Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet

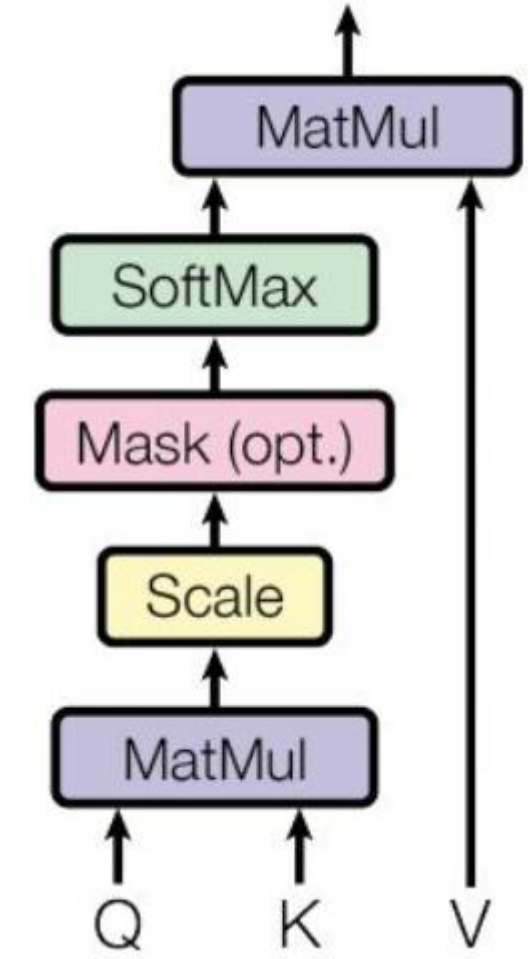
预备知识

1.self-attention

2.transformer

3VIT

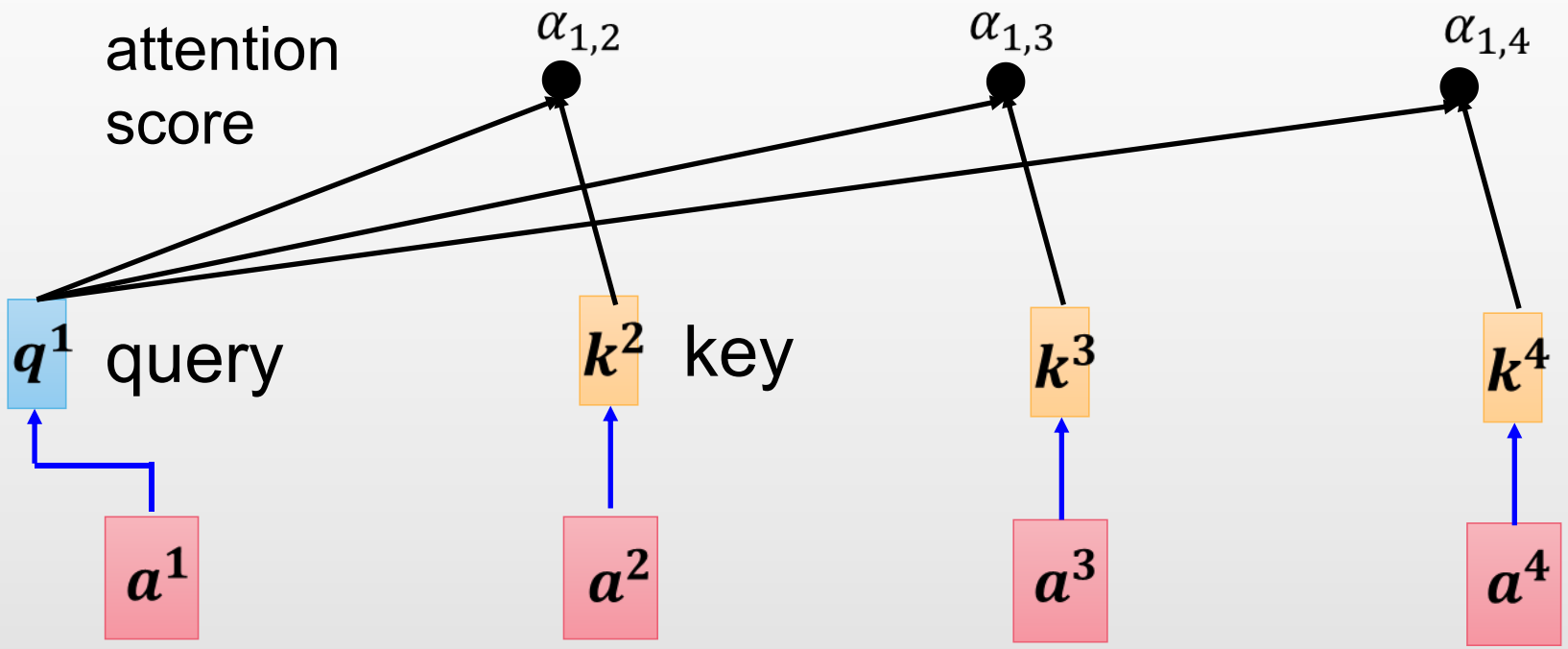
SELF-ATTENTION



- 2005年，Bahdanau等人在论文《Neural Machine Translation by Jointly Learning to Align and Translate》
- Google 机器翻译团队在NIPS 2017上发表的《Attention is All You Need》
- GoogleMind 2014年发表《Recurrent Models of Visual Attention》

SELF-ATTENTION

$$\alpha_{1,2} = \boldsymbol{q^1} \cdot \boldsymbol{k^2} \qquad \alpha_{1,3} = \boldsymbol{q^1} \cdot \boldsymbol{k^3} \qquad \alpha_{1,4} = \boldsymbol{q^1} \cdot \boldsymbol{k^4}$$

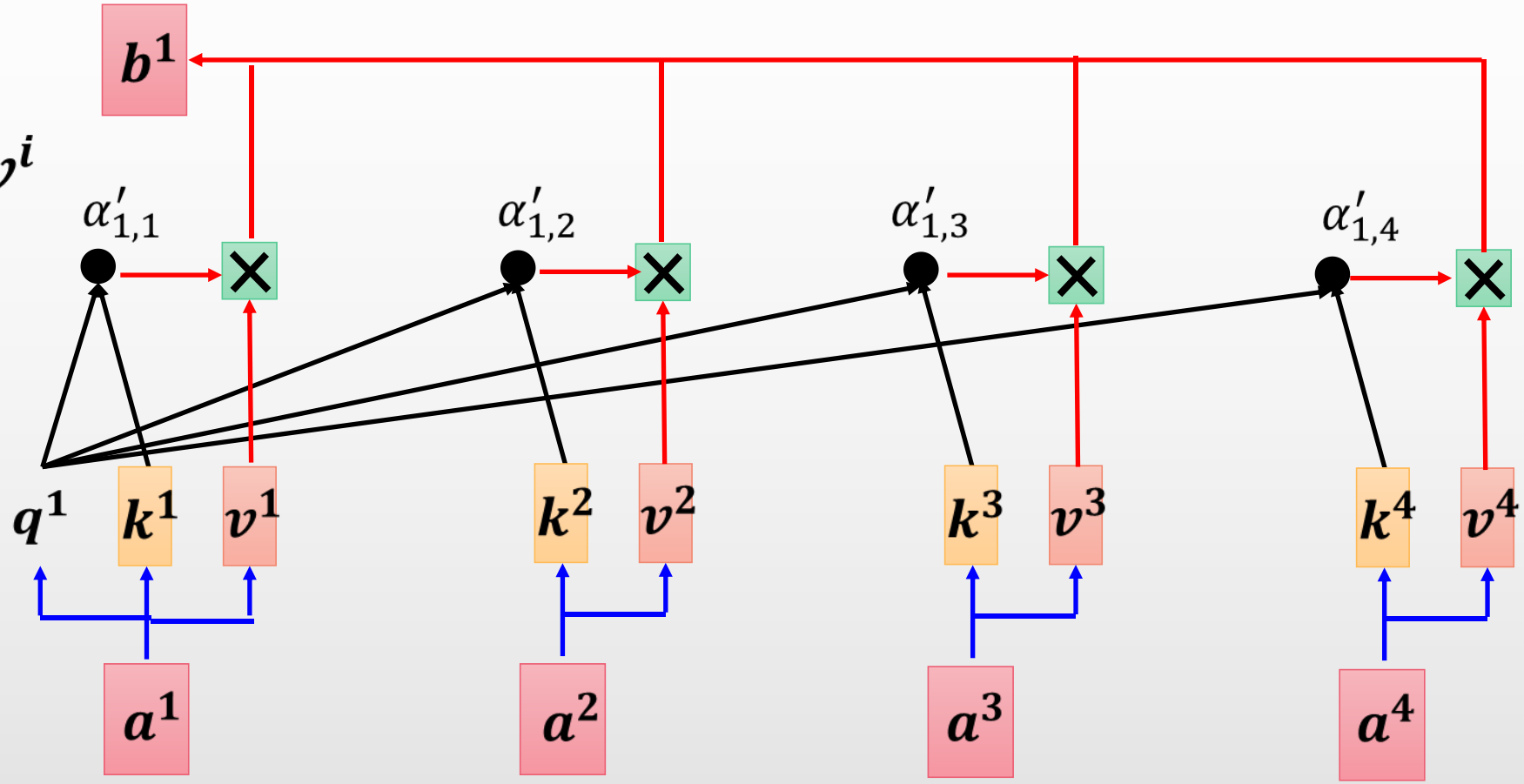


$$\boldsymbol{q^1} = W^q \boldsymbol{a^1} \qquad \boldsymbol{k^2} = W^k \boldsymbol{a^2} \qquad \boldsymbol{k^3} = W^k \boldsymbol{a^3} \qquad \boldsymbol{k^4} = W^k \boldsymbol{a^4}$$

Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet

李玉光

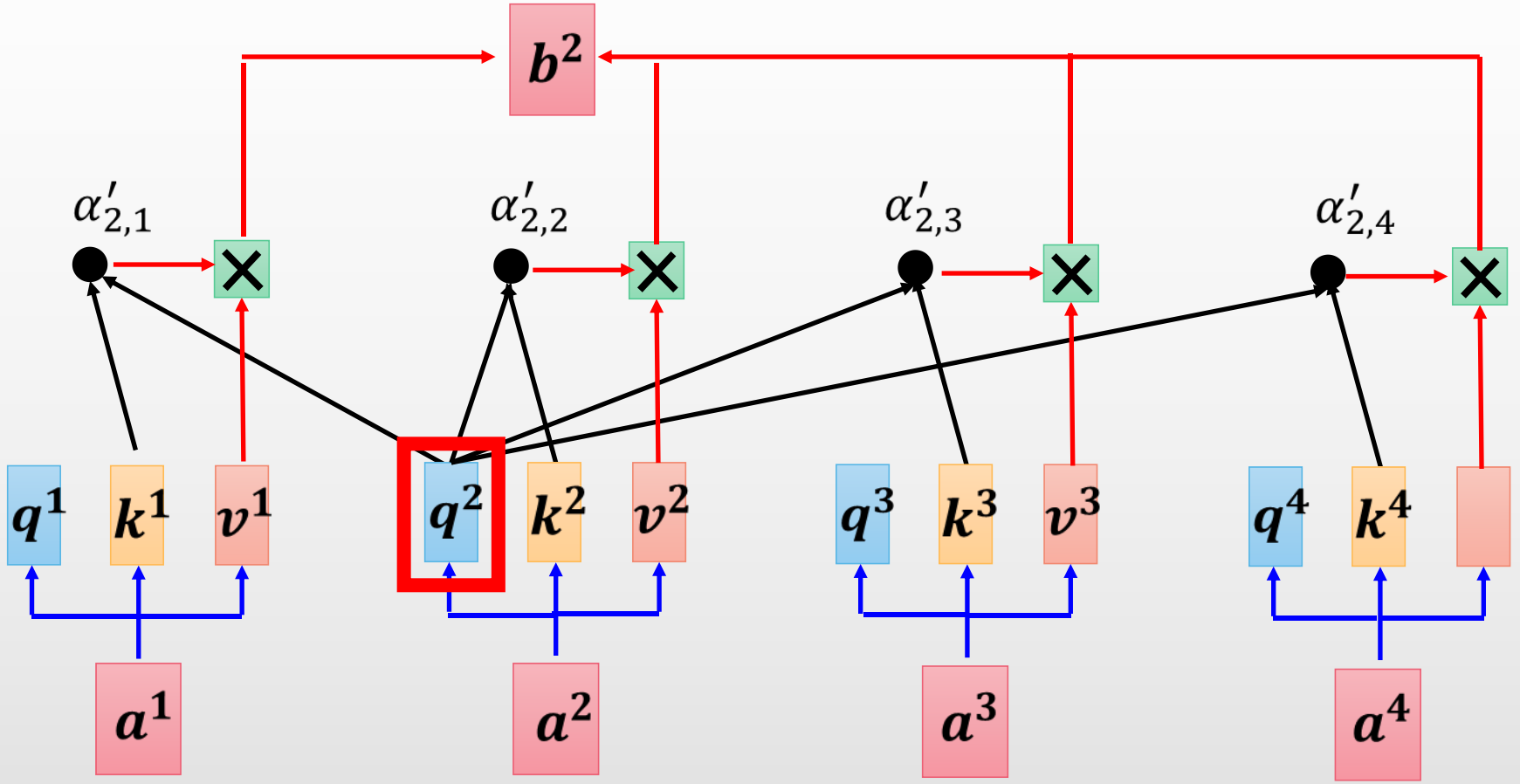
$$b^1 = \sum_i \alpha'_{1,i} v^i$$



$$v^2 = W^v a^2 \qquad v^3 = W^v a^3 \qquad v^4 = W^v a^4$$

Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet

李玉光



Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet

李玉光

$$q^i = W^q a^i \qquad q^1 q^2 q^3 q^4 = W^q \; a^1 a^2 a^3 a^4$$

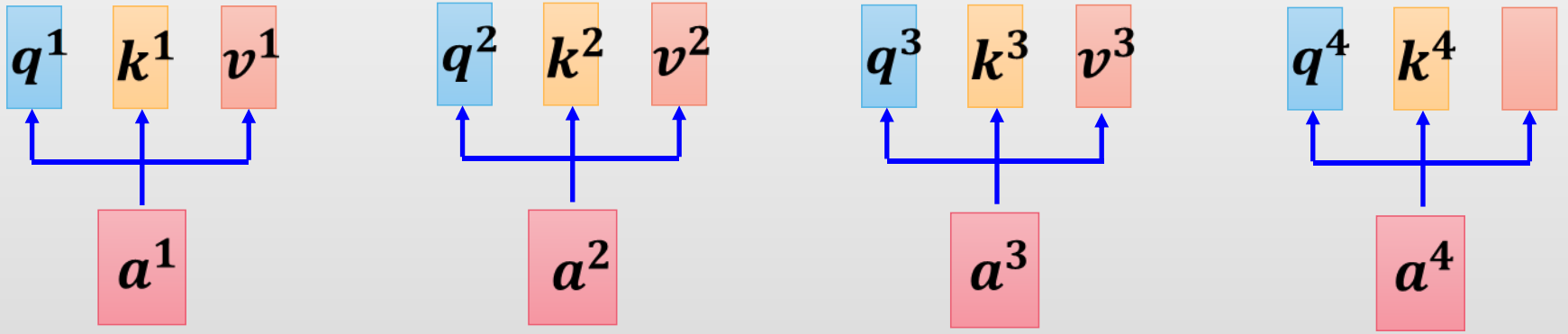
Q $\quad$ I

$$k^i = W^k a^i \qquad k^1 k^2 k^3 k^4 = W^k \; a^1 a^2 a^3 a^4$$

K $\quad$ I

$$v^i = W^v a^i \qquad v^1 v^2 v^3 v^4 = W^v \; a^1 a^2 a^3 a^4$$

V $\quad$ I

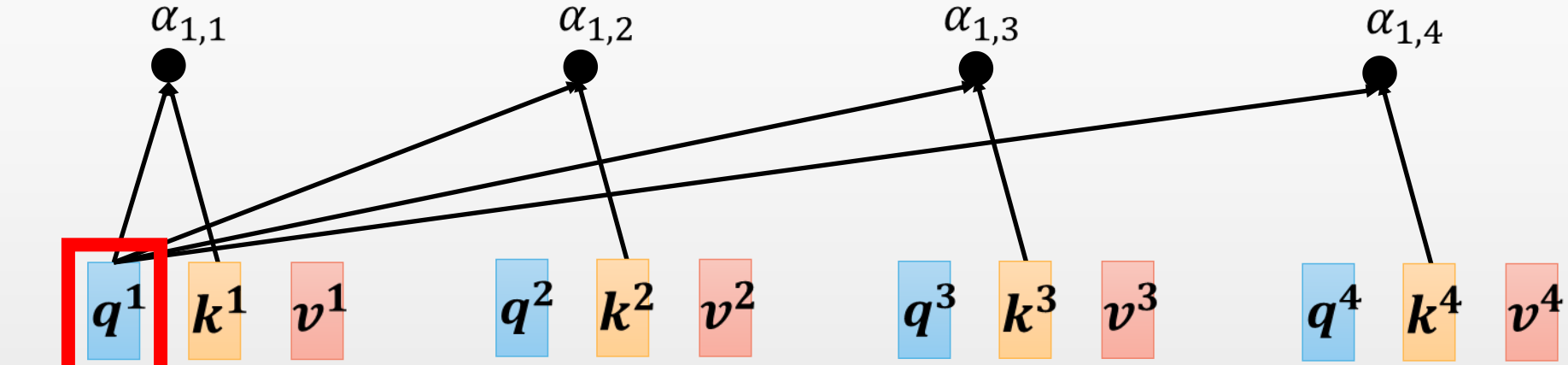


李玉光

$$\alpha_{1,1} = \boldsymbol{k^1}\,\boldsymbol{q^1} \qquad \alpha_{1,2} = \boldsymbol{k^2}\,\boldsymbol{q^1}$$

$$\alpha_{1,3} = \boldsymbol{k^3}\,\boldsymbol{q^1} \qquad \alpha_{1,4} = \boldsymbol{k^4}\,\boldsymbol{q^1}$$

$$\begin{bmatrix}\alpha_{1,1}\\ \alpha_{1,2}\\ \alpha_{1,3}\\ \alpha_{1,4}\end{bmatrix} = \begin{bmatrix}\boldsymbol{k^1}\\ \boldsymbol{k^2}\\ \boldsymbol{k^3}\\ \boldsymbol{k^4}\end{bmatrix} \boldsymbol{q^1}$$

