Vision Transformer

ViT

Vision Transformers Series

Recap of previous videos:

Evolution of Self-Attention in Images

(1)

Augmenting conv. neural networks with the attention mechanism

E.g., AAConv

(2)

Building fully-attentional models

E.g., SASA

What makes transformers so powerful?

1. **Attention mechanism** to learn long-range dependencies
2. Scalability in Pre-training on Large Datasets for **Transfer Learning**
3. Efficiency in Leveraging **Self-Supervised** Learning on **Unlabeled Data**

ViT overview

- Convert input image to a sequence of image patches (aka tokens)
- Applying standard transformer (with minimal alterations) to the sequence
- Design objective: minimal inductive bias, learn everything from scratch

Key Design Strategy:

- ➢ Make the least use of 2D structure
- ➢ Learn everything from scratch

ViT step-by-step

Input 2D image: $X \in \mathbb{R}^{H \times W \times C}$ → Sequence of image patches: $X_p \in \mathbb{R}^{N \times (P^2 C)}$

$X \in \mathbb{R}^{H \times W \times C}$

Each patch: $P \times P$

$$N = \frac{H \times W}{P^2} \Rightarrow \text{number of patches}$$

(**effective** sequence length)

$X \in \mathbb{R}^{H \times W \times C}$

$$N = \frac{H \times W}{P^2} \quad \Rightarrow \quad \text{number of patches}$$

(**effective** sequence length)

