8주차 Vision adv 논문 리뷰 과제

VIT : An IMAGE IS WORTH 16x16 WORDS : TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

1. Introduction

- **Inductive Bias (귀납적 편향)**

  의미 : training에서 보지 못한 데이터에 대해서도 적절한 귀납적 추론이 가능하도록 하기 위해 모델이 가지고 있는 가정들의 집합

  DNN의 기본적인 요소들의 inductive bias

  1. FC (fully connected) : 입력 및 출력 element가 모두 연결되어 있어 구조적으로 특별한 relational inductive bias 가정 x
  2. Convolutional NN : 작은 크기의 커널로 이미지를 지역적으로 보며, 동일한 커널로 이미지 전체를 본다는 점에서 locality와 transitional invariance(입력의 위치가 변해도 출력은 변하지 않는다.) 특성 지님
  3. Recurrent NN : 입력한 데이터들이 시간적 특성을 가지고 있다고 가정하므로 sequentiality와 temporal invariance 특성 지님

  💡 **Transformer는 self attention 기반으로, CNN 및 RNN 보다 상대적으로 inductive bias가 낮 충분하지 못한 양의 데이터를 학습할 때 일반화가 잘 되지 않는다.**

Component	Entities	Relations	Rel. inductive bias	Invariance
Fully connected	Units	All-to-all	Weak	-
Convolutional	Grid elements	Local	Locality	Spatial translation
Recurrent	Timesteps	Sequential	Sequentiality	Time translation
Graph network	Nodes	Edges	Arbitrary	Node, edge permutations

Table 1: Various relational inductive biases in standard deep learning components. See also Section 2.

- **Vision Transformer (VIT) 개요**
  - 본 연구에서 NLP에서 사용되는 self-attention 기반 아키텍처인 Transformer가 주요 모델를 이미지에 그대로 적용시킨 Vision Transformer (VIT) 제안
  - VIT → 이미지를 패치로 분할 하는데, 이는 NLP의 단어와 같다고 볼 수 있다. 분할한 각 패치의 linear embedding sequence를 Transformer의 input 값으로 넣어서 이미지를 분류한다.
  - 중간 규모의 ImageNet과 같은 dataset에서 training된 모델은 강한 정규화 없이도 비슷한 크기의 ResNet보다 약간 낮은 정확도를 나타낸다. → VIT가 일부 CNN보다 inductive bias가 낮아 데이터 양이 부족한 상황에서는 잘 일반화가 되지 않는다는 것을 확인할 수 있다.
  - 하지만 더 큰 dataset 이미지로 훈련시킨다면, 충분한 규모로 사전 훈련이 진행되고, 데이터 포인트가 적은 작업으로 전이될 때 훌륭한 결과를 얻는다.
  - ImageNet-21k 데이터셋이나 JFT-300M 데이터셋에서 VIT가 SOTA 성능을 도출하는 것을 통해 large scale 학습이 낮은 inductive bias로 인한 성능 저하를 해소시키는 것을 알 수 있다.

