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

Presentation by: Amir Karami Fini

Table Of Contents

The Idea

Research Background

ViT

Explain the Architecture

Training

Fine Tuning

Results

The Idea

Section 1

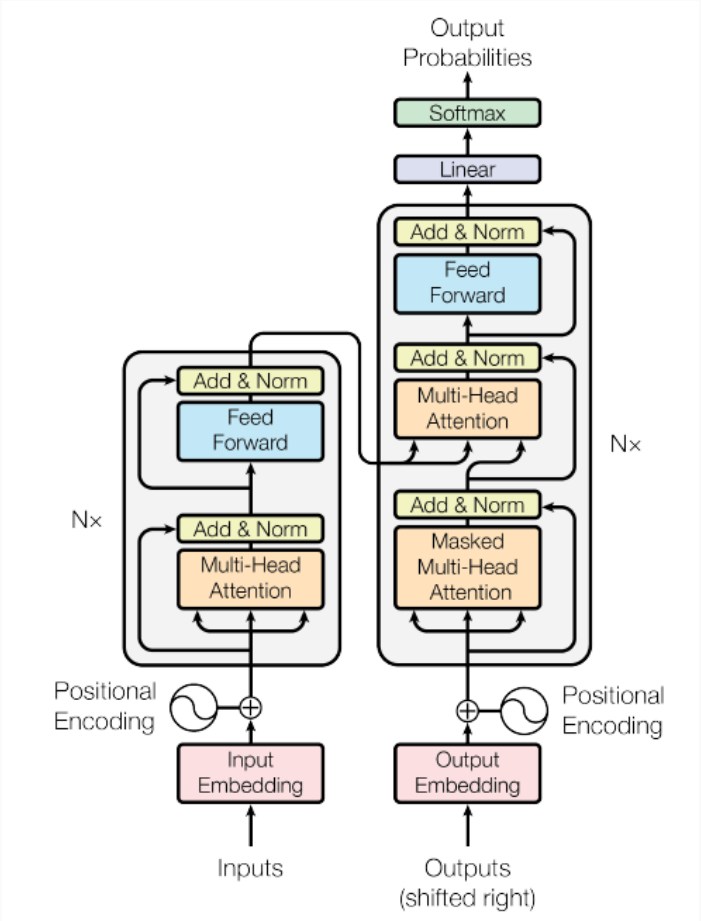
The Idea

- Transformers have become the model of choice in natural language processing (NLP)
- Its applications to computer vision remain limited
- Attention use in CNNs, but overall structure is same

- Reliance on CNNs is not necessary

Transformer

- Encoder/Decoder
- Self attention
- Positional Encoding

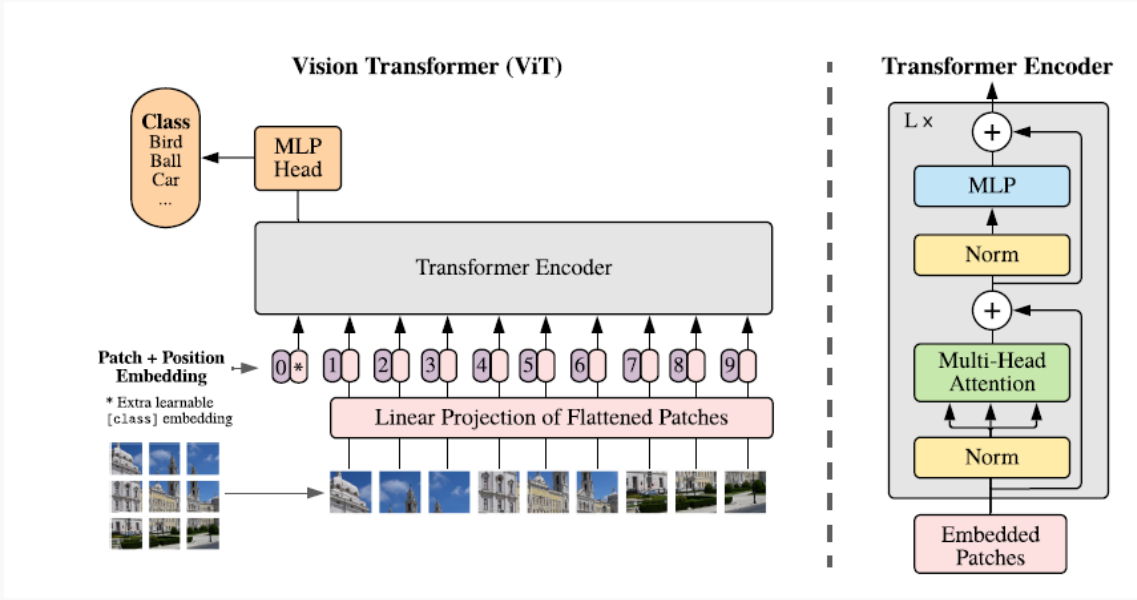


ViT (Vision Transformer)

