An image is worth 16x16 words: transformers for image recognition[1]

[1] https://openreview.net/forum?id=YicbFdNTTy

Introduction

Why?

- Well, transformers..
- Dominant approach in NLP: pre-train on large dataset fine-tune on smaller task
    - This is possible due to Transformer’s computational efficiency

Presenter: Eloy Geenjaar

Multi-headed attention

A quick recap/introduction

Introduction

Multi-headed attention

- Single-headed attention:

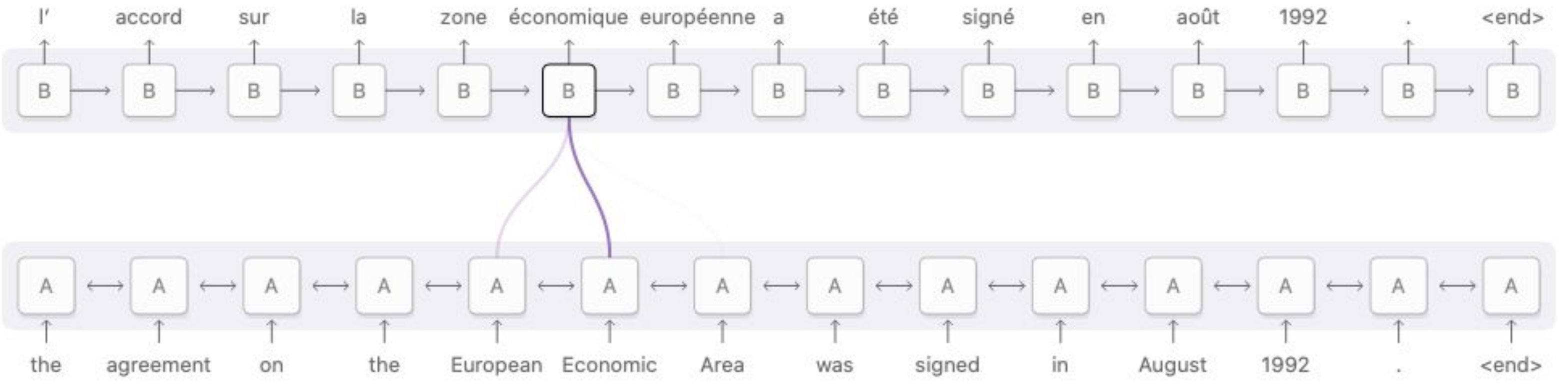


Diagram derived from Fig. 3 of Bahdanau, *et al.* 2014

