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

ICLR 2021 (under review)

Sungman Cho.

Introduction

- **Self-attention** based architectures, in particular Transformers, have become the model of choice in NLP.

- Inspired by NLP successes, multiple works try combining CNN-like architectures with self-attention.

    - Non-local Neural Network (CVPR, 2018)
    - DETR (ECCV, 2020)
    - Stand-alone self-attention in vision models (NeurIPS, 2019)
    - Stand-alone axial-attention for panoptic segmentation (ECCV, 2020)
    - Exploring the limits of weakly supervised pretraining (ECCV, 2018)
    - Self-training with noisy student improves imagenet classification (CVPR, 2020)
    - Big transfer (BiT): General visual representation learning (ECCV, 2020)

Introduction

- Such models yield modest results when trained on mid-sized datasets such as ImageNet, **achieving accuracies of a few percentage points below ResNets of comparable size.**

  : Transformers lack some inductive biases inherent to CNNS, such as translation equivariance and locality.

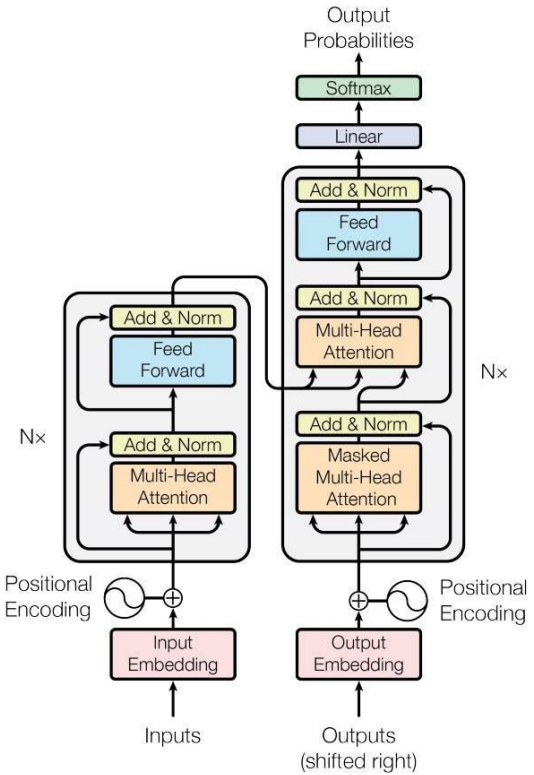
- We find that **large scale training trumps inductive bias.**

Related Works

- Transformers were proposed by Vaswani et al. (2017) for machine translation.

- Large Transformer-based models are often pre-trained on large corpora and then fine-tuned for the task at hand

