Not All Images are Worth 16x16 Words: Dynamic Vision Transformers with Adaptive Sequence Length

—— Tsinghua University & Huawei Technologies, arxiv.2021

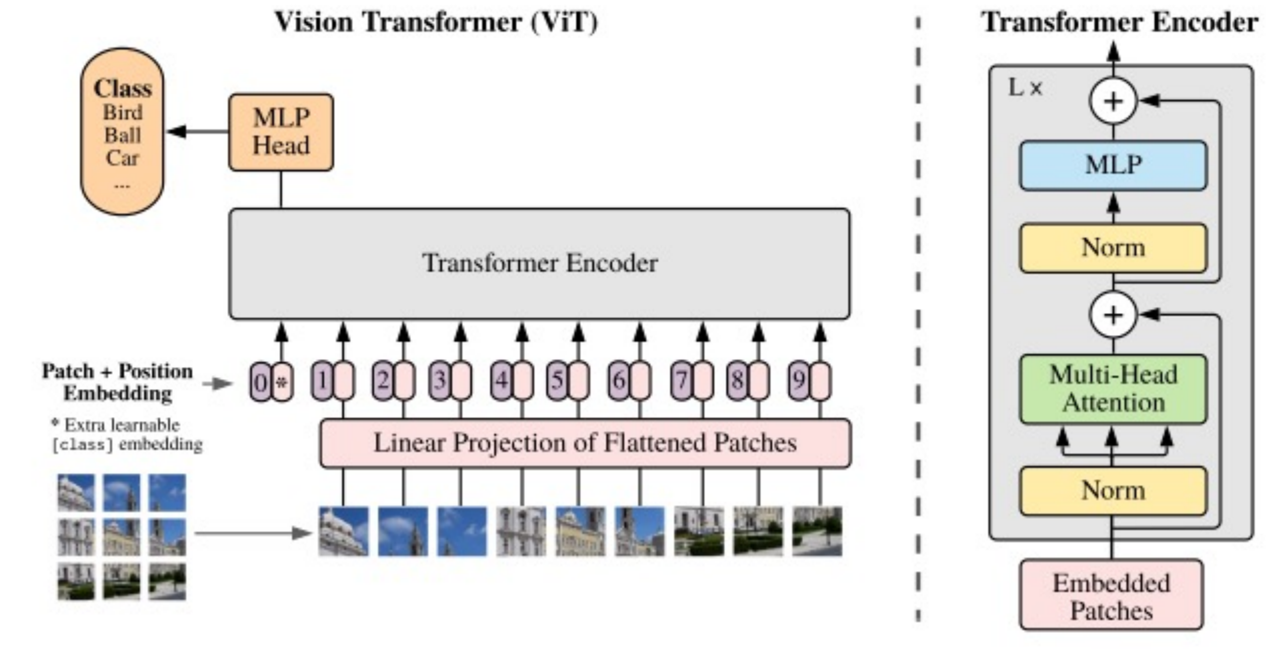
Fang Zhiyuan

2021.7.23

Prior knowledge

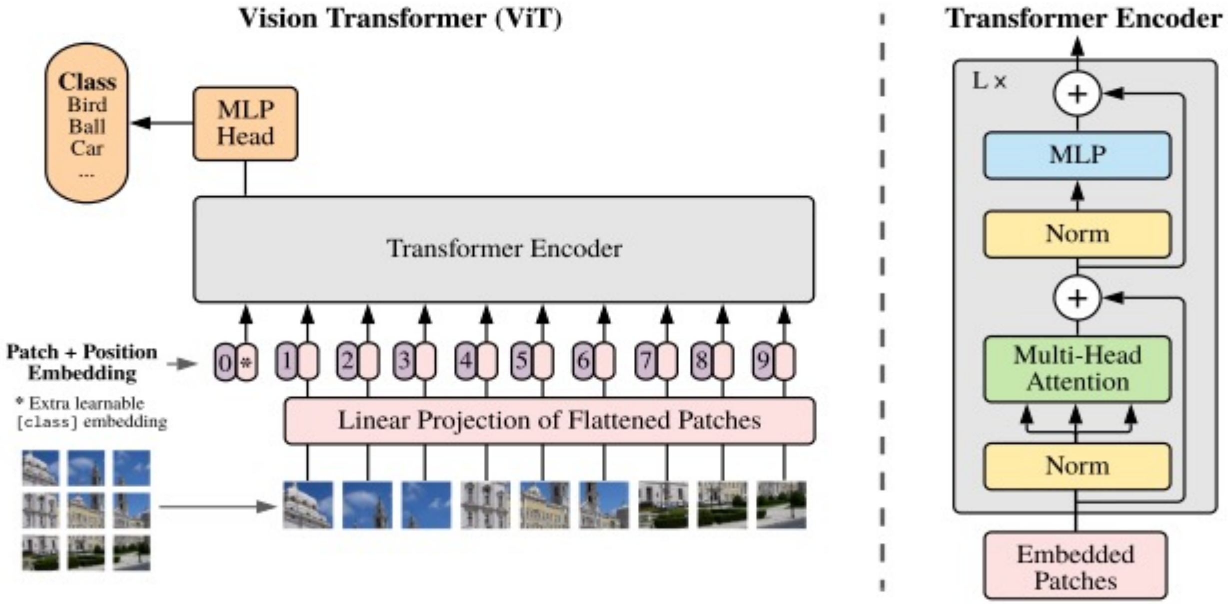
An Image Is Worth 16x16 Words: Transformers For Image Recognition At Scale, ICLR 2021