[2] https://distill.pub/2016/augmented-rnns/

Introduction

Multi-headed attention

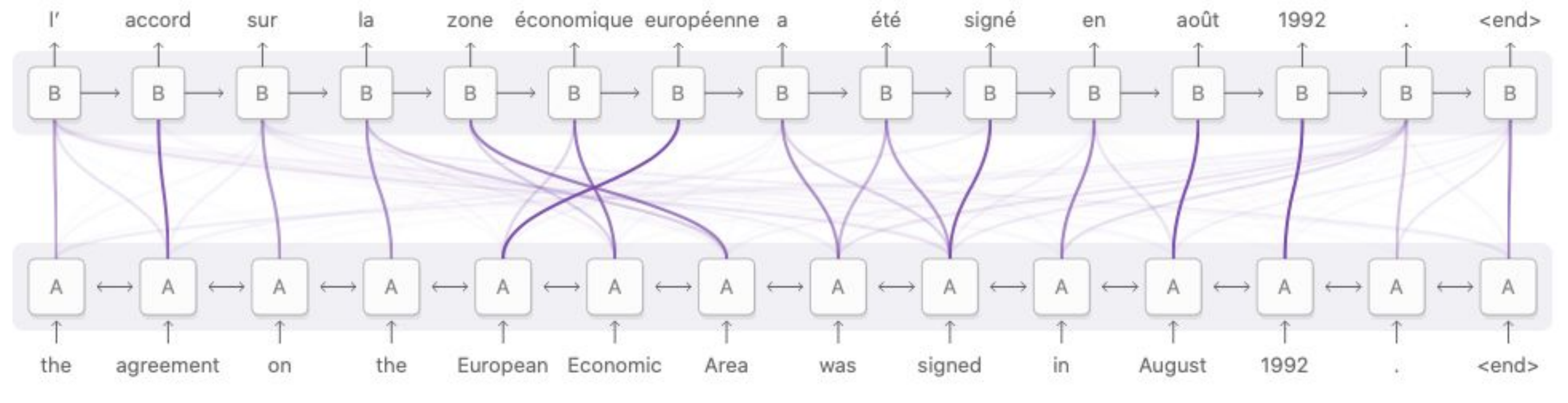


Diagram derived from Fig. 3 of Bahdanau, *et al.* 2014

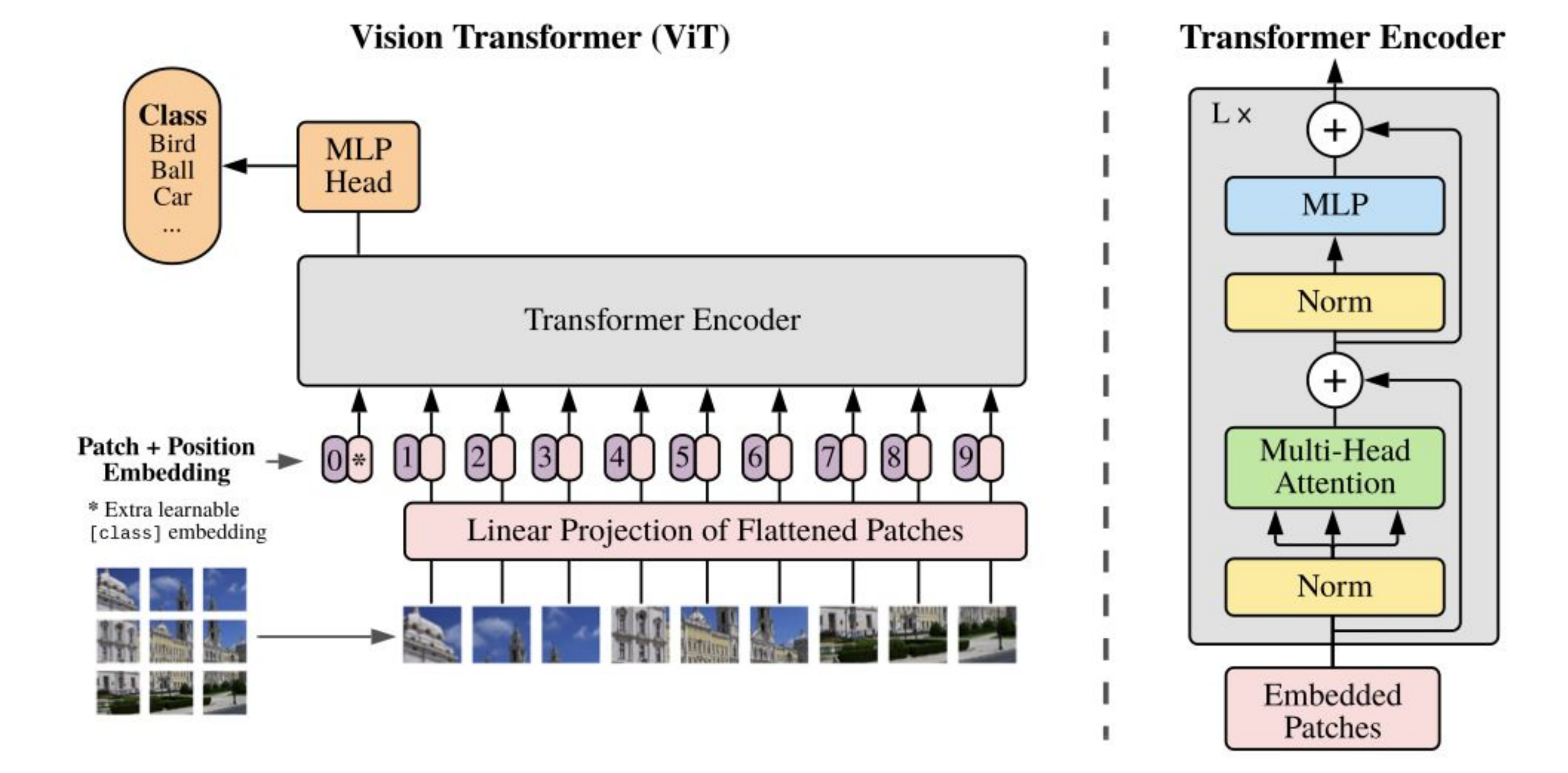
[2] https://distill.pub/2016/augmented-rnns/

ViT

Here we go

ViT

The architecture



The patches in the Visual Transformer are used in the same way as words are in NLP tasks

Why not earlier?

- Mid-sized datasets such as ImageNet require an inductive bias in the model to get a good performance on
- Transformers do not generalize well when trained on mid-sized datasets

