Vision *Transformer*

plain concepts

Vision Transformers

Computer vision dominating architecture

Convolutional Neural Networks (CNN) — 1989

Transformers — 2017

Vision Transformes (ViT) — 2020

DINOv2
DeiT
CLIP

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

Alexey Dosovitskiy*,†, Lucas Beyer*, Alexander Kolesnikov*, Dirk Weissenborn*, Xiaohua Zhai*, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby*,†

*equal technical contribution, †equal advising

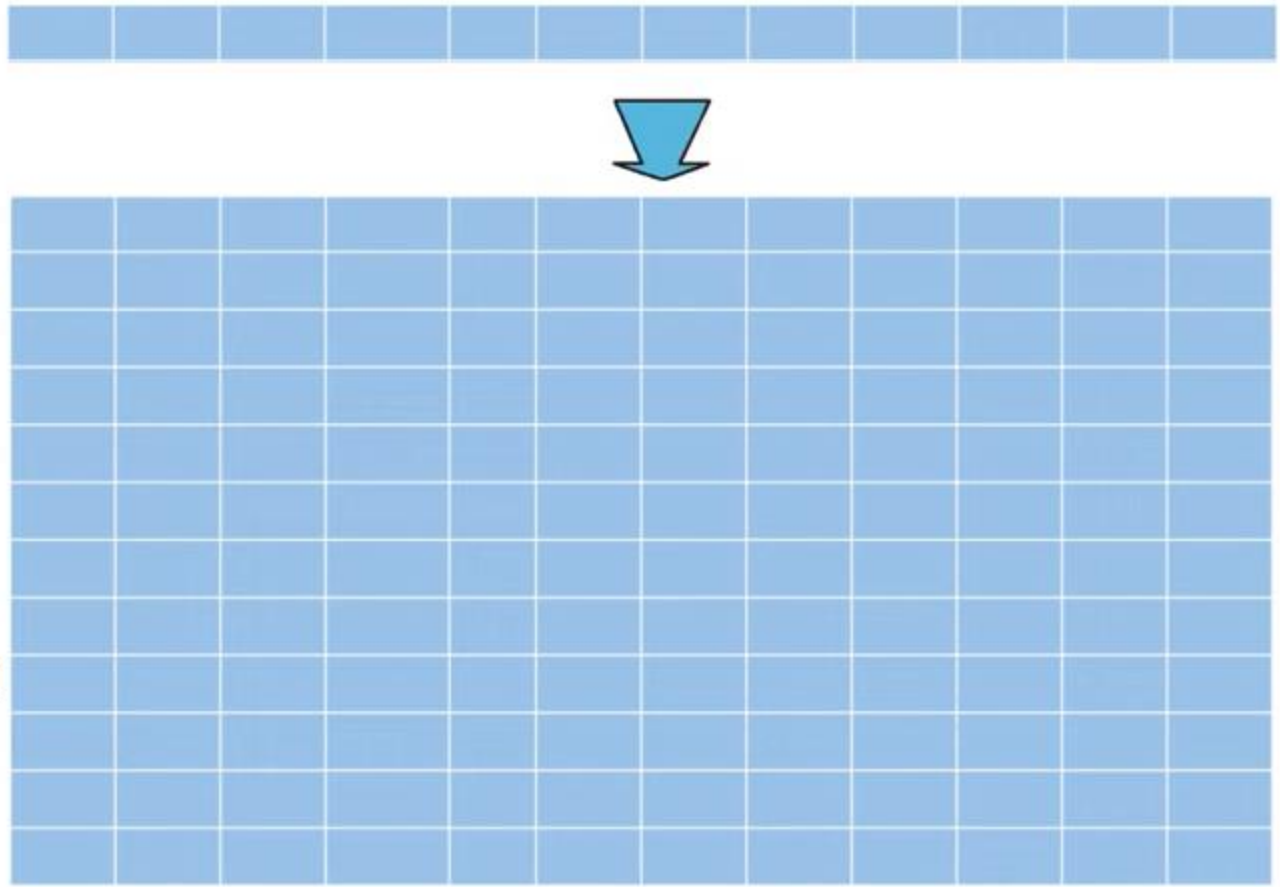
Google Research, Brain Team

{adosovitskiy, neilhoulsby}@google.com

Using a Transformer as-is?

