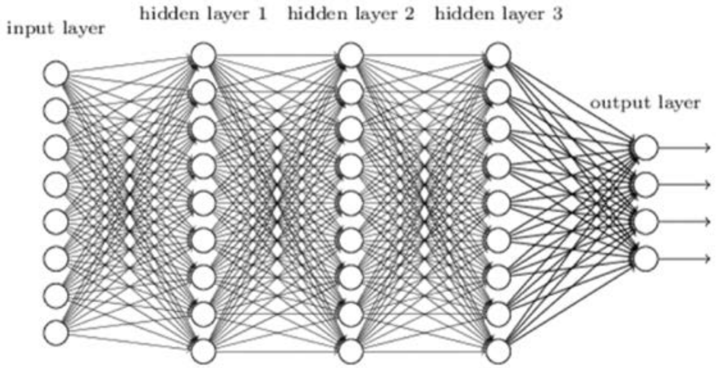
Vision Transformer(ViT)

발표자: 문경환

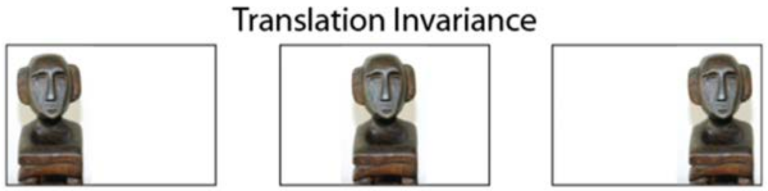
Index

1. About CNN
2. About ViT
3. Attention
4. Self-Supervision
5. Hyperparameters
6. Q&A

About CNN