  : BERT
  : GPT



Related Works

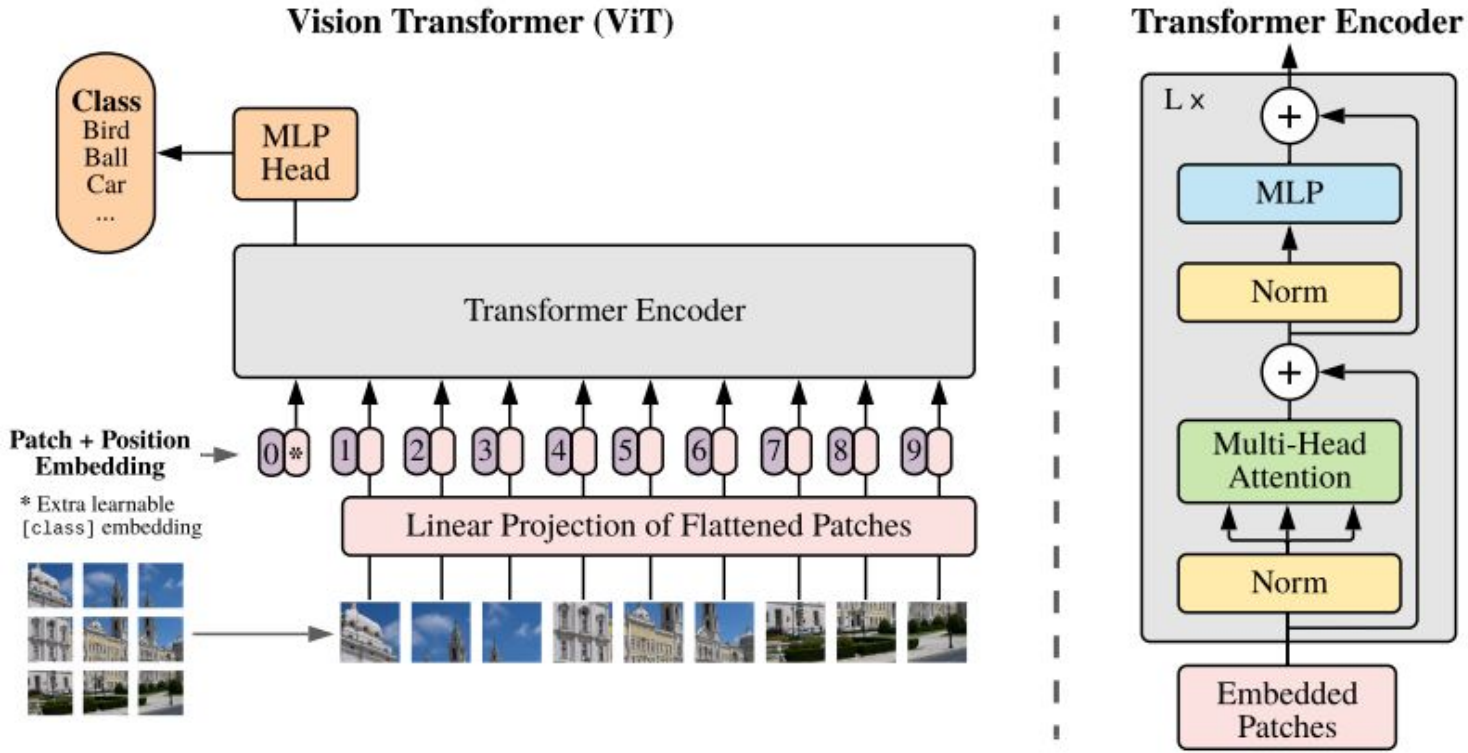
- **Naive application** of self-attention to images would **require that each pixel attends** to every other pixel.

  : Quadratic cost in the number of pixels !

- **Several approximations have been tried in the past.**
  (Require complex engineering)

  : Image Transformer (ICML, 2018)
  : Stand-alone self-attention in vision models (NeurIPS, 2019)
  : On the relationship between self-attention and convolutional layers (ICLR, 2020)
  : Exploring self-attention for image recognition (CVPR, 2020)
  : Generating long sequences with sparse transformers (arXiv, 2019)

Vision Transformer (VIT)



Datasets

- ImageNet (1k classes, 1.3M images)
- ImageNet-21k (21k classes, 14M images)
- JFT (18k classes, 303M images)
- **Validation**
  : CIFAR 10/100, Oxford-IIIT Pets, Oxford Flowers-102

Training & Fine-tuning

- Optimizer : Adam
- Batch size : 4096
- Weight decay of 0.1

Experiments: SOTA

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

	Ours (ViT-H/14)	Ours (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	88.36	87.61 ± 0.03	87.54 ± 0.02	$88.4/\mathbf{88.5}^*$
ImageNet ReaL	**90.77**	90.24 ± 0.03	90.54	90.55
CIFAR-10	$\mathbf{99.50} \pm 0.06$	99.42 ± 0.03	99.37 ± 0.06	–
CIFAR-100	$\mathbf{94.55} \pm 0.04$	93.90 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	$\mathbf{97.56} \pm 0.03$	97.32 ± 0.11	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	$\mathbf{99.74} \pm 0.00$	99.63 ± 0.03	–
VTAB (19 tasks)	$\mathbf{77.16} \pm 0.29$	75.91 ± 0.18	76.29 ± 1.70	–
TPUv3-days	2.5k	0.68k	9.9k	12.3k

Experiments: SOTA

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

	Ours (ViT-H/14)	Ours (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	88.36	87.61 ± 0.03	87.54 ± 0.02	$88.4/\mathbf{88.5}^*$
ImageNet ReaL	**90.77**	90.24 ± 0.03	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04	93.90 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03	97.32 ± 0.11	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	**99.74** ± 0.00	99.63 ± 0.03	–
VTAB (19 tasks)	**77.16** ± 0.29	75.91 ± 0.18	76.29 ± 1.70	–
TPUv3-days	2.5k	0.68k	9.9k	12.3k

