Transformer and Self-Attention

hugging-face illustrated- transformer

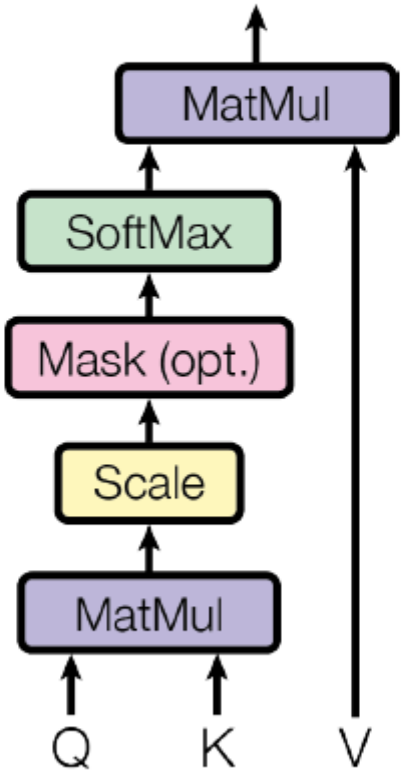
self-attention

RNN structure is hard to parallel(GPU acceleration), CNN is hard to grab long-range info.

scaled dot-product attention

- Q query
- K key
- V value

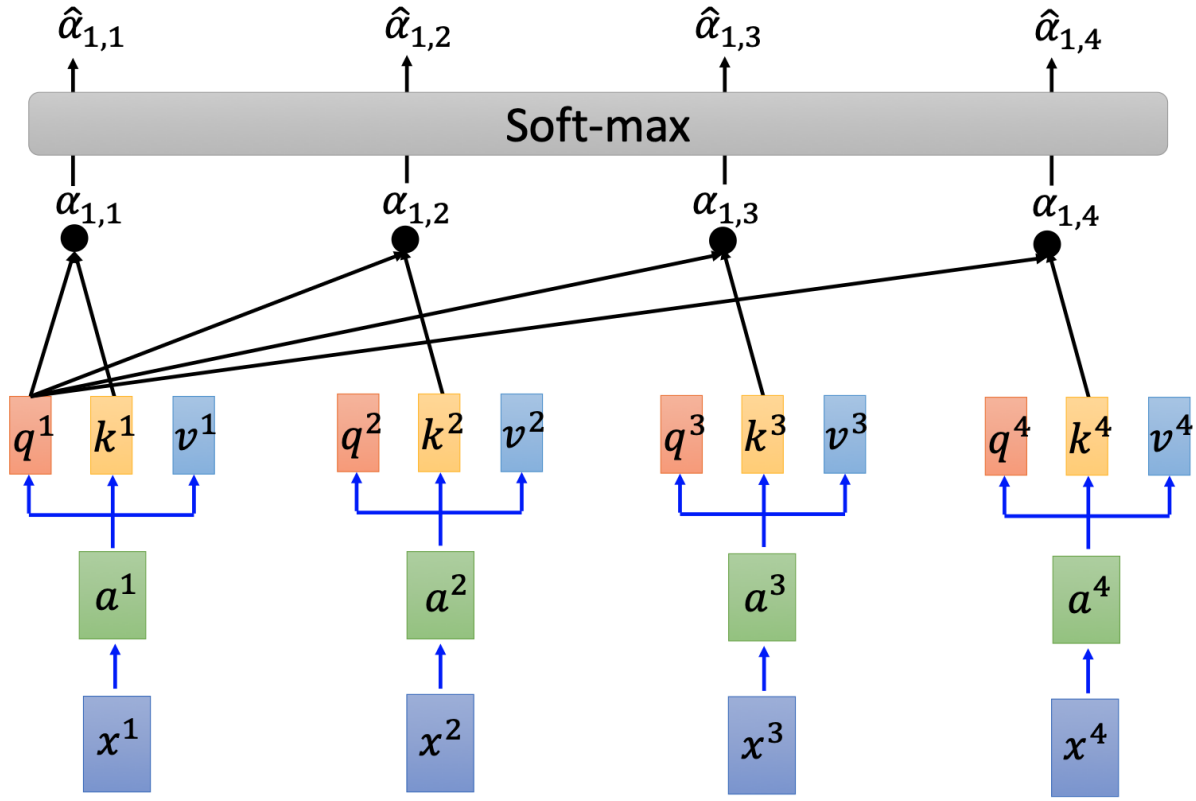
Scaled Dot-Product Attention



$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Self-attention

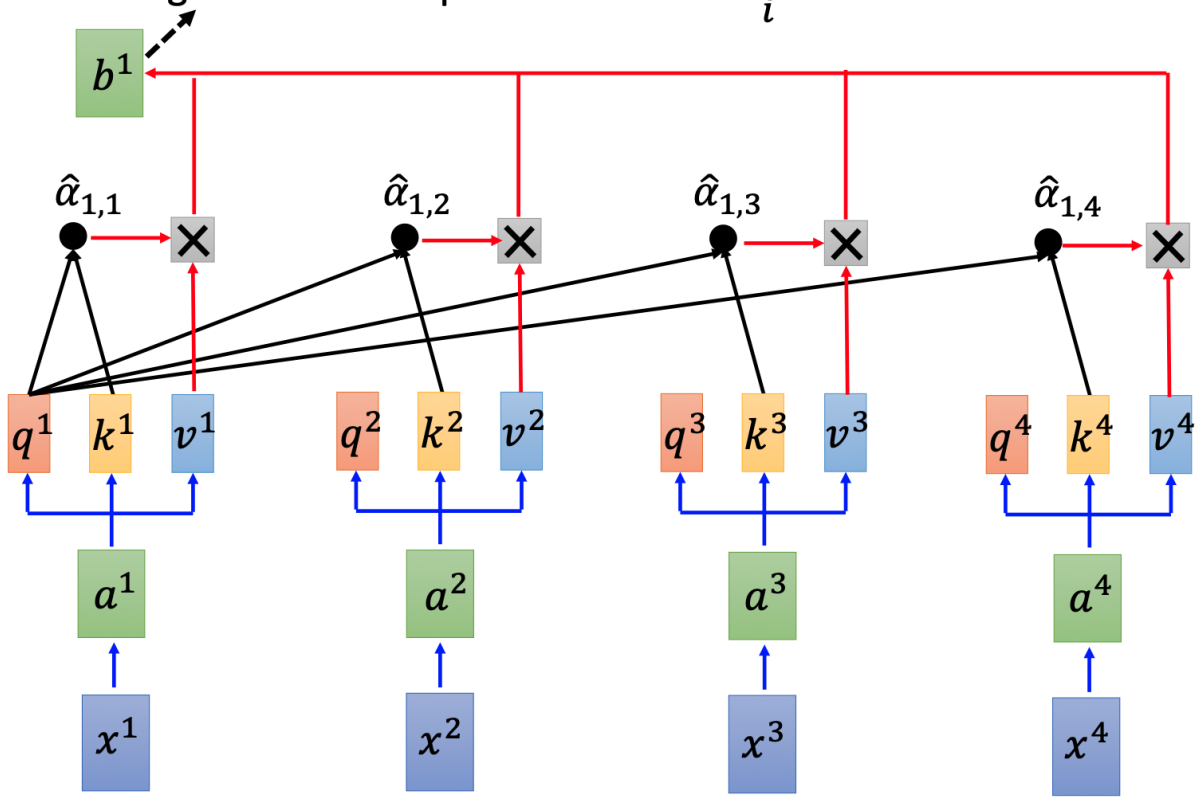
$$\hat{\alpha}_{1,i} = exp(\alpha_{1,i}) / \sum_j exp(\alpha_{1,j})$$