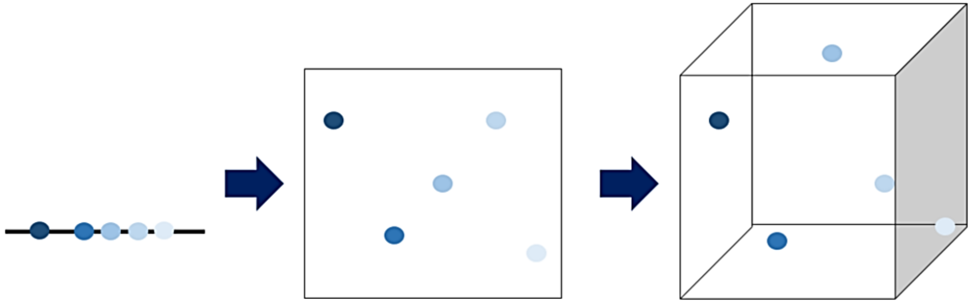
출처 : DeepLearning Wiki



출처 : 데이콘

DNN — 모든 퍼셉트론이 연결되어 있는 구조

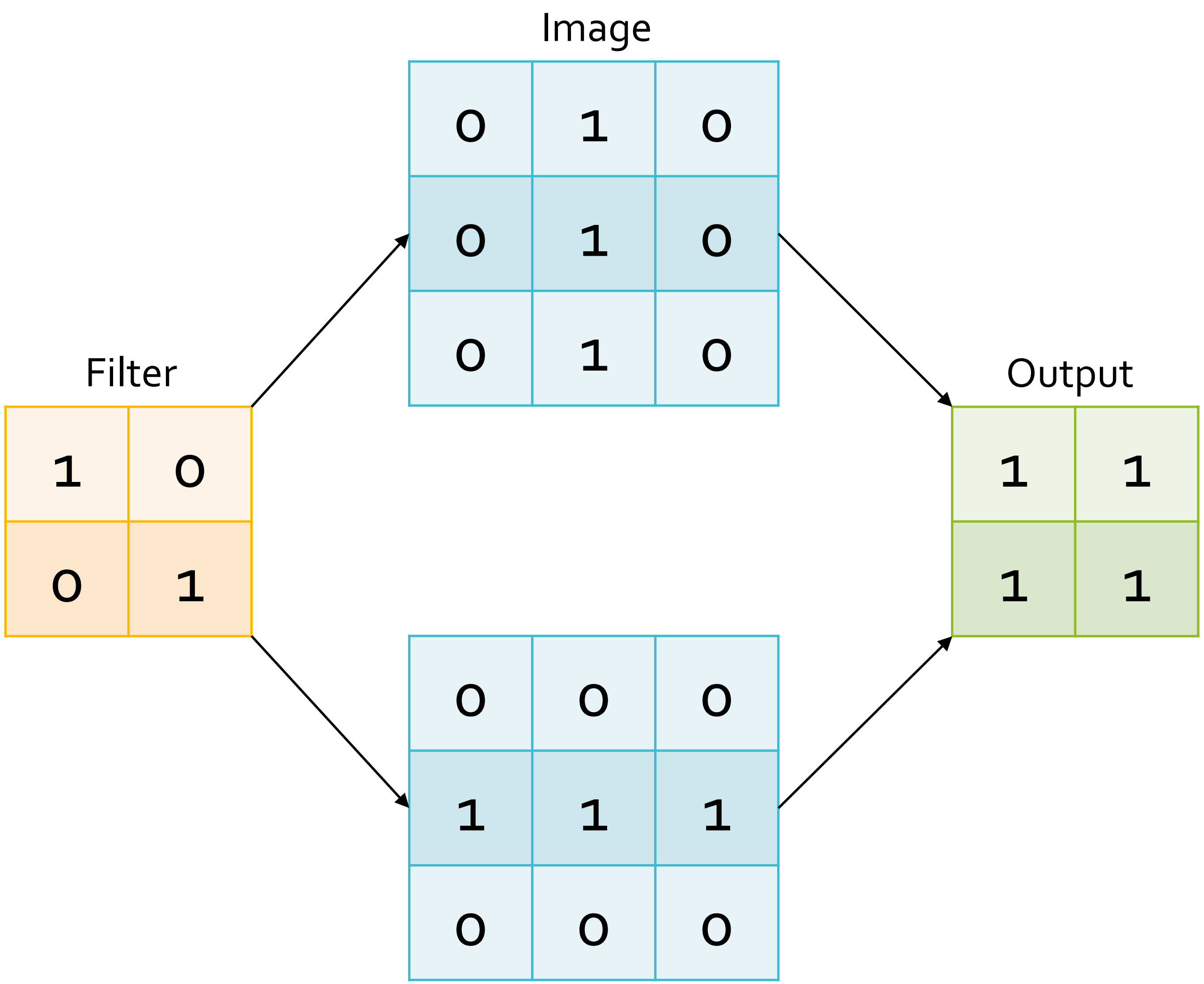
특징 — 이미지의 학습 과정에서 이미지가 이동하면 예측 정확도가 좋지 않음.



Made by: ta-daa

차원의 저주 — 차원이클수록예측이어려우며,DNN의문제점을개선할방안이필요하였음.