- Replace convolution layers with standard transformer



Prior knowledge

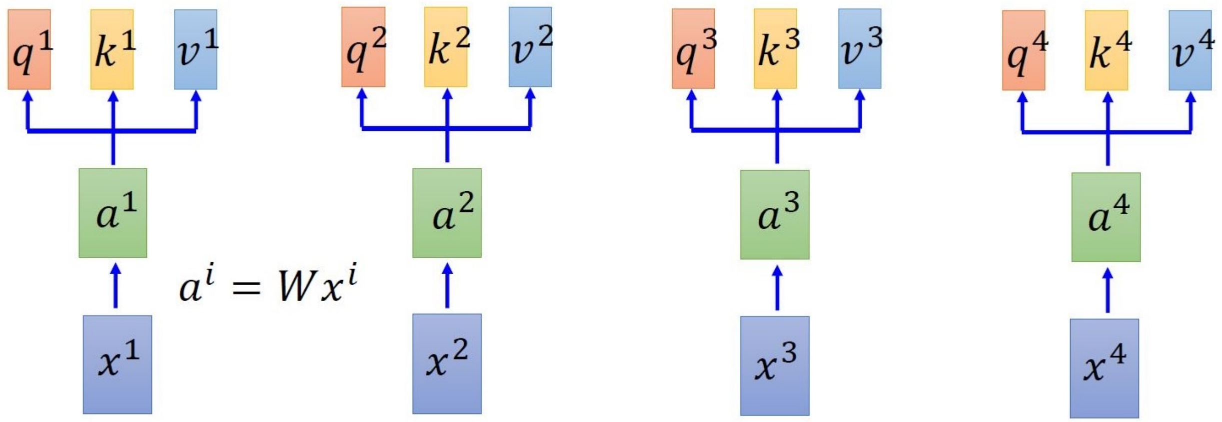
- Split an image into patches, $x \in H \times W \times C \rightarrow x_p \in N \times (P^2 \cdot C)$
- provide the sequence of linear embeddings, $x_p \in N \times (P^2 \cdot C) \rightarrow x_p \in N \times D$
- Position embeddings
- Additional learnable embedding (class token)



Prior knowledge

Self-attention

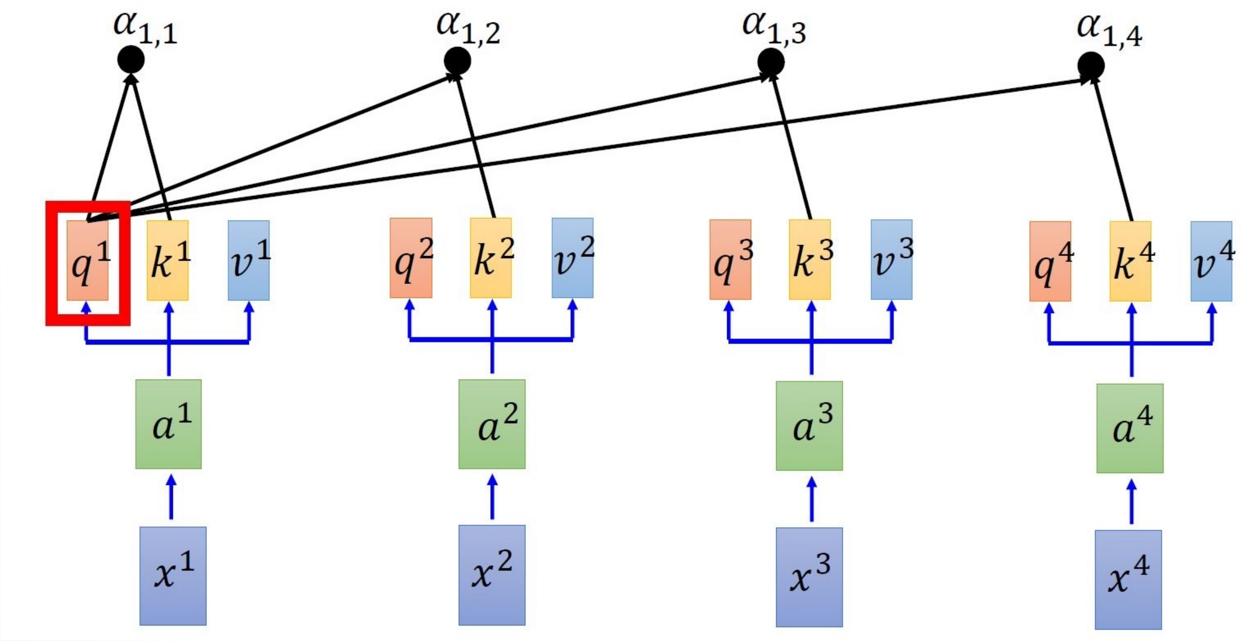
$$q^i = W^q a^i \qquad k^i = W^k a^i \qquad v^i = W^v a^i$$



