Vision Transformer (ViT)

기초심화CV팀
김수란

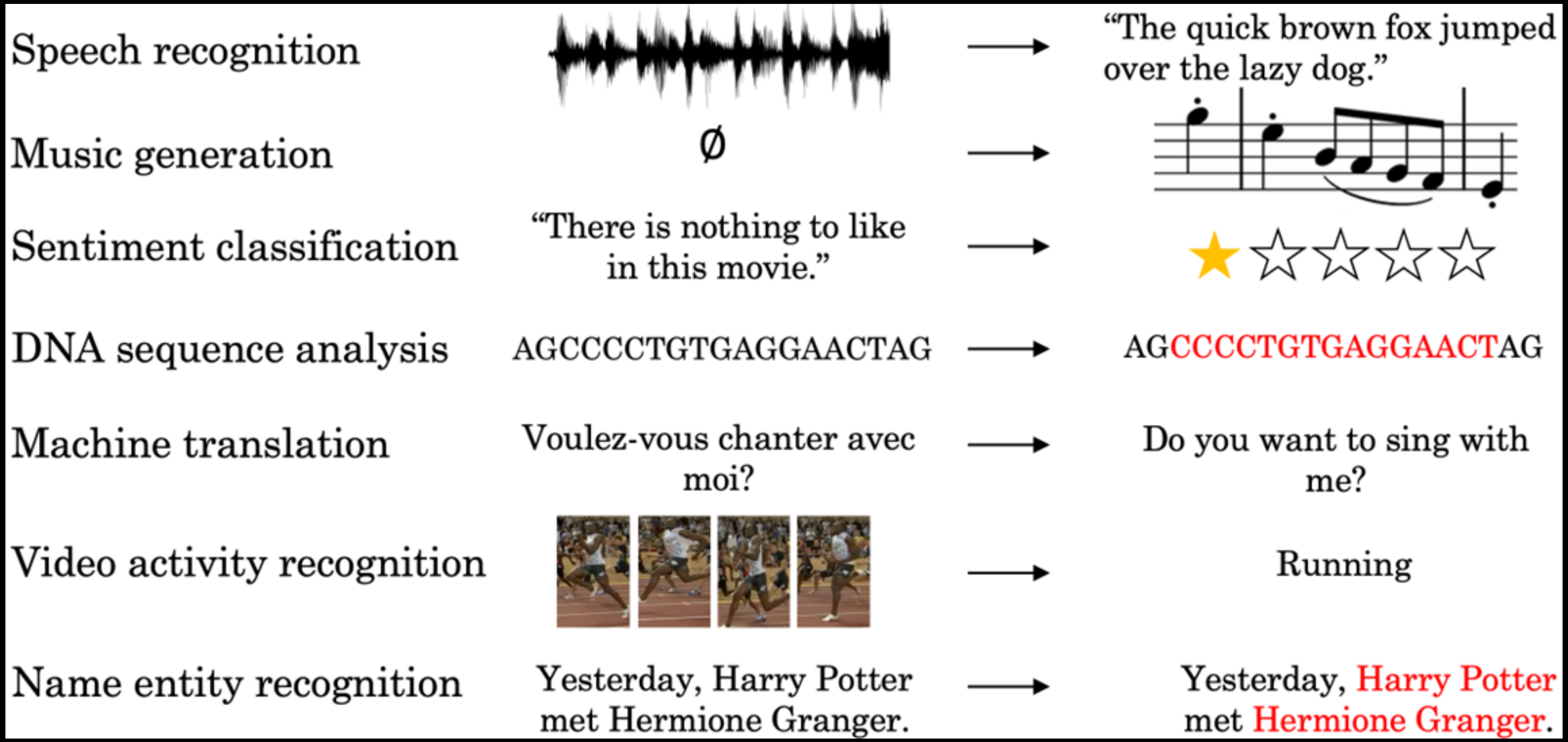
Table of contents

01

Background of ViT

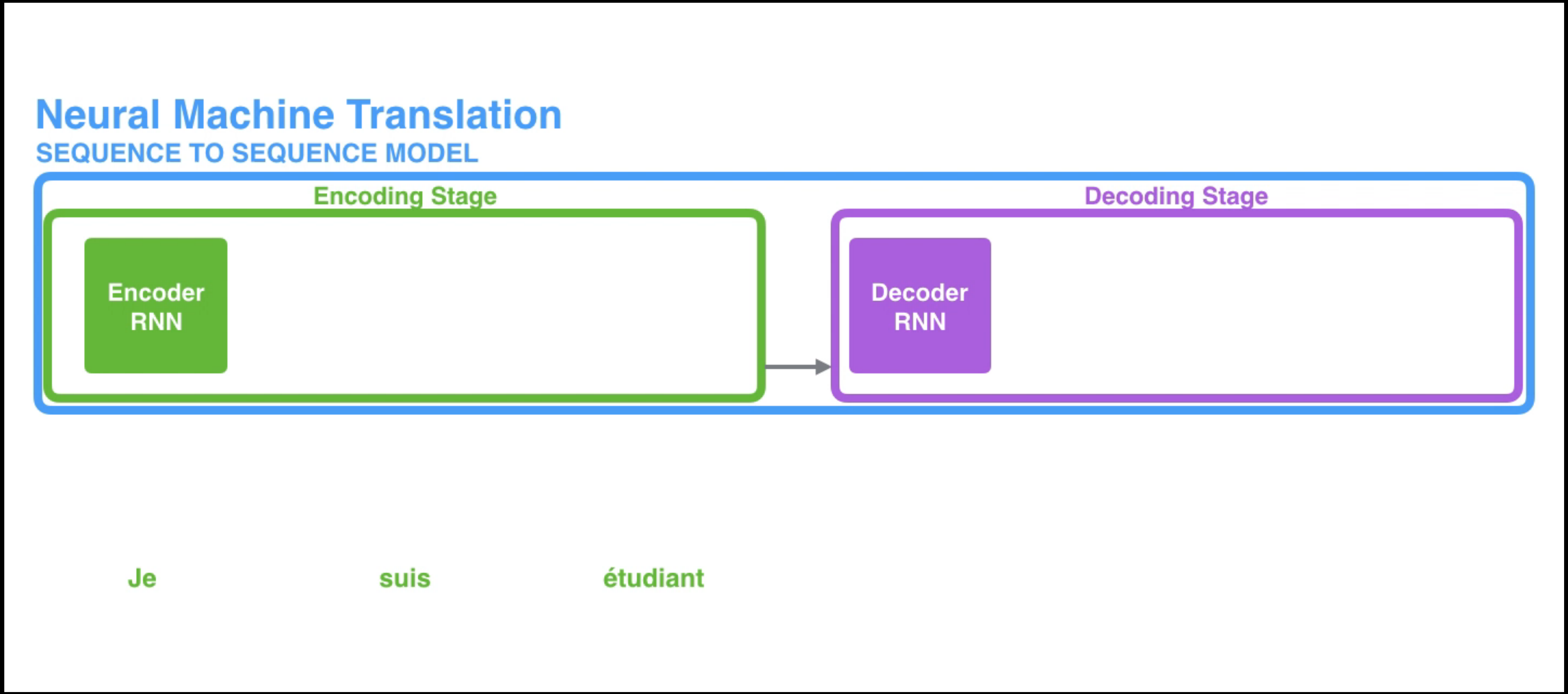
Sequential Data

Sequential data is common in various fields.



