



视觉变压器 好耶（狗头）

Dive into Vision
Transformers

Self-attention

<https://arxiv.org/abs/1706.03762>



q : query (to match others)

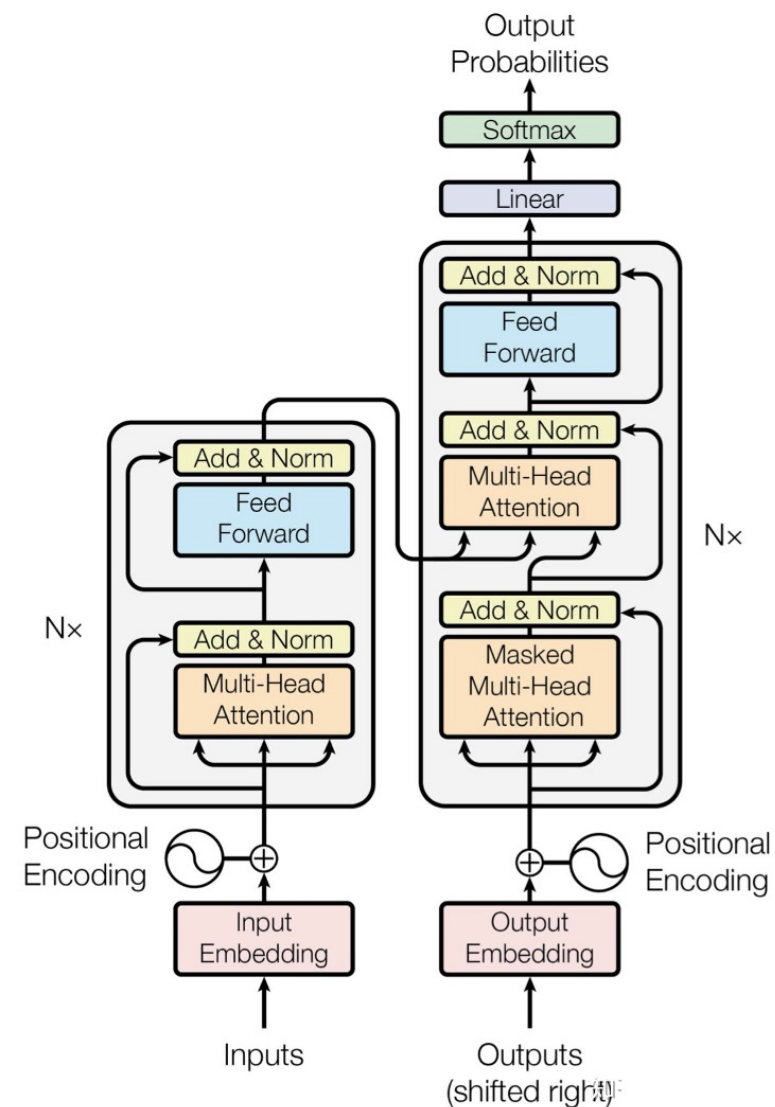
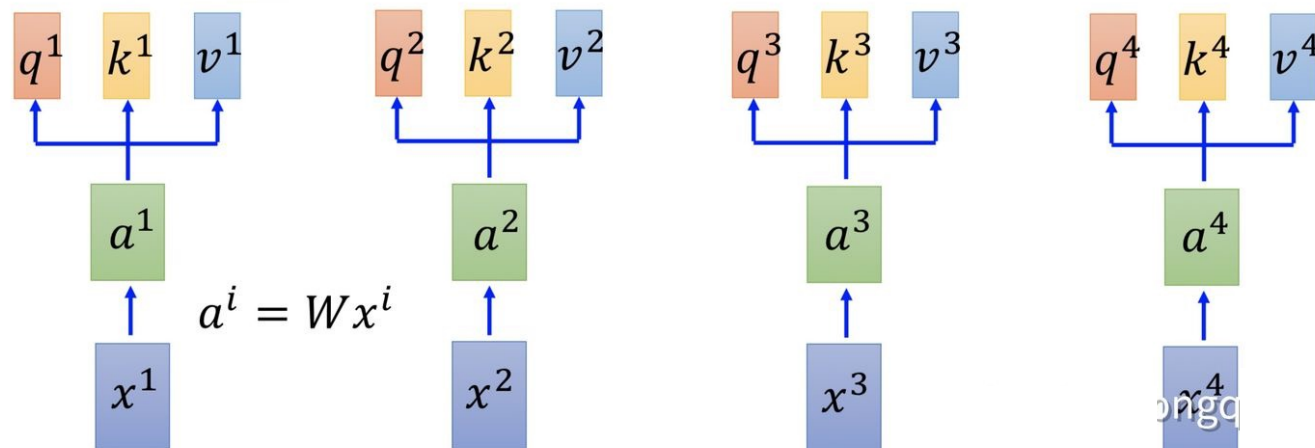
$$q^i = W^q a^i$$

k : key (to be matched)

