An Image is Worth 16x16 Words:

Transformers for Image Recognition at Scale

Binder Jonas

Overview

- Introduction
- Vision Transformer
- Experiments
- State Today
- Conclusion
- Questions

Transformers

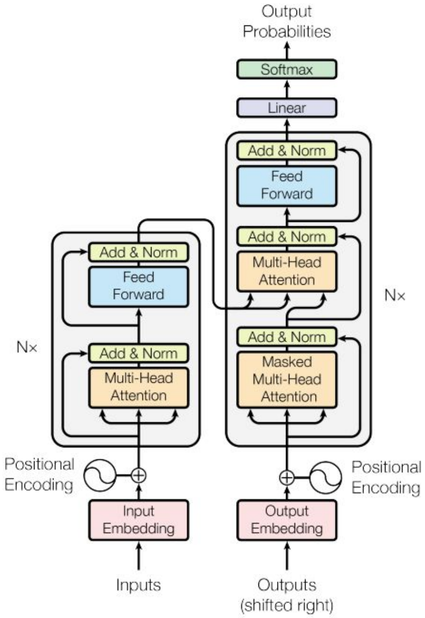


Figure 1: The Transformer - model architecture.

Source: Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. Attention is all you need. In NIPS 2017

Transformers

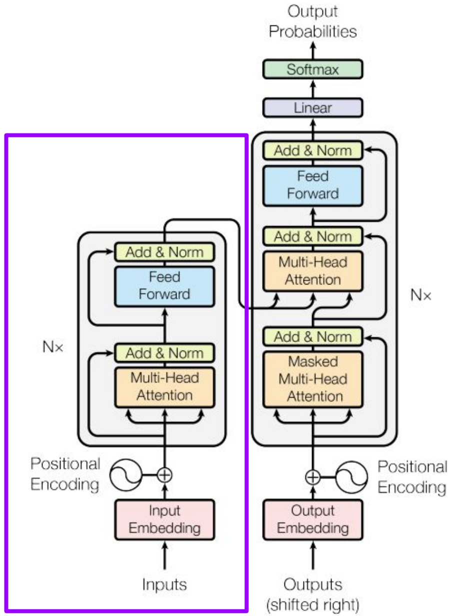
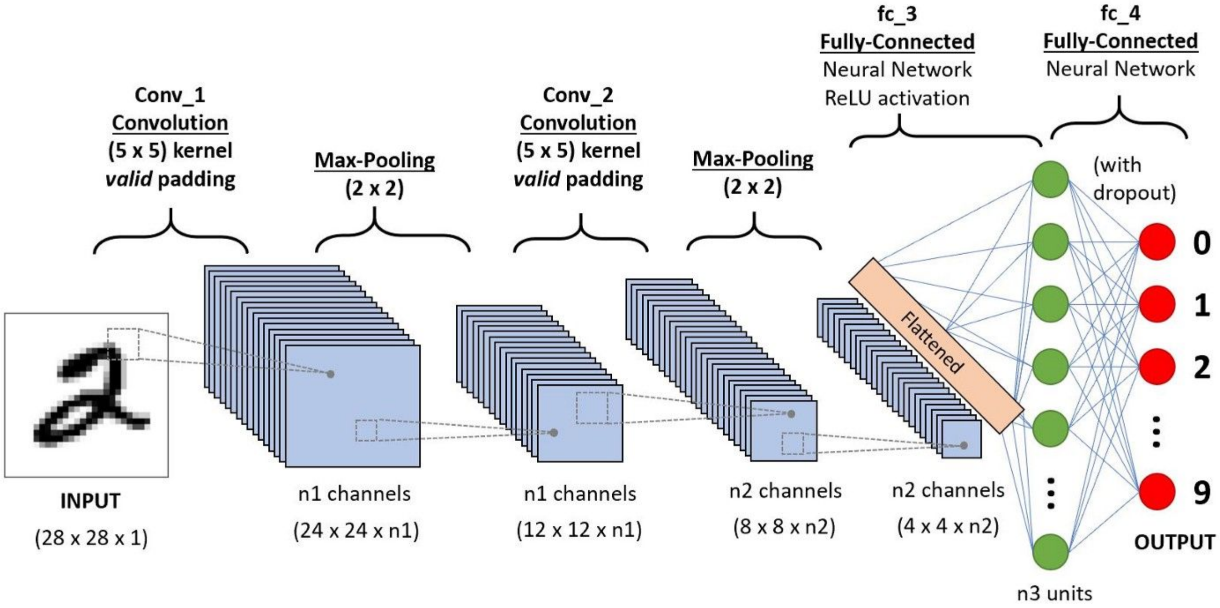


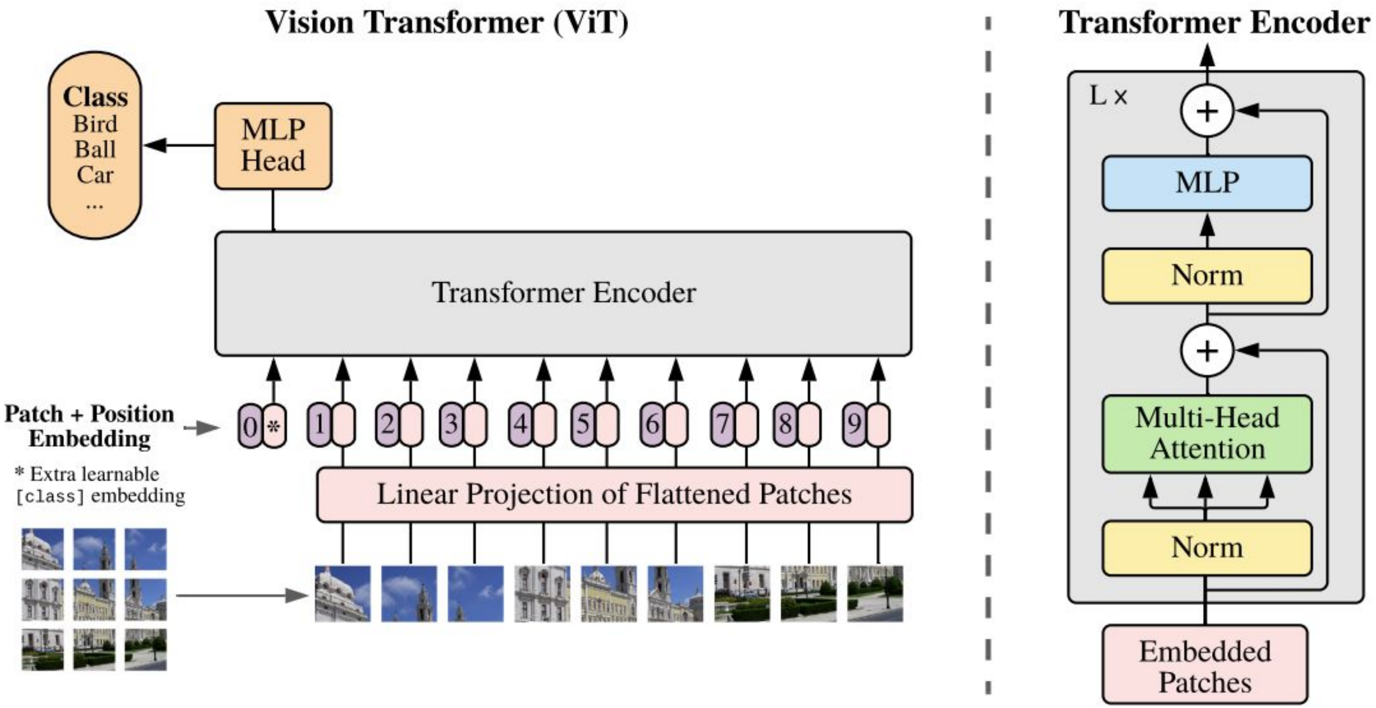
Figure 1: The Transformer - model architecture.

Source: Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. Attention is all you need. In NIPS 2017

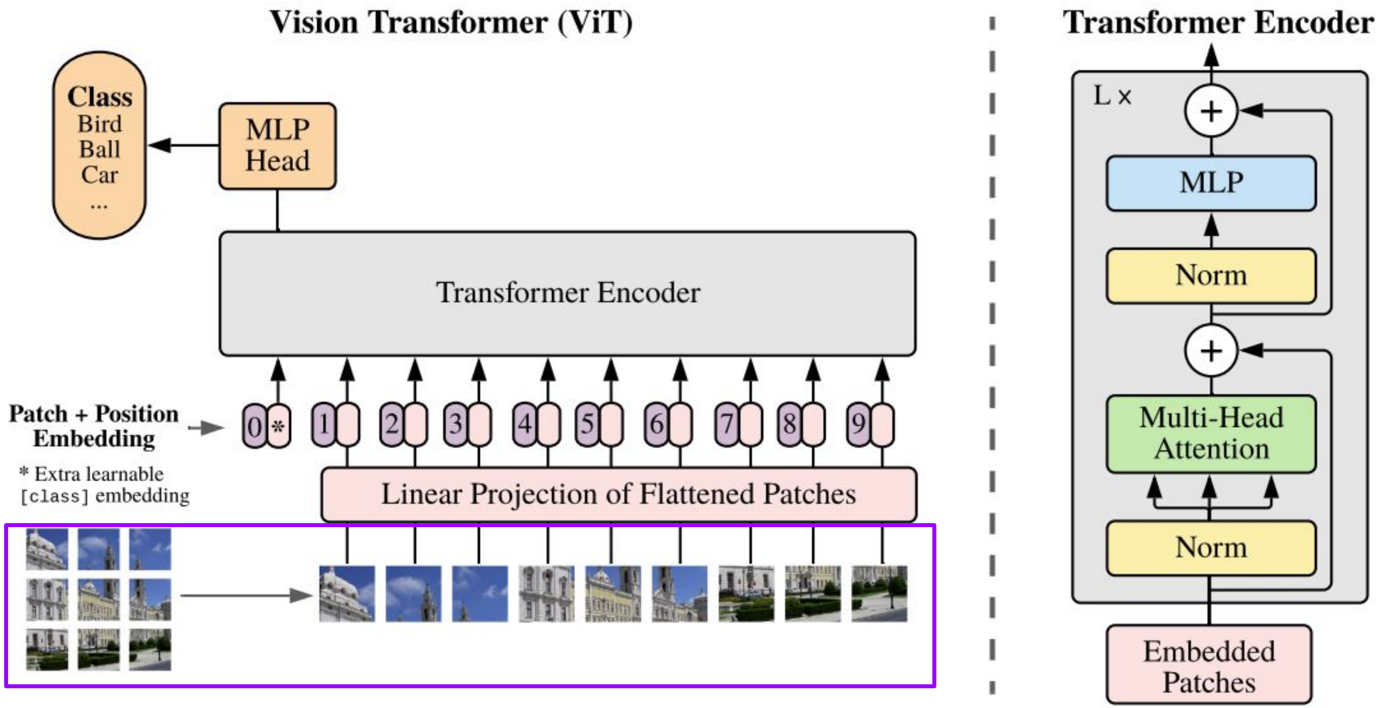
Convolutional Neural Networks (CNN)