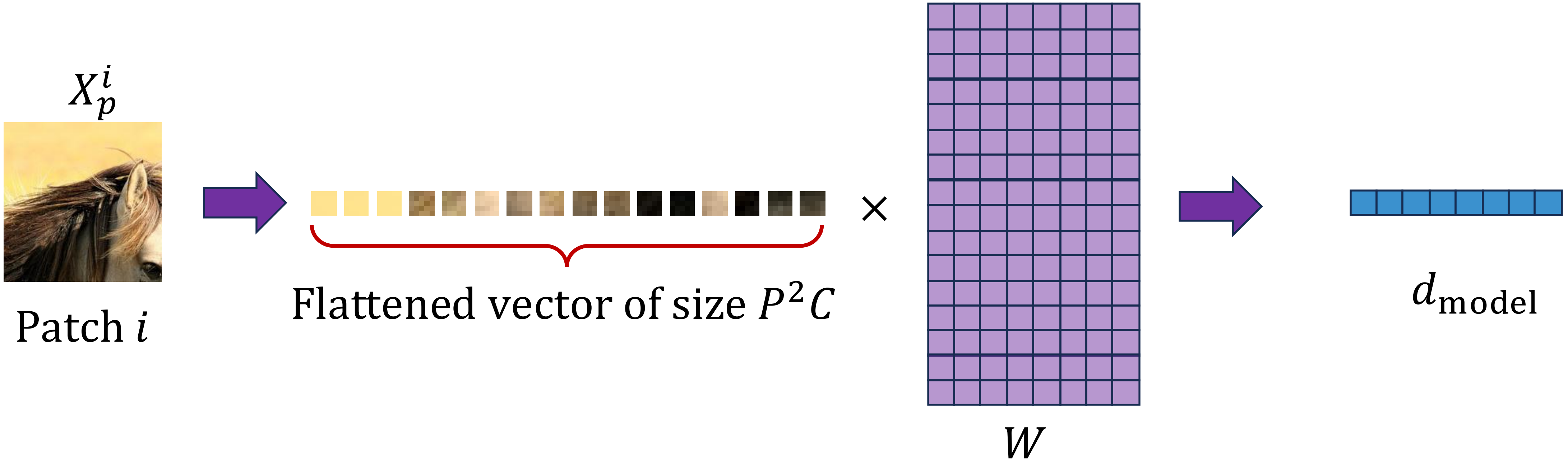
X_p^i

Patch i

Flattened vector of size P^2C

ViT step-by-step

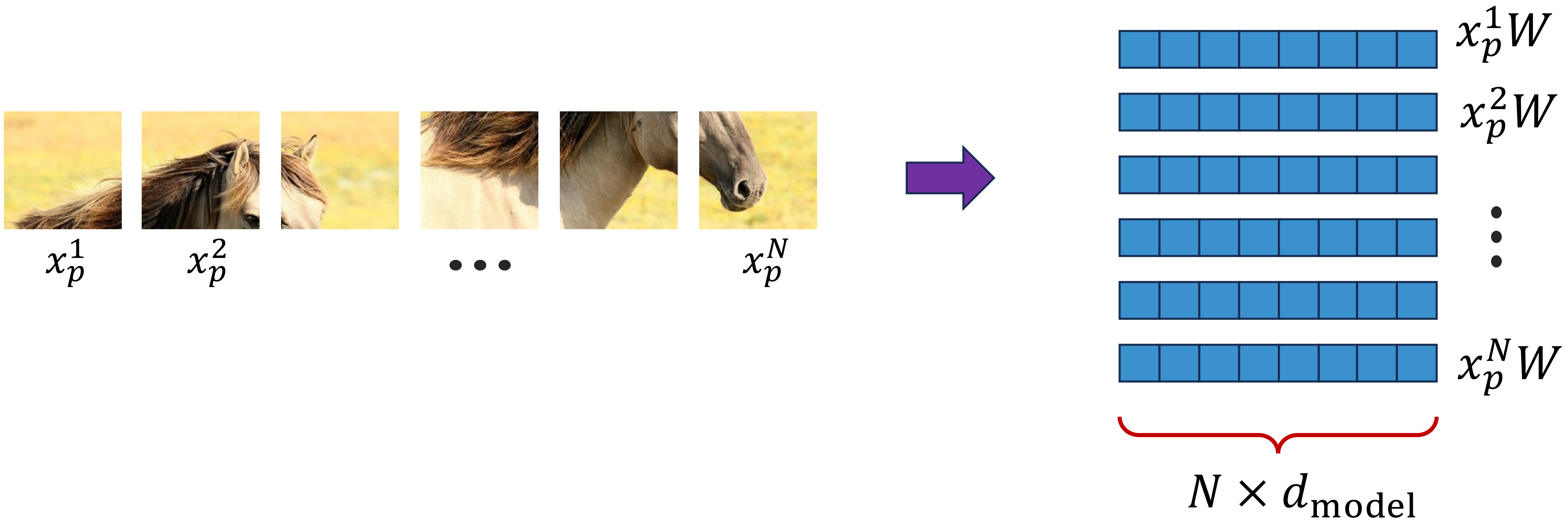
Flatten and project patches linearly $\Rightarrow$ $X_pW \in \mathbb{R}^{N \times d_{\text{model}}}$