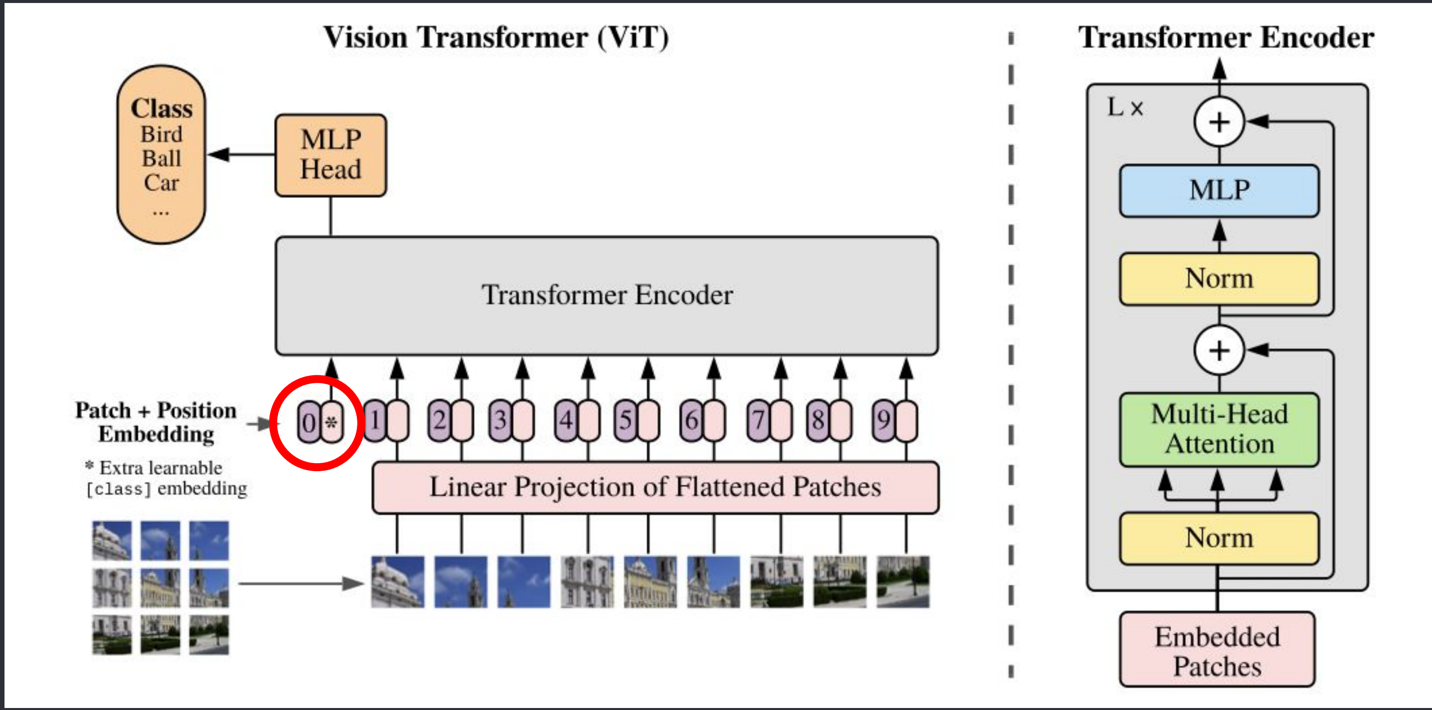
So what does this mean?

- Large scale training is more important than an inductive bias for SOTA results

- This is the logical next steps following a trend in image recognition at increasingly larger scales

How does it do classification

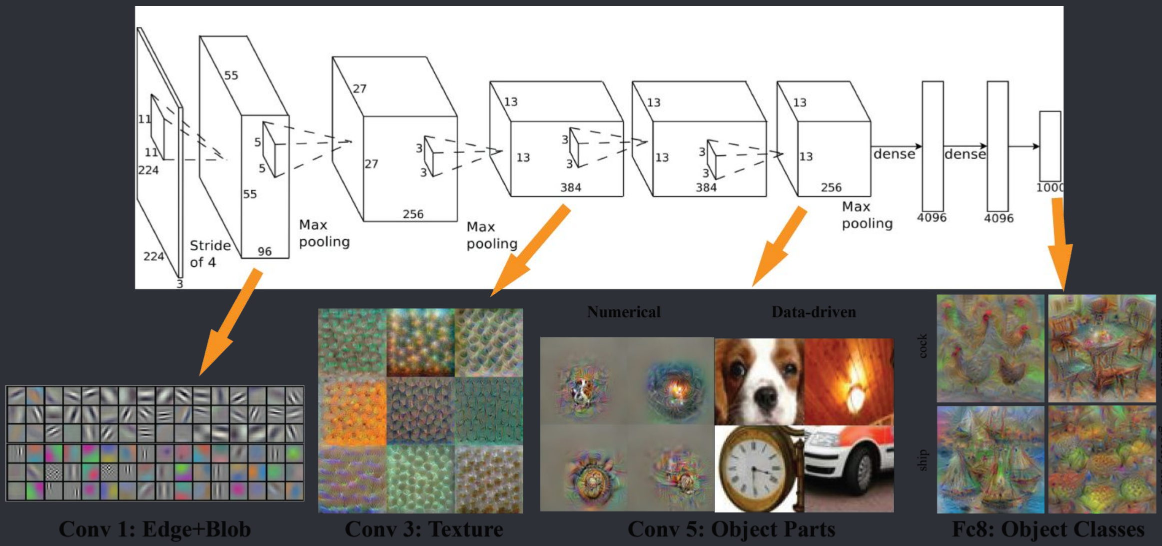
- Append a learnable embedding to the patches, this patch is used to predict the class



ViT

Hybrid architecture

- Use patches from feature map in early layers of a ResNet



[3] https://towardsdatascience.com/understanding-ssd-multibox-real-time-object-detection-in-deep-learning-495ef744fab

Fine-tuning

- Fine-tune on smaller task with images of a higher resolution with the same patch size
  - Leads to longer sequence of patches
  - Need to interpolate positional embeddings according to original pre-training resolution
  - Better performance

Configurations

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: Configuration of our different model variants.

