An Image is Worth 16x16 words: Transformers for Image Recognition at Scale

Alexey Dosovitskiy et al., 2021

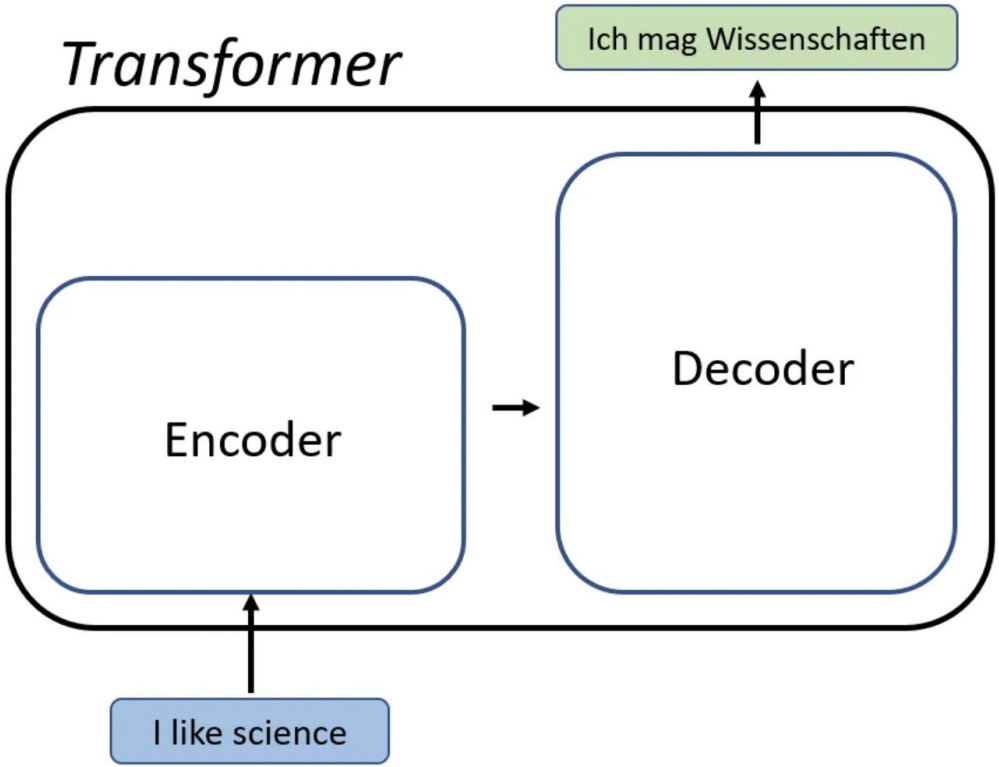
Presentation by Tiha Nikolova, 3th November 2023

Outline

1. Motivation
2. Architecture
3. Experiments
4. Inspecting the model

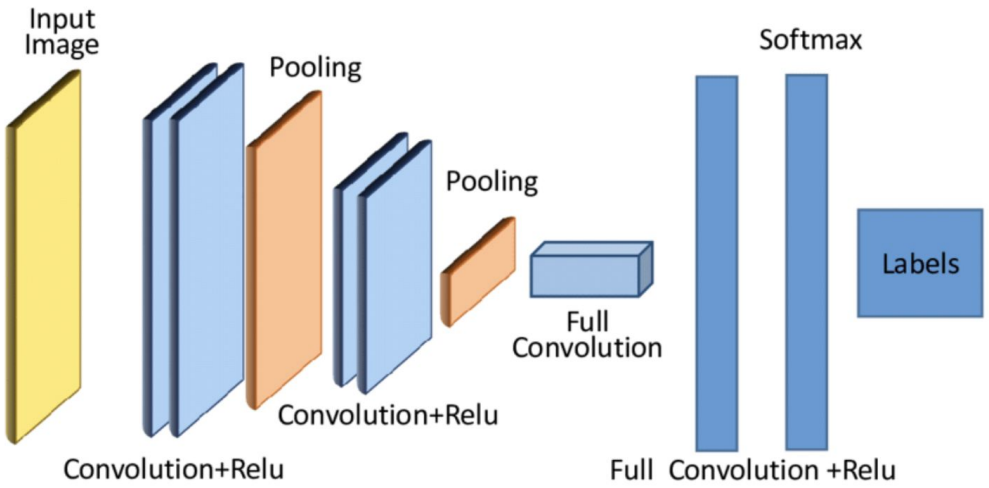
Motivation

- Transformers - model of choice for NLP tasks
- Pre-trained on large datasets and fine-tuned for the specific task
- Examples - BERT, GPT



Motivation

- In computer vision CNNs remain dominant
- Some works try to combine CNNs architectures with self-attention
  - naive application - each pixel attends to every other
  - attention only in local neighbours
  - attention to blocks of varying size



→ could not scale well

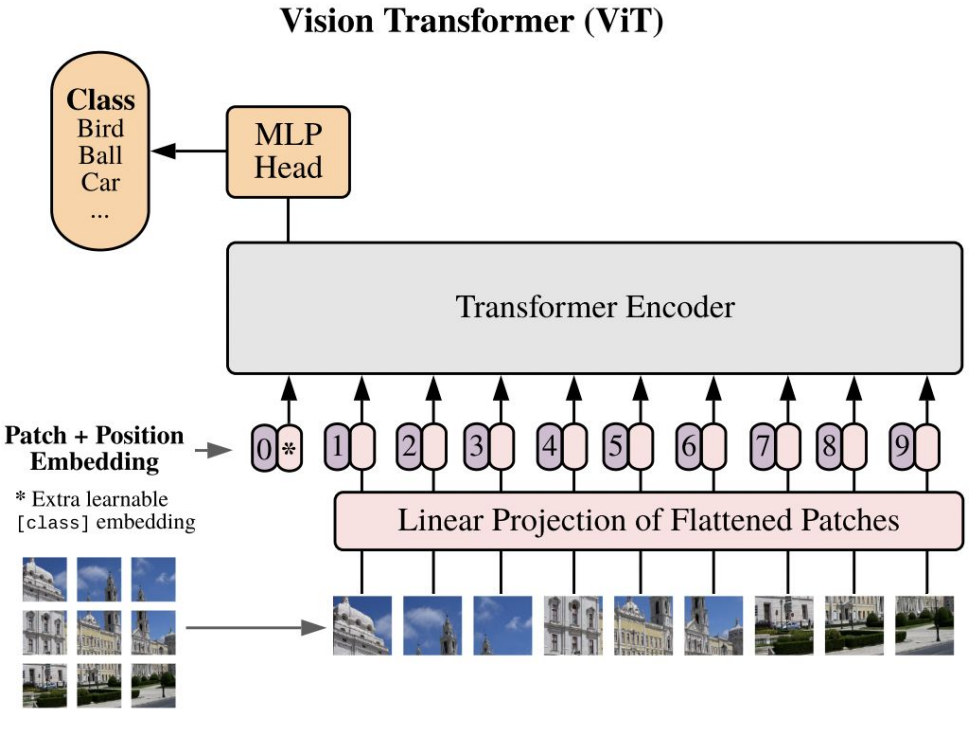
Motivation


Apply a standard Transformer directly to images, with the fewest possible modifications.

