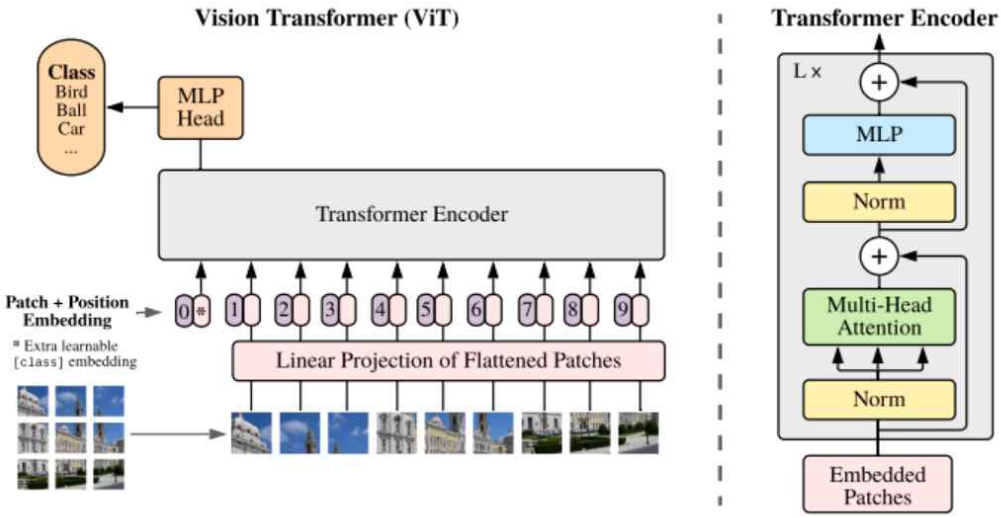
[AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE]

1. Model



이미지는 고정된 크기의 패치로 쪼개고, 각각을 선형적으로 임베딩한 후 위치 임베딩을 더하여 결과 백터를 일반적인 Transformer 인코더의 인풋으로 입력한다. 분류 과제를 수행하기 위해 추가적으로 학습되는 "classification token"을 만들어 시퀀스에 더한다.

2. Vision Transformer (ViT)

[이미지 인풋]

- 일반적인 Transformer은 토큰 임베딩에 대한 1차원의 시퀀스를 입력으로 받음

- 2차원의 이미지를 다루기 위해 논문에서는 이미지를 flatten된 2차원의 패치의 시퀀스로 변환함

- 즉, H x W x C → N x (P^2 x C) 로 변환

  > (H, W)는 원본 이미지의 크기, C는 채널 개수를 의미

  > (P, P)는 이미지 패치의 크기

  > N = HW/P^2 = 패치의 개수

- Transformer은 모든 레이어에서 고정된 벡터 크기 D를 사용하기 때문에 이미지 패치는 펼친 다음 D차원 벡터로 linear projection 시킴

- BERT의 [CLS]토큰과 비슷하게 임베딩 된 패치의 시퀀스에 z0 = x_class 임베딩을 추가로 붙여 넣음

- 이후 이 패치에 대해 나온 인코더 아웃풋은 이미지 representation으로 해석하여 분류에 사용

[위치 임베딩]

- 각각의 패치 임베딩에 위치 임베딩을 더하여 위치 정보를 활용할 수 있도록 함

- 학습 가능한 1차원의 임베딩을 사용

- 2차원 정보를 유지하는 위치 임베딩도 활용해 보았으나, 유의미한 성능 향상은 없었기 때문

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}}; \mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \cdots; \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}, \quad \mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D}, \mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D} \tag{1}$$

$$\mathbf{z}'_\ell = \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \quad \ell = 1 \ldots L \tag{2}$$

$$\mathbf{z}_\ell = \text{MLP}(\text{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, \quad \ell = 1 \ldots L \tag{3}$$

$$\mathbf{y} = \text{LN}(\mathbf{z}_L^0) \tag{4}$$

[하이브리드 아키텍처]

