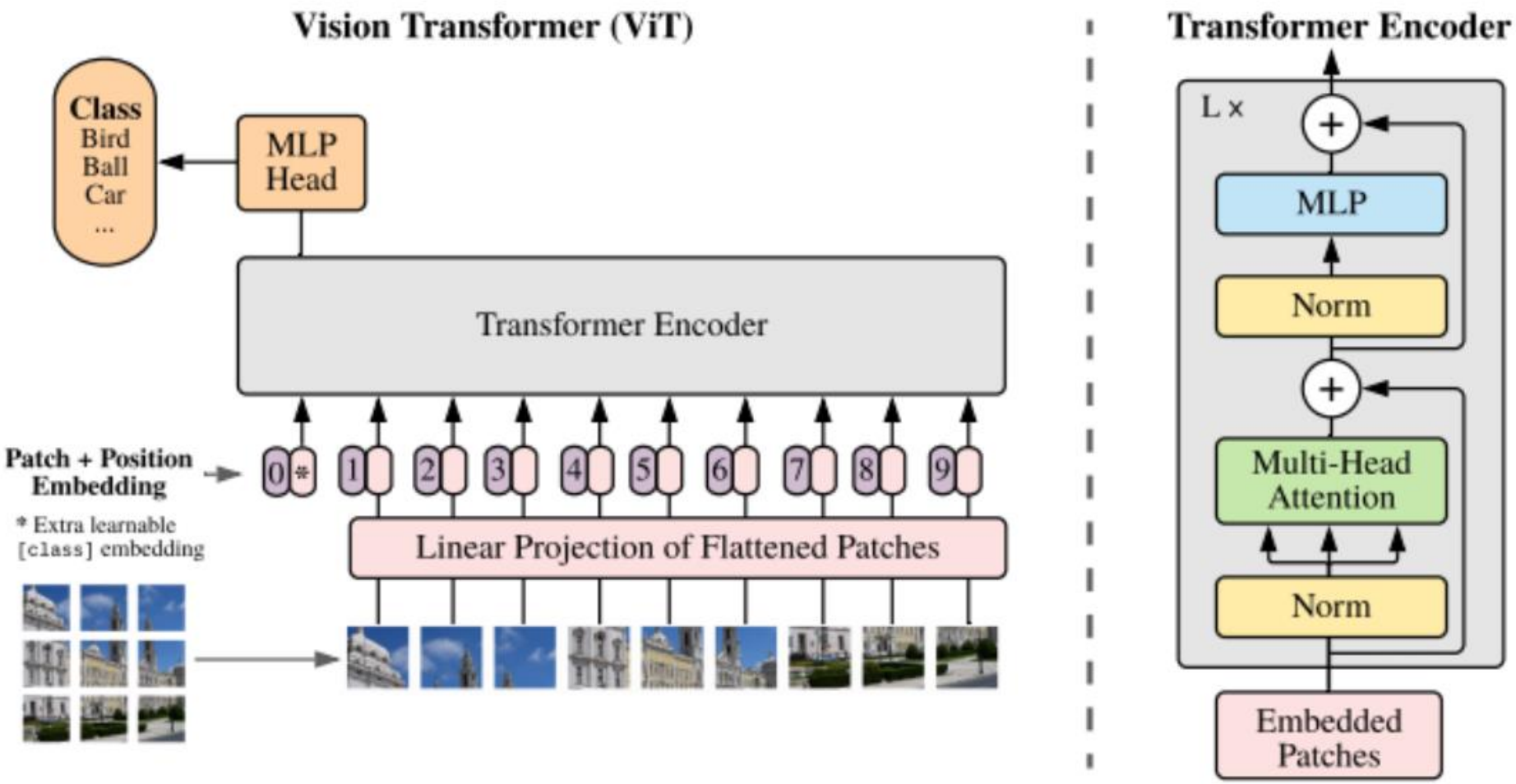
Vision Transformer



- Motivation
- Methodology
- Result

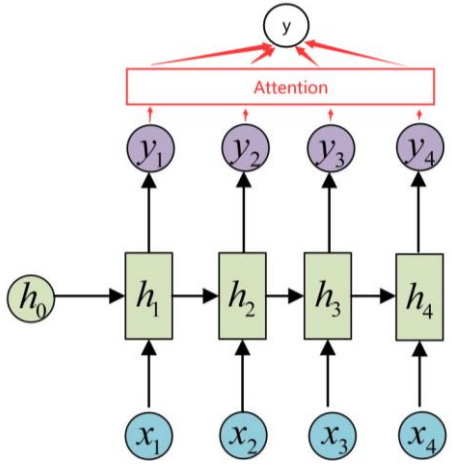
Motivation

We have shown that the standard **BERT** recipe **is effective** on a wide range of model sizes