Self-attention

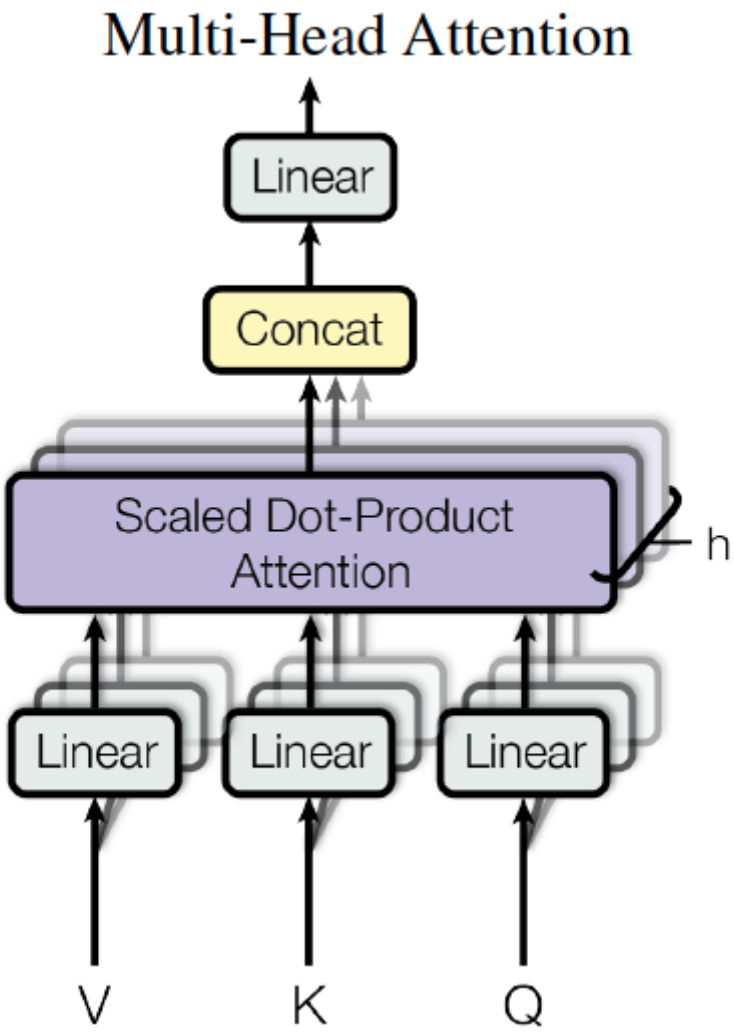
Considering the whole sequence

$$b^1 = \sum_i \hat{\alpha}_{1,i} v^i$$



multi-head attention

different head focus on different types of relevance
$MultiHead(Q, V, K) = Concat(head_1, \ldots, head_h)$ where
$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$