About CNN



Filter

1	0
0	1

Image

0	1	0
0	1	0
0	1	0

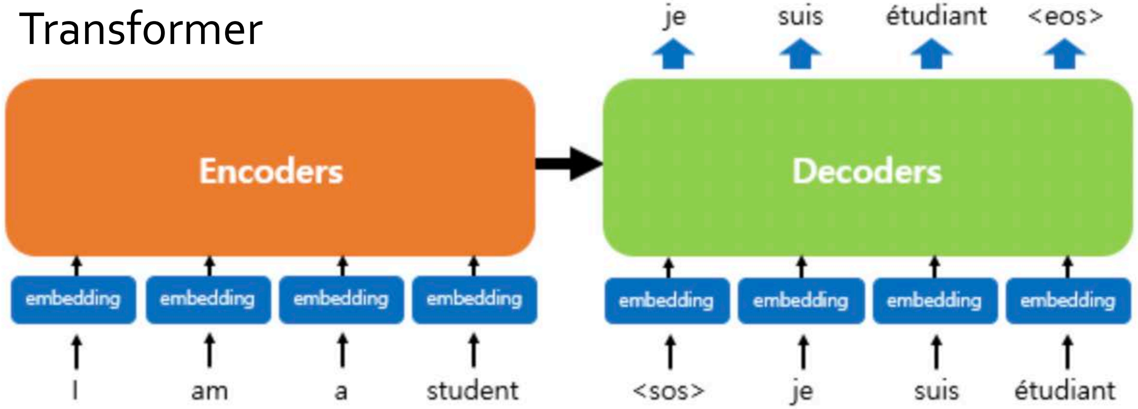
0	0	0
1	1	1
0	0	0

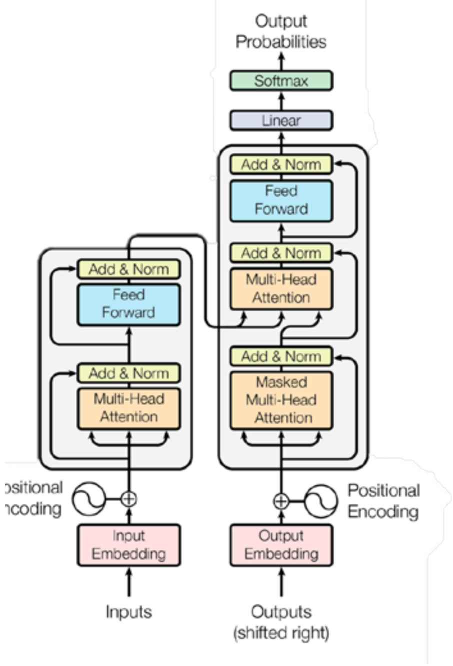
Output

1	1
1	1

About ViT

Transformer

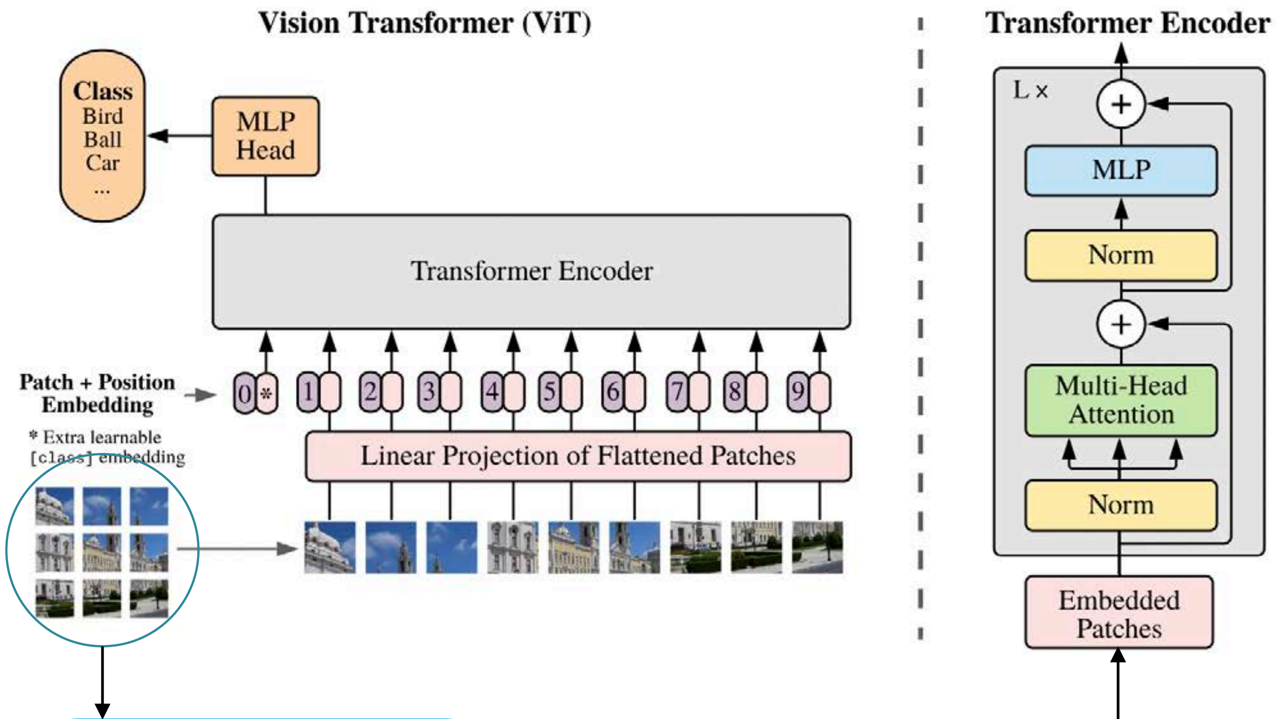




Transformer 자연어 분야에서 문장 번역을 위해 고안된 모델.

