Week4_예습과제_김도희

An Image is Worth 16x16 Words

: Transformers for Image Recognition at Scale

> 💡 We show that this reliance on CNNs is not necessary and apure transformer applied directly to sequences of image patches can perform very well on image classification tasks.
>
> When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.

⇒ **image patches**를 **sequence**로 생각해, **Transformer** 사용했더니 **computing resource**를 줄이는 등의 성능 향상

1. Introduction

NLP에서 주로 사용되던 트랜스포머 모델을 이미지 인식에 도입하려는 시도로, CV 분야에서는 CNN이 주로 사용되었지만 트랜스포머를 이미지 패치에 적용하여 CNN을 대체 가능하다. 특히, 대규모 데이터셋에서 트랜스포머를 사전 학습시켰을 경우 효과가 더 커진다.

2. Related work

Vision 분야에서는, Attention 기법을 이미지의 각 픽셀과 모든 픽셀의 어텐션을 계산하여 직접적으로 이미지에 적용하는 방법을 생각해볼 수 있다. 하지만, 이런 방식은 실질적인 input size에는 적용할 수 없는 한계가 있다.

CNN이 가지고 있는 locality과 translation equivariance 같은 특성이 트랜스포머에서는 자연스럽게 나타나지 않기 때문에, 이를 해결하려는 접근들이 많았다.

⇒ **CNN**은 지역적 패턴을 인식하고, 이미지가 이동해도 동일하게 인식할 수 있는 구조적 강점이 있지만, **ViT**는 이 특성들을 가지지 않으며, 이를 데이터에서 학습하게 되어 유연성을 가진다.

3. Method

3.1. Vision Transformer

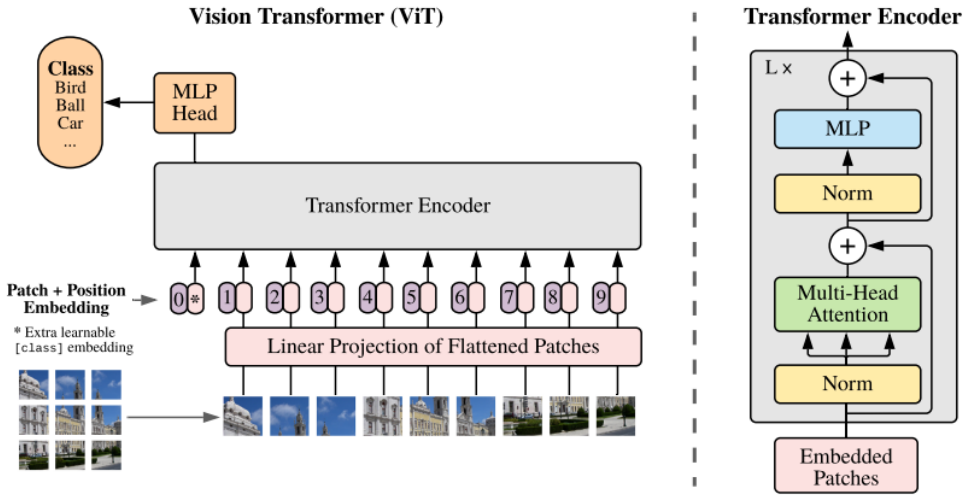

Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification token" to the sequence. The illustration of the Transformer encoder was inspired by Vaswani et al. (2017).

→ 이미지를 작은 패치로 나누어 각각의 패치를 하나의 token으로 처리합니다. 이 패치들은 NLP에서 트랜스포머가 단어를 처리하는 방식과 유사하게 다루어진다. 즉, 이미지 전체를 패치 단위로 나누어 그 관계를 학습한다.

1. 이미지 패치로 나누기
2. 패치 임베딩
3. 위치 임베딩
4. 트랜스포머 입력

5. [CLS] 토큰

3.2 Fine-Tunning and higher resolution

ViT를 large dataset에 pre-trained하고 down stream tasks에 fine-tuning을 진행.

prediction head를 없애고, $D \times K$의 feed forward layer로 변경하는데 고해상도로 down-stream task에 fine-tuning 시에 효과적이다. 이 때 pre-trained 단계에서 사용했던 patch size와 동일한 size를 사용해 **더 긴 sequence length**를 사용하게 되어, positional embedding은 효과가 없어지기 때문에 길이에 맞춰 2D interpolation을 진행한다

4. Experiments

4.1 Setup

- Dataset

  large-scael dataset(**pre-trained) ⇒** 벤치마크 task**(transfer)**

- Model Variants

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Table 1: Details of Vision Transformer model variants.

ViT는 BERT 모델에서 사용된 구성을 기반으로 설계되었으며, "Base", "Large", "Huge" 모델이 포함

패치 크기가 작아질수록 트랜스포머의 시퀀스 길이가 길어지기 때문에, 모델의 계산 복잡도가 증가

- Training & Fine-tuning
  - Optimizer : **Adam** 고정($\beta1$=0.9,$\beta2$=0.999)
  - (high)weight decay : 0.1

- linear learning rate warmup and decay

Dataset	Steps	Base LR
ImageNet	20 000	{0.003, 0.01, 0.03, 0.06}
CIFAR100	10 000	{0.001, 0.003, 0.01, 0.03}
CIFAR10	10 000	{0.001, 0.003, 0.01, 0.03}
Oxford-IIIT Pets	500	{0.001, 0.003, 0.01, 0.03}
Oxford Flowers-102	500	{0.001, 0.003, 0.01, 0.03}
VTAB (19 tasks)	2 500	0.01

Table 4: Hyperparameters for fine-tuning. All models are fine-tuned with cosine learning rate decay, a batch size of 512, no weight decay, and grad clipping at global norm 1. If not mentioned otherwise, fine-tuning resolution is 384.

