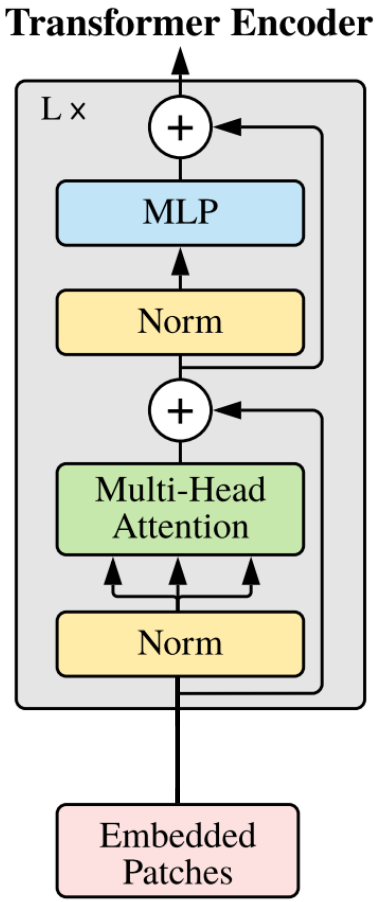
Vision Transformer (ViT)

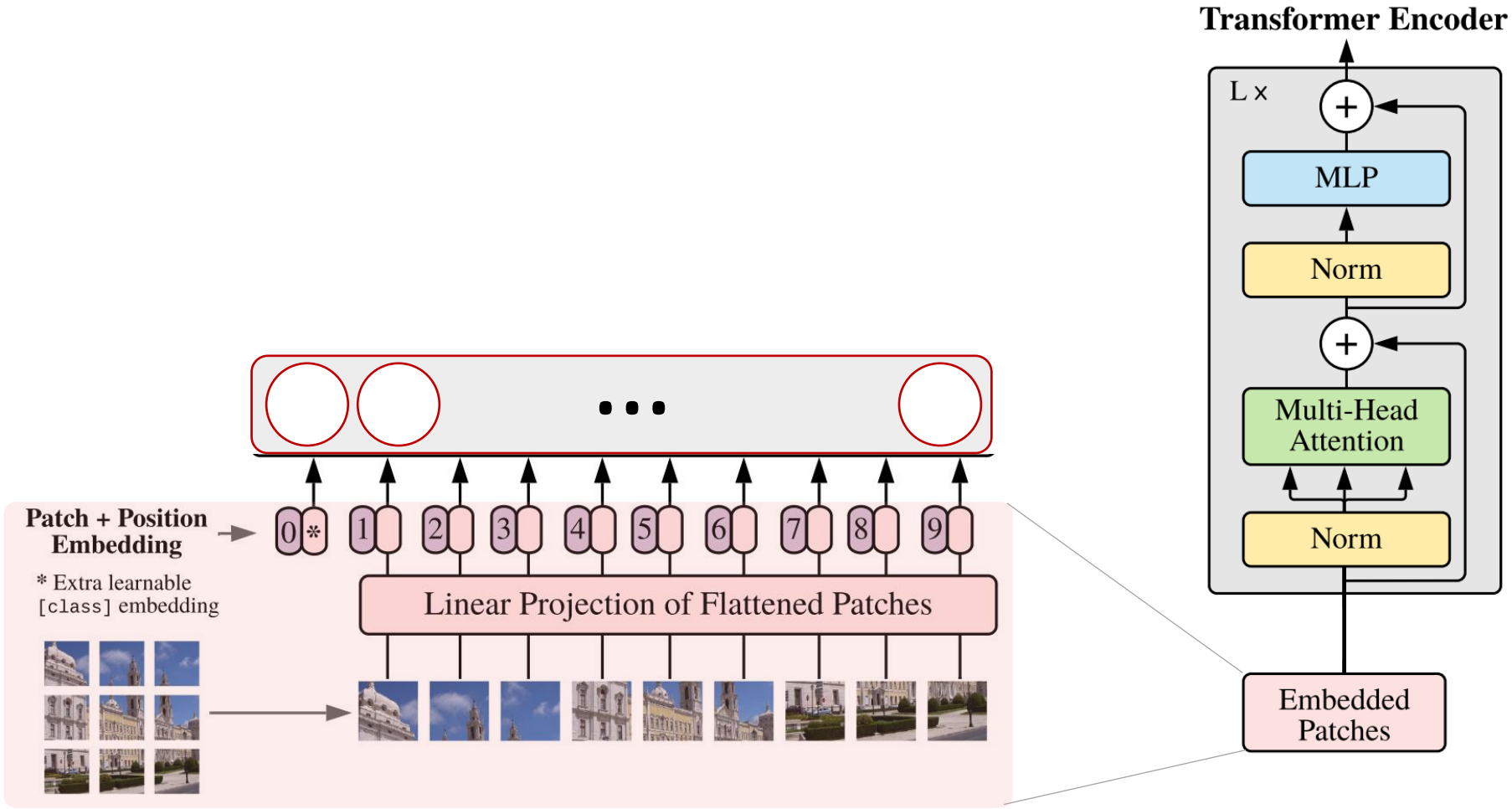
Arman Afrasiyabi

PHD Student at Université Laval

