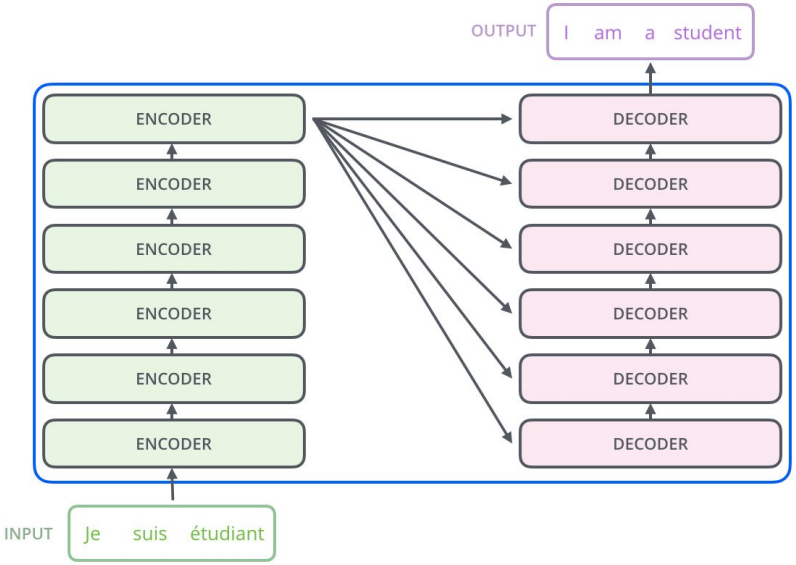
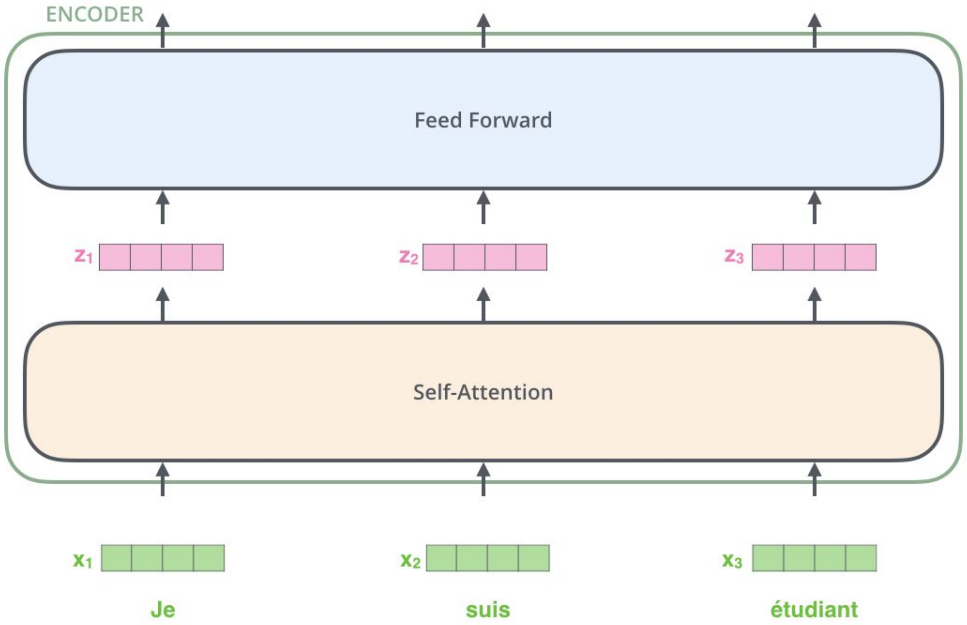
Paper Overview: AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

October 2020, Parsa Torabian, Deep Learning Reading Club

- Transformer Review (Attention is all you need)
- Transformer for images
- Results
- Discussion time