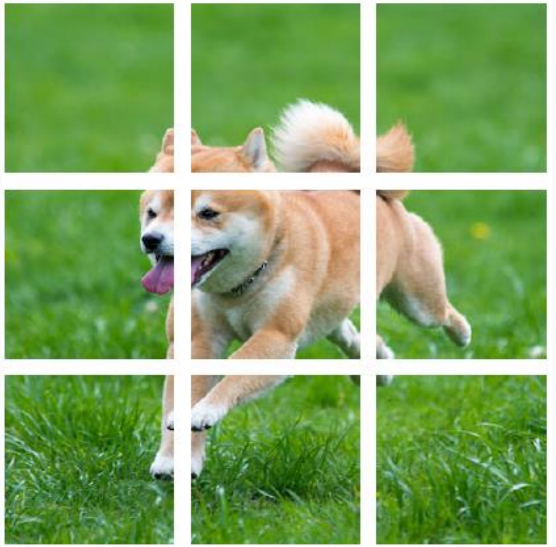
Section 2

ViT

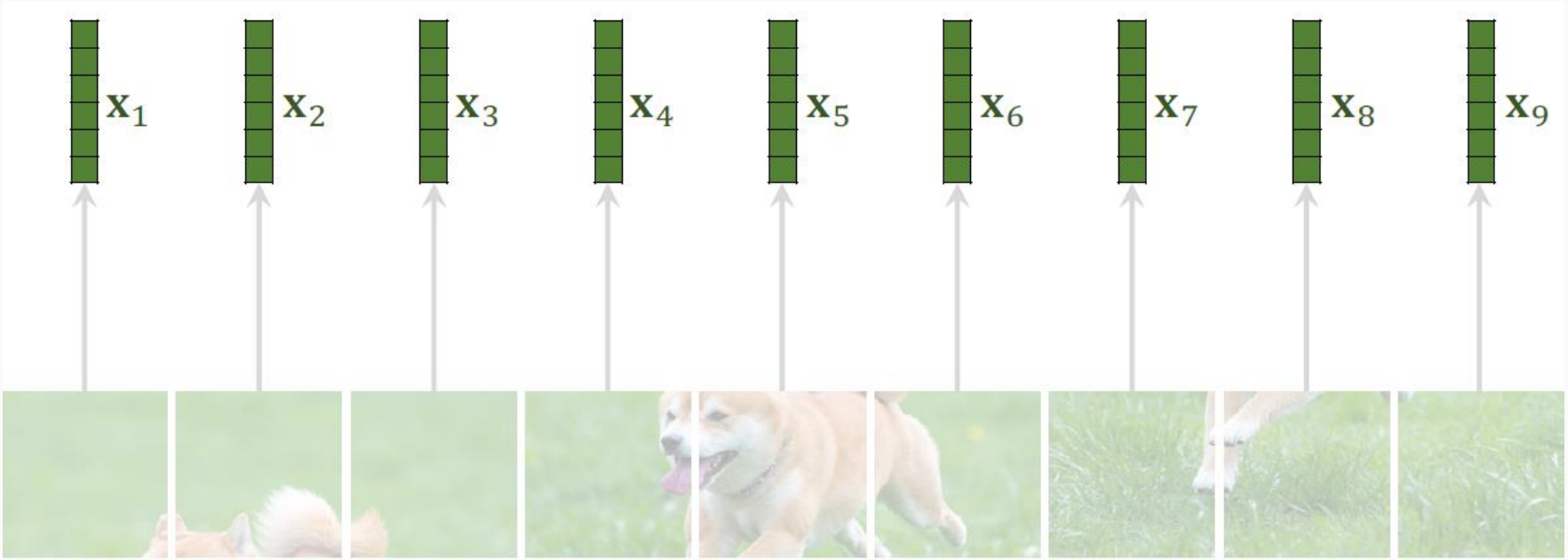
- Patches are Words!
- Flatten
- Cls token



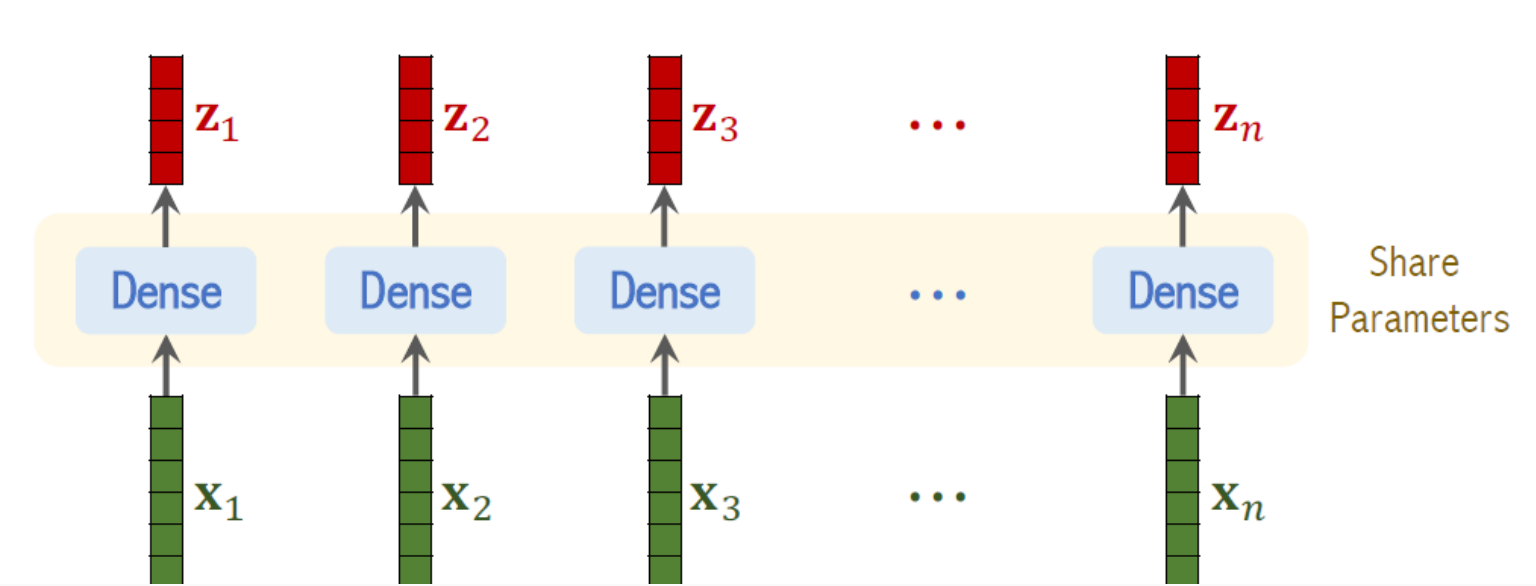
Patches are Words!