2. Related Work and Proposed METHOD

- **VIT 모델 구조 - 총 5 Step으로 진행**

>> Step 1 : 이미지 x 에 대해서 (PxP) 크기의 패치 N개로 분할하여 패치 sequence x_p 구축

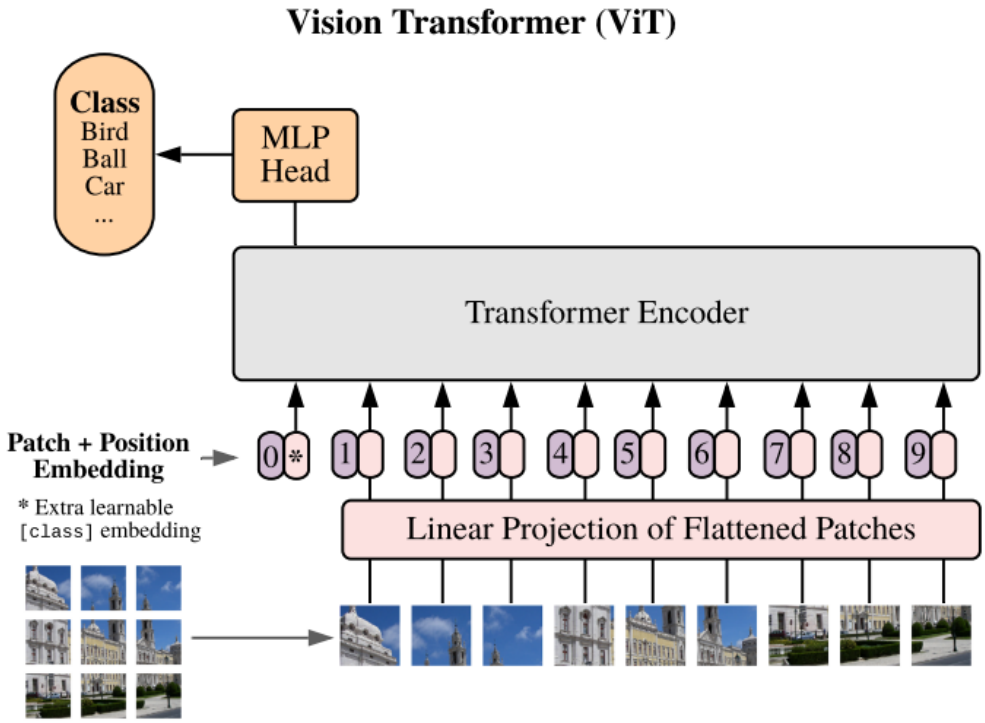
>> Step 2 : Trainable linear projection을 통해 x_p의 각 패치를 flatten한 벡터를 D 차원으로 변환하여 이를 patch embedding으로 사용 (하나 하나의 patch → NLP의 token으로 간주)

learnable positional embedding과 element-wise sum으로 결합된 embedding을 사용하여 0번째 patch + position embedding에는 class을 부여 → patch embedding

>> Step 3 : learnable class embedding과 patch embedding에 learnable position embedding을 더한다,

>> Step 4 : 임베딩 → Vanilla Transformer encoder에 input 값으로 넣어 마지막 layer에서 class embedding에 대한 output인, image representation 도출 (BERT에서 동일하게, class token을 넣고 각 token에 대한 representation 사용)

>> Step 5 : MLP Head에 image representation을 input 으로 넣어 최종 이미지의 class 분류



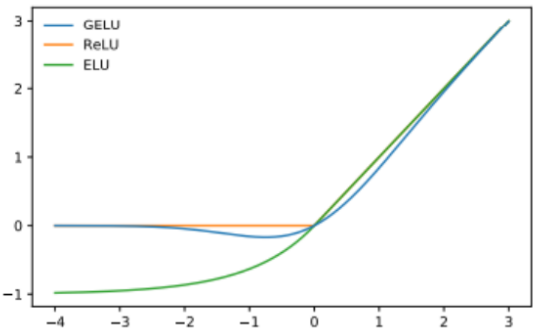
- **Positional Embedding**

  ViT에서는 아래 4가지 position embeding을 시도한 후, 최종적으로 가장 효과가 좋은 1D position embedding을 ViT에 사용한다.

  1. No positional information : Considering the inputs as a bag of patches.
  2. 1-dimenstional positional embedding : Considering the inputs as a sequence of patches in the raster order(배열된 순서).
  3. 2-dimenstional positional embedding : Considering the inputs as a grid of patches in two dimensions(공간적 배열 체계적으로 이해).
  4. Relative positional embeddings : Considering the relative dinstance between patches to encode the spatial information as intead of their absolute position.