- Transformer Self-Attention – each token attends all other tokens
- Feeding an image to traformer – each pixel attends all other pixels

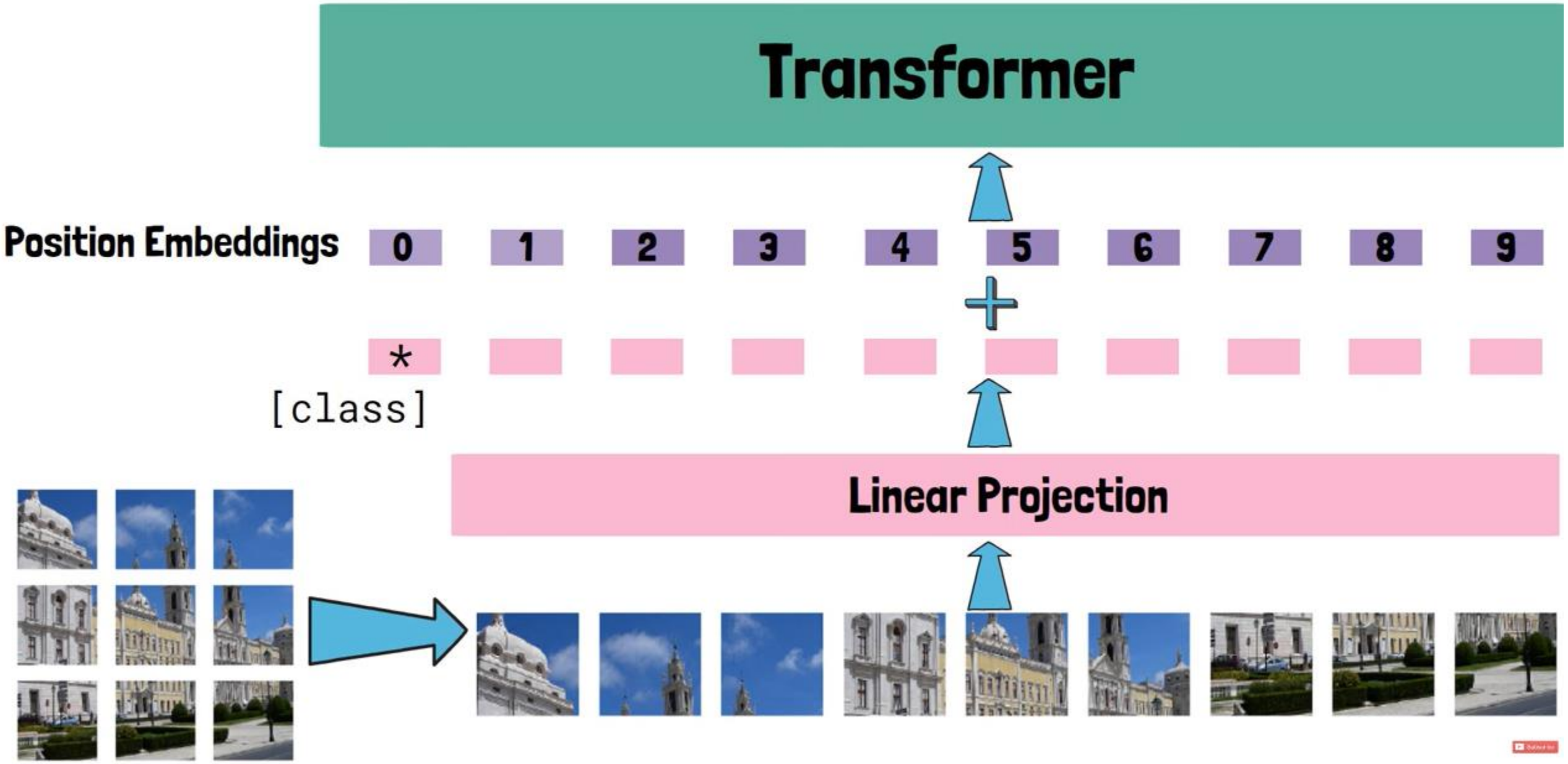
	256 X 256 image	512 X 512 image
Input sequence	65k pixels	262k pixels
Attention matrix	65k X 65k attention matrix	262k X 262k attention matrix