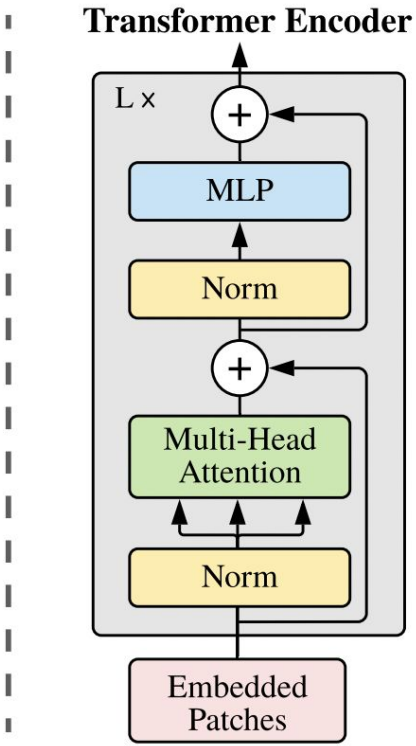
Motivation

- Basic idea: Split image into patches (words) and feed them in the Transformer
- Image recognition at *larger scales*
- Promising results
  - when trained on mid-sized dataset, the accuracies were a bit below these of ResNet
  - when trained on larger datasets, excellent results achieved

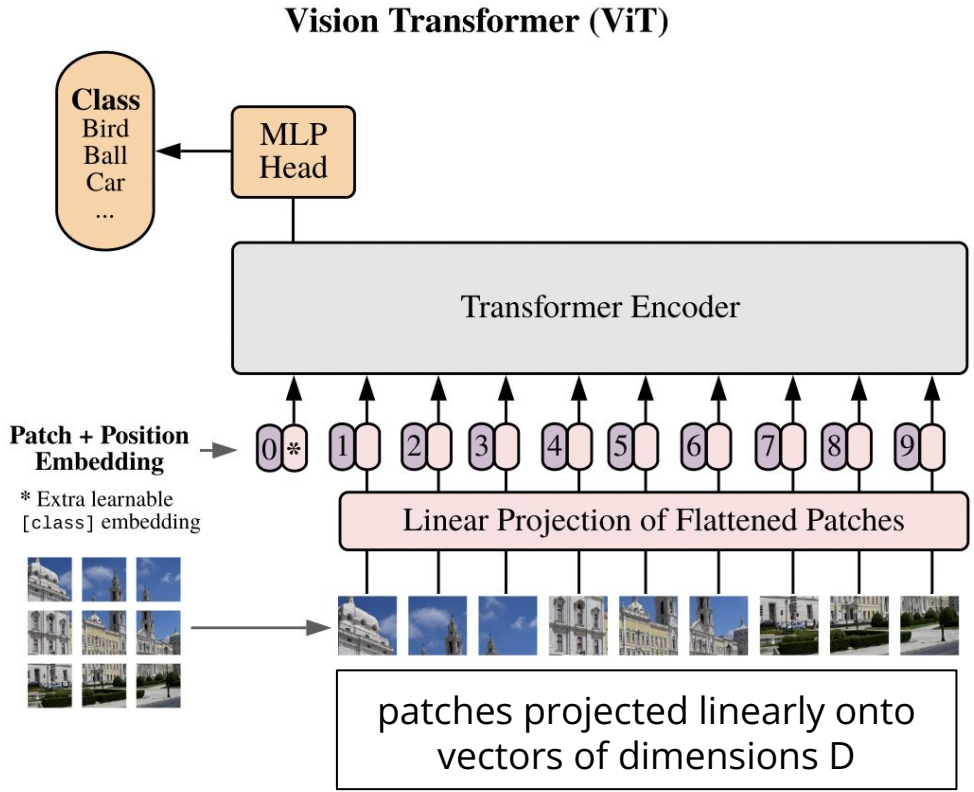
Architecture - Vision Transformer (ViT)