Prior knowledge

Self-attention

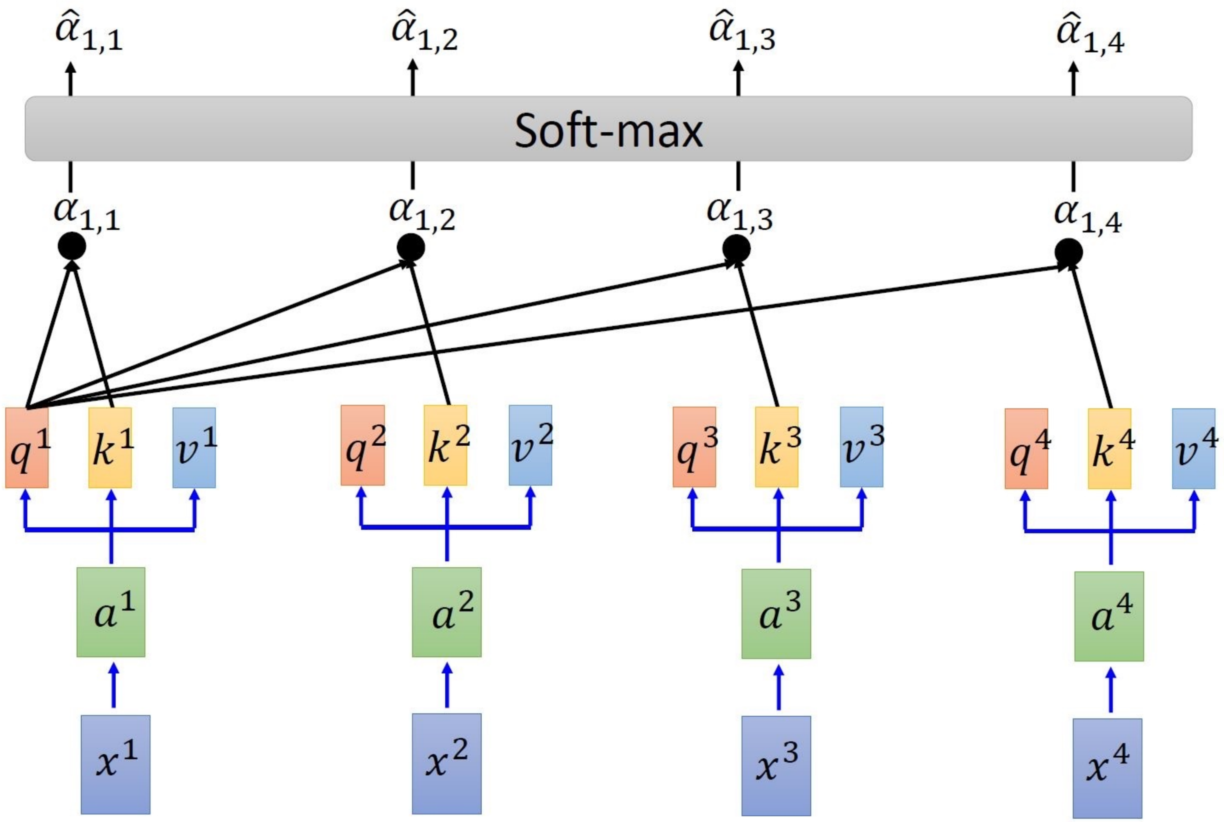
$a_{1,i} = q^1 \cdot k^1/\sqrt{d}$, d is the dim of q and k



Prior knowledge

Self-attention

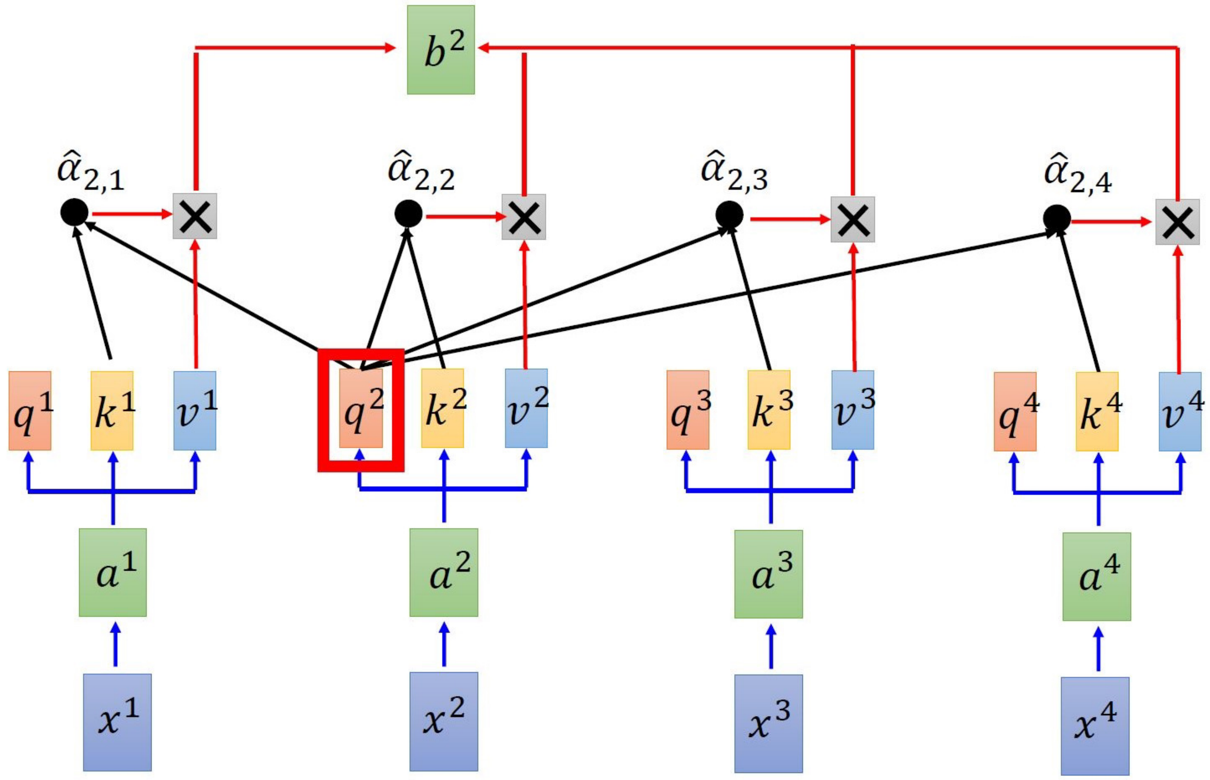
$$\hat{a}_{1,i} = \exp(a_{1,i}) \,/\, \sum_j \exp(a_{i,j})$$