Experiments: VTAB

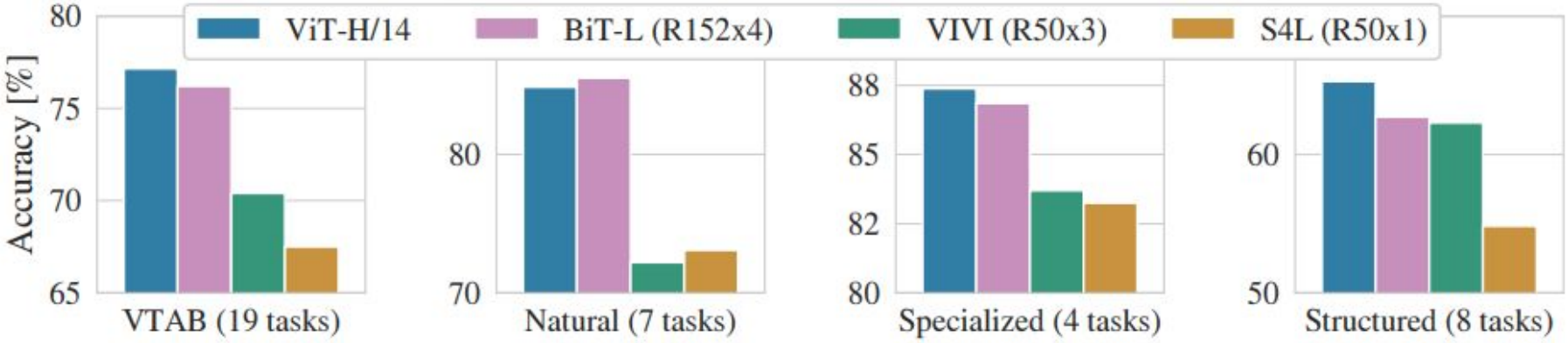
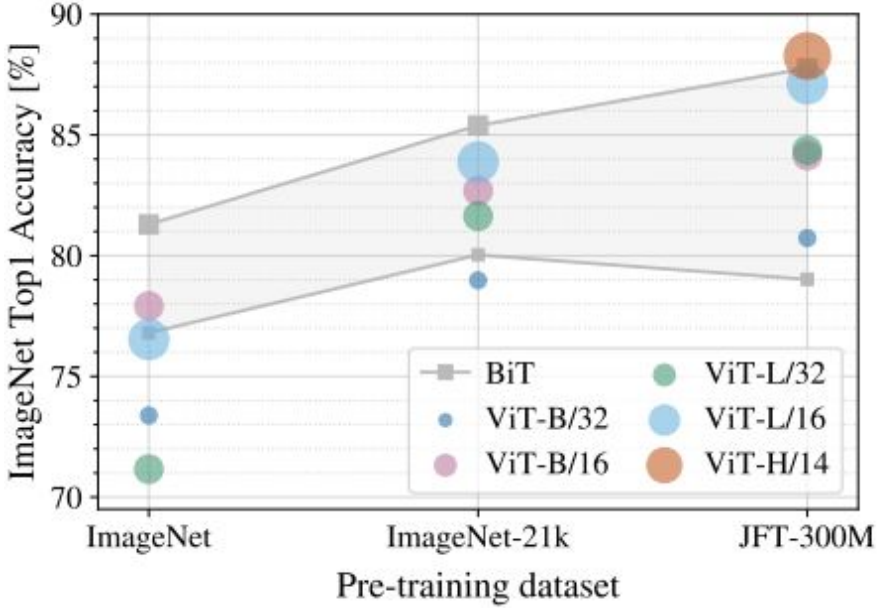
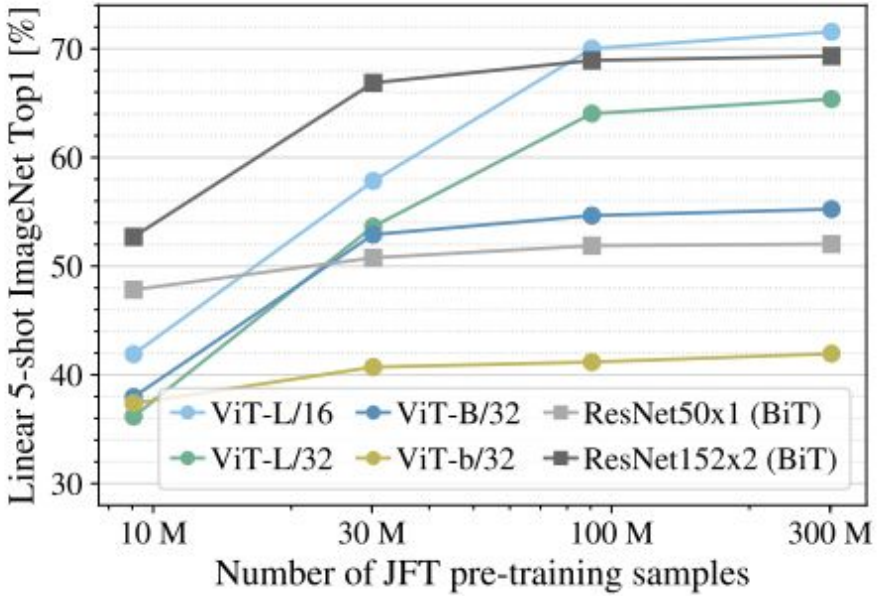


