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

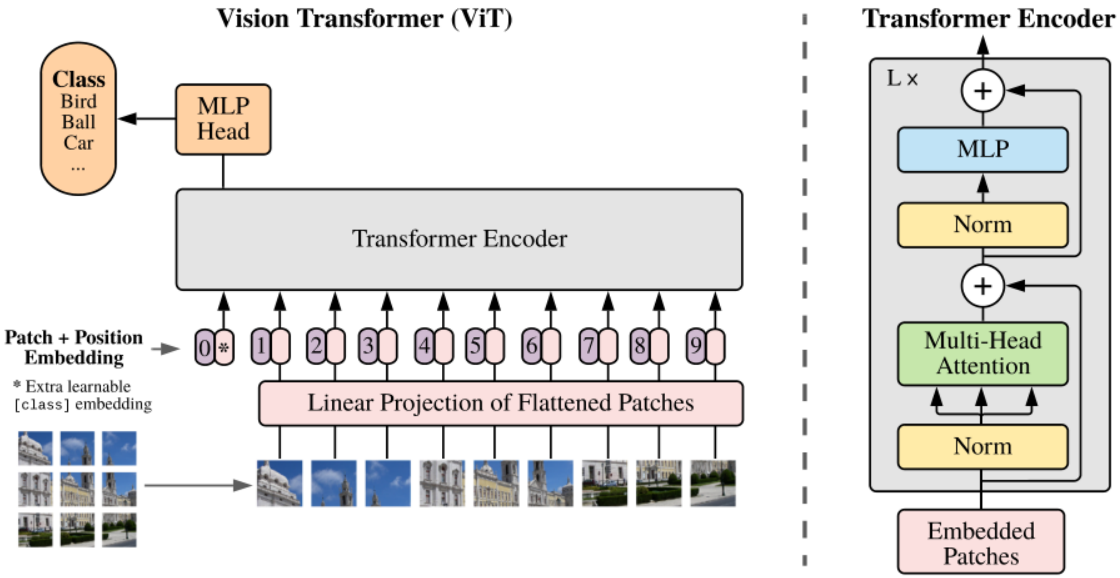
Dosovitskiy Et. Al.

What is a vision transformer (ViT)?

What is a vision transformer (ViT)?

- In this paper they propose an architecture to replace a conventional CNN architecture
- Can be combined with a CNN -> They argue causes complex engineering and performance challenges

What is a vision transformer (ViT)?

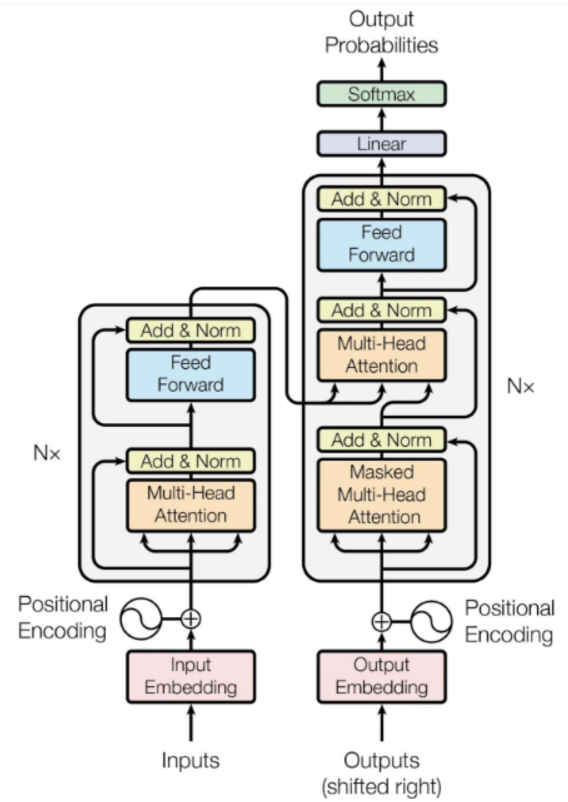


What is a vision transformer (ViT)?

Architecture

1. Split images into fixed-size patches
2. Linearly embed the image patches
3. Add position embeddings
4. Feed resultant sequence of vectors to a standard Transformer encoder
5. Add an extra learnable “classification token” to the sequence in order to perform classification

What is a Transformer?



https://jinglescode.github.io/2020/05/27/illustrated-guide-transformer/

What is a Transformer?

- Comes from NLP research
- They state that it is considered the norm for NLP tasks
- Has an encoder stage and a decoder stage
- Vectorises the input
- Embedding to represent meaning
- Has a positional vector for each component

What is a Transformer?