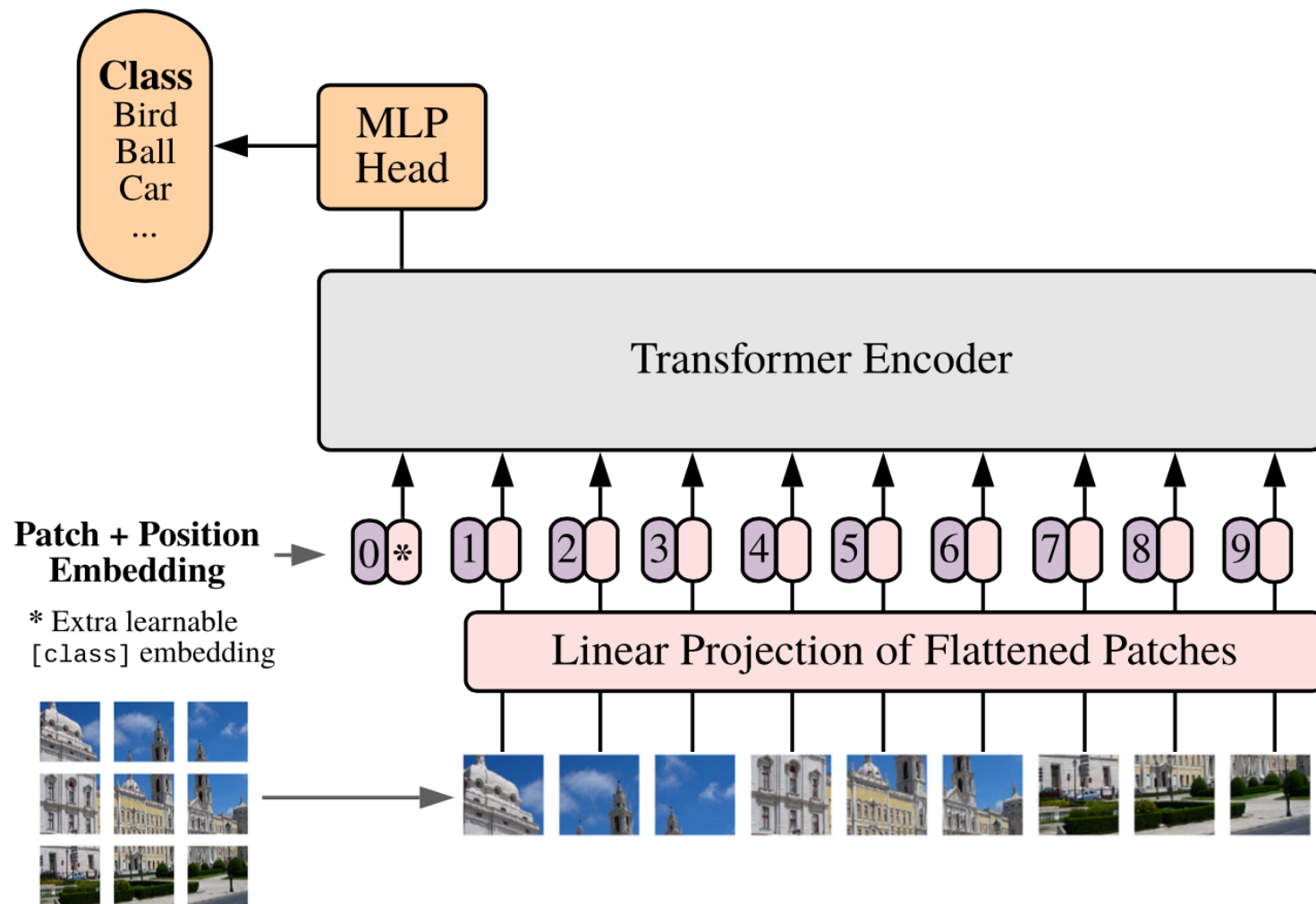
$$k^i = W^k a^i$$

v : information to be extracted

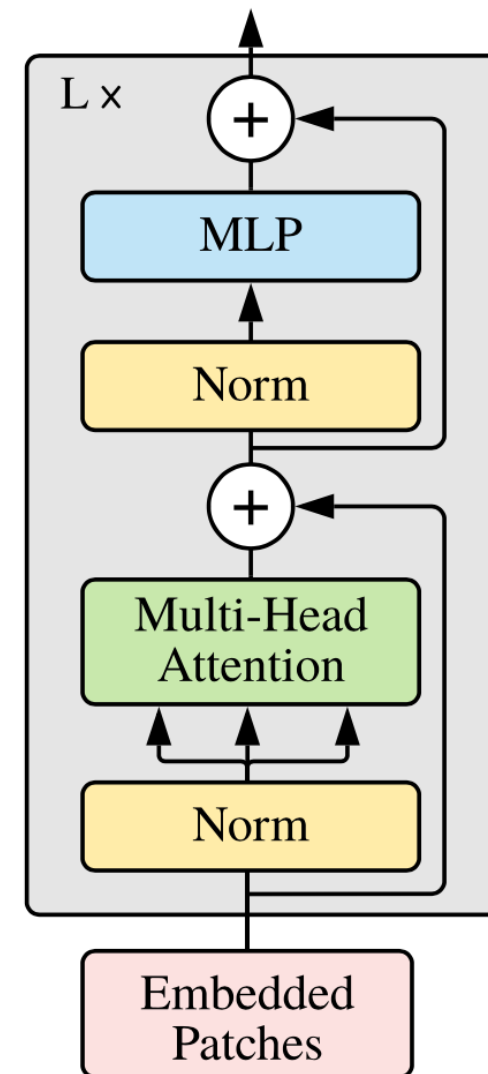
$$v^i = W^v a^i$$



Vision Transformer (ViT)