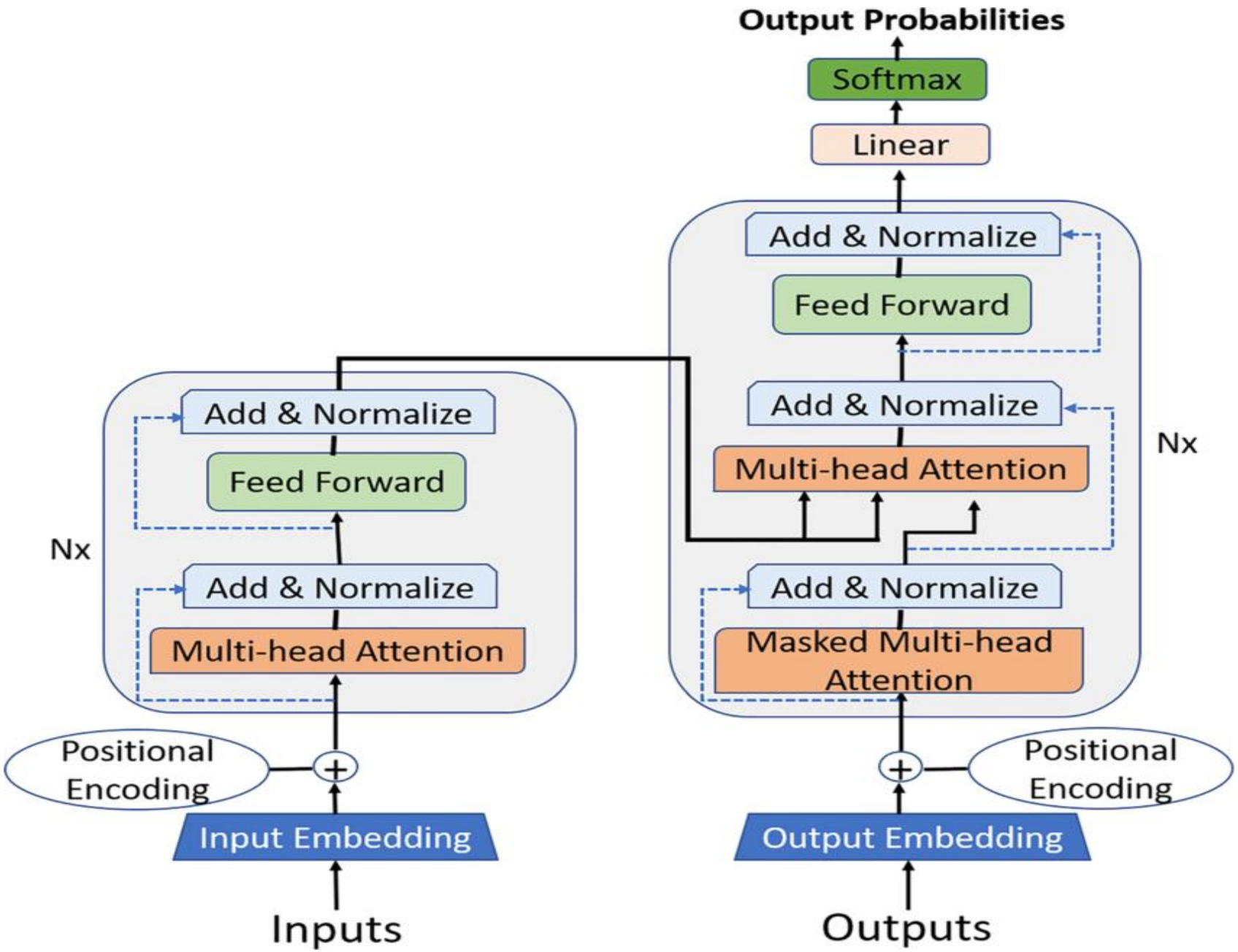


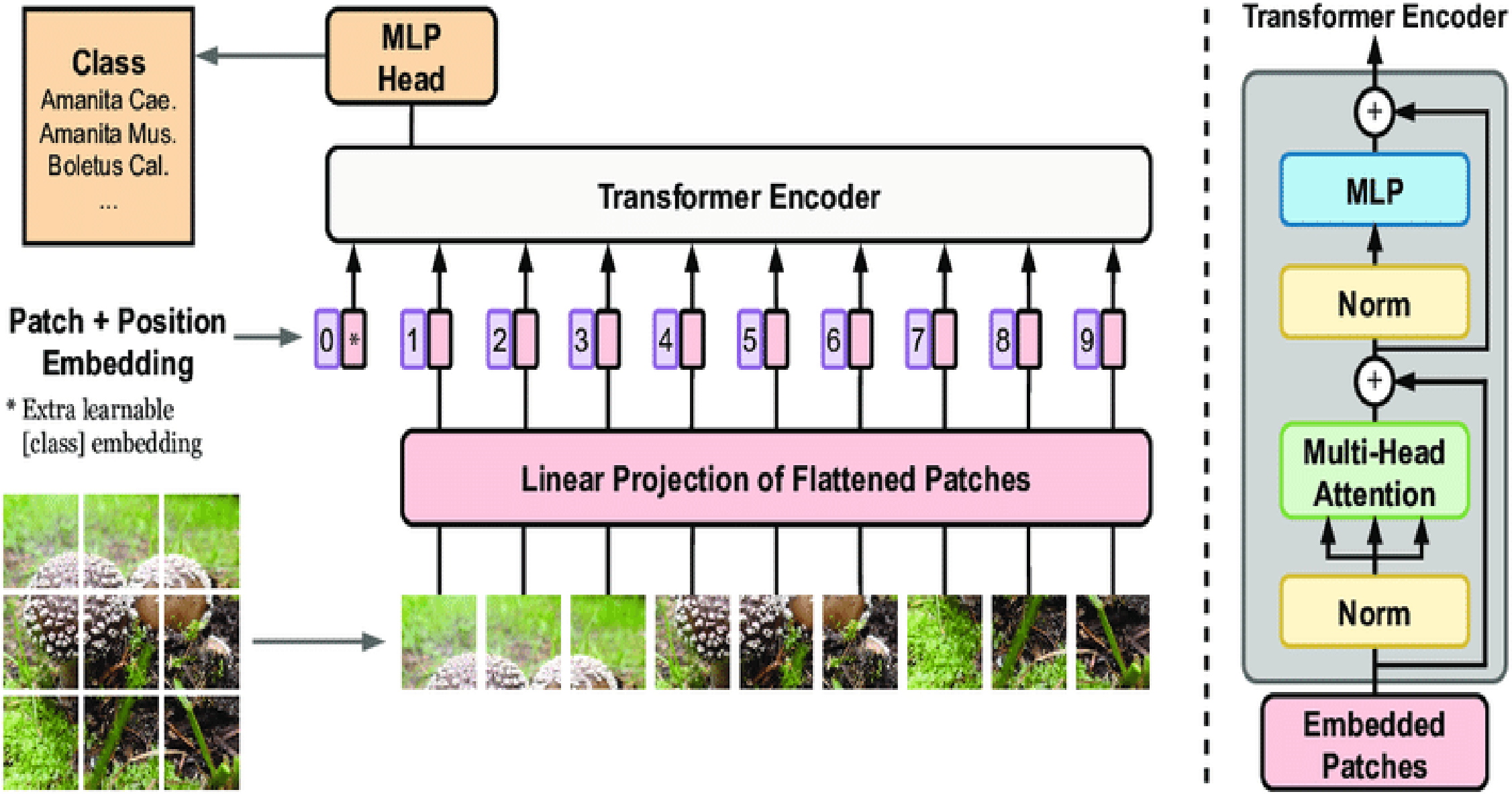
Quadratic dependency on the sequence length

- Handling images as-is results in too long sequence lenghts

How Vision Trasnformer Works?




Output Probabilities
Softmax
Linear
Add & Normalize
Feed Forward
Add & Normalize
Multi-head Attention
Add & Normalize
Masked Multi-head Attention
Nx
Add & Normalize
Feed Forward
Add & Normalize
Multi-head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs


Class
Amanita Cae.
Amanita Mus.
Boletus Cal.
...
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
MLP
Norm
Multi-Head Attention
Norm
Embedded Patches

Vision Transformers (ViTs)

224*224

patch size = 16*16

Number of patches = (224 / 16) * (224 / 16) = 14 * 14 = 196 patches

Stride =16

Vision Transformers (ViTs)

