EUROPEAN CONFERENCE ON COMPUTER VISION
MILANO 2024

PYRA: Parallel Yielding Re-Activation for Training-Inference Efficient Task Adaptation

Yizhe Xiong[1,2], Hui Chen[2*], Tianxiang Hao[1,2], Zijia Lin[1], Jungong Han[2,3], Yuesong Zhang[4], Guoxin Wang[4], Yongjun Bao[4], Guiguang Ding[1,2]

1 清华大学软件学院 School of Software, Tsinghua University

2 北京信息科学与技术国家研究中心 BEIJING NATIONAL RESEARCH CENTER FOR INFORMATION SCIENCE AND TECHNOLOGY

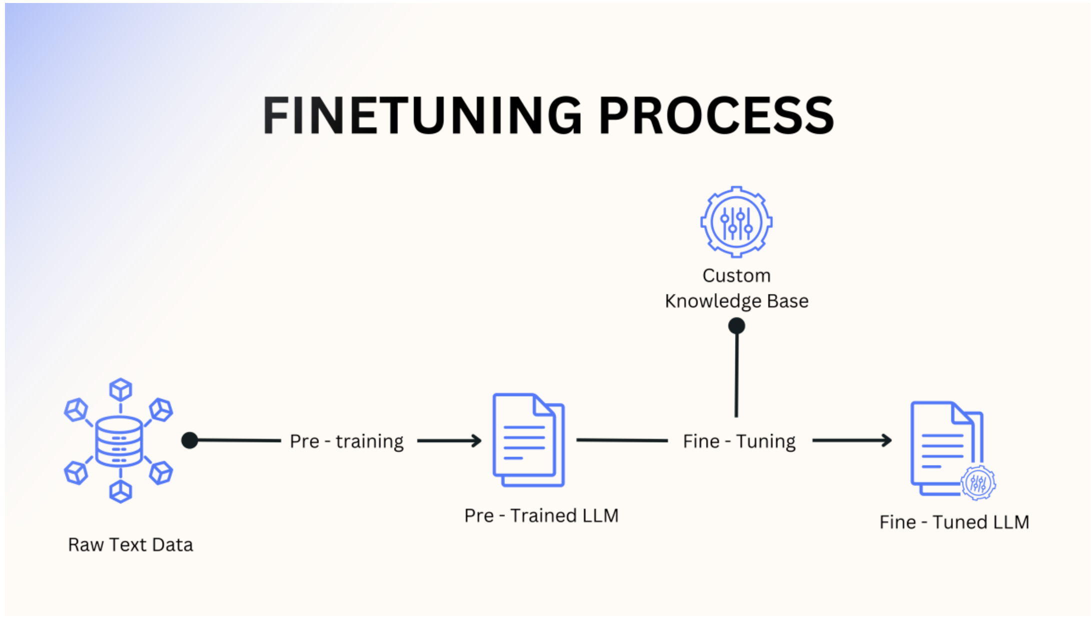
3 清華大學自動化系 TSINGHUA AUTOMATION

4 JD.COM

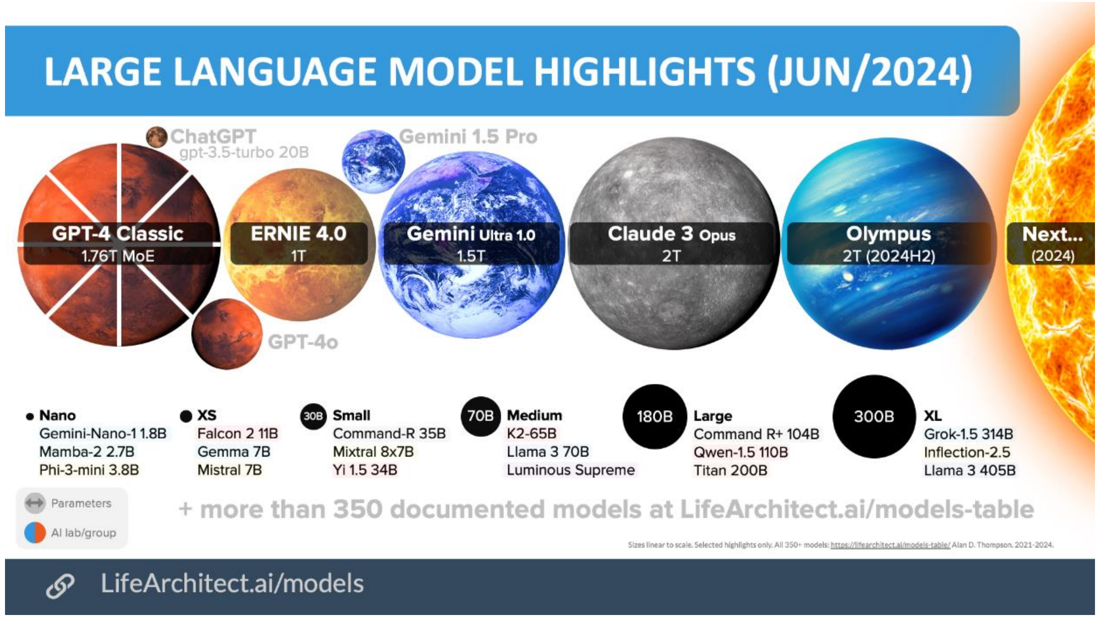
Application of Foundation Models

- Foundation Models (LLaMA, ViT, CLIP, etc.) are fine-tuned on downstream task data for real-world applications.
- The size of Foundation models are rapidly growing.

1



2



Credits:
1: https://www.linkedin.com/pulse/what-supervised-fine-tuning-tagx-yq8if/
2: https://lifearchitect.ai/models/