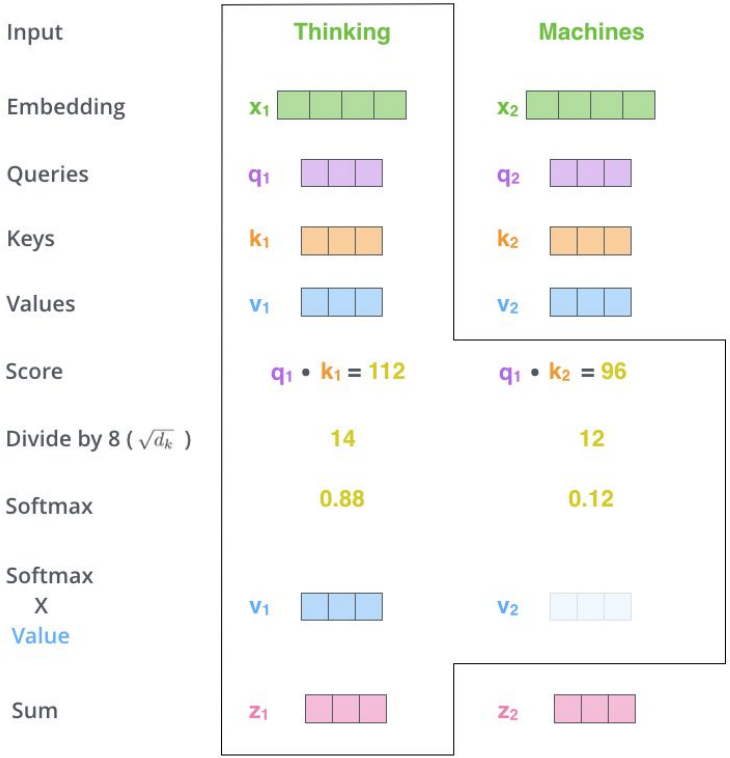
Transformers Review

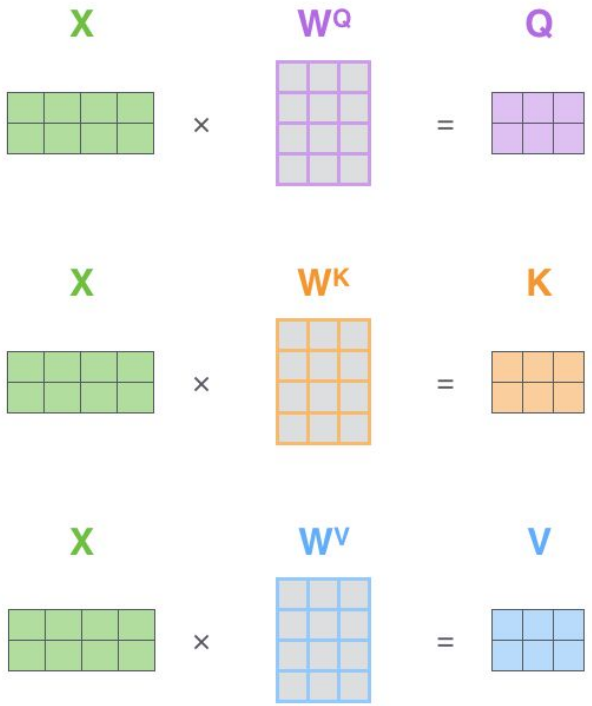




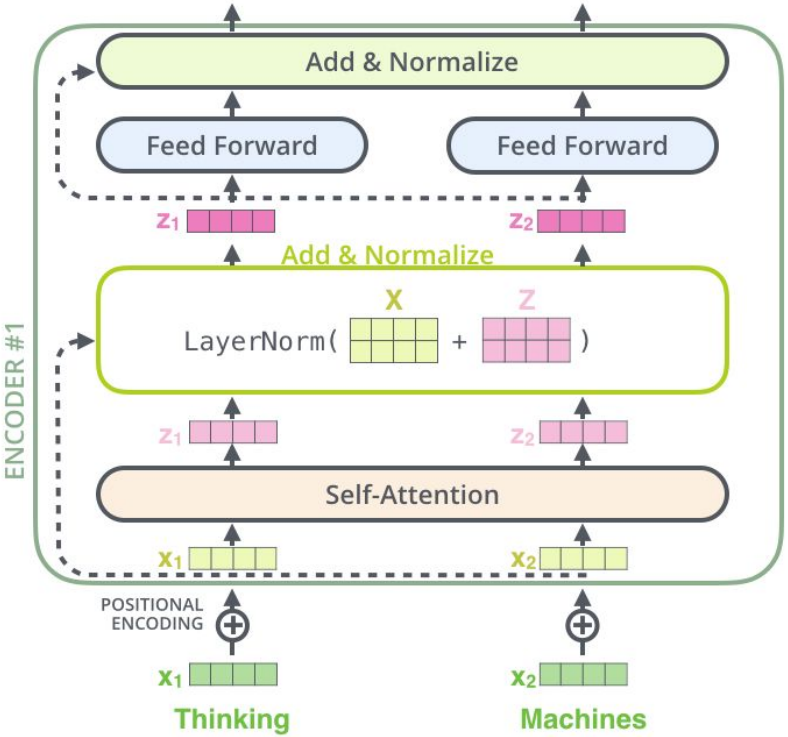
Images from: http://jalammar.github.io/illustrated-transformer/