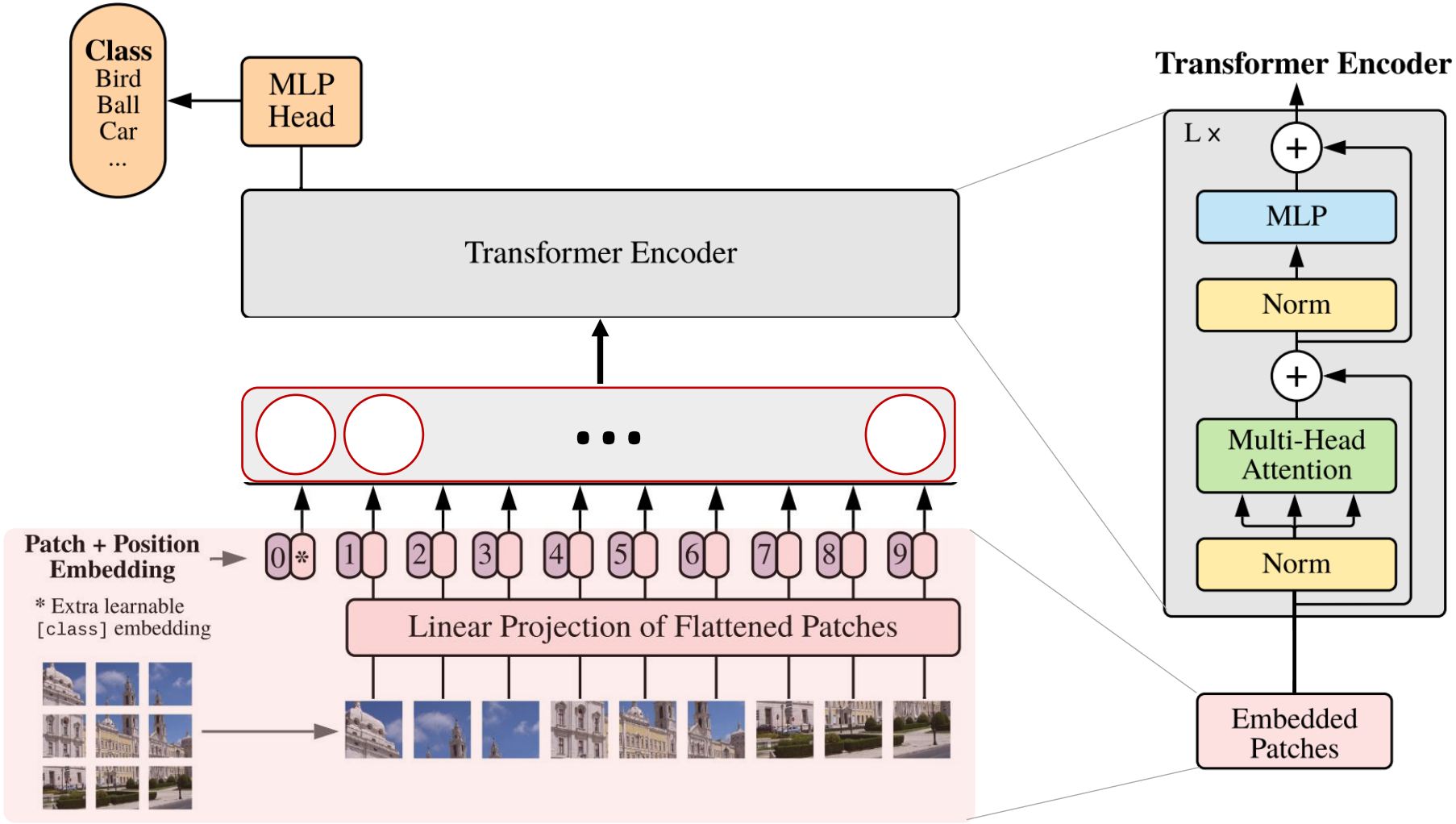
Vision Transformer (ViT)



Vision Transformer (ViT)



Vision Transformer (ViT)



Input shape

[b, 3, 84, 84]