- Fed into an Encoder attention block
- Generates attention vectors for every word

What is a Transformer?

- The decoder block is fed the input of data you want to transform the working set into i.e. The output language in a translation problem
- The attention vectors are also fed in
- The meaning of each word is encoded at the embedding layer
- Maps attention vectors between Encoder and Decoder
- Predicts next output (classification etc) using a feed-forward network
- Repeats until the end of the sentence or input is reached

Why propose this architecture?

They argue that for a minimal drop in accuracy with mid-sized datasets (i.e. ImageNet) over conventional networks liek ResNets.

However at large scale (14M - 300M images) they show that their architecture approaches or beats multiple benchmarks they have run

Why propose this architecture?

Scale

Results

	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

Implementation

- **They state that ViT has much less image-specific inductive bias over CNNs**

- **A hybrid architecture exists where instead of input images, feature maps are used**

Fine tuning

Large resolution images use the same patch size which creates longer sequenc

Fine tuning

The pre-trained position embeddings loose their meanings with higher resolution images

Solved by 2D interpolation of the positions

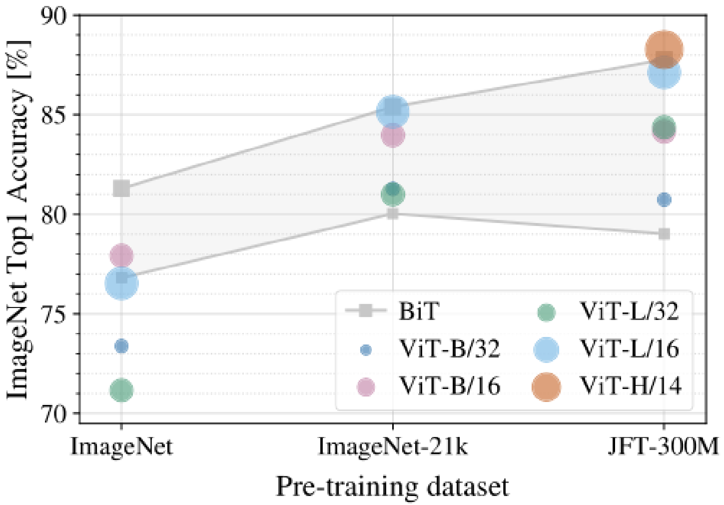
Experimental Setup

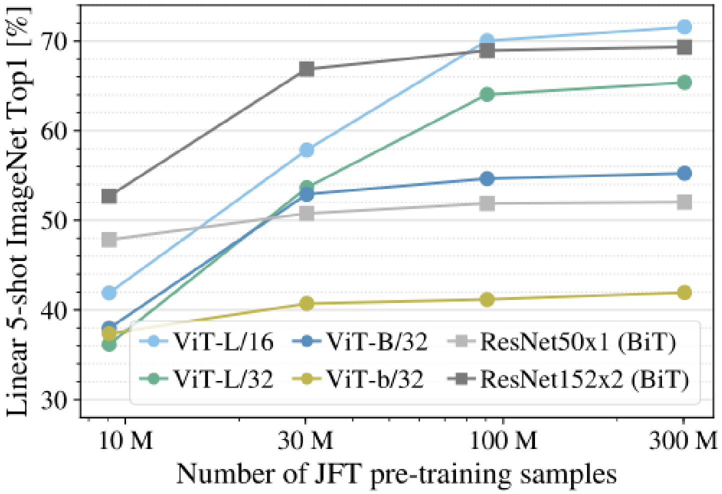
- Compare ResNet, Vision Transformer and a hybrid model
- 16 X 16 patch sizes (except ViT-H/14)
- Use varying sized classification datasets
- Several (ViT) variants are used

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.