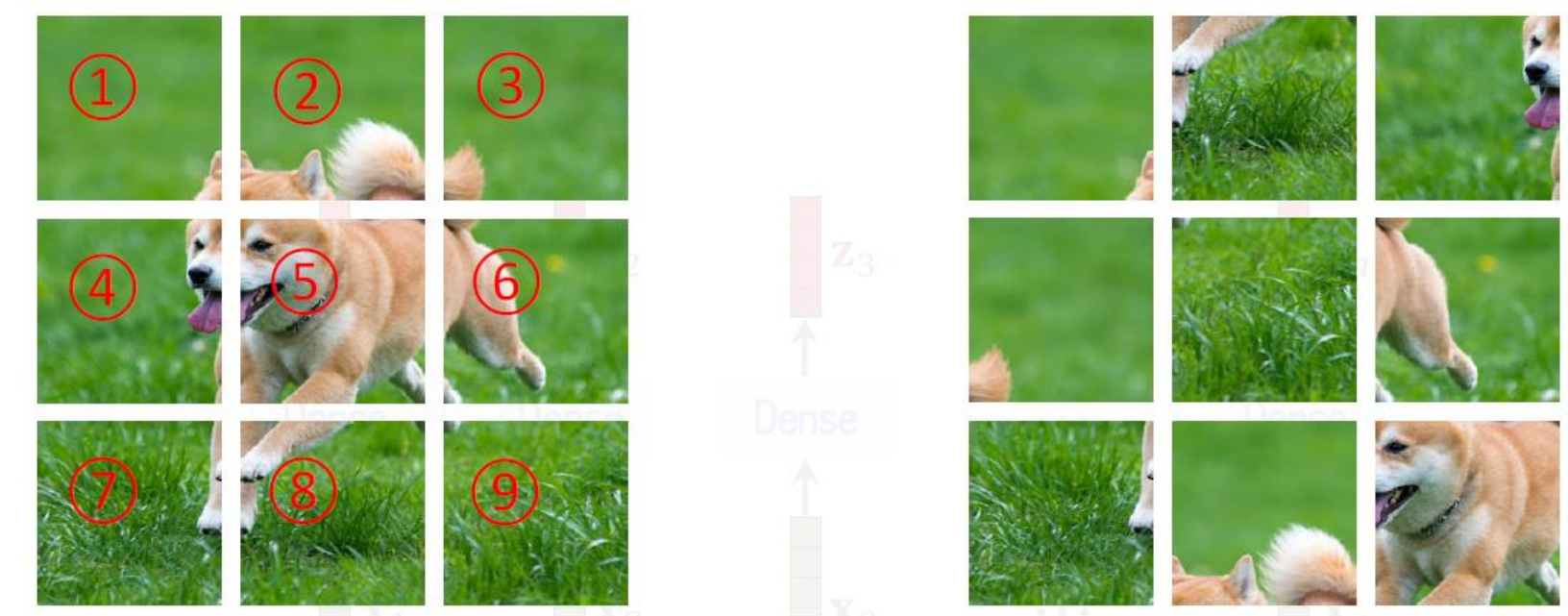
Flatten



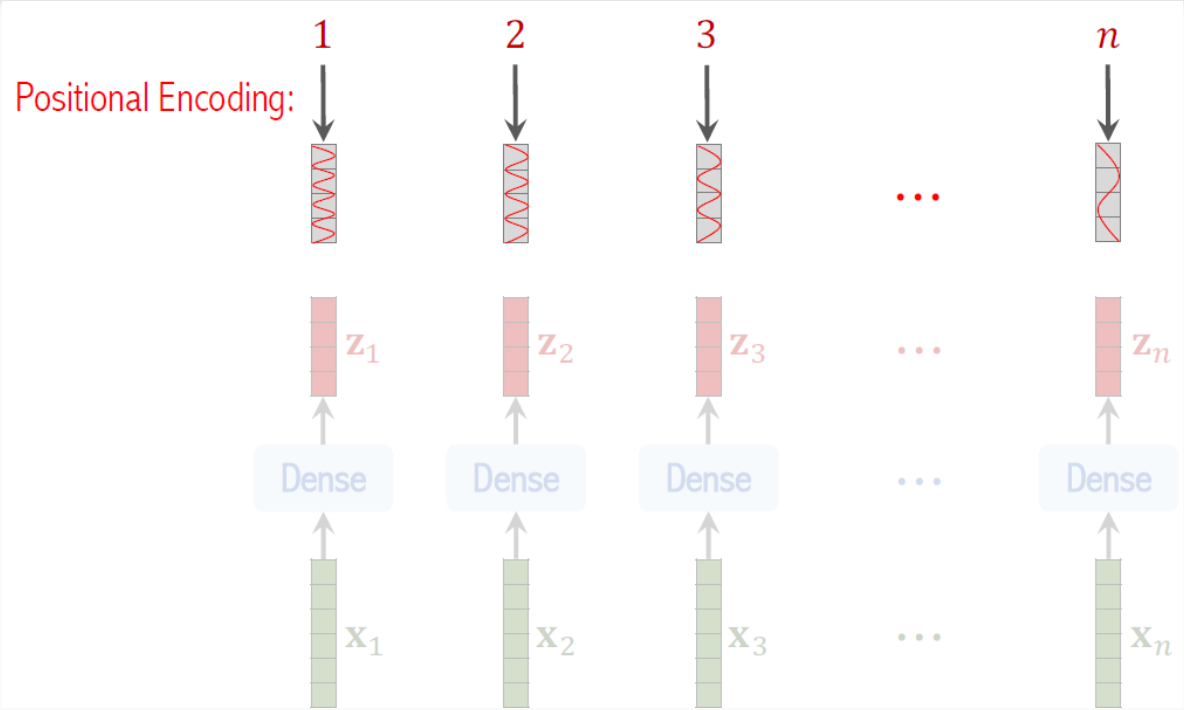
Flatten


z_1
z_2
z_3
z_n
Dense
Dense
Dense
Dense
Share Parameters
x_1
x_2
x_3
x_n