masked multi-head attention

predict the next word in a sequence given the previous words

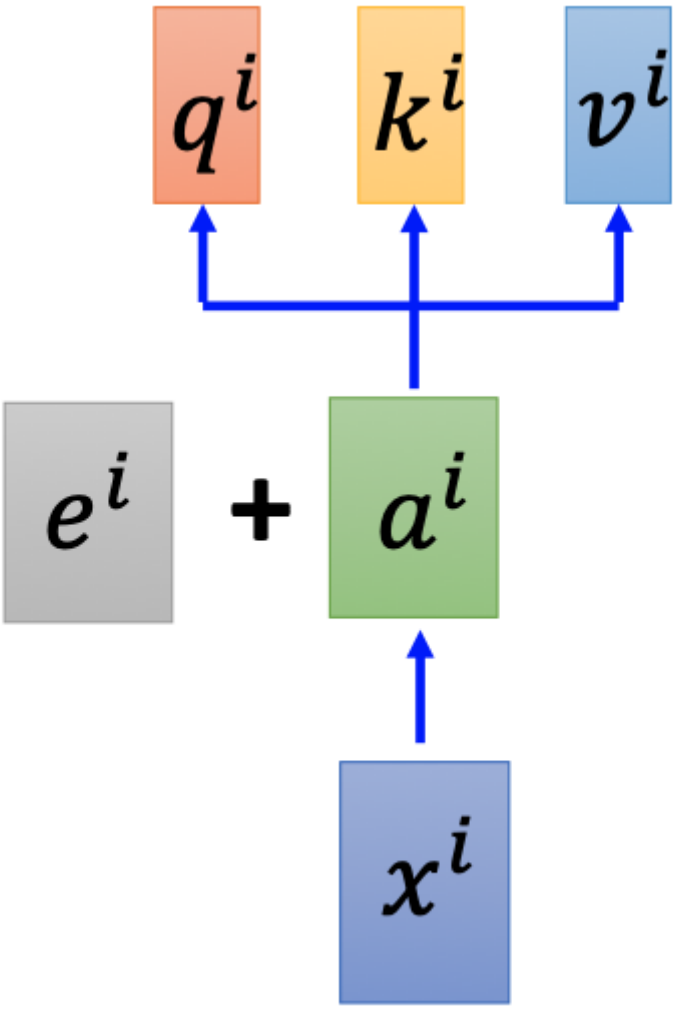
FFN

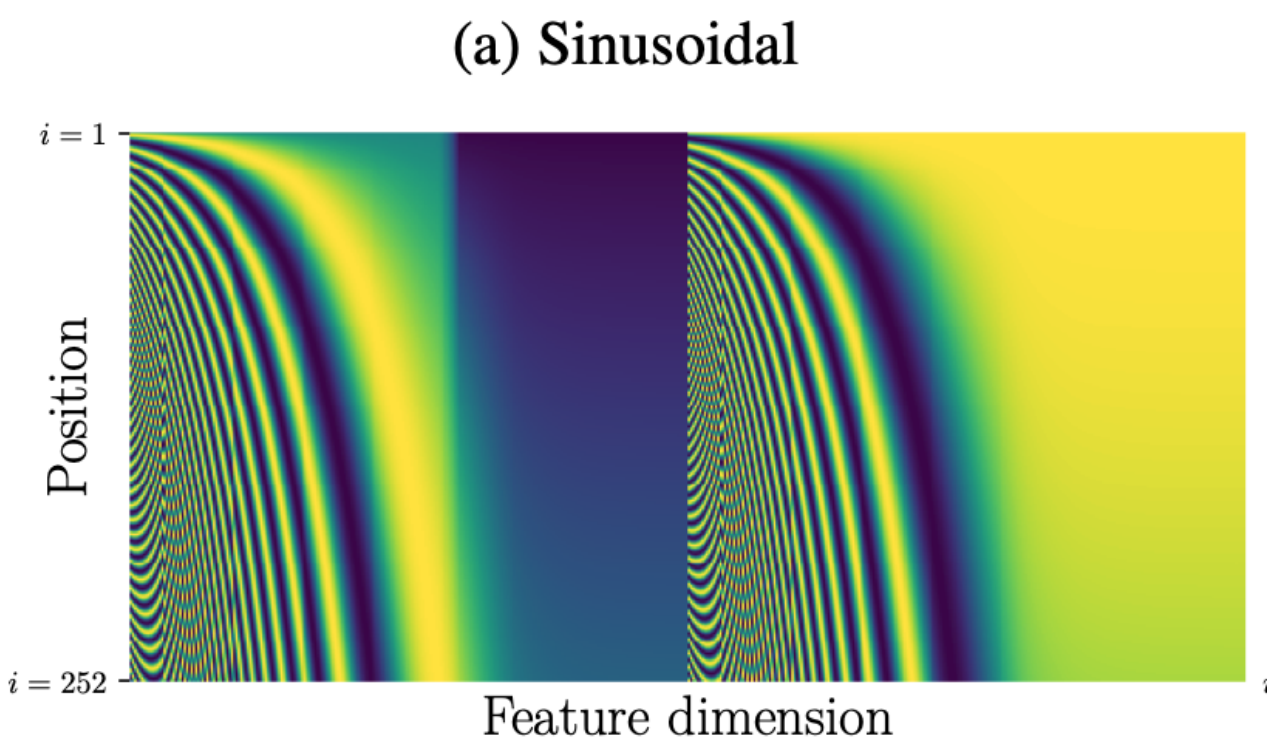
ReLU

positional embedding

self-attention have non positional info

$PE(pos, 2i) = sin(pos/10000^{2i/d_{model}})$





