

# Lab4. Final Report

## Vision Transformers 모델 이해

정보컴퓨터공학부 202355514 강지원

### 1 주요 내용

본 Lab에서는 사전 학습된 Vision Transformers(ViT)를 사용하여 이미지 분류 작업을 수행하고 결과를 시각화 및 분석한다.

### 2 ViT

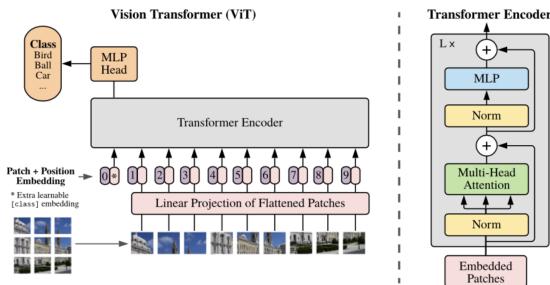


Figure 1: ViT 모델 구조

ViT(Vision Transformers)는 트랜스포머 모델을 이미지 인식에 적용한 모델이다. 이미지를 여러 개의 작은 패치로 나누어 각 패치를 하나의 단어처럼 처리하고, 이를 트랜스포머에 입력하여 전체 이미지를 이해한다. 이 방식은 CNN과 달리 긴 거리의 의존성을 효과적으로 학습할 수 있으며, 대규모 데이터와 사전 학습을 통해 높은 성능을 보인다.

### 3 Vision Transformer inference 파이프라인

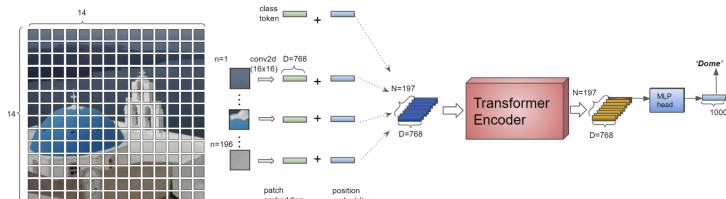


Figure 2: ViT detail

### 3.1 이미지 패치 단위 나누기

입력 이미지를 N개의 패치로 나누는 단계로, 16x16 kernel size, (16, 16) stride를 가지는 2D convolutional filter를 이용하여 224x224의 이미지를 14x14의 패치로 나눈다. 시각화한 모습은 다음과 같다.

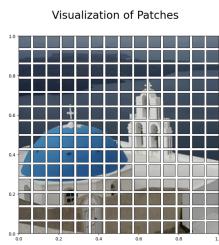


Figure 3: Visualization of Patches

### 3.2 Position Embeddings 더하기

각 패치의 위치를 알려주기 위해, 학습 가능한 position embedding을 patch embedding vector에 더하는 단계이다. position embedding vector는 이미지 내에 높은 유사성을 보이는 distance를 학습한다. 시각화한 모습은 다음과 같다.

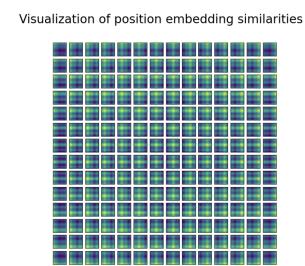


Figure 4: Visualization of Position Embedding similarity

### 3.3 Transformer Input 생성

입력 벡터에 학습 가능한 클래스 토큰을 맨 앞에 추가하고 각 패치 임베딩에 위치 임베딩을 더해 Transformer 입력을 생성하는 단계이다. 이렇게 만들어진 197개의 벡터(1개의 클래스 토큰 + 14 x 14개의 패치)가 Encoder로 전달된다.

### 3.4 Transformer Encoder

Transformer Encoder는 197개의 입력 벡터가 12개의 인코더 블록을 순차적으로 통과하는 구조로, 각 벡터는 fully connected layer를 거쳐 q,k,v로 분리되고, 12개의 head로 나누어 Multi-Head attention을 수행한다. 이후 attention 결과를 결합하고 layer normalization과 두 개의 fully connected layer를 거쳐 동일한 차원의 출력을 생성한다. attention matrix를 시각화 한 모습은 다음과 같으며, 해당 시각화는 Vision Transformer의 각 Attention Head가 이미지의 어느 부분에 주목 하는지를 보여준다. 주변 픽셀과 유사도가 높을수록 노란색 계열로 나타난다.

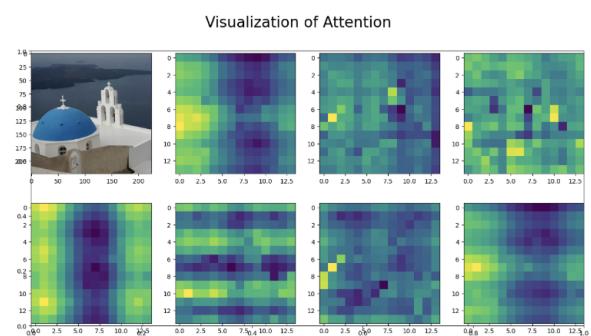


Figure 5: Visualization of Attention

### 3.5 MLP(classification) Head

Transformer Encoder의 출력 중 첫 번째 벡터(cls\_token)가 MLP Head에 입력되어 최종 1,000차원의 분류결과를 생성한다. 결과는 다음과 같다.

Classification Head: Linear( $in\_features = 768$ ,  $out\_features = 1000$ ,  $bias = True$ )

Inference Result:  $id = 497$ , label name = church, church\_building

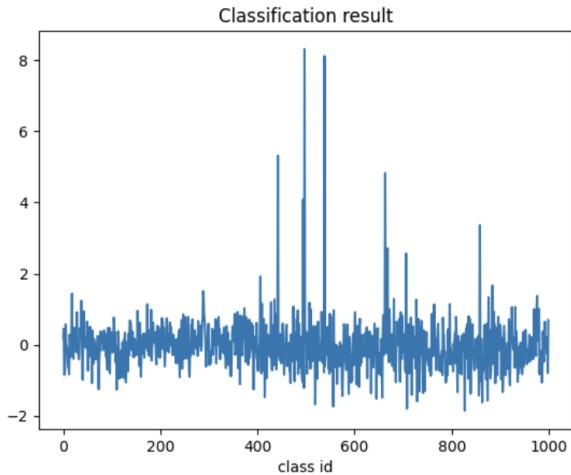


Figure 6: Classification Result

## 4 결론

해당 Lab을 통해 Vision Transformer(ViT)의 전체 구조와 내부 동작 과정을 직접 구현하고 이해할 수 있었다. 특히 ViT pipeline을 단계적으로 분석하며, 모델이 이미지를 어떻게 인식하는지 시각적으로 확인하였다. 이를 통해 ViT가 CNN과는 다른 방식으로 이미지의 전역적 관계를 학습한다는 점을 이해할 수 있었다.

- 
- [1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, arXiv preprint arXiv:2010.11929, 2020. <https://arxiv.org/abs/2010.11929>
  - [2] IkSun, *Chapter 4. Attention Value Matrix in Transformer*, Velog, 2023. <https://velog.io/@alstjsdlr0321/Chapter-4.-Attention-Value-Matrix-in-Transformer>