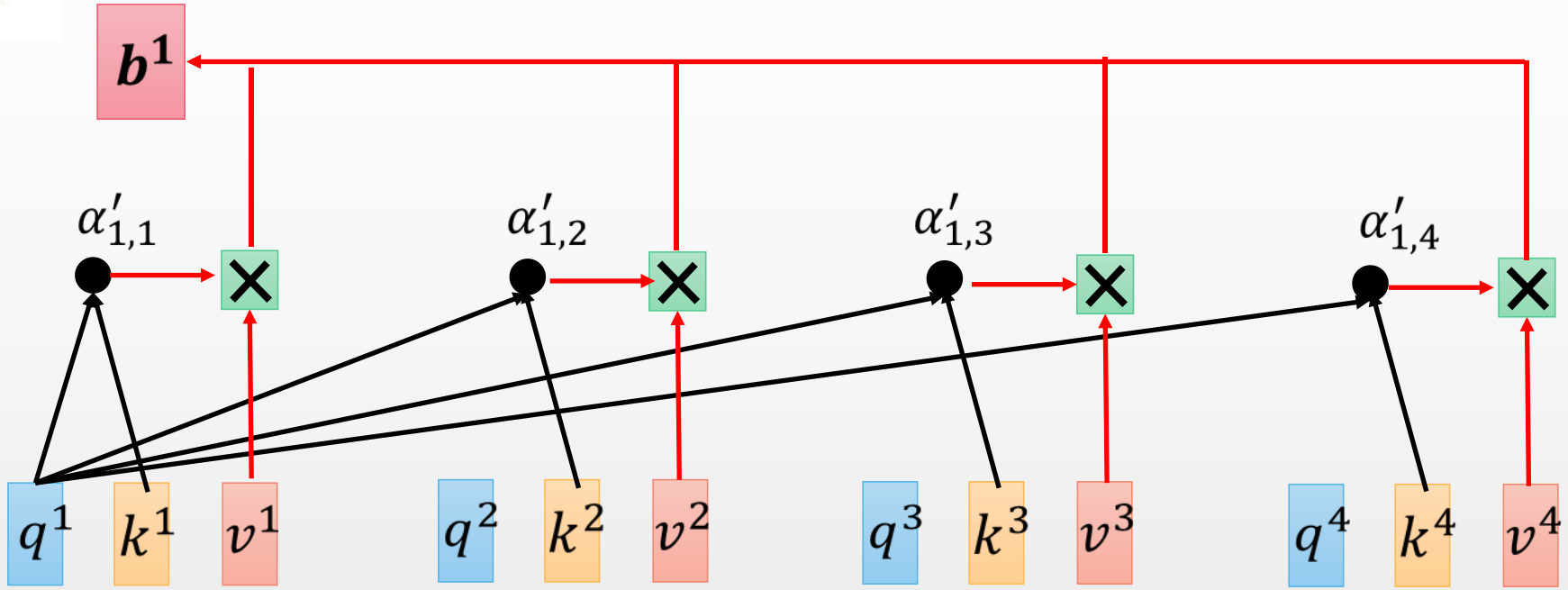


$$\begin{bmatrix}\alpha'_{1,1} & \alpha'_{2,1} & \alpha'_{3,1} & \alpha'_{4,1}\\ \alpha'_{1,2} & \alpha'_{2,2} & \alpha'_{3,2} & \alpha'_{4,2}\\ \alpha'_{1,3} & \alpha'_{2,3} & \alpha'_{3,3} & \alpha'_{4,3}\\ \alpha'_{1,4} & \alpha'_{2,4} & \alpha'_{3,4} & \alpha'_{4,4}\end{bmatrix} \quad A' \qquad \text{softmax} \qquad \begin{bmatrix}\alpha_{1,1} & \alpha_{2,1} & \alpha_{3,1} & \alpha_{4,1}\\ \alpha_{1,2} & \alpha_{2,2} & \alpha_{3,2} & \alpha_{4,2}\\ \alpha_{1,3} & \alpha_{2,3} & \alpha_{3,3} & \alpha_{4,3}\\ \alpha_{1,4} & \alpha_{2,4} & \alpha_{3,4} & \alpha_{4,4}\end{bmatrix} \quad A$$

$$= \begin{bmatrix}\boldsymbol{k^1}\\ \boldsymbol{k^2}\\ \boldsymbol{k^3}\\ \boldsymbol{k^4}\end{bmatrix} \quad K^T \qquad \begin{matrix}\boldsymbol{q^1} & \boldsymbol{q^2} & \boldsymbol{q^3} & \boldsymbol{q^4}\end{matrix} \quad Q$$

Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet

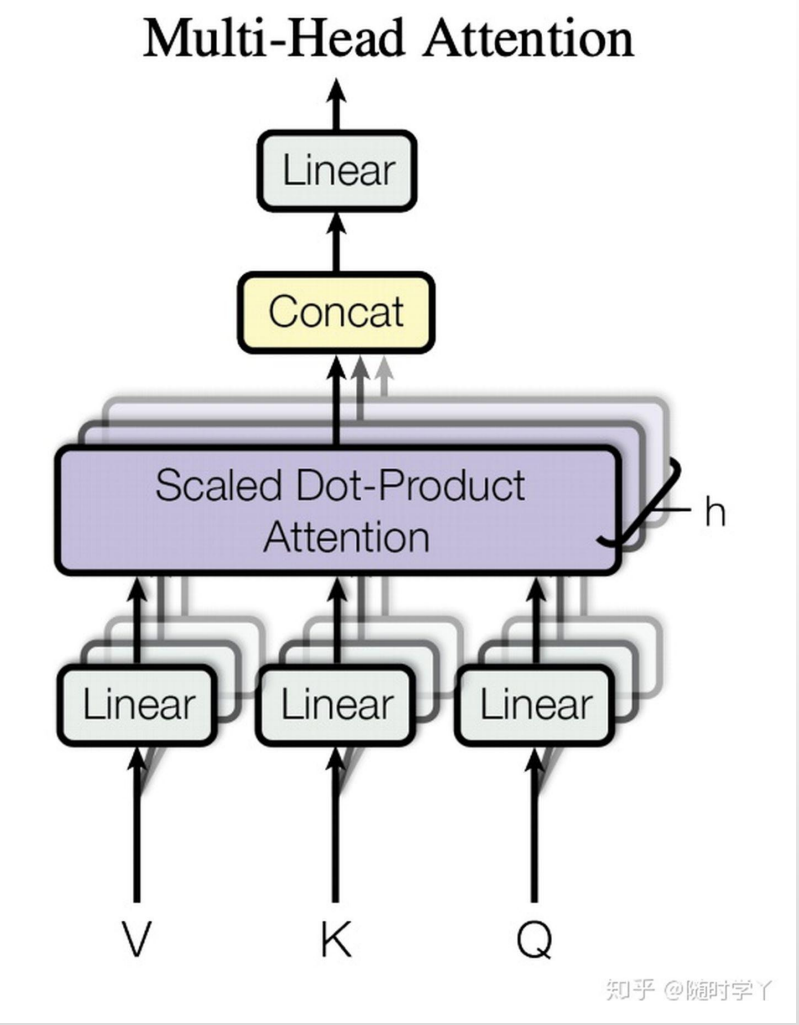
李玉光

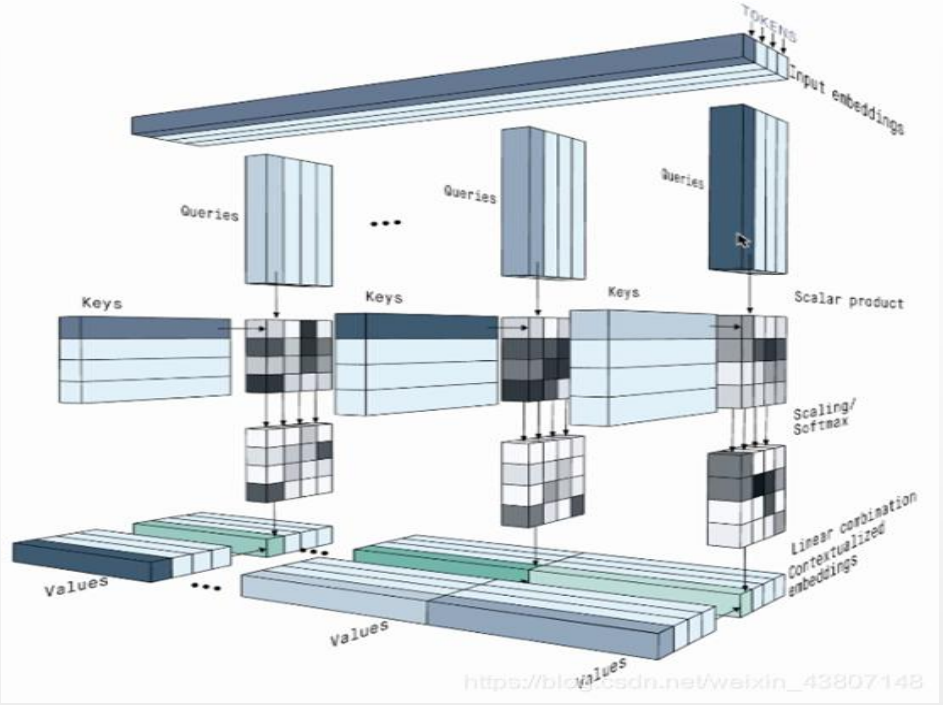


$$\underbrace{b^1 b^2 b^3 b^4}_{O} = \underbrace{v^1 v^2 v^3 v^4}_{V} \underbrace{\begin{bmatrix} \alpha'_{1,1} & \alpha'_{2,1} & \alpha'_{3,1} & \alpha'_{4,1} \\ \alpha'_{1,2} & \alpha'_{2,2} & \alpha'_{3,2} & \alpha'_{4,2} \\ \alpha'_{1,3} & \alpha'_{2,3} & \alpha'_{3,3} & \alpha'_{4,3} \\ \alpha'_{1,4} & \alpha'_{2,4} & \alpha'_{3,4} & \alpha'_{4,4} \end{bmatrix}}_{A'}$$

Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet

李玉光





$$Q_i = QW_i^Q, K_i = KW_i^K, V_i = VW_i^V, i = 1, \ldots, 8$$

$$head_i = Attention(Q_i, K_i, V_i), i = 1, \ldots, 8$$

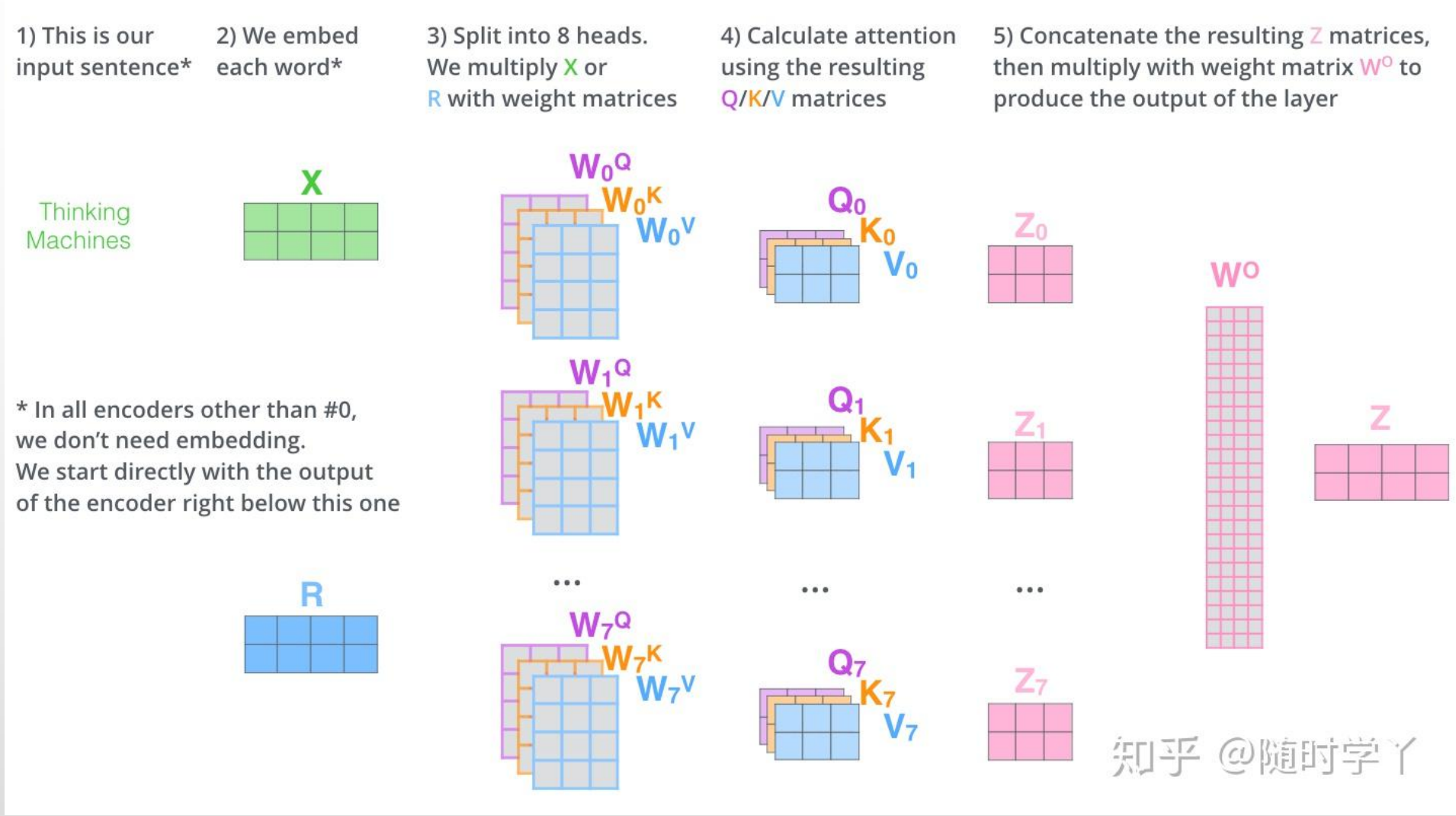
$$MultiHead(Q, K, V) = Concact(head_1, \ldots, head_8)W^O$$

这里，我们假设

$$Q, K, V \in R^{512}, W_i^Q, W_i^K, W_i^V \in R^{512 \times 64}, W^O \in R^{512 \times 512}, head_i \in R^{64}$$

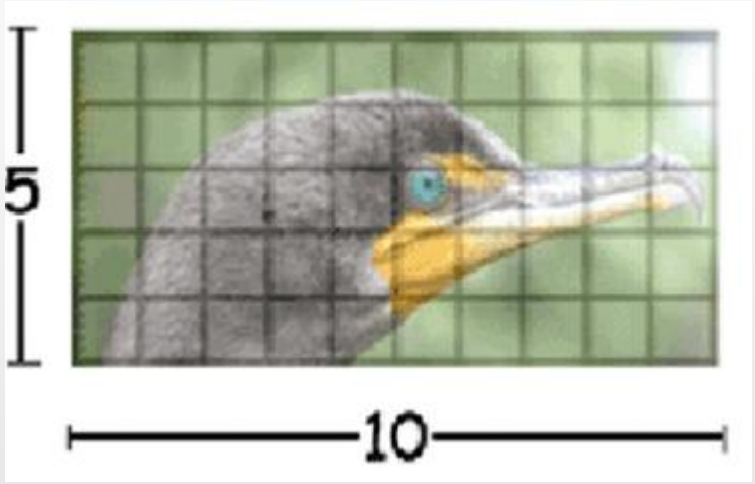
Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet

李玉光

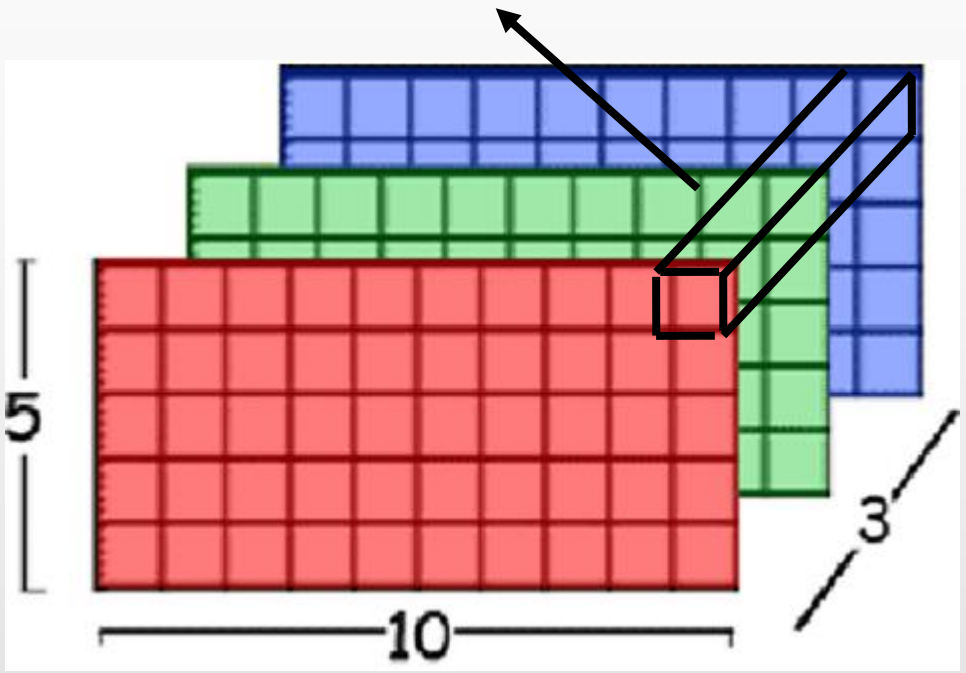


Self-attention for Image

An **image** can also be considered as a **vector set**.

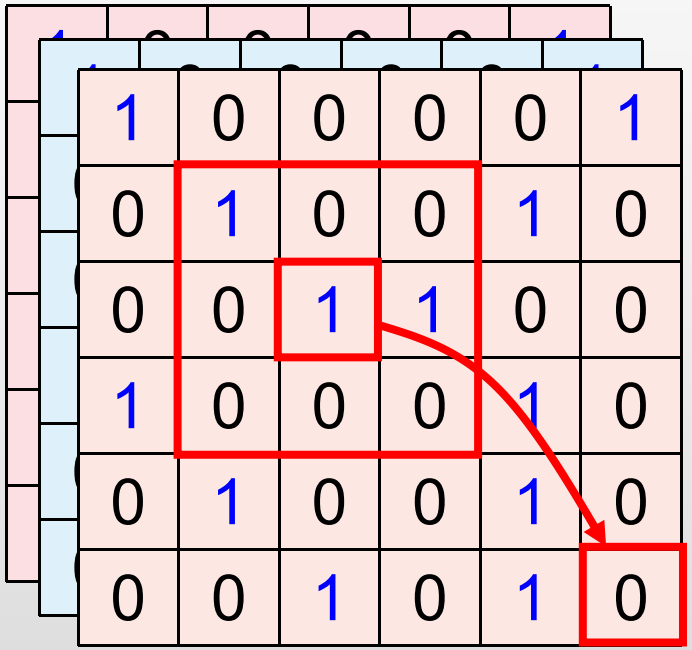


This is a vector.