ViT

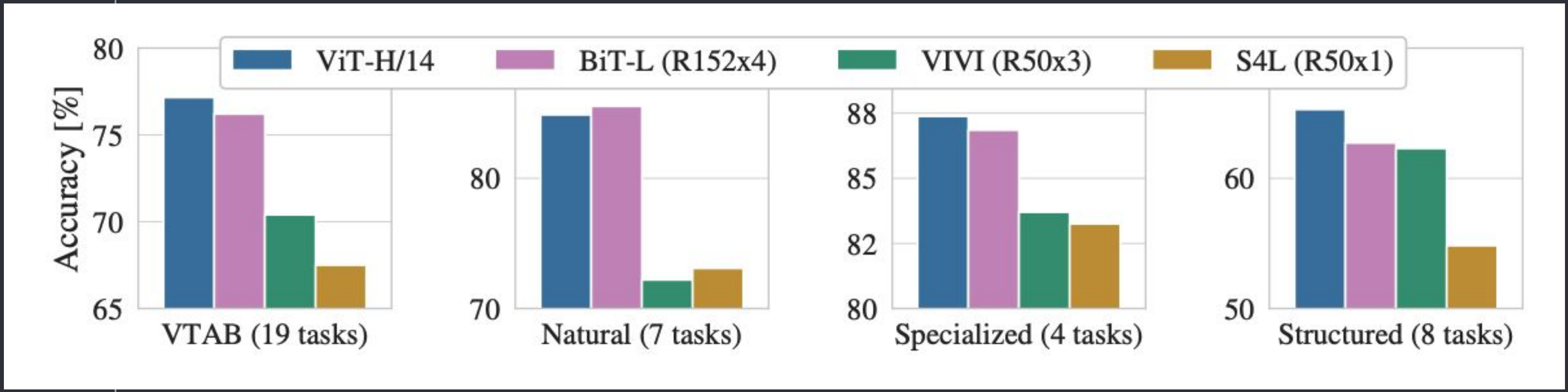
Results

	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

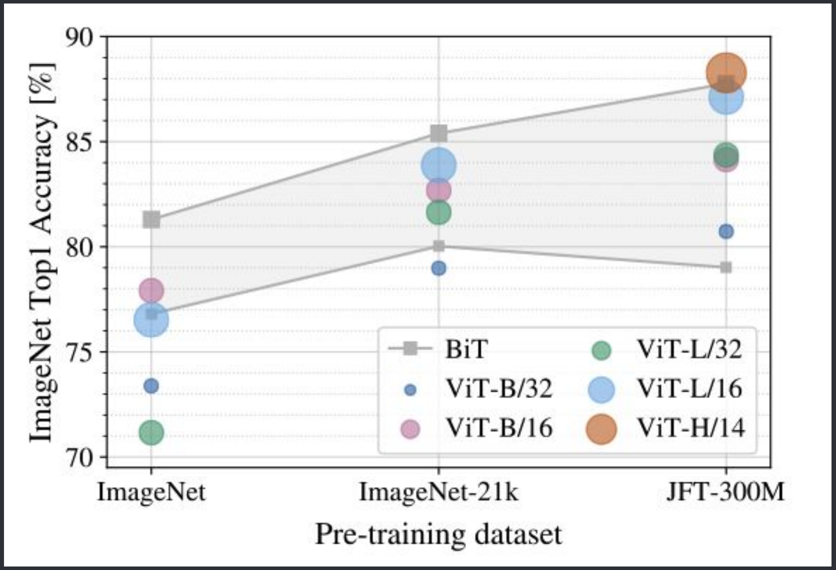
Pre-training efficiency may be affected by hyperparameters, architecture choice -> controlled study

VTAB

- 19 tasks, low data transfer: 1000 examples, 3 types of tasks:
  - Natural images: Pets, CIFAR-like task
  - Specialized: Medical, Satellite imagery
  - Structured: Tasks that require geometric understanding or localization