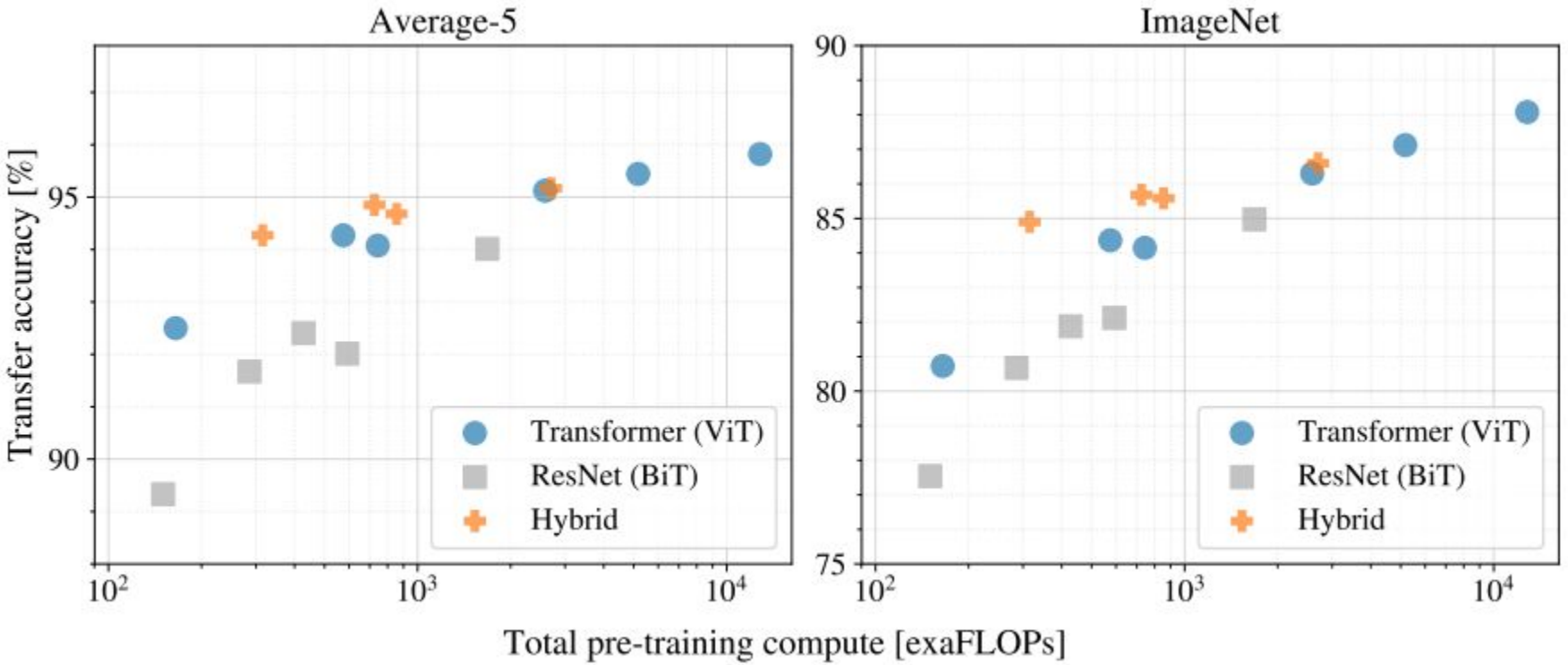
Figure 2: Breakdown of VTAB performance in *Natural*, *Specialized*, and *Structured* task groups.