structure of transformer

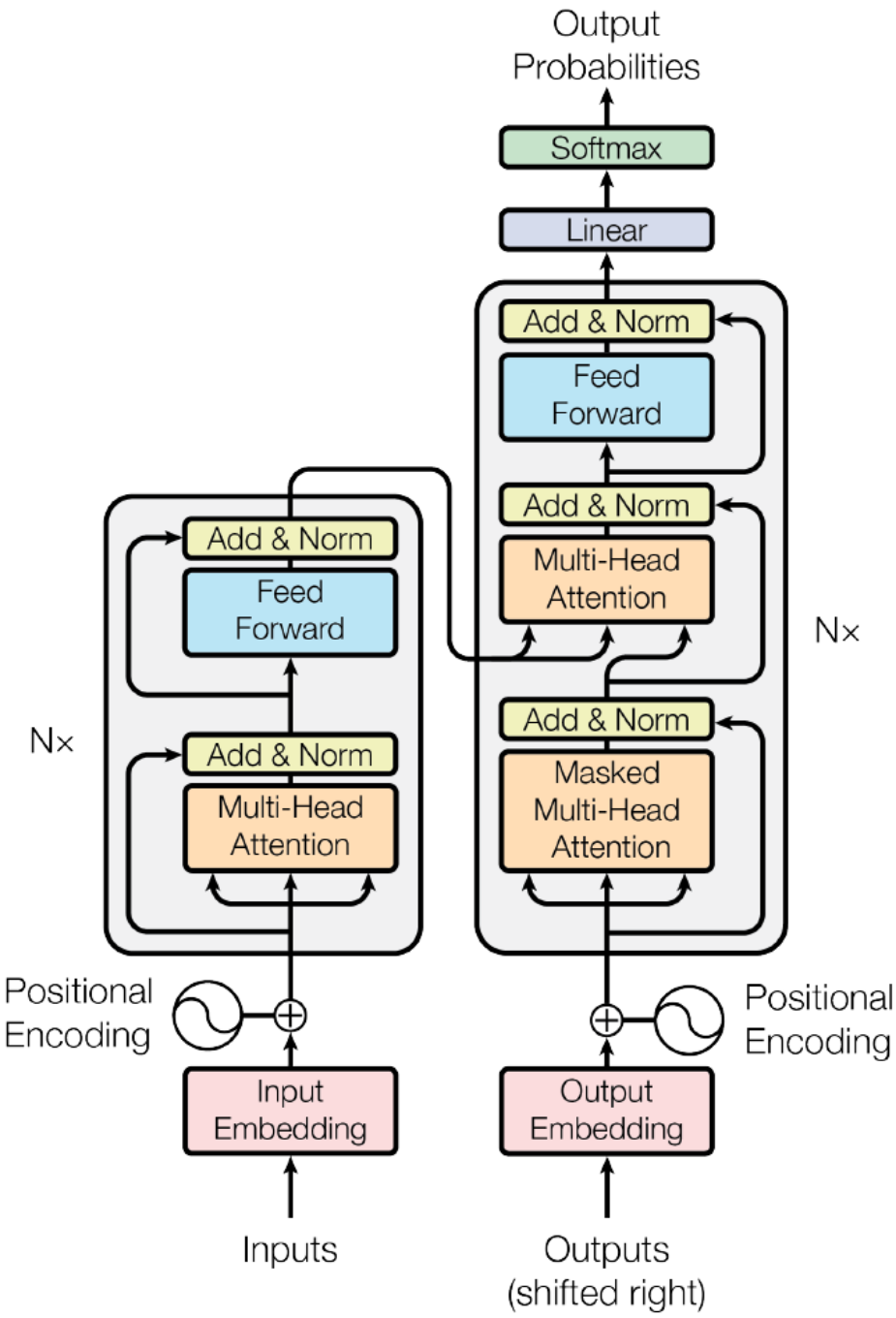


Figure 1: The Transformer - model architecture.

application

CNN

transformer is the complex version of CNN, In large model, transformer outperforms, but at smaller dataset, CNN performs better.