196 patches

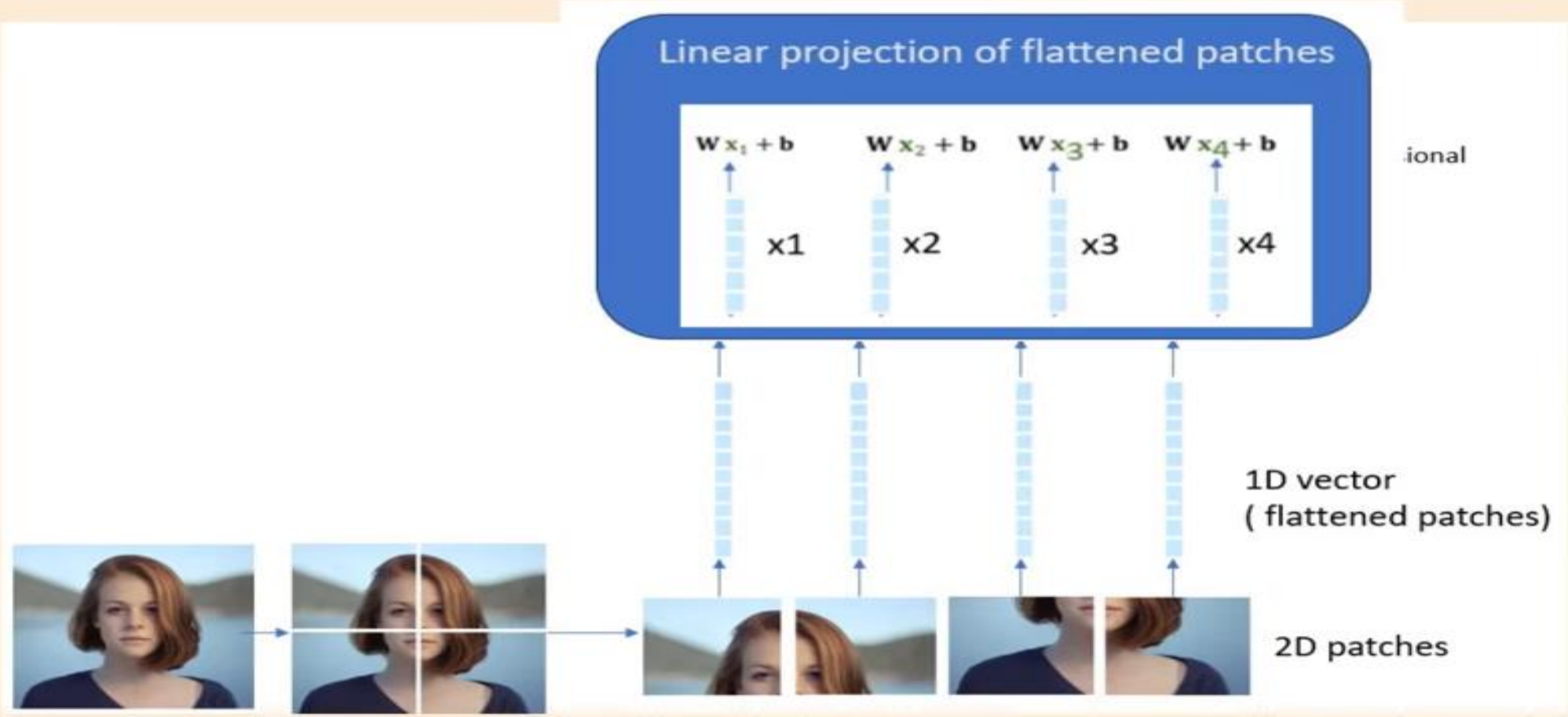
Total patches are 196 and each patch is having size of 16×16×3, assuming an RGB image