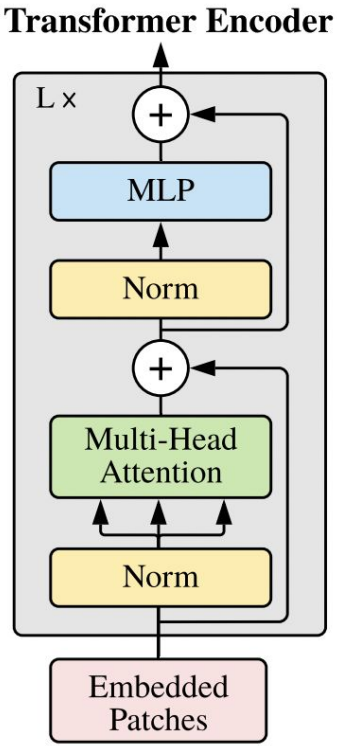


split image into patches



Architecture

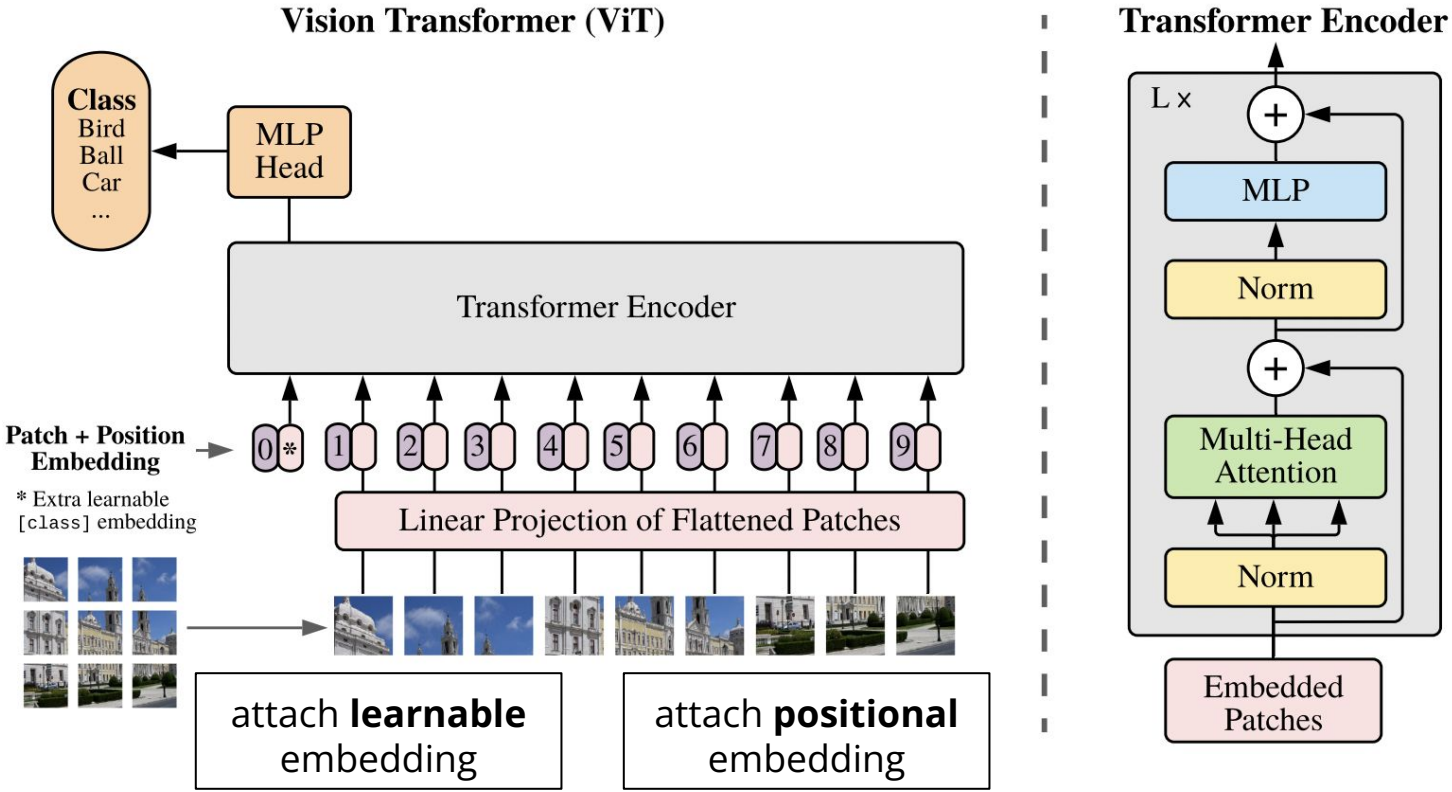




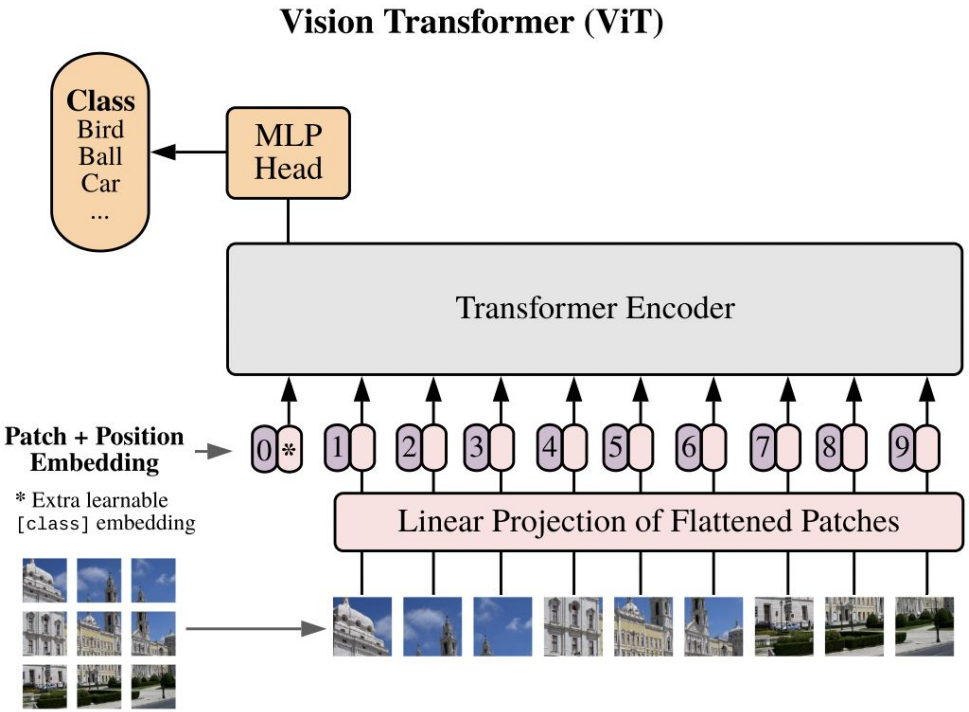
embed the patches

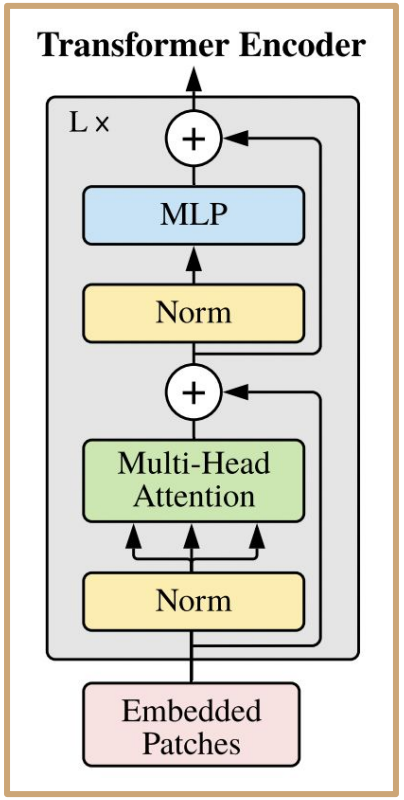
patches projected linearly onto vectors of dimensions D

Architecture



Architecture





hidden size D

Experiments - the datasets

Training datasets

- ImageNet - 1k classes, 1.3M images
- ImageNet-21k - 21k classes, 14M images
- JFT - 18k classes, 303M images

Examples from the ImageNet dataset

Experiments - the datasets

Transfer

- ImageNet
- Oxford-IIIT Pets
- Oxford Flowers-102
- VTAB classification suite with 19 downstream tasks

Examples from the Oxford Pets dataset

Experiments - model variants

- ViT

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.

- Baseline - modified version of ResNet - “ResNet (BiT)”

Experiments - the SOTAs

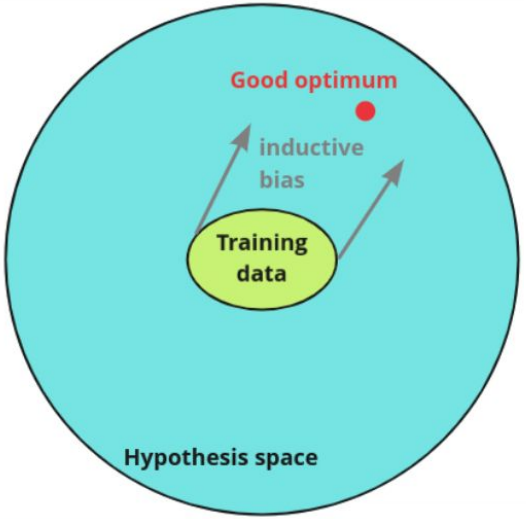
- **Big Transfer** - BiT - supervised transfer learning with large ResNets
- **Noisy Student** - semi-supervised approach trained on ImageNet and JFT-300M
- report average accuracy

Other type of Noisy Student(s)

CNN vs ViT

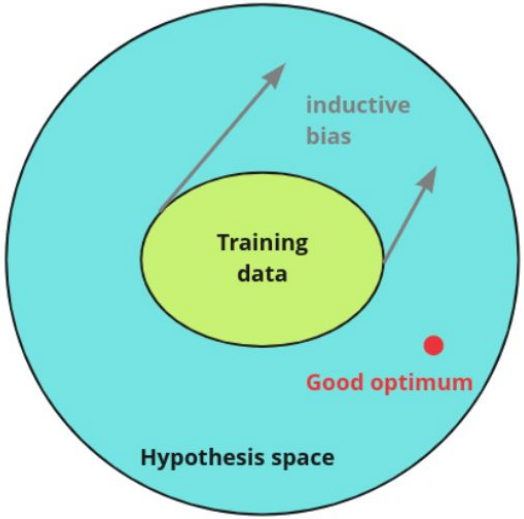
- CNNs have image-specific inductive bias

