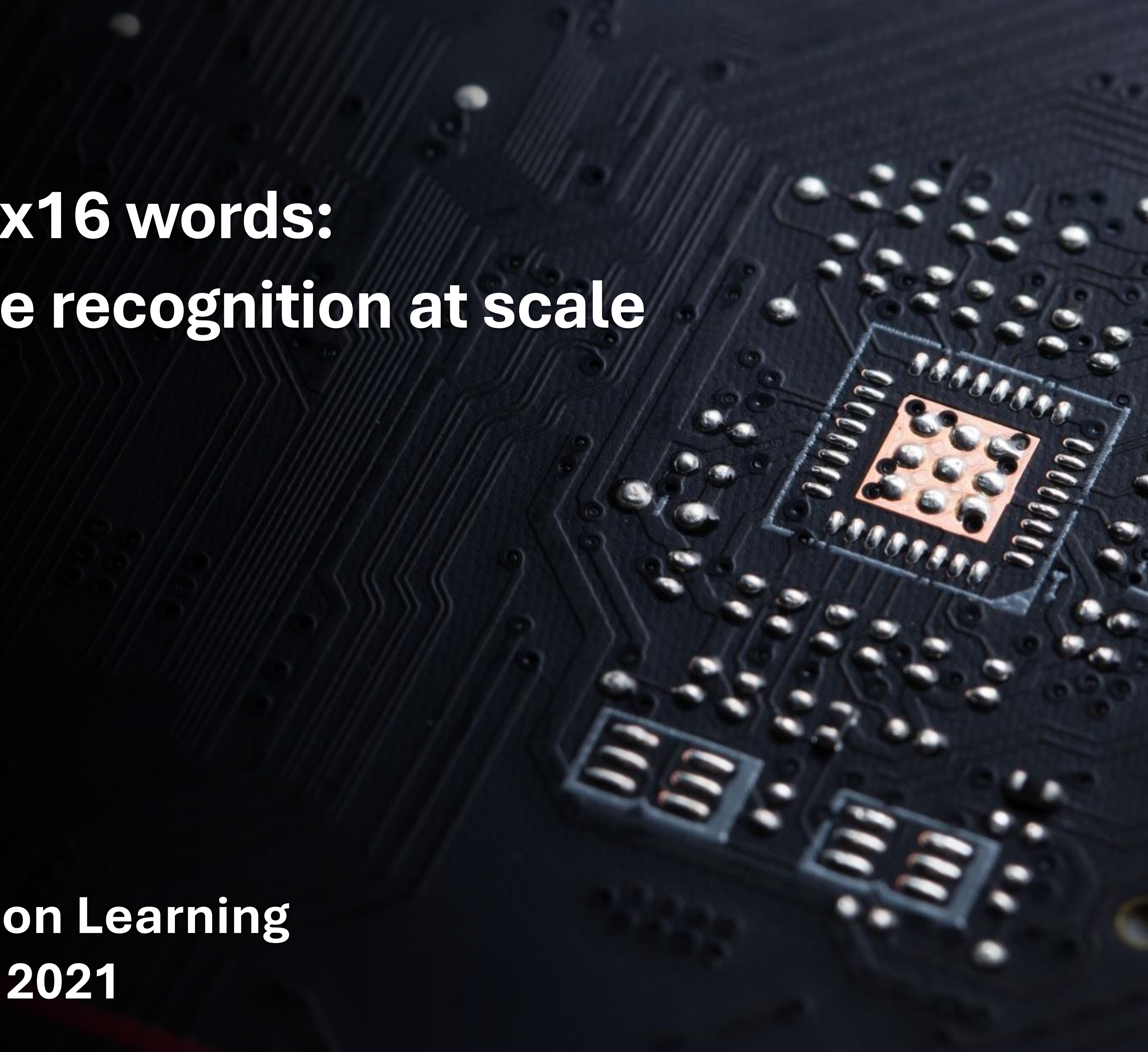
An Image is worth 16x16 words: Transformer for image recognition at scale

By Alexey Dosovitskiy



At

International Conference on Learning Representations (ICLR) in 2021

Outline

1. Motivation of the paper
2. Method
3. Experiment and Result
4. Conclusion
5. Discussion

Motivation

1 Limitations of CNNs

- Computationally intensive

2 Success of transformer in NLP tasks

- Breakthrough performance in NLP
- Motivate to apply transformer to solve visual problem

