추가하고 단위 입력 간의 자연스러운 결합을 유도할 수 있다. 결과적으로 패치 단위의 입력을 사용하는 MLP-Mixer에 푸리에 임베딩을 적용하였을 때 영상 분류 정확성이 최대 1.23% 증가함을 확인할 수 있었다. 또한, 푸리에 임베딩을 사용하여 픽셀 단위의 입력을 받는 Perceiver에서 입력을 패치 단위로 실험하였을 때, 영상 분류 과업에서 정확성이 최대 82.75% 향상됨을 확인할 수 있었다.

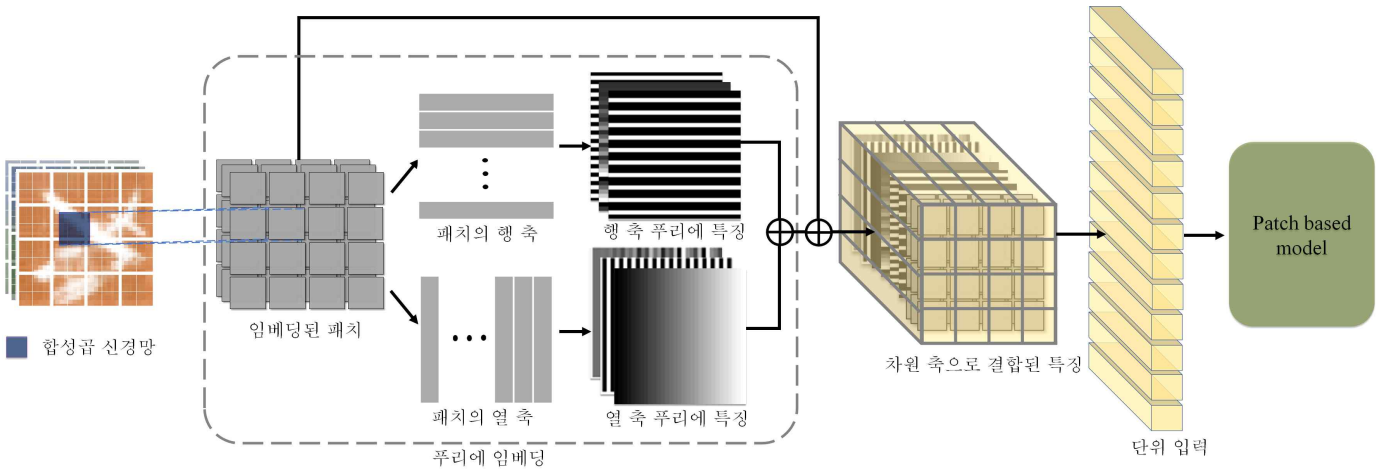


그림 1.
원본 영상에 패치 임베딩과 푸리에 임베딩을 적용하여 모델의 입력으로 전달하는 과정이다. 합성곱 신경망을 이용하여 패치 임베딩을 진행하고, 푸리에 임베딩 과정에서 행 축과 열 축의 주파수 신호를 추출한 후 패치 임베딩과 결합하여 모델에 단위 입력으로 들어가게 된다.

## Ⅰ. 서 론

심층학습(deep learning)에서 패치 임베딩(patch embedding)을 적용하여 단위 입력(token)을 사용하는 최신 모델들은 점차 다양한 과업에서 사용되고 있다[1]. 특히 자연어 처리 과업에서 단위 입력을 입력으로 받는 어텐션 메커니즘(attention mechanism)[1]을 이용한 모델들이 좋은 성능을 보여주었다. 컴퓨터 비전 영역에서도 성능향상을 위해 패치 임베딩을 사용한 단위 입력을 받음으로써 어텐션 메커니즘의 사용이 가능해졌다[1]. 추가로 영상에서의 패치 임베딩은 유도 편향(inductive bias)인 지역성 정보라는 추가적인 정보를 주어 모델의 학습을 윤활하게 한다.

패치 임베딩은 위치 또는 순서 정보를 명시적으로 인코딩하지 않고, 모델에서 독립적으로 계산되기 때문에 순서 정보를 담고 있을 수 있으므로 모델의 성능을 크게 좌우한다[1]. 따라서 위치 인코딩(positional encoding)[1]을 사용하여 패치에 지역성 정보를 넣어주는 것이 일반적이다. 본 논문에서는 패치 단위의 입력을 받는 모델에 위치정보를 주입하고, 패치 간의 자연스러운 결합을 위해 푸리에 임베딩(fourier embedding)을 이용한 위치 인코딩을 연구한다.

