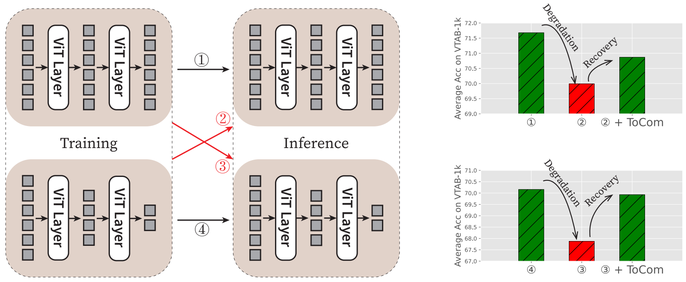
Token Compensator: Altering Inference Cost of Vision Transformer without Re-Tuning

Shibo Jie† Yehui Tang‡ Jianyuan Guo‡ Zhi-Hong Deng† Kai Han‡ Yunhe Wang‡

† School of Intelligence Science and Technology, Peking University ‡ Huawei Noah's Ark Lab

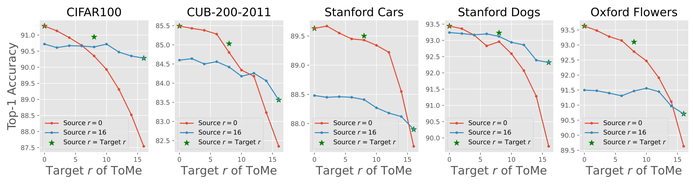
Abstract

- Token compression methods focus on scenario when training and inference compression degrees are consistent (① & ④).
- Performance significantly degrades when compression degrees in training and inference are not equal (② & ③).
- After directly inserting our `ToCom` without re-training, the performance is recovered.



Motivation

The greater the disparity between the compression degrees during training (source degree) and inference (target degree), the more the performance degradation.



However, **the gap between models with different source degrees is transferable across tasks.** $\mathcal{M}_m^{\mathcal{D}_\mathcal{A}}$ denotes model trained with source degree m on dataset $\mathcal{D}_\mathcal{A}$, then

$$\mathcal{M}_m^{\mathcal{D}_\mathcal{A}} - \mathcal{M}_n^{\mathcal{D}_\mathcal{A}} + \mathcal{M}_n^{\mathcal{D}_\mathcal{B}} \approx \mathcal{M}_m^{\mathcal{D}_\mathcal{B}}. \tag{1}$$

Evidence: Evaluated on $\mathcal{D}_\mathcal{B}$ with target degree 16 (ToMe merges 16 tokens per layer). $\mathcal{D}_\mathcal{A}$ is CIFAR100.

Dataset $\mathcal{D}_\mathcal{B}$	$\mathcal{M}_0^{\mathcal{D}_\mathcal{B}}$	$\mathcal{M}_{16}^{\mathcal{D}_\mathcal{A}} - \mathcal{M}_0^{\mathcal{D}_\mathcal{A}} + \mathcal{M}_0^{\mathcal{D}_\mathcal{B}}$
CUB-200-2011	82.4	**83.2**
Stanford Cars	87.6	**87.9**
Stanford Dogs	89.5	**90.3**
Oxford Flowers	89.6	**91.2**

Method

We intend to find a universal plugin to compensate for the gap between models with different source degrees on any datasets.

$$\mathcal{M}_m \oplus \mathcal{P}_{m\to n} = \mathcal{M}'_n \approx \mathcal{M}_n. \tag{2}$$

Challenge: 16×17 choices of (m, n) in Eq. 2 for DeiT-B + ToMe, significant training and storage overheads for all $\mathcal{P}_{m\to n}$.

- Use LoRA as $\mathcal{P}_{m\to n}$

$$\begin{cases} \mathbf{W}_{\mathcal{M}'_n} = \mathbf{W}_{\mathcal{M}_m} + s \cdot \mathbf{AB}, & \text{if } \mathbf{W} \in \{\mathbf{W}_q, \mathbf{W}_v\}, \\ \mathbf{W}_{\mathcal{M}'_n} = \mathbf{W}_{\mathcal{M}_m}, & \text{otherwise}, \end{cases} \tag{3}$$

- Estimate the gap between models only at adjacent compression degrees. When $n > m$:

$$\mathcal{M}_m \oplus \left(\bigoplus_{i=m}^{n-1} \mathcal{P}_{i\to i+1} \right) = \mathcal{M}'_n \approx \mathcal{M}_n. \tag{4}$$