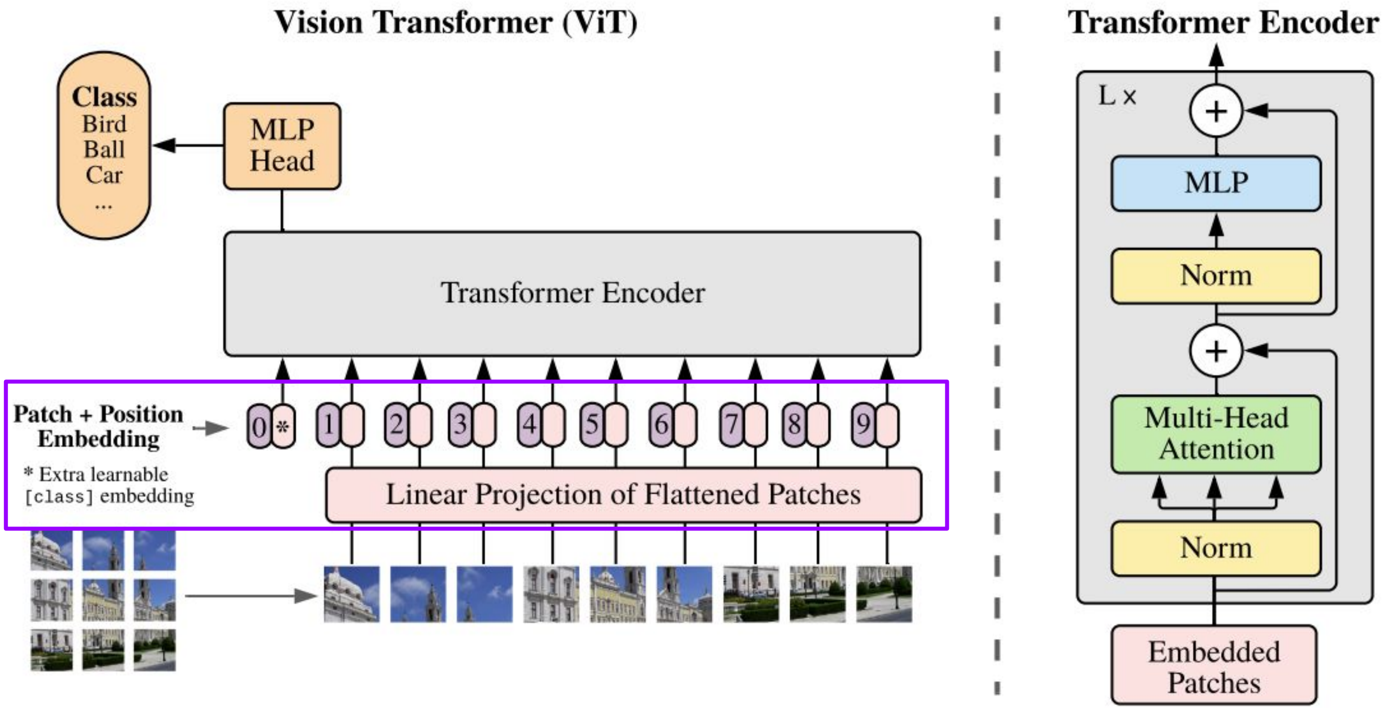
Vision Transformer



Vision Transformer



Vision Transformer



Patch Embeddings

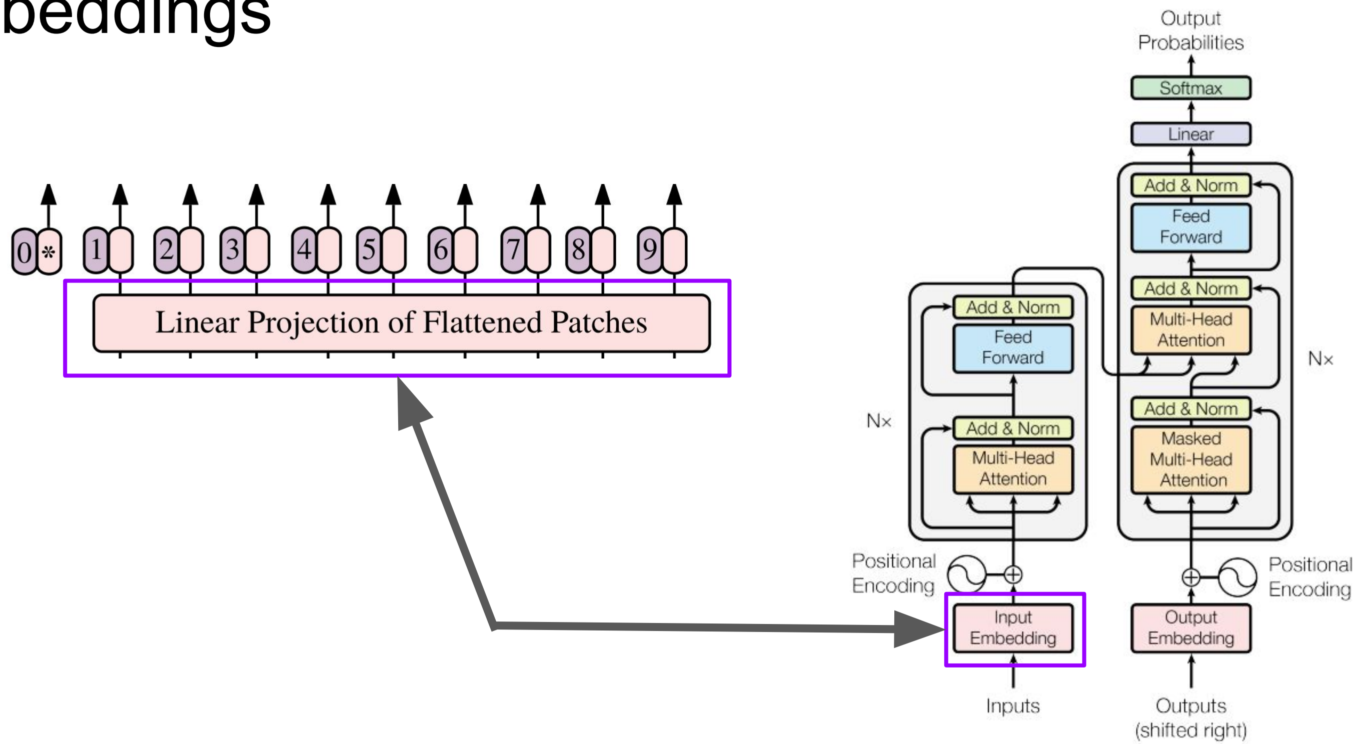
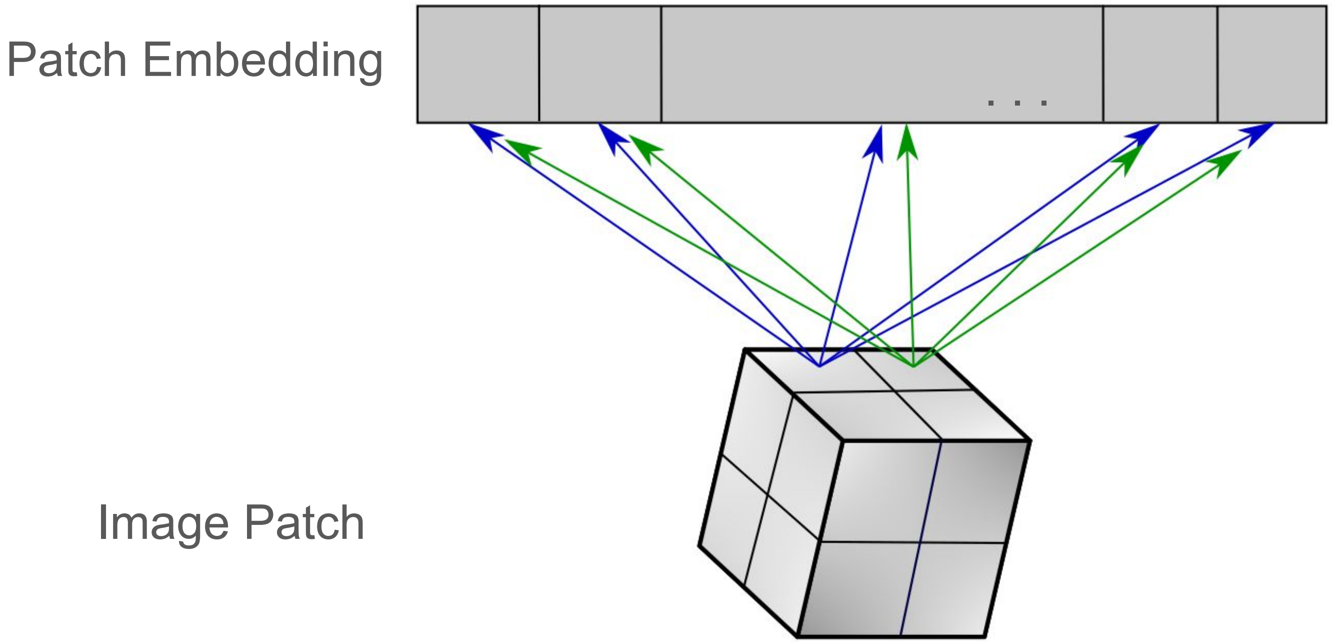
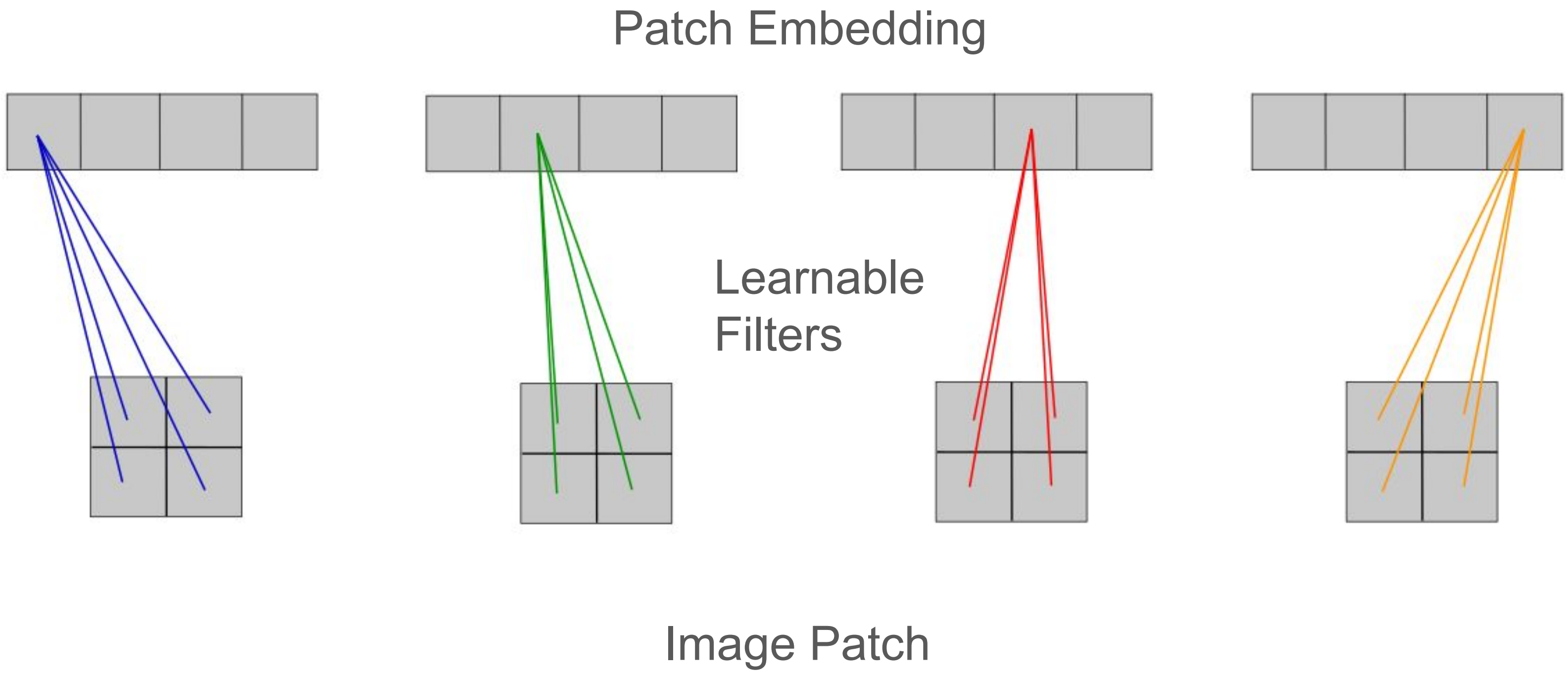