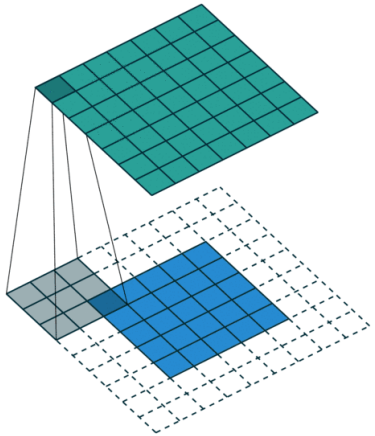
Time Sequential relation： Text

Spatial relation： Image

We have shown that the standard **BERT** recipe **is effective** on a wide range of model sizes



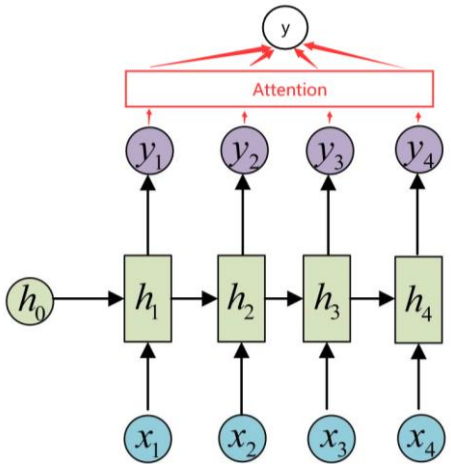
Attention，Transformer



Convolution

Since Transformer yields considerable result in Time Sequential data,
Can Transformer be applied to images?