3 Explore the use of transformer in visual tasks

- Pure transformer has not been applied to visual tasks before
- Only self-attention mechanism has been integrated

Method

1 **Construct a Vision Transformer model (ViT)**

- By passing patches of image as an input for the transformer

2 **Build other models for comparison on image classification tasks**

Modified CNN (BiT)

- Built on ***ResNet*** architecture
- ***Baseline model*** for comparison

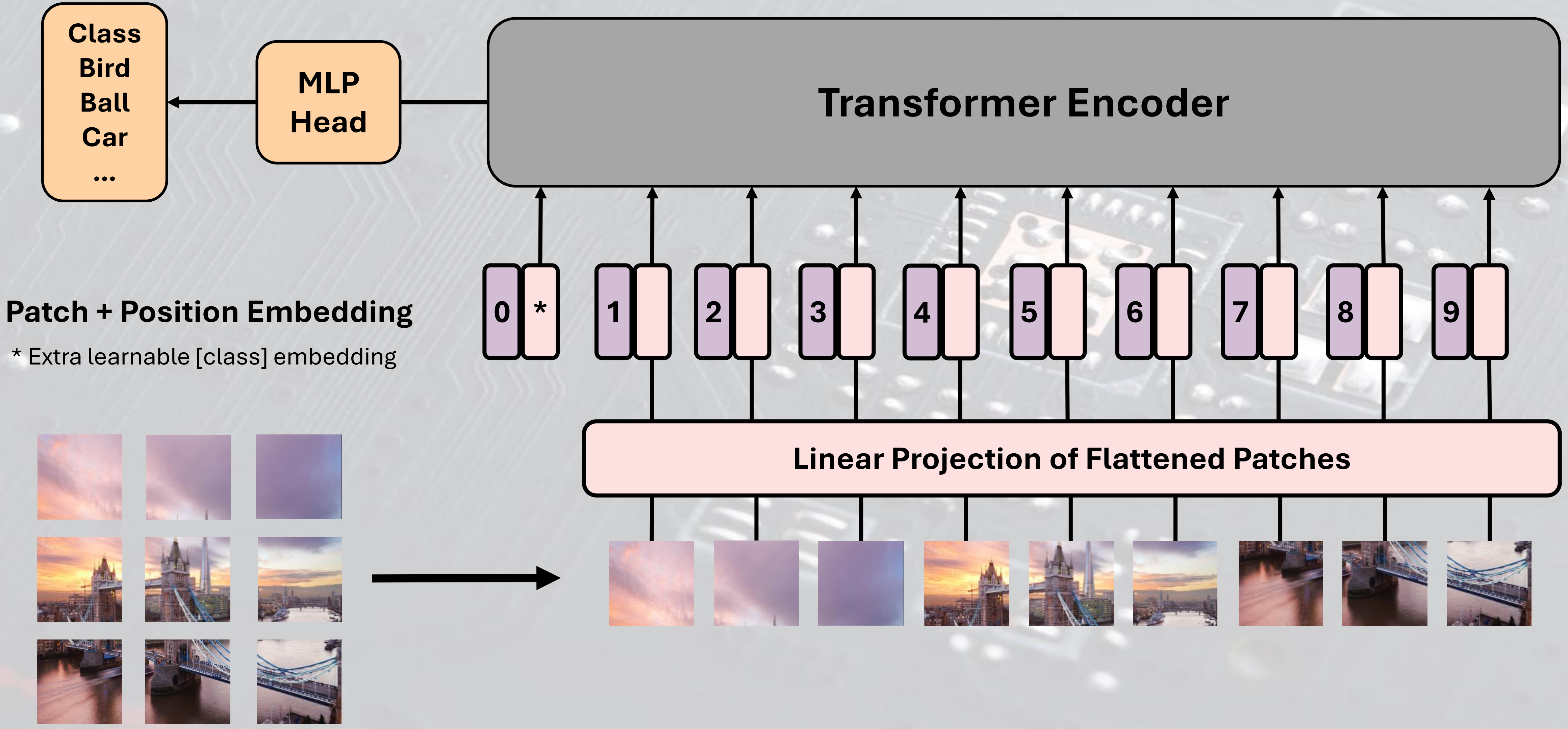
Vision Transformer (ViT)

- ***ViT-Base*** – denoted as ViT-B
- ***ViT-Large*** – denoted as ViT-L
- ***ViT-Huge*** – denoted as ViT-H