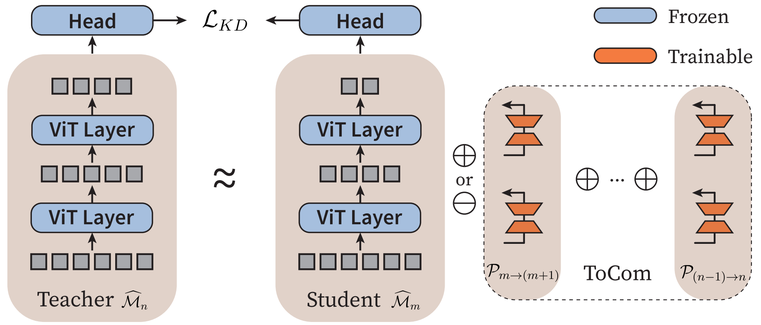
- Suppose the gap between models is invertible, $\mathcal{P}_{n\to m} = \ominus\mathcal{P}_{m\to n}$. When $n < m$:

$$\mathcal{M}_m \ominus \left(\bigoplus_{i=n}^{m-1} \mathcal{P}_{i\to i+1} \right) = \mathcal{M}'_n \approx \mathcal{M}_n, \tag{5}$$

Pre-training with self-distillation on ImageNet for only 10 epochs. Only the plugin $\mathcal{P}$ is trainable.

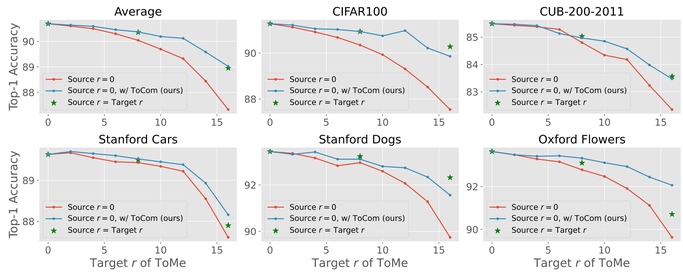
$$\mathcal{L} = \begin{cases} \mathcal{L}_{KD}\left(\widehat{\mathcal{M}}_m \oplus \left(\bigoplus_{i=m}^{n-1} \mathcal{P}_{i\to(i+1)} \right), \widehat{\mathcal{M}}_n \right), & \text{if } n > m \\ \mathcal{L}_{KD}\left(\widehat{\mathcal{M}}_m \ominus \left(\bigoplus_{i=n}^{m-1} \mathcal{P}_{i\to(i+1)} \right), \widehat{\mathcal{M}}_n \right), & \text{if } n < m \end{cases} \tag{6}$$

in which m and n are randomly sampled in each step satisfying $m \neq n$.



Experiments

DeiT-B, ToMe, source r < target r



VTAB-1K	Natural Avg. Acc.	Δ	Specialized Avg. Acc.	Δ	Structured Avg. Acc.	Δ
Target $r = 16$ (1.9× inference speedup)						
Source $r = 0$	73.7		80.0		56.7	
+`ToCom`	75.8	+2.1	83.6	+3.6	57.9	+1.2
Source $r = 4$	74.0		81.0		57.2	
+`ToCom`	75.7	+1.7	83.8	+2.8	58.0	+0.8
Source $r = 16$	75.5		83.6		58.8	
Target $r = 12$ (1.5× inference speedup)						
Source $r = 0$	75.9		82.8		58.6	
+`ToCom`	76.6	+0.7	84.5	+1.7	59.1	+0.5
Source $r = 4$	76.0		83.5		58.8	
+`ToCom`	76.5	+0.5	84.5	+1.0	59.2	+0.4
Source $r = 12$	76.5		84.1		59.3	

DeiT-B, ToMe, source r > target r

VTAB-1K	Natural Avg. Acc.	Δ	Specialized Avg. Acc.	Δ	Structured Avg. Acc.	Δ
Source $r = 16$ (1.9× training speedup)						
Target $r = 0$	76.0		84.0		57.8	
+`ToCom`	76.4	+0.4	84.6	+0.6	59.2	+1.4
Target $r = 4$	76.0		84.2		57.9	
+`ToCom`	76.3	+0.3	84.6	+0.4	59.2	+1.3
Target $r = 16$	75.5		83.6		58.8	
Source $r = 12$ (1.5× training speedup)						
Target $r = 0$	76.9		84.5		59.1	
+`ToCom`	77.0	+0.1	84.9	+0.4	59.4	+0.3
Target $r = 4$	76.8		84.6		59.2	
+`ToCom`	76.9	+0.1	84.9	+0.3	59.5	+0.3
Target $r = 12$	76.5		84.1		59.3	

arXiv: https://www.arxiv.org/pdf/2408.06798

ECCV 2024 - Milan

Contact: parsley@pku.edu.cn