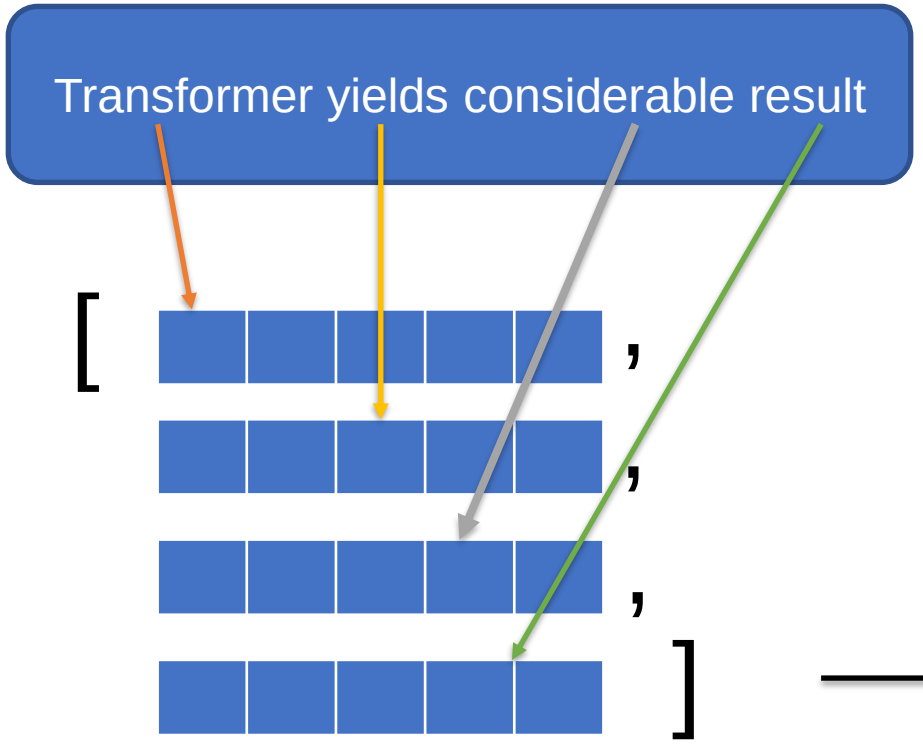
+

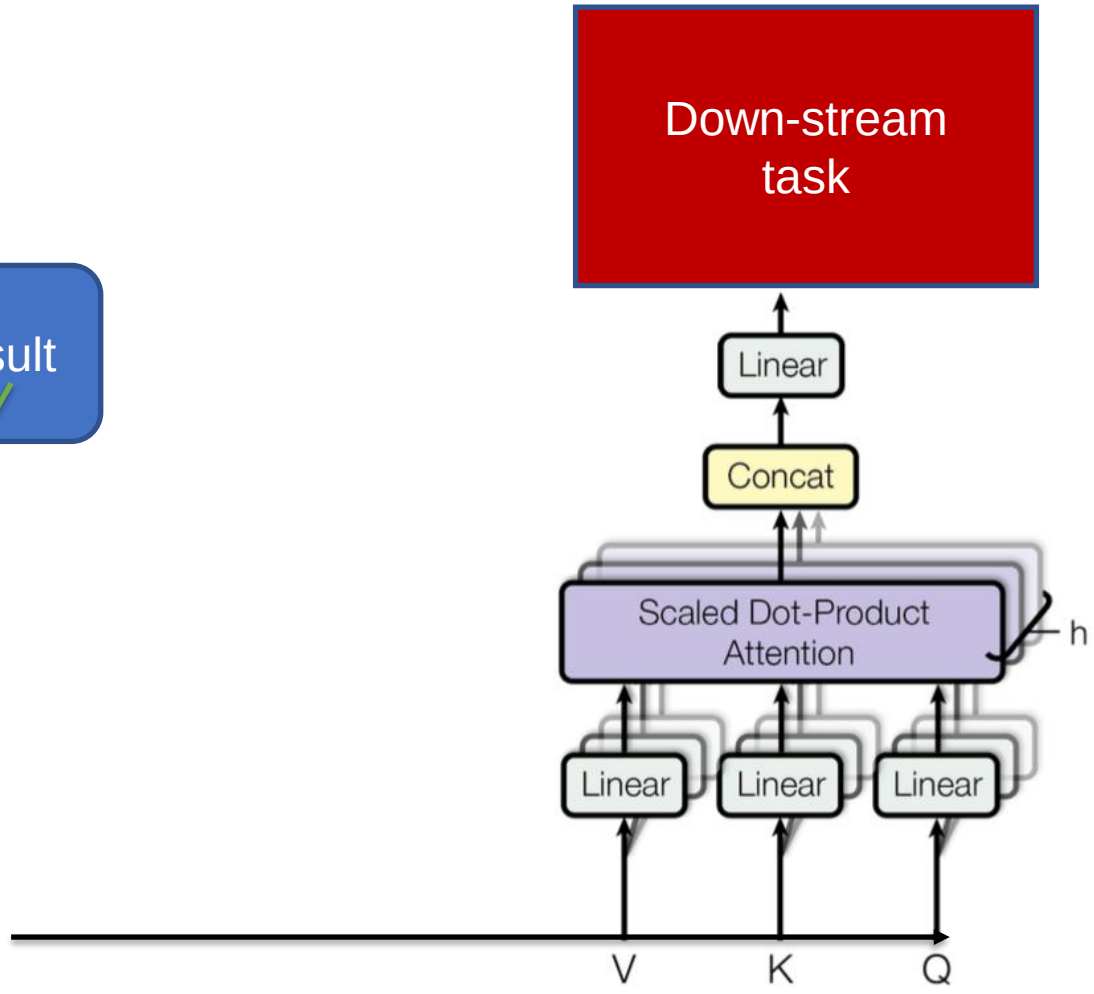