## Ⅱ. 본 론

패치 단위의 입력을 사용하는 모델에서 위치정보를 모델에 학습시키기 위해 사용하는 다양한 인코딩 방법이 있다[1, 2]. 기존 방식은 정현파를 이용하여 패치끼리의 상대적인 위치정보를 모델에 전달하거나, 학습가능한 파라미터를 이용하여 위치정보를 학습하도록 하는 방법이 있다[1].

푸리에 임베딩은 주파수 차원에서 입력 데이터를 주파수 신호로 표현할 수 있게 한다[3]. 해당 특성은 위치정보와 주파수 신호를 모델에 입력하는 역할을 하게 된다. 푸리에 임베딩은 입력의 낮은 주파수부터 높은 주파수까지의 조합의 특성을 모델에 반영할 수 있게 해주고, 이를 통해 추가적인 정보를 학습하도록 한다[3]. 본 논문에서는 [그림 1]과 같이 푸리에 임베딩을 패치 단위의 학습에 적용하여 모델이 주파수 신호와 위치정보를 동시에 학습할 수 있도록 한다. 입력

| 패치크기 | MLP-Mixer[4] | | Perceiver[5] | |
|---|---|---|---|---|
| | CIFAR-10[6] | SVHN[7] | CIFAR-10[6] | SVHN[7] |
| 1 | 77.88 | 94.87 | 48.08 | 19.59 |
| 1 with FE | 79.50 (+ 2.08%) | 94.92 (+0.05%) | 64.53 (+34.21%) | 91.36 (+366.36%) |
| 2 | 87.56 | 96.39 | 63.21 | 52.63 |
| 2 with FE | **88.64 (+1.23%)** | 95.99 (-0.42%) | 72.35 (+14.46%) | **96.18 (+82.75%)** |
| 4 | 90.06 | 96.85 | 60.51 | 79.26 |
| 4 with FE | 90.56 (+0.56%) | 96.88 (+0.58%) | **73.22 (+21.00%)** | 97.04 (+22.43%) |
| 8 | 86.07 | 95.77 | 50.76 | 91.01 |
| 8 with FE | 86.54 (+0.55%) | **96.33 (+0.56)** | 62.93 (+0.58%) | 94.65 (+4.00%) |
| 16 | 72.62 | 95.47 | 61.83 | 91.27 |
| 16 with FE | 73.14 (+0.72%) | 95.59 (+0.13%) | 62.19 (+0.36) | 93.99 (+2.98%) |

표 1. 패치 단위의 입력에서 푸리에 임베딩을 적용했을 때 분류성능을 알아보기 위한 정확성 측정 실험. 푸리에 임베딩 사용 여부를 표기하기 위해 푸리에 임베딩을 적용했을 때, FE(Fourier Embedding)을 표기한다. 가장 높은 성능변화를 가진 결과를 굵은 글씨로 표시한다. 증감율은 푸리에 임베딩을 사용하지 않았을 때와의 비교이며 괄호 안에 표기한다.

영상에 대해 패치별로 값을 뽑아내기 위해 합성곱 신경망을 이용한다. 그 후 패치 단위에서 주파수 신호를 추출하기 위해 임베딩된 패치들을 행 축과 열 축으로 분해하여 푸리에 임베딩한다. 입력 영상에서 열 축의 크기가 $M$이고, 행 축의 크기가 $N$인 영상에서의 푸리에 임베딩은 [식 1]과 같다.

$$F(u,v) = \frac{1}{NM}\sum_{x=0}^{N-1}\sum_{y=0}^{M-1} f(x,y)e^{-2\pi i(\frac{ux}{N}+\frac{vy}{M})} \quad (1)$$

[식 1]에서 $x$와 $y$는 원본 영상의 픽셀값을 의미하고, $u$와 $v$는 주파수 신호의 차원을 의미한다. 행 축과 열 축의 푸리에 임베딩은 $x$ 혹은 $y$번째에 해당하는 열과 행을 각각 1차원 데이터로 간주하여 $u$와 $v$ 차원으로 변환한다.

[식 1]을 통해 구한 행 주파수 신호와 열 주파수 신호는 [그림 1]과 같이 임베딩된 패치들과 결합하여 하나의 특징행렬을 만든다. 이렇게 만들어진 특징행렬을 단위 입력으로 분해한 후 모델의 입력으로 사용한다.