Source of image: https://www.researchgate.net/figure/Color-image-representation-and-RGB-matrix_fig15_282798184

Self-attention v.s. CNN



CNN: self-attention that can only attends in a receptive field

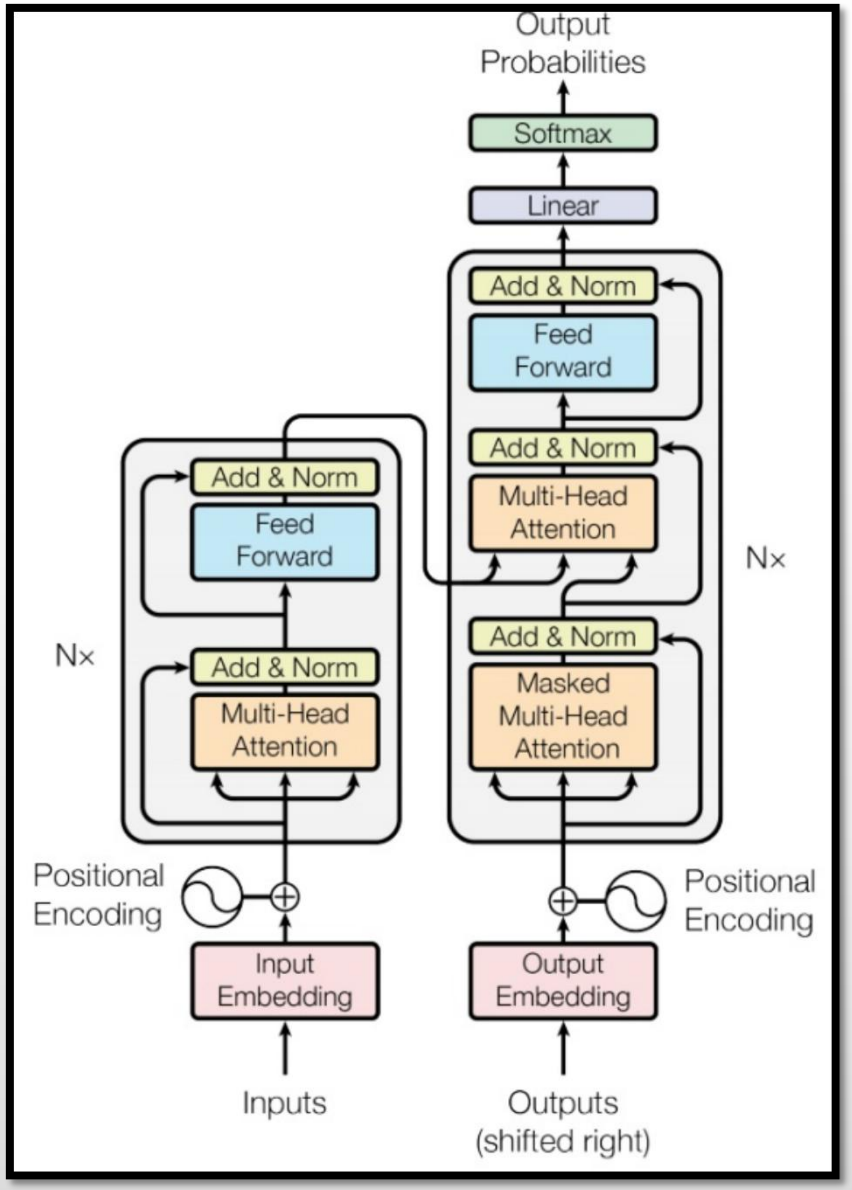
- CNN is simplified self-attention.

Self-attention: CNN with learnable receptive field

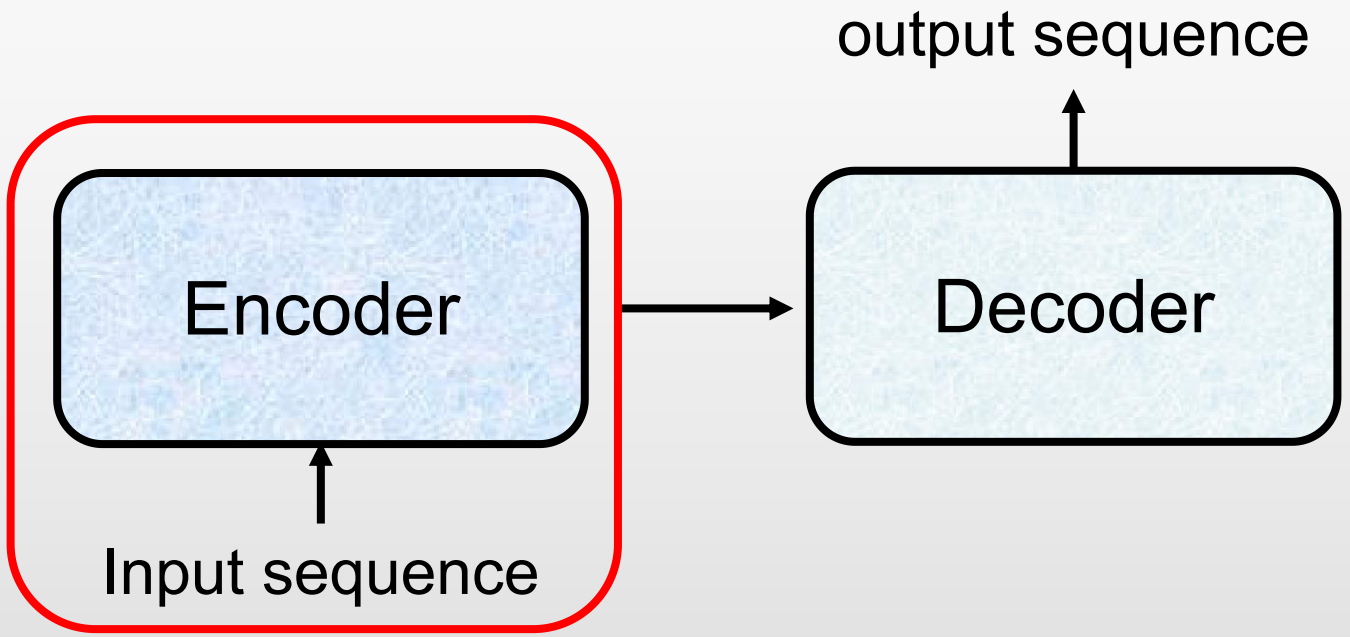
- Self-attention is the complex version of CNN.

TRANSFORMER

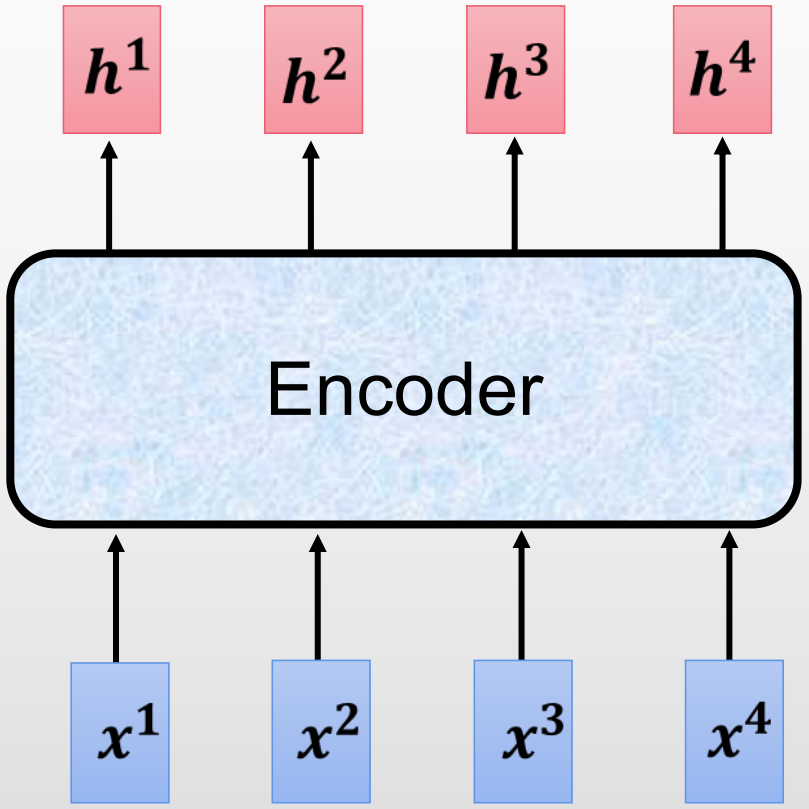
Transformer is the first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence aligned RNNs or convolution。

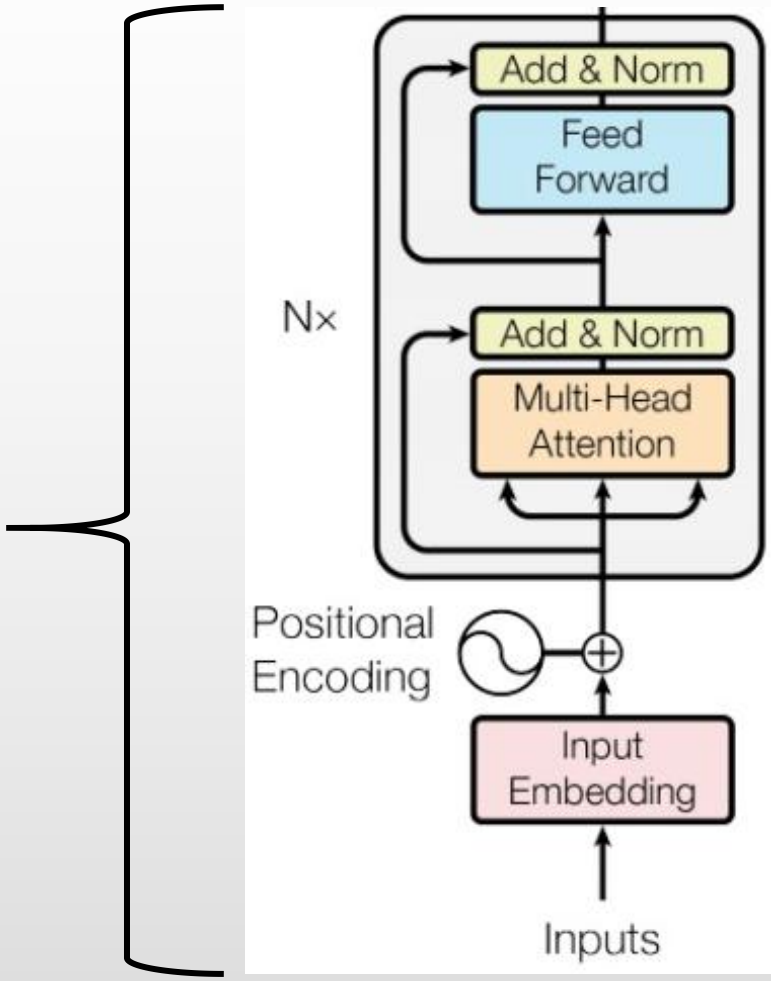


Encoder



You can use **RNN** or **CNN**.