Experiments: Transfer, Few-shot

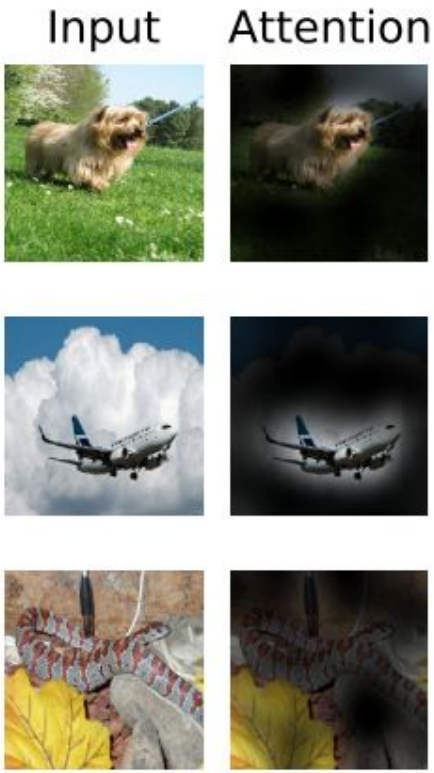


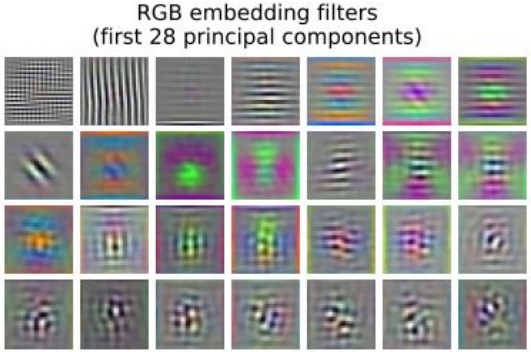


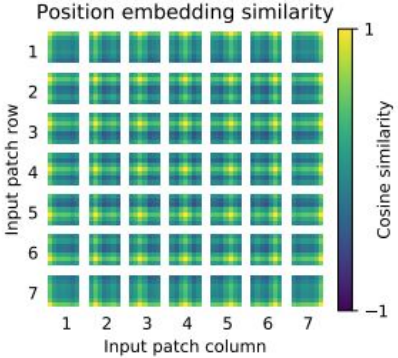
Experiments: Performance vs. Cost

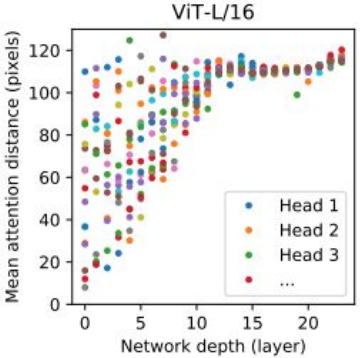


Experiments: Inspect









Conclusion

- Simple, yet Scalable, Strategy works surprisingly well when coupled with pre-training on large datasets.

- Vision Transformer matches or exceeds the SOTA on many image classification datasets, whilst being relatively cheap pre-train.

- Challenges
  : NOT detection and segmentation
  : NOT Self-supervised pre-training methods

