2025 한양대학교

SW/ICT/AI 종합학술대회

[HYAI : PAPER TO CODE CHALLENGE]

제목: 밑바닥부터 시작하는 ViT

김예준(2025091621), 고희권(2022013627), 김한나(2023026835)

2025. 11. 28

한양대학교 ERICA 소프트웨어융합대학 / 한양대학교 ERICA SW중심대학사업단

구현하려는 논문

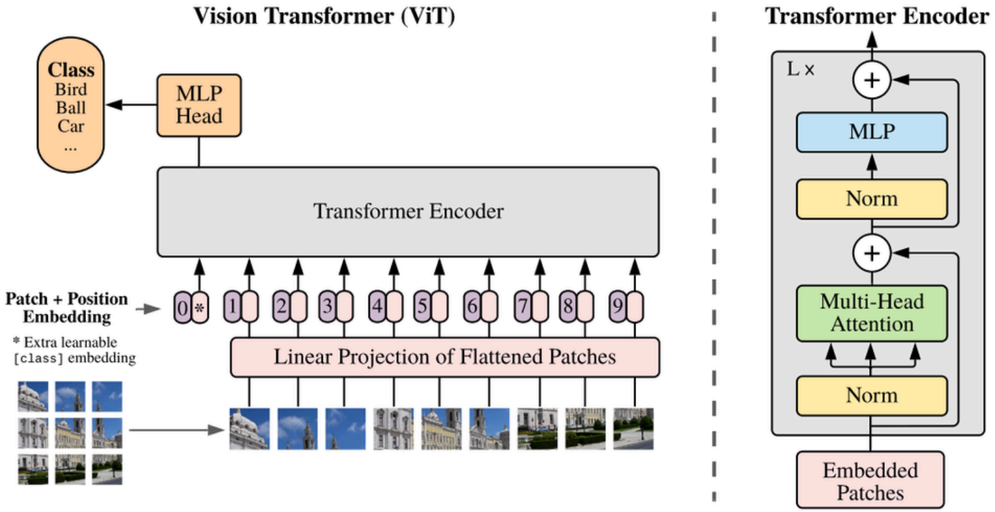
AN IMAGE IS WORTH 16X16 WORDS:
TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy*,†, Lucas Beyer*, Alexander Kolesnikov*, Dirk Weissenborn*,
Xiaohua Zhai*, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer,
Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby*,†
*equal technical contribution, †equal advising
Google Research, Brain Team
{adosovitskiy, neilhoulsby}@google.com

원 논문이 해결하려는 문제

자연어에서 쓰이는 표준 Transformer를,
복잡한 트릭이나 CNN 구조 없이
그대로 이미지 분류에 써도 경쟁력이 있을 수 있는가?
그리고 그 때 필요한 조건 (data size...)은 무엇인가?

참고 Image



구체적 문제

기존 비전 분야에서는 CNN이 사실상 표준이었고, self-attention은
보통 CNN에 일부만 섞거나 일부 레이어만 대체하는 식으로 쓰였다.

순수 Transformer는 CNN이 갖는 국소성·평행이동
불변성 같은 귀납적 편향이 없어서,
중간 크기 데이터셋(ImageNet 정도)에서만 보면
성능이 떨어지는 것으로 알려져 있었음

INTRO.

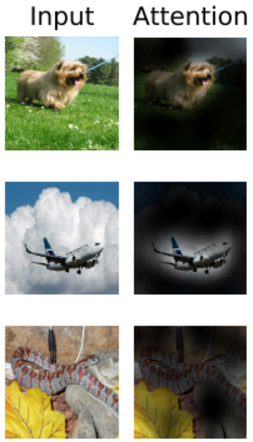
목표 : Vision Transformer를 저수준에서 직접 구현하고 구조·PE·Attention을 재현해 custom 데이터셋에서 성능과 시각화를 검증하였다.