Prior knowledge

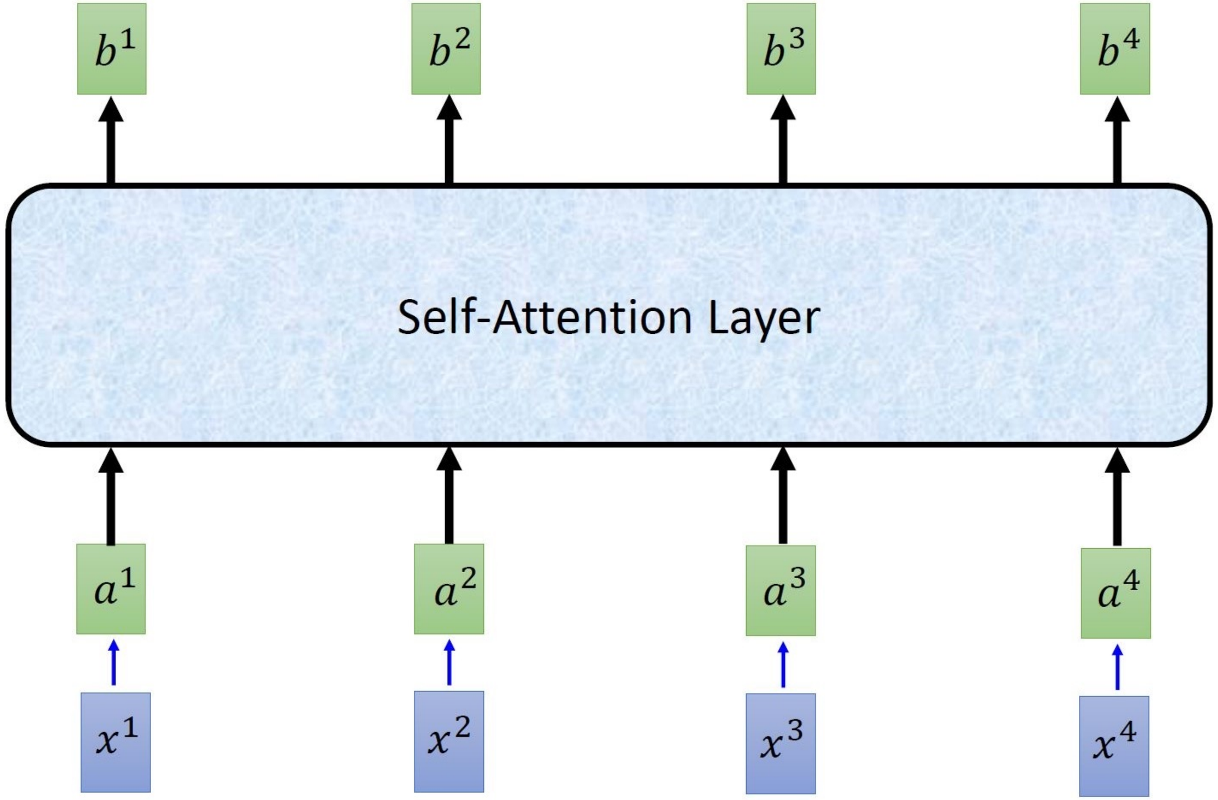
Self-attention

$$b^2 = \sum_i \hat{a}_{2,i}\, v^i$$



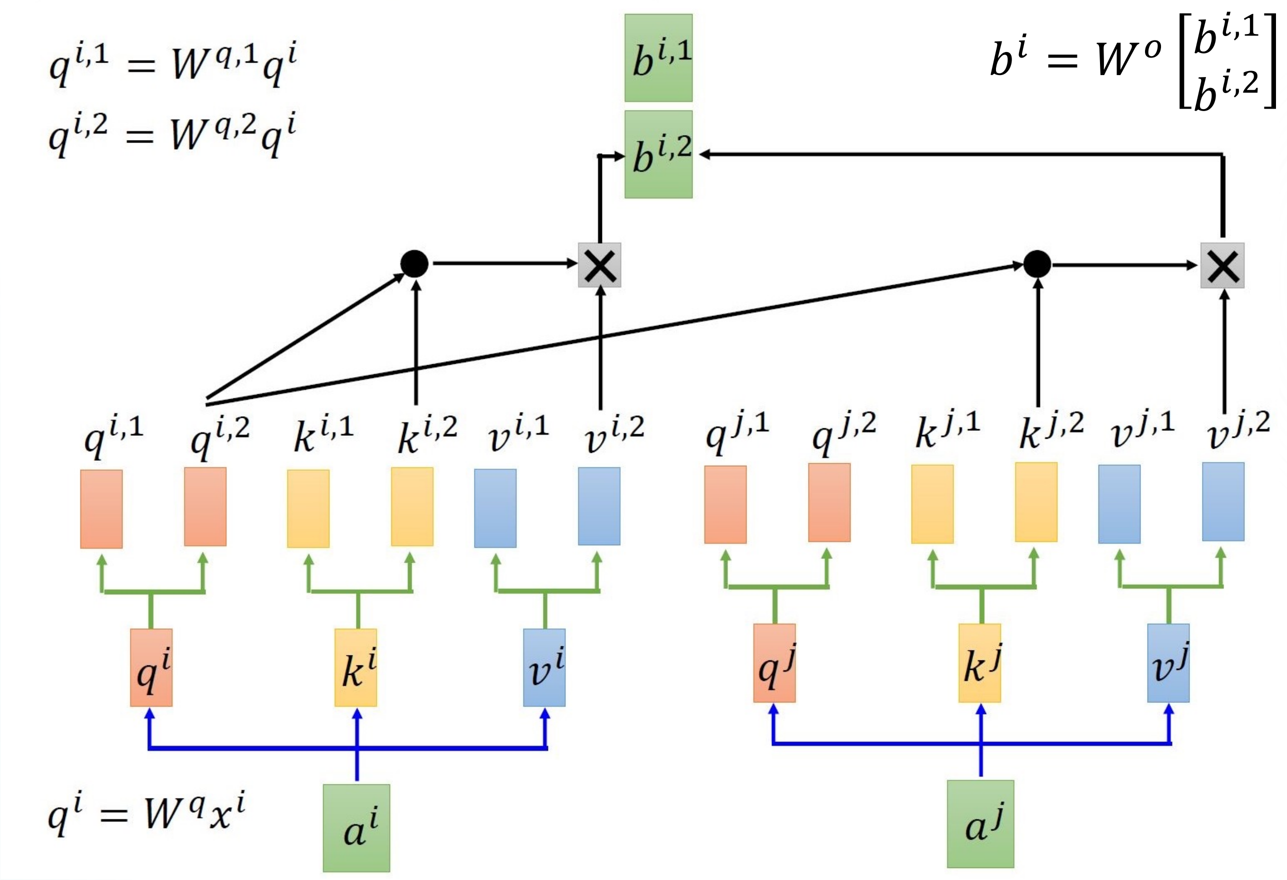
Prior knowledge

Self-attention



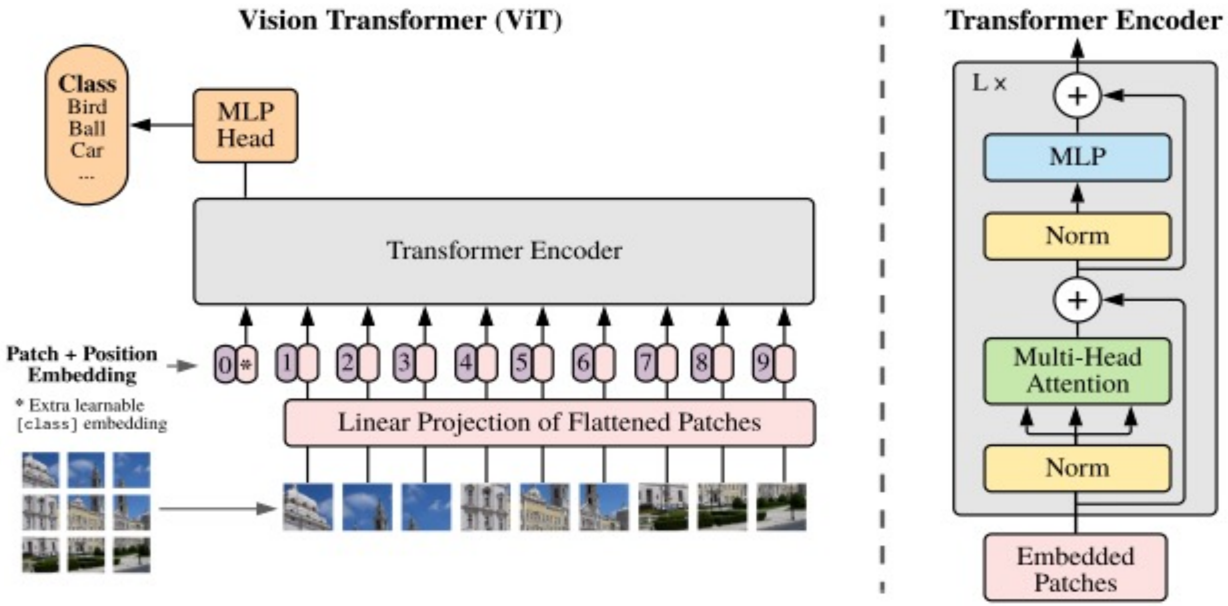
Prior knowledge

Multi-head Self-attention (2 heads)



Motivation

- Not be optimal to treat all samples with the same number of tokens
- 14x14 v.s 4x4