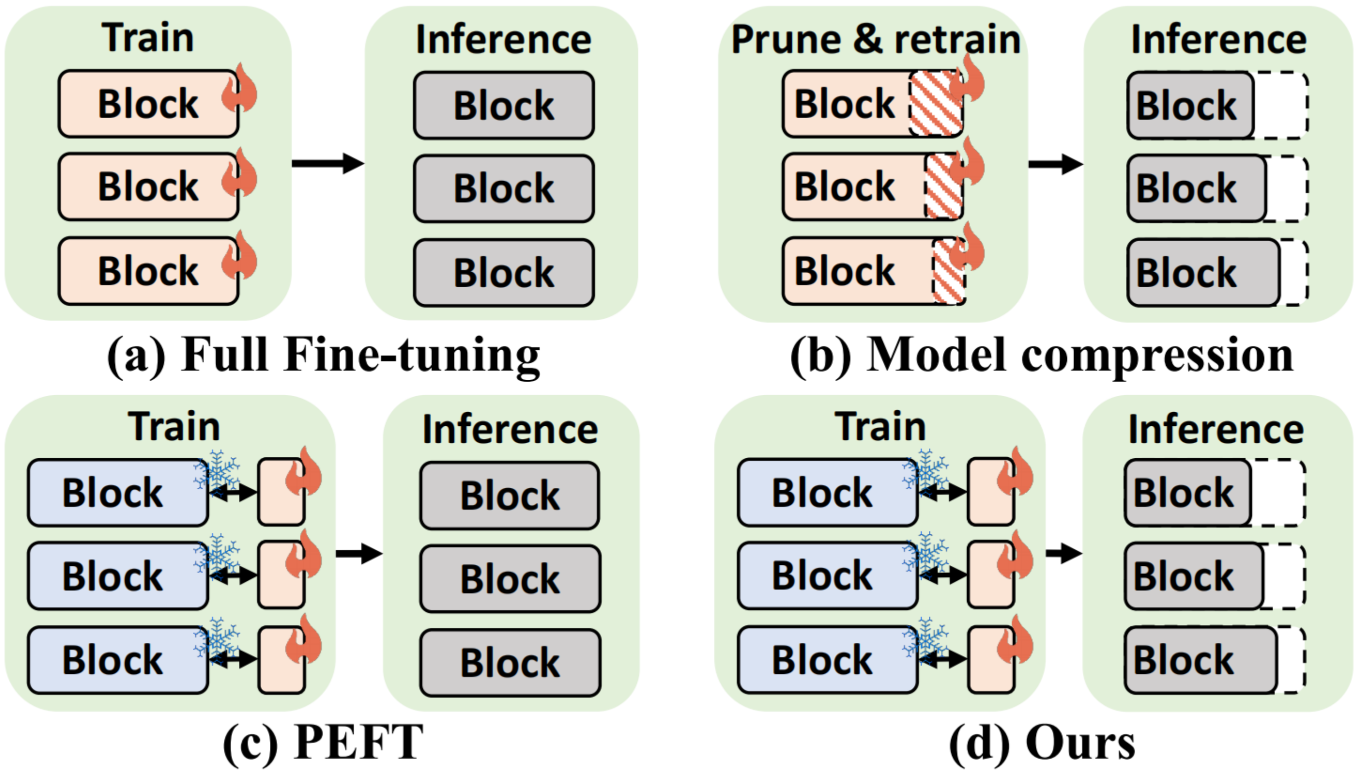
Application of Foundation Models

- Primary obstacles in implementing Foundation Models on downstream tasks:
  - Training overhead when fine-tuning on downstream tasks
  - Inference efficiency after model deployment

- The above two topics are investigated separately:
  - Parameter-Efficient Fine-Tuning (PEFT): train small amount of parameters during fine-tuning