- ViT has much less image-specific inductive bias



less training data

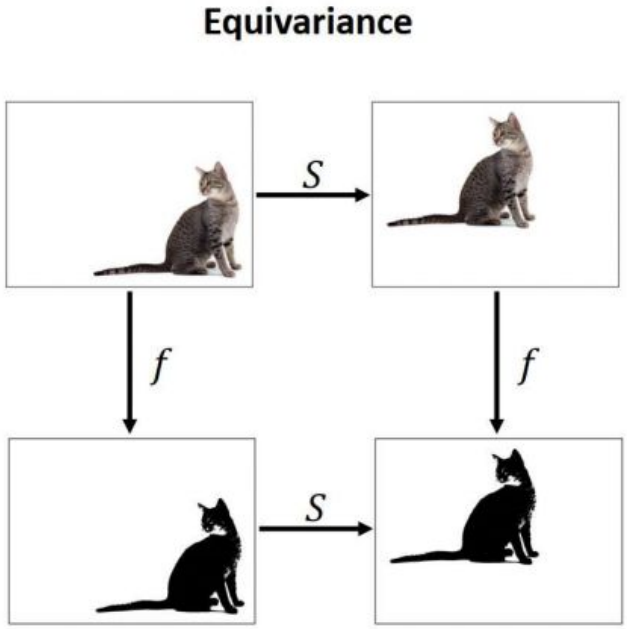
more training data



CNN vs ViT

- In CNNs locality, 2D neighbourhood structure and translation equivariance are baked into each layer

- In ViT only MLP layers are locally and translationally equivariant while self-attentions are global



Experiments - comparison to SOTA

	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

the smaller model ViT-L/16
already outperforms BiT
and is much faster!

Experiments - comparison to SOTA

	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

further improvement of the huge model

Experiments - comparison to SOTA

	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

performs well while computational time is extremely low

Experiments - comparison to SOTA

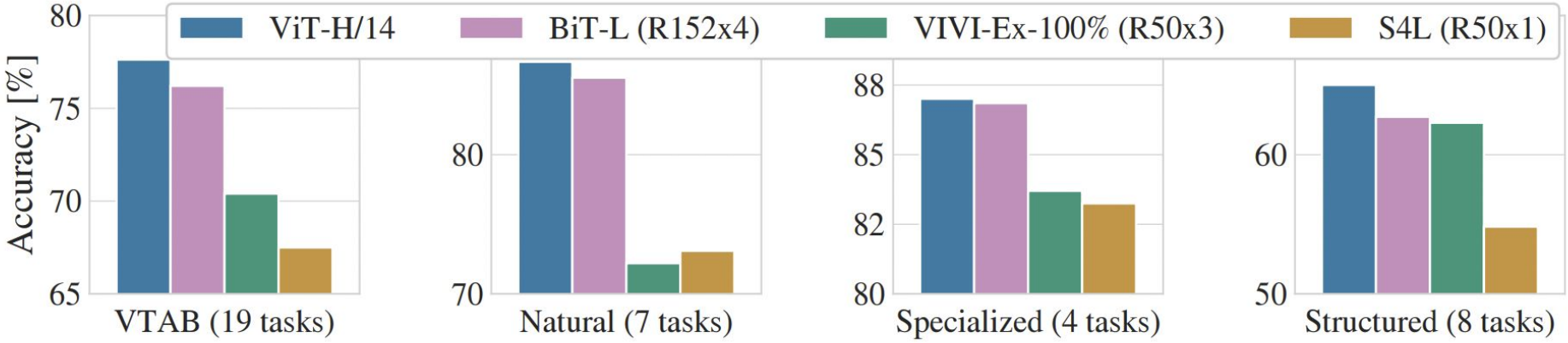
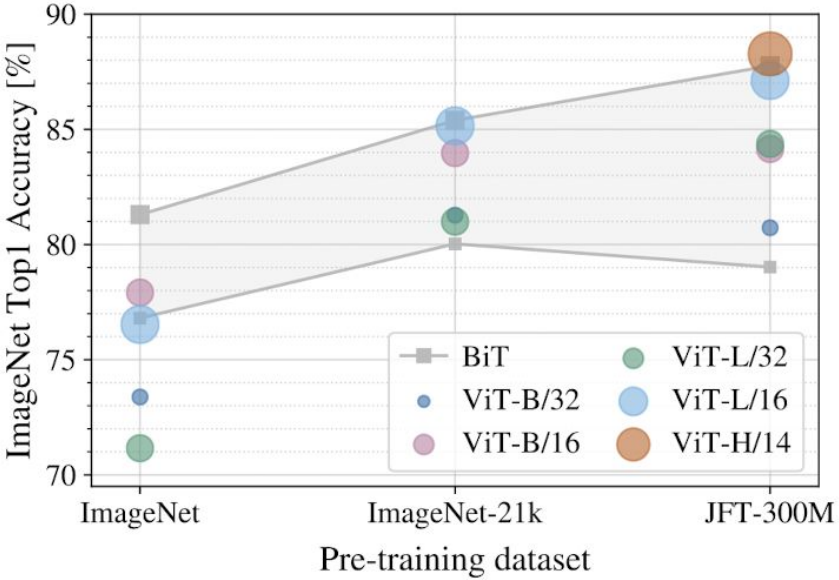


Figure 2: Breakdown of VTAB performance in *Natural*, *Specialized*, and *Structured* task groups.

Experiments - pre-training data requirements

- How important is the size of the dataset?

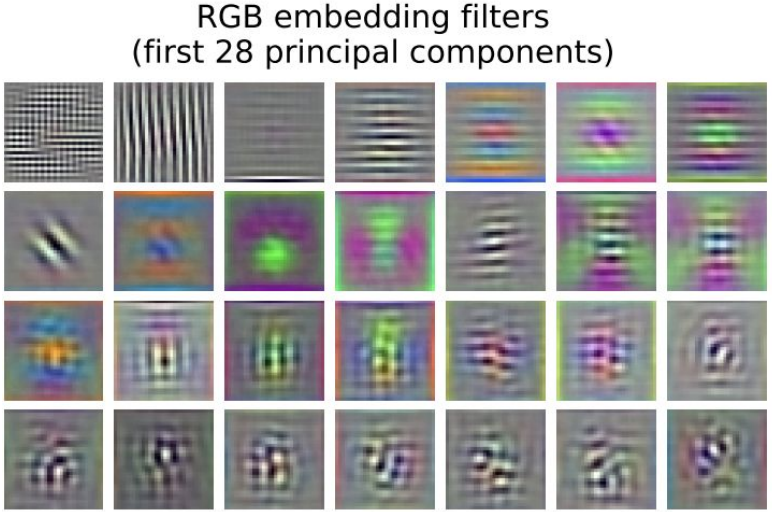


When pretrained on the smallest dataset, ImageNet, ViT-Large underperforms compared to ViT-Base.