Figure 1: The Transformer - model architecture.

Linear Projection



Analogy to Convolution



Patch Embeddings

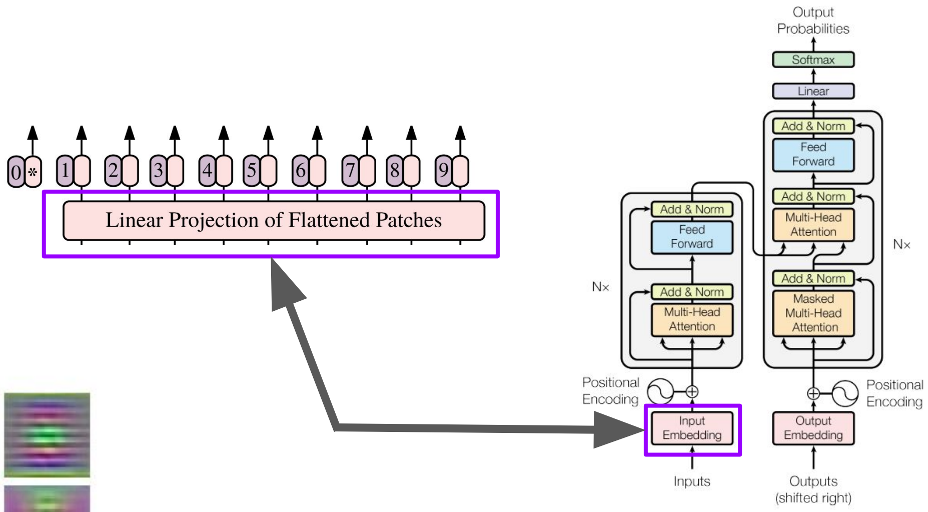
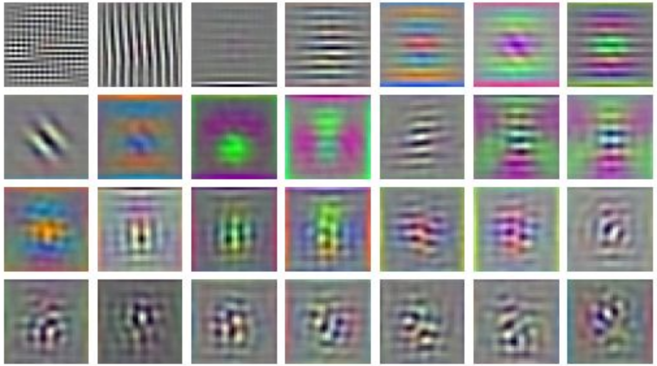
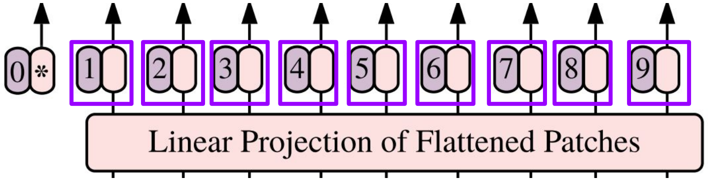


Figure 1: The Transformer - model architecture.

RGB embedding filters
(first 28 principal components)



Position Embeddings



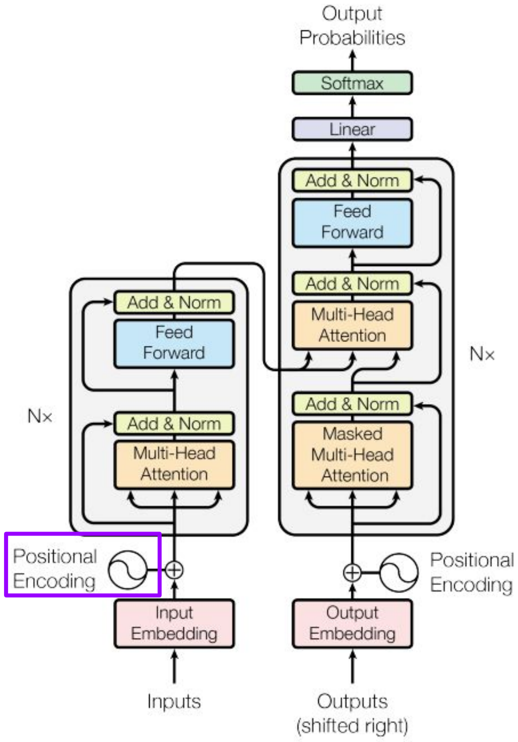
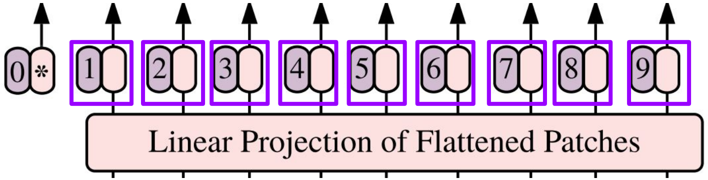


Figure 1: The Transformer - model architecture.

Position Embeddings



$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

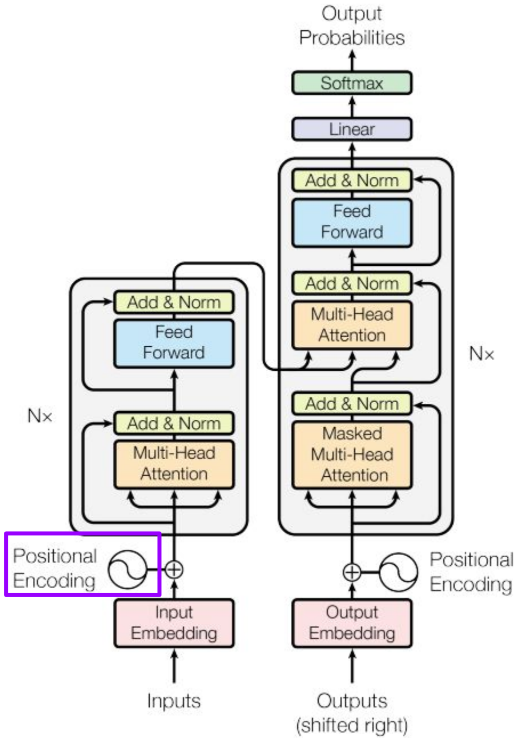
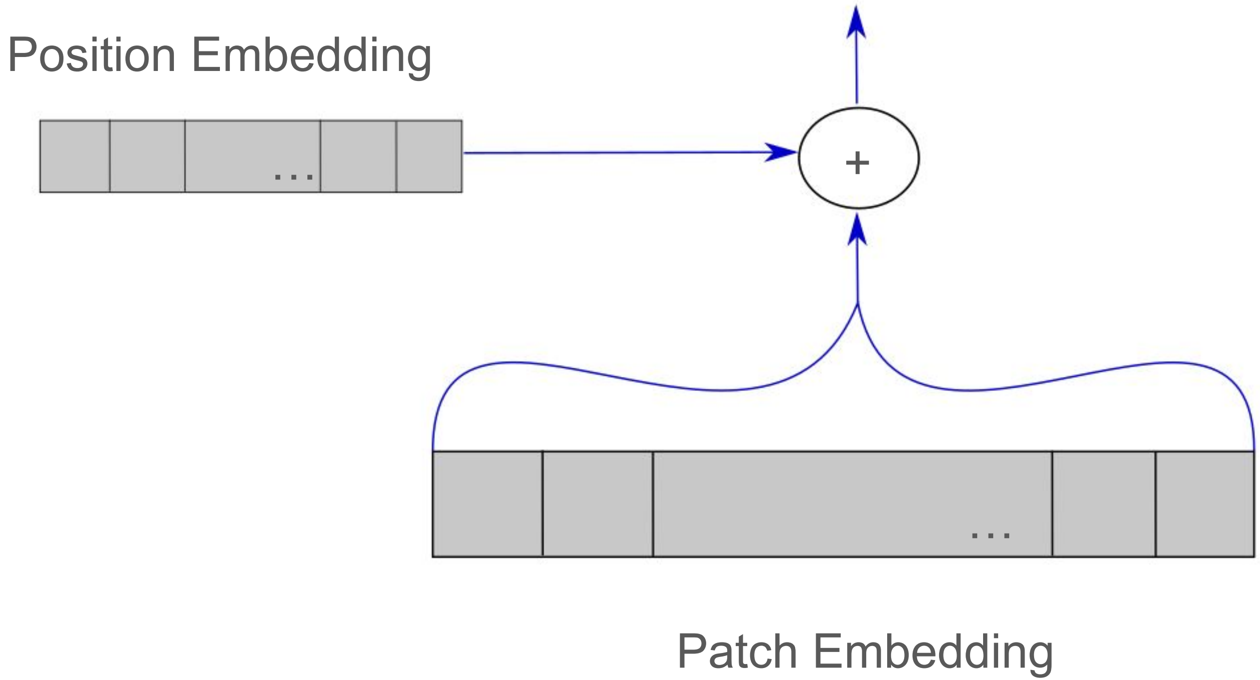
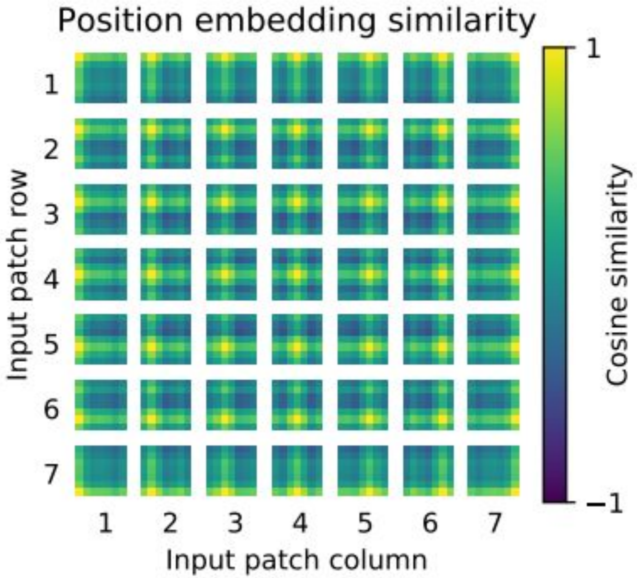


Figure 1: The Transformer - model architecture.