- 이미지 패치를 그대로 사용하는 대신, CNN의 결과 나온 feature map을 인풋 시퀀스로 사용할 수 있음

- 하이브리드 모델에서는 패치 임베딩 프로젝션을 CNN feature map에서 결과로 나온 패치에 대해 적용함

- 특수한 케이스로 패치는 1x1 크기를 가질 수 있는데, 이 경우는 인풋 시퀀스를 단순히 feature map에 대한 차원으로 flatten 한 후 Transformer의 차원으로 projection 한 결과임

- [CLS]에 해당하는 인풋 임베딩과 위치 임베딩은 기존 모델과 동일하게 적용함

3. fine-tuning과 높은 해상도 이미지 다루기

일반적으로 large-scale dataset에 대해 ViT를 pre-train하고 downstream task에 대해 fine-tuning을 수행한다. 이를 위해 pre-trained prediction head를 제거하고 0으로 초기화된 D×K feedforward layer를 추가한다. 여기서 K는 downstream class의 개수이다.

Pre-train 보다 높은 resolution으로 fine-tuning하는것은 종종 도움이 된다. 더 높은 resolution의 이미지를 feed할 때 patch 크기를 동일하게 유지하므로 sequence length가 더 길어진다. Vision Transformer는 임의의 sequence length를 처리할 수 있지만 pre-trained position embedding은 의미가 없을 수 있다. 따라서 원본 이미지에서의 위치에 따라 pre-trained position embedding의 2D interpolation을 수행한다. Resolution 조정 및 patch 추출은 이미지의 2D 구조에 대한 inductive bias가 Vision Transformer에 수동으로 주입되는 유일한 지점이다.

4. Result

모델 사이즈

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

벤치마크 데이터셋 성능

	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

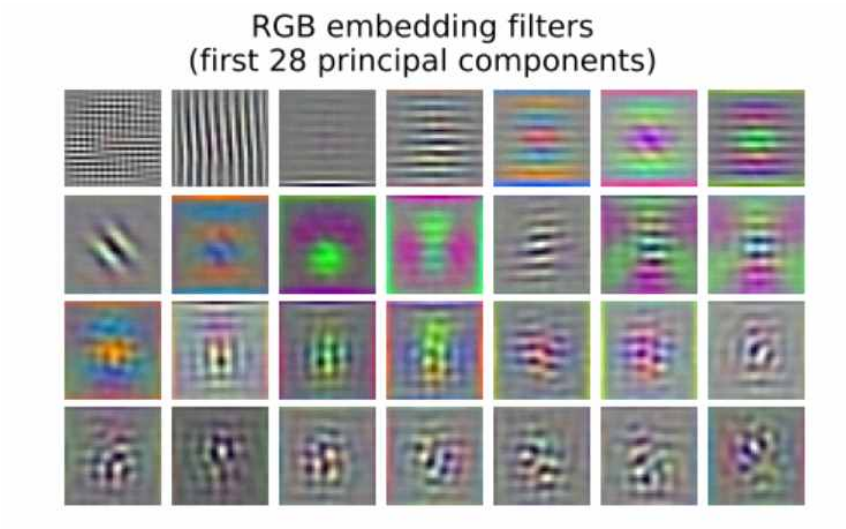
Table 2: Comparison with state of the art on popular image classification benchmarks. We report mean and standard deviation of the accuracies, averaged over three fine-tuning runs. Vision Transformer models pre-trained on the JFT-300M dataset outperform ResNet-based baselines on all datasets, while taking 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).

5. Inspecting Vision Transformer

[임베딩 프로젝션]

- ViT는 펼쳐진 패치를 더 낮은 차원의 공간으로 매핑한다.