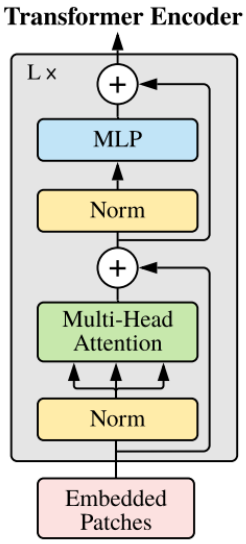
Pos. Emb.	Default/Stem	Every Layer	Every Layer-Shared
No Pos. Emb.	0.61382	N/A	N/A
1-D Pos. Emb.	0.64206	0.63964	0.64292
2-D Pos. Emb.	0.64001	0.64046	0.64022
Rel. Pos. Emb.	0.64032	N/A	N/A

- **Transformer Encoder**
  - ViT → Multi-head Self Attention (MSA)와 MLP block으로 이루어져 있다.
  - MLP는 2개의 layer를 가지며, GELU activation function을 사용한다.



출처 : GAUSSIAN ERROR LINEAR UNITS (GELUS)

  - 각 block의 앞에는 Layer Norm (LN)을 적용하고, 각 block의 뒤에는 residual connection (+ 부분 해당)을 적용한다.



수식

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

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

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

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

- **ViT 관점에서의 Inducitve Bias**

- ViT에서 MLP는 locality와 translation equivariance(입력이 비뀜에 따라 출력도 바뀐다.)가 있지만, MSA는 global하기 때문에 CNN보다 image - specific inductive bias가 낮다.
- 떠라서 아래 두 가지 방법을 통해 ViT에 inductive bias의 주입을 시도한다.

1. Patch extraction : cutting the image into patches (패치 단위로 이미지를 자른다.
2. Resolution adjustment(해상도 조정) : adjusting the position embeddings for images of different resolution at fine-tuning.

- **Hybrid Architecture**
  - ViT → raw image가 아닌 CNN으로 추출한 raw image의 feature map을 활용할 수 있다.
  - feature map은 이미 raw image의 공간적 정보를 포함하고 있다. 따라서 hybrid architecture는 패치 크기를 1x1로 설정해도 무방하다.
  - 1x1 크기의 patch를 사용할 경우 feature amp의 공간 차원을 flatten하여 각 vector에 linear projection을 적용하면 된다.
- **Fine-tuning and Higher Resolution**
  - Large scale로 ViT를 사전 훈련한 후, 해당 모델을 downstream task에 fine-tuning하여 사용할 수 있다.
  - ViT를 fine-tuning 할 때, ViT의 pre-trained prediction head를 zero-initialized feedforward layer로 대체한다.
  - 또한 ViT를 fine-tuning 할 때, pre-training과 동일한 패치의 크기를 사용하기 때문에 고해상도의 이미지로 fine-tuning을 하면 sequence 길이가 더 길어진다.
  - ViT는 가변적 길이의 patch들을 처리할 수 있지만, pre-trained position embedding은 의미가 사라지므로 pre-trained position embedding을 원본 이미지의 위치에 따라 2D interpolation(보간법)하여 사용한다.

→ Fine- tuning 시 MLP head만 변경을 해서 모델 구축한다.

3. Experiments

- **Datasets**
  - ViT는 아래와 같이 class와 이미지의 개수가 각각 다른 3개의 데이터셋을 기반으로 pre-train 진행한다.

- 아래의 benchmark tasks를 downstream task로 하여 pre-trained ViT의 representation 성능을 검증한다.
  - ReaL labels (Beyer et al., 2020), CIFAR-10/100 (Krizhevsky, 2009), Oxford-IIIT Pets (Parkhi et al., 2012), and Oxford Flowers-102 (Nilsback & Zisserman, 2008)
  - 19-task VTAB classification suite

Pre-trained Dataset	# of Classes	# of Images
ImageNet-1k	1k	1.3M
ImageNet-21k	21k	14M
JFT	18k	303M (High resolution)

- **Model Variants**
  - ViT는 아래와 같이 총 3개의 volume에 대해 실험 진행, 다양한 패치 크기에 대해 실험을 진행했다.
  - Baseline CNN은 batch normalization layer를 group normalization으로 변경하고 standarized convolutional layer를 사용하여 transfer learning에 적합한 Big Transformer (BiT) 구조의 ResNet을 사용한다.

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.