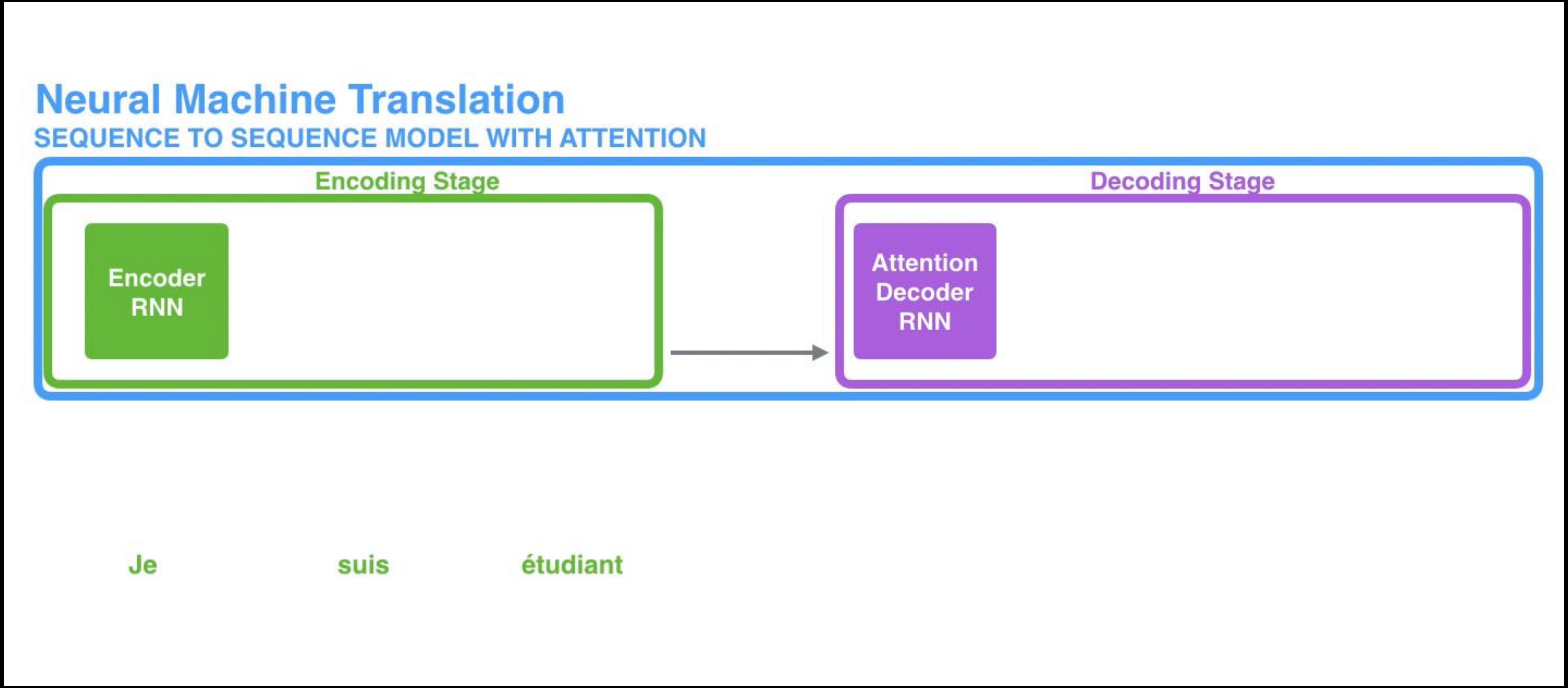
https://wooono.tistory.com/241

Seq2Seq Model with RNN



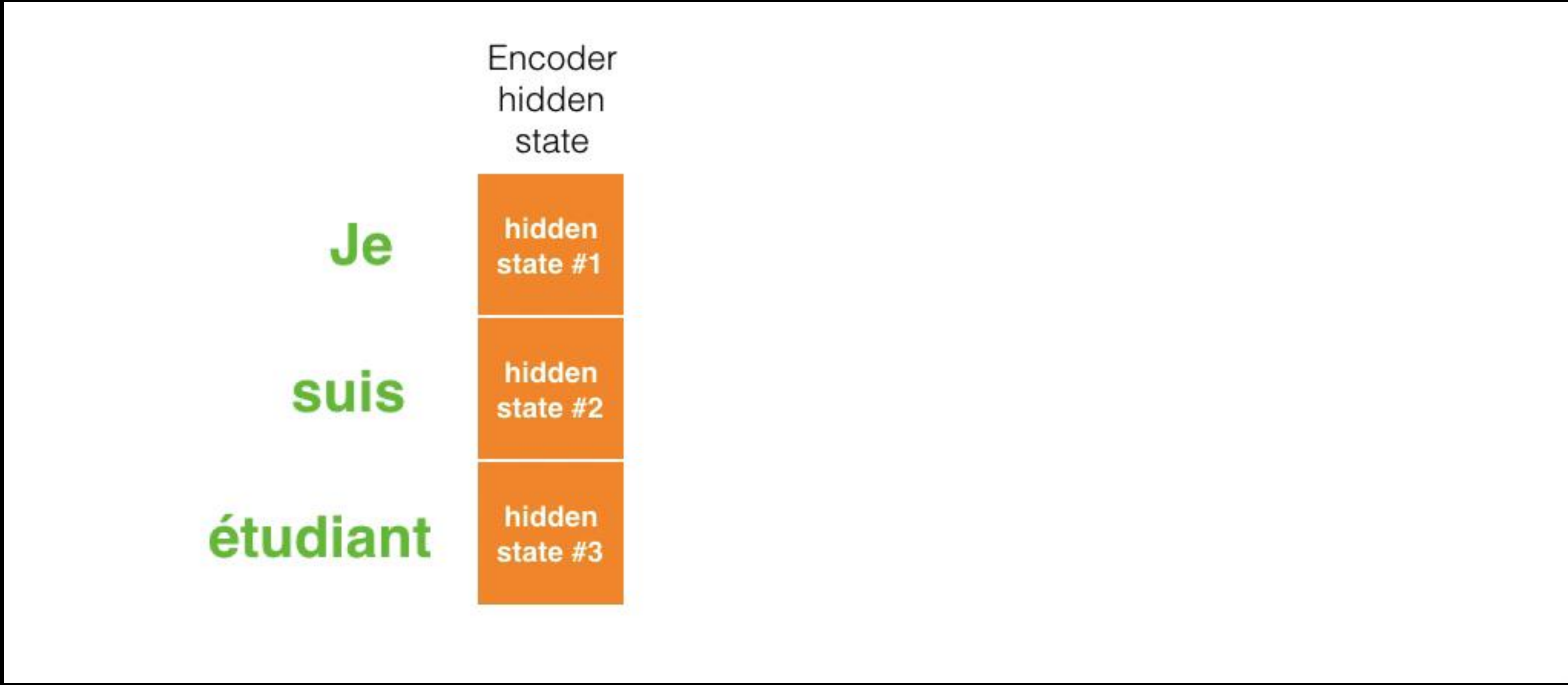
https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/

Seq2Seq Model with Attention



https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/

Seq2Seq Model with Attention



https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/

Attention is great

- Attention significantly improves performance (in many applications)
  - It’s very useful to allow decoder to focus on certain parts of the source
- Attention solves the bottleneck problem
  - Attention allows decoder to look directly at source; bypass bottleneck
- Attention helps with vanishing gradient problem
  - Provides shortcut to faraway states
- Attention provides some interpretability
  - By inspecting attention distribution, we can see what the decoder was focusing on

Abigail See, Stanford CS224n

Transformer

Attention Is All You Need

Ashish Vaswani[*]
Google Brain
avaswani@google.com

Noam Shazeer[*]
Google Brain
noam@google.com

Niki Parmar[*]
Google Research
nikip@google.com

Jakob Uszkoreit[*]
Google Research
usz@google.com

Llion Jones[*]
Google Research
llion@google.com

Aidan N. Gomez[*] [†]
University of Toronto
aidan@cs.toronto.edu

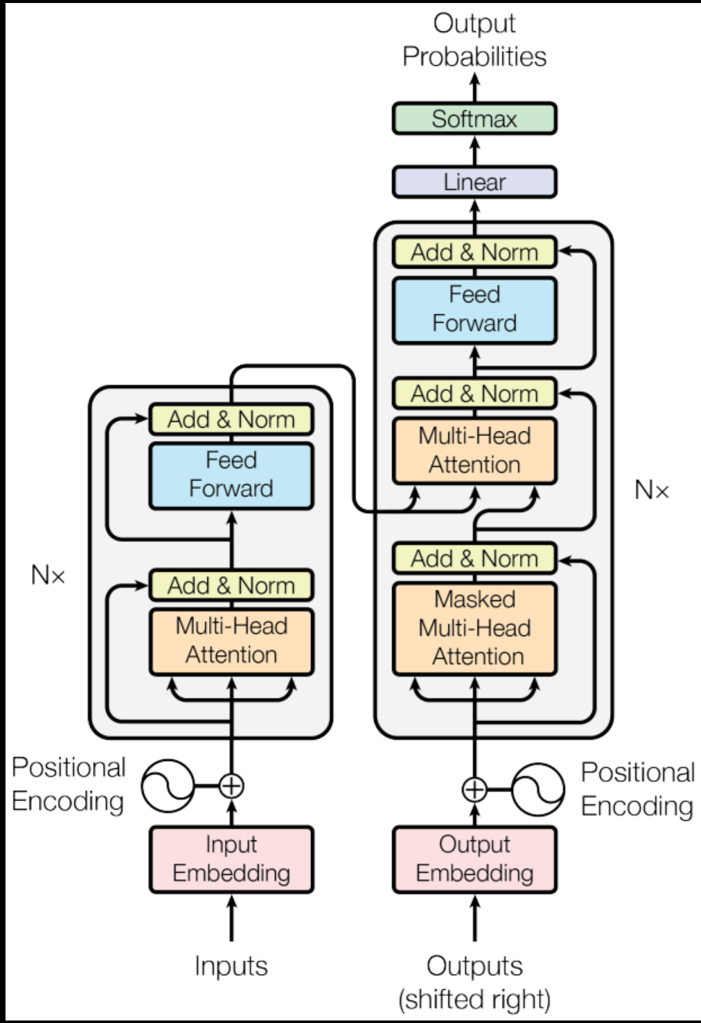
Łukasz Kaiser[*]
Google Brain
lukaszkaiser@google.com

Illia Polosukhin[*] [‡]
illia.polosukhin@gmail.com

https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Transformer

Encoder ->



<- Decoder

https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Transformer

Transformers have become the model of choice in NLP.

In CV, however, classic CNN architectures are still SOTA.

Chat GPT

Vision Transformer!

Published as a conference paper at ICLR 2021

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

Alexey Dosovitskiy[*,†], Lucas Beyer[*], Alexander Kolesnikov[*], Dirk Weissenborn[*], Xiaohua Zhai[*], Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby[*,†]

[*]equal technical contribution, [†]equal advising

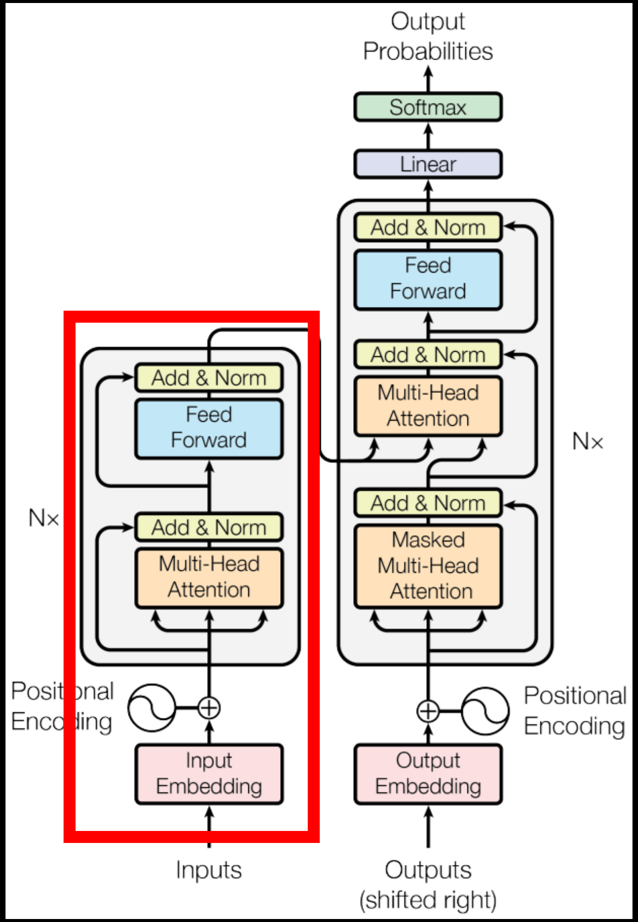
Google Research, Brain Team

{adosovitskiy, neilhoulsby}@google.com

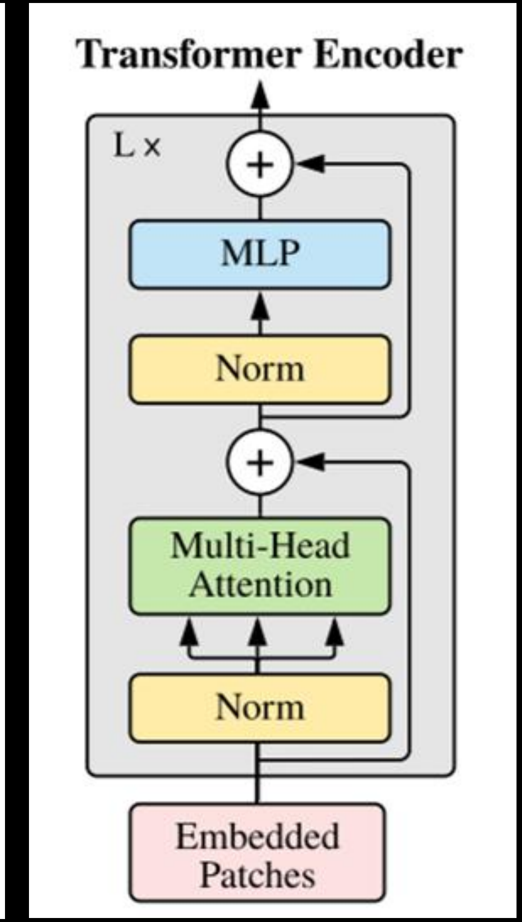
02

ViT Architecture and Workflow

Transformer -> Vision Transformer (ViT)