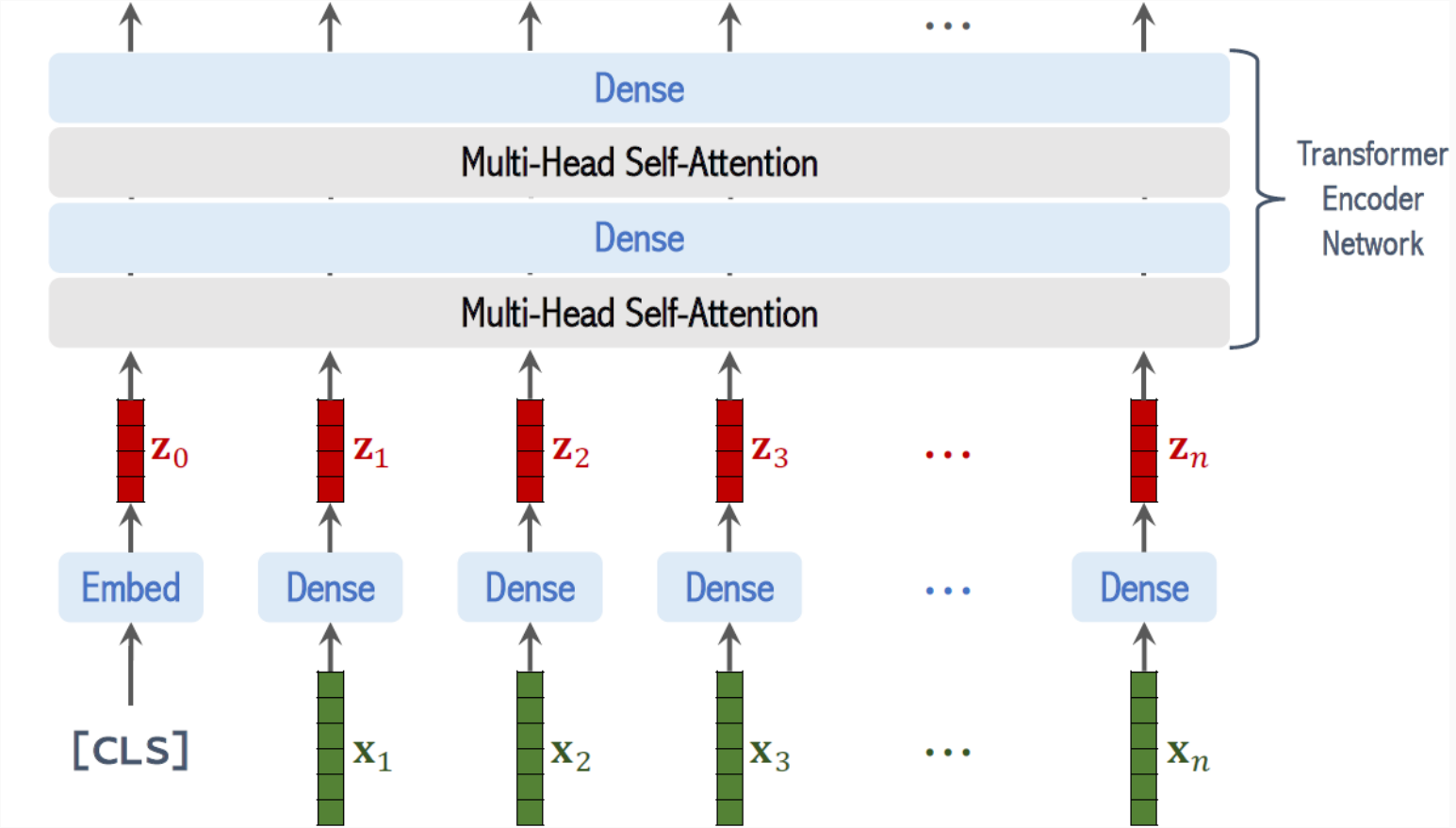
Positional Encoding



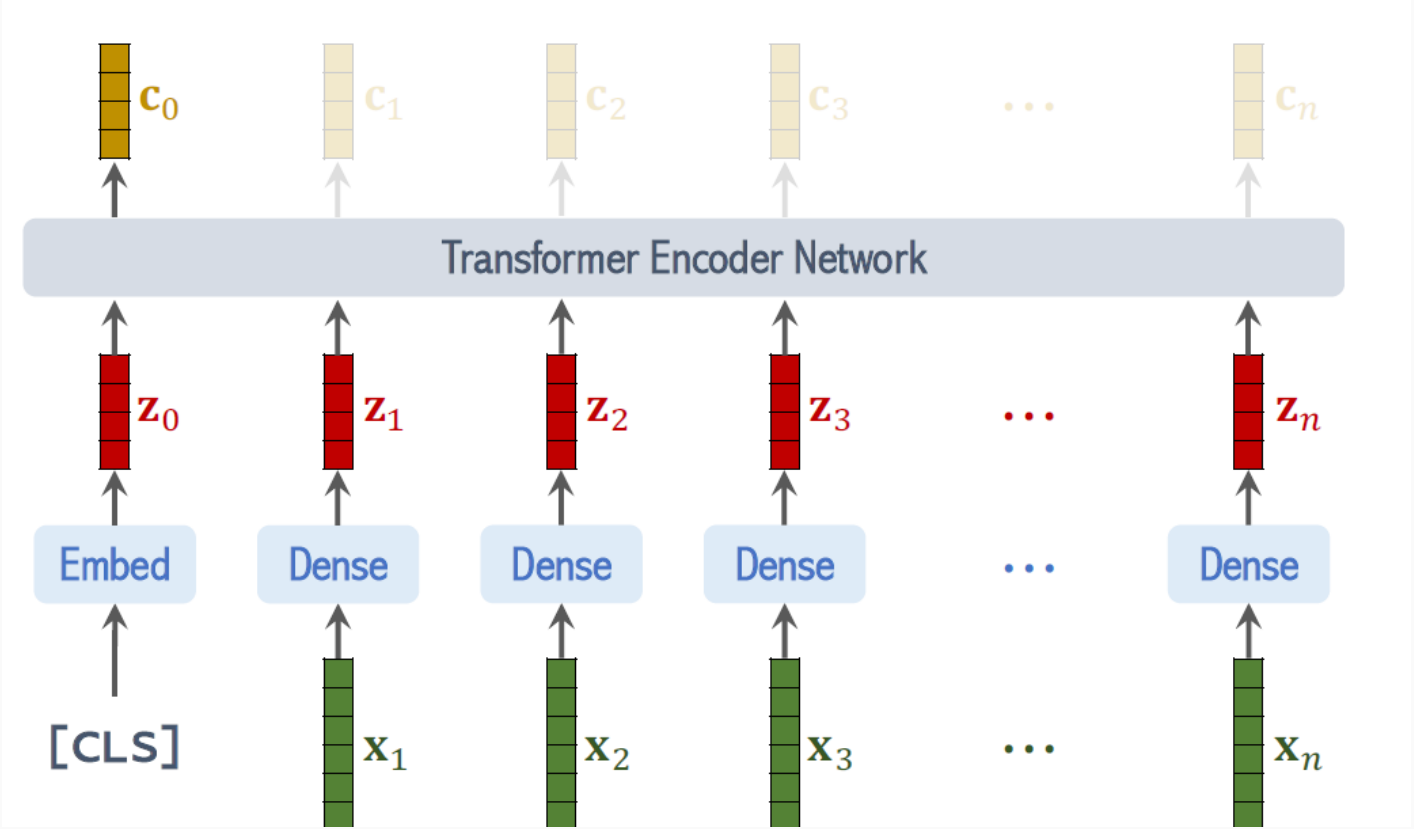
Positional Encoding



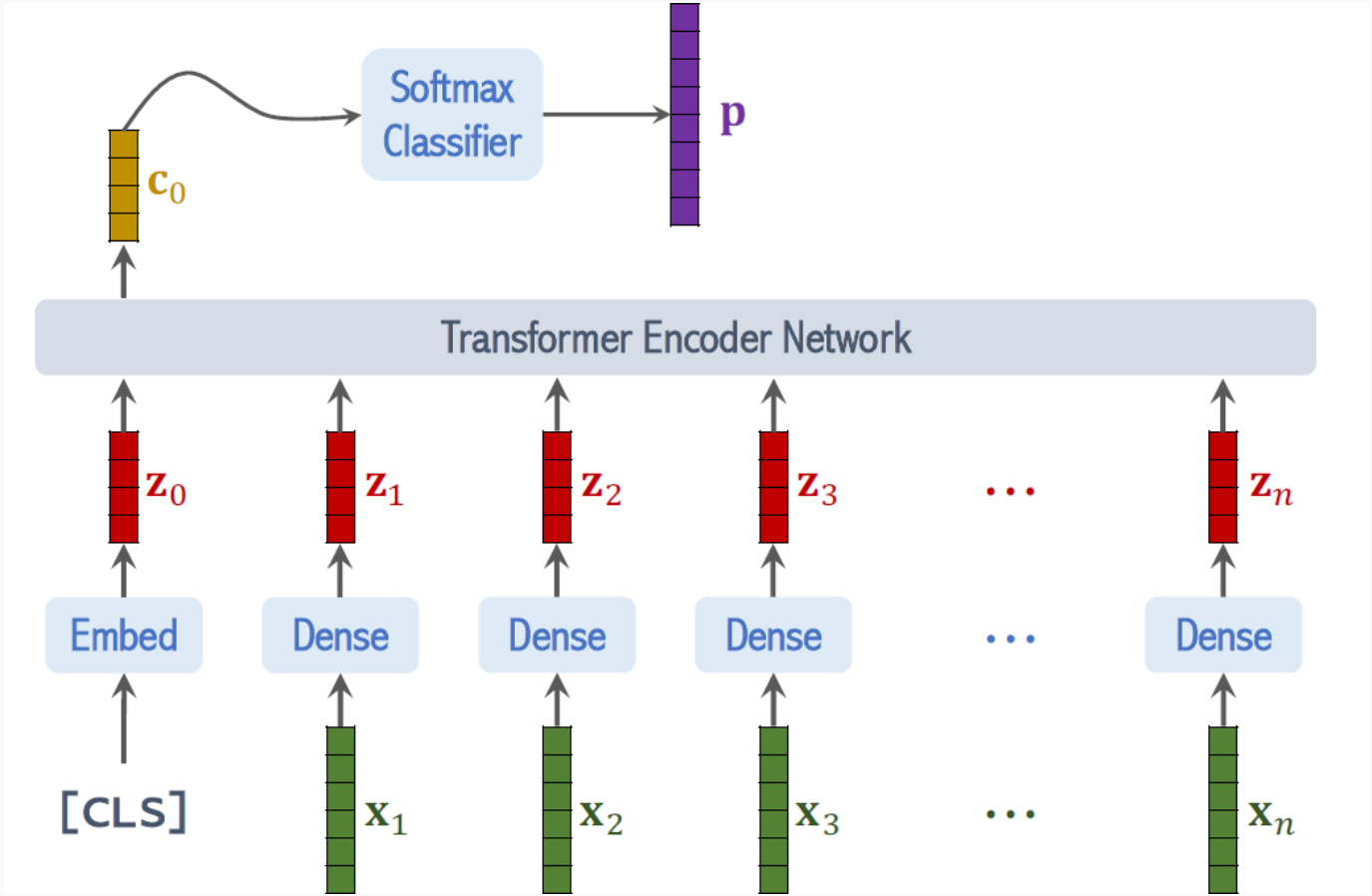
The Architecture