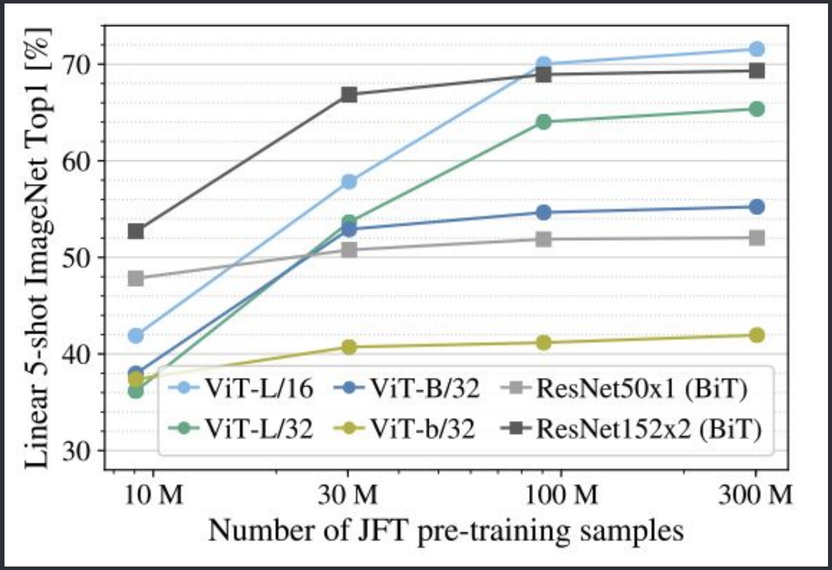


Evaluate pre-train size importance



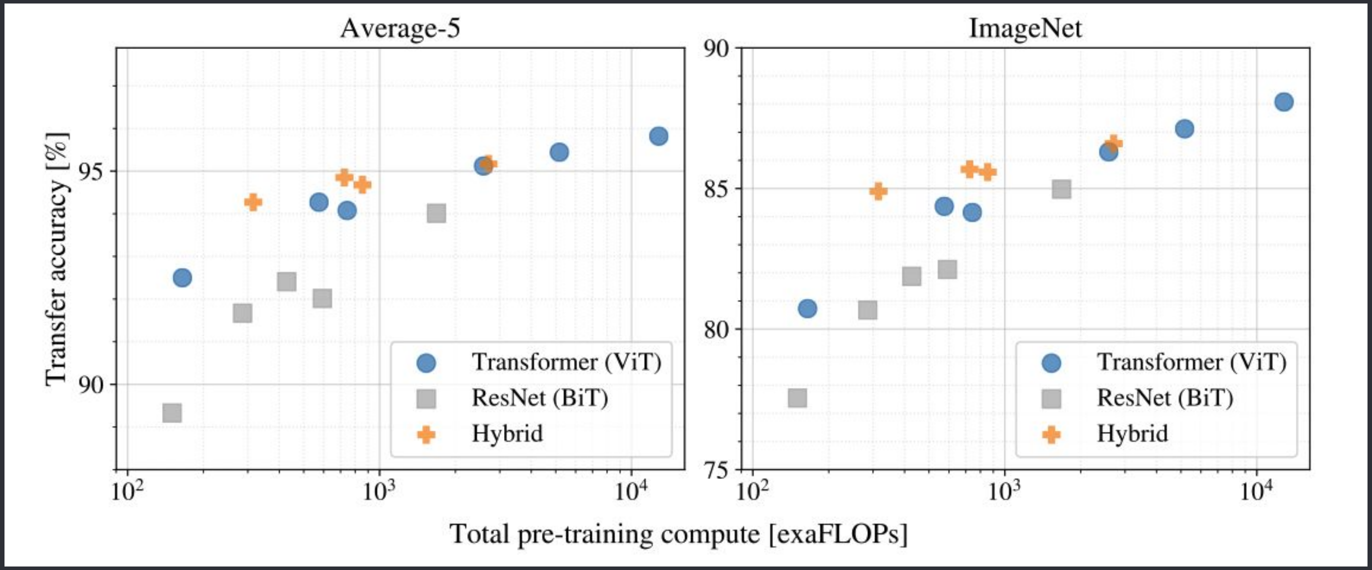
- Fine-tuning to ImageNet with hyperparam search and regularization optimization (Bigger ViTs get outperformed by smaller ViTs)

Evaluate pre-train size importance Pt. 2



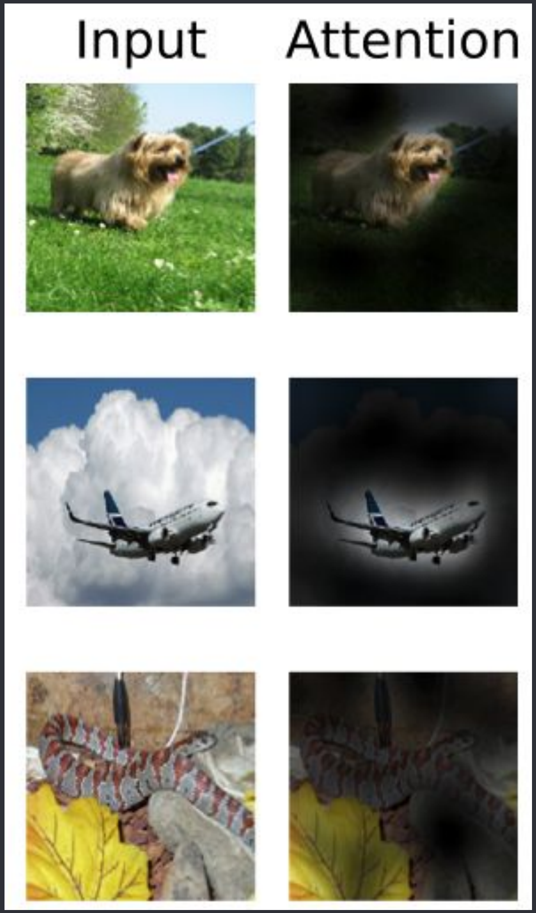
- Linear few-shot evaluation on ImageNet (no hyperparameter optimization nor regularization optimization) -> ViT overfits on smaller training subsets of JFT.