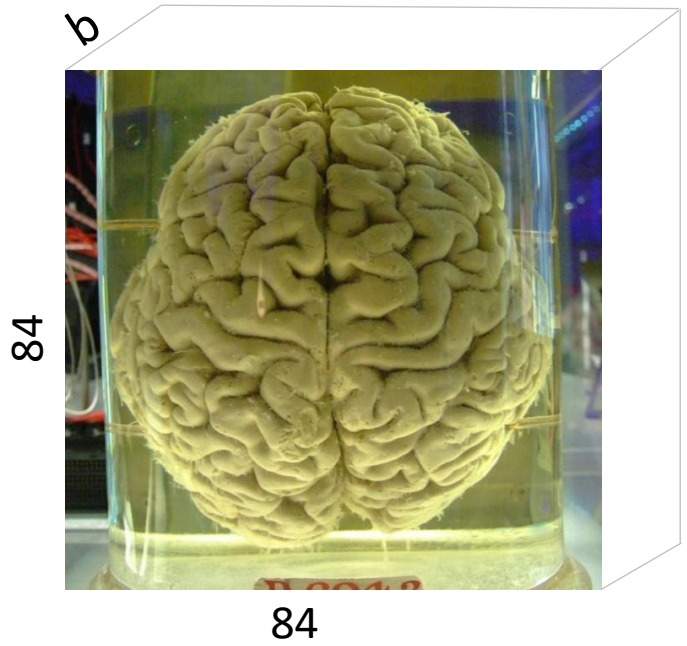
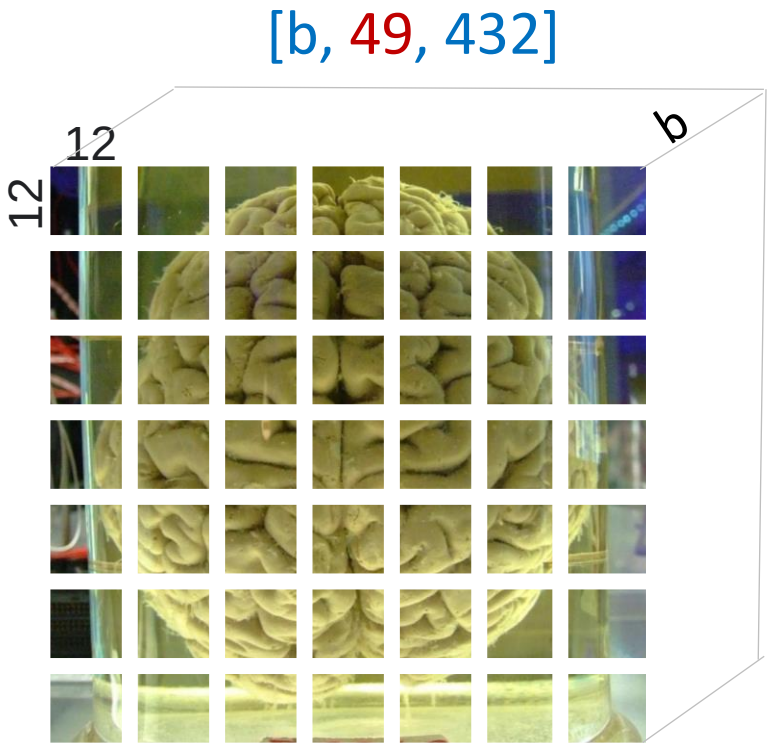
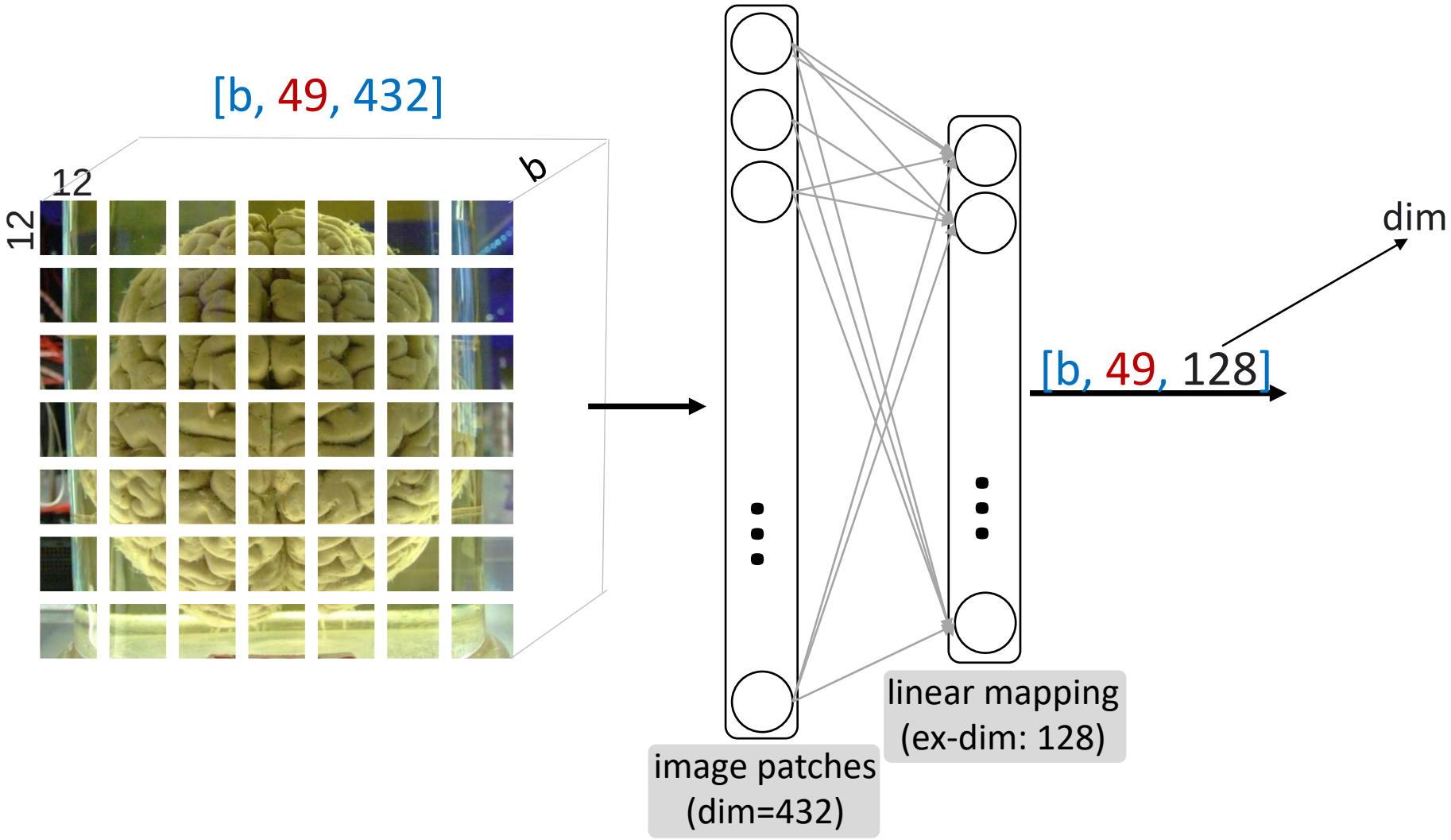