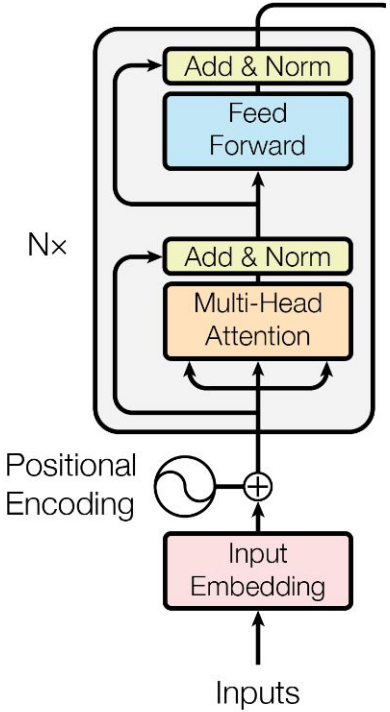
Self Attention (Transformers Review)



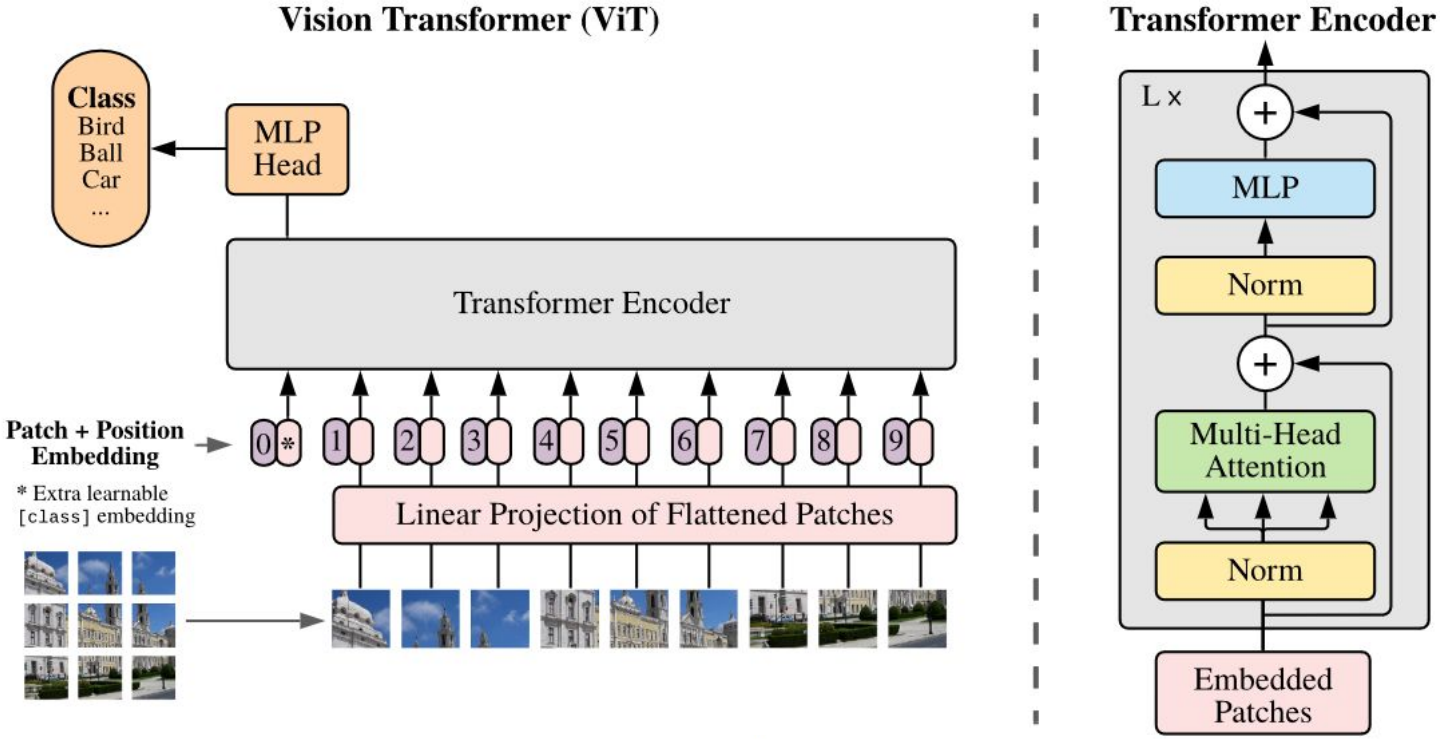


Putting it all together (transformers review)





Transformers for Images

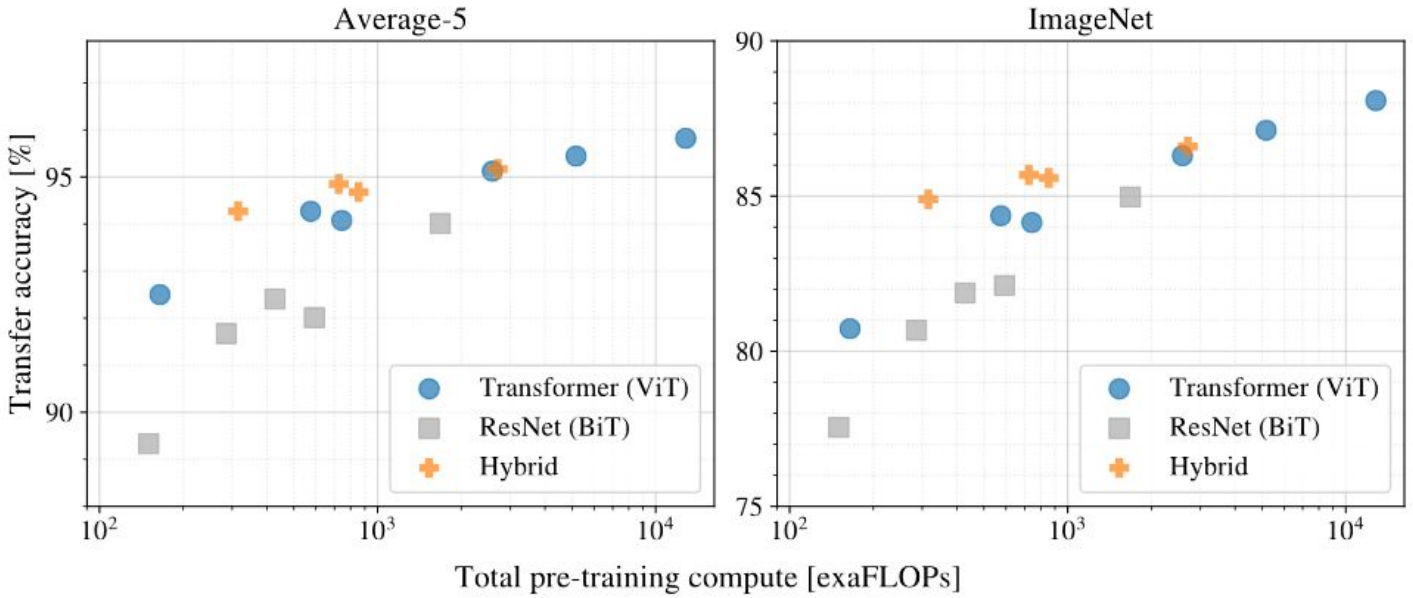


$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}}; \mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \cdots; \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}, \qquad \mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D}, \ \mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D}$$