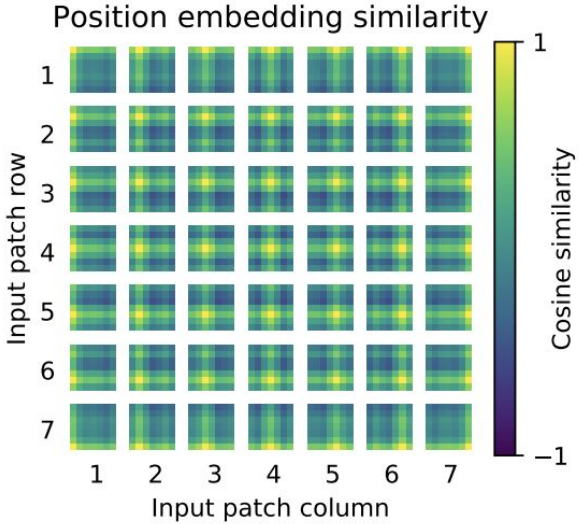
With JFT-300 we see the advantage of the large models.

To beat the baseline CNNs we need larger pret-raining sets.

Inspecting ViT



top principal components of
the learned embedding filters
→ plausible functions



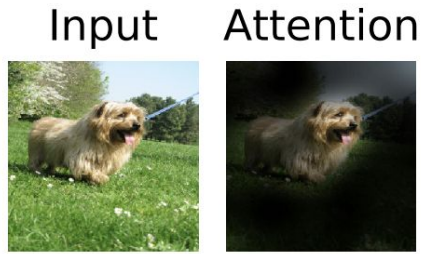
closer patches have
closer embeddings

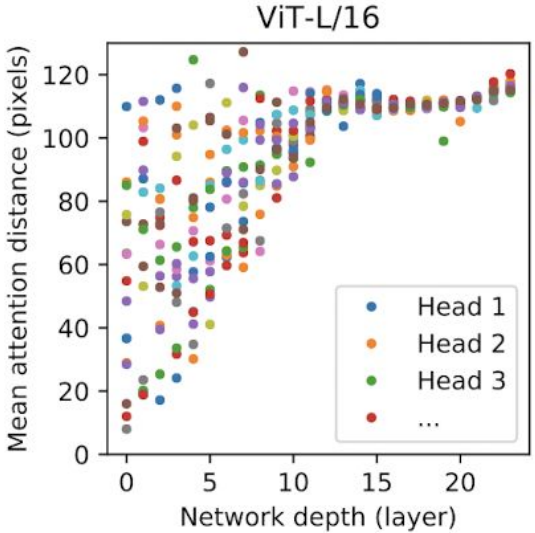
Inspecting ViT

- self-attention allows to integrate information across the entire image

→ information is integrated globally

- the model attends to image regions that are semantically relevant for classification





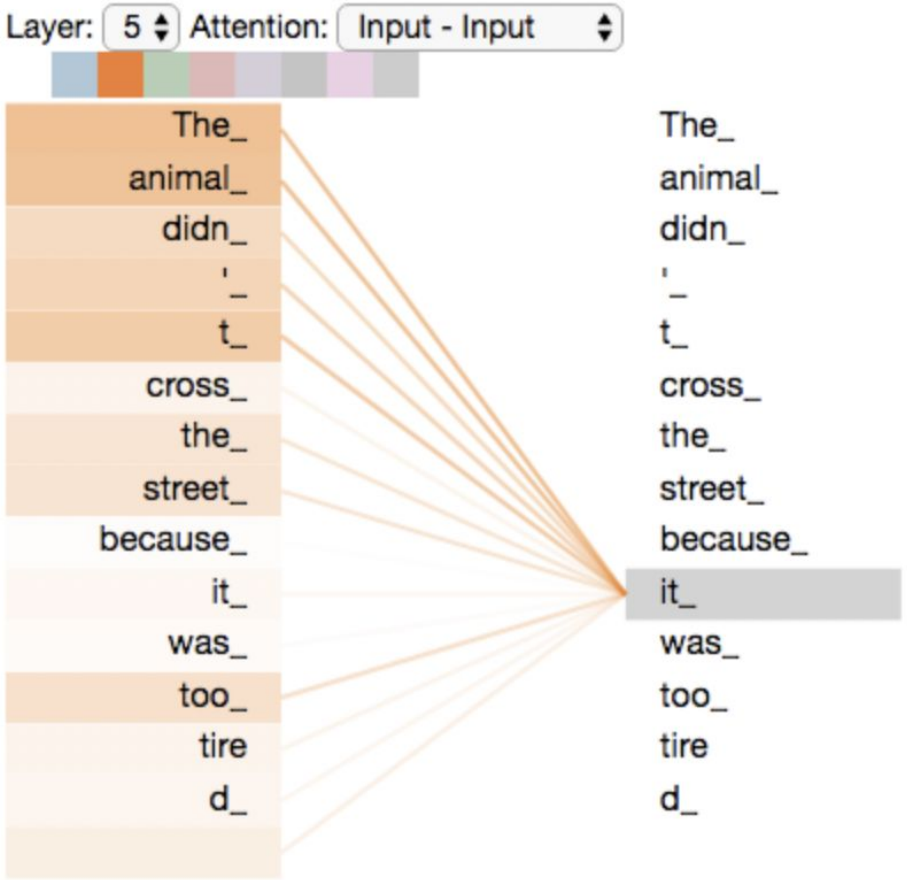
Conclusion

- ViT proves that in computer vision, the reliance on CNN is not necessary
- Excellent results compared to CNN SOTAs with fewer computational resources
- Important applications - image classification, object detection, image segmentation and many more

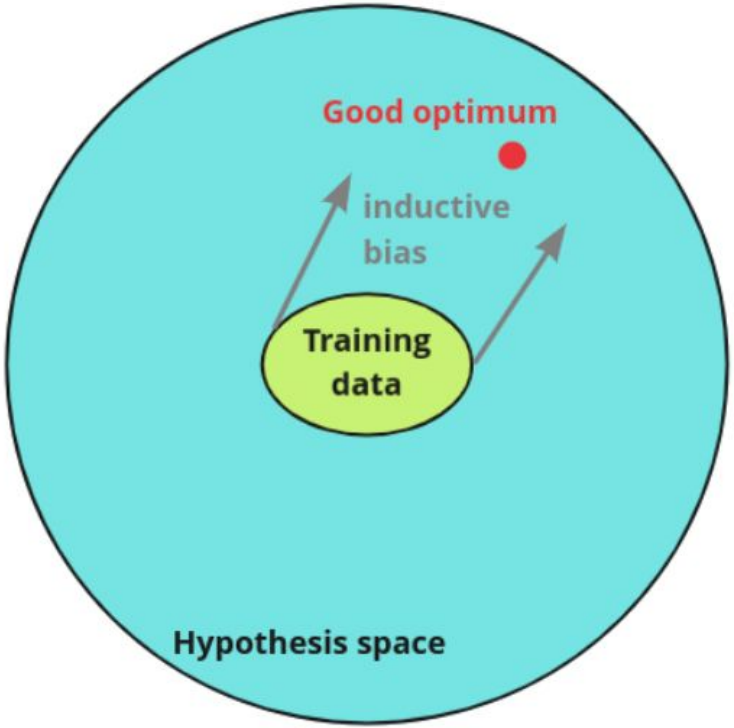
Thank you for your *attention*!