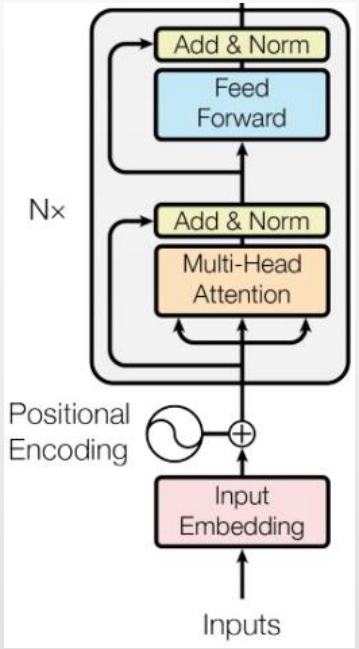




Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet

李玉光

POSITIONAL ENCODING



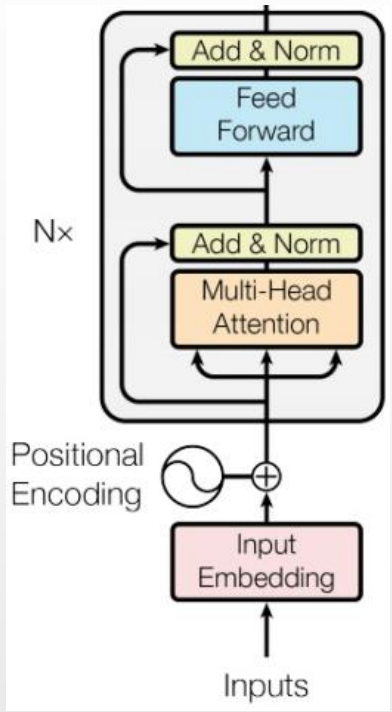
$$PE = pos = 0, 1, 2, \cdots, T-1$$

$$PE = pos/(T-1)$$

$$PE(pos) = \sin\left(\frac{pos}{\alpha}\right)$$

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right)$$

$$PE(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right)$$



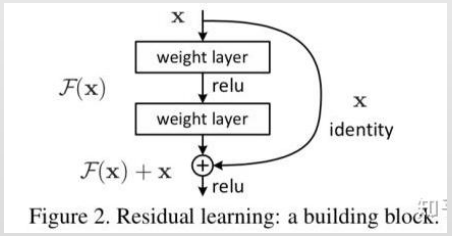


Figure 2. Residual learning: a building block.

norm

$a+b$

b + a

$$\begin{bmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_K \end{bmatrix} \qquad x'_i = \frac{x_i - m}{\sigma}$$