  - Model Compression: model pruning, knowledge distillation, model quantization, etc.

Motivation: Efficency in Training & Inference?

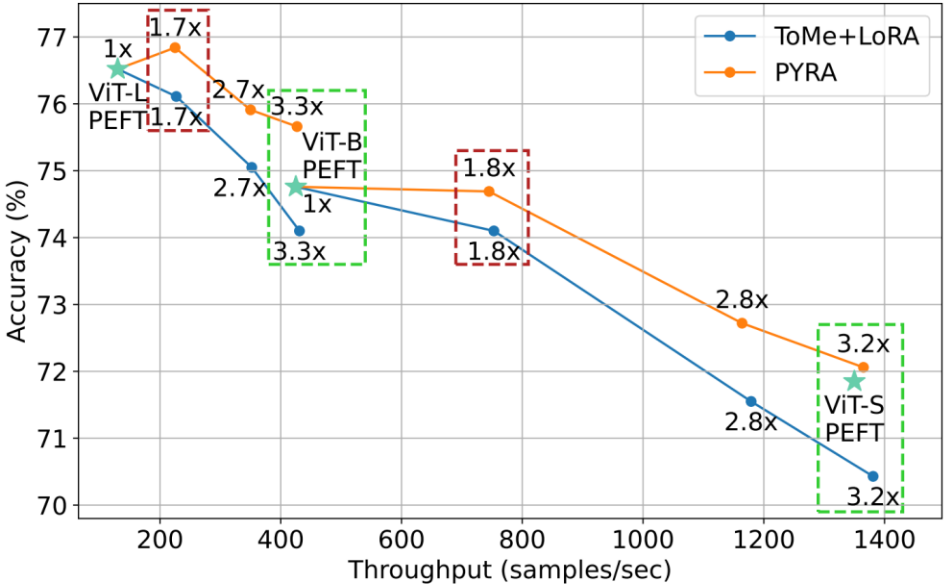
- However, neither PEFT or Model Compression solves both problems!
- Can we achieve both training efficiency and inference efficiency simultaneously?
- We formalize this challenge as **Training-Inference Efficient Task Adaptation**.