  - Accuracy improve 15.9%
  - Increase the FLOPs by 8.5 times



# of tokens	14x14	7x7	4x4
Accuracy	76.7%	70.3%	60.8%
FLOPs	1.78G	0.47G	0.21G

Motivation

- Automatically configure a decent token number
- Main Mechanism
  - Using increasing number of tokens
  - Early terminate
  - Feature-wise reuse
  - Relation-wise reuse

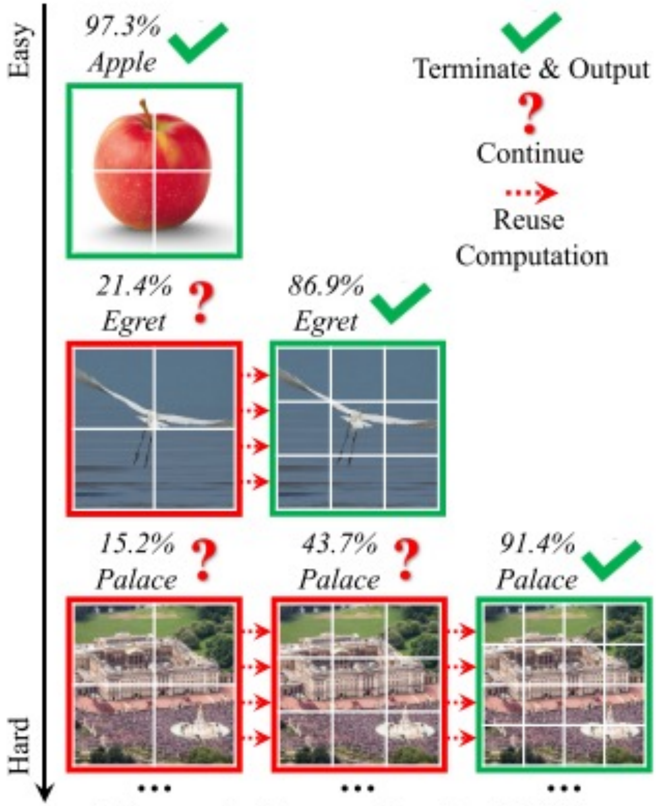
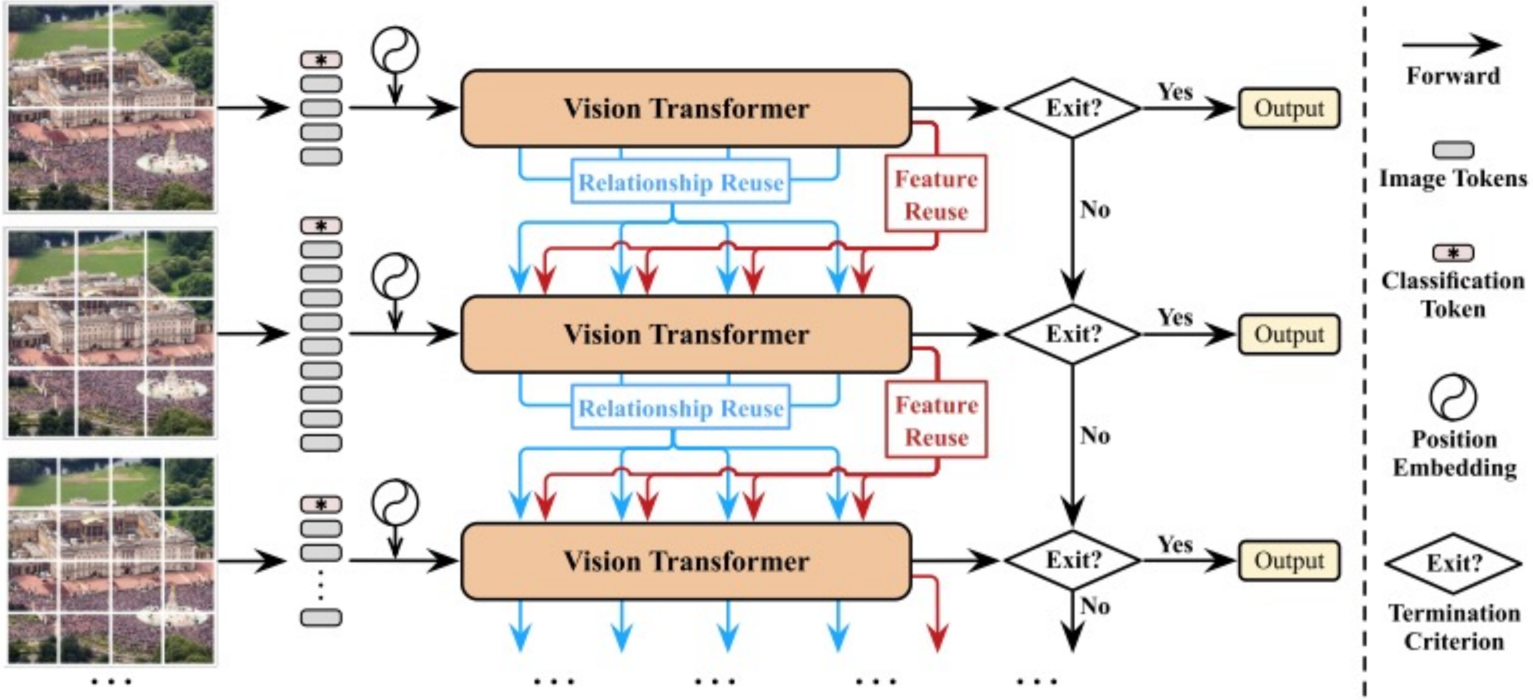