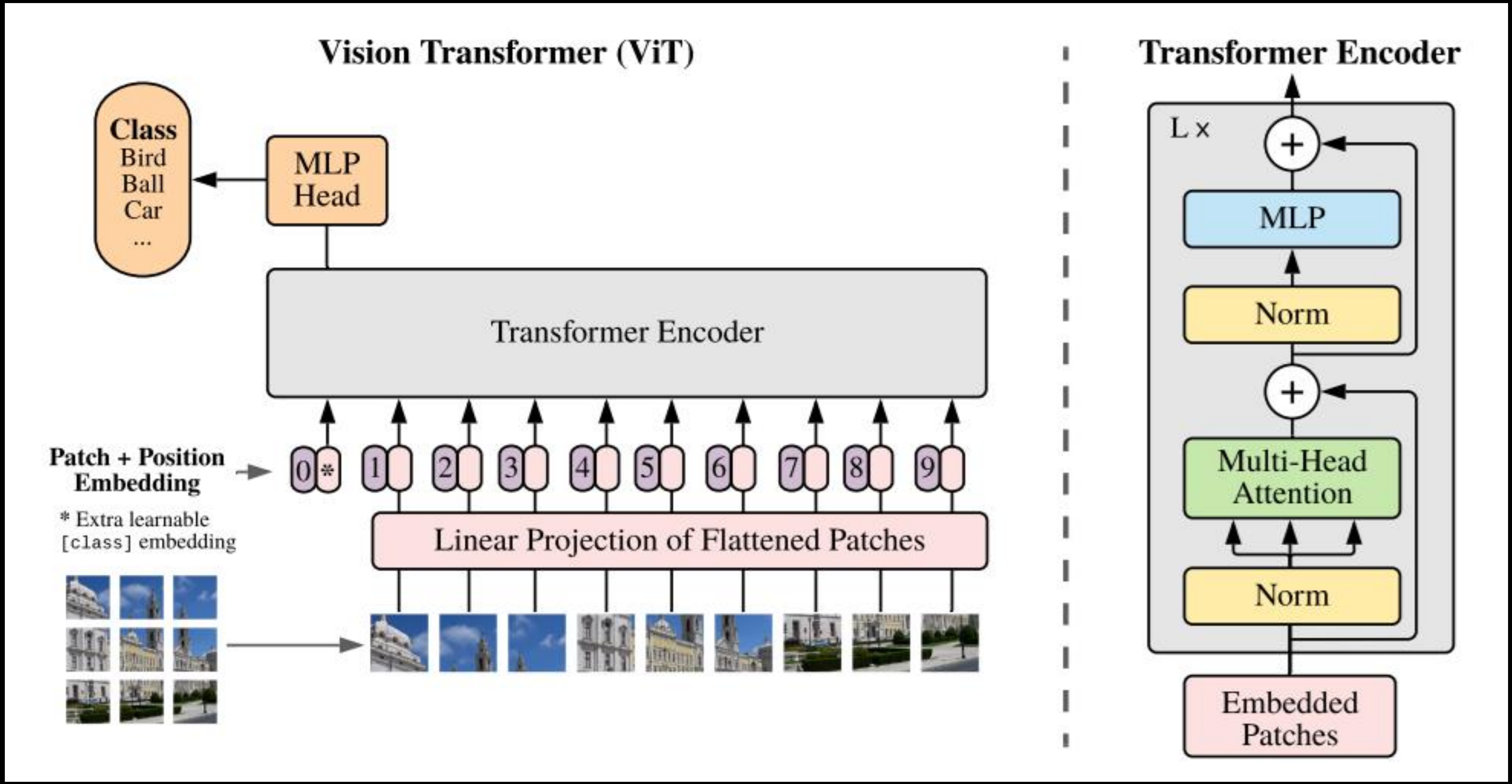
Transformer



ViT

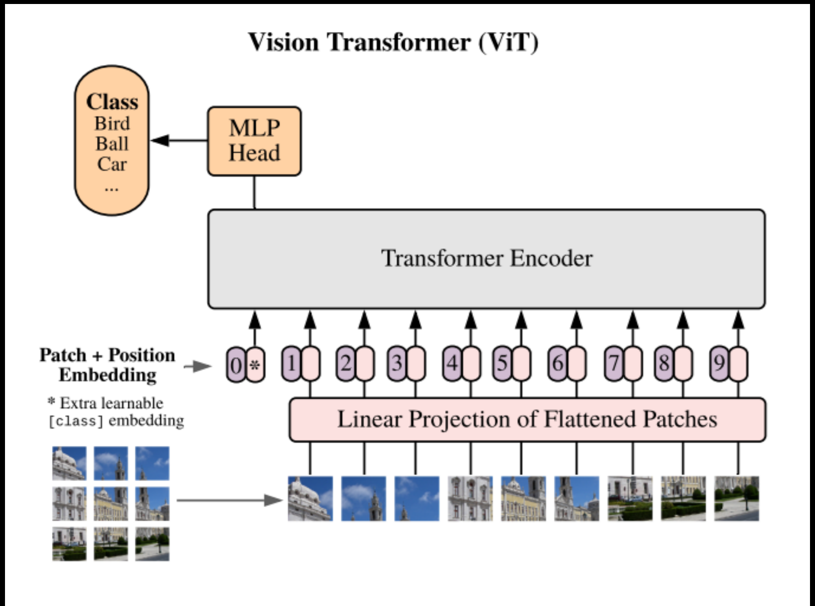
- Note that small difference compared to the transformer encoder.
- Norm comes before the MHA and MLP.

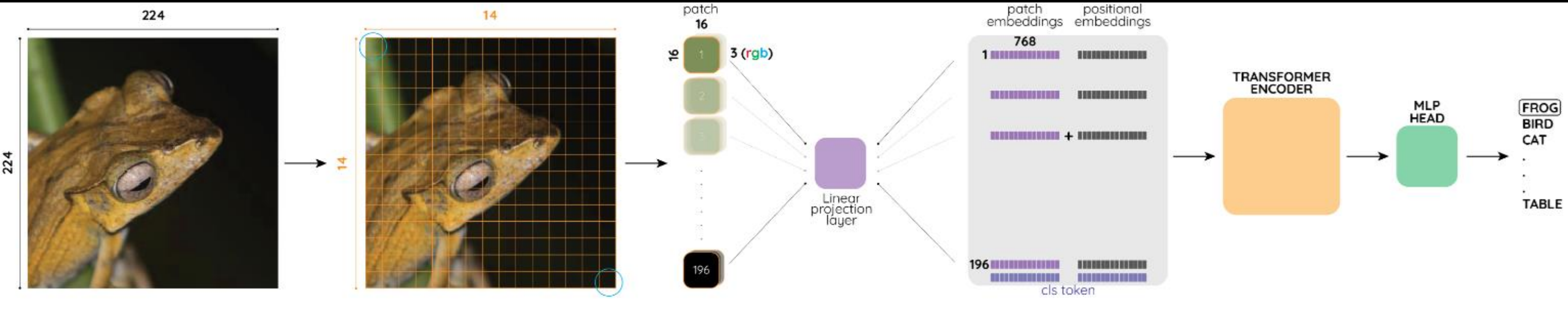
Vision Transformer (ViT)



https://arxiv.org/pdf/2010.11929

ViT: Step-by-Step Example





https://amaarora.github.io/posts/2021-01-18-ViT.html

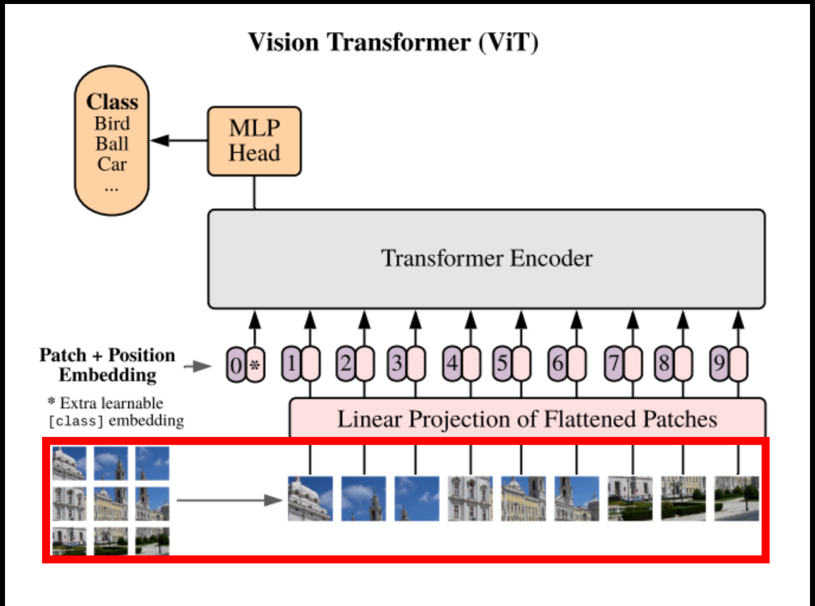
Patches

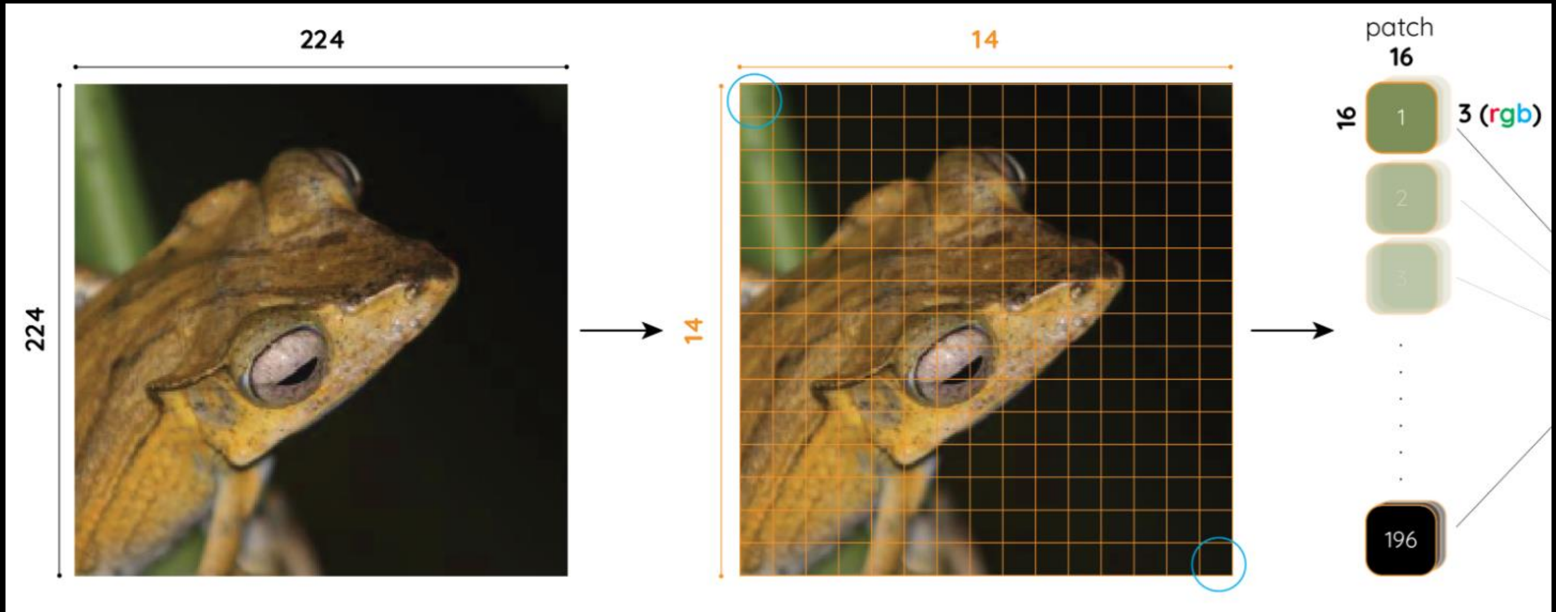
- Using an image as a sequence of patches.
- Partition an image into equal sized patches.

    224*224*3 image to 14 patches in each dimension.

    Thus, patch size = 16*16*3

    Total of 196 = 14*14 patches

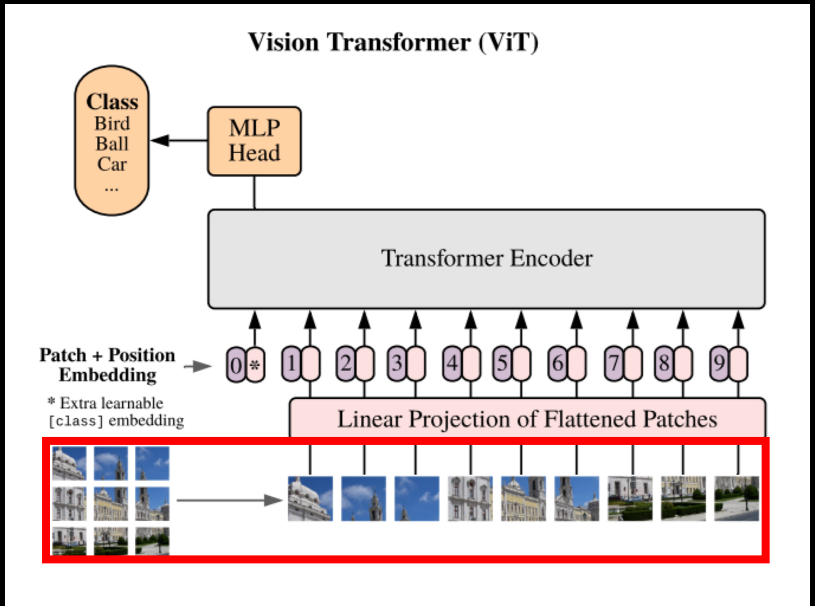


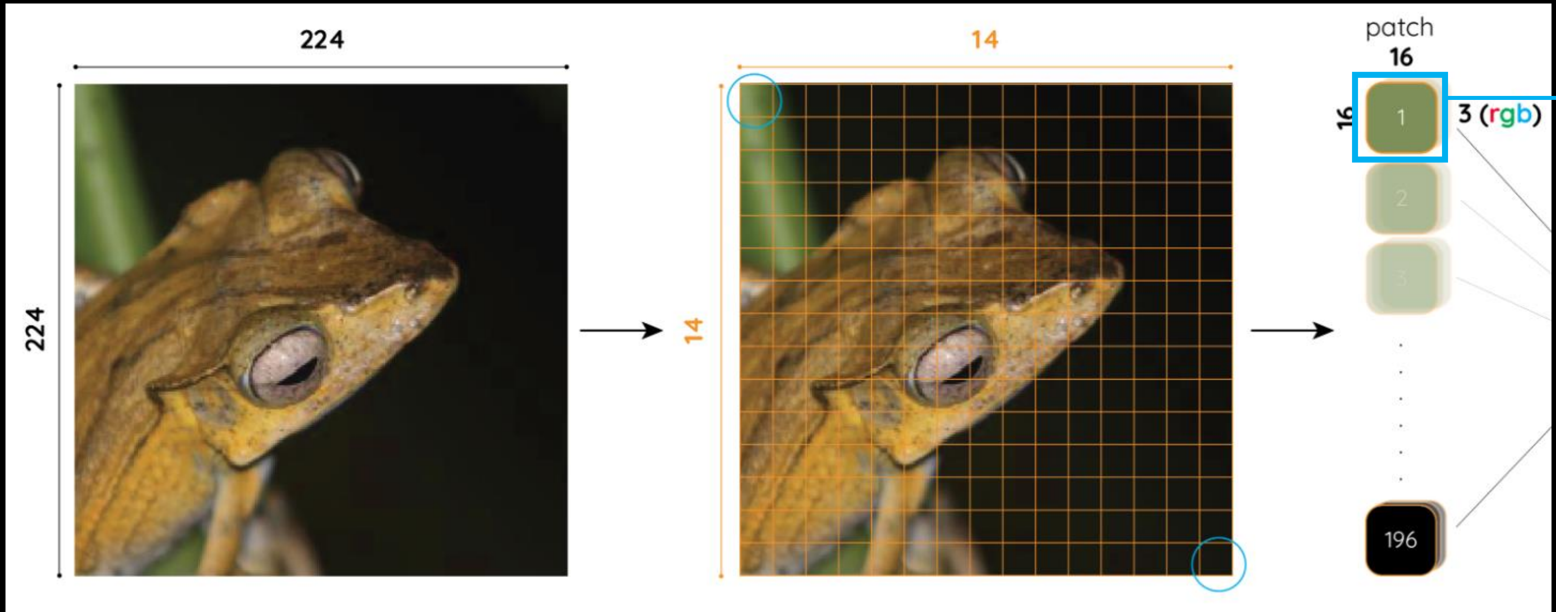


Flattening

- Each patch is technically 16*16*3 dimension.
- Flattening (a.k.a. vectorization):