Scaling study

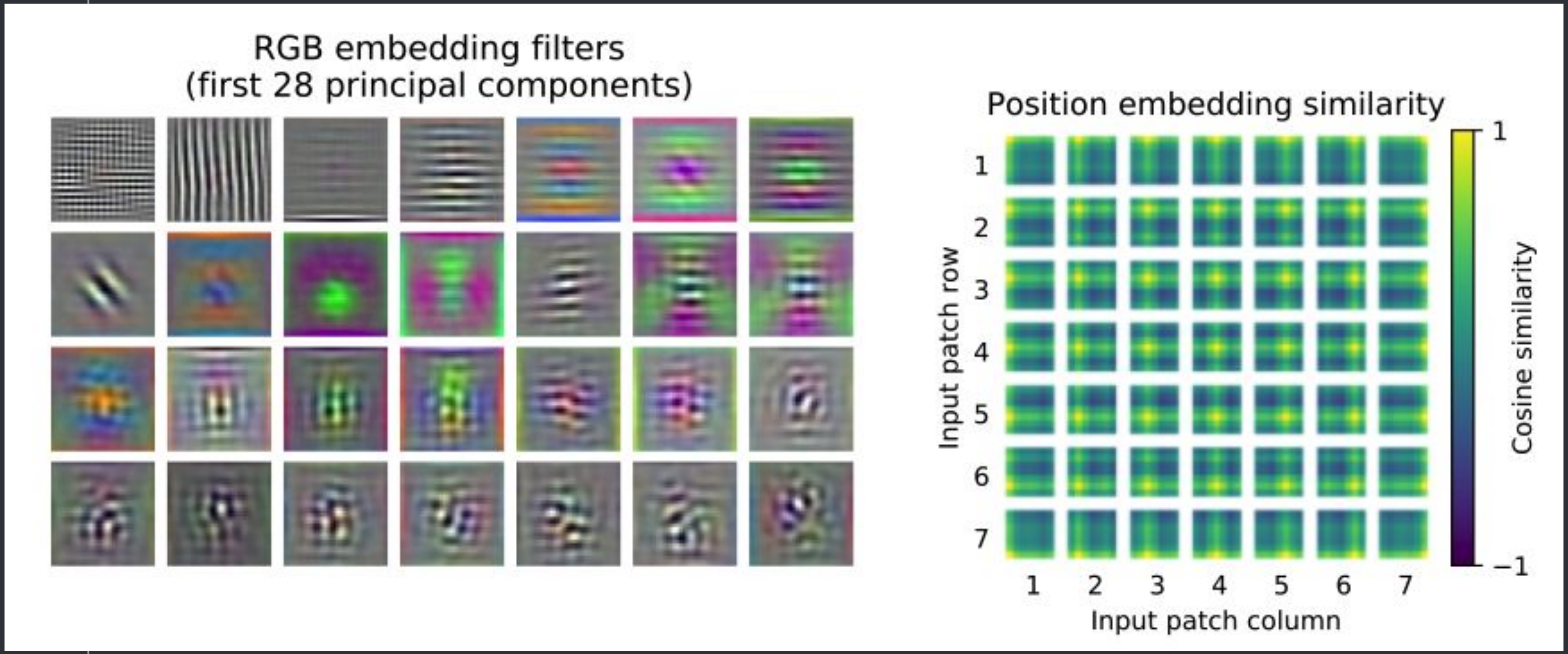


- ViTs outperform ResNets on performance/compute trade-off
- Hybrids outperform ViT on small computational budgets

What does it attend to?



What kind of embeddings does it learn?



All you need is depth?