구조 구현 근거

4.5 Inspecting Vision Transformer

To begin to understand how the Vision Transformer processes image data, we analyze its internal representations. The first layer of the Vision Transformer linearly projects the flattened patches into a lower-dimensional space (Eq. 1). Figure 7 (left) shows the top principal components of the the learned embedding filters. The components resemble plausible basis functions for a low-dimensional representation of the fine structure within each patch.

After the projection, a learned position embedding is added to the patch representations. Figure 7 (center) shows that the model learns to encode distance within the image in the similarity of position embeddings, i.e. closer patches tend to have more similar position embeddings. Further, the row-column structure appears; patches in the same row/column have similar embeddings. Finally, a sinusoidal structure is sometimes apparent for larger grids (Appendix D). That the position embeddings learn to represent 2D image topology explains why hand-crafted 2D-aware embedding variants do not yield improvements (Appendix D.4).



Self-attention allows ViT to integrate information across the entire image even in the lowest layers

A. 입력 Image를 patch로 분할 수 linear Projection하는
VIT 구조를 명확히 설명함.
B. Self-attention은 초기 계층부터 전역 정보를 통합함을 언급하며,
이는 우리가 구현한 Multi-Head Attention 구조와 일치한다.

Positional Encoding 재현 근거

Pos. Emb.	Default/Stem	Every Layer	Every Layer-Shared
No Pos. Emb.	0.61382	N/A	N/A
1-D Pos. Emb.	0.64206	0.63964	0.64292
2-D Pos. Emb.	0.64001	0.64046	0.64022
Rel. Pos. Emb.	0.64032	N/A	N/A

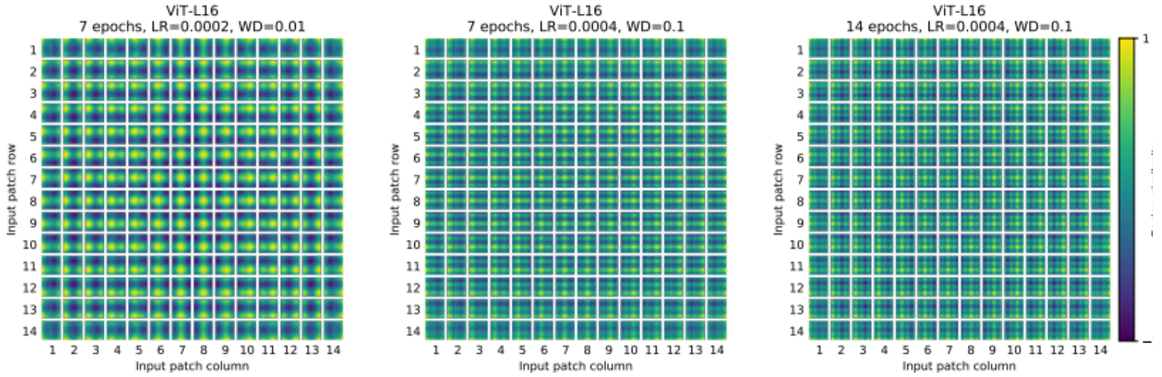


Figure 10: Position embeddings of models trained with different hyperparameters.

A. Appendix D.4는 다양한 positional embedding 방식을 비교해 ViT가
1D/2D 임베딩을 통해 공간적 정보를 학습함
B. Figure 10의 시각화는 positional embedding이
어떻게 패치 위치 정보를 반영하는지를 명확히 나타낸다.

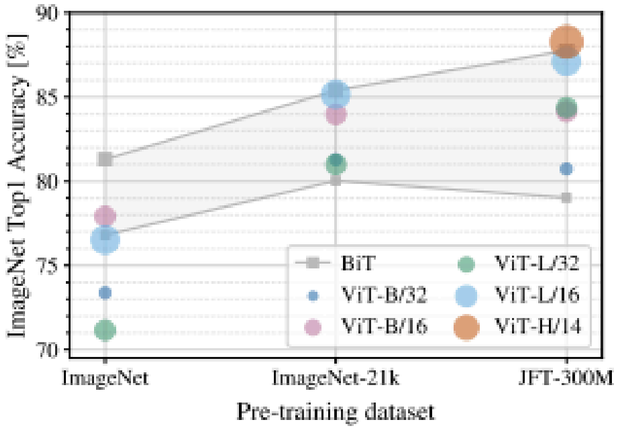


Figure 3: Transfer to ImageNet. While large ViT models perform worse than BiT ResNets (shaded area) when pre-trained on small datasets, they shine when pre-trained on larger datasets. Similarly, larger ViT variants overtake smaller ones as the dataset grows.