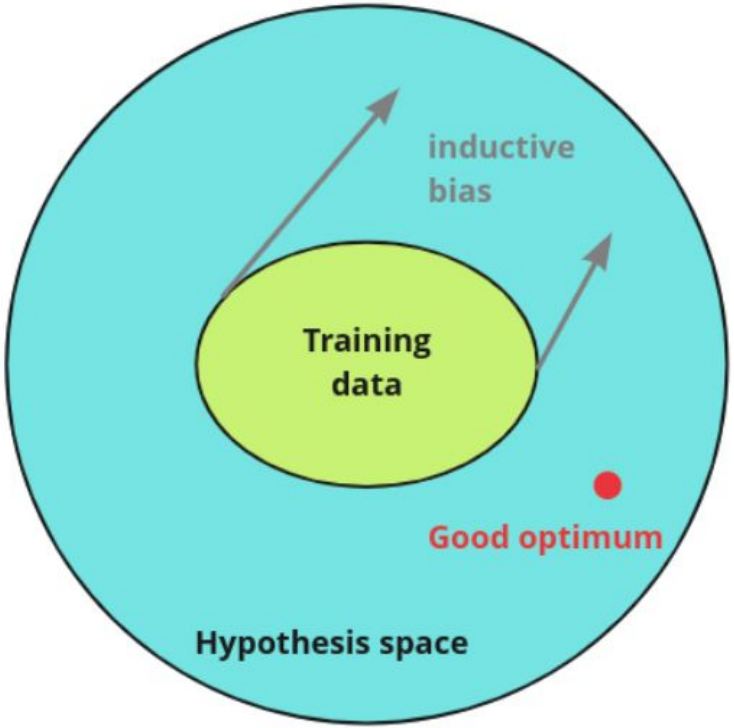
APPENDIX

What is attention?

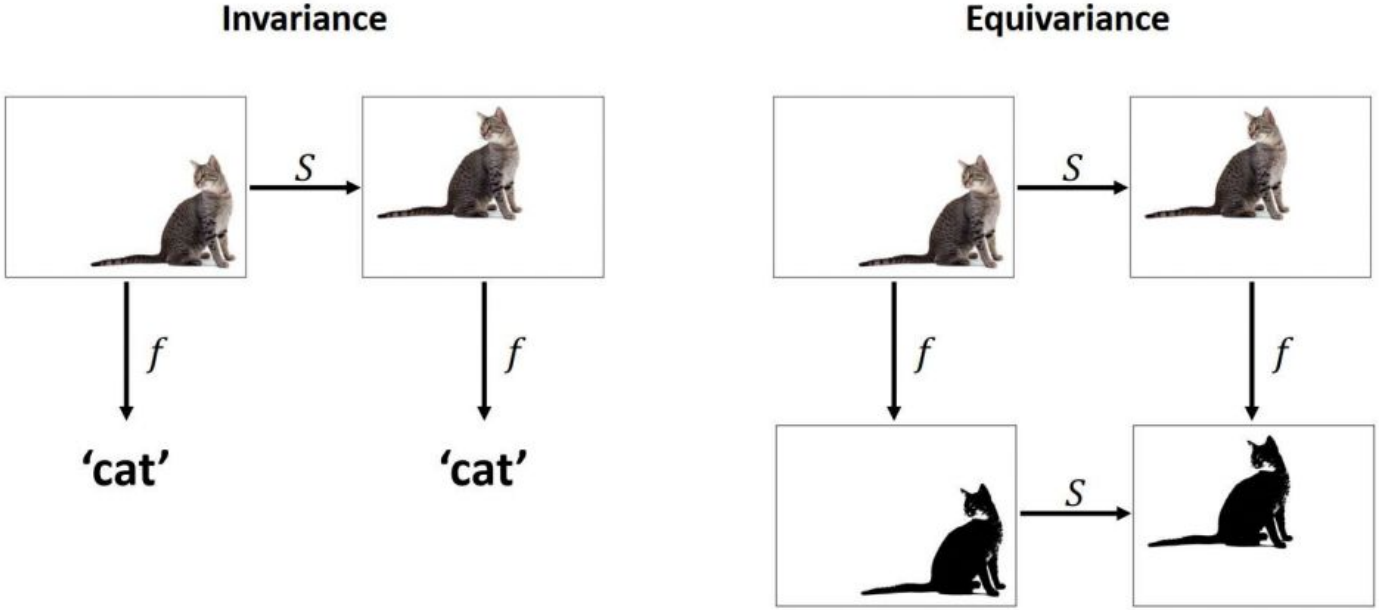


What is inductive bias?





What is translational equivariance?



ViT and baseline training details

Training

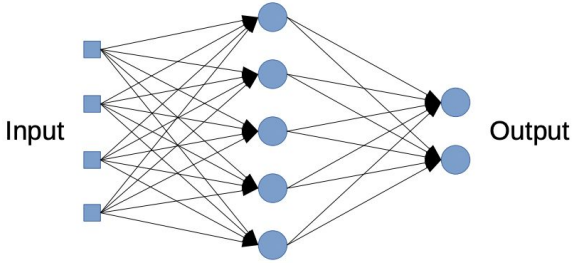
- Adam
- batchsize 4096
- weight decay 0.1

Fine-tuning

- SGD with momentum
- batchsize 512

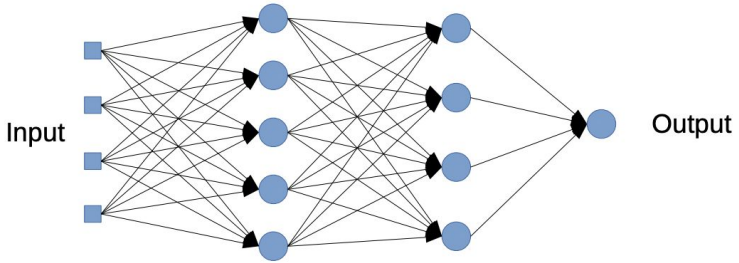
What is MLP?

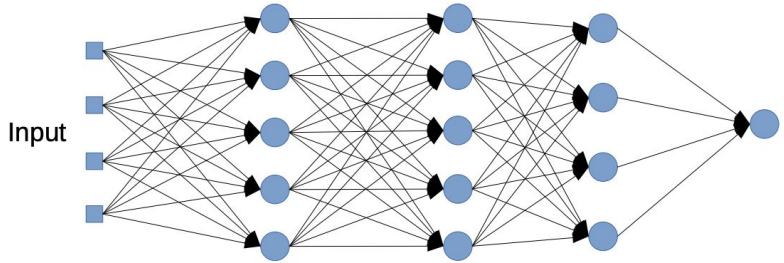
We need the Multi-layer perceptron (MLP)