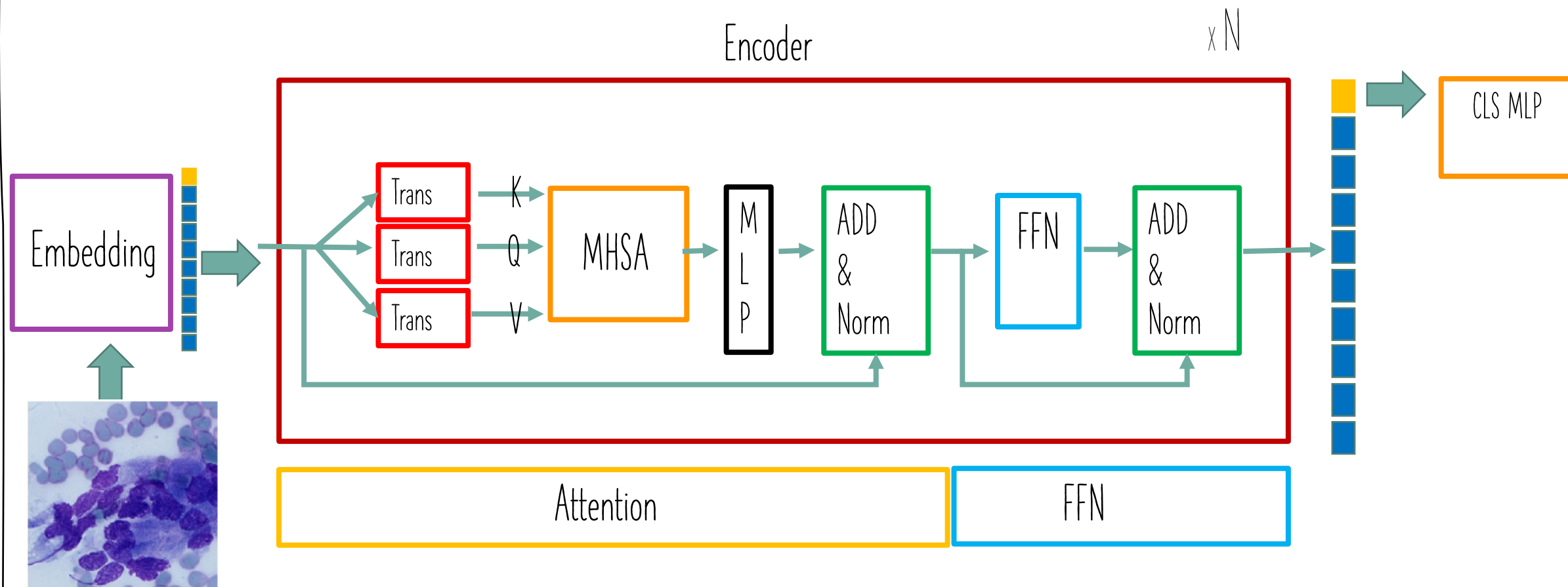
Transformer Encoder



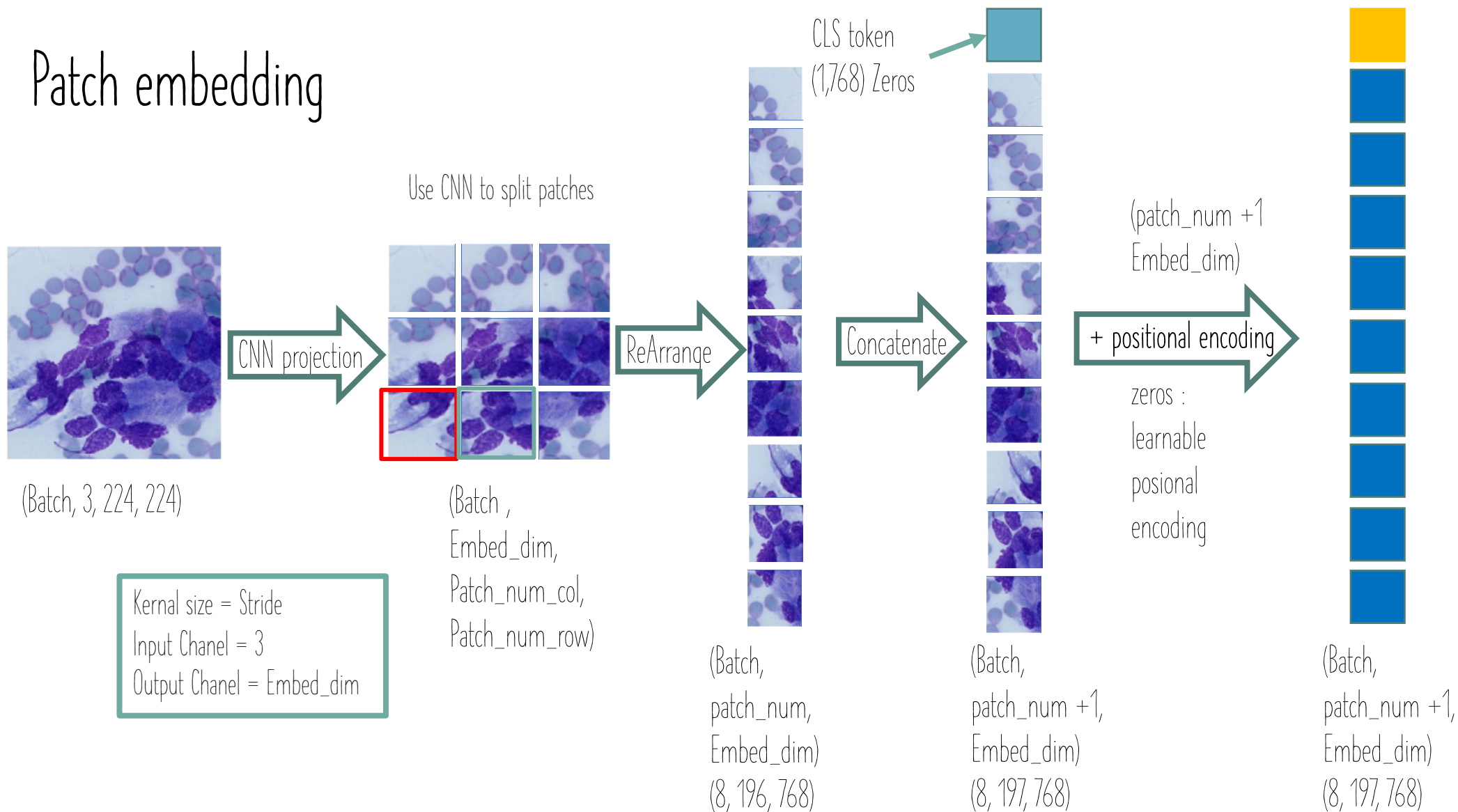


ViT Vision Transformer