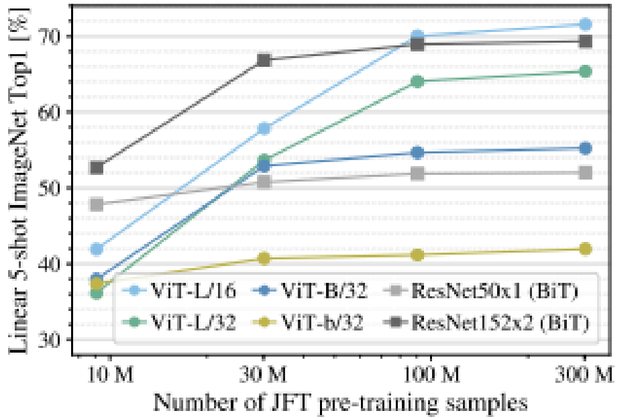


Figure 4: Linear few-shot evaluation on ImageNet versus pre-training size. ResNets perform better with smaller pre-training datasets but plateau sooner than ViT, which performs better with larger pre-training. ViT-b is ViT-B with all hidden dimensions halved.

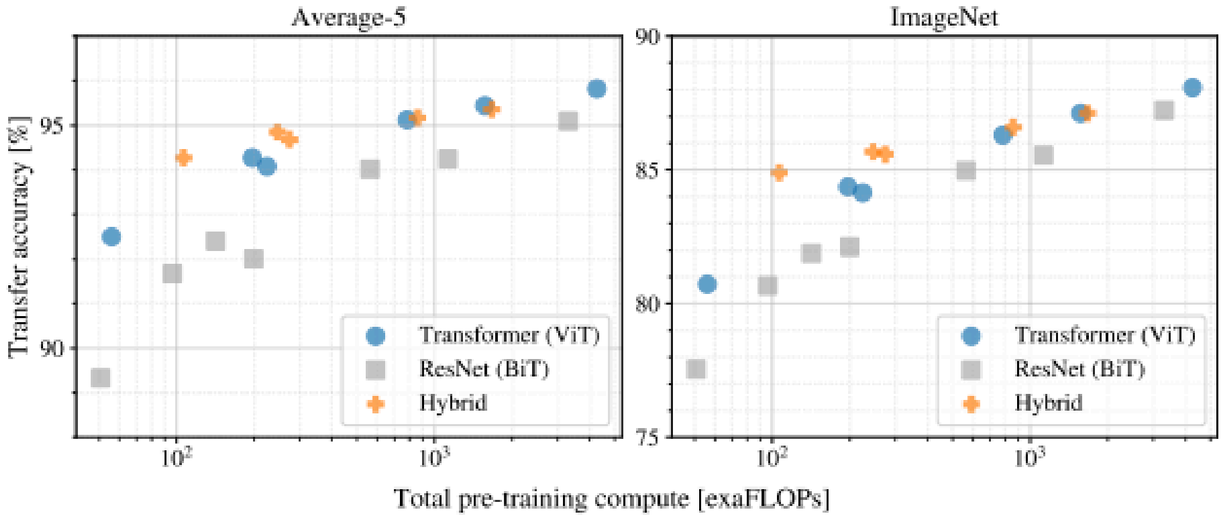
[Figure3]

- **Small data → ResNet 우위**
- **Large data → ViT 성능이 빠르게 상승하여 ResNet을 능가함**

ViT-L/16 > ViT-B/16

[Figure4]

- **ResNet은 적은 pre-training에서 빠르게 성능이 출력되지만, 금방 plateau에 도달함.**
- **ViT는 큰 데이터에서 큰 효율을 보이며, few-shot 성능도 더 높아짐.**



결론

ViT는 대규모 데이터에서 좋은 성능을 발휘하며 ResNet을 넘어서는 모델임을 보여줌

METHOD

구현 모델의 구조

ViT 구조 요약

본 연구에서는 Vision Transformer(ViT)를 저수준부터 직접 구현하여 Patch Embedding, Positional Encoding, Multi-head Attention을 모두 수식 기반으로 재현하였다.