One hidden layer

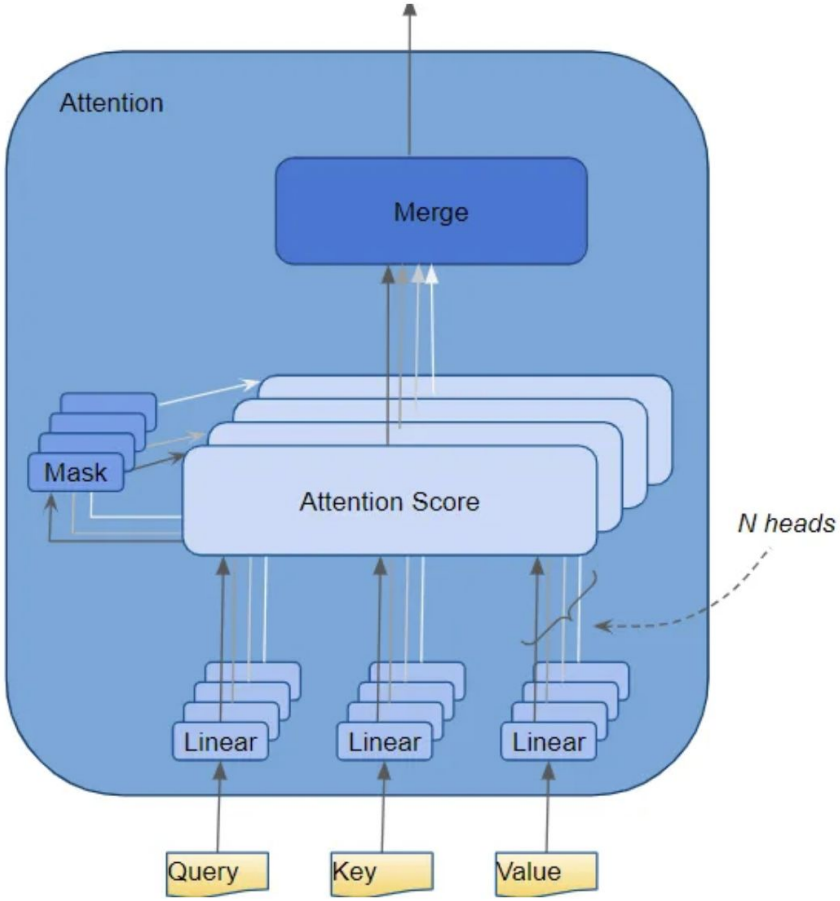
Two hidden layers





Three hidden layers (towards deep MLP)

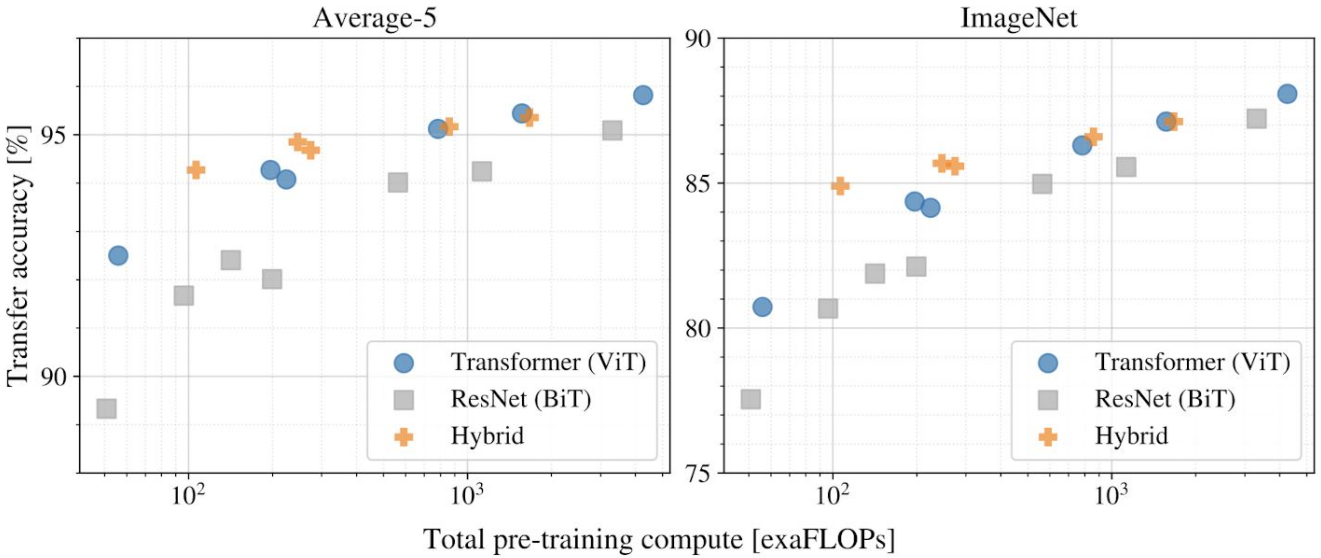
What is Multi-Head Self-Attention?



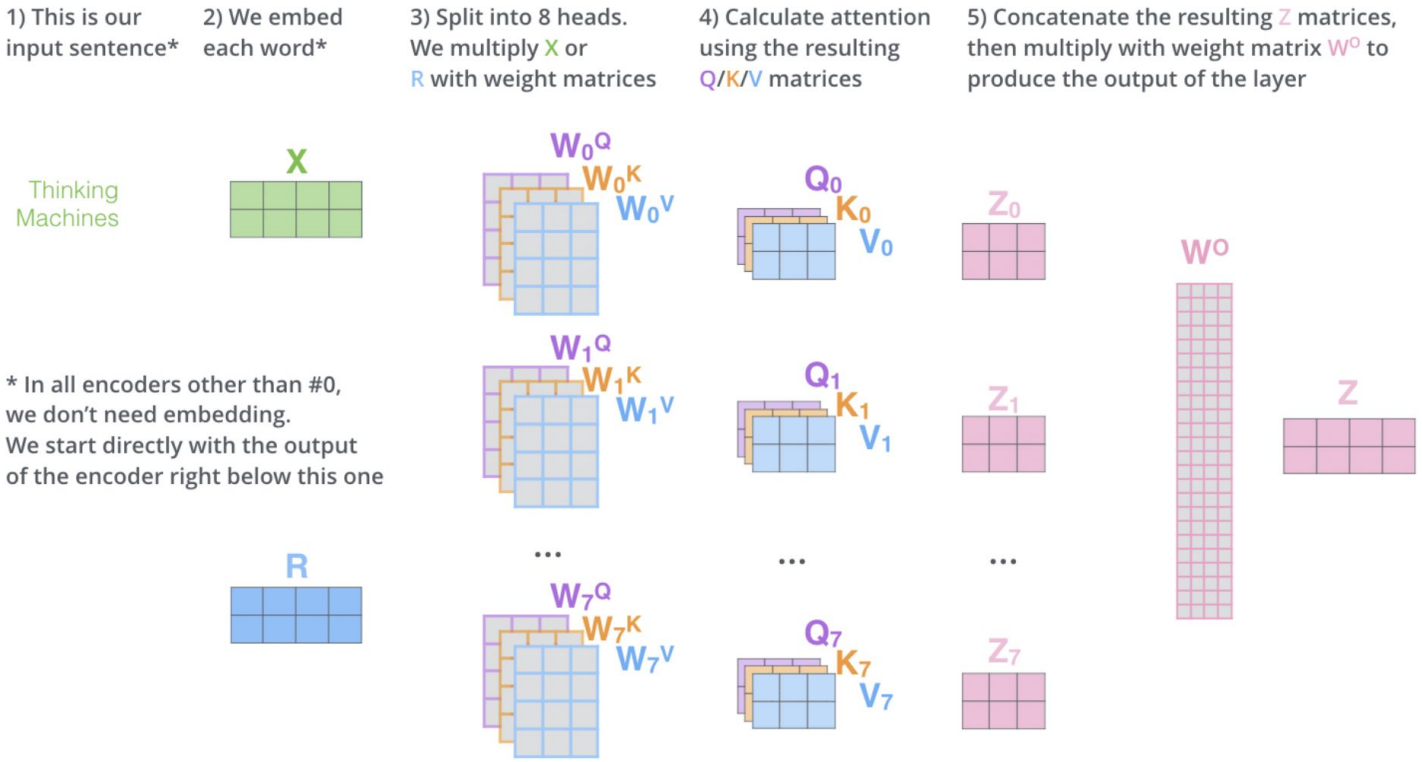
(Image by Author)