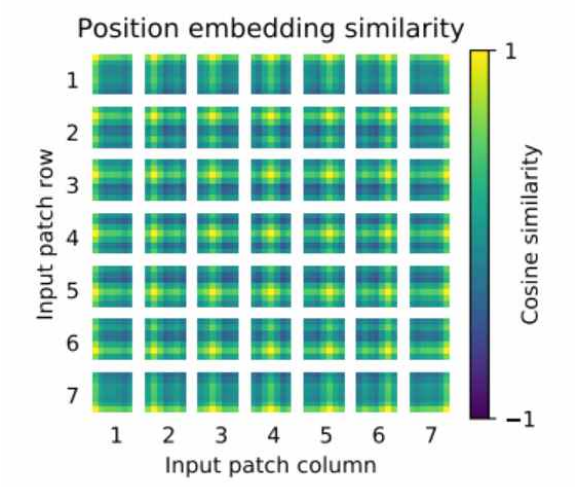
- 아래 그림은 학습된 임베딩 필터 중 중요한 몇 가지 구성 요소들을 시각화한 것이다.



[위치 임베딩]

- 선형 프로젝션 이후, 각각의 패치 representation에는 위치 임베딩이 더해지게 된다.

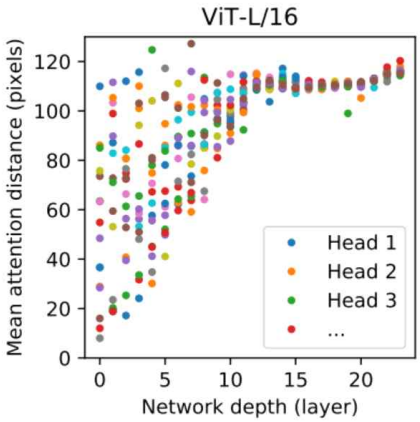
- 시각화 결과를 보면, 모델은 이미지 내의 거리 개념을 인코딩하여 위치 임베딩에서 유사성이 나타난다는 것을 알 수 있다.



[Self-attention]

- Self-attention은 가장 밑단에 있는 레이어에서부터도 ViT가 전체 이미지에 있는 정보를 통합하도록 돕는다.

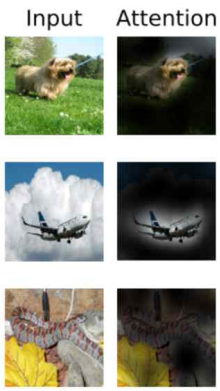
- 이러한 능력을 어느 정도까지 활용할 수 있는지 조사하기 위해 attention weight에 기반하여 이미지 공간 상에서 정보가 취합되는 평균 거리를 구해보았다. (여기서 attention distance는 CNN에서 receptive field와 비슷하게 해석할 수 있다.)



- 실험 결과 attention head 중 일부는 가장 낮은 레이어에서부터 대부분의 이미지에 attend 하고 있고, 이렇게 글로벌하게 정보를 통합하는 능력을 모델이 활용하는 것으로 보인다.

- 또 다른 attention head는 밑단 레이어에서 일관적으로 작은 거리의 패치에 집중하는 모습을 보였는데, 이렇게 지역적인 attention은 하이브리드 모델에서는 좀처럼 나타나지 않았다. 즉, 이 attention head는 CNN의 밑단에서 일어나는 것과 비슷한 작용을 하는 것이라고 유추할 수 있다.

- attention이 일어나는 거리는 네트워크의 깊이가 깊어질수록 늘어난다는 것도 찾을 수 있었으며, 전체적으로 모델은 의미적으로 분류 과제에 필요한 부분에 attend 하는 것을 찾을 수 있었다.



6. Self-supervision

Transformer는 NLP task에서 인상적인 성능을 보여주었다. 그러나 대부분의 성공은 확장성 뿐만 아니라 self-supervised pre-training에서 비롯된다. 또한 BERT에서 사용되는 MLM(Masked Language Modeling)을 모방하여 self-supervision을 위한 Masked Patch Prediction에 대한 예비탐색을 수행한다.

Self-supervised pre-training을 수행한 ViT-B/16 모델은 ImageNet에서 79.9%의 정확도를 달성하여 scratch로 부터 train을 수행한것보다 2%가 향상되었지만 supervised pre-training보다는 4% 떨어졌다.