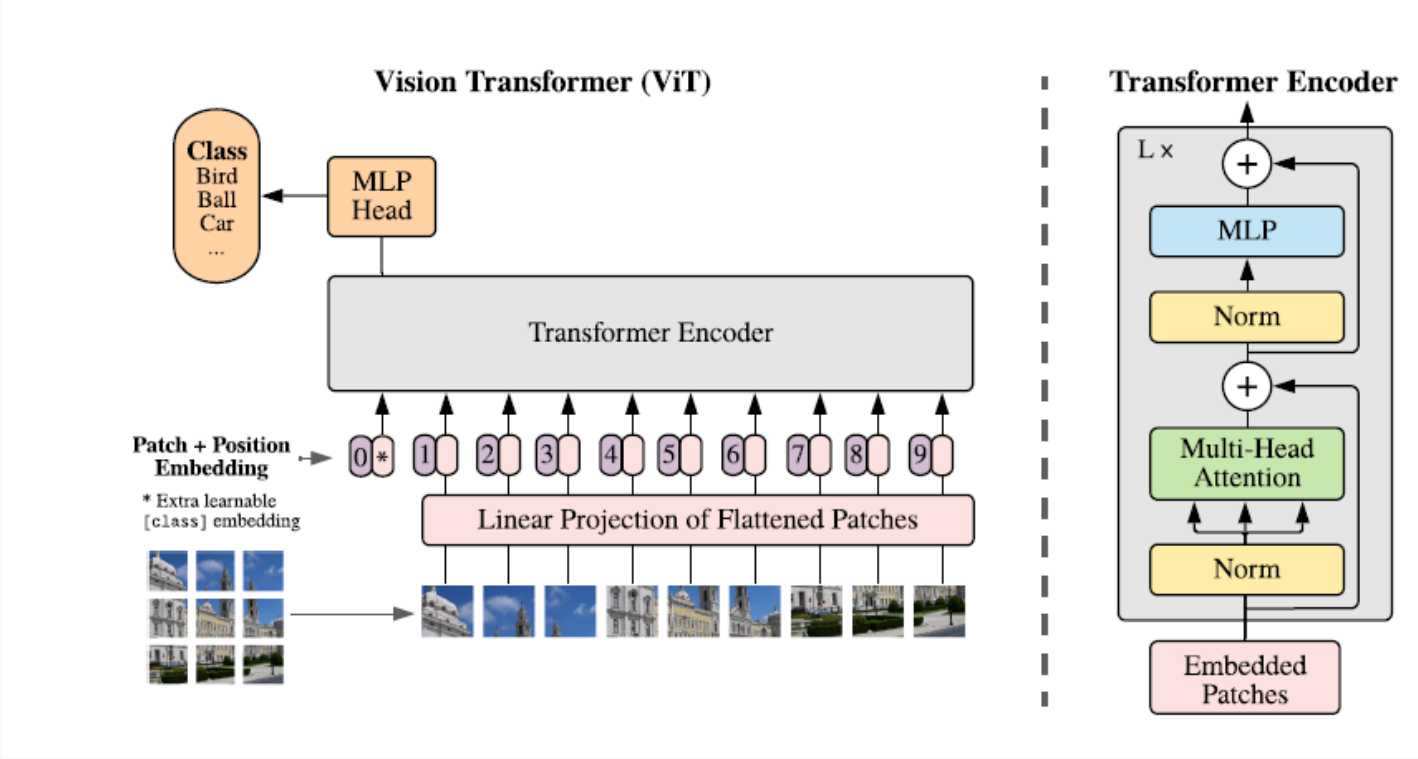
The Architecture



The Architecture



The Architecture



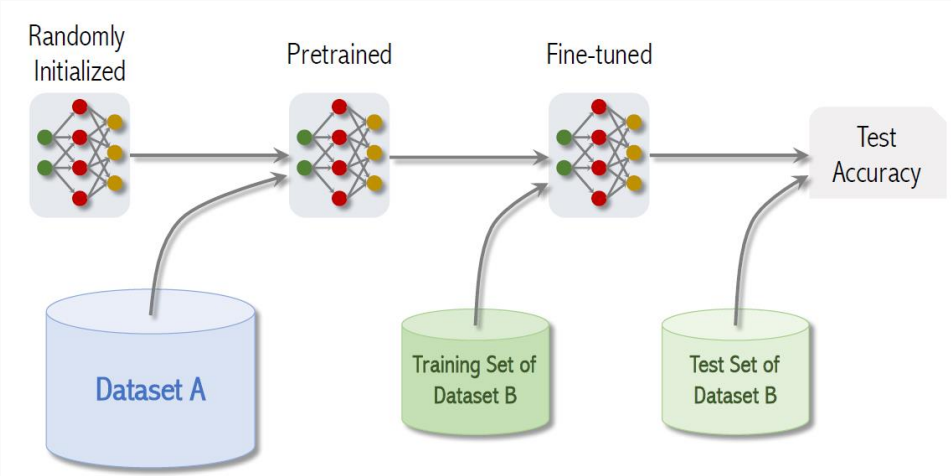
Training

Section 3

Training

- ViT needs large amount of data
- So, lets Fine tune our models!

	# of Images	# of Classes
ImageNet (Small)	1.3 Million	1 Thousand
ImageNet-21K (Medium)	14 Million	21 Thousand
JFT (Big)	300 Million	18 Thousand