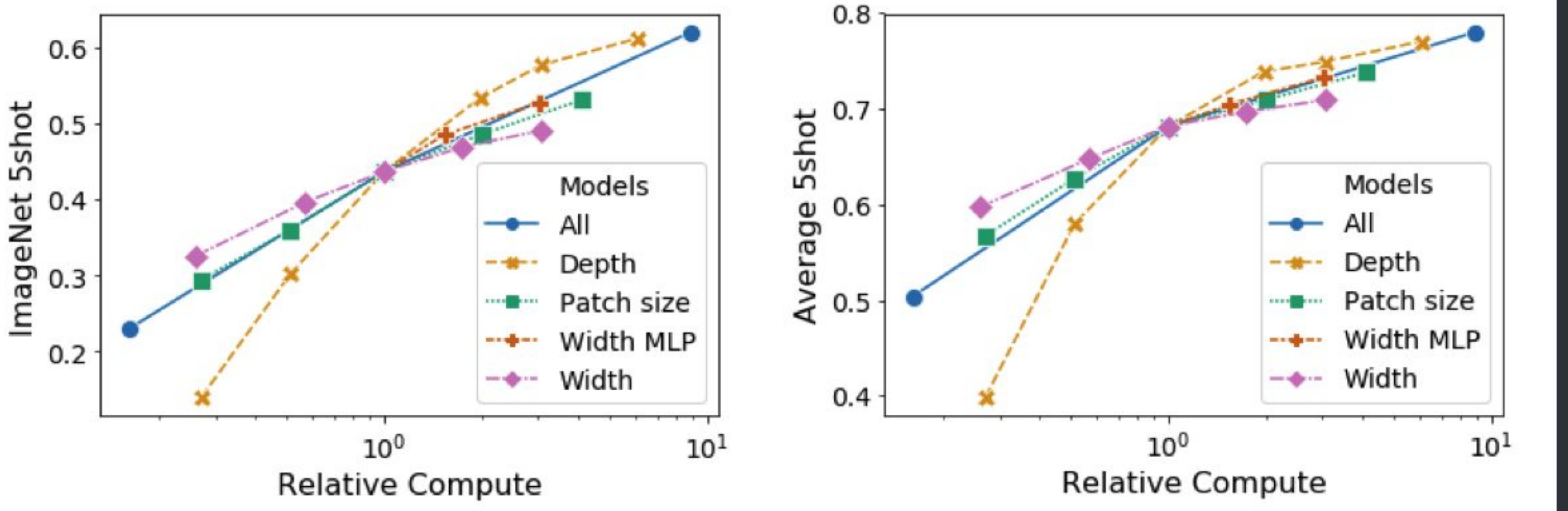
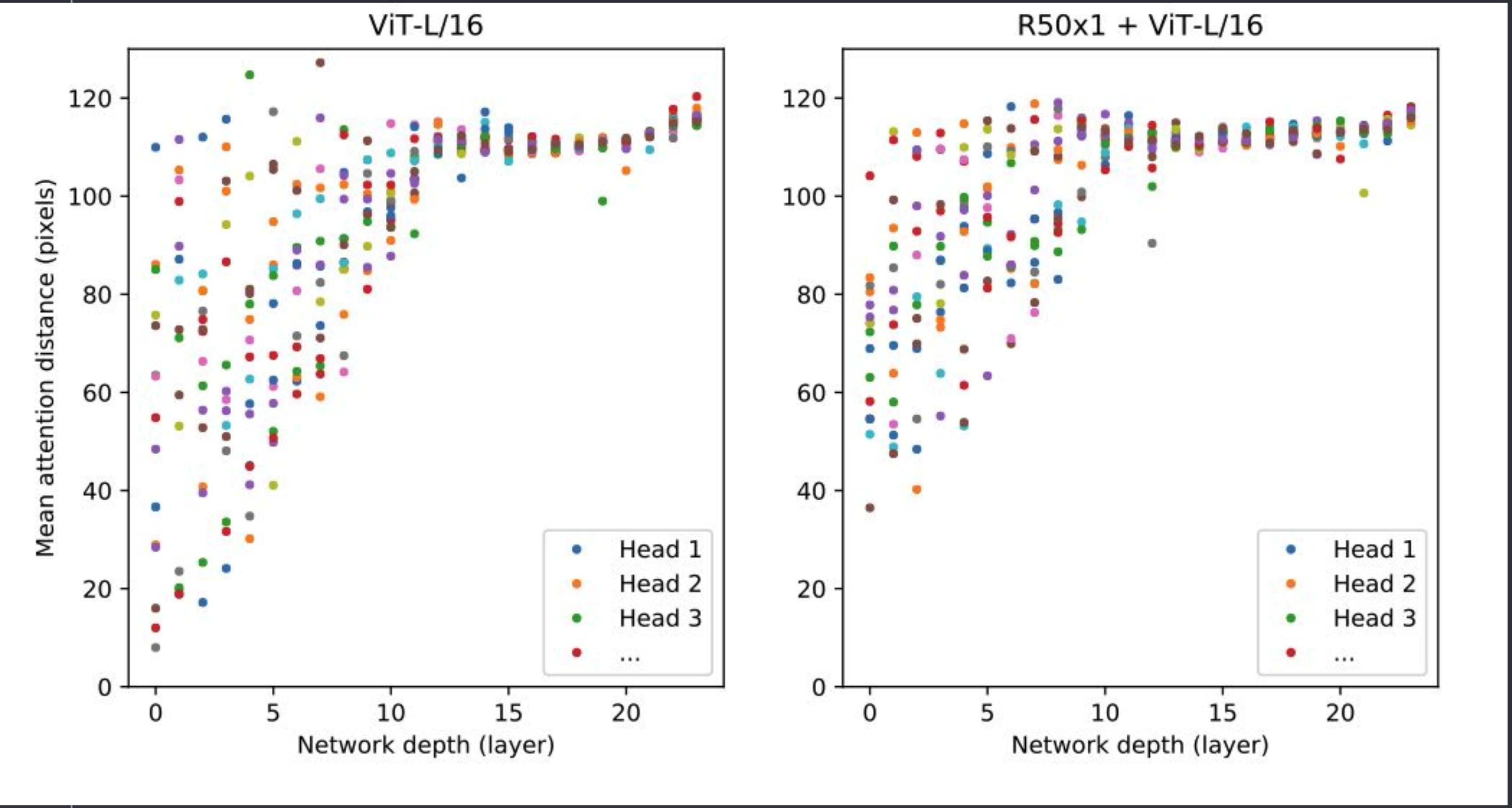


Figure 8: Scaling different model dimensions of the Vision Transformer.