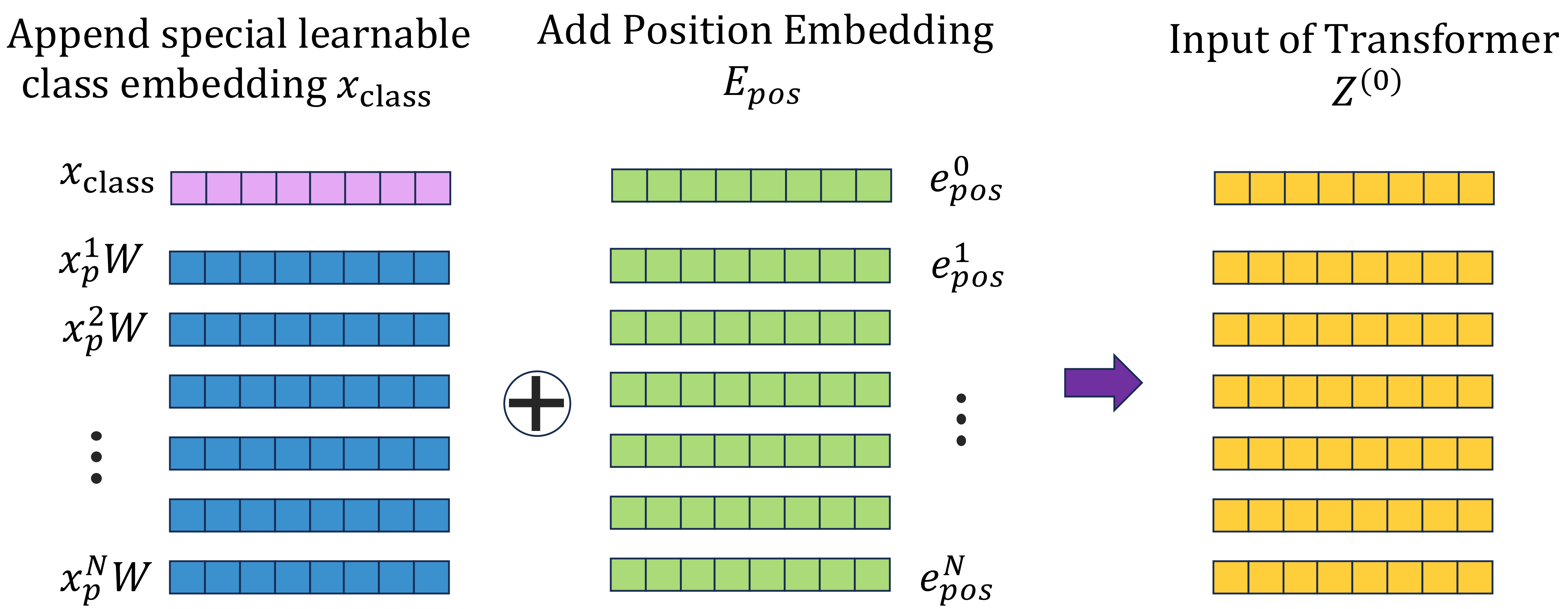
size $= P^2C \times d_{\text{model}}$

ViT step-by-step

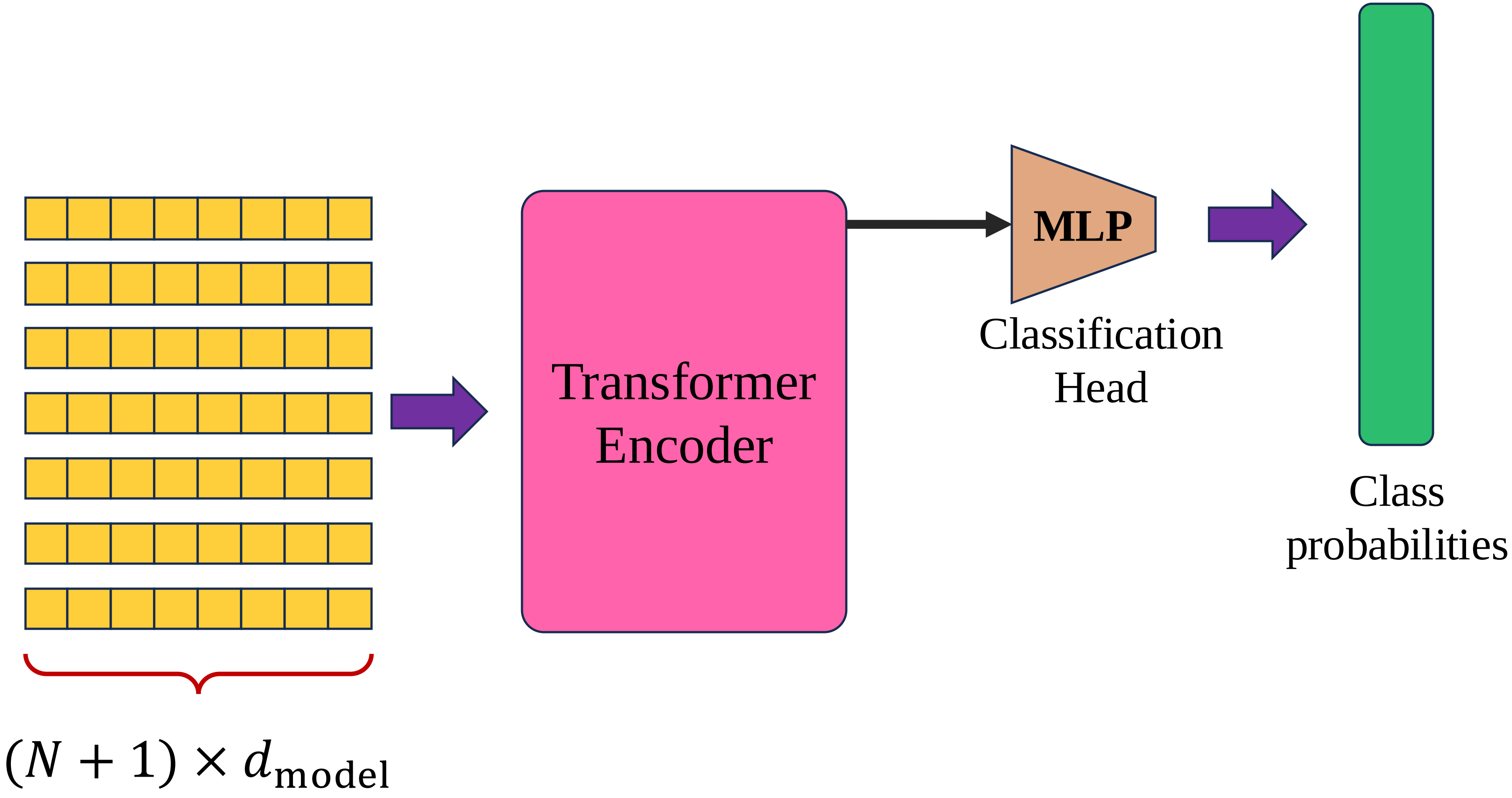
Flatten and project patches linearly $\Rightarrow$ $X_p W \in \mathbb{R}^{N \times d_{\text{model}}}$