Custom Dataset 설명

모델은 직접 촬영한 ‘bibi’와 ‘hanyang’ 두 클래스로 구성된 경량 이미지 데이터셋을 기반으로 학습되었으며, 소규모 데이터 조건에서 ViT 구조의 동작을 분석하는 데 목적을 두었다.



Model dataset

Implementation Details

학습 설정

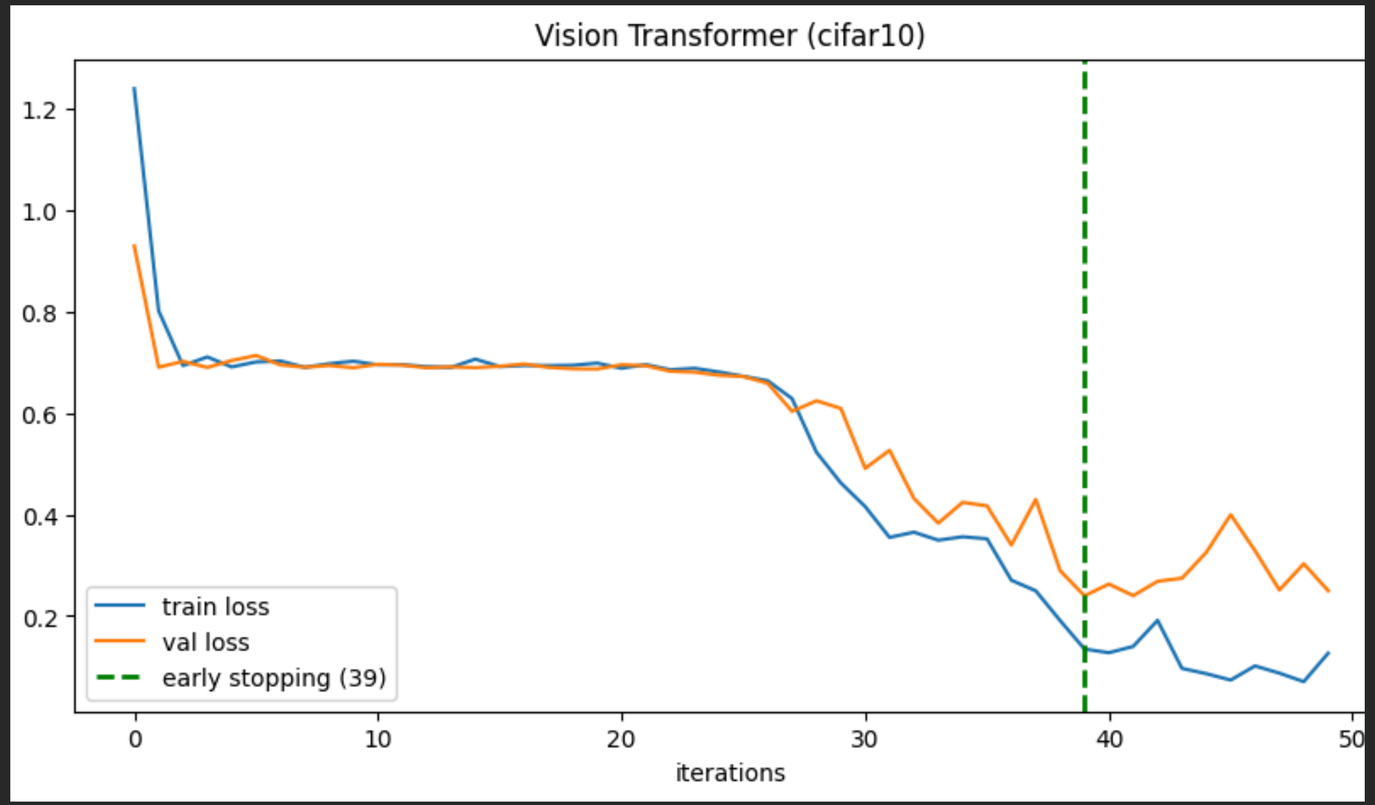
- Epoch: 50
- Batch size: 256
- Optimizer: Adam (lr=0.001, weight_decay=5e-5)
- Loss: CrossEntropyLoss
- Seed: 77