- **Comparison to SOTA (State of the Art)**
  - 본 실험에서 14x14 패치 크기를 사용한 ViT-Huge와 16x16 패치 크기를 사용한 ViT-Large의 성능을 baseline과 비교했다.

- JFT dataset에서 pre-training한 ViT-L/16 모델이 모든 downstream task에 대하여 BiT-L보다 높은 성능을 도출한다.
- ViT-L/14 모델은 ViT-L/16 모델보다 향상된 성능을 도출하였으며, BiT-L 모델보다 학습 시간 또한 훨씬 짧았다.

	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

- 19- task VTAB classification suite를 아래와 같이 3가지 그룹으로 나누어 추가 실험을 진행했다.
  - Natural : task like Pets, CIFAR, etc
  - Speicalized : Medical and satellite imagery
  - Structured : tasks that require geometric understanding like localization
- 전체 데이터뿐만 아니라 각 그룹에서도 ViT-H/14가 좋은 결과를 도출했다.

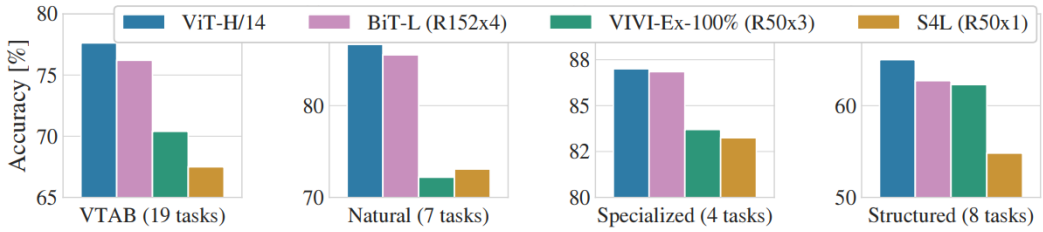
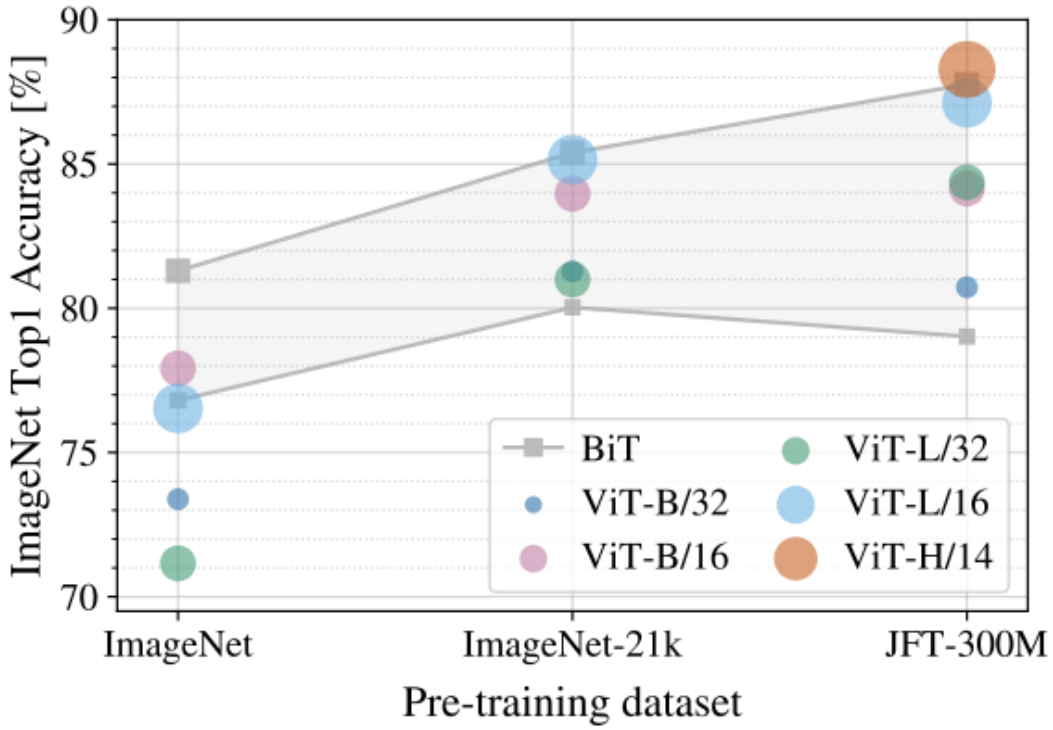
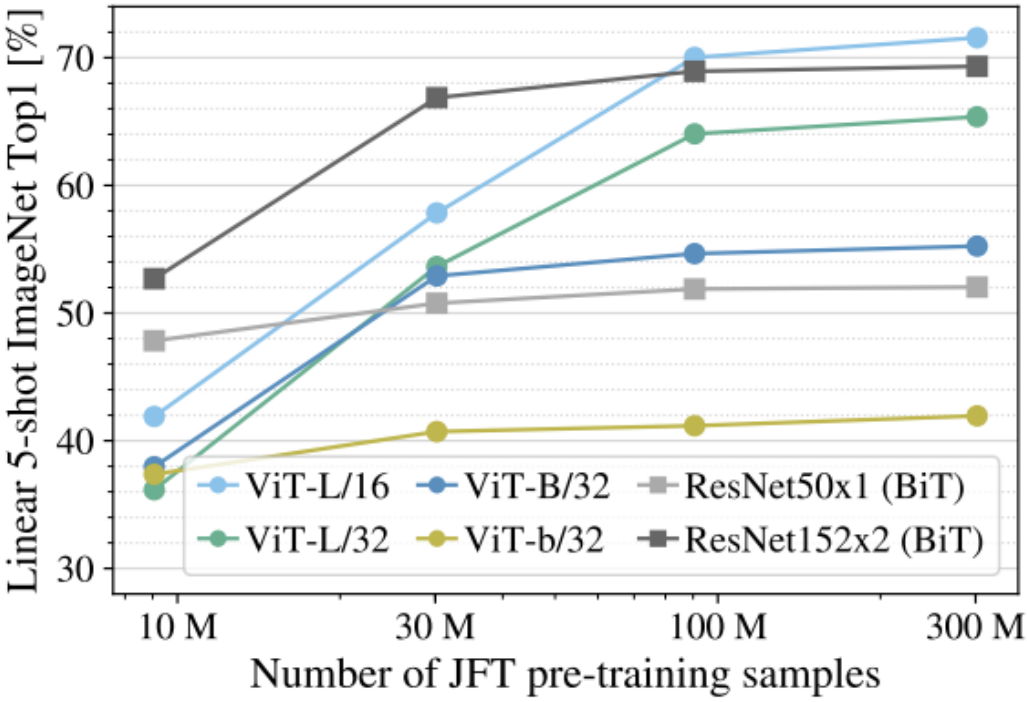