Vision Transformer

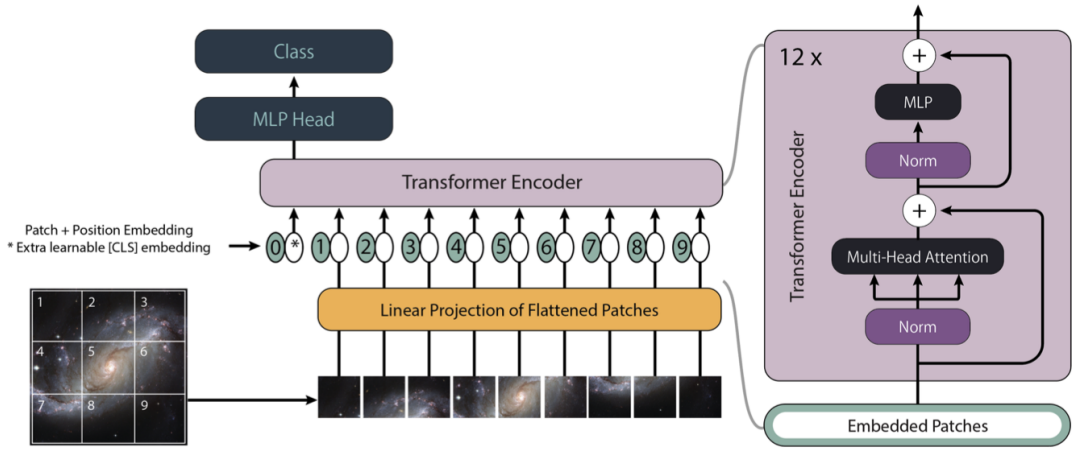


Figure 1: The architecture overview of Vision Transformer. This diagram is adapted from [34].

DEtection TRansformer

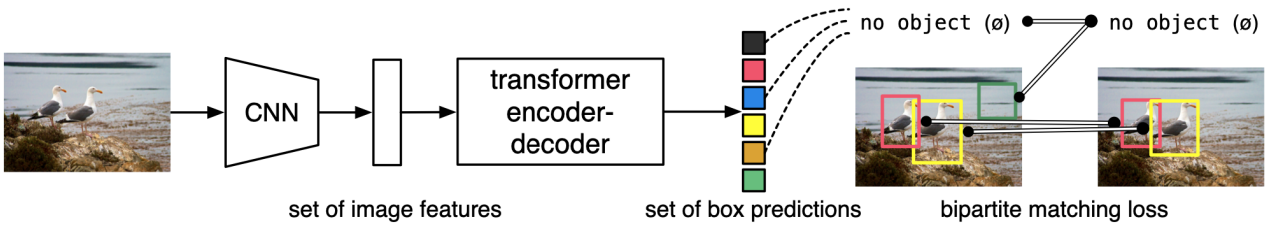


Fig. 1: DETR directly predicts (in parallel) the final set of detections by combining a common CNN with a transformer architecture. During training, bipartite matching uniquely assigns predictions with ground truth boxes. Prediction with no match should yield a "`no object`" ($\varnothing$) class prediction.

CNN backbone+ transformer⇒ detection (class, bounding box)

GPT

generative pre-trained transformer

unsupervised learning form large dataset+ fine tuning

Reference

Lee Hung-yi Machine learning lecture speech.ee.ntu.edu.tw/~tlkagk/courses/ML_2019/Lecture/Transformer (v5).pdf

Attention is all your need Learning to Encode Position for Transformer with Continuous Dynamical Model End-to-End Object Detection with Transformers Training language models to follow instructions with human feedback AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Appendix

additive attention