norm

Layer Norm

https://arxiv.org/abs/1607.06450

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_K \end{bmatrix}$$

mean m

standard deviation σ

norm

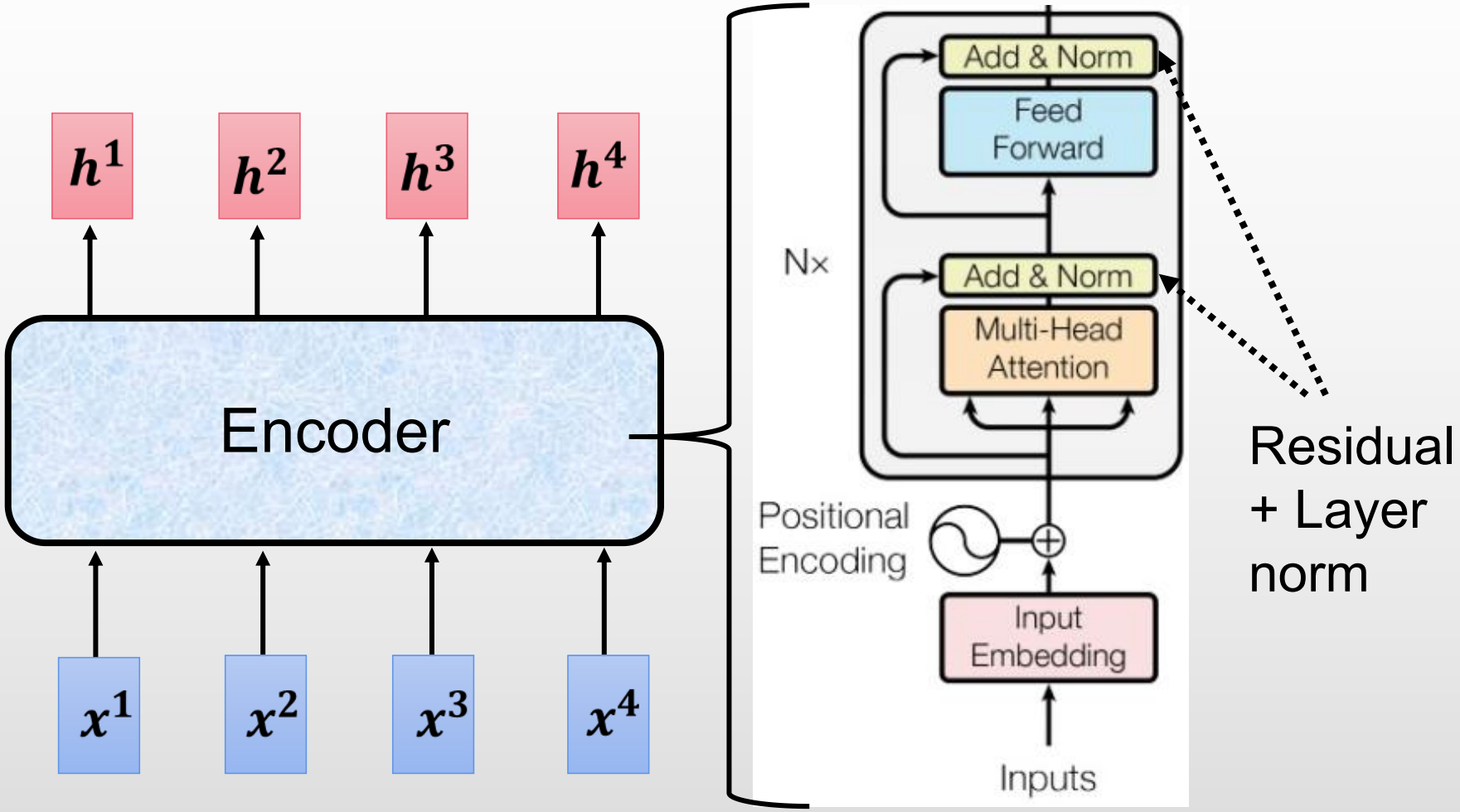
+

FC

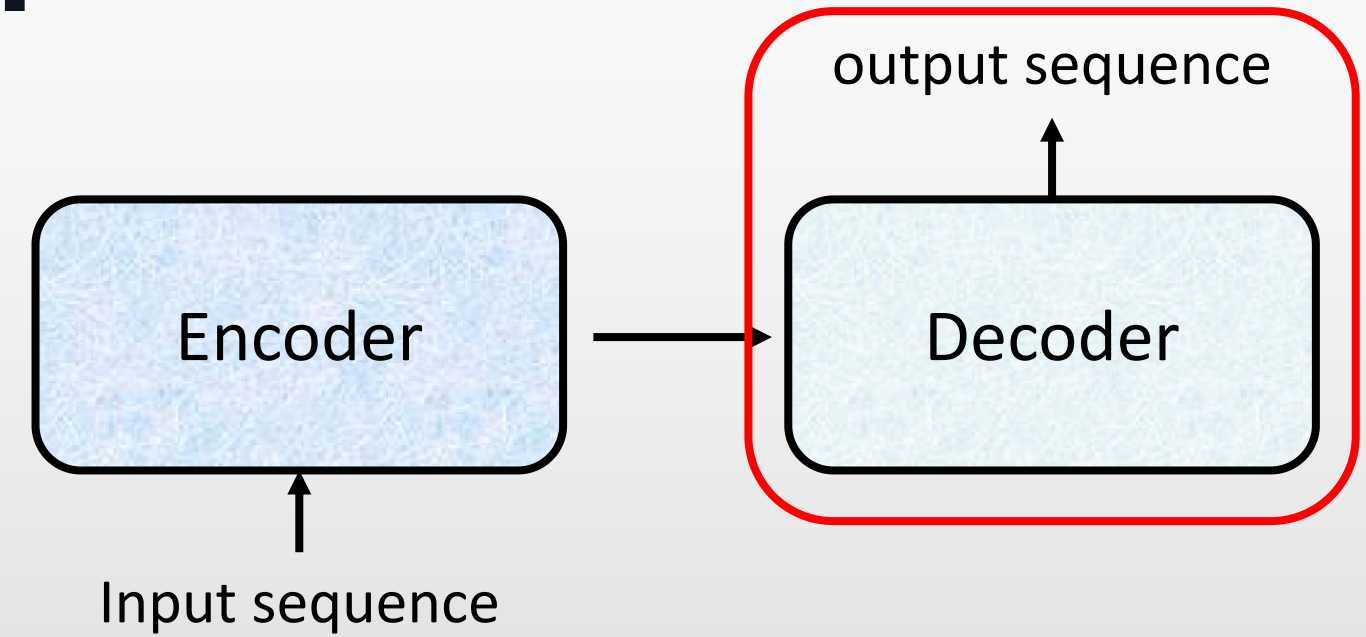
Self-attention

Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet

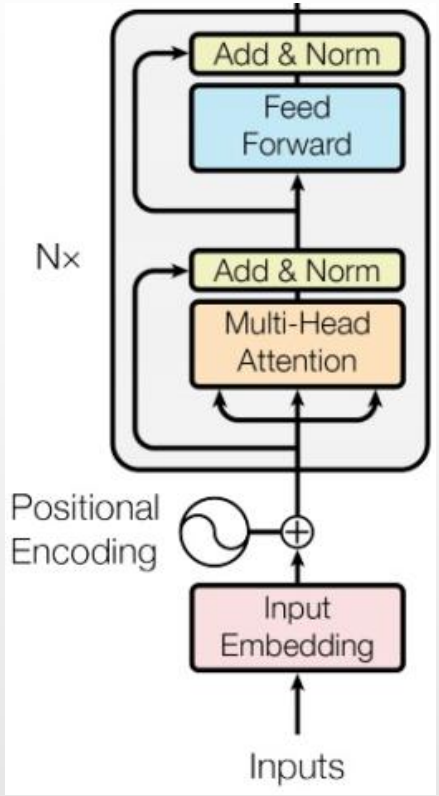
李玉光



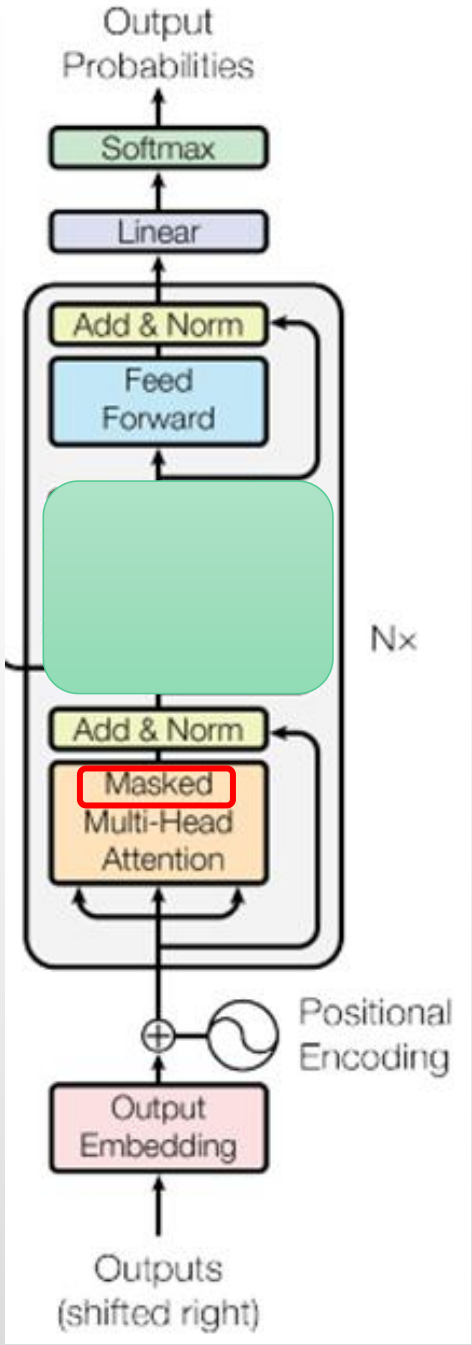
Decoder



李玉光



Encoder

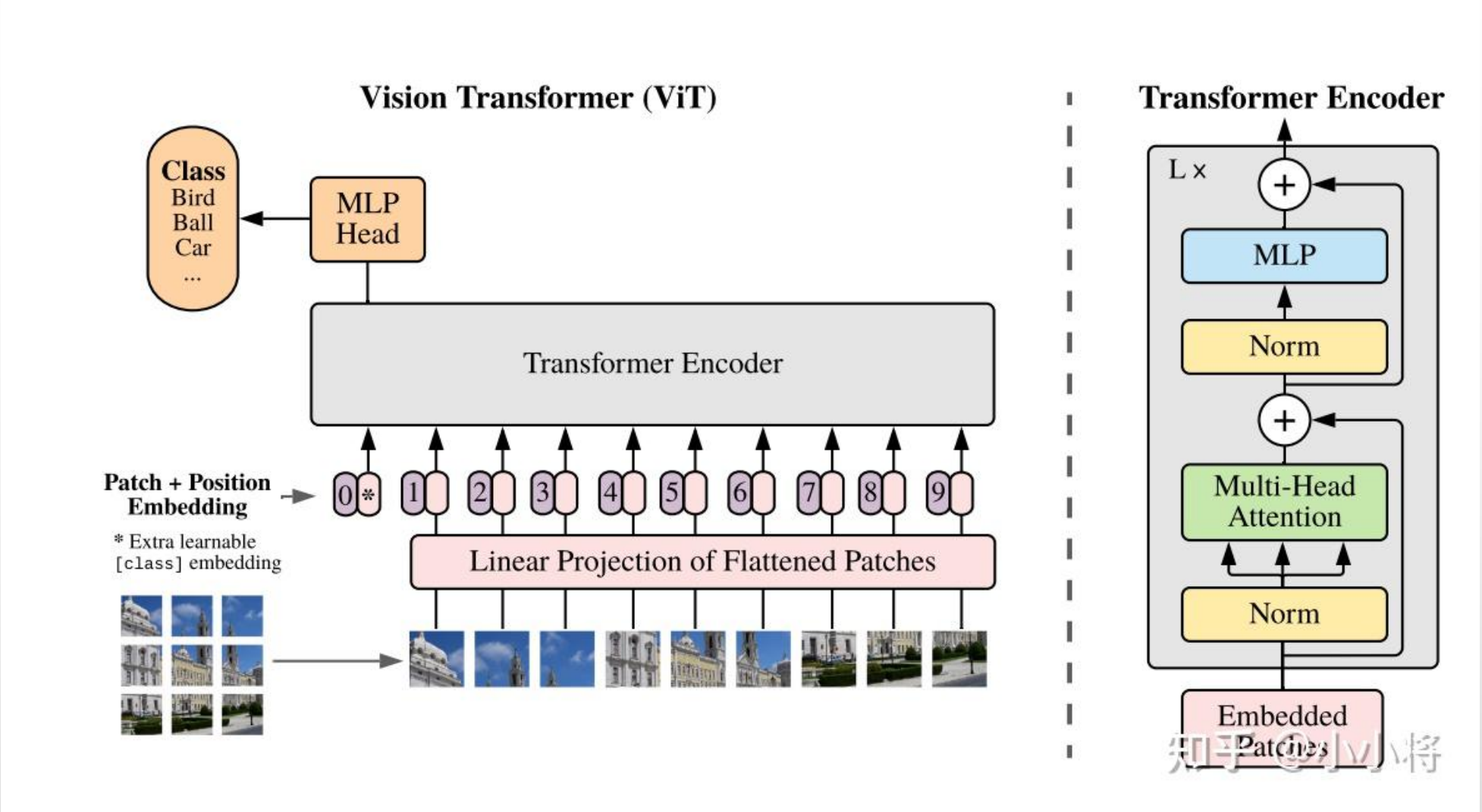


Decoder

Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet

李玉光

VIT



Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet

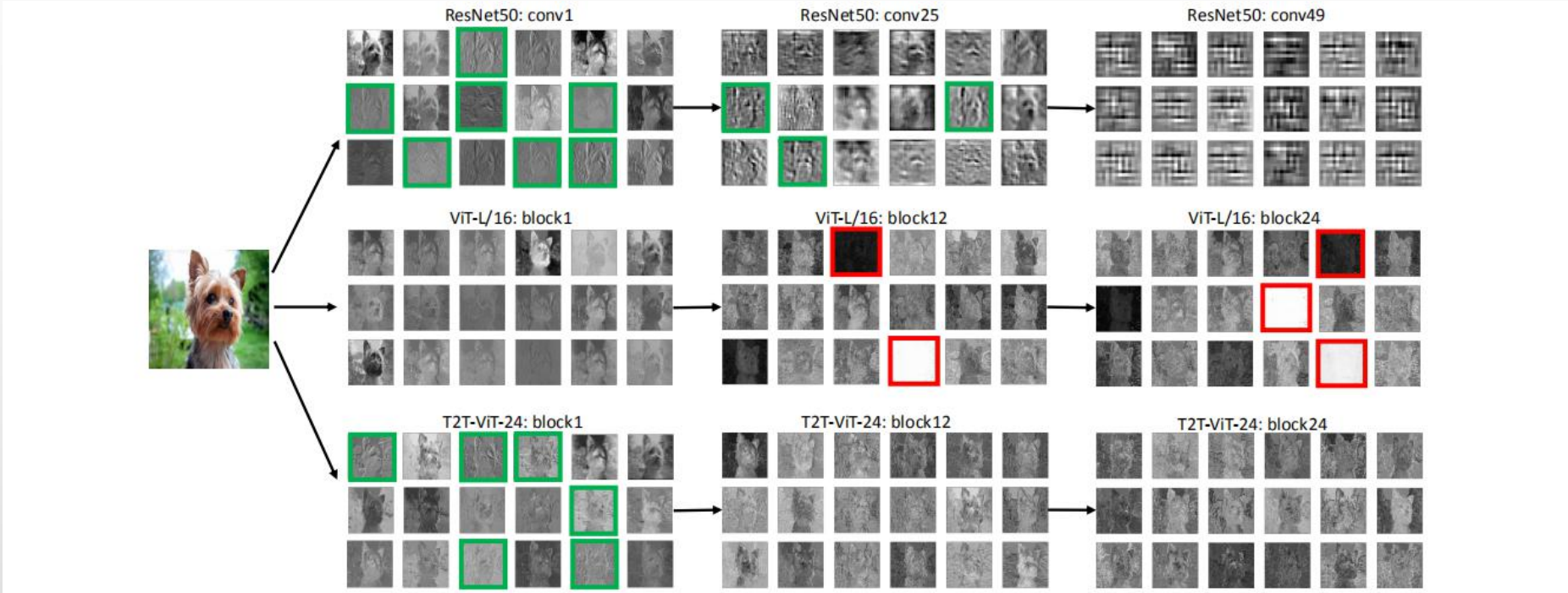
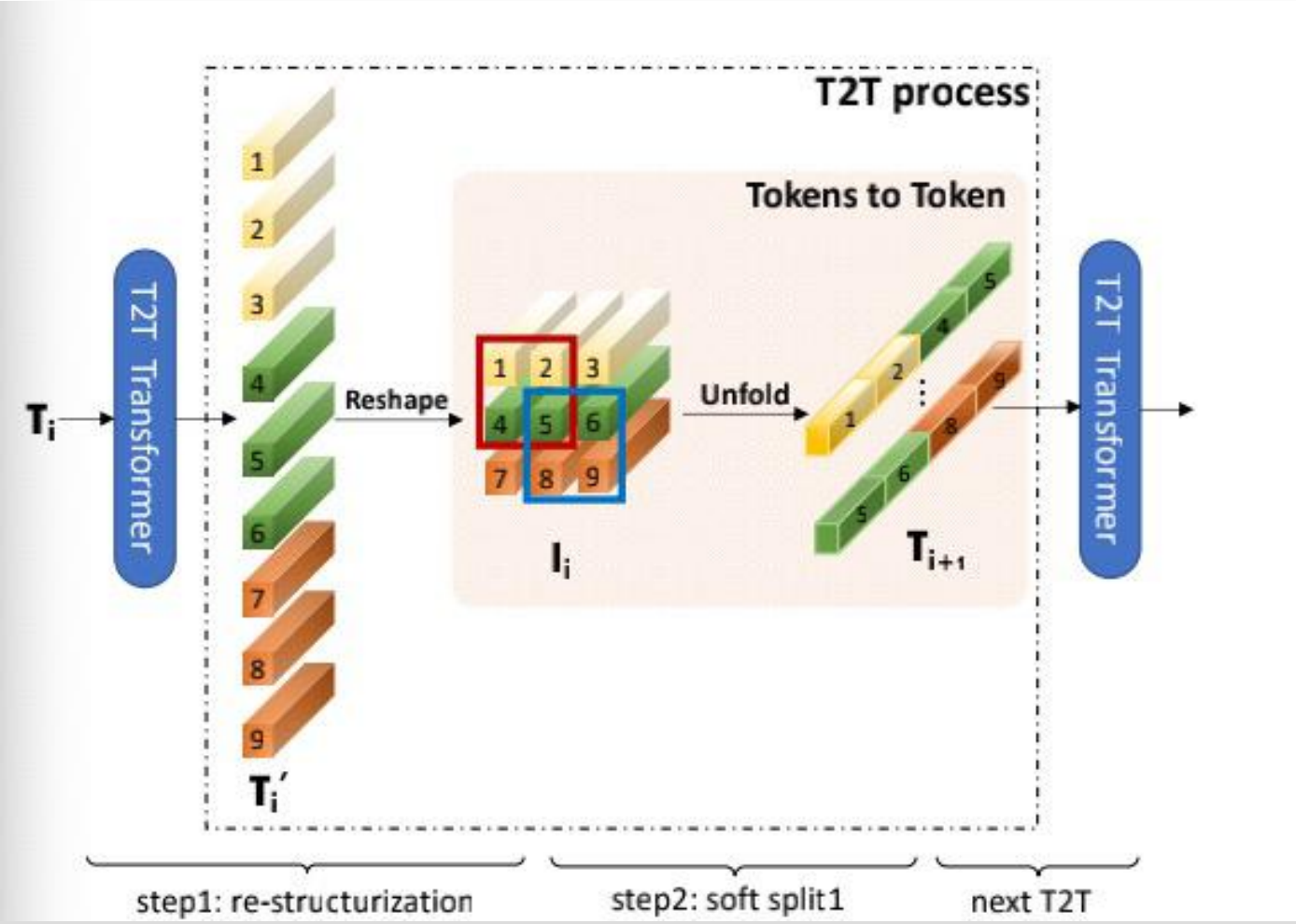


Figure 2. Feature visualization of ResNet50, ViT-L/16 [12] and our proposed T2T-ViT-24 trained on ImageNet. The green boxes highlight learned low-level structure features such as edges and lines. The red boxes highlight invalid feature maps with zero or too large values. Note the feature maps visualized here for ViT and T2T-ViT are not attention maps, but image features reshaped from tokens. For better visualization, we scale input image to size 1024×1024.

Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet

李玉光

T2T-ViT



1）Restructurization

T' = MLP(MSA(T))

I =reshape (T')

2）Soft Split

$$T_{i+1} = SS(I_i)\ , i = 1, \cdots, (n-1)$$

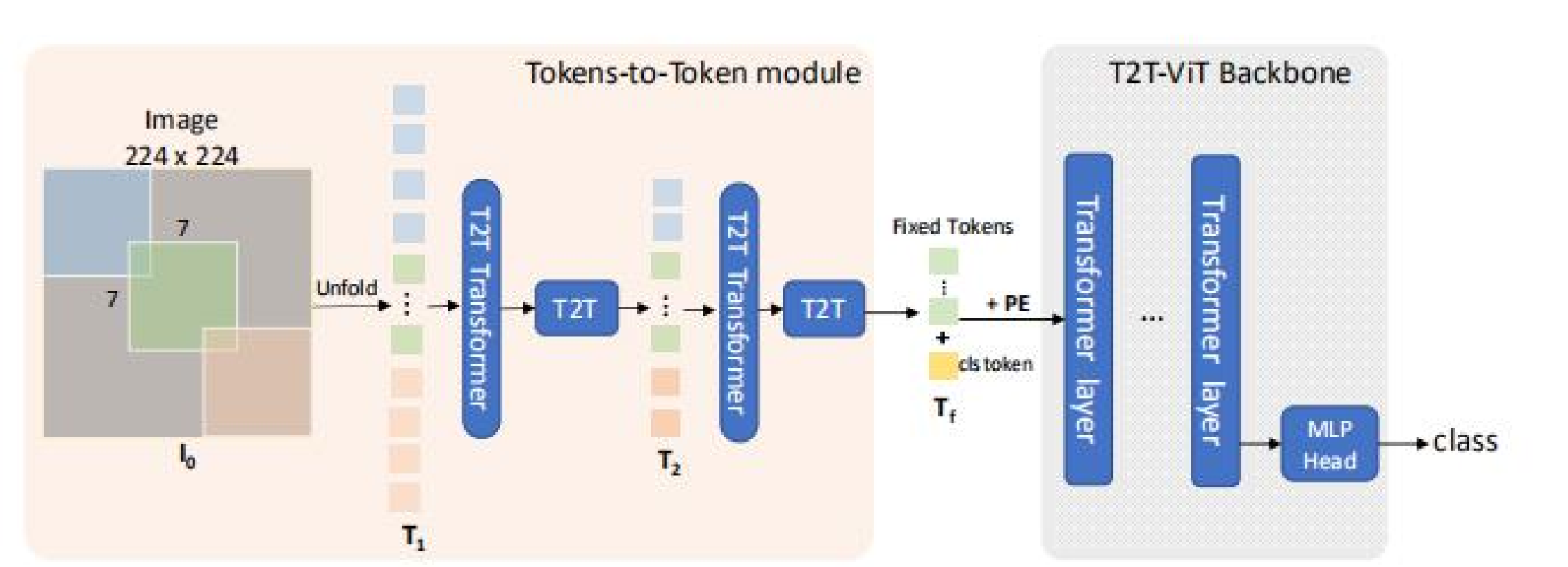
Backbone

- Dense Connection，类似于DenseNet；
- Deep-narrow vs shallow-wide结构，类似于Wide-ResNet一文的讨论；
- Channel Attention，类似SENet；
- More Split Head，类似ResNeXt；
- Ghost操作，类似GhostNet。

结论：Deep-Narrow结构可以在通道层面通过减少通道维度减少冗余，可以通过提升深度提升特征丰富性。

Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet

李玉光



Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet

李玉光

Models	Top1-Acc (%)	Params (M)	MACs (G)
ResNet50 [15]	76.2	25.5	4.3
ResNet50*	79.1	25.5	4.3
T2T-ViT-14	**80.6**	21.4	4.8
T2T-ViT$_t$-14	**80.7**	21.5	5.2
ResNet101 [15]	77.4	44.6	7.9
ResNet101*	79.9	44.6	7.9
T2T-ViT-19	**81.2**	39.0	8.0
T2T-ViT$_t$-19	**81.4**	39.0	8.4
ResNet152 [15]	78.3	60.2	11.6
ResNet152*	80.8	60.2	11.6
T2T-ViT-24	**81.8**	63.9	12.6
T2T-ViT$_t$-24	**82.2**	64.1	13.2

感谢大家！