Figure 2: Breakdown of VTAB performance in *Natural*, *Specialized*, and *Structured* task groups.

- **Pre-training Data Requirements**
  - 본 실험에서는 pre-training 데이터셋의 크기에 따른 fine-tuning 성능을 확인했다.



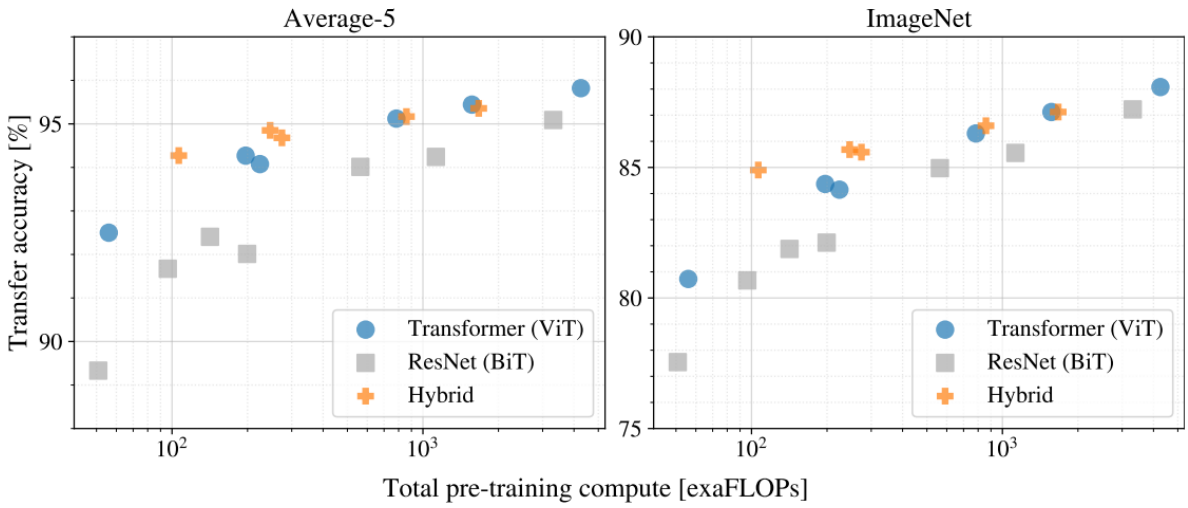
각 데이터셋에 대하여 pre-training한 ViT를 ImageNet에 transfer learning한 정확도를 확인한 결과, 데이터가 클수록 ViT가 BiT보다 성능이 좋고 크기가 큰 ViT 모델이 효과가 있다.



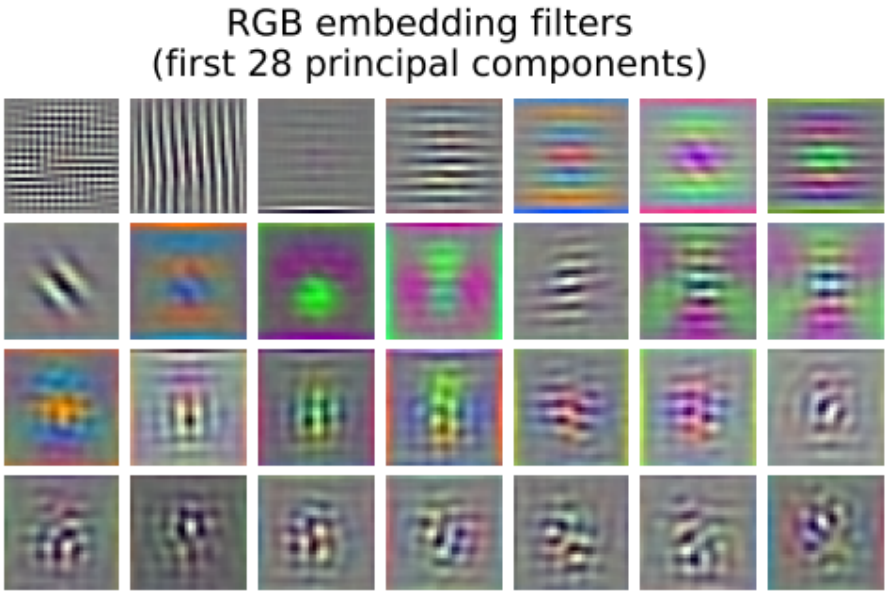
JFT를 각각 다른 크기로 랜덤 샘플링한 dataset을 활용하여 실험을 진행한 결과, 작은 데이터셋에서 CNN의 inductive bias가 효과가 있었으나, 큰 데이터셋에서는 데이터로부터 패턴을 학습하는 것만으로 충분함을 알 수 있다.