- Metrics
  - few-shot accuracy

    ⇒ recover the exact solution in closed form
  - fine-tuning accuracy

4.2 Comparision to SOTA

- Largest model (ViT-H/14 and ViT-L/16– to state-of-the-art CNNs)
  - point 1: supervised transfer learning with large ResNets
  - point 2: Noisy Student

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	**99.74** ± 0.00	99.61 ± 0.02	99.63 ± 0.03	–
VTAB (19 tasks)	**77.63** ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

Table 2: Comparison with state of the art on popular image classification benchmarks. We report mean and standard deviation of the accuracies, averaged over three fine-tuning runs. Vision Transformer models pre-trained on the JFT-300M dataset outperform ResNet-based baselines on all datasets, while taking substantially less computational resources to pre-train. ViT pre-trained on the smaller public ImageNet-21k dataset performs well too. *Slightly improved 88.5% result reported in Touvron et al. (2020).

⇒ ViT-L/16 가 BiT-L에 비해 모든 task에서 우수하며, ViT-H/14가 위의 모델 중에서 가장 우수하다고 할 수 있다. 그리고 transformer기반의 model이 ResNet기반의 model에 비해, 계산량이 더 적다는 것을 알 수 있다.

4.3 Pretraining Data Requirement

1. ViTmodels을 size가 증가하는 방향으로 학습
2. random subsets of 9M, 30M, and 90M 에서 학습

4.4 Scaling Study

ViT 모델의 크기와 성능 간의 관계를 분석한다. 더 큰 데이터셋과 더 많은 모델 파라미터를 사용할수록 성능이 향상된다는 점을 강조된다. 특히, CNN은 작은 데이터셋에서는 우수한 성능을 보이지만, 대규모 데이터셋에서는 transformer가 더 좋은 성능을 보인다.

⇒ 데이터 크기가 성능에 중요한 영향을 미친다.

4.5 Inspecting Vision Transformer

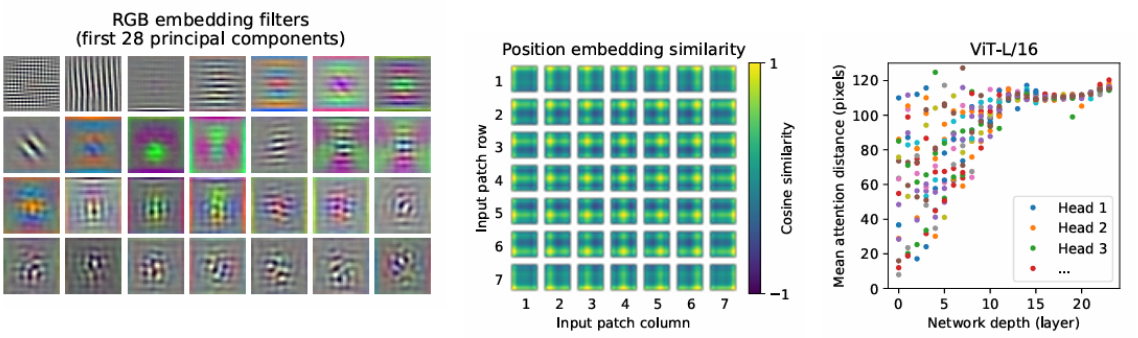

Figure 7: **Left:** Filters of the initial linear embedding of RGB values of ViT-L/32. **Center:** Similarity of position embeddings of ViT-L/32. Tiles show the cosine similarity between the position embedding of the patch with the indicated row and column and the position embeddings of all other patches. **Right:** Size of attended area by head and network depth. Each dot shows the mean attention distance across images for one of 16 heads at one layer. See Appendix D.7 for details.

et al., 2019; Radford et al., 2018). We also perform a preliminary exploration on *masked patch prediction* for self-supervision, mimicking the masked language modeling task used in BERT. With self-supervised pre-training, our smaller ViT-B/16 model achieves 79.9% accuracy on ImageNet, a significant improvement of 2% to training from scratch, but still 4% behind supervised pre-training. Appendix B.1.2 contains further details. We leave exploration of contrastive pre-training (Chen et al., 2020b; He et al., 2020; Bachman et al., 2019; Hénaff et al., 2020) to future work.

- 좌측 : **flattened patches**를 저차원으로 투영
- 중앙 : 위치임베딩을 더함
- 우측 : 이미지 공간에서 평균 거리를, attention weights를 기반으로 나타낸 모습.

4.6 Self-Supervision

large scale self-supervised pre-training라는 점이 ViT의 성능에 영향을 줬다.

5. Conclusion

향후 탐색할 과제로는 객체 검출 및 분할 등 다양한 컴퓨터 비전 작업에 트랜스포머를 적용하는 것. 또한 Self-Supervision과 같은 방법들을 통해 추가적인 성능 향상이 가능할 것이라는 점을 시사하고 있다.

2010.11929v2.pdf