Method	(a)	(b)	(c)	(d)
Non-vast Data Scale	✅	❌	✅	✅
Training Efficient	❌	❌	✅	✅
Inference Efficient	❌	✅	❌	✅

Problems of the Baseline approach

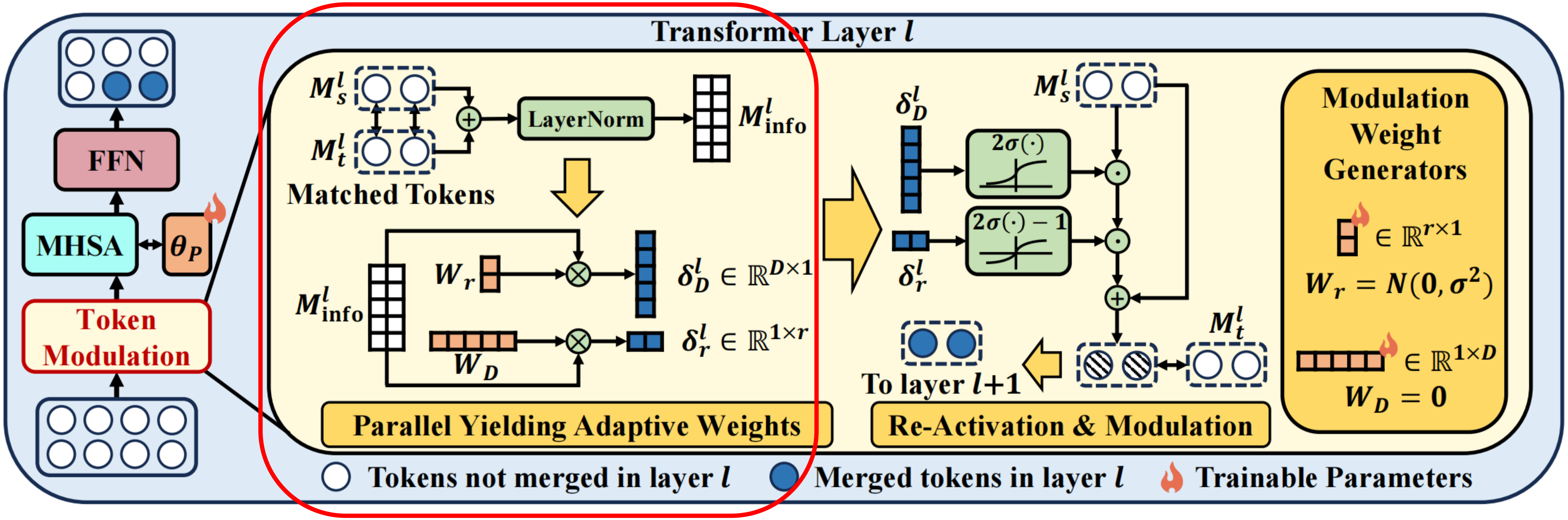
- Baseline solution: a mergeable PEFT method + a training-free compression method
  - Here we choose LoRA[1] + Token merging[2] (a strong baseline according to our experiments)
- Quick evaluation of the baseline approach:
  - **Un-ignorable performance drops** under low compression rates (1.7 to 1.8 times speedup)
  - "**Adverse Compression**": Under-performs "smaller backbone with similar throughput + PEFT" under high compression rates (over 3 times speedup)



Credits:
1: LoRA: Low-Rank Adaptation of Large Language Models (ICLR 2022)
2: Token Merging: Your ViT But Faster (ICLR 2023)

PYRA: Parallel Yielding Re-Activation

- Parallel Yielding Adaptive Weights:
  - For the merging token pairs, we group the source M_s^l, the target M_t^l, then creating
$$M_{\text{info}}^l = \text{LayerNorm}(M_s^l + M_t^l) \in \mathbb{R}^{D\times r}$$
  - Then, we use learnable Modulation Weight Generators to create adaptive weights:
$$\delta_D^l = M_{\text{info}}^l W_r^l \in \mathbb{R}^{D\times 1} \quad \delta_r^l = W_D^l M_{\text{info}}^l \in \mathbb{R}^{1\times r}$$