Figure 1: Examples for DVT.

Method

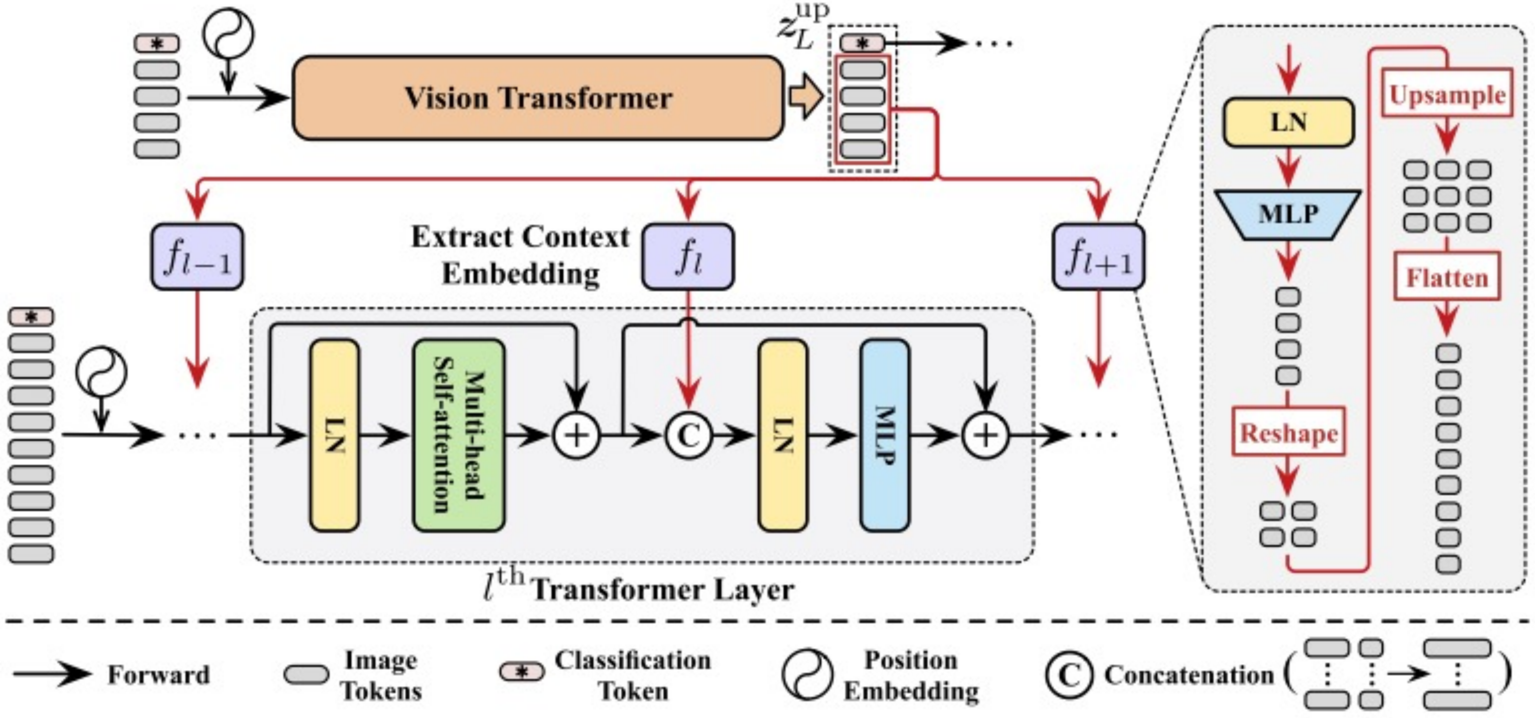


$$minimize \frac{1}{|D_{train}|} \sum_{(x,y)\in D_{train}} \left[\sum_{i} L_{CE}(p_i, y) \right]$$