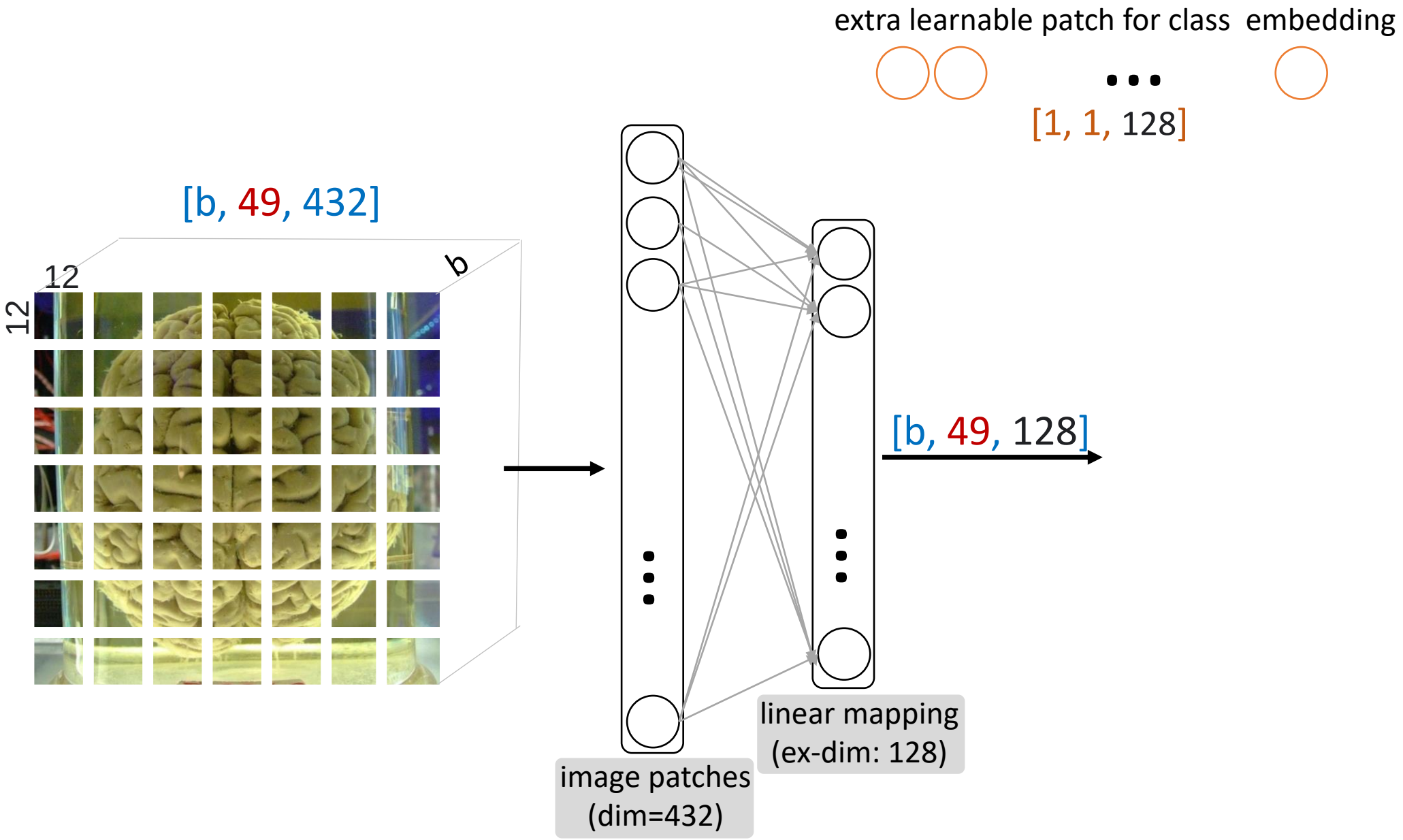
Image to Patches



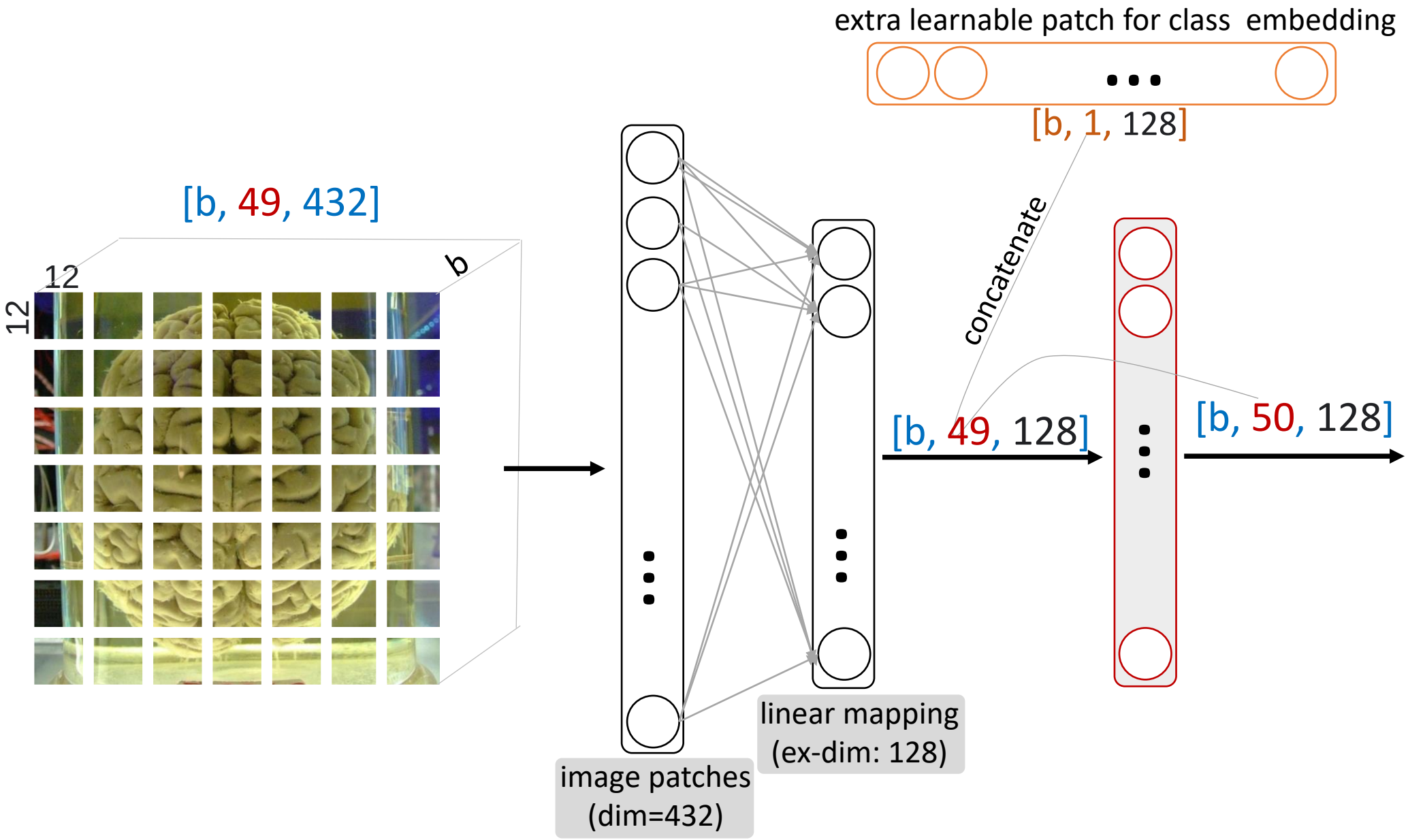
Patch to Token



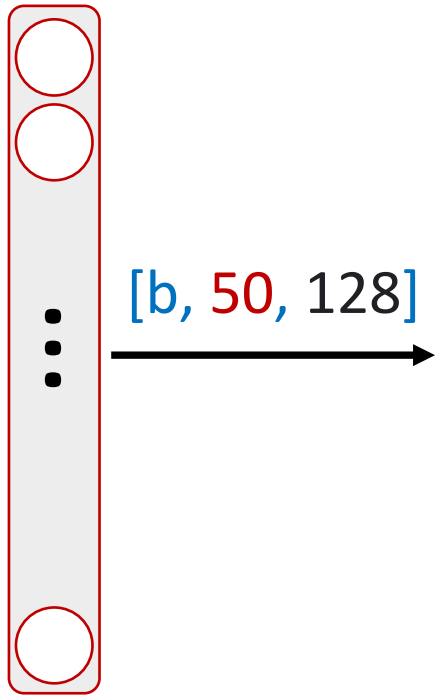
Concat. learnable Class Token



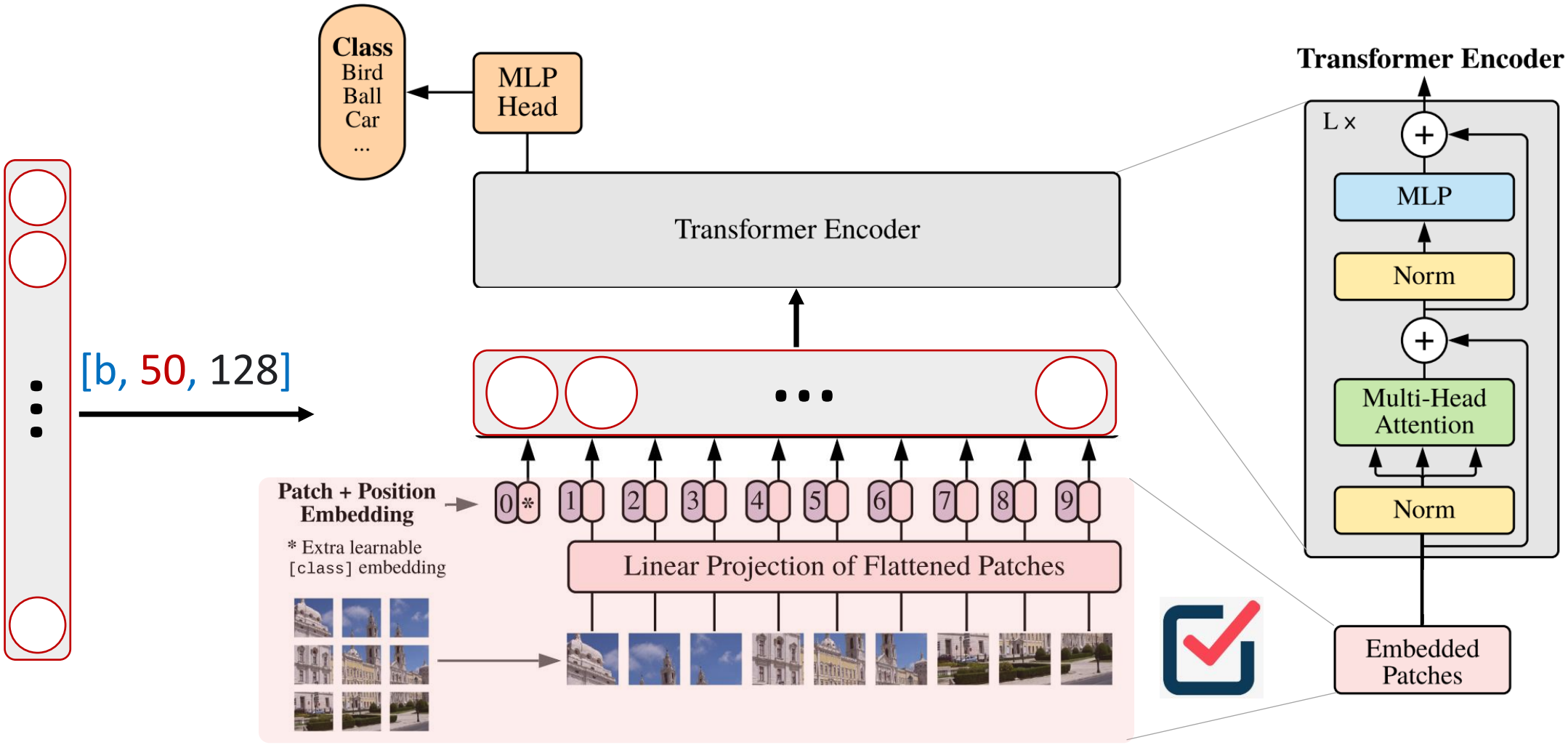