APPENDIX

Experiments: SOTA

Pre-training Data	Dataset	ViT-B/16	ViT-B/32	ViT-L/16	ViT-L/32	ViT-H/14
ImageNet	CIFAR-10	98.13	97.77	97.86	97.94	-
	CIFAR-100	87.13	86.31	86.35	87.07	-
	ImageNet	77.91	73.38	76.53	71.16	-
	ImageNet ReaL	83.57	79.56	82.19	77.83	-
	Oxford Flowers-102	89.49	85.43	89.66	86.36	-
	Oxford-IIIT-Pets	93.81	92.04	93.64	91.35	-
ImageNet-21k	CIFAR-10	98.95	98.79	98.97	99.13	-
	CIFAR-100	91.67	91.97	91.73	93.04	-
	ImageNet	83.97	81.28	79.45	80.99	-
	ImageNet ReaL	88.35	86.63	83.38	85.65	-
	Oxford Flowers-102	99.38	99.11	99.09	99.19	-
	Oxford-IIIT-Pets	94.43	93.02	93.18	93.09	-
JFT-300M	CIFAR-10	99.00	98.61	99.42	99.19	99.50
	CIFAR-100	91.87	90.49	93.90	92.52	94.55
	ImageNet	84.15	80.73	87.61	84.37	88.36
	ImageNet ReaL	88.85	86.27	90.24	88.28	90.77
	Oxford Flowers-102	99.56	99.27	99.74	99.45	99.68
	Oxford-IIIT-Pets	95.80	93.40	97.32	95.83	97.56

Experiments: Hyperparameters (train)

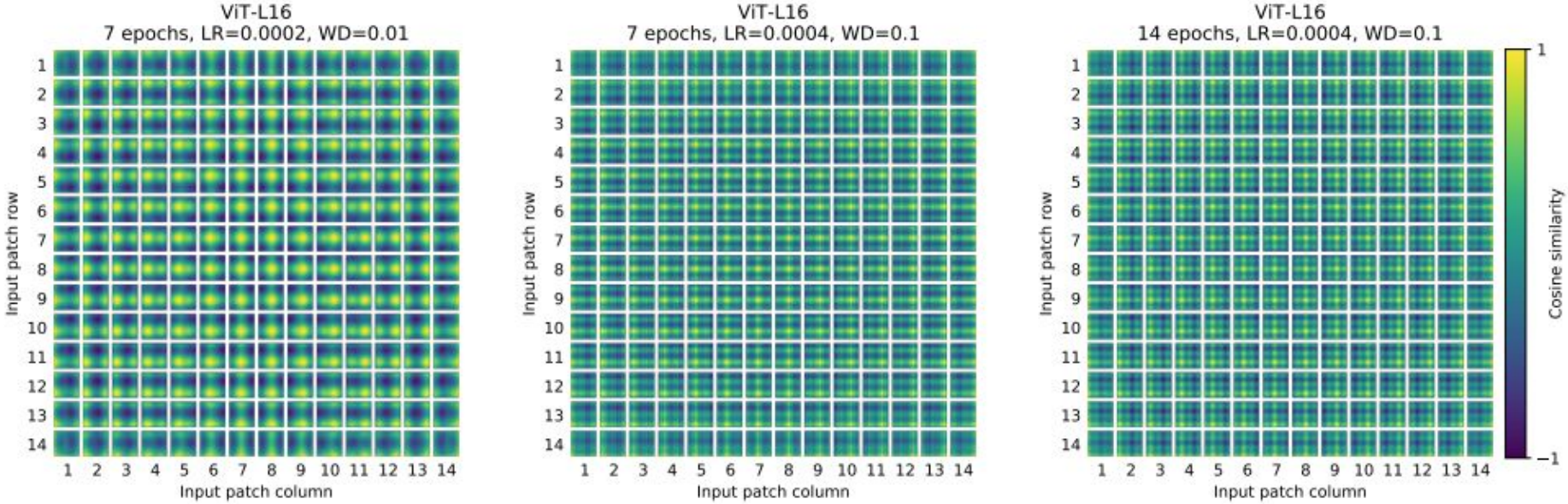
Models	Dataset	Epochs	Base LR	LR decay	Weight decay	Dropout
ViT-B-{16,32}	JFT-300M	7	$8 \cdot 10^{-4}$	linear	0.1	0.0
ViT-L-32	JFT-300M	7	$6 \cdot 10^{-4}$	linear	0.1	0.0
ViT-L-16	JFT-300M	7/14	$4 \cdot 10^{-4}$	linear	0.1	0.0
ViT-H-14	JFT-300M	14	$3 \cdot 10^{-4}$	linear	0.1	0.0
R50x{1,2}	JFT-300M	7	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
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

Experiments: Hyperparameters (fine-tune)