  Tensor becomes a 1D-vector

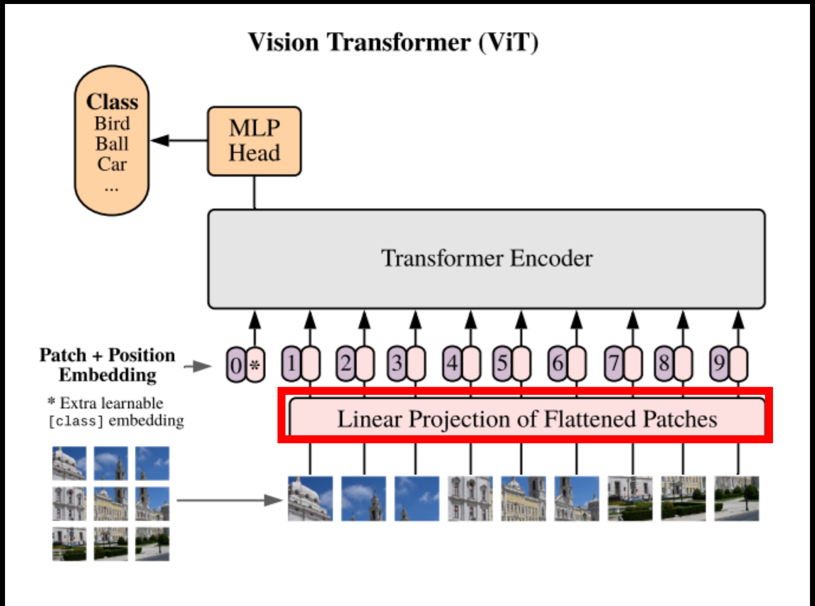


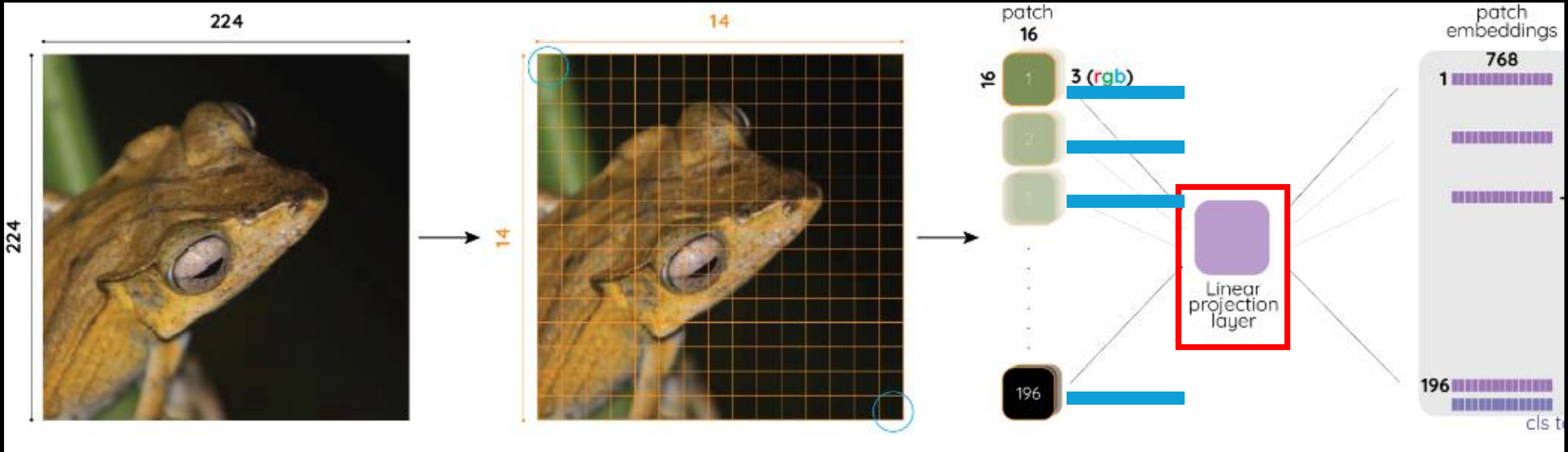


768 = 16x16x3

Patch Embedding

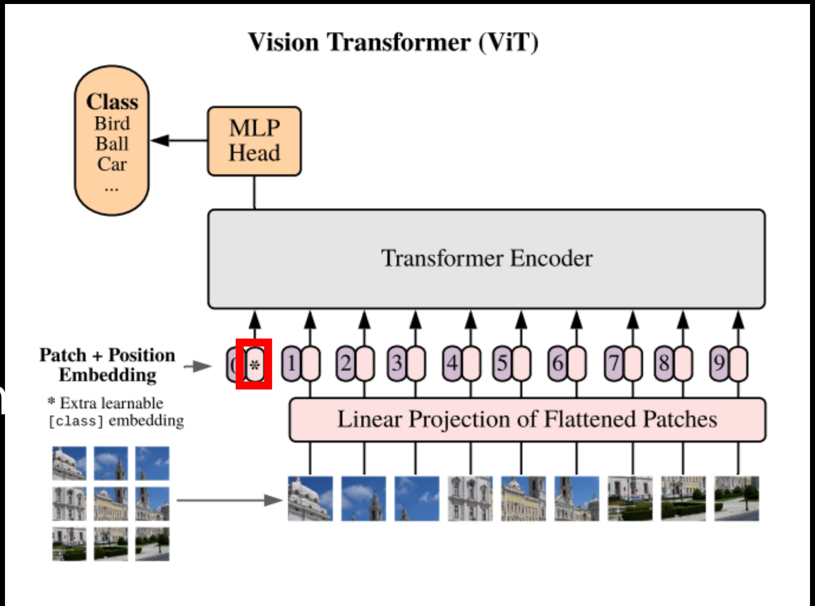
- Each 768*1 flattened patch goes through a fully connected layer to become a new 768D embedding.

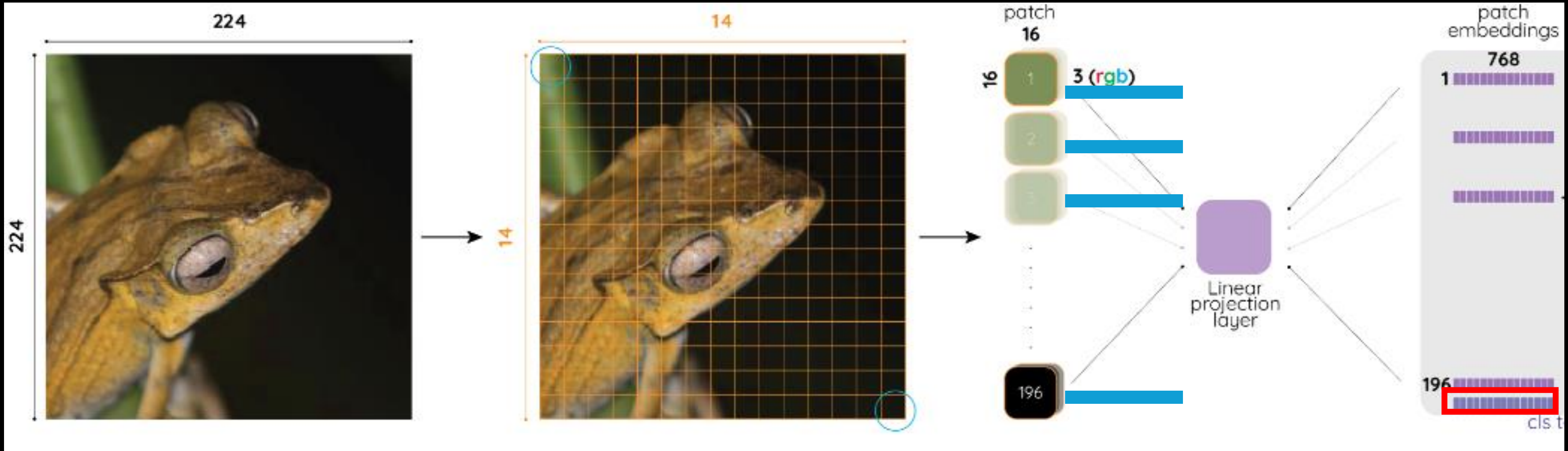




Patch Embedding

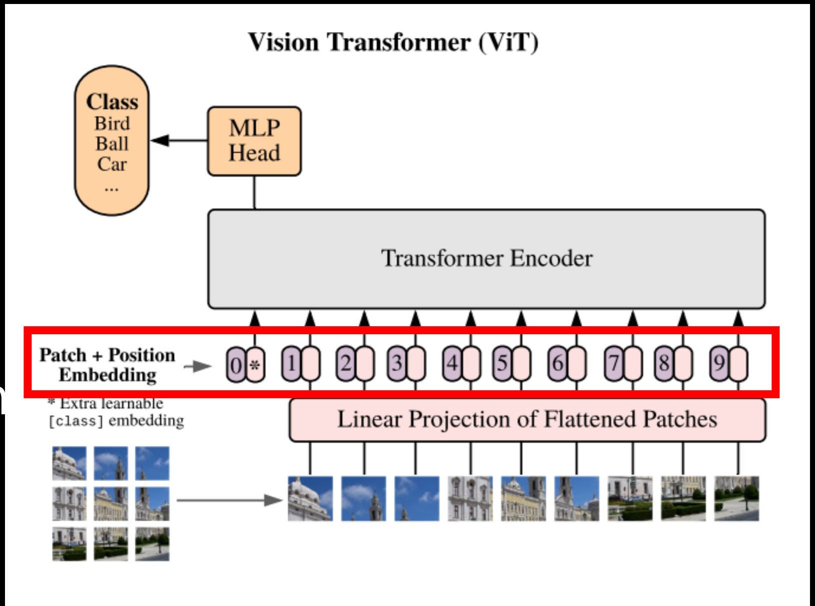
- Extra learnable class embedding with non-patch information
- 197 = 196(patch embeddings) + 1(class embedding)