Concat. learnable Class Token



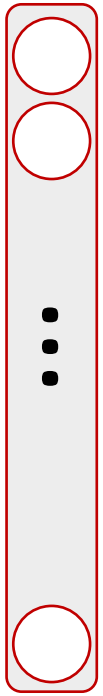
Patch embedding



Big picture

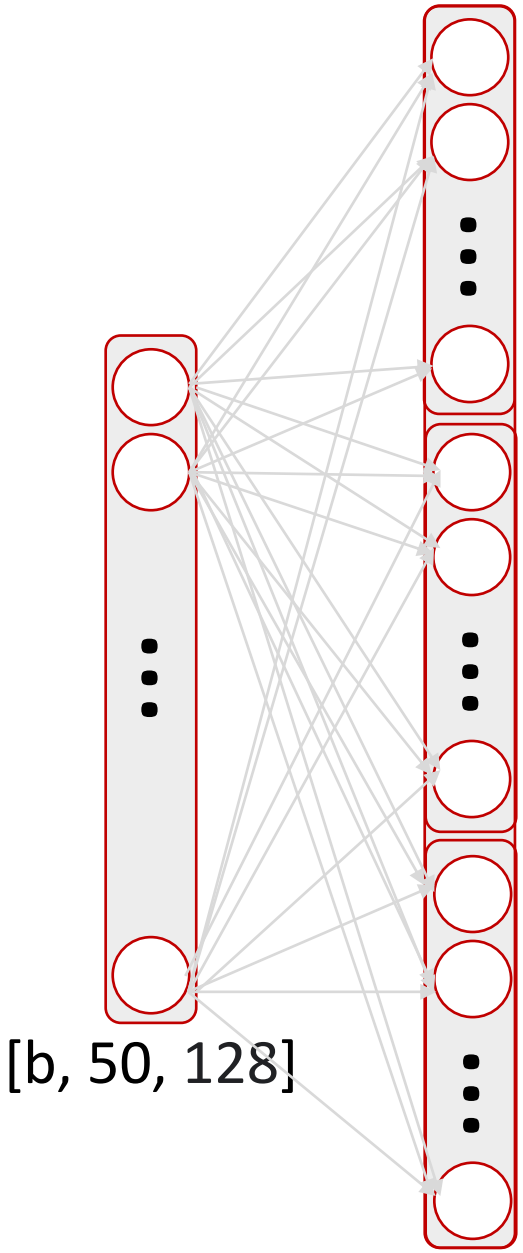

Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
[b, 50, 128]
Patch + Position Embedding
* Extra learnable [class] embedding
0 * 1 2 3 4 5 6 7 8 9
Linear Projection of Flattened Patches
Transformer Encoder
L x
MLP
Norm
Multi-Head Attention
Norm
Embedded Patches