Flatten the patches from 2D to 1D

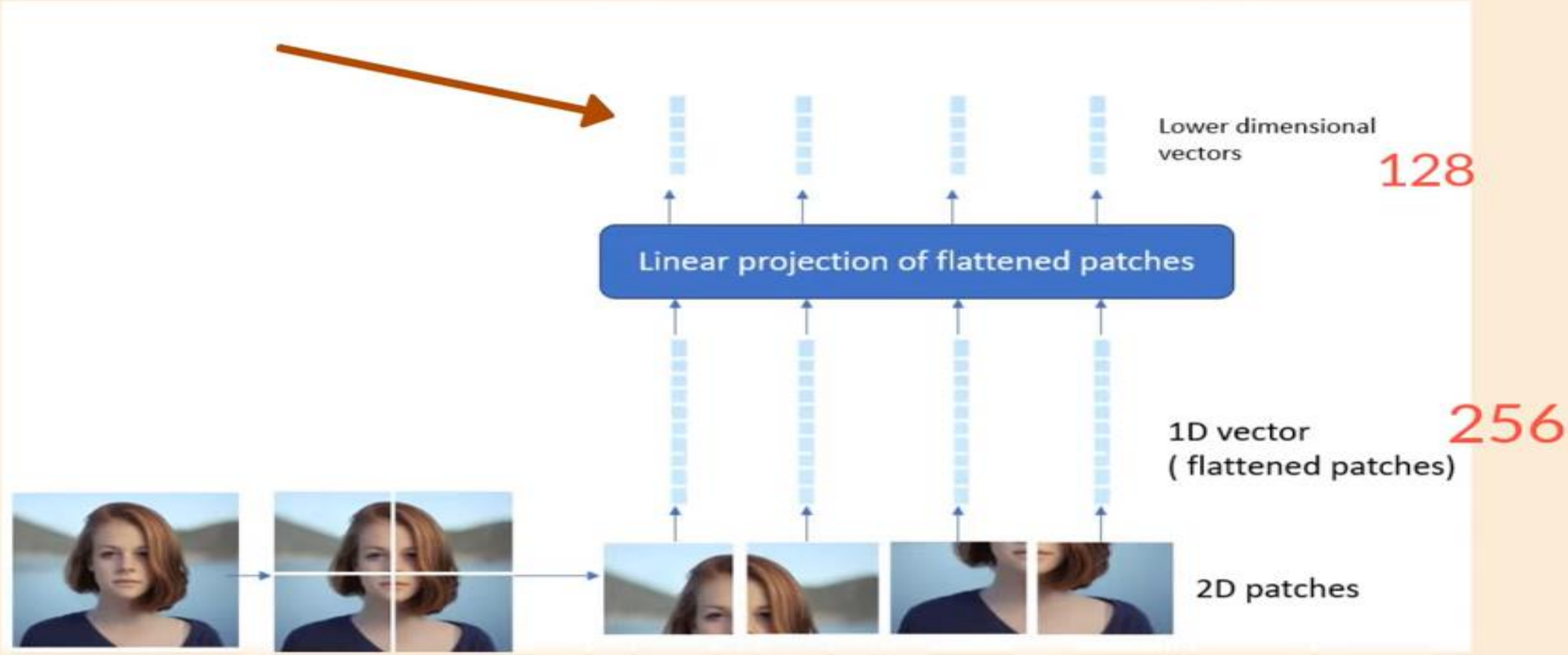
Each patch is flattened into a 1D vector

1D vector

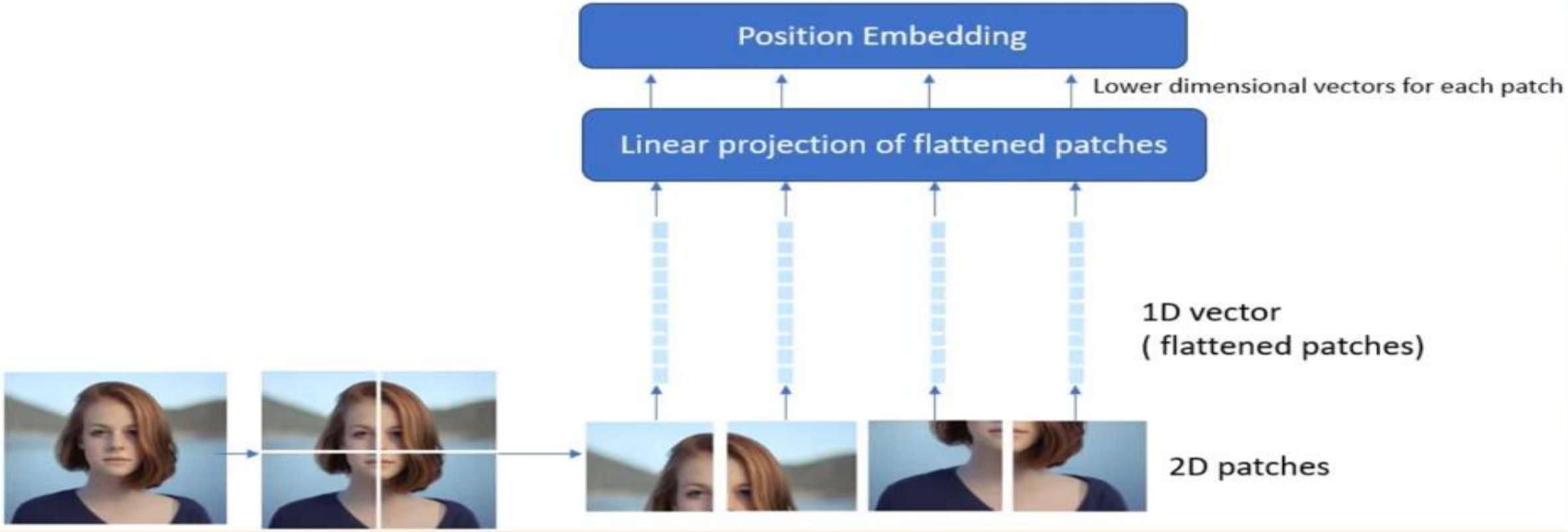
Vision Transformers (ViTs)