Hybrid Vision Transformer

- Use CNN to ***extract feature***
- ***Feed extracted features*** to ViT

Vision Transformer



Vision Transformer

Input Image

- **Original size:** 48x48
- **Number of patches:** 9
- **Patch size:** 16x16
- Arrange patches in order, from left to right and top to bottom

Vision Transformer

Input Image

- **Original size:** 48x48
- **Number of patches:** 9
- **Patch size:** 16x16
- Arrange patches in order, from left to right and top to bottom

1 2 3 4 5 6 7 8 9

Vision Transformer

Flattening Patches

- **Grayscale:** Each patch will become a vector with a size of (16x16x1) = **256**
- **RGB:** Each patch will become a vector with a size of (16x16x3) = **768**

1 2 3 4 5 6 7 8 9

[...(768)**...]** → **[...**(768)**...]** → **[...**(768)**...]**

9 Vectors

Vision Transformer

Linear Projection

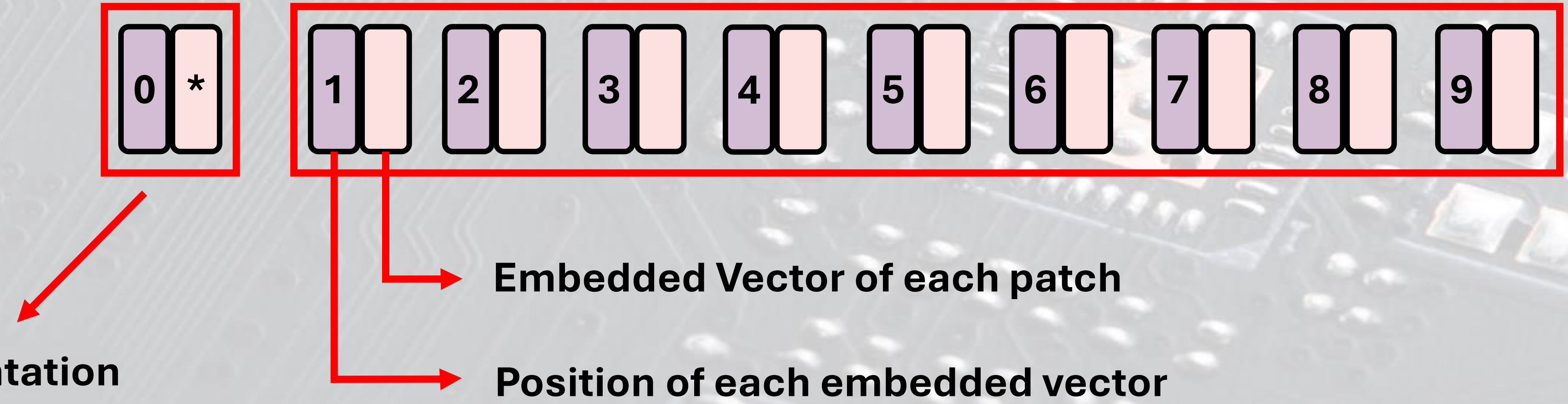
- Each vector will be transformed into **“Embedded vector”** by multiplying it with **weight matrix**
- Weight matrix has size **(length of vector, embedding dimension)**
- Where each weight is **trainable**, and embedding dimension is **hyperparameter**

		[..................] **[..................]** . . **[..................]**		
[..................]	**X**		**=**	**[..................]**
Patch Vector (1 x 768)		**Weight Matrix (768 x d)**		**Embedded Vector (1 x d)**

Vision Transformer

Patch and Position Embedding

- Embedding class and position of class
- Embedding position for each embedded vector



Representation

0 - Position of class vector

*** -** Class vector(learnable)

[.................] **+** **[.................]**

(Element-wise)

Position Vector (1 x d) + **Embedded Vector (1 x d)**

Vision Transformer

Transformer Encoder

- Each embedded vector with embedded position will be passed through the encoder individually