Transformer: MH-Attension

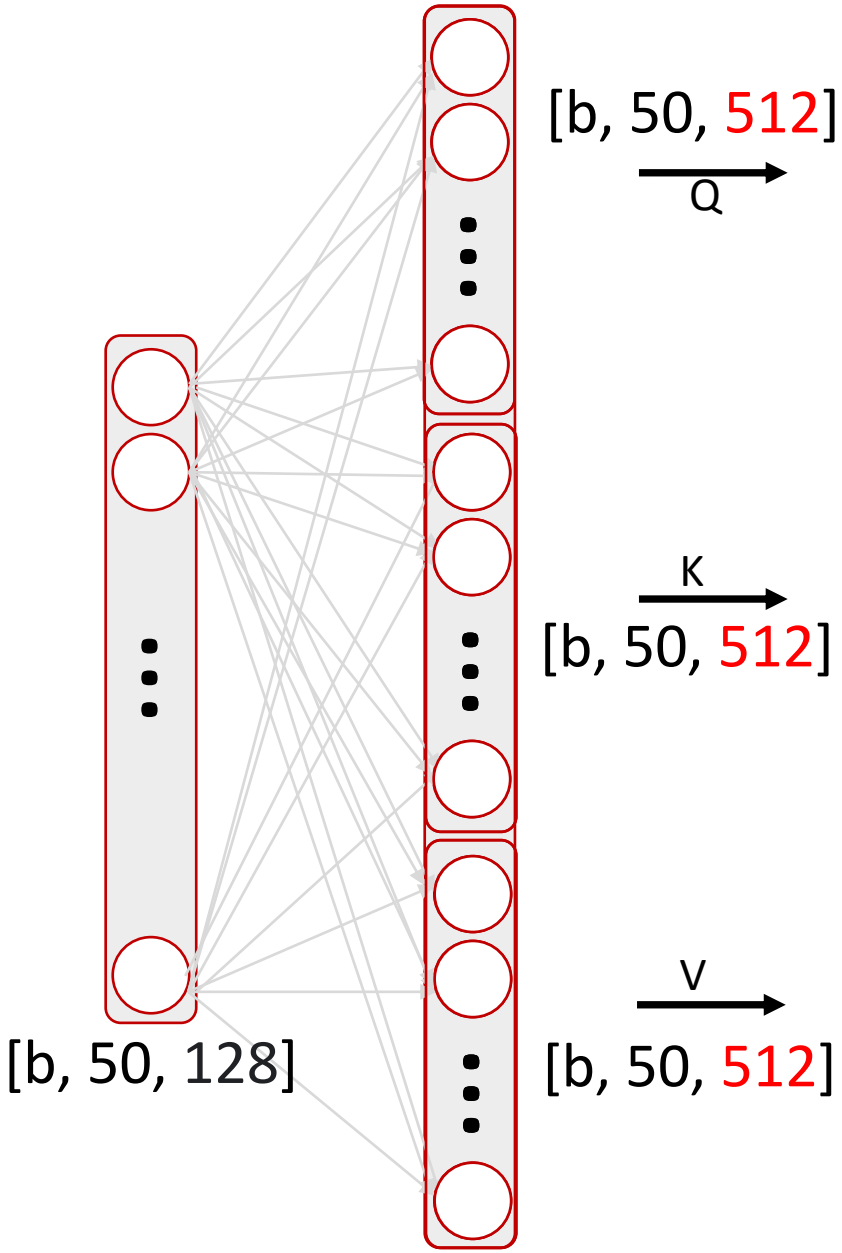


[b, 50, 128]