특징 학습 데이터의 양이 증가할수록 모델의 성능 개선

About ViT



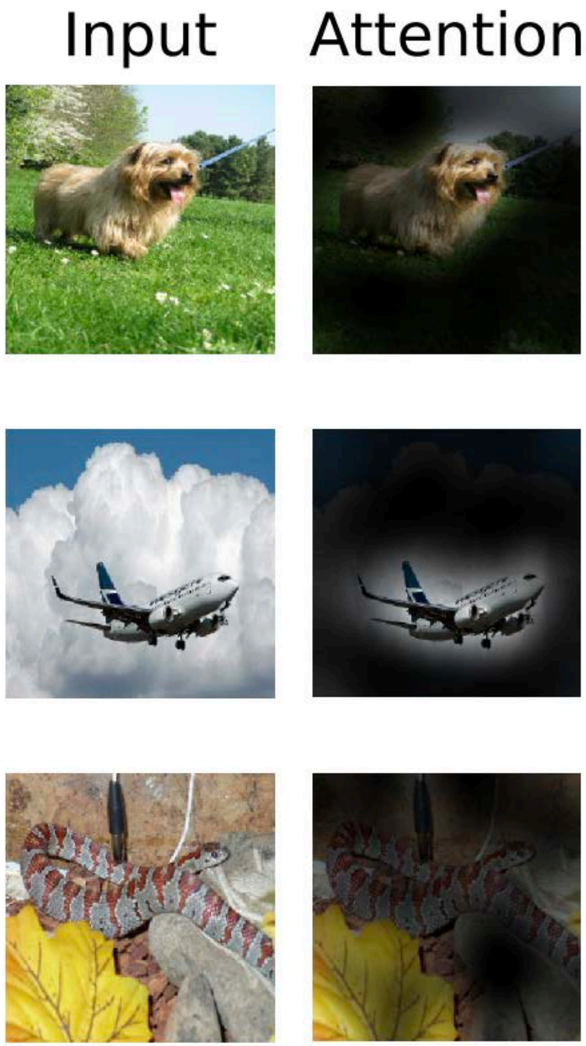
$$\mathbf{x} \in \mathbb{R}^{H \times W \times C}$$

$$\mathbf{x}_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$$

$$N = HW/P^2$$

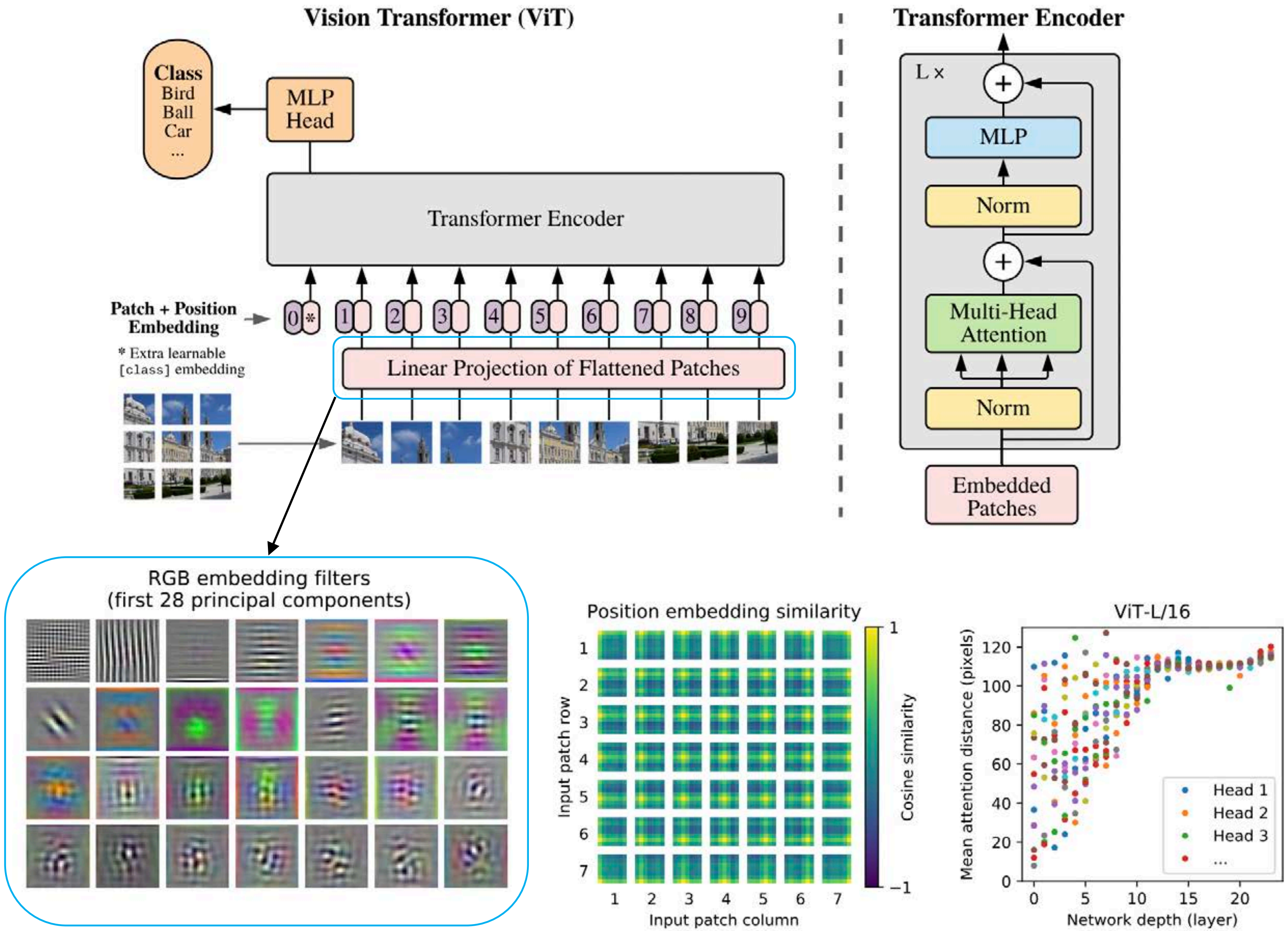
$$\begin{aligned}
\mathbf{z}_0 &= [\mathbf{x}_{\text{class}};\, \mathbf{x}_p^1\mathbf{E};\, \mathbf{x}_p^2\mathbf{E};\, \cdots;\, \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}, \\
\mathbf{z'}_\ell &= \mathrm{MSA}(\mathrm{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \\
\mathbf{z}_\ell &= \mathrm{MLP}(\mathrm{LN}(\mathbf{z'}_\ell)) + \mathbf{z'}_\ell, \\
\mathbf{y} &= \mathrm{LN}(\mathbf{z}_L^0)
\end{aligned}$$