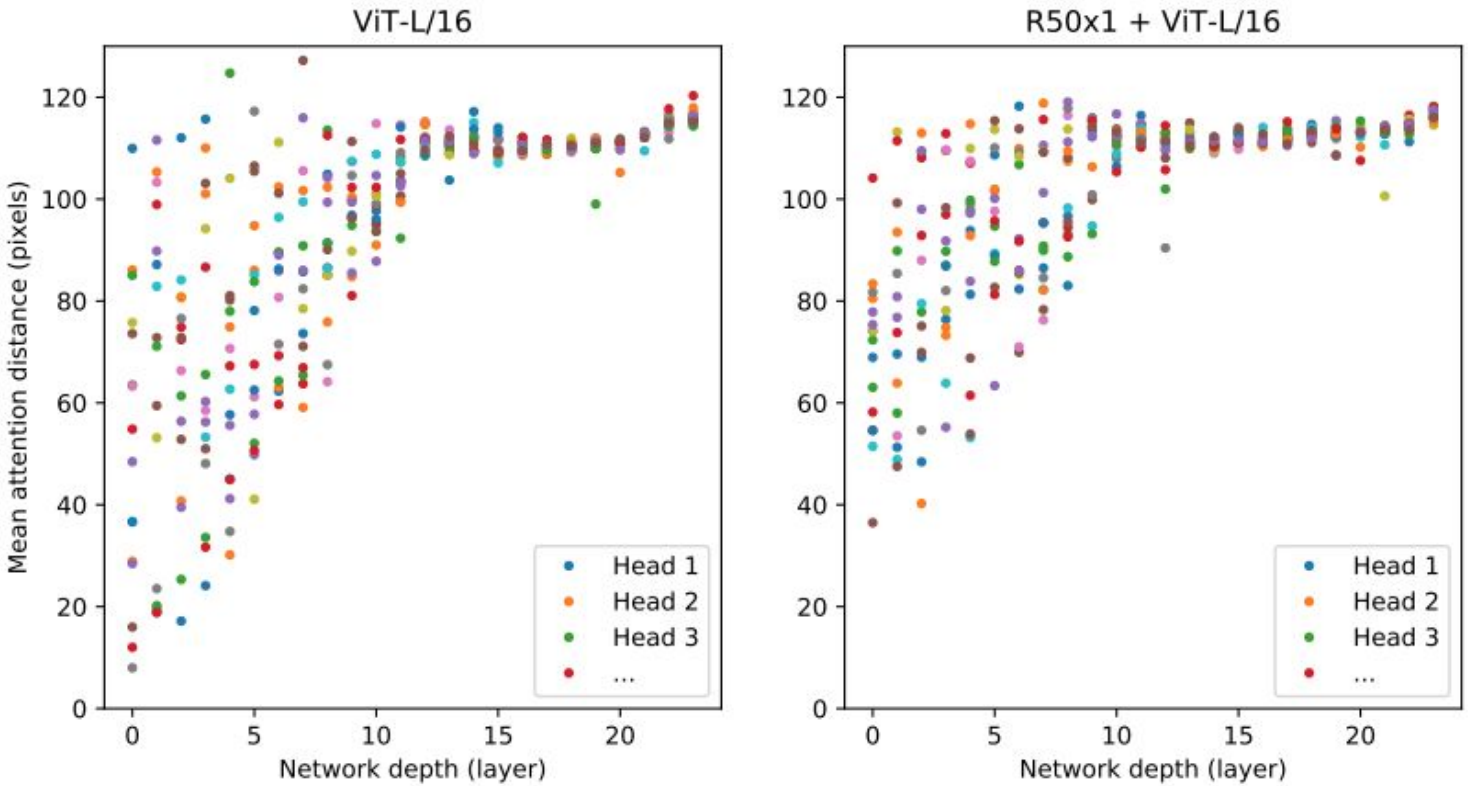
Dataset	Steps	Base LR
ImageNet	20 000	{0.003, 0.01, 0.03}
CIFAR100	10 000	{0.001, 0.003, 0.01, 0.03}
CIFAR10	10 000	{0.001, 0.003, 0.01, 0.03}
Oxford-IIIT Pets	500	{0.001, 0.003, 0.01, 0.03}
Oxford Flowers-102	500	{0.001, 0.003, 0.01, 0.03}
VTAB (19 tasks)	{1 000, 10 000}	{0.01, 0.1}

- Batch size: 512
- Resolution: 384
- Grad clipping at global norm 1.

Experiments: Positional Encoding



Experiments: Attention Distance



Thank You.