## Ⅲ. 실 험

본 실험에서는 푸리에 임베딩 적용 여부와 패치 크기에 따른 분류성능을 교차 실험한다. 푸리에 임베딩은 최대 주파수 112에서 64개의 주파수 대역을 가지는 푸리에 임베딩을 사용한다. 실험에서 패치 개수에 따른 계산 비용이 효율적인 모델인 MLP-Mixer[3]를 사용한다. 추가로 푸리에 임베딩을 사용하여 픽셀 단위의 입력을 받는 Perceiver[4]에서 입력을 패치 단위로 확장하여 실험한다. [그림 2]에서 볼 수 있듯이 패치 단위의 입력을 사용하는 MLP-Mixer의 경우 비용이 지수적으로 증가하는 어텐션 메커니즘 대신 패치별 채널별로 분리하여 MLP(multi layer perceptron)를 통과하는 방식으로 입력 개수에 따른 비용을 감소시킨다. 픽셀 단위의 입력을 사용하는 Perceiver의 경우 학습 가능한 파라미터와 입력 간의 교차 어텐션(cross attention)으로 병목현상을 유도하여 계산 비용을 줄여 효율적으로 계산한다.

MLP-Mixer[4], Perceiver[5]은 사전 학습되지 않은 상태에서 CIFAR-10[5]과 SVHN[6] 데이터셋에서 학습한다. 본 실험에서 설정한 패치 크기에 따라 패치 임베딩을 진행한 후 푸리에 임베딩을 적용하여 분류 정확성을 측정한다.

[표 1]에서 볼 수 있듯이 MLP-Mixer[4]에 푸리에 임베딩

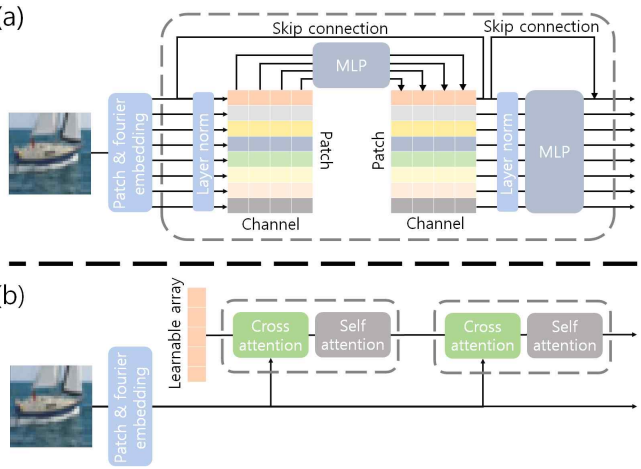


그림 2. 패치 단위의 입력을 사용하는 모델 MLP-Mixer[4](a)와 Perceiver[5](b). (a) MLP를 이용하여 데이터를 채널 단위와 패치 단위로 학습시키는 MLP-Mixer[4] (b) 학습 가능한 배열(learnable array)과 입력의 이용한 교차 어텐션을 진행하는 Perceiver[5].

을 적용하였을 때 CIFAR-10[6] 데이터셋에서 최대1.23%, SVHN[7] 데이터 셋에서 최대 0.56% 정확성 향상을 확인할 수 있었다. 추가로 Perceiver[5]에서 진행한 실험에서 패치 단위의 입력을 사용하였을 때, 푸리에 임베딩을 적용하면 CIFAR-10[5]과 SVHN[6]에서 각각 최대 21.00%, 82.75%의 정확성 향상이 있었다.

## Ⅳ. 결 론

패치를 입력으로 사용하는 모델들은 원본 영상을 패치 단위로 분해하고 재결합하는 과정에서 위치정보를 잃게 된다. 따라서 모델에 적절한 위치정보를 주입하고 패치의 자연스러운 결합을 위해 위치 인코딩 등의 다양한 기법이 필요하다. 본 논문에서는 패치 단위에서 푸리에 임베딩을 적용하여 패치 크기에 따른 성능을 비교하였다. 그 결과 패치 단위에서의 푸리에 임베딩이 성능향상에 도움이 되는 것을 확인하였다.

## ACKNOWLEDGMENT

이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.RS-2022-00167194,미션 크리티컬 시스템을 위한 신뢰 가능한 인공지능)

## 참 고 문 헌

[1] Vaswani, Ashish, et al. Attention is all you need. Advances in neural information processing systems 30. 2017.

[2] Shaw, P., Uszkoreit, J., & Vaswani, Self-attention with relative position representations. 2018.

[3] Rahaman, Nasim, et al. On the spectral bias of neural networks. In: International Conference on Machine Learning. 2019.

[4] Tolstikhin, Ilya O., et al. MLP-mixer: An all-MLP architecture for vision. Advances in Neural Information Processing Systems. 2021.

[5] Jaegle, Andrew, et al. Perceiver: General perception with iterative attention. In: International conference on machine learning. 2021.

[6] Krizhevsky,Alex, et al. Learning multiple layers of features from tiny images. 2009.

[7] Netzer, Yuval, et al. Reading digits in natural images with unsupervised feature learning. 2011.