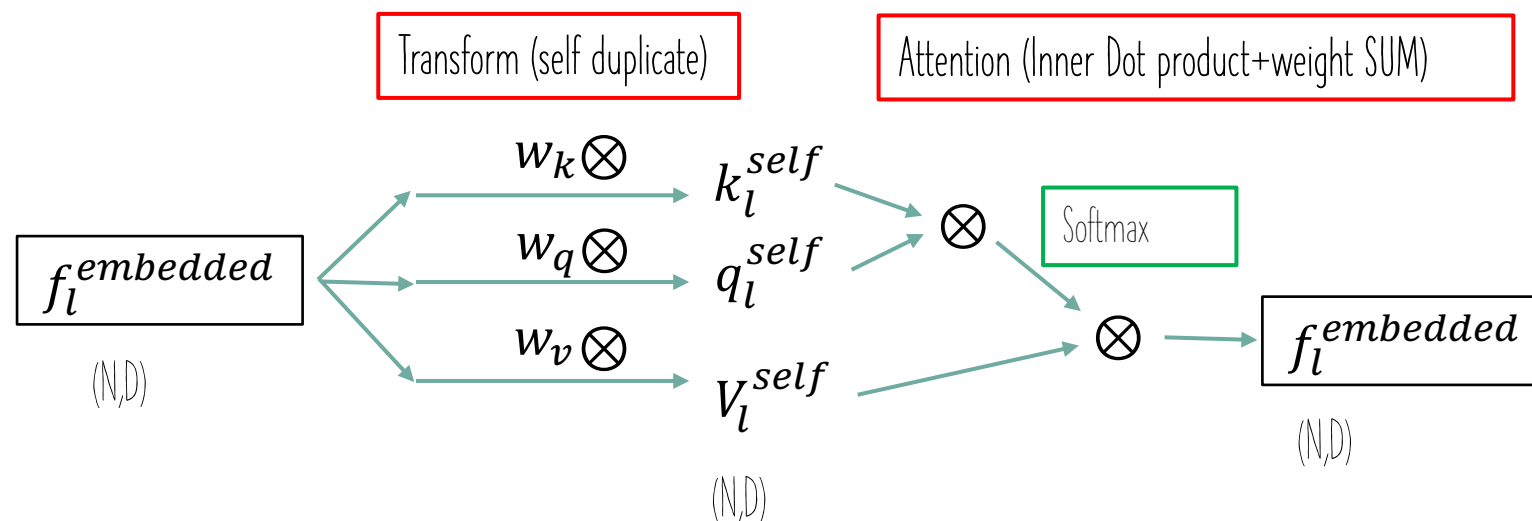
AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE
Arxiv 2010.11929