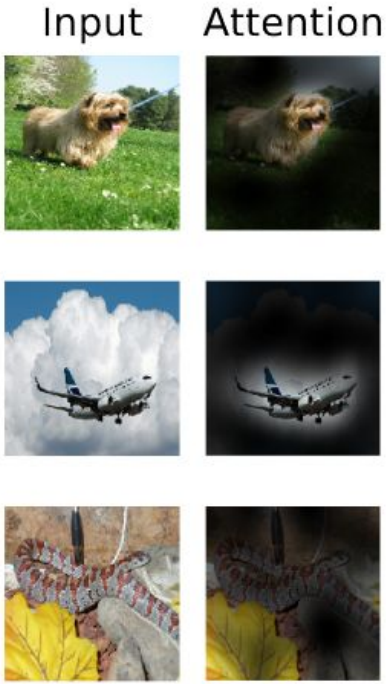
Results

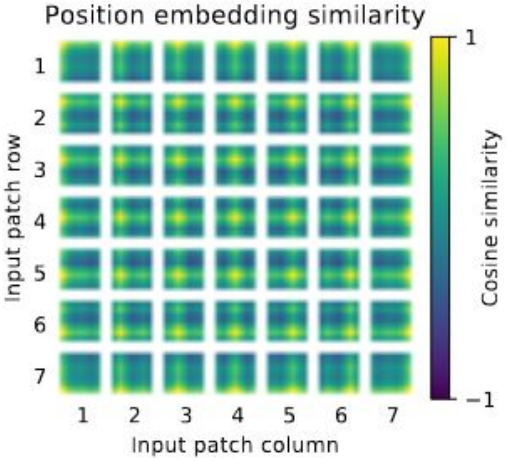
	Ours (ViT-H/14)	Ours (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	88.36	87.61 ± 0.03	87.54 ± 0.02	$88.4/\mathbf{88.5}^{*}$
ImageNet ReaL	**90.77**	90.24 ± 0.03	90.54	90.55
CIFAR-10	$\mathbf{99.50} \pm 0.06$	99.42 ± 0.03	99.37 ± 0.06	–
CIFAR-100	$\mathbf{94.55} \pm 0.04$	93.90 ± 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

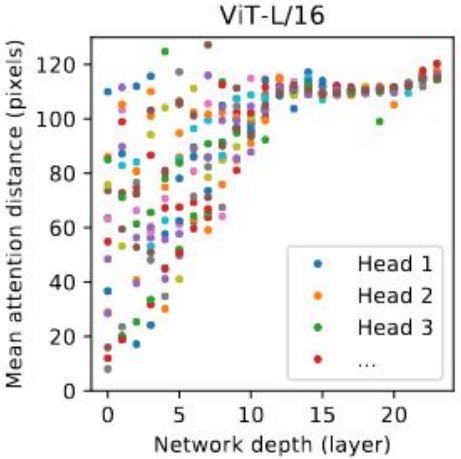
Results

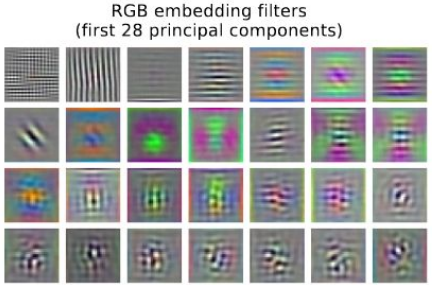

Average-5
ImageNet
Transfer accuracy [%]
95
90
90
85
80
75
10^2
10^3
10^4
Transformer (ViT)
ResNet (BiT)
Hybrid
Total pre-training compute [exaFLOPs]

Results


Input
Attention


Position embedding similarity
Input patch row
Input patch column
Cosine similarity
1
2
3
4
5
6
7
1
−1


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


RGB embedding filters
(first 28 principal components)

Takeaways/Discussion

- Data can make up for inductive biases
- Transformers are general enough for images