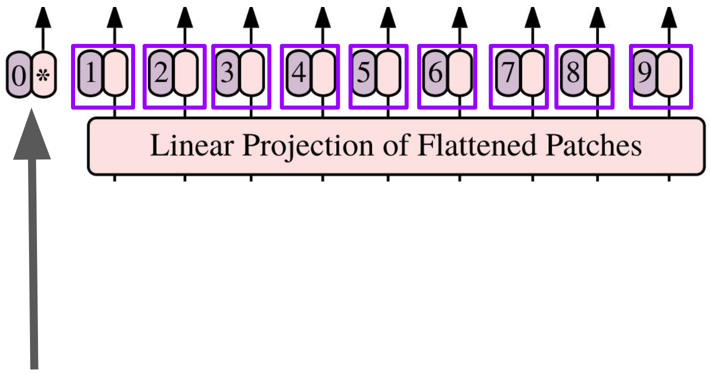
Position Embedding



Position Embeddings



Position Embeddings



Class Token

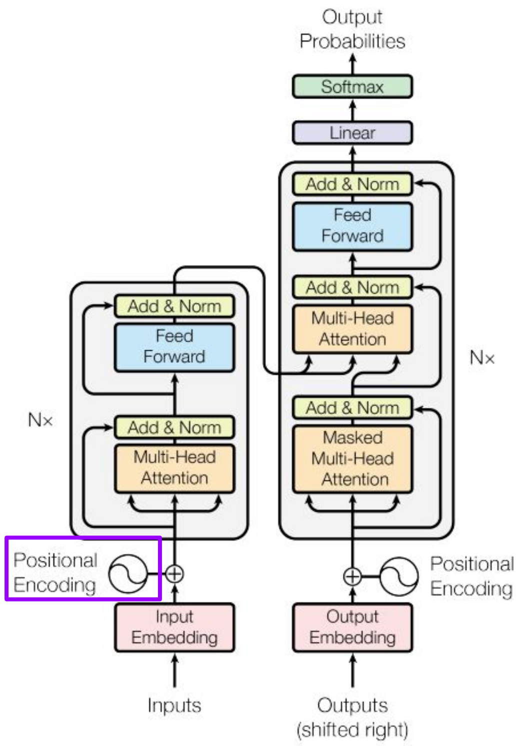
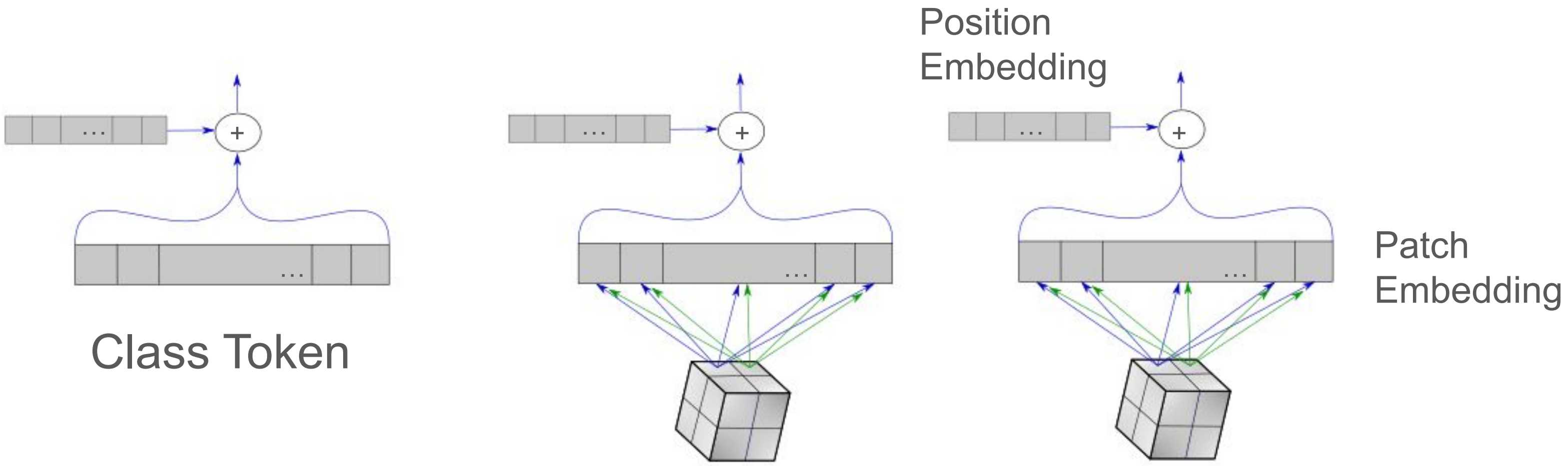


Figure 1: The Transformer - model architecture.

Whole Pipeline



Position Embeddings

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}};\, \mathbf{x}_p^1\mathbf{E};\, \mathbf{x}_p^2\mathbf{E};\, \cdots;\, \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}$$