Main Components

1 **Multi-Head Attention**

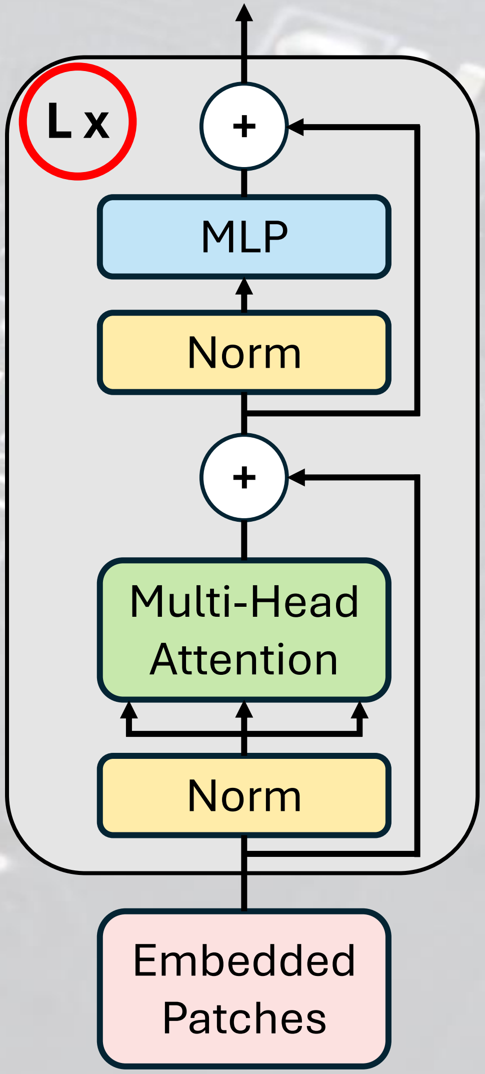
- Process input in parallel
- To capture complex relationship between data

2 **MLP**

- Apply non-linear transformation to the data

3 **Layer of transformer (L x)**

- Number of identical encoders stacked on top of each other



Vision Transformer

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

1 **Layers**

- Number of identical encoders stacked on top of each other in the encoder block

2 **Hidden size D**

- Number of dimension used in linear projection

3 **MLP size**

- The size of MLP in the encoder

4 **Heads**

- Number of multi-head attention used in each encoder

5 **Params**

- Total number of trainable parameters in the vision transformer model

Experiment

Experiment 1: Image classification tasks on various dataset

Models

Name	Type	Trained on (dataset)	Patch size	CNN architecture
ViT-H/14(JFT)	Vision transformer	JFT-300M	14	-
ViT-L/16(JFT)	Vision transformer	JFT-300M	16	-
ViT-L/16(I21k)	Vision transformer	ImageNet – 21k	16	-
BiT-L	CNN	-	-	ResNet152
Noisy Student	CNN	-	-	EfficientNet-L2