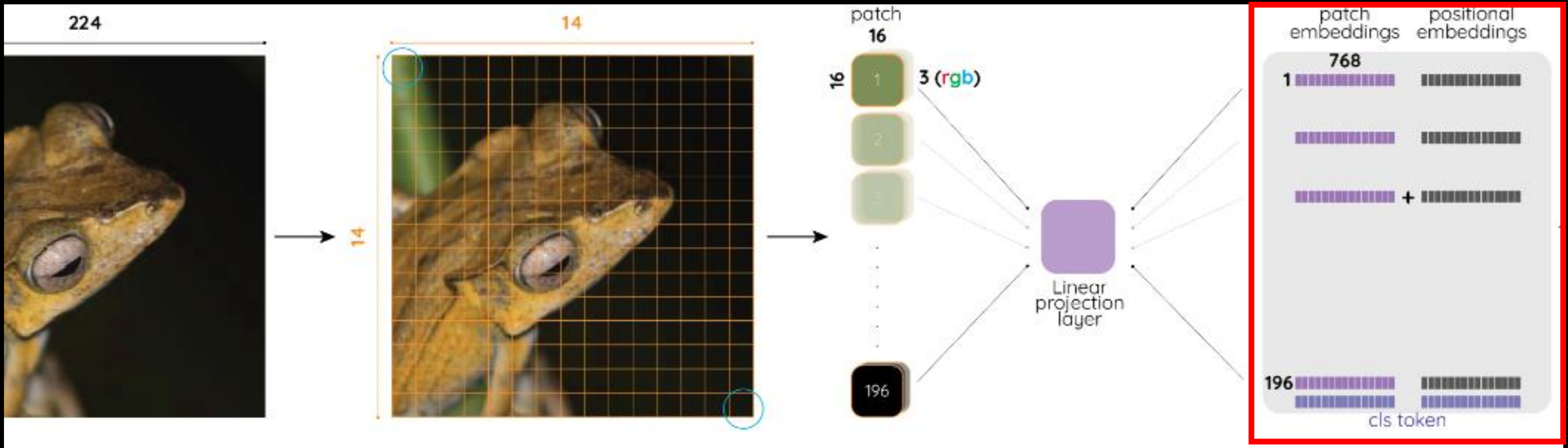




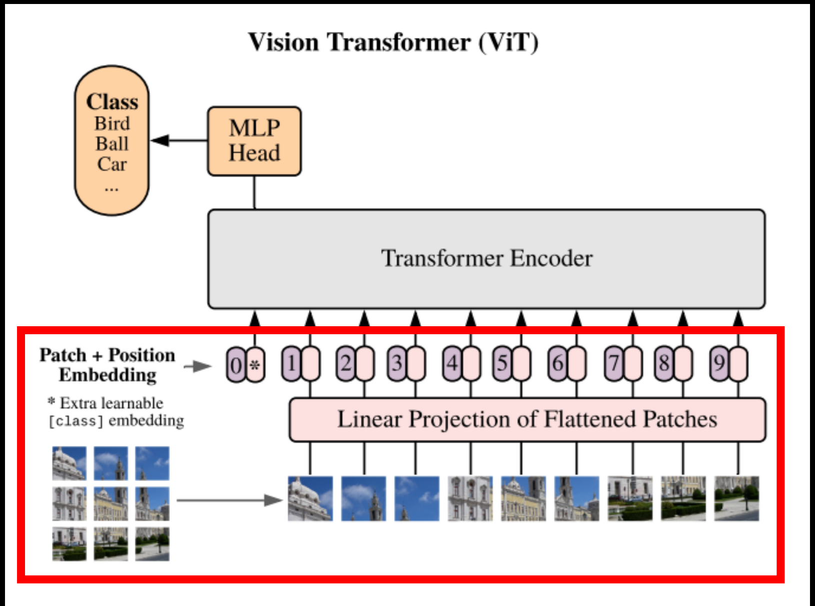
Positional Embedding

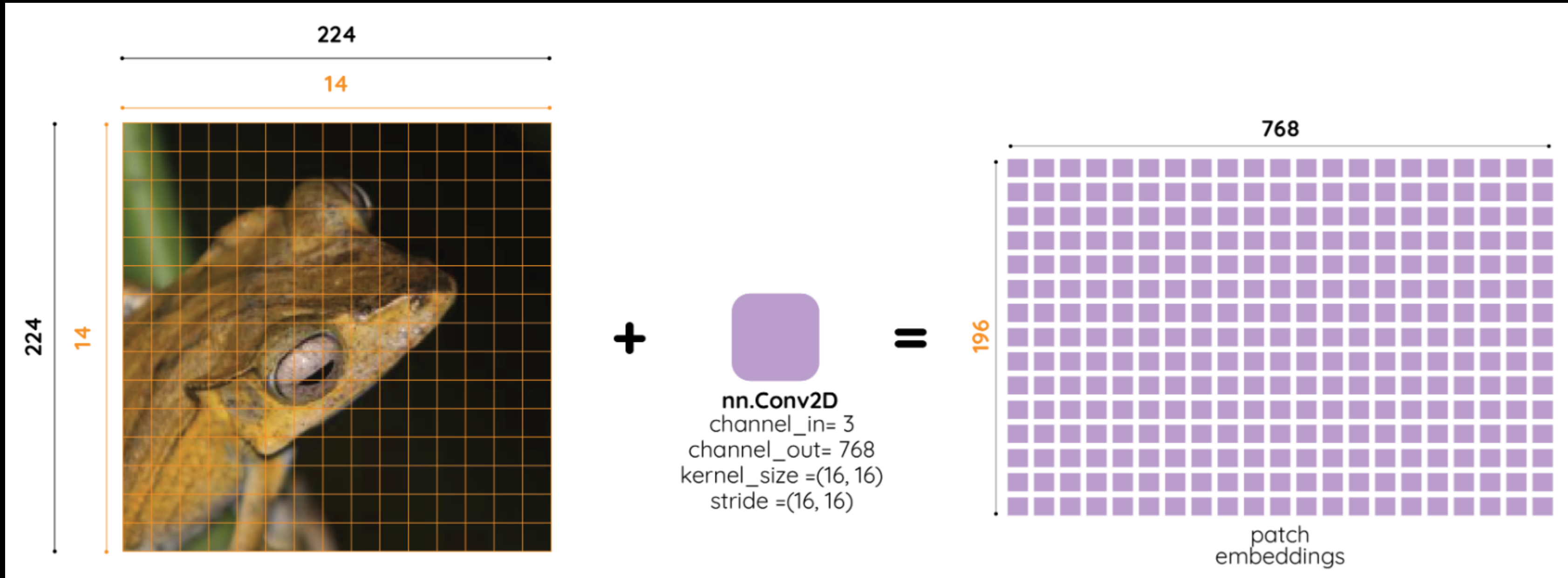
- Extra learnable class embedding with non-patch information
- 197 = 196(patch embeddings) + 1(class embedding)