Patch embedding

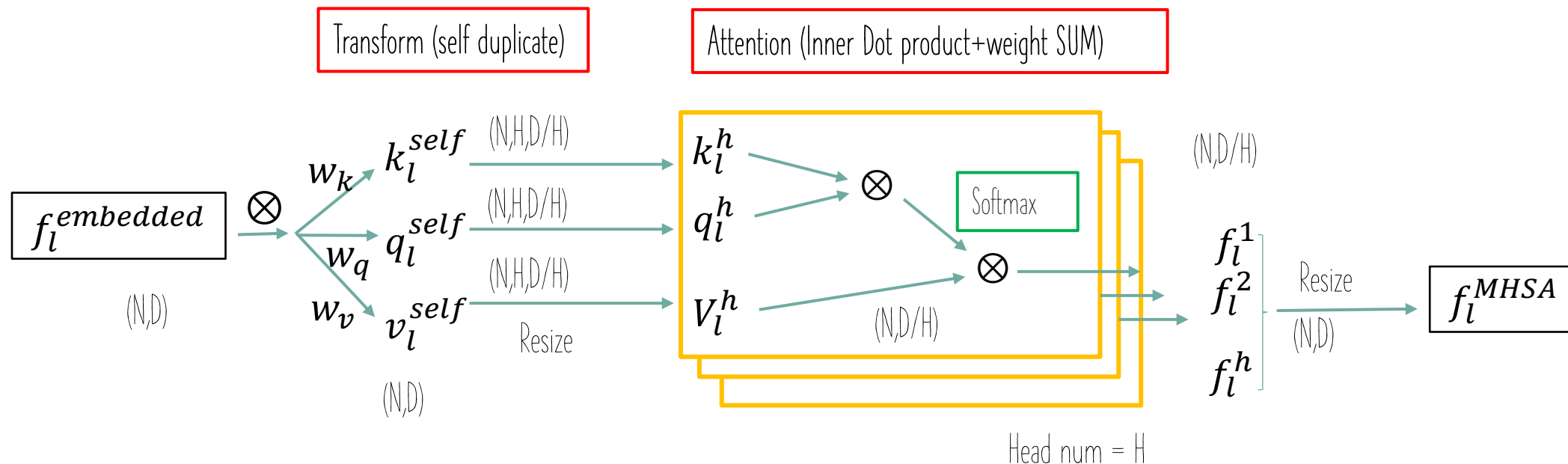


SA self attention



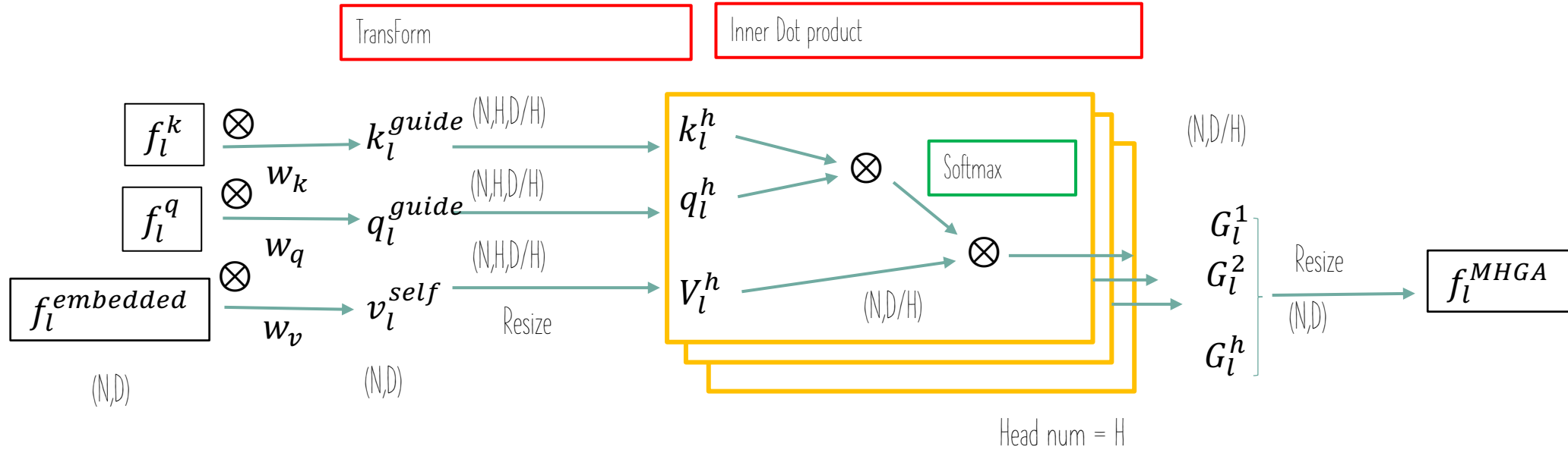
For a given embedded patch $f_l^{embedded}$, $f_l^{MHSA} = SoftMax(q_l^{self^T} \cdot k_l^{self})^T \cdot V_l^{self}$,
where $q_l^{self} = w_q \cdot f_l^{embedded}$, $k_l^{self} = w_k \cdot f_l^{embedded}$, $V_l^{self} = w_v \cdot f_l^{embedded}$.

MHSA multi-head self attention



For a given embedded patch in each head $f_l^{embedded}$, $f_l^{MHSA} = \text{SoftMax}(q_l^{self^T} \cdot k_l^{self})^T \cdot V_l^{self}$,
 where $q_l^{self} = w_q \cdot f_l^{embedded}$, $k_l^{self} = w_k \cdot f_l^{embedded}$, $V_l^{self} = w_v \cdot f_l^{embedded}$.