Result

Experiment 1: Image classification tasks on various dataset

Result

	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

Experiment

Experiment 2: Using different pre-trained dataset

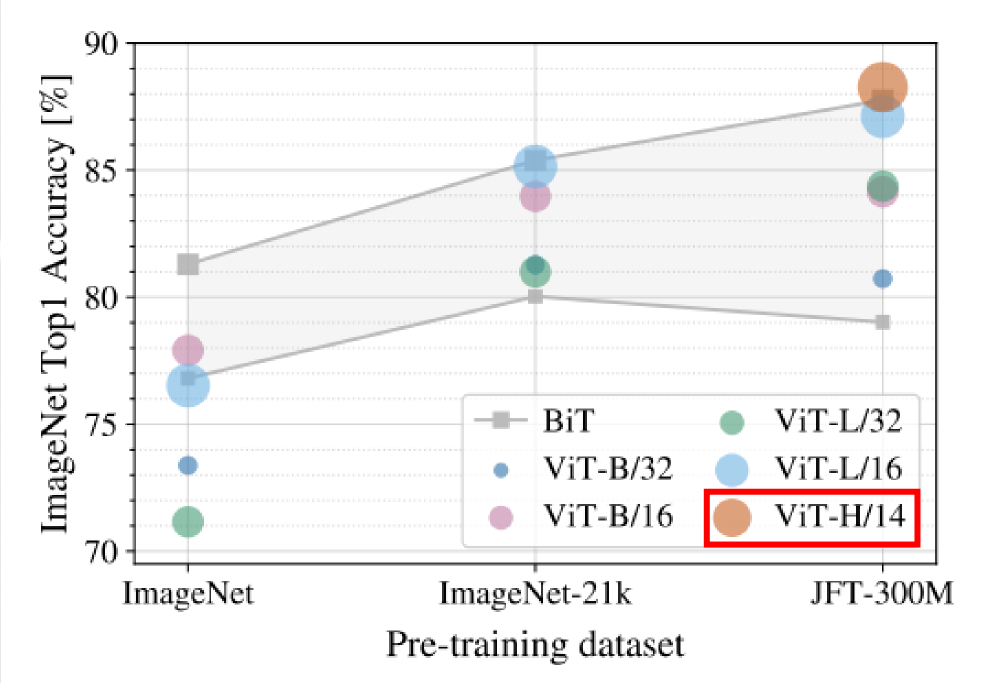
Models

Name	Type	Patch size	CNN architecture
ViT-H/14	Vision transformer	14	-
ViT-L/16	Vision transformer	16	-
ViT-L/32	Vision transformer	32	-
ViT-B/16	Vision transformer	16	-
ViT-B/32	Vision transformer	32	-
BiT	CNN	-	ResNet152

Result

Experiment 2: Using different pre-trained dataset

Result



**Test the accuracy on ImageNet dataset*

ImageNet
- Images – 1.2M
- Classes – 1k

ImageNet-21k
- Images - 14M
- Classes – 21k

JFT-300M
- Images - 300M
- Classes – 18k

" Vision transformer performs better when being trained on a larger dataset "

Experiment

Experiment 3: Using the same pre-trained dataset, but different sample size

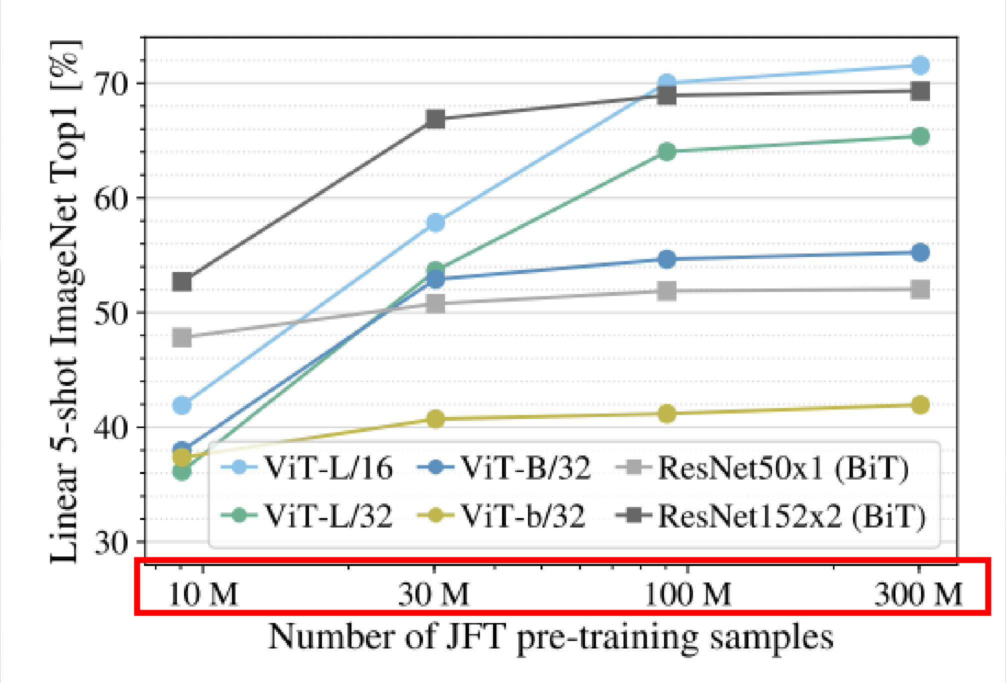
Models

Name	Type	Hidden dimension	Patch size	CNN architecture
ViT-L/16	Vision transformer	1024	16	-
ViT-L/32	Vision transformer	1024	32	-
ViT-B/32	Vision transformer	768	16	-
ViT-b/32	Vision transformer	384	32	-
BiT	CNN	-	-	ResNet152
BiT	CNN	-	-	ResNet50

Result

Experiment 3: Using the same pre-trained dataset, but different sample size

Result



**Few-shot evaluation on ImageNet dataset*

- ***ResNet152*** model outperform when being trained on ***10M and 30M samples.***
- ***Large vision transformer*** surpasses when the samples size increase to ***100M and 300M***

Experiment

Experiment 4: Using the same computational budget

Models

Modified CNN (BiT)