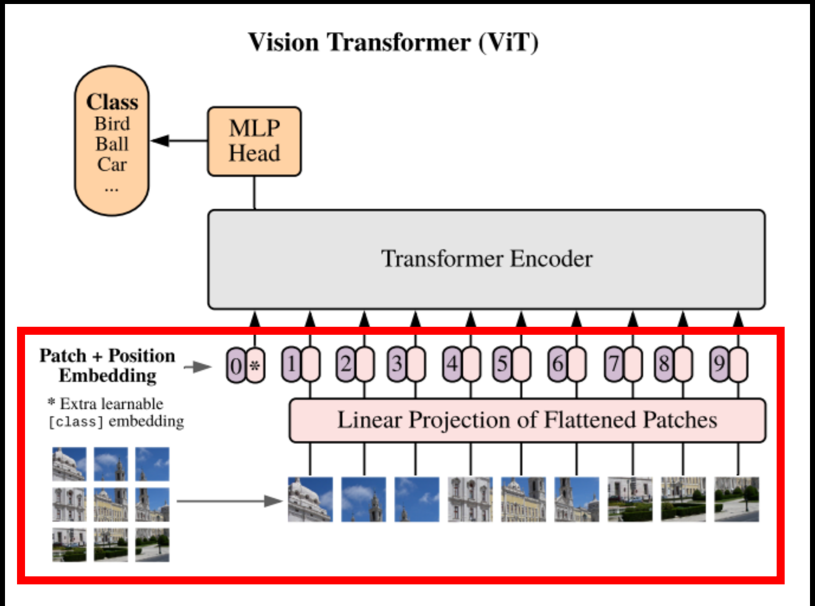


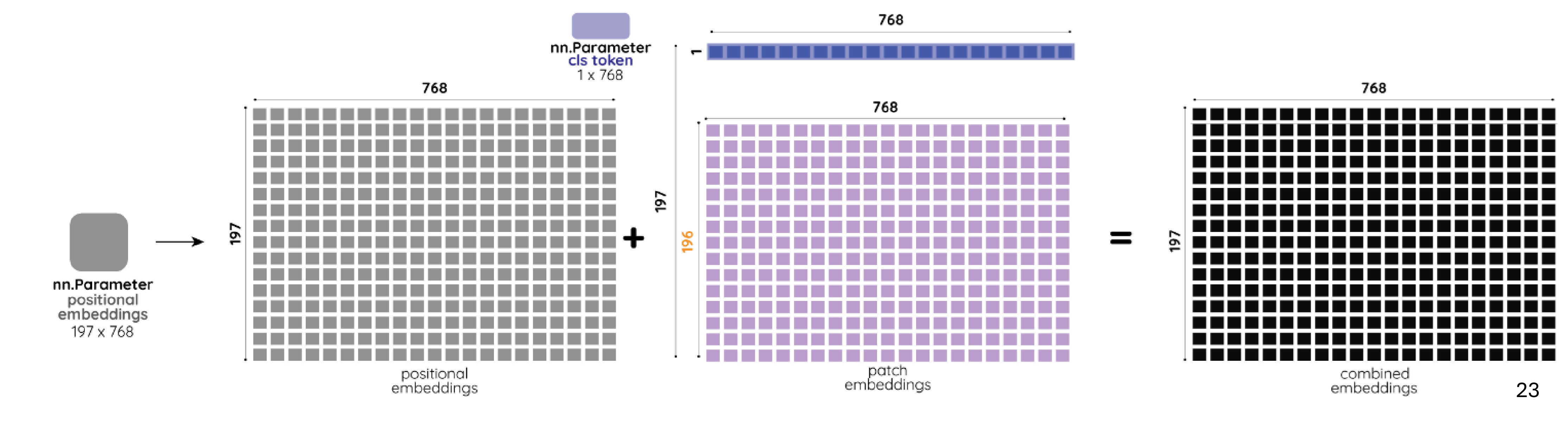
Summary So Far



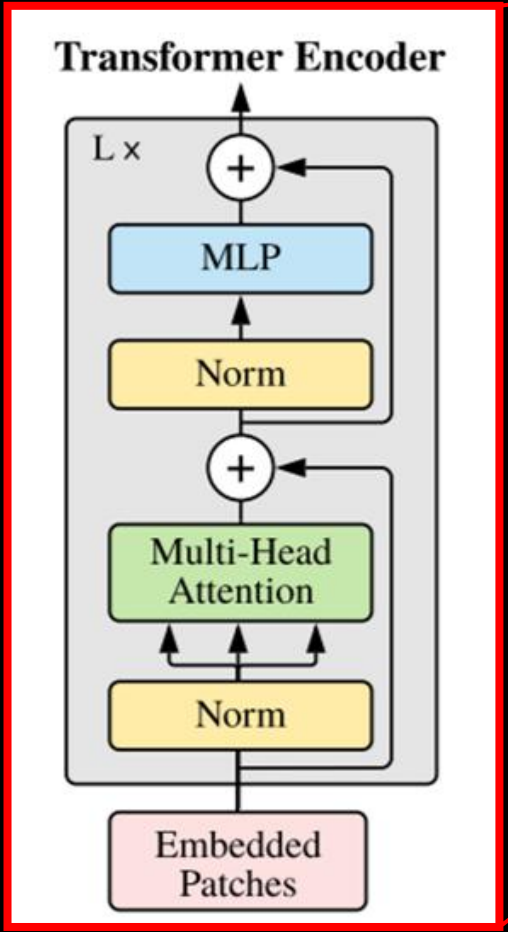


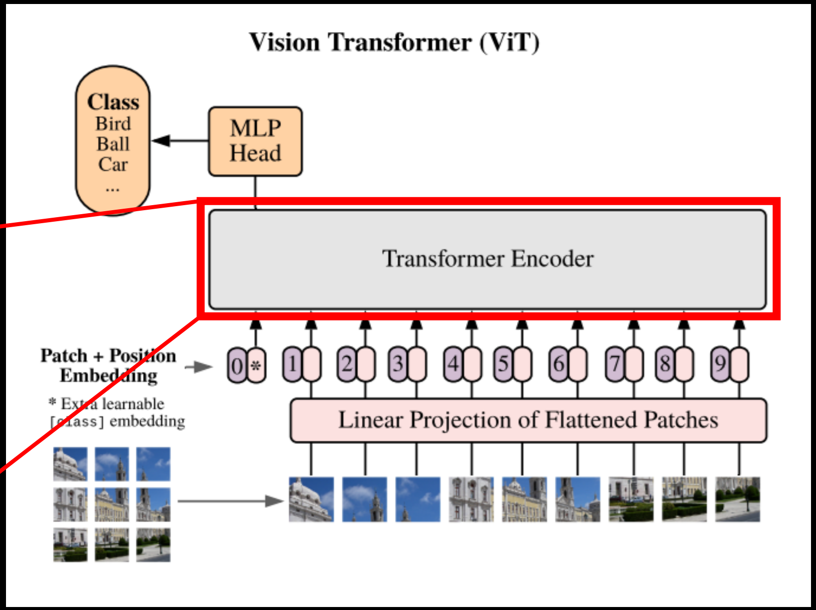
Summary So Far



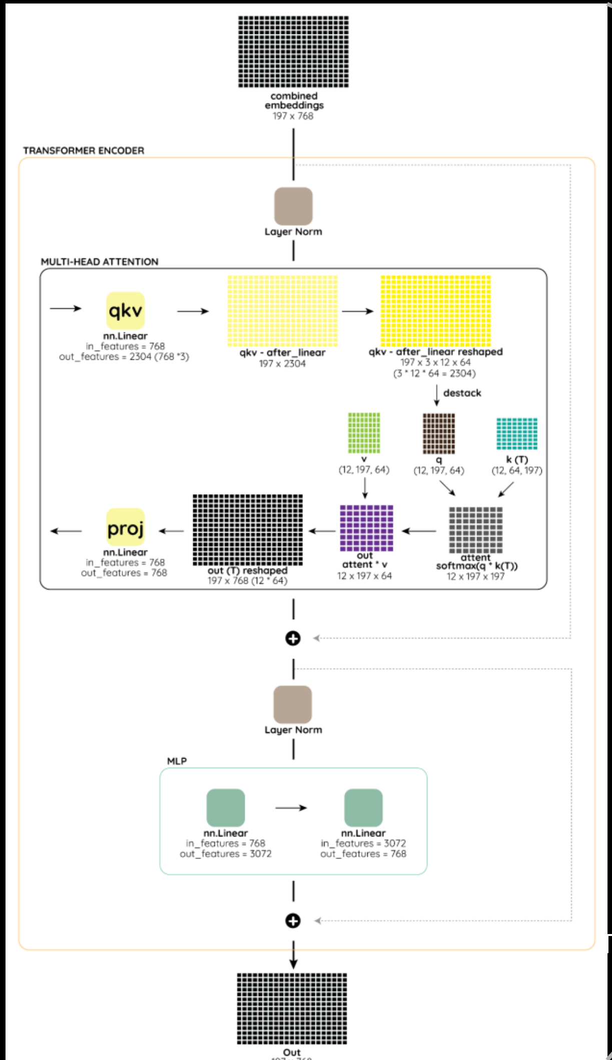


Transformer Encoder

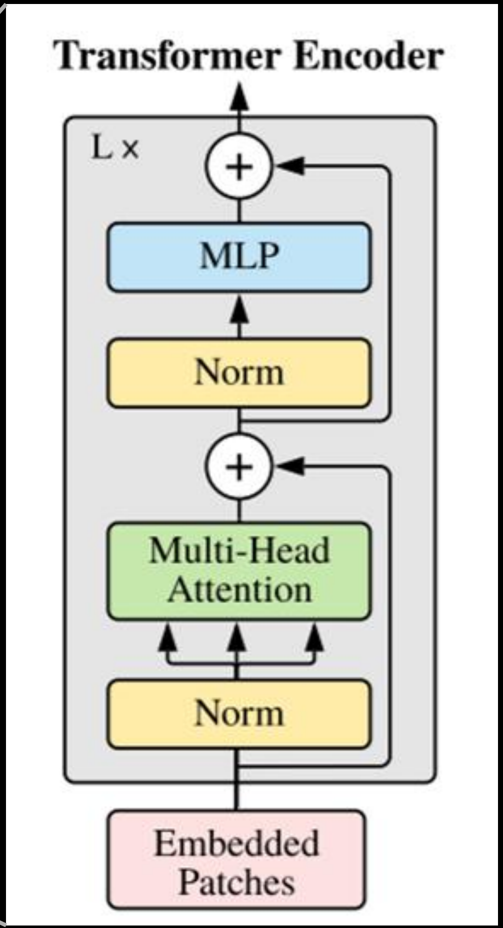




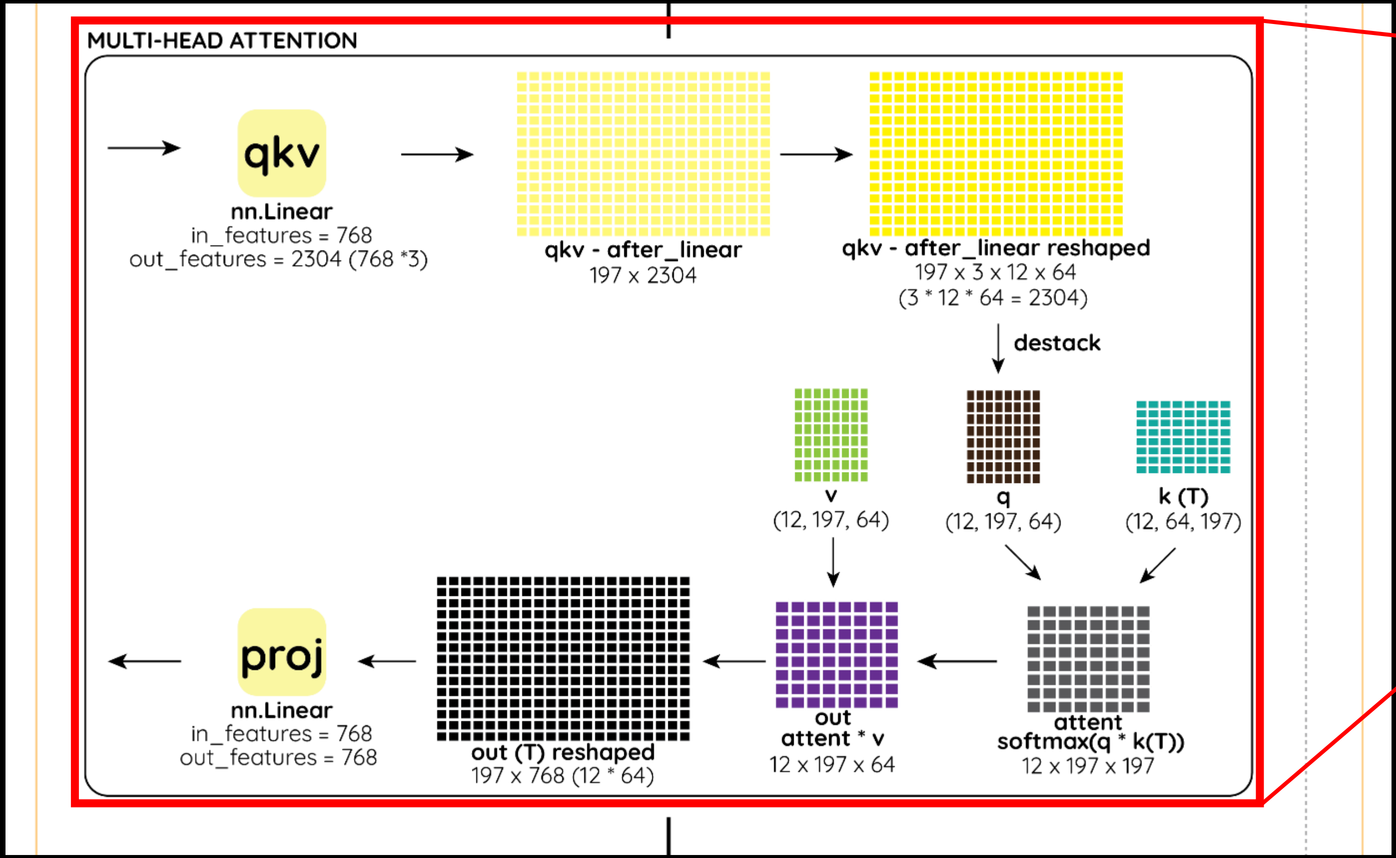
Transformer Encoder

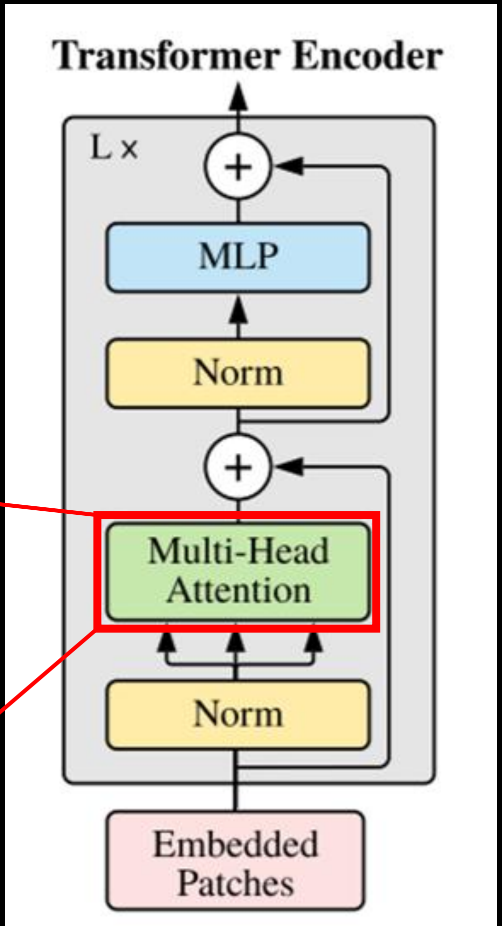


T.html



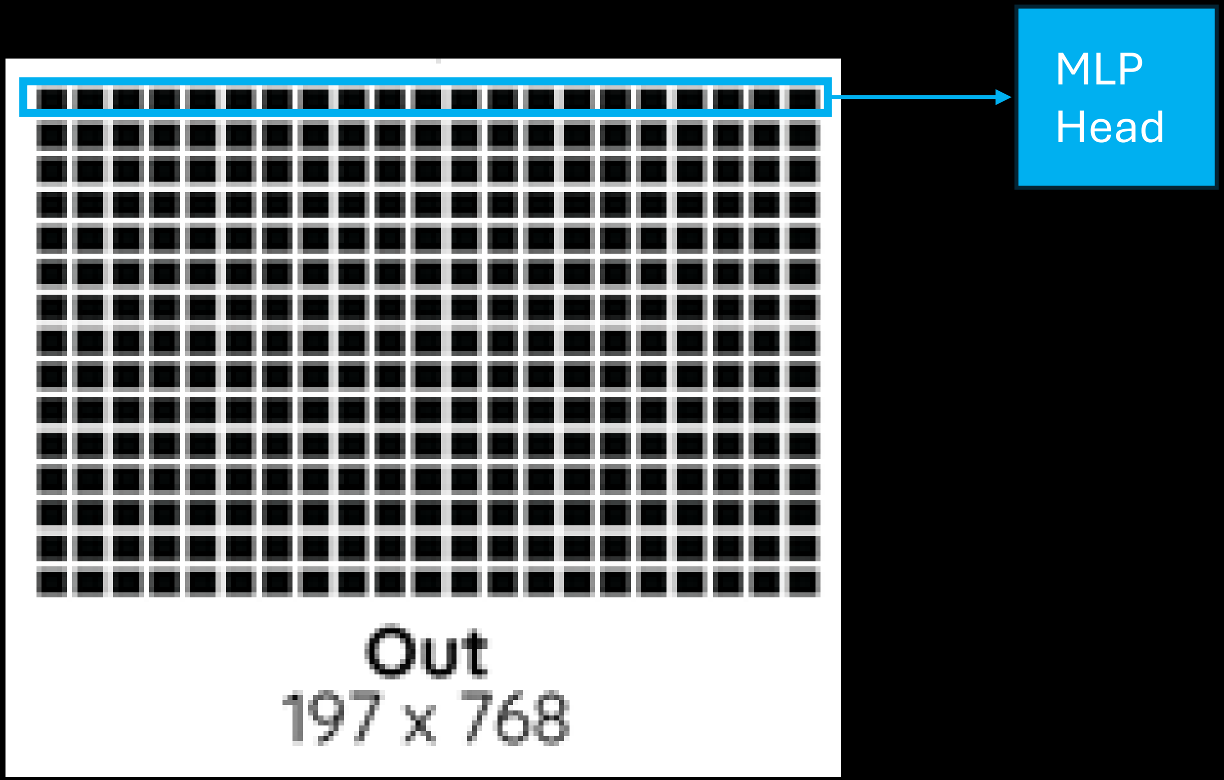
Transformer Encoder

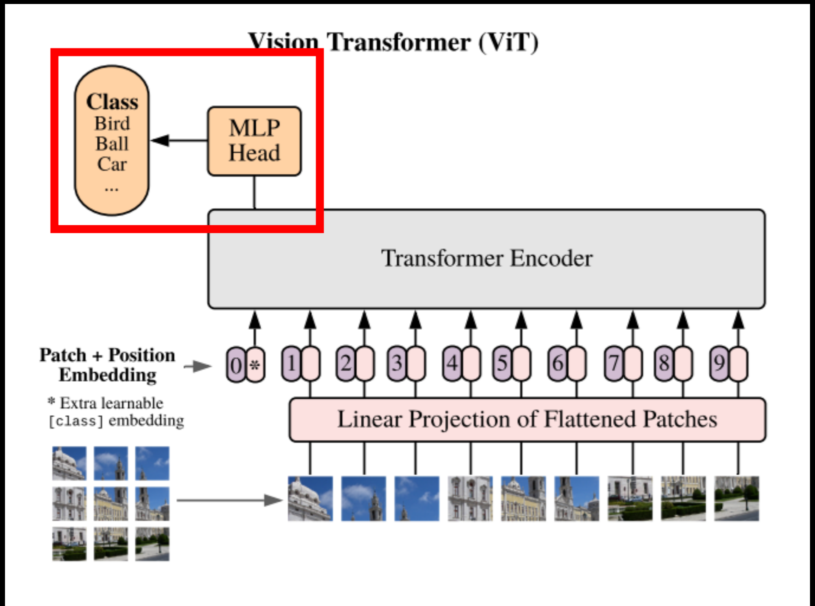




https://amaarora.github.io/posts/2021-01-18-ViT.html

Classification

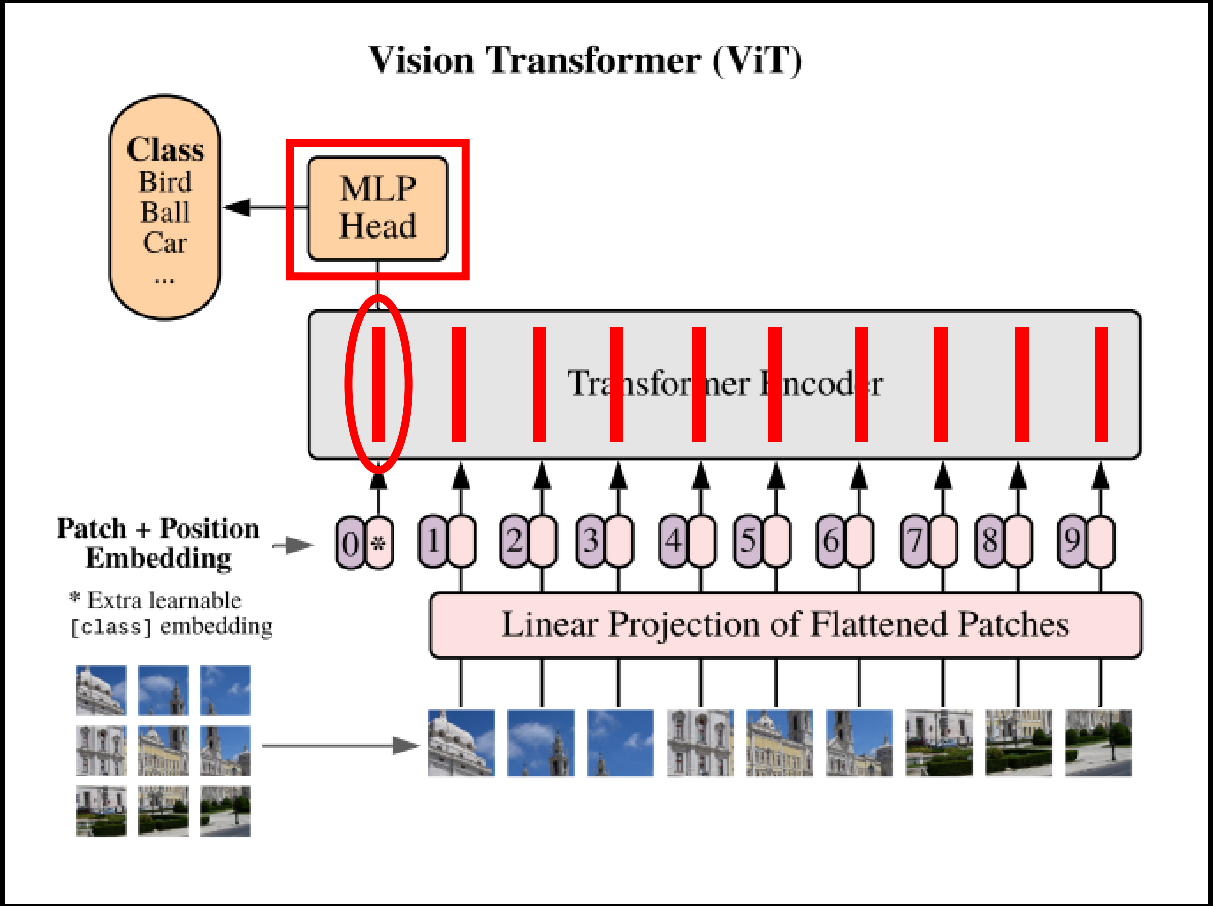




Only the [class token]'s output from the Transformer Encoder layer is used as the image representation for class prediction.

https://amaarora.github.io/posts/2021-01-18-ViT.html