Transformer: MH-Attention



Transformer: MH-Attention

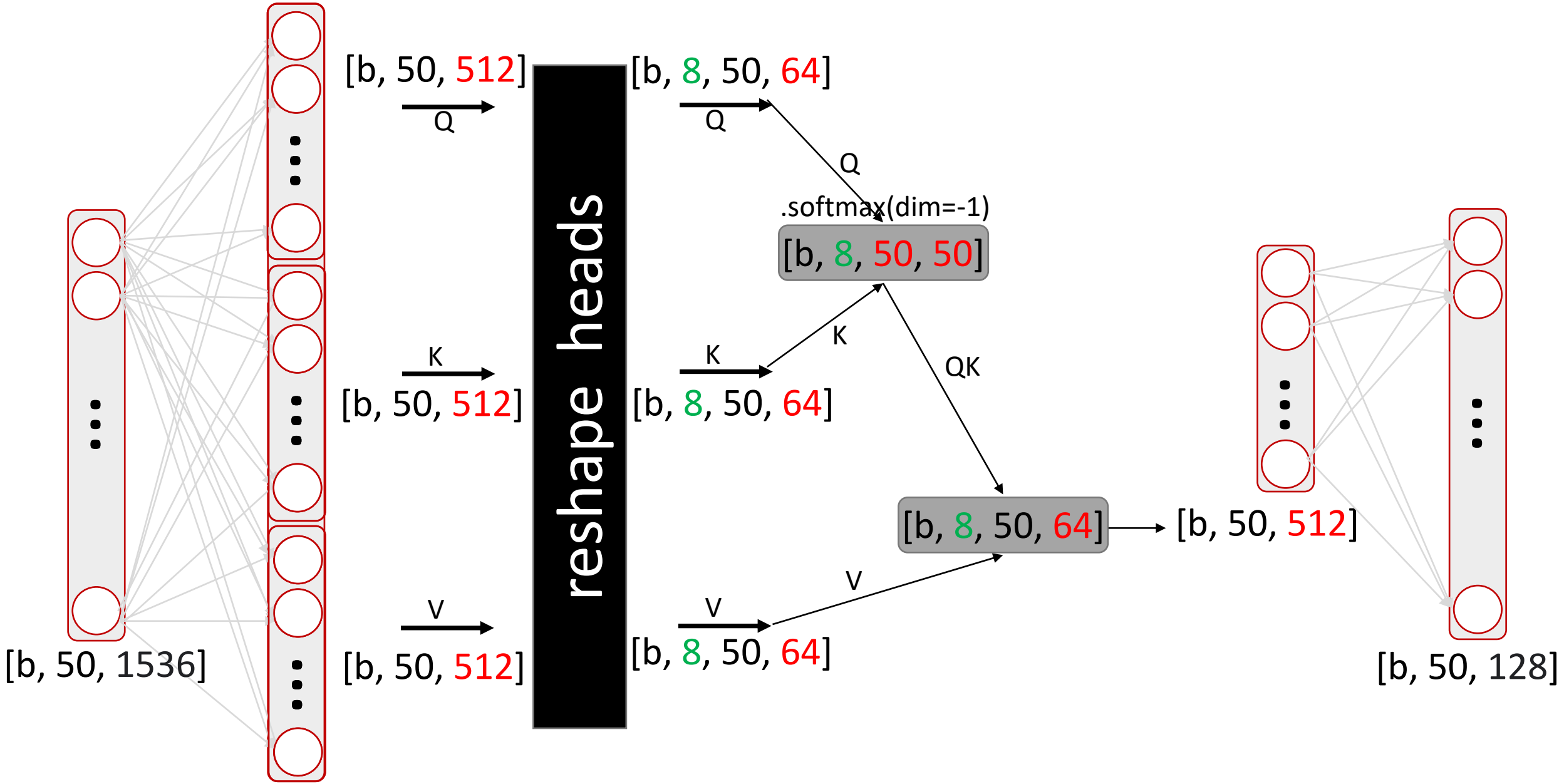


512=64*3

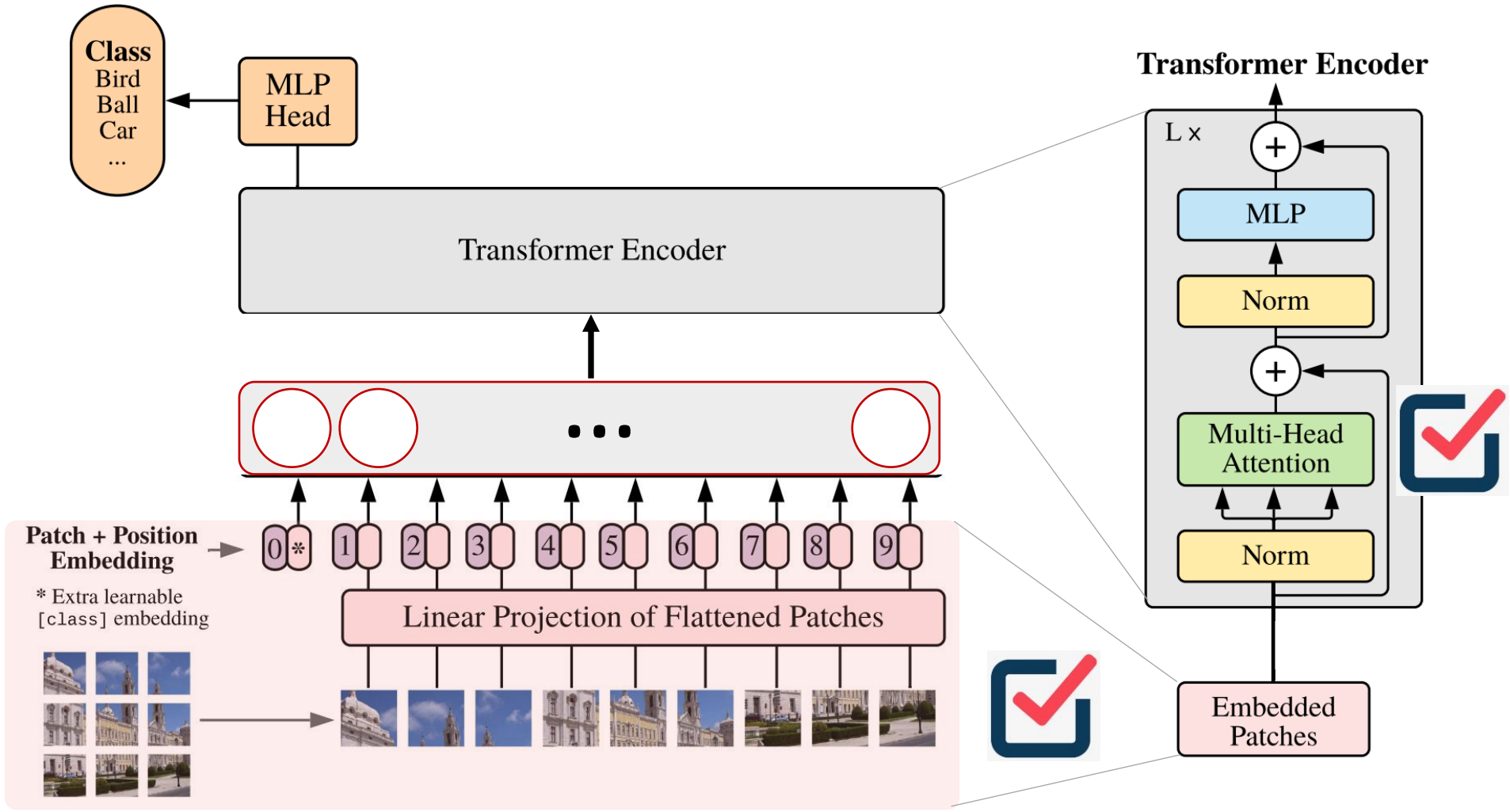
n-heads

head_dim

Transformer: MH-Attention (n_head = 8)



Big picture


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 x
MLP
Norm
Multi-Head Attention
Norm
Embedded Patches

Transformer: MLP