> 💡 **CNN, 작은 데이터셋에서는 inductive bias 덕분에 좋은 결과를 도출할 수 있지만, 많은 데이터셋에서는 ViT가 훨씬 더 나은 결과를 도출할 수 있다.**

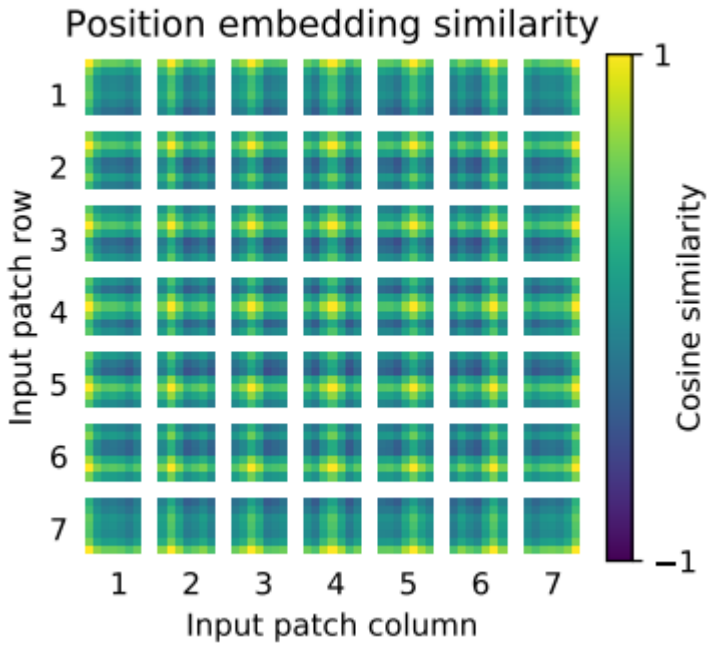
- **Scaling Study**
  - JFT를 기반으로 pre-training cost 대비 transfer 성능을 검증하여 모델들의 scaling study를 진행했다.
    - Pre-training cost : TPUv3 accelerator에서 모델의 inference 속도 관련 지표 의미
  - ViT가 성능, cost의 trade-off에서 ResNet (BiT)보다 우세한 것을 검증했다.
  - Cost가 증가할 수록 Hybrid와 ViT의 성능과 cost의 trade-off 차이가 감소한다.



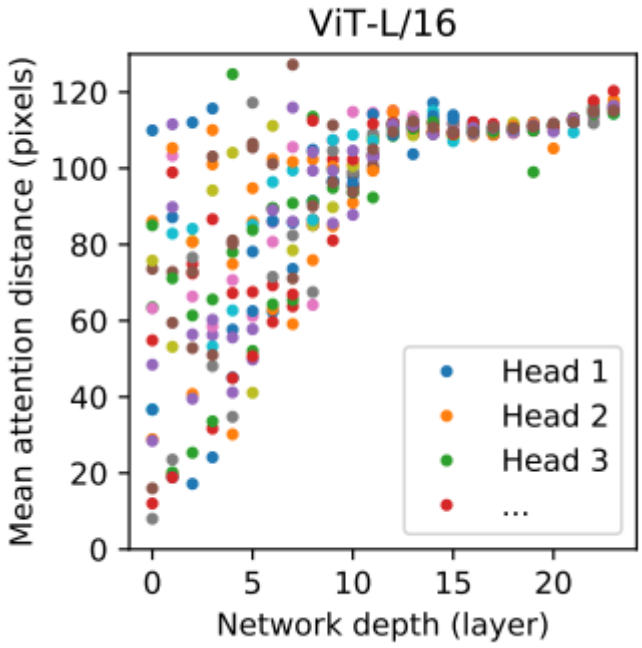
- Inspecting Vision Transformer
  - 본 실험에서는 ViT가 어떻게 이미지를 처리하는지 이해하기 위한 실험 진행



flatten 패치를 patch embedding으로 변환하는 linear projection의 principal components를 분석했다.



패치 간 position embedding의 유사도를 통해 가까운 위치에 있는 패치들의 position embedding이 유사한지 확인했다.



ViT의 layer별 평균 attention dinstance를 확인한 결과, 초반 layer에서도 attention을 통해 이미지 전체의 정보를 통합하여 사용함을 알 수 있다.

참고자료

https://www.youtube.com/watch?v=0kgDve_vC1o&t=57s&pp=ygUEdml0IA%3D%3D

inductive bias

https://velog.io/@euisuk-chung/Inductive-Bias란

https://moon-walker.medium.com/transformer는-inductive-bias이-부족하다라는-의미는-무엇일까-4f6005d32558