Classification



https://amaarora.github.io/posts/2021-01-18-ViT.html

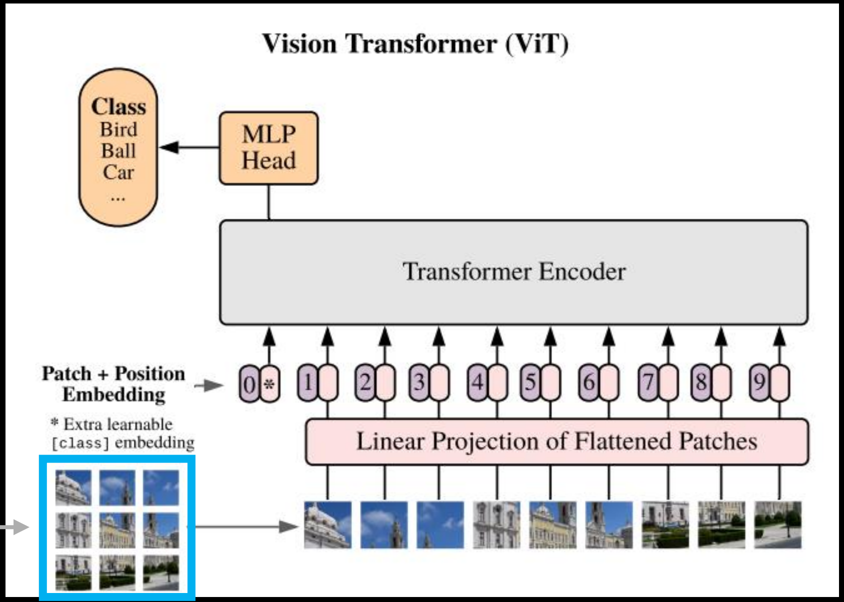
03

ViT Performance Analysis

Experiments

We evaluate the representation learning capabilities of

- ResNet (BiT)
- ViT
- Hybrid



Feature map of CNN

Model Variants

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.

For instance, ViT-L/16 means the "Large" variant with 16x16 input patch size.

https://arxiv.org/pdf/2010.11929

Datasets

- ImageNet: 1k classes with 1.3M images
- ImageNet-21k: 21k classes with 14M images
- JFT: 18k classes and 303M high-resolution images
- 19-task VTAB: 1000 training examples per task

Metrics

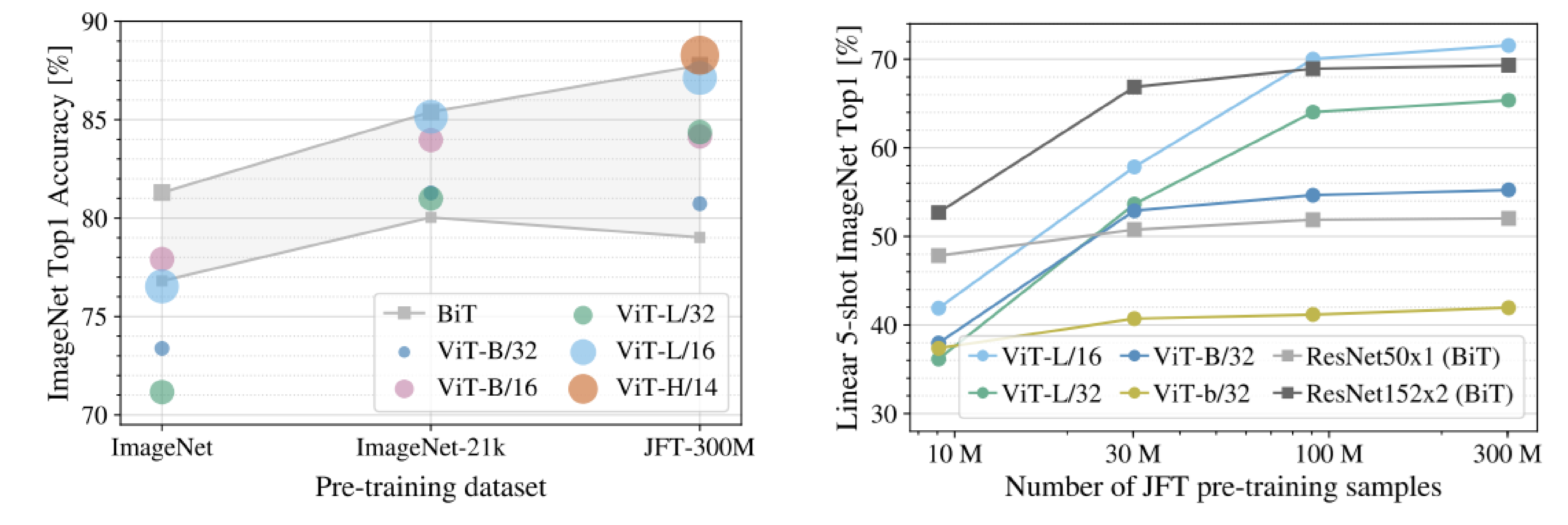
We report results on downstream datasets either through few-shot or finetuning accuracy.