Method

Feature reuse

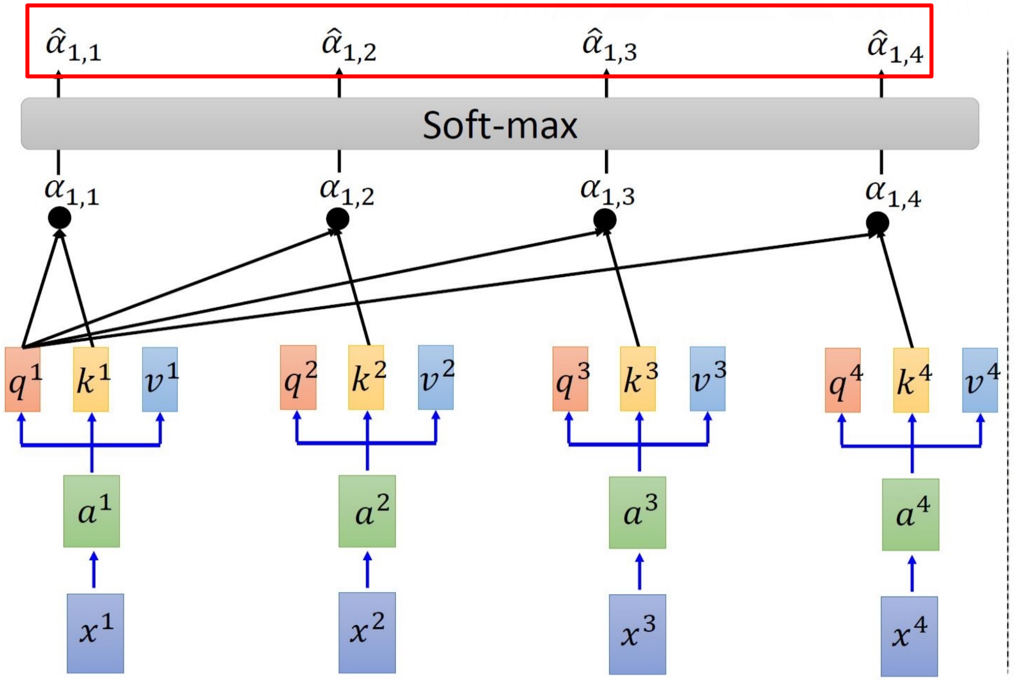
- leverage the image tokens output Z_L^{up} to learn a embedding E_l
  - $E_l = f_l(Z_L^{up}) \in \mathbb{R}^{N \times D'}$



Method

Relationship reuse

- Each layer learns a group of attention maps of the relationships among tokens
- These relationships are helpful in learning the downstream Transformer
- Algorithm
  - $Q_l = z_l W^Q,\ K_l = z_l W^K,\ V_l = z_l W^V$
  - $Attention(z_l) = Softmax(A_l)V_l, A_l = Q_l K_l^T / \sqrt{d}$
  - $A^{up} = Concat\left(A_1^{up}, A_2^{up}, \cdots, A_L^{up}\right) \in \mathbb{R}^{N_{up} \times N_{up} \times N_{up}^{Att}}$
  - $Attention(z_l) = Softmax\left(A_l + r_l(A^{up})\right)V_l$



Method

Early termination

- The i^{th} exit that produces the softmax prediction p_i
- If $\max_j p_{ij} \geq \eta_i$, the inference will stop by adopting the p_i as output
- How to get the values of $\{\eta_1, \eta_2, \cdots\}$
  - Given a computational budget $B > 0$, the optimal thresholds can be optimized by:

$$\underset{\eta_1,\eta_2,\cdots}{\mathrm{max}imize}\, Acc(D_{val}, \{\eta_1, \eta_2, \cdots\}), s.t. FLOPs(D_{val}, \{\eta_1, \eta_2, \cdots\}) \leq B$$

  - genetic algorithm

Experiments

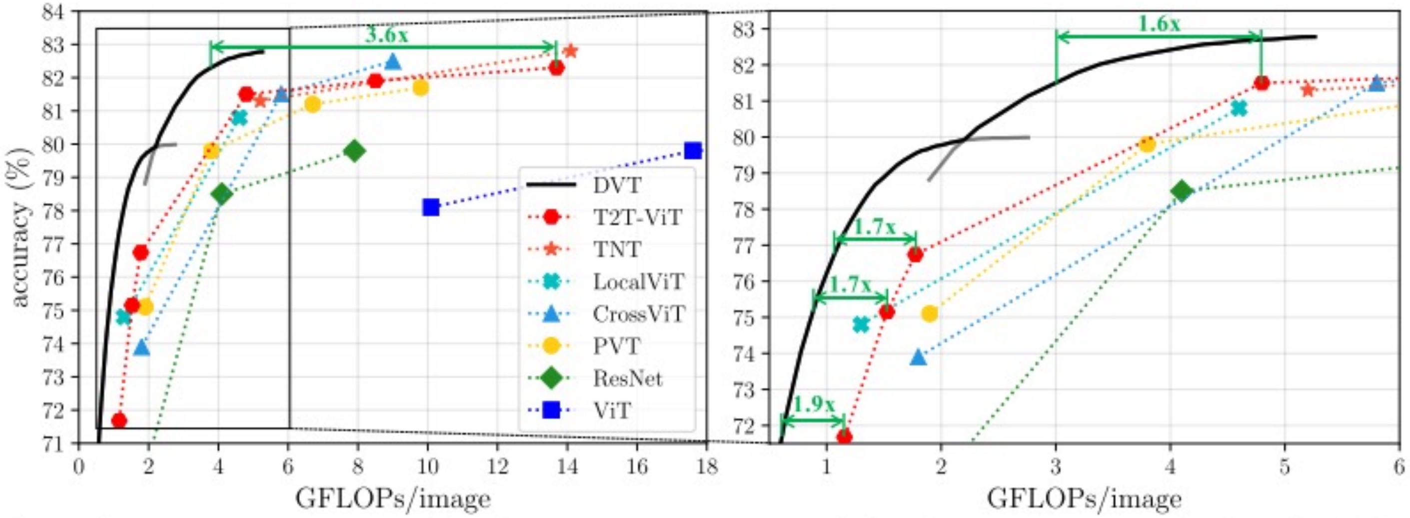


Figure 6: Top-1 accuracy v.s. GFLOPs on ImageNet. DVT is implemented on top of T2T-ViT-12/14.