ViT Hybrid architecture

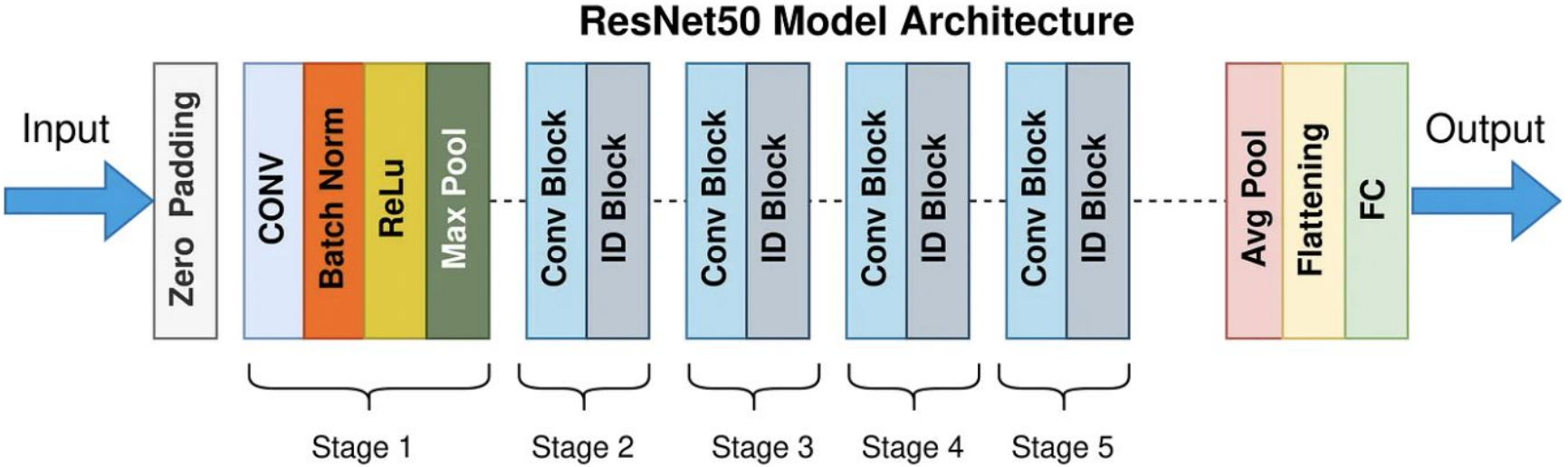
- the input sequence formed from feature maps of a CNN
- the patch embedding is applied to patches extracted from a CNN feature map



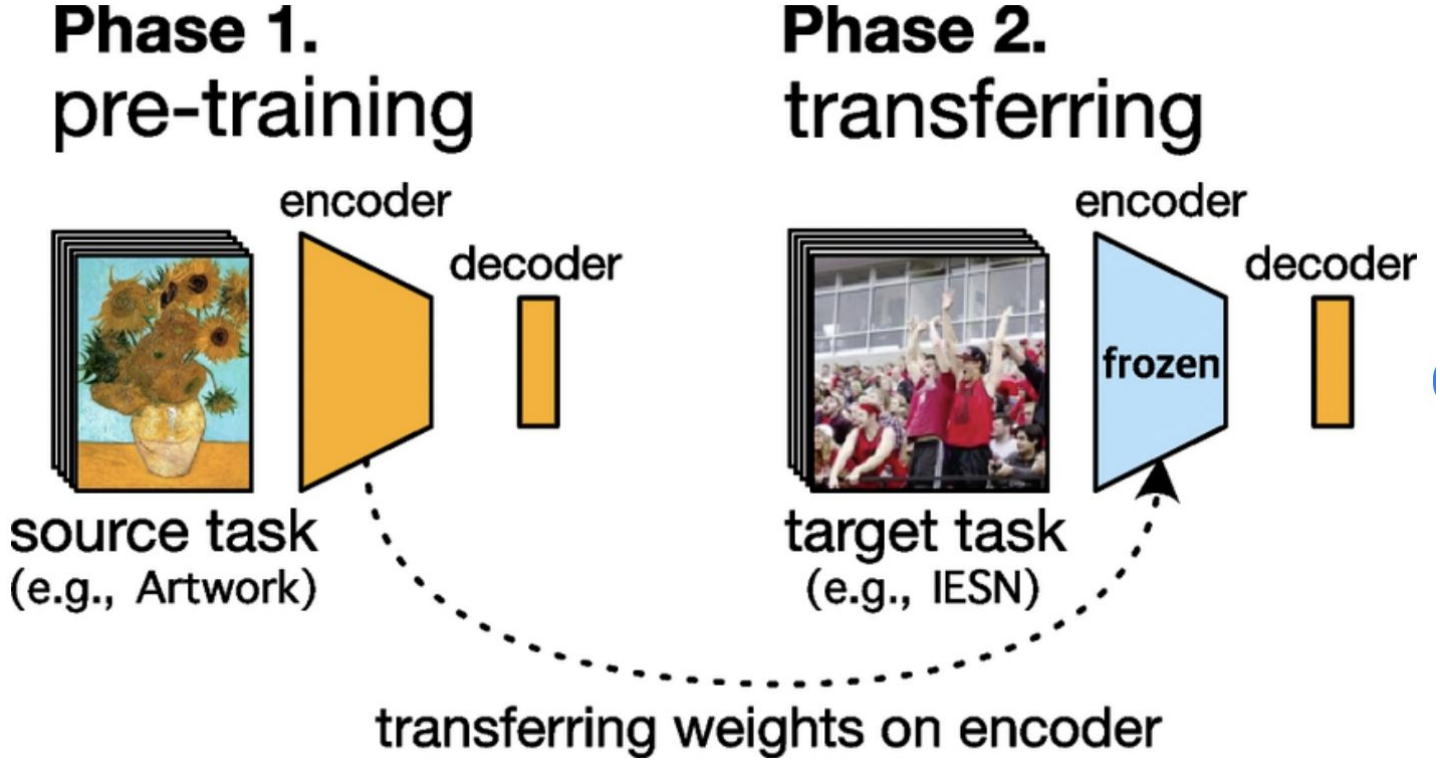
What is multi-headed attention?



What is the ResNet Architecture?



What is Transfer Learning?



CNN vs ViT

- CNNs have image-specific inductive bias
- locality, 2D neighbourhood structure and translation equivariance are baked into each layer

- ViT has much less image-specific inductive bias
- only MLP layers are locally and translationally equivariant while self-attentions are global
- position embeddings carry no information about the 2D positions of the patches → spatial informations have to be learned by scratch