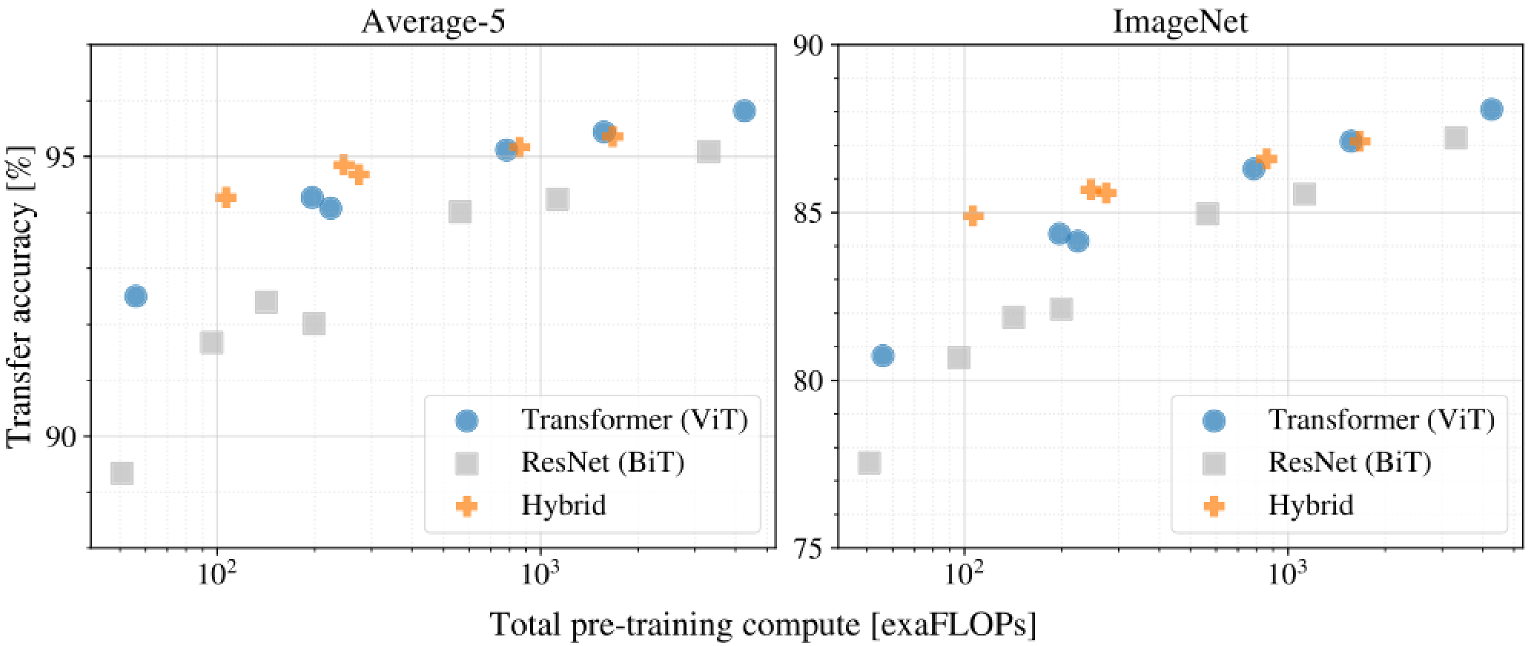
More Results





At smaller datasets ResNet beats ViT accuracy but it overtakes in performance as the datasets grow

More Results - How do the models scale?



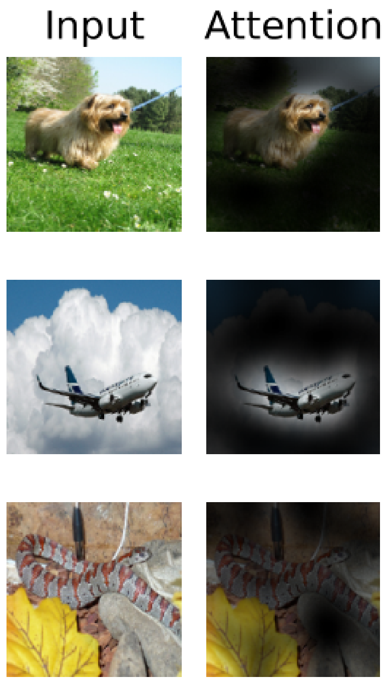


Figure 6: Representative examples of attention from the output token to the input space. See Appendix D.7 for details.

The Code

- All code is provided
- Also Collabratory links
- Can supply your own datasets

Some Links:

https://github.com/google-research/vision_transformer

https://github.com/google-research/vision_transformer/blob/master/vit_jax/models.py

https://colab.research.google.com/github/google-research/vision_transformer/blob/master/vit_jax.ipynb

https://colab.research.google.com/github/google-research/vision_transformer/blob/master/vit_jax_augreg.ipynb

The Code

- The code is nice and simple
- The architecture used is moderately simple
- Re-implemented ResNet using Jax and Flax (Google specific library)
- Other implementations of some models are available:
- https://github.com/google-research/big_transfer
- Excludes ViT

Conclusion

- An interesting application of a new kind of architecture
- Ridiculous scale (at the moment) is required for model performance gains
- Only looked at classification problems
- Simplicity is always welcome
- Open source code is a plus
- Some more in-depth sanity checks on the network (beyond just visualising the attention would have been a nice bonus)

Fin.