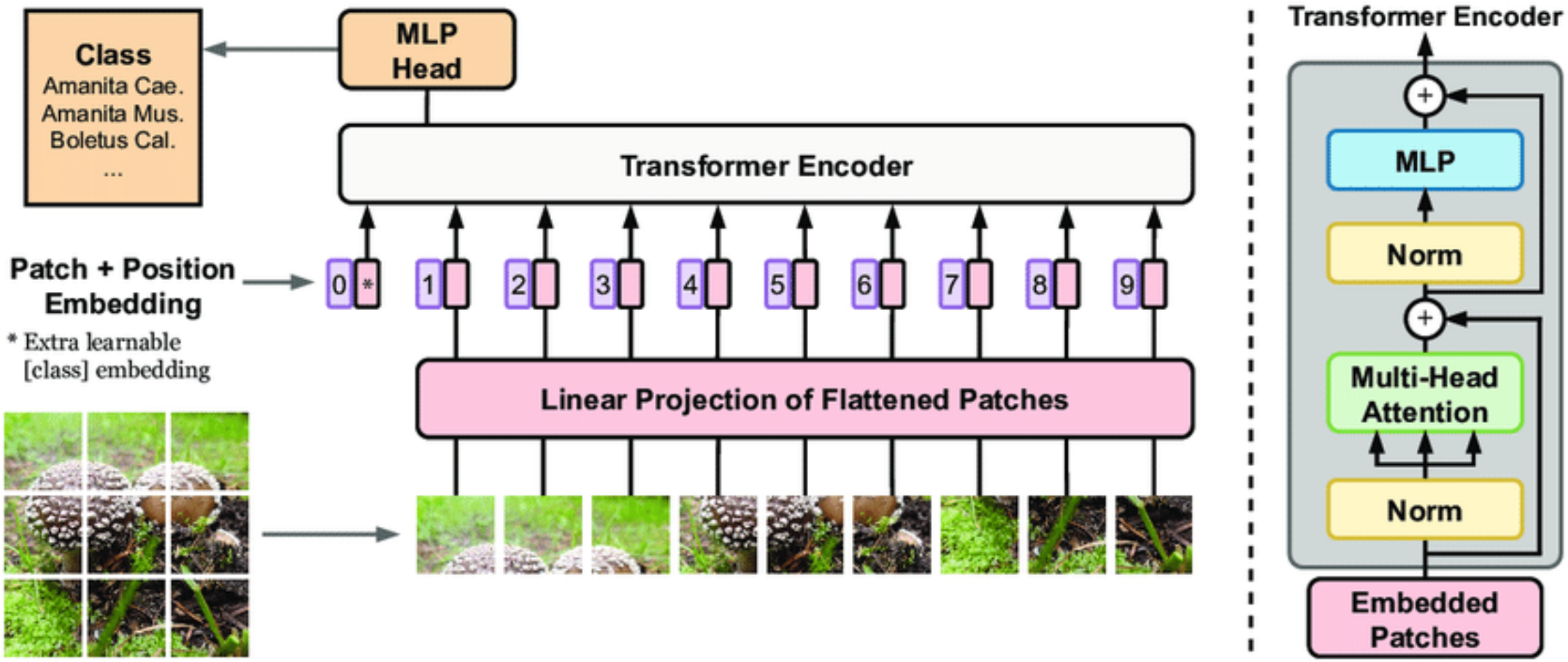


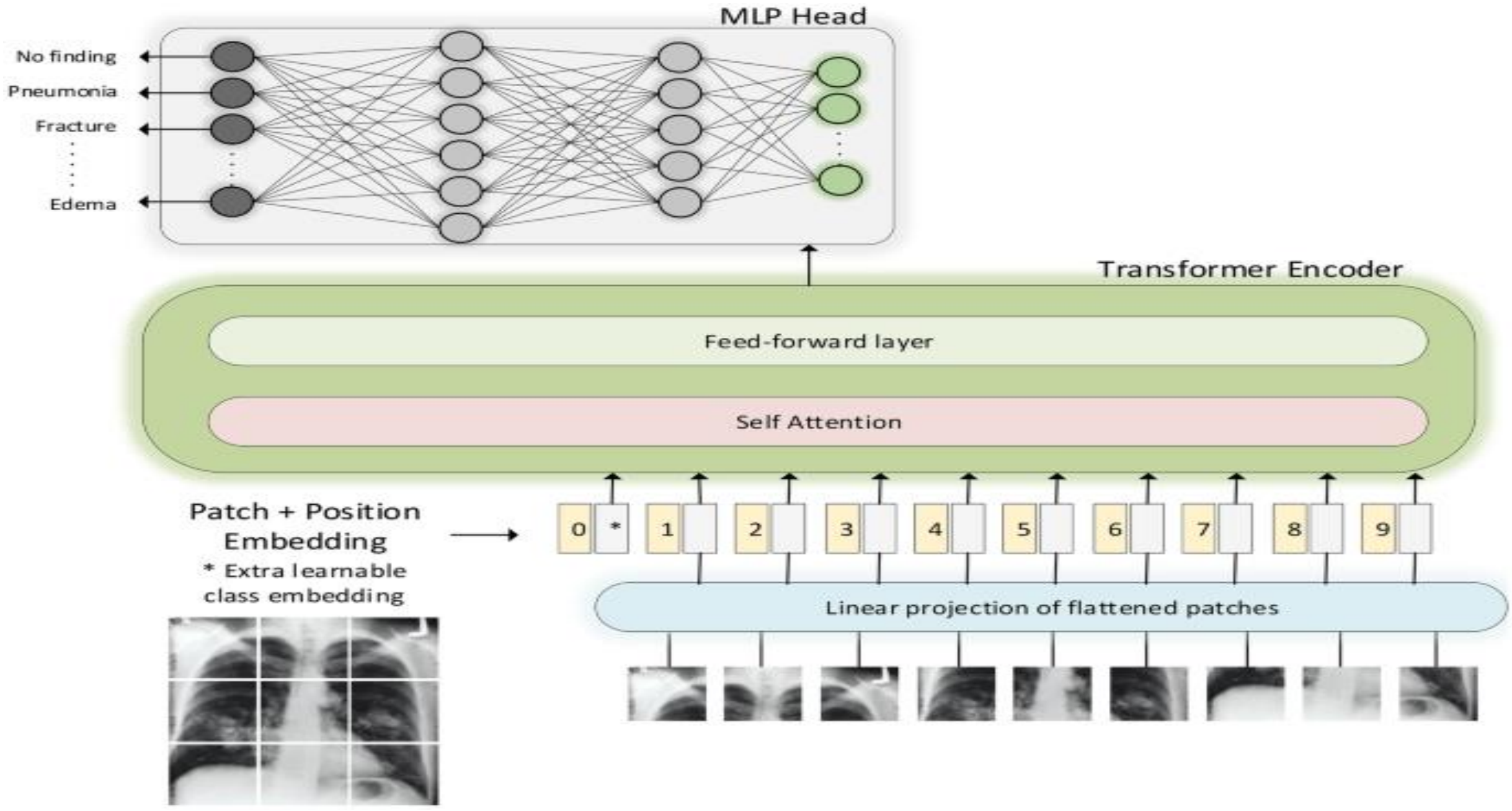
Vision Transformers (ViTs)



Vision Transformers (ViTs)




Class
Amanita Cae.
Amanita Mus.
Boletus Cal.
...
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
MLP
Norm
Multi-Head Attention
Norm
Embedded Patches


MLP Head
No finding
Pneumonia
Fracture
Edema
Transformer Encoder
Feed-forward layer
Self Attention
Patch + Position Embedding
* Extra learnable class embedding
0
*
1
2
3
4
5
6
7
8
9
Linear projection of flattened patches

Inductive Bias

CNN – high inductive bias

A lot of guidance is given to the model

Transfomer – each token can attend to any token in every layer

Much less guidance