Attention

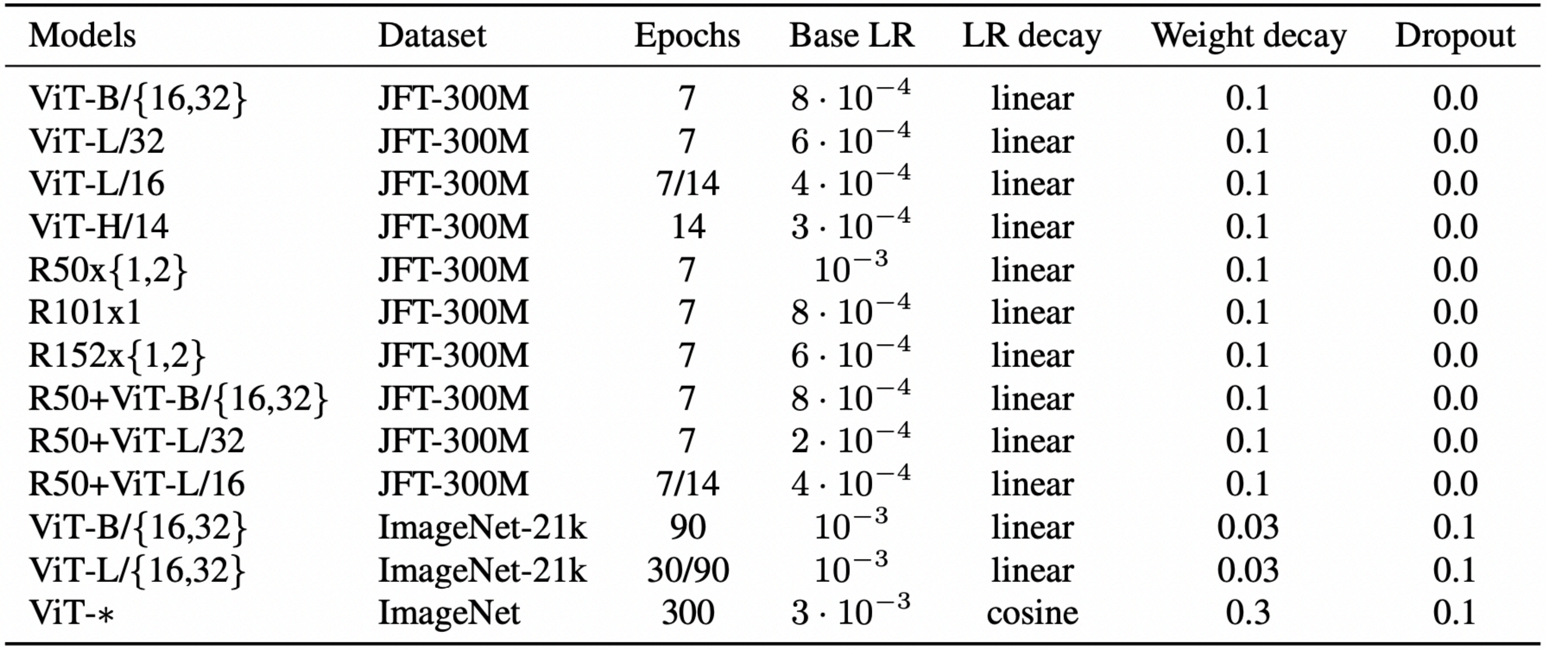


- A learned Position embedding is added to the patch representation.
- Closer patches tend to have more simmilar position embeddings.
- Patches in the same row/column have simmilar embeddings.
- This allows VIT to integrate information across the entire image even in the lowest layers

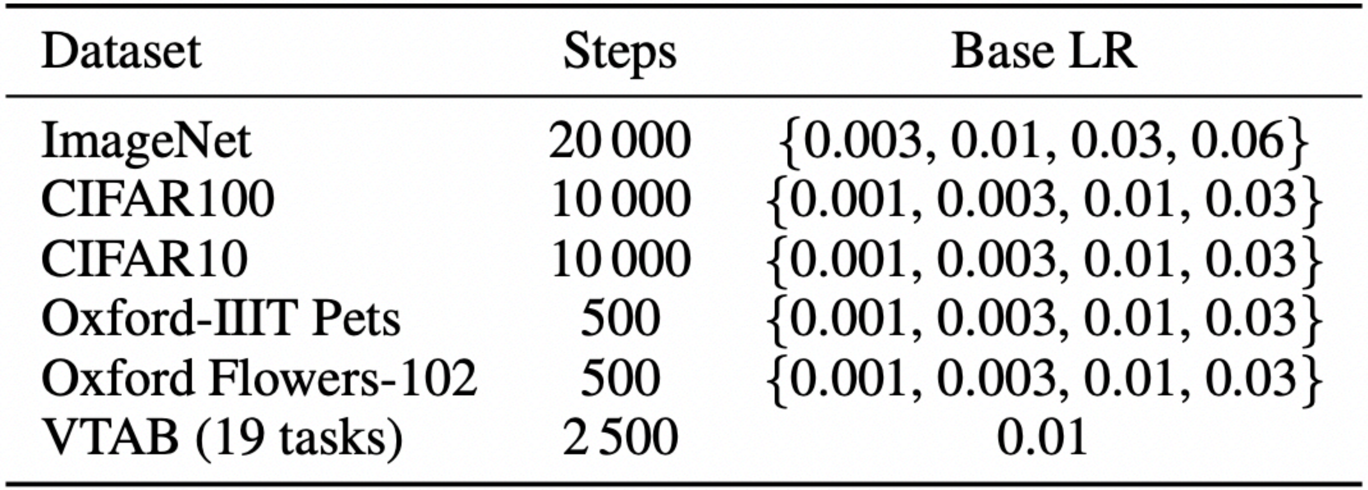
Self-Supervision



Hyperparameters



Models	Dataset	Epochs	Base LR	LR decay	Weight decay	Dropout
ViT-B/{16,32}	JFT-300M	7	$8 \cdot 10^{-4}$	linear	0.1	0.0
ViT-L/32	JFT-300M	7	$6 \cdot 10^{-4}$	linear	0.1	0.0
ViT-L/16	JFT-300M	7/14	$4 \cdot 10^{-4}$	linear	0.1	0.0
ViT-H/14	JFT-300M	14	$3 \cdot 10^{-4}$	linear	0.1	0.0
R50x{1,2}	JFT-300M	7	10^{-3}	linear	0.1	0.0
R101x1	JFT-300M	7	$8 \cdot 10^{-4}$	linear	0.1	0.0
R152x{1,2}	JFT-300M	7	$6 \cdot 10^{-4}$	linear	0.1	0.0
R50+ViT-B/{16,32}	JFT-300M	7	$8 \cdot 10^{-4}$	linear	0.1	0.0
R50+ViT-L/32	JFT-300M	7	$2 \cdot 10^{-4}$	linear	0.1	0.0
R50+ViT-L/16	JFT-300M	7/14	$4 \cdot 10^{-4}$	linear	0.1	0.0
ViT-B/{16,32}	ImageNet-21k	90	10^{-3}	linear	0.03	0.1
ViT-L/{16,32}	ImageNet-21k	30/90	10^{-3}	linear	0.03	0.1
ViT-$*$	ImageNet	300	$3 \cdot 10^{-3}$	cosine	0.3	0.1



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

Q&A