모델 설정

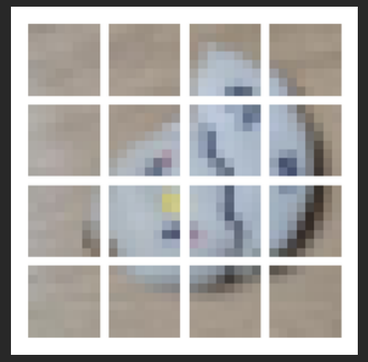
- Patch size: 4
- Embedding dim: 192
- Encoder layers: 12
- Attention heads: 12
- Positional Encoding: 직접 구현(수식 기반)
- Patch Embedding: Rearrange To
  Linear projection 직접 구현

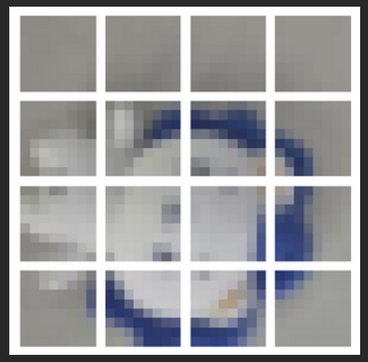
실행 환경

- RTX 5060 Ti



훈련이 진행될수록 손실이 안정적으로 감소하며,
약 39 epoch에서 모델 성능이 최대로 수렴하였다.





Image를 8x8 패치로 분할 후
128차원 임베딩 벡터로 변환



```
predicted:
bibi bibi bibi bibi
label:
bibi bibi bibi bibi
```

테스트셋에서 bibi 클래스 이미지 4장을
모두 정확하게 분류하며,
학습된 패턴이 안정적으로 유지됨을 확인.

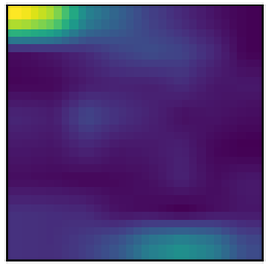
```
torch.Size([1, 3, 32, 32])
```



```
bibi hanyang hanyang bibi
```



bibi



Attention Map



Attentioned

CLS token 기반 Self-Attention을 시각화
모델이 캐릭터의 중앙부·윤곽 등
핵심 영역에 집중했음을 확인

Quantitative Results

우리 ViT 모델은 소규모 custom dataset(bibi/hanyang 2-class)에서 약
37.5%의 accuracy를 기록하였다. 데이터 규모가 매우 작음에도 불구하고
모델이 두 클래스의 패턴적 특징을 부분적으로 학습한 것을 확인할 수 있었다.

```
Accuracy of the network on the 10000 test images: 37.5 %
```

DISCUSSION & CONCLUSION

Discussion

- 소규모 custom 데이터셋에서도 ViT의 Attention이 의미 있는 영역(캐릭터 중심부)을 일관적으로 포착함을 확인함
- Self-attention 기반 구조가 데이터 규모에 민감하지만, 적은 데이터에서도 기본 패턴을 학습함을 검증함

Conclusion

- ViT를 저수준부터 직접 구현하여 Patch Embedding·Positional Encoding·Attention의 동작을 성공적으로 재현
- 데이터 규모는 제한적이었으나, ViT 구조가 소규모 환경에서도 일관된 Attention 패턴을 학습한다는 점을 확인하였다.