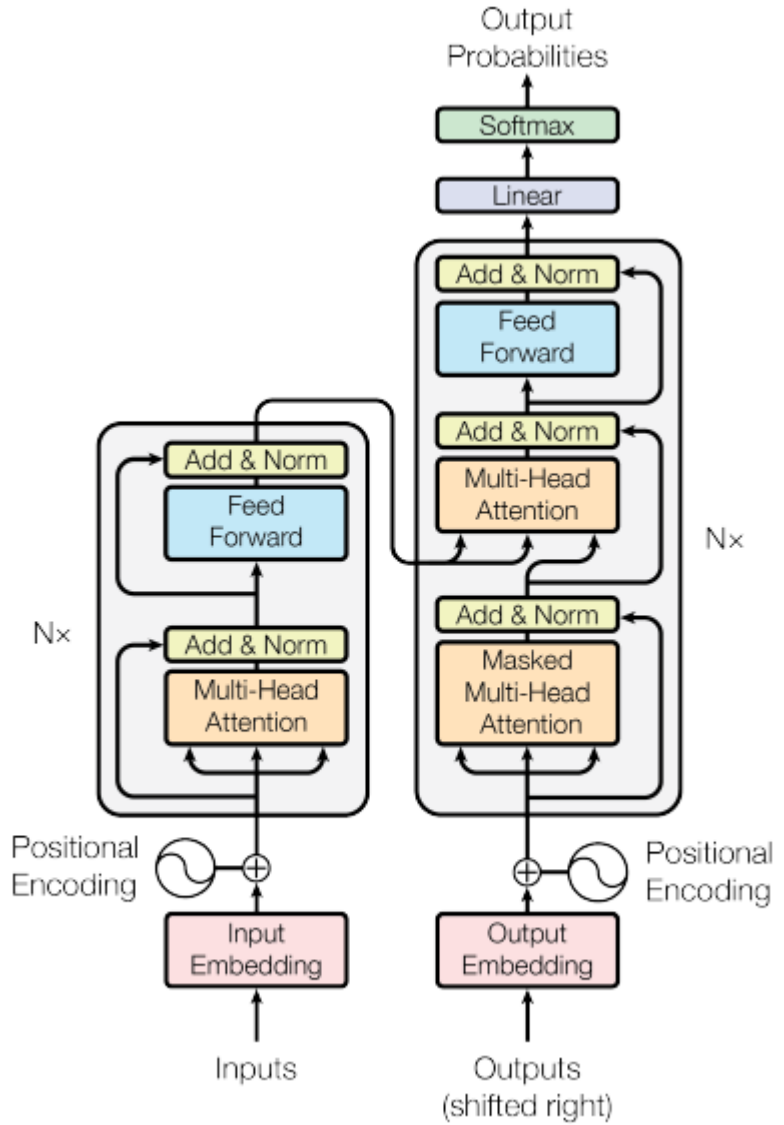
Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

