Dividing input image into patches

Flattening the patches

Linear Projection Of Flattened Patches

* Extra learnable [class] embedding

Patch + Position Embedding

0 * 1 2 3 4 5 6 7 8 9

Transformer Encoder

Norm

Multi-Head Attention

Norm

MLP

MLP Head

Class Prediction

Creamy Paste

Diced

Grated

Juiced

Julienne

Sliced

Whole