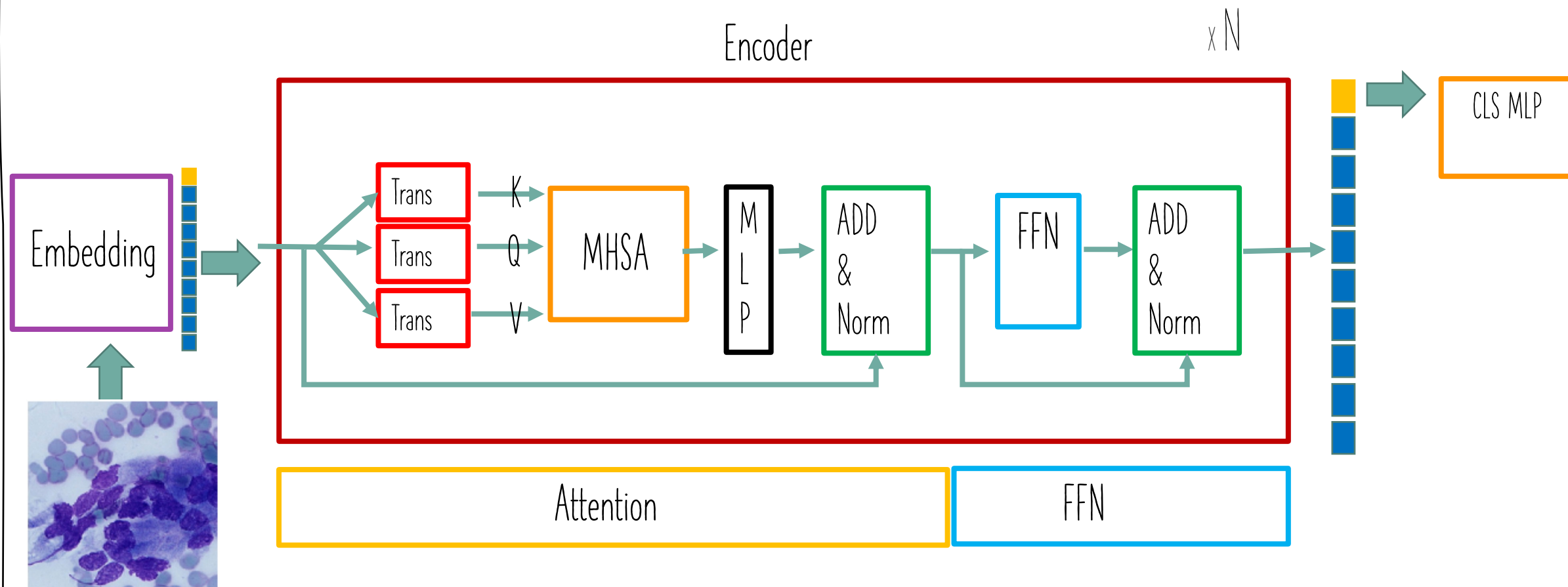
MHGA multi-head guided attention



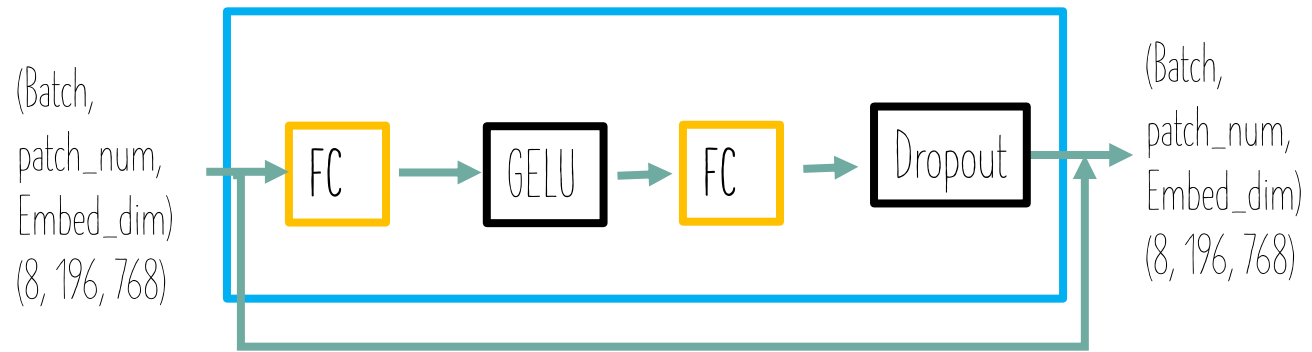
For a given embedded patch in each head $f_l^{embedded}$, $f_l^{MHSA} = \text{SoftMax}(q_l^{self^T} \cdot k_l^{self})^T \cdot V_l^{self}$,
 where $q_l^{self} = w_q \cdot f_l^{embedded}$, $k_l^{self} = w_k \cdot f_l^{embedded}$, $V_l^{self} = w_v \cdot f_l^{embedded}$.