Корягин Никита

Transformers Preview

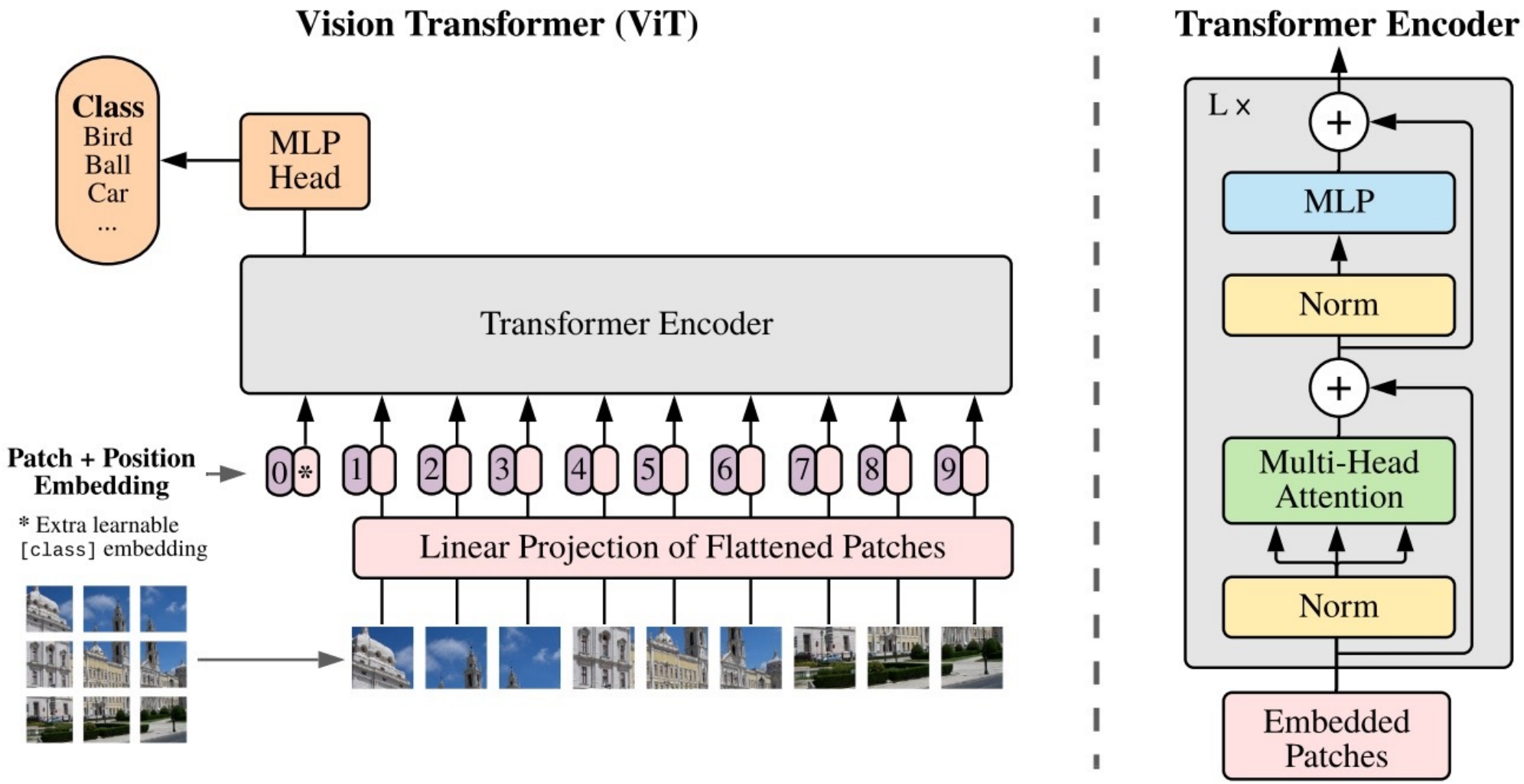
Attention Is All You Need

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\mathsf{T}}{\sqrt{d_k}}\right)V$$



Transformers Preview

Vision Transformers



Swin Transformer

Publication & Authors

Authors: Microsoft Research Asia

Published at **ICCV2021 (**won the Best Paper Award**)**

Swin Transformer

Results

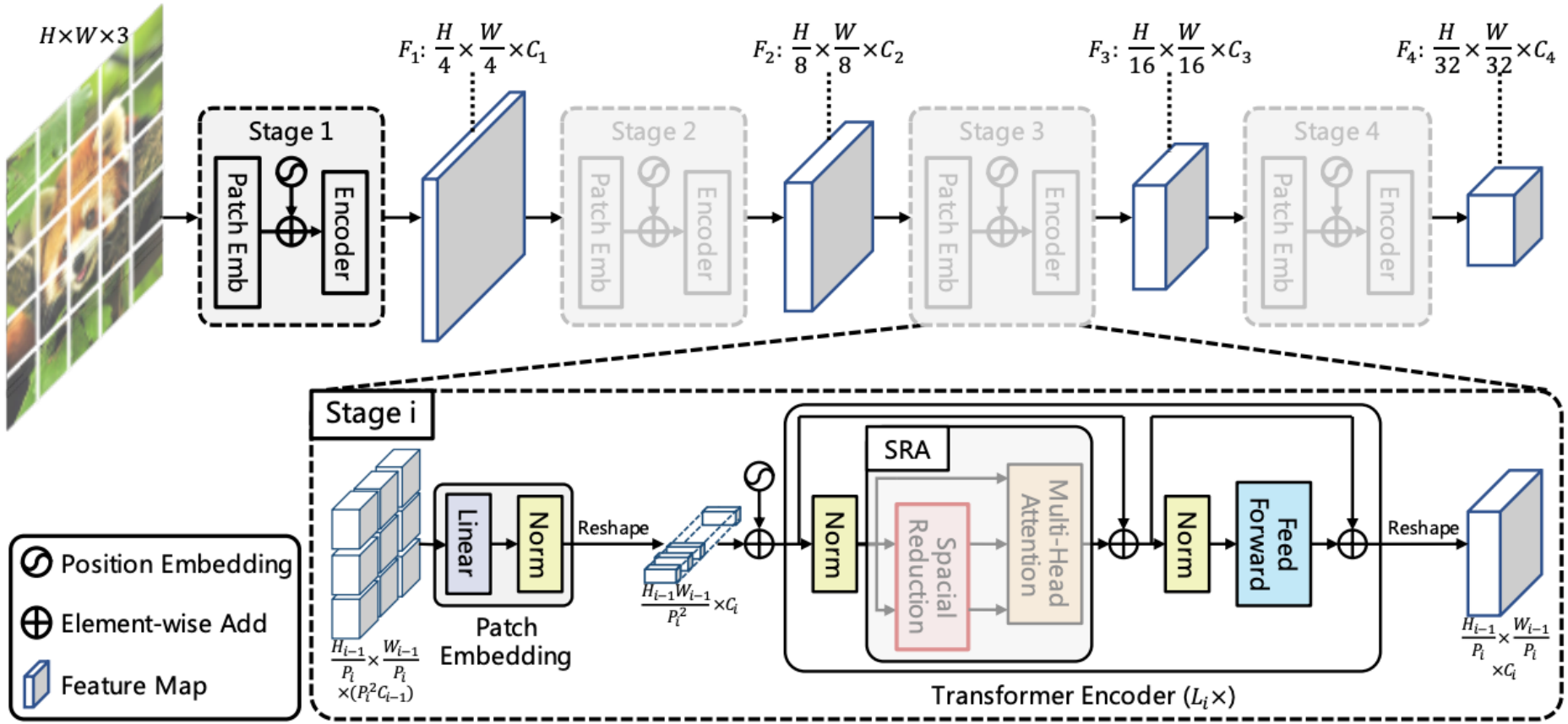
SOTA:

- COCO detection
- ADE20K semantic segmentation

Swin Transformer

Competitors

Pyramid ViT



Swin Transformer

Competitors

ViTDet

- Better quality because of MAE
- Takes longer to train and inference

Swin Transformer

Following works

Swin -2

Same authors, scaling Swin to 3 billion parameters

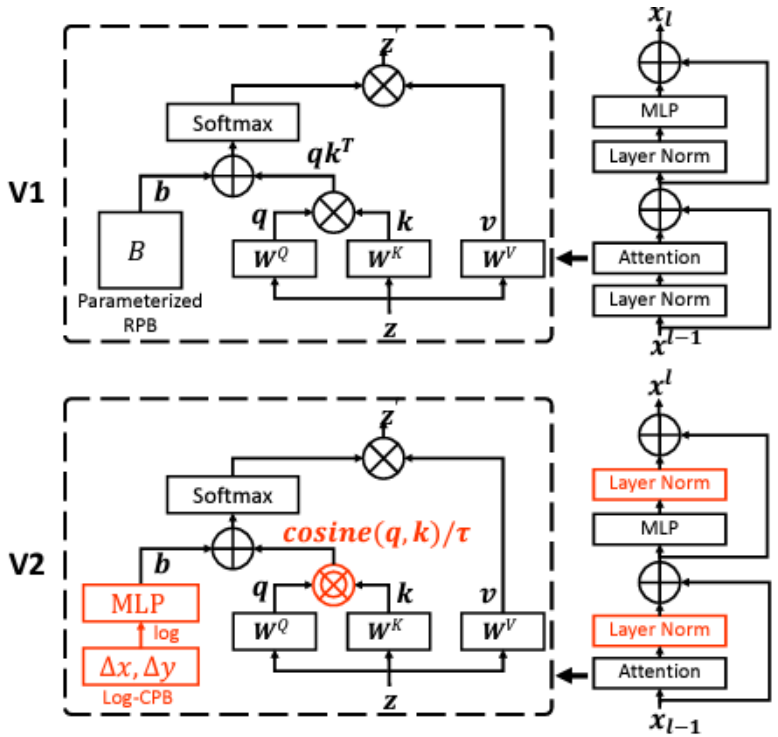


Figure 1. To better scale up model capacity and window resolution, several adaptions are made on the original Swin Transformer architecture (V1): 1) A *res-post-norm* to replace the previous *pre-norm* configuration; 2) A *scaled cosine attention* to replace the original *dot product attention*; 3) A *log-spaced continuous* relative position bias approach to replace the previous *parameterized* approach. Adaptions 1) and 2) make it easier for the model to scale up capacity. Adaption 3) makes the model to be transferred more effectively across window resolutions. The adapted architecture is named Swin Transformer V2.