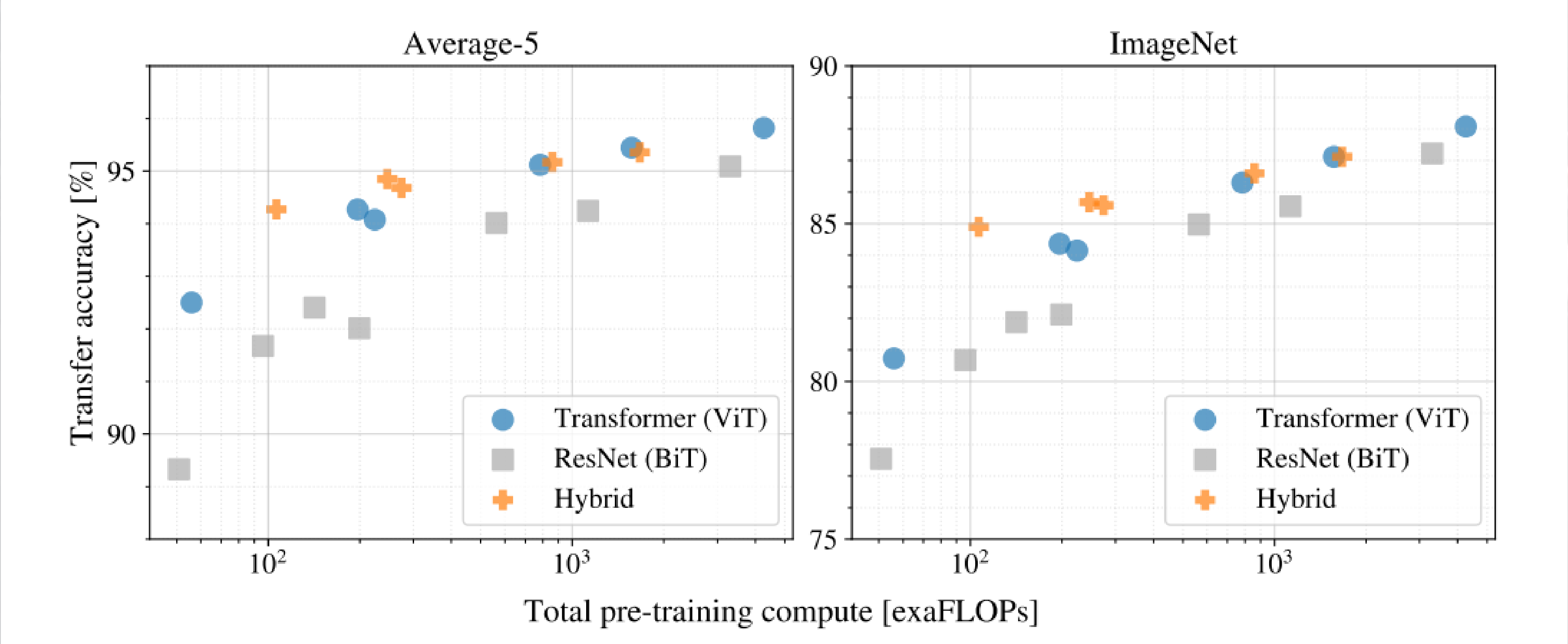
Vision Transformer (ViT)

Hybrid Vision Transformer

Result

Experiment 4: Using the same computational budget

Result



Conclusion

1. Vision Transformers (ViT) ***achieve state-of-the-art performance*** on image classification tasks, outperforming conventional CNNs when pre-trained on large datasets.

2. ViT models ***scale effectively*** with the size of the pre-training dataset

3. Introduced ***various ViT models***, including ***Base, Large, and Huge***, each with increasing capacity and complexity

4. Within the ***same computing budget***, ViT achieves ***better performance*** in comparison to CNN model

5. ViT learns to recognise image ***without relying on the assumption***, which is suitable for new kind of image task

Discussion

1. **Model Scalability:** Vision Transformers show improved performance with increased model complexity, indicating potential for ***even greater accuracy with future refinements*** in architecture and training techniques.

2. **Data Demands:** As models grow in complexity, they ***require larger datasets for optimal training***, highlighting the need for more efficient data utilization strategies or semi-supervised learning approaches

3. **Application Potential:** The adaptability of Vision Transformers to various scales of data ***suggests extensive applications***, from healthcare diagnostics to automated systems in transportation and manufacturing

Reference

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N., 2020. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arXiv [Online]. Available from: https://arxiv.org/abs/2010.11929 [14 April 2024].

THANK YOU FOR LISTENING