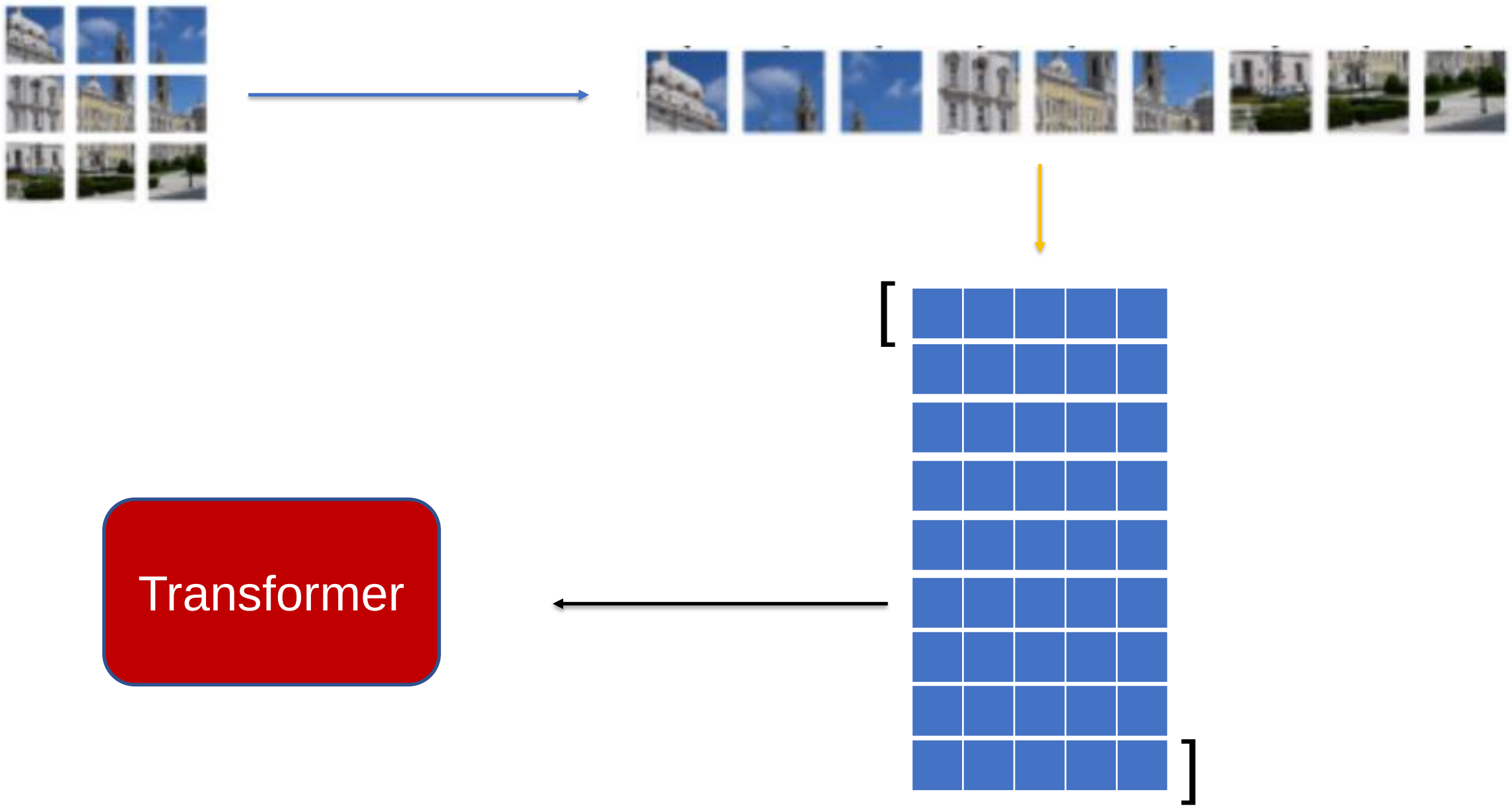
= ?