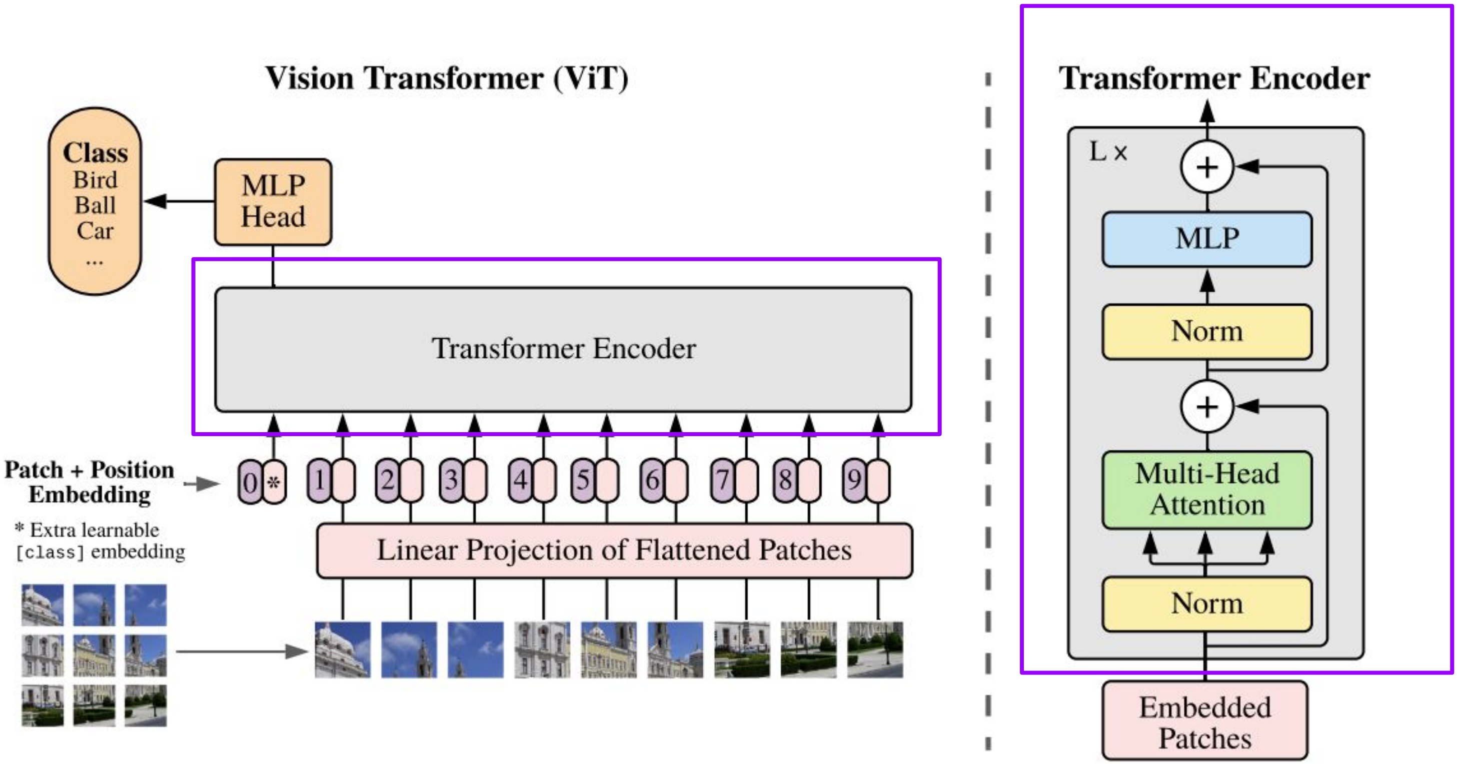
Class Token

Image Patch

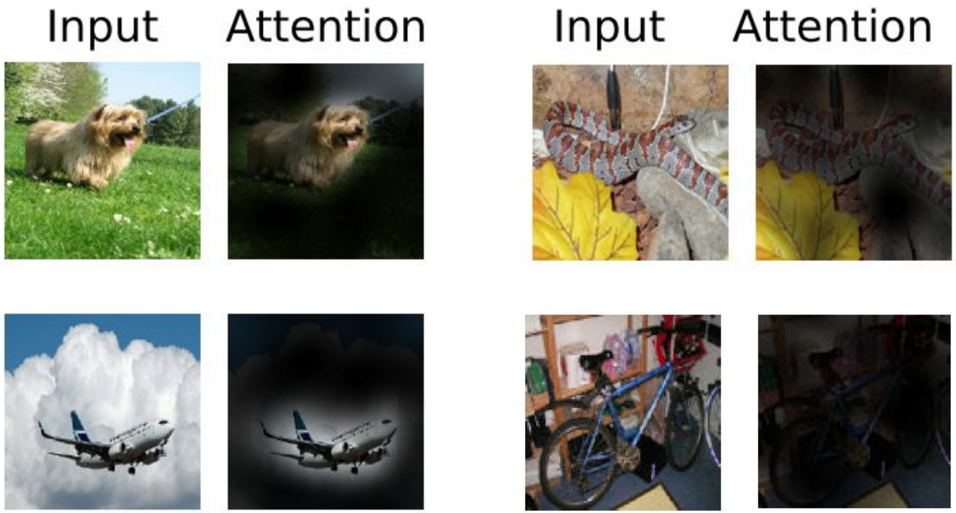
Embedding Filter

Positional Embedding

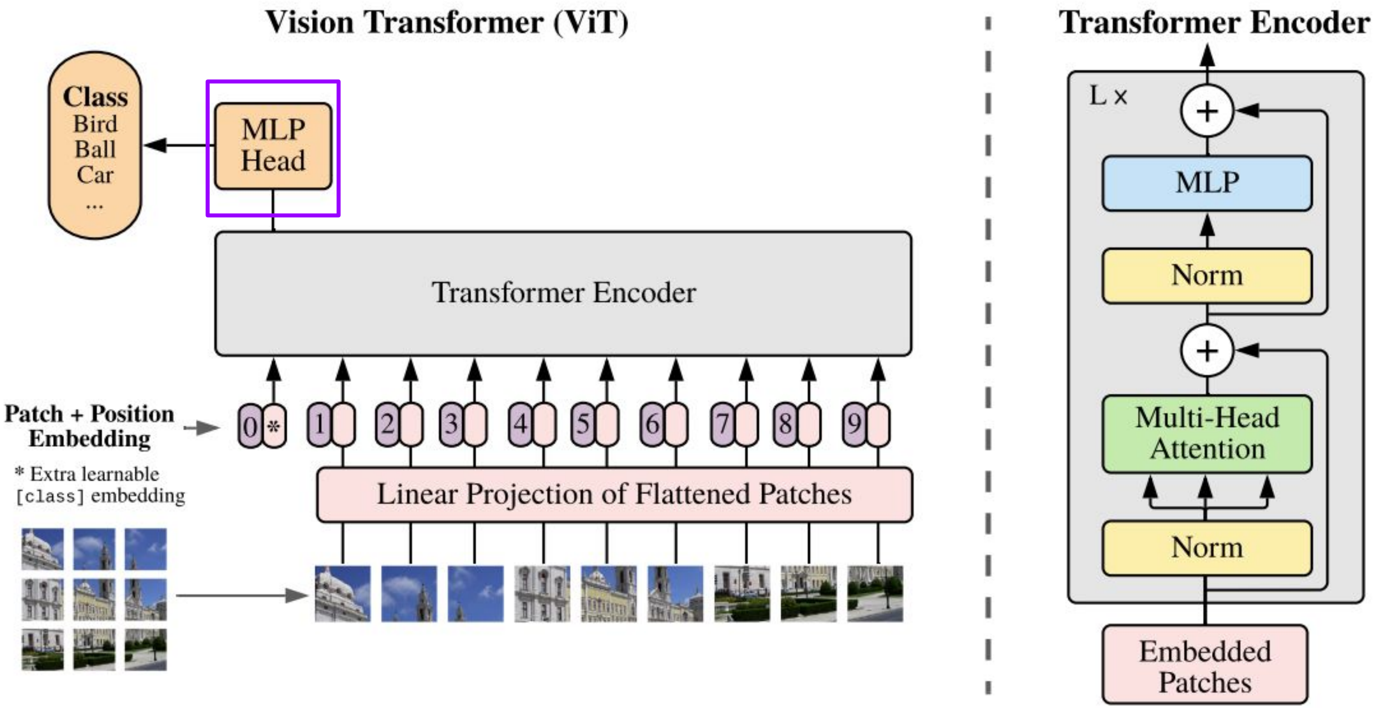
Transformer Encoder



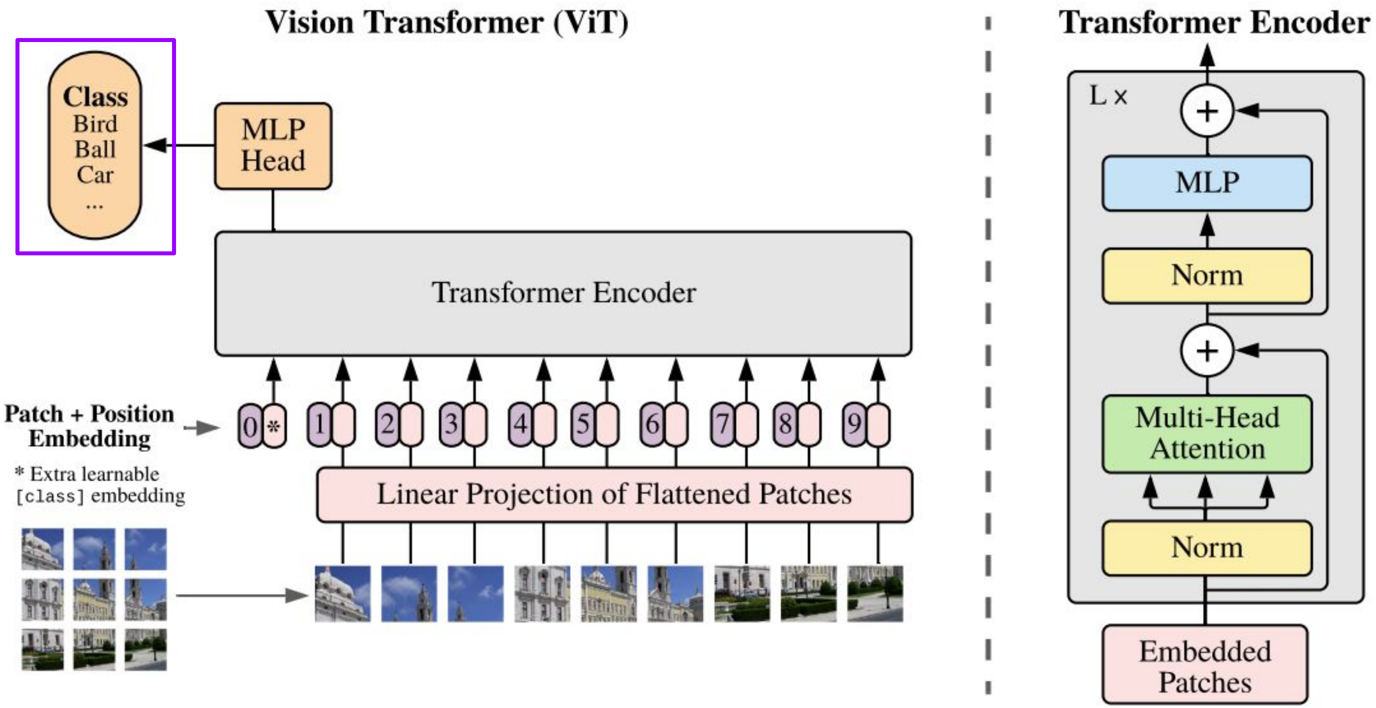
Visualization Attention



Multilayer Perceptron



Prediction



Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP
Head
Transformer Encoder
Patch + Position
Embedding
* Extra learnable
[class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches
Transformer Encoder
L ×
MLP
Norm
Multi-Head
Attention
Norm
Embedded
Patches

Models

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

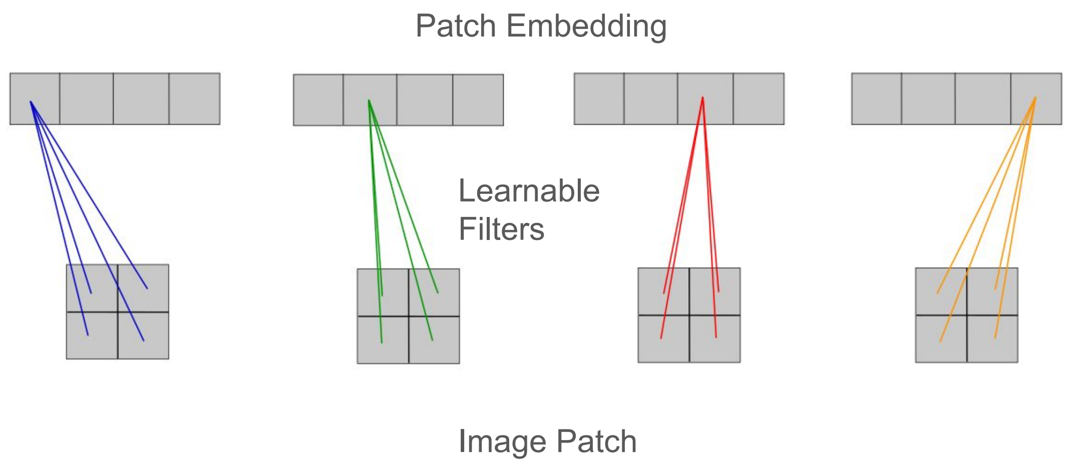
Models

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	$\boxed{768}$	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Hidden Size Intuition

Patch Size: 16 x 16 x 3 ⟶ $16 * 16 * 3 = \boxed{768}$

⟶ 1 Filter per pixel



Models

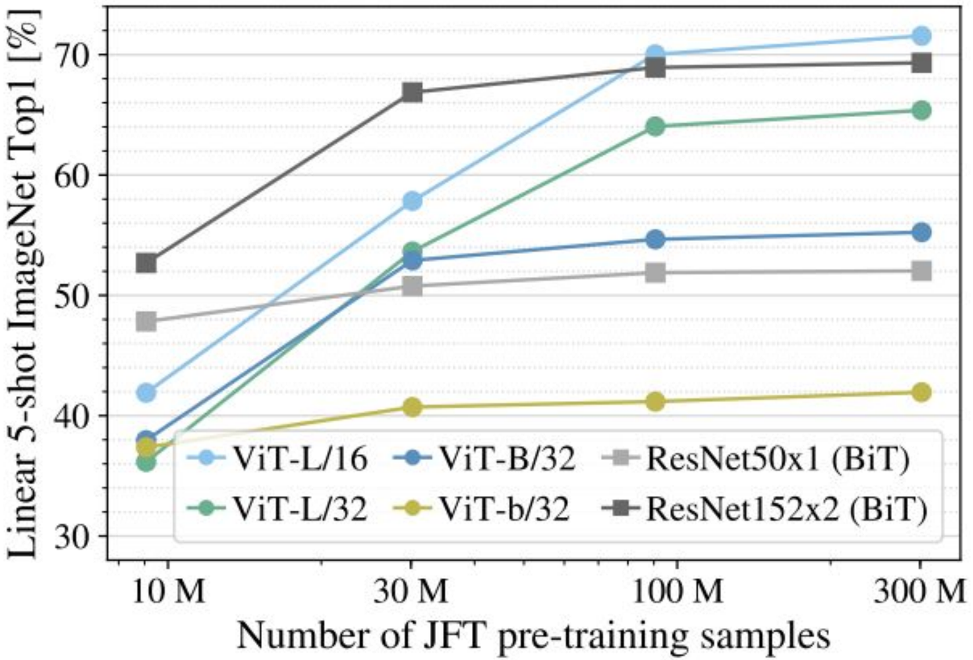
Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Results