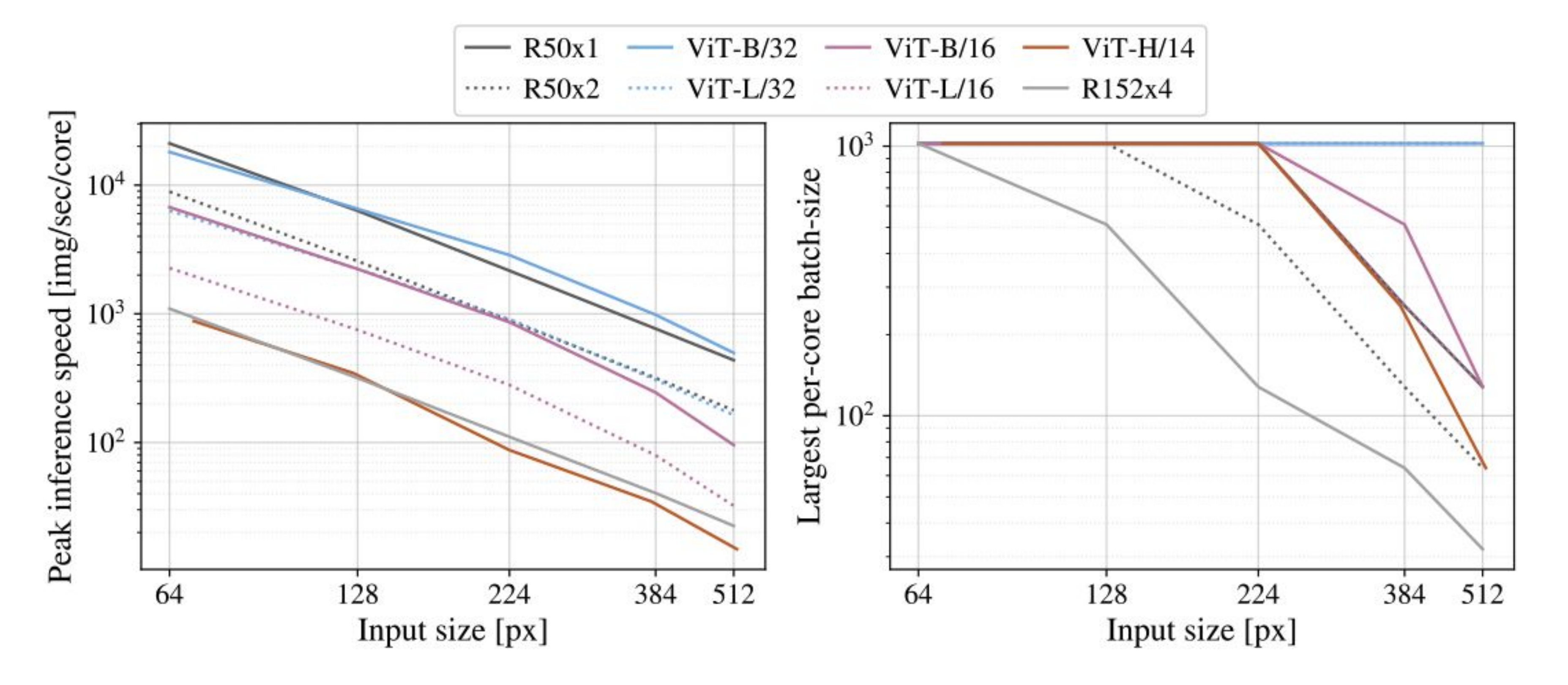
[4] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).

ViT

Attention distance



Inference


R50x1
ViT-B/32
ViT-B/16
ViT-H/14
R50x2
ViT-L/32
ViT-L/16
R152x4
Peak inference speed [img/sec/core]
10^4
10^3
10^2
64
128
224
384
512
Input size [px]
Largest per-core batch-size
10^3
10^2
64
128
224
384
512
Input size [px]

Self-supervised pre-training

- They do a test with masked patch prediction
  - Accuracy on ImageNet is 4% behind supervised pre-training

Future work

- Explore self-supervised pre-training instead of supervised pre-training
- Explore larger ViT models

Thanks!

ANY QUESTIONS?

Let's keep discussing the ideas and looking for ways to learn them deeper by applying them in unexpected ways