Table 2: The practical speed of DVT.

Models	ImageNet (NVIDIA 2080Ti, bs=128) Top-1 acc.	Throughput
T2T-ViT-7	71.68%	1574 img/s
DVT	**78.48%** (↑6.80%)	1574 img/s
T2T-ViT-10	75.15%	1286 img/s
DVT	**79.74%** (↑4.59%)	1286 img/s
T2T-ViT-12	76.74%	1121 img/s
DVT	**80.43%** (↑3.69%)	1128 img/s
T2T-ViT-14	81.50%	619 img/s
DVT	81.50%	**877 img/s** (↑1.42x)
T2T-ViT-19	81.93%	382 img/s
DVT	81.93%	**666 img/s** (↑1.74x)

Table 3: Performance of DVT on CIFAR-10/100.

Models	CIFAR-10 Top-1 acc.	CIFAR-10 GFLOPs	CIFAR-100 Top-1 acc.	CIFAR-100 GFLOPs
T2T-ViT-10	97.21%	1.53	85.44%	1.53
DVT	97.21%	**0.50** (↓3.1x)	85.45%	**0.54** (↓2.8x)
T2T-ViT-12	97.45%	1.78	86.23%	1.78
DVT	97.46%	**0.52** (↓3.4x)	86.26%	**0.61** (↓2.9x)
T2T-ViT-14	98.19%	4.80	89.10%	4.80
DVT	98.19%	**0.77** (↓6.2x)	89.11%	**1.62** (↓3.0x)
T2T-ViT-19	98.43%	8.50	89.37%	8.50
DVT	98.43%	**1.44** (↓5.9x)	89.38%	**1.74** (↓4.9x)
T2T-ViT-24	98.53%	13.69	89.62%	13.69
DVT	98.53%	**1.49** (↓9.2x)	89.63%	**1.86** (↓7.4x)

Ablation study

- The three-exit DVT based on T2T-ViT12
- Both the two reuse mechanisms are able to improve the accuracy at the 2nd and 3rd exits
- Computation reusing slightly hurts the accuracy at the 1st exit

Table 4: Effects of feature (F) and relationship (R) reuse. The percentages in brackets denote the additional computation compared to baselines involved by the reuse mechanisms.

Reuse F	Reuse R	1st Exit (7x7) Top-1 acc.	1st Exit (7x7) GFLOPs	2nd Exit (10x10) Top-1 acc.	2nd Exit (10x10) GFLOPs	3rd Exit (14x14) Top-1 acc.	3rd Exit (14x14) GFLOPs
		70.33%	0.47	73.54%	1.37	76.74%	3.15
✓		69.42%	0.47	75.31%	1.43 (4.4%)	79.21%	3.31 (5.1%)
	✓	69.03%	0.47	75.34%	1.41 (2.9%)	78.86%	3.34 (6.0%)
✓	✓	69.04%	0.47	**75.65%**	1.46 (6.6%)	**80.00%**	3.50 (11.1%)

$$\underset{\eta_1,\eta_2,\cdots}{\text{maximize}}\ Acc(D_{val},\{\eta_1,\eta_2,\cdots\}),\ s.t.\ FLOPs(D_{val},\{\eta_1,\eta_2,\cdots\}) \leq B$$

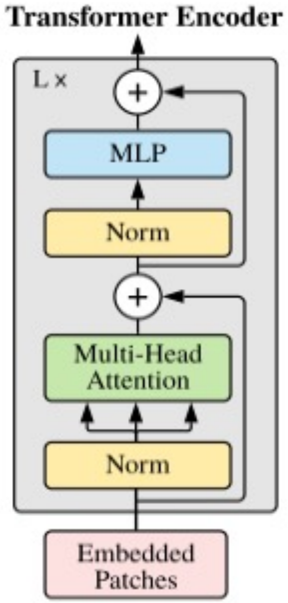
Ablation study

Table 5: Ablation studies for feature reuse.

Ablation	1st Exit (7x7)	2nd Exit (10x10)	
	Top-1 acc.	Top-1 acc.	GFLOPs
w/o reuse	**70.08%**	73.61%	1.37
Layer-wise feature reuse	69.84%	74.31%	1.43
Reuse classification token	69.79%	74.70%	1.43
Remove $f_l(\cdot), l \geq 2$	69.33%	74.73%	1.38
Remove LN in $f_l(\cdot)$	69.63%	75.05%	1.42
Ours	69.44%	**75.23%**	1.43

Table 6: Ablation stuc reuse.

Ablation	0x10) GFLOPs
w/o reuse	1.37
Layer-wise relationship reuse	1.38
Reuse final-layer relationships	1.39
MLP→Linear	1.38
Naive upsample	1.41
Ours	1.41



Ablation study

Table 7: Comparisons of early-termination criterions. The accuracy under each budget is reported.

Ablation	Top-1 acc. 0.75G	1.00G	1.25G	1.50G
Randomly Exit	70.19%	71.66%	72.61%	73.59%
Entropy-based	73.41%	75.21%	77.08%	78.40%
Confidence-based (ours)	**73.70%**	**76.22%**	**77.89%**	**78.89%**