Patch Size

	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** $\pm$ 0.04	87.76 $\pm$ 0.03	85.30 $\pm$ 0.02	87.54 $\pm$ 0.02	88.4/88.5*
ImageNet ReaL	**90.72** $\pm$ 0.05	90.54 $\pm$ 0.03	88.62 $\pm$ 0.05	90.54	90.55
CIFAR-10	**99.50** $\pm$ 0.06	99.42 $\pm$ 0.03	99.15 $\pm$ 0.03	99.37 $\pm$ 0.06	—
CIFAR-100	**94.55** $\pm$ 0.04	93.90 $\pm$ 0.05	93.25 $\pm$ 0.05	93.51 $\pm$ 0.08	—
Oxford-IIIT Pets	**97.56** $\pm$ 0.03	97.32 $\pm$ 0.11	94.67 $\pm$ 0.15	96.62 $\pm$ 0.23	—
Oxford Flowers-102	99.68 $\pm$ 0.02	**99.74** $\pm$ 0.00	99.61 $\pm$ 0.02	99.63 $\pm$ 0.03	—
VTAB (19 tasks)	**77.63** $\pm$ 0.23	76.28 $\pm$ 0.46	72.72 $\pm$ 0.21	76.29 $\pm$ 1.70	—
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

Pre-training vs. Performance

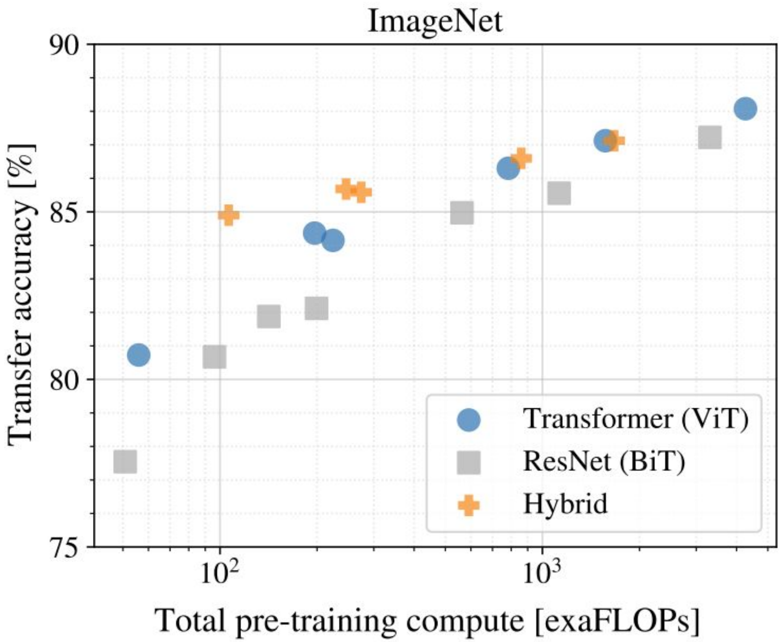


Results

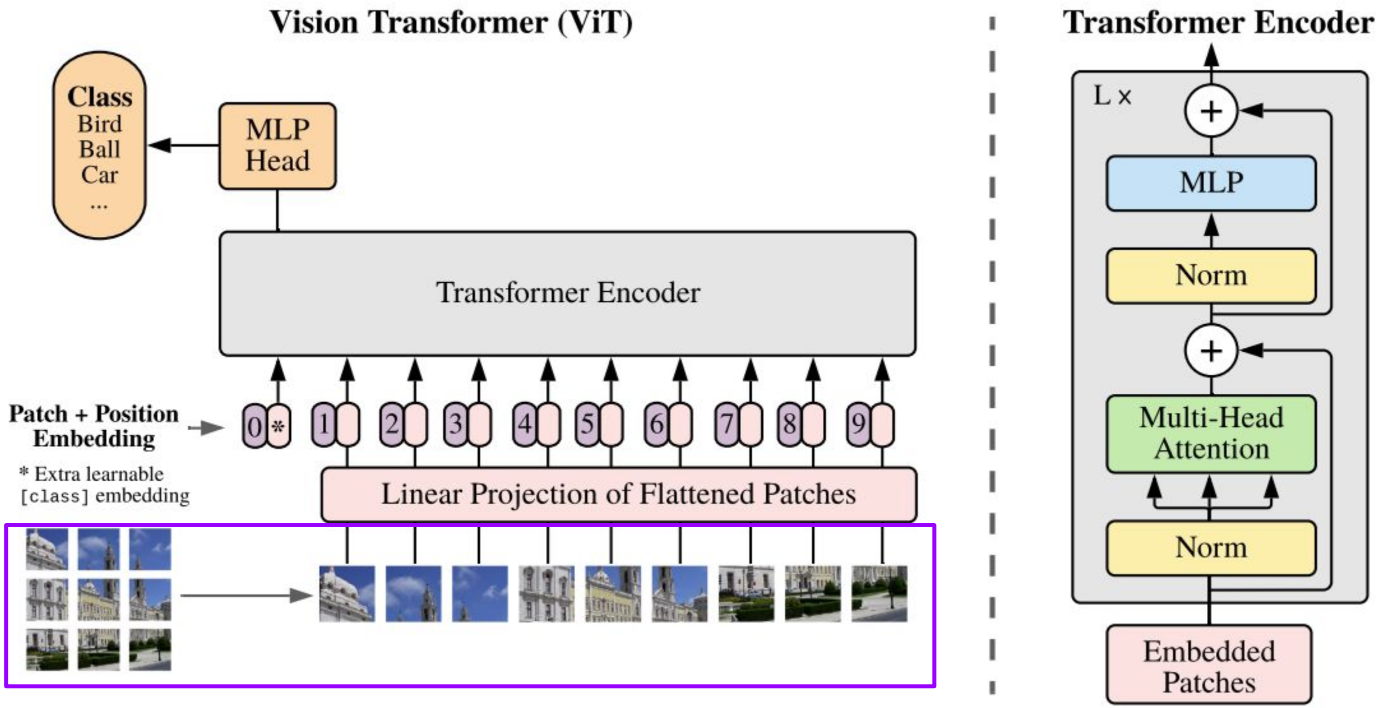
Patch Size

	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** $\pm$ 0.04	87.76 $\pm$ 0.03	85.30 $\pm$ 0.02	87.54 $\pm$ 0.02	88.4/88.5*
ImageNet ReaL	**90.72** $\pm$ 0.05	90.54 $\pm$ 0.03	88.62 $\pm$ 0.05	90.54	90.55
CIFAR-10	**99.50** $\pm$ 0.06	99.42 $\pm$ 0.03	99.15 $\pm$ 0.03	99.37 $\pm$ 0.06	–
CIFAR-100	**94.55** $\pm$ 0.04	93.90 $\pm$ 0.05	93.25 $\pm$ 0.05	93.51 $\pm$ 0.08	–
Oxford-IIIT Pets	**97.56** $\pm$ 0.03	97.32 $\pm$ 0.11	94.67 $\pm$ 0.15	96.62 $\pm$ 0.23	–
Oxford Flowers-102	99.68 $\pm$ 0.02	**99.74** $\pm$ 0.00	99.61 $\pm$ 0.02	99.63 $\pm$ 0.03	–
VTAB (19 tasks)	**77.63** $\pm$ 0.23	76.28 $\pm$ 0.46	72.72 $\pm$ 0.21	76.29 $\pm$ 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

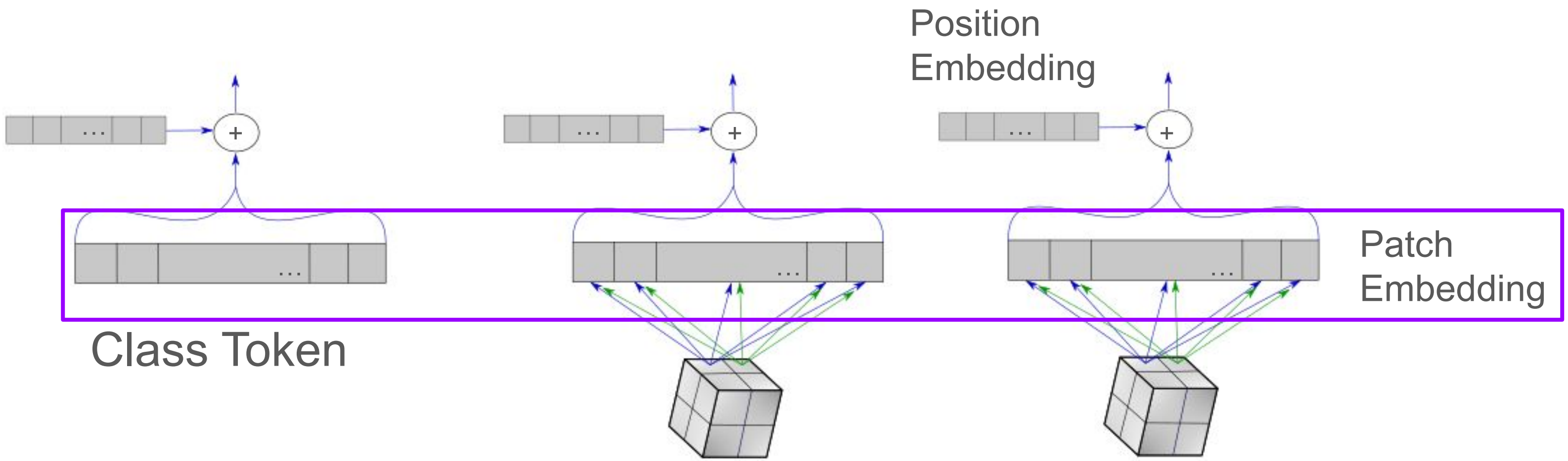
Pre-Training vs. Accuracy



Vision Transformer



Self Supervision



Results

- Only tested on smallest model (ViT-B/16)
- Faster convergence
- Worse performance than pretraining:
  - Training from scratch: 77.9%
  - With self-supervision:79.9%
  - With pre-training: 84.15%