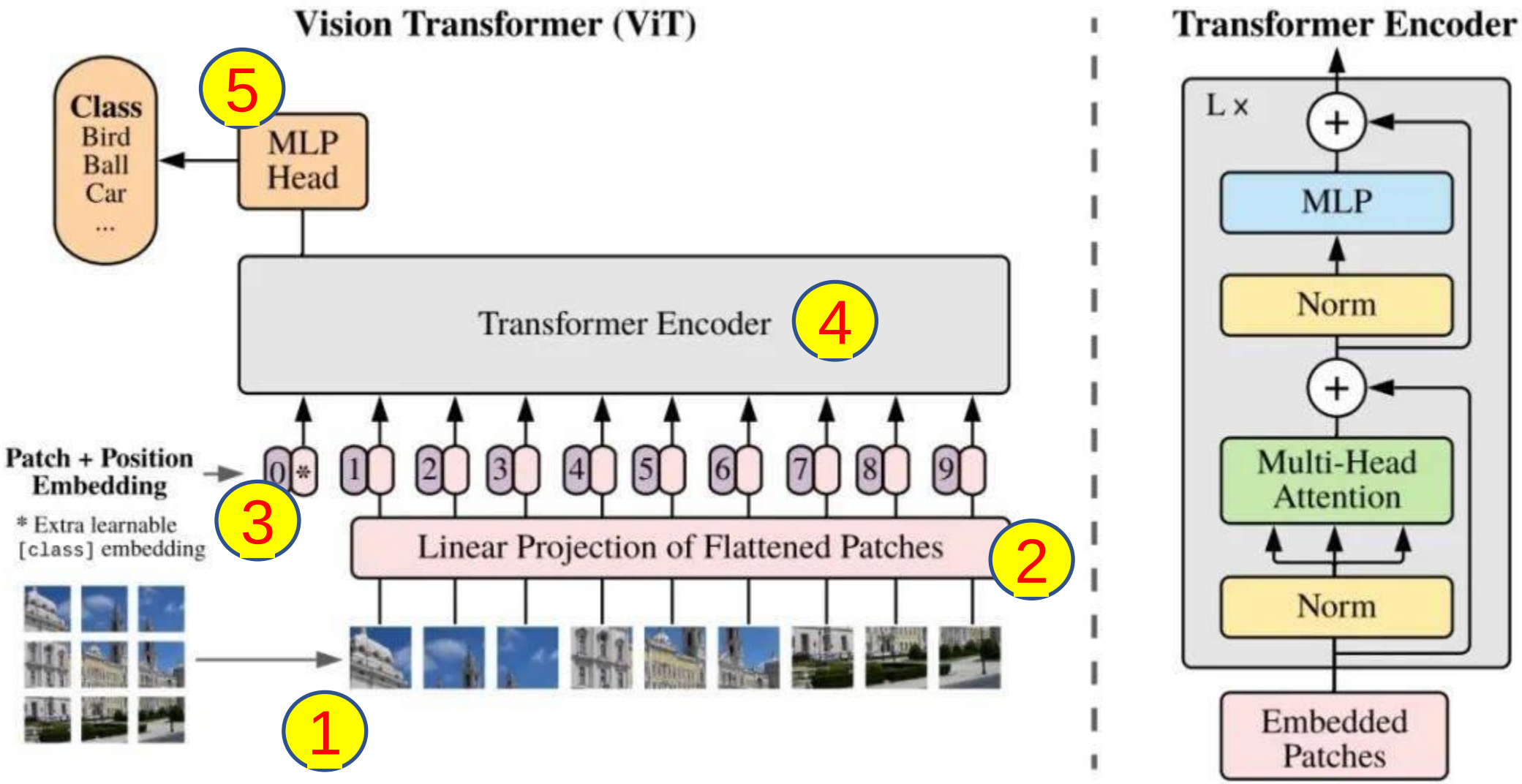
Methodology

Review： Transformer






Transformer


Vision Transformer (ViT)
Class
Bird
Ball
Car
...
5
MLP
Head
Transformer Encoder
4
Patch + Position
Embedding
* Extra learnable
[class] embedding
3
0 * 1 2 3 4 5 6 7 8 9
Linear Projection of Flattened Patches
2
1
Transformer Encoder
L x
MLP
Norm
Multi-Head
Attention
Norm
Embedded
Patches

Result

CIFAR100 Top1 Accuracy

1	**EffNet-L2** (SAM)	96.08	✓	Sharpness-Aware Minimization for Efficiently Improving Generalization
2	**ViT-H/14**	94.55±0.04	✓	An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale
3	**ViT-B-16** (ImageNet-21K-P pretrain)	94.2	✓	ImageNet-21K Pretraining for the Masses
4	**CvT-W24**	94.09	✓	CvT: Introducing Convolutions to Vision Transformers
5	**ViT-L/16**	93.90±0.05	✓	An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

ImageNet Top1 Accuracy

1	**CoAtNet-7**	90.88%		2440M	✓	CoAtNet: Marrying Convolution and Attention for All Data Sizes			2021	CNN Conv+Transformer JFT-3B
2	**ViT-G/14**	90.45%		1843M	✓	Scaling Vision Transformers			2021	Transformer JFT-3B
3	**CoAtNet-6**	90.45%		1470M	✓	CoAtNet: Marrying Convolution and Attention for All Data Sizes			2021	Conv+Transformer JFT-3B
4	**ViT-MoE-15B** (Every-2)	90.35%		14700M	✓	Scaling Vision with Sparse Mixture of Experts			2021	Transformer JFT-3B
5	**Meta Pseudo Labels** (EfficientNet-L2)	90.2%	98.8%	480M	✓	Meta Pseudo Labels			2021	EfficientNet JFT-300M