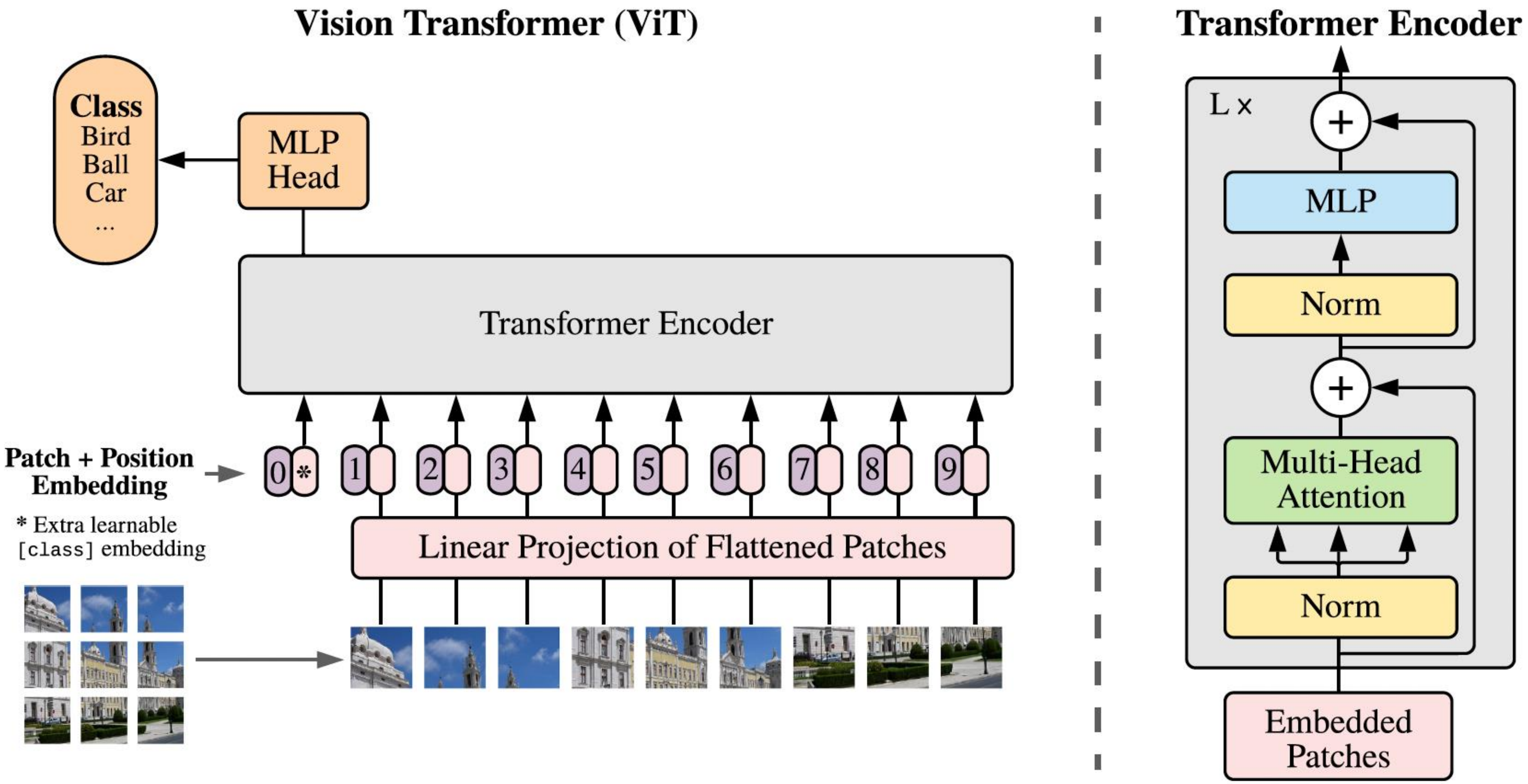


ViT step-by-step



ViT step-by-step




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

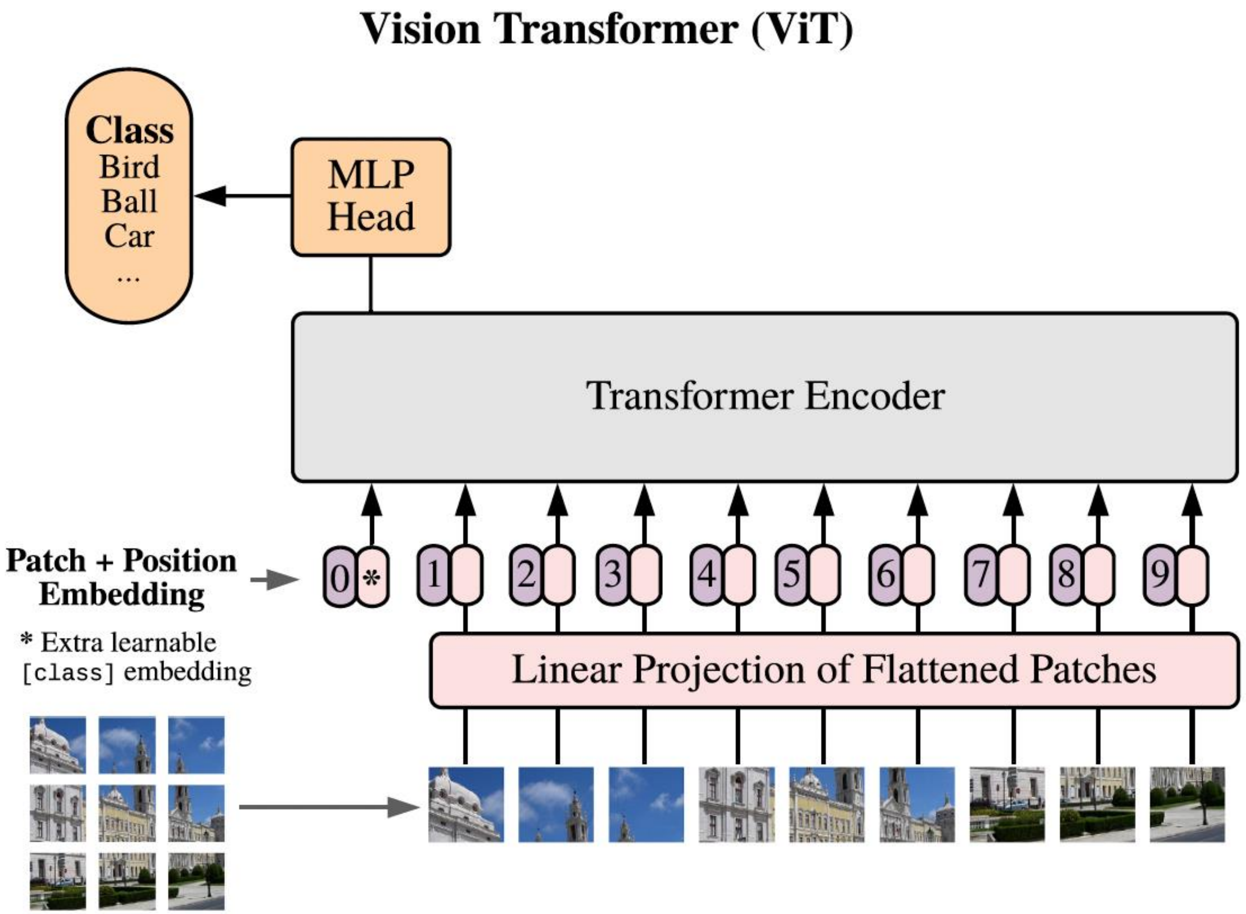
ViT Architecture

- **The first token:** special learnable classification token

Adapted from BERT

➔ The role of this token is to aggregate information from the entire sequence

➔ The final representation corresponding to this token is used in the final classification head



Training ViT

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

ViT-B/16 → Vit-Base with 16×16 patch

Pre-train on a large dataset

(JFT-300M)

Supervised Learning

↓

Fine-tune on smaller downstream tasks

Minimal inductive bias

- → Everything has to be learned from scratch
- → Requires lots of training data

Fine-tuning ViT

Replace the MLP head with a newly initialized linear layer

Effective to pre-train at low-resolution and then fine-tune at higher resolution

- Pre-train at 224×224
- Fine-tune at 384×384

Fine-tuning at higher resolution

- Maintain the same patch size as in pre-training

➔ Results in a higher number of patches (N)

- Problem: no learned position embeddings for $i > N_{224}$

Solution:
2D interpolation of learned position embeddings

ViT 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

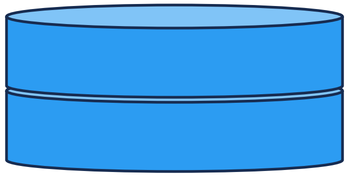
Mid-sized Training Data (ImageNet-21k)

ResNet-based models outperform ViT

Sufficient Training Data (JFT-300M)

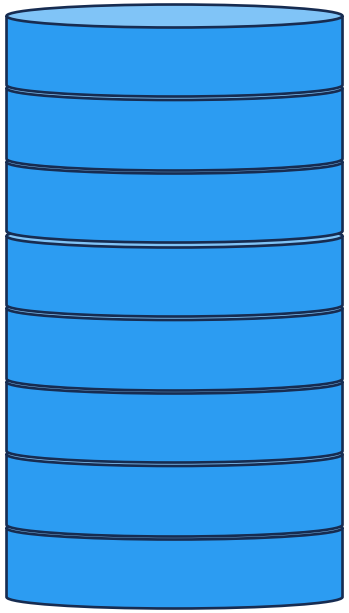
ViT achieves SOTA

Inductive Bias
vs.
Large-scale Training



Mid-sized training data
(e.g., ImageNet – 1M)

Inductive bias plays an important role



Very large training data
(e.g., JFT300M)

Large-scale training is superior to inductive bias

Self-Supervised Learning

Masking random patches

Similar to masked-language modeling (e.g., BERT)

Explored 3 mask-prediction strategies

Mean (RGB) pixel prediction	Predict a 4×4 downsized version	Predict the entire patch $(L2)$
0.71 0.62 0.47		

Key Take-aways

- ViT: Convert an input image into a sequence of image patches and applying standard Transformer
- Leveraging the key properties Transformers: Pre-train on large datasets
  - ➢ Pre-training on large data trumps inductive bias
- The paper mostly covered supervised learning ➔ need labeled data
- Preliminary exploration of self-supervised learning

Thanks for watching