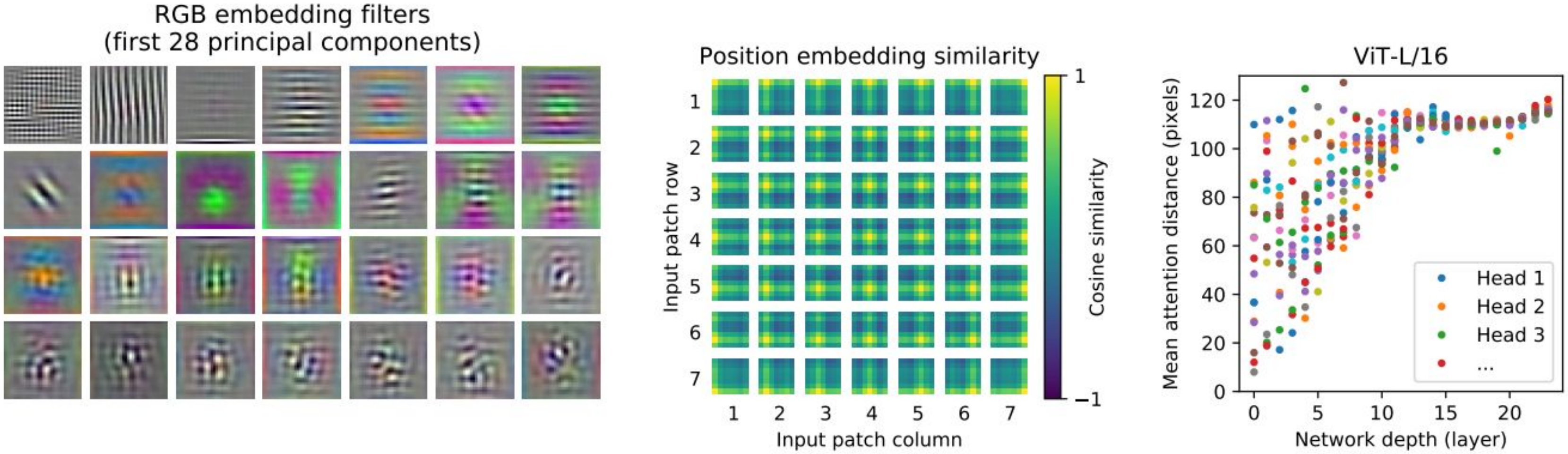
State Today

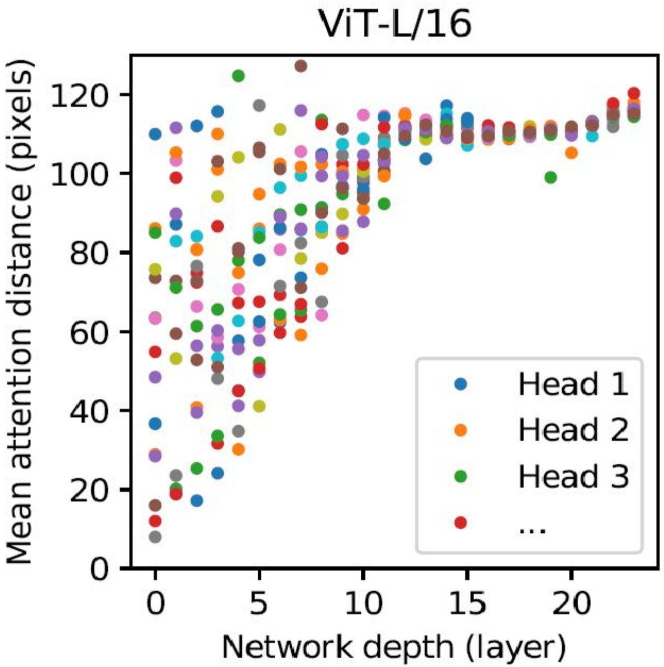
- Transformer methods are at the top of Leaderboards
- Lager Vision Transformer Vit-G/14 with larger pre-train set
- Transformers can also be used for detection and segmentation tasks
- Self-supervision in the next talk

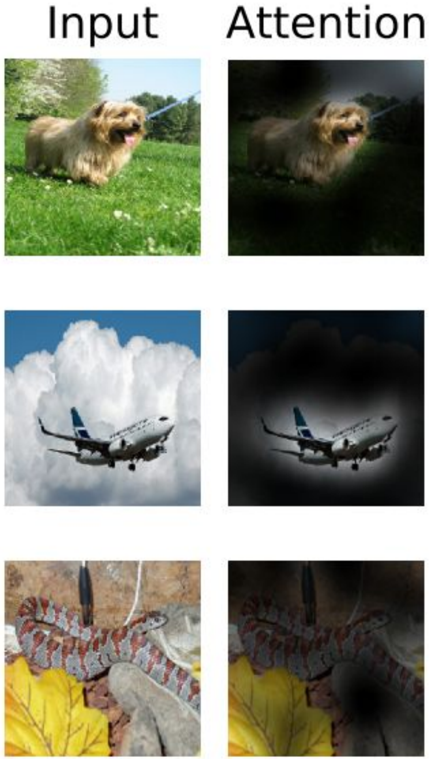
Conclusion

- It is possible to use Transformers for Image Classification
- Large pre-train datasets are needed
- Perform better than comparable ResNets with the same amount of compute
- Self-supervision is also possible

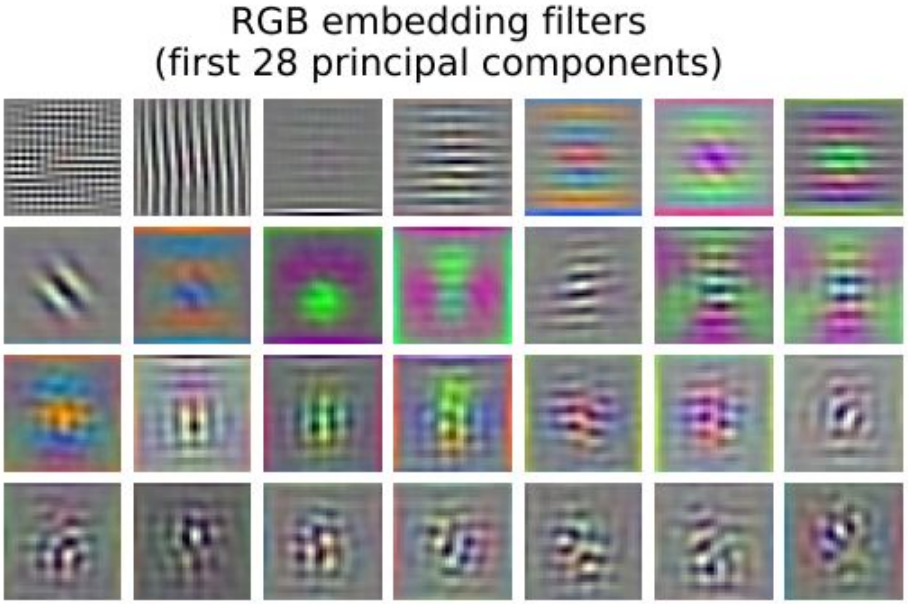
Appendix

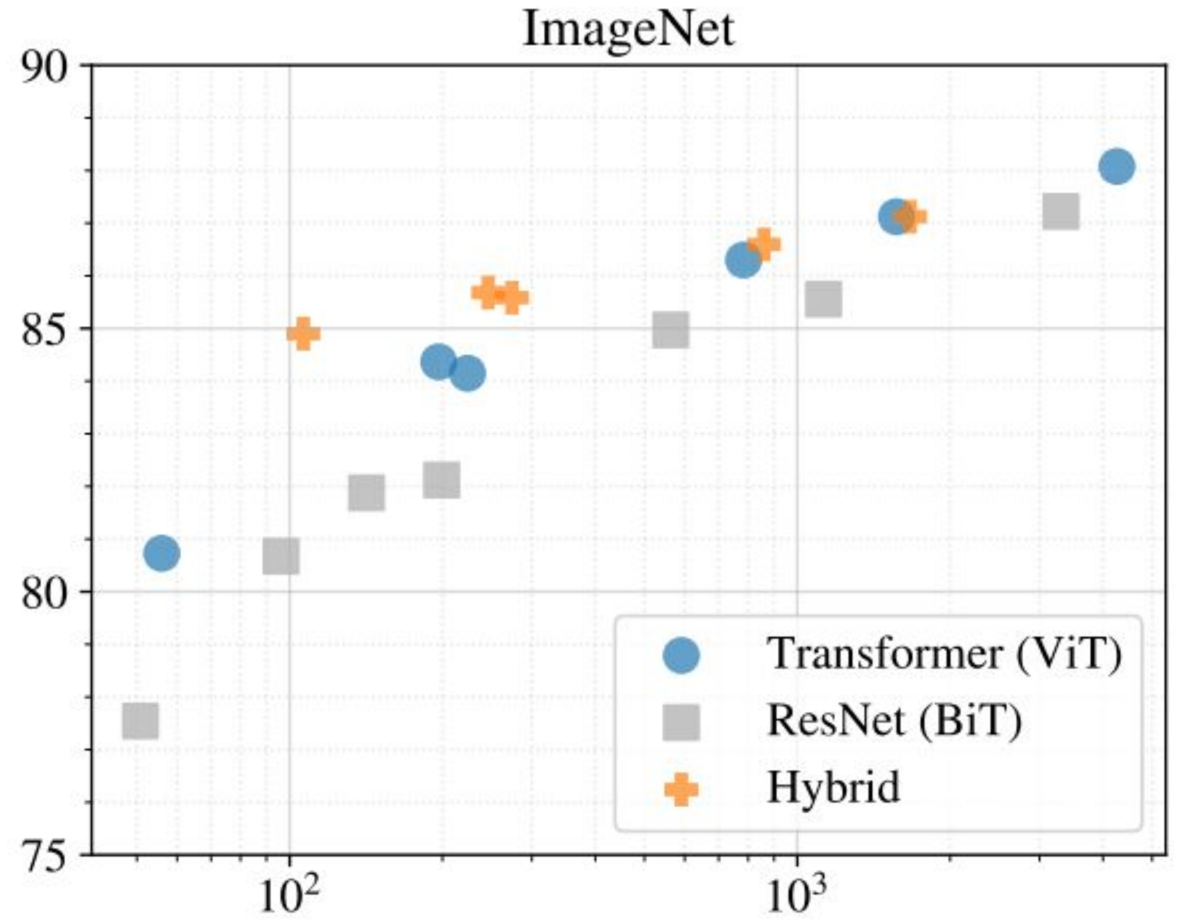

RGB embedding filters
(first 28 principal components)
Position embedding similarity
Input patch row
Input patch column
Cosine similarity
1
−1
ViT-L/16
Mean attention distance (pixels)
Network depth (layer)
Head 1
Head 2
Head 3
...