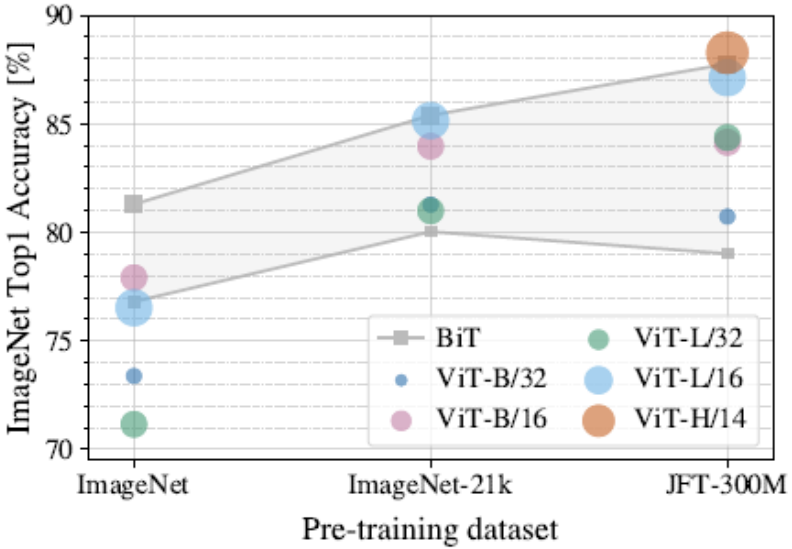
Results

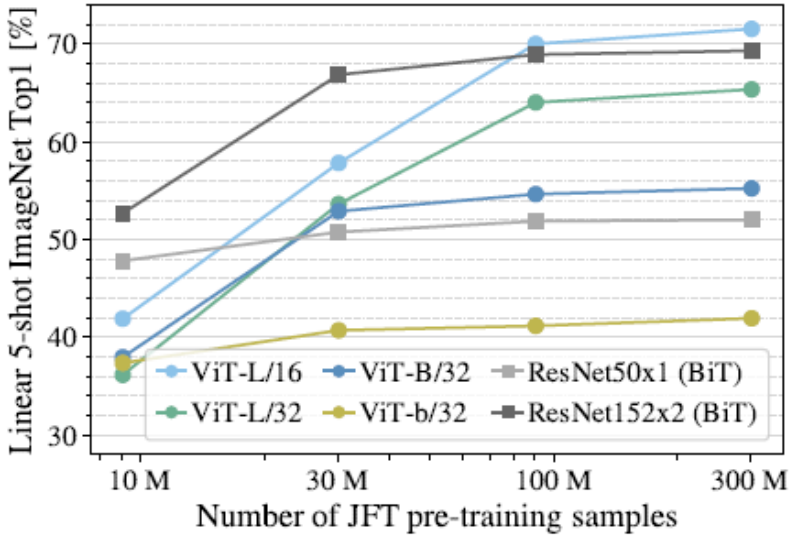
Section 4

Results

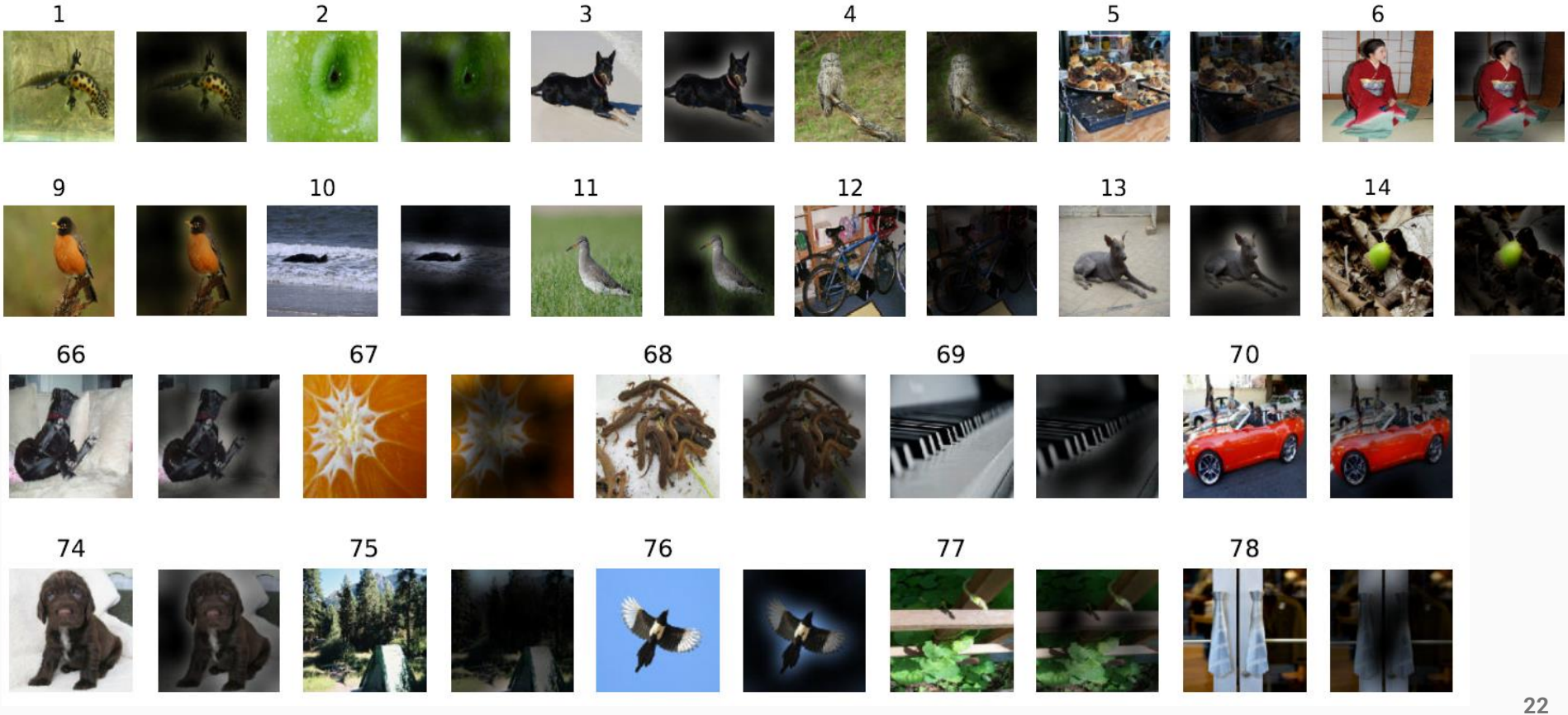
	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 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

Results


ImageNet Top1 Accuracy [%]
90
85
80
75
70
ImageNet
ImageNet-21k
JFT-300M
Pre-training dataset
BiT
ViT-B/32
ViT-B/16
ViT-L/32
ViT-L/16
ViT-H/14


Linear 5-shot ImageNet Top1 [%]
70
60
50
40
30
10 M
30 M
100 M
300 M
Number of JFT pre-training samples
ViT-L/16
ViT-B/32
ResNet50x1 (BiT)
ViT-L/32
ViT-b/32
ResNet152x2 (BiT)

Results



References

- A. Vaswani et al., “Attention Is All You Need,” Jun. 2017
- A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” Oct. 2020
- https://github.com/wangshusen/DeepLearning
- https://github.com/NielsRogge/Transformers-Tutorials/tree/master/VisionTransformer

Thanks for joining us

src: unsplash.com