ViT Vision Transformer

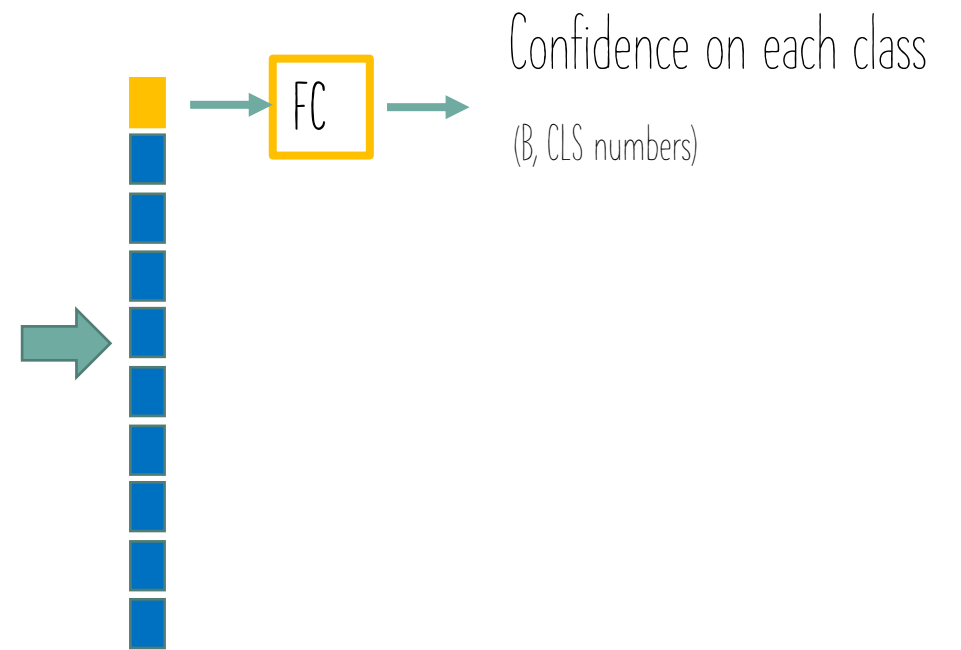
AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE
Arxiv 2010.11929



FFN



CLS MLP

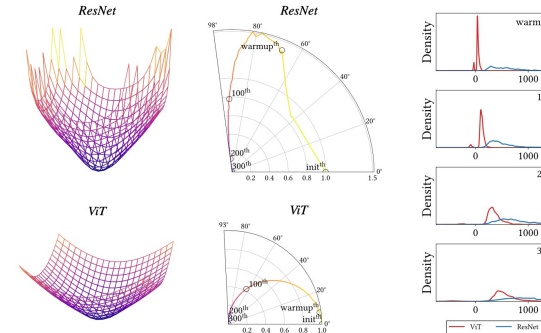


POINTS TO DECLEAR

1. Who's learning the most of the hard works ?
2. Who's the heaviest ?
3. Why its soooo stirring ?
4. A slight view of the future ? (Nay, actually my works lol)

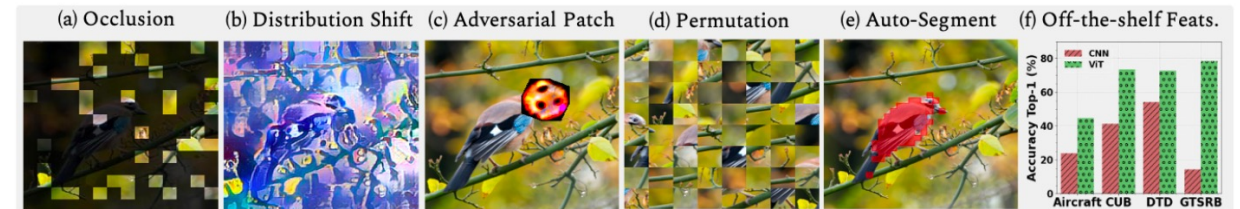
Hard Math

HOW DO VISION TRANSFORMERS WORK?
Arxiv 2202.06709



Inspiring Experiments

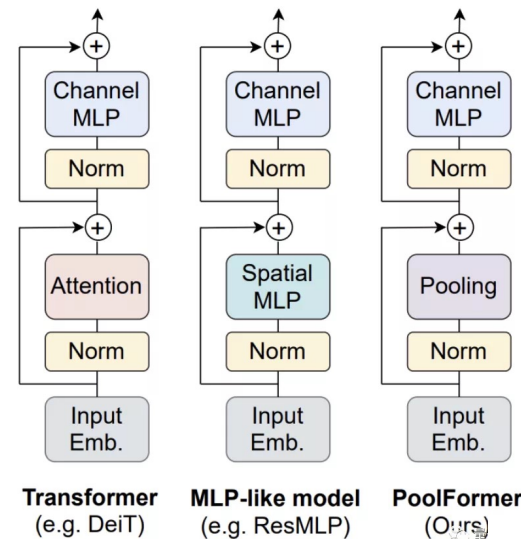
NeurPS2021 Intriguing Properties of Vision Transformers
Arxiv 2105.10497