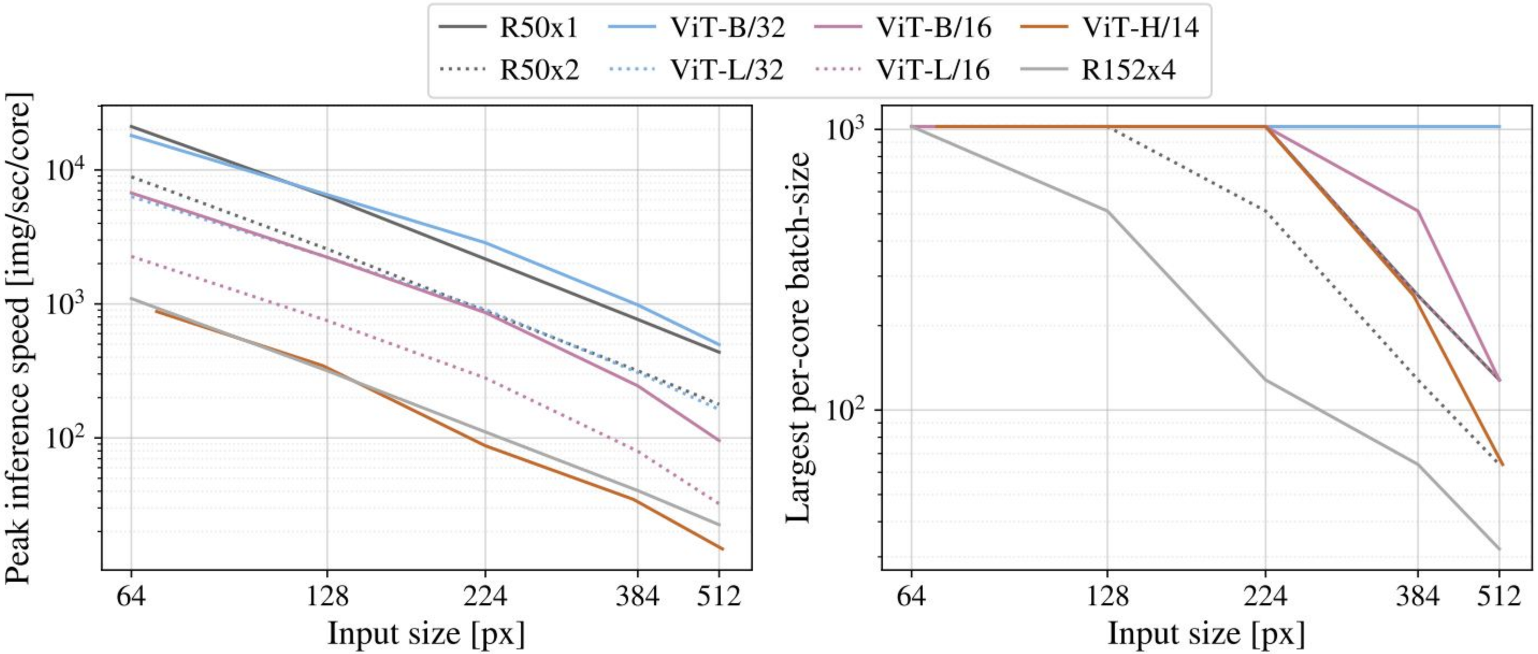

ViT-L/16
Mean attention distance (pixels)
Network depth (layer)
0
20
40
60
80
100
120
0
5
10
15
20
Head 1
Head 2
Head 3
...


Input
Attention

Patch Embeddings




ImageNet
90
85
80
75
10^2
10^3
Transformer (ViT)
ResNet (BiT)
Hybrid



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]

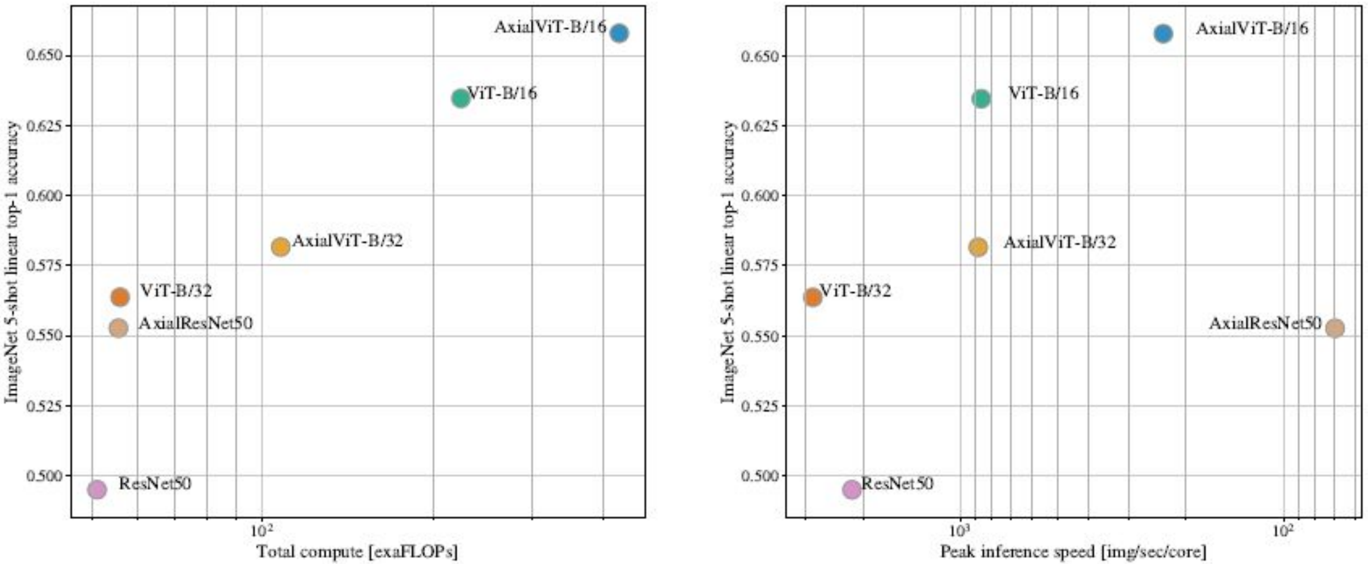
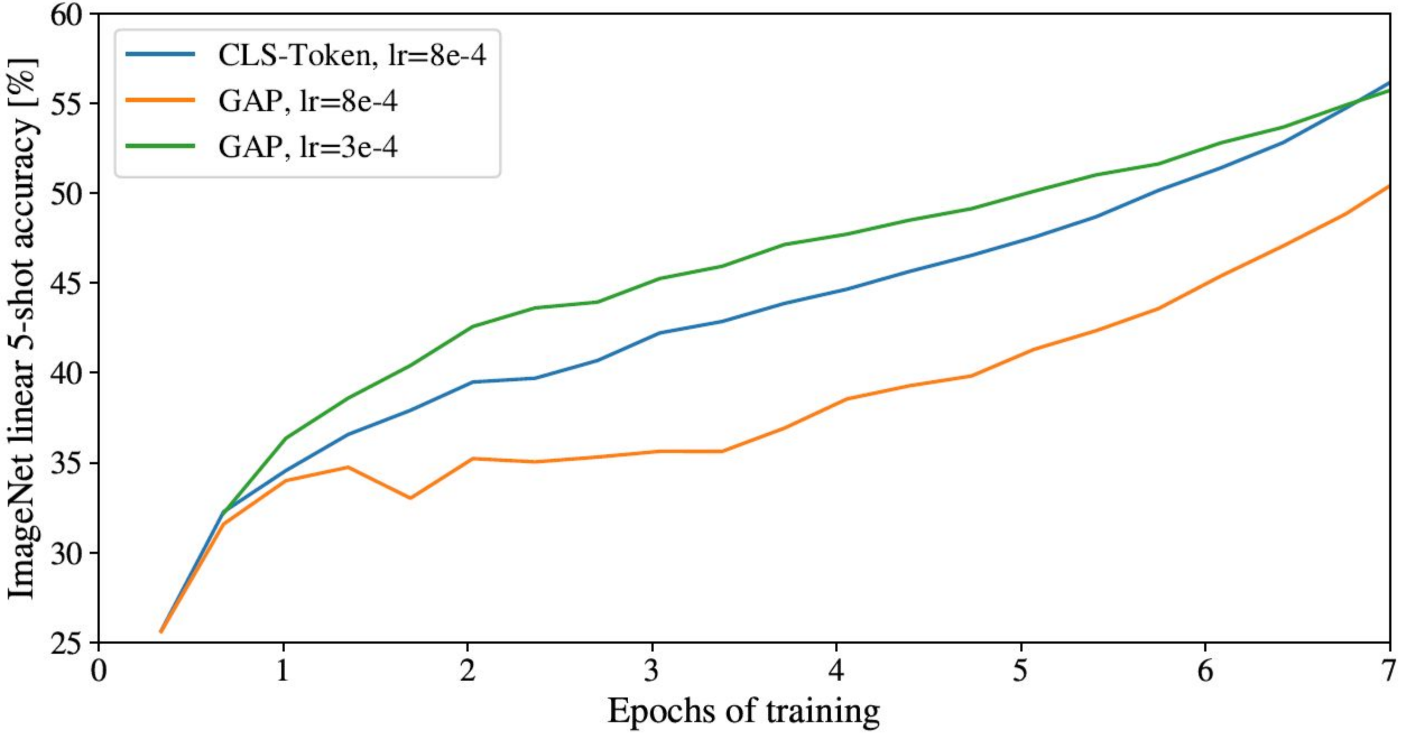


Figure 13: Performance of Axial-Attention based models, in terms of top-1 accuracy on ImageNet 5-shot linear, versus their speed in terms of number of FLOPs (**left**) and inference time (**left**).

		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	99.16	99.13	99.27
	CIFAR-100	91.67	91.97	93.44	93.04	93.82
	ImageNet	83.97	81.28	85.15	80.99	85.13
	ImageNet ReaL	88.35	86.63	88.40	85.65	88.70
	Oxford Flowers-102	99.38	99.11	99.61	99.19	99.51
	Oxford-IIIT-Pets	94.43	93.02	94.73	93.09	94.82
JFT-300M	CIFAR-10	99.00	98.61	99.38	99.19	99.50
	CIFAR-100	91.87	90.49	94.04	92.52	94.55
	ImageNet	84.15	80.73	87.12	84.37	88.04
	ImageNet ReaL	88.85	86.27	89.99	88.28	90.33
	Oxford Flowers-102	99.56	99.27	99.56	99.45	99.68
	Oxford-IIIT-Pets	95.80	93.40	97.11	95.83	97.56

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



CLS-Token, lr=8e-4
GAP, lr=8e-4
GAP, lr=3e-4
ImageNet linear 5-shot accuracy [%]
Epochs of training
60
55
50
45
40
35
30
25
0
1
2
3
4
5
6
7