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** $\pm$ 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

Pretraining Data Requirements

Good scalability: more data, higher performance

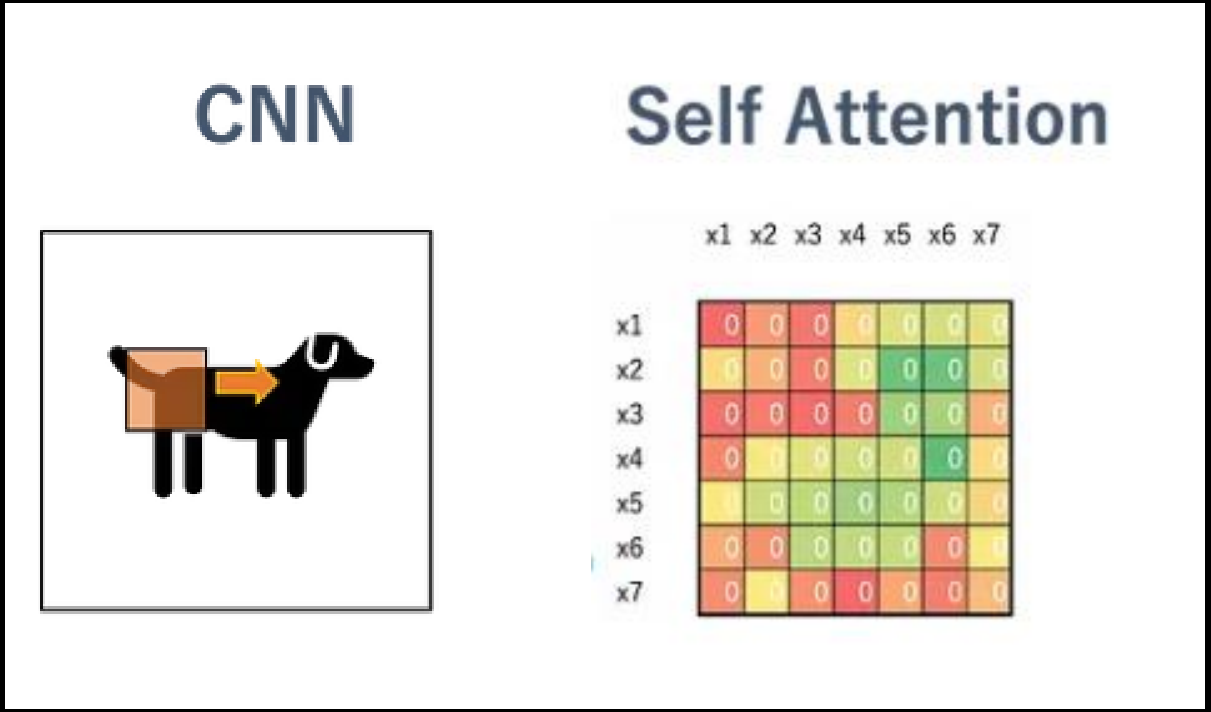


https://arxiv.org/pdf/2010.11929

Inductive Bias: CNN vs. ViT

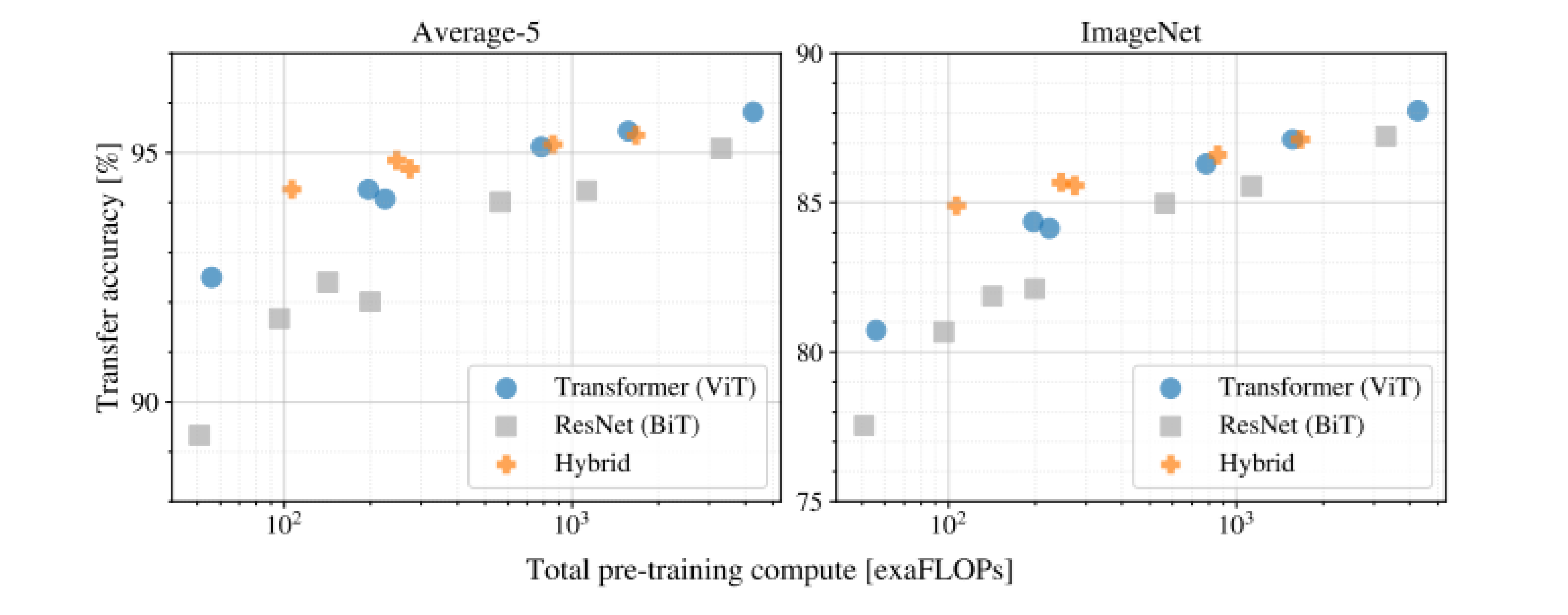
CNN, which has a strong inductive bias that the information is locally aggregated.

ViT, which has a relatively weak inductive bias because it only correlates all features.



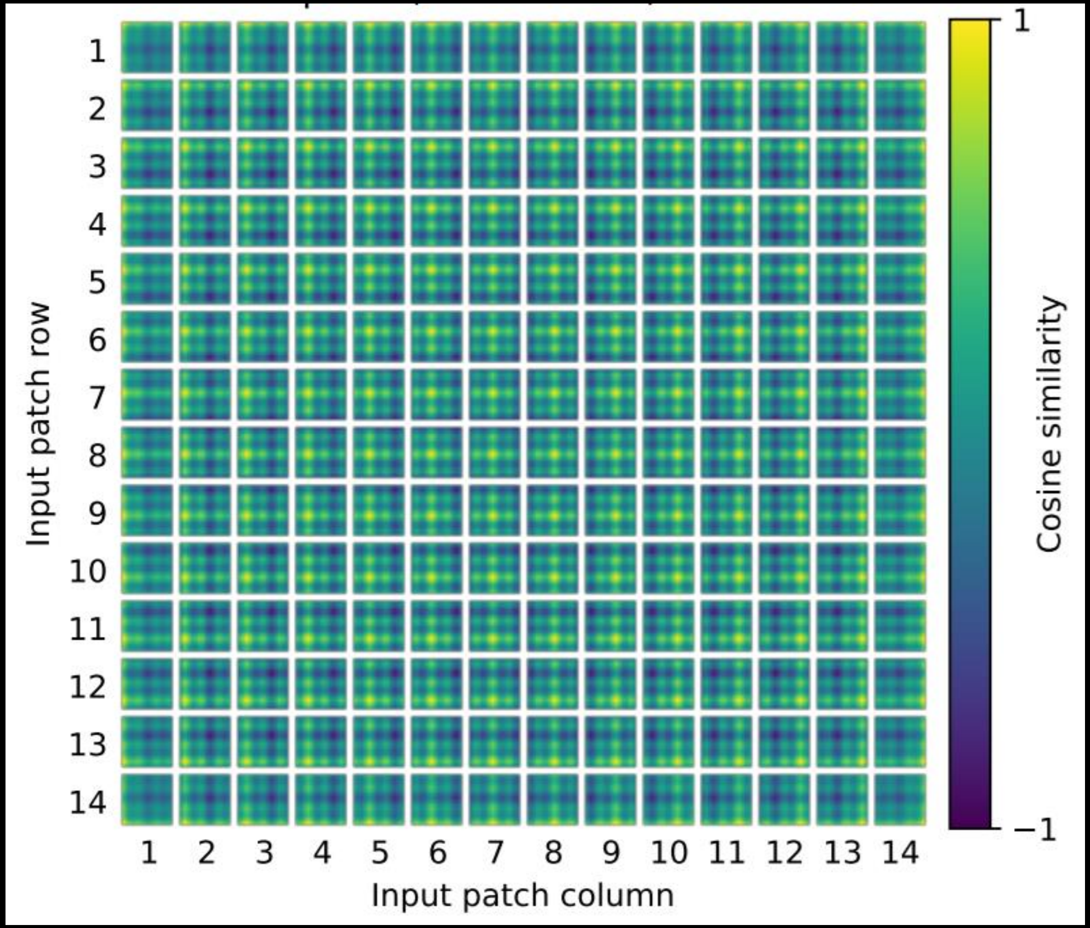
ViTs vs. ResNets: Efficiency Edge

ViTs dominate ResNets on the performance/compute tradeoff!



https://arxiv.org/pdf/2010.11929

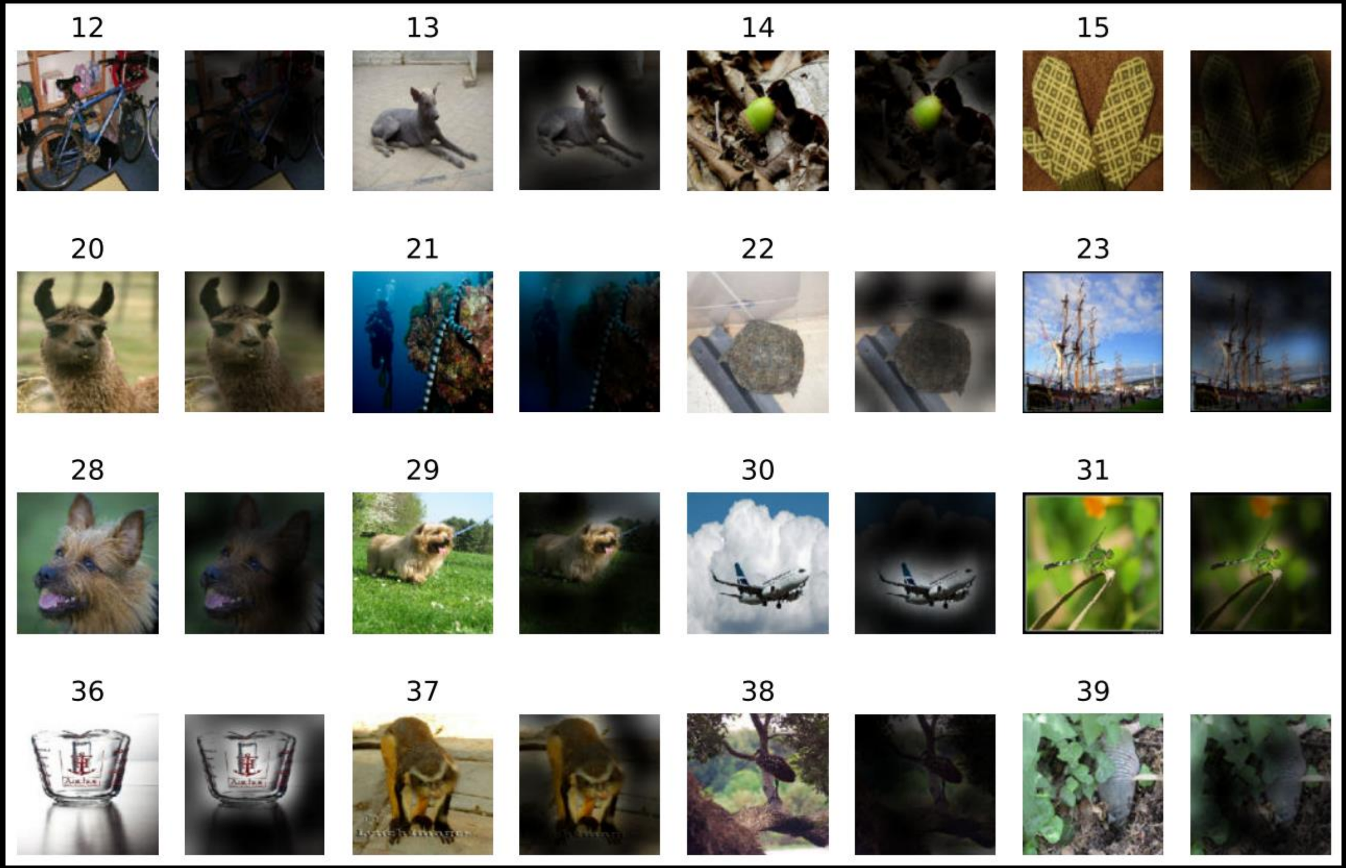
Position Embedding Similarity



https://arxiv.org/pdf/2010.11929

The learned positional embedding happen to distinguish their positions!

Attention map



Examples of attention from the output token to the input space.

https://arxiv.org/pdf/2010.11929

Significance of ViT

- Direct application of Transformers to image recognition.
- ViT matches or exceeds the SOTA of many image classification datasets and is cost-effective to pre-train.

- Many challenges remain…

Thank You!
Any Questions?