Inspiring Model play

MetaFormer is Actually What You Need for Vision
Poolformer Arxiv 2111.11418

Patches Are All You Need
<https://openreview.net/pdf?id=TVHS5Y4dNvM>



Under review as a conference paper at ICLR 2022

CONVOLUTIONS ATTENTION MLPs
PATCHES ARE ALL YOU NEED? 🙋

Anonymous authors
Paper under double-blind review



Model structure

Learning strategy

Initiallization

MSHT: Multi-stage Hybrid Transformer for the ROSE Image Analysis of Pancreatic Cancer

<https://github.com/sagizty/Multi-Stage-Hybrid-Transformer>

Arxiv: 2112.13513

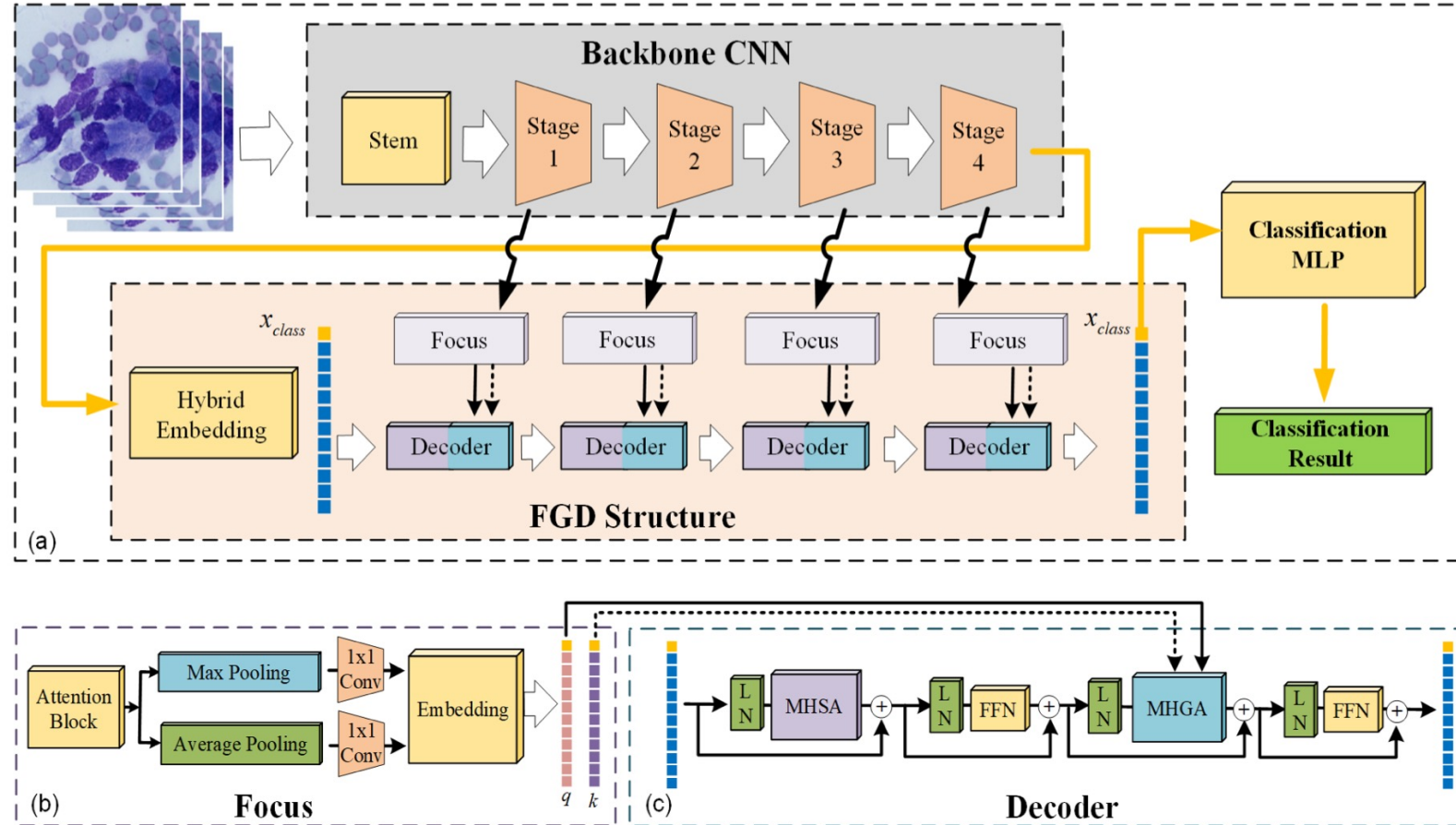


Fig. 1. The architecture of the proposed Multi-stage Hybrid Transformer (MSHT) model for the classification of ROSE images of pancreatic cancer. (a) The architecture of MSHT. (b) The focus block of the FGD structure (c) The decoder of the FGD structure. MHSA denotes multi-head self-attention, MHGA denotes multi-head guided-attention, LN denotes layer norm block, FFN denotes the feed-forward network, and MLP denotes multi-layer perceptron.

Thanks

PPT: <https://github.com/sagizty/Insight/blob/main/Dive%20into%20Transformer.pdf>

Code: https://github.com/sagizty/Insight/blob/main/ZTY_Torch/ViT.py