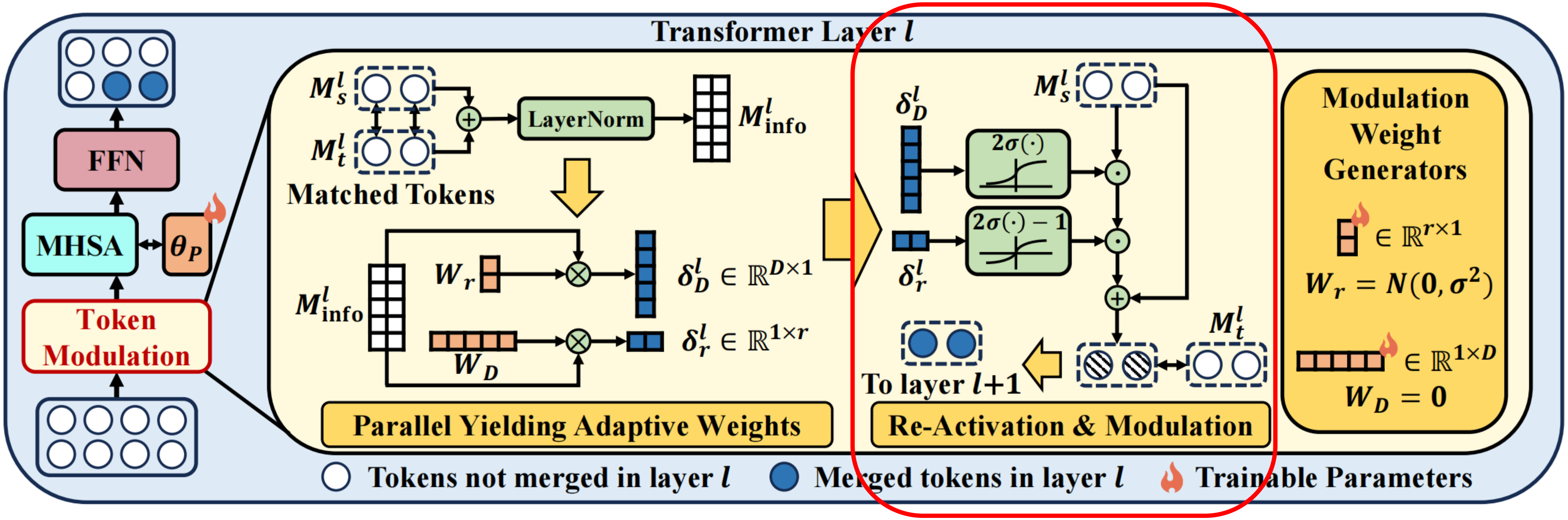


PYRA: Parallel Yielding Re-Activation

- Re-Activation & Modulation:
  - To regularize the adaptive weights and create non-linearity for expression capacity, we use a re-activation before token modulation. We add a residual connection for better gradients.

$$\widehat{M}_s^l = 2\sigma(\hat{\delta}_D^l)\odot M_s^l$$

$$M_s^l \leftarrow M_s^l + \left(2\sigma\left(\hat{\delta}_r^l\right) - 1\right)\odot \widehat{M}_s^l$$



PYRA: Parallel Yielding Re-Activation

- PYRA is extremely parameter-efficient and compute-efficient:
  - Low compression rate: **50% FLOPs decrease, less than 0.01% trainable parameters** (Table 1)
  - High compression rate: **>75% FLOPs decrease, less than 0.01% trainable parameters**

Table 1: The complexity comparisons between conducting PEFT with and without PYRA. The FLOPs metric is obtained during inference.

Model	Metric	Total	PEFT w/o PYRA	(%)	PEFT w. PYRA	(%)
ViT-Base	# params	86M	0.29M	0.34%	0.30M	0.35%
	FLOPs	16.37G	16.37G	100%	8.15G	49.79%
ViT-Large	# params	303M	1.18M	0.39%	1.20M	0.40%
	FLOPs	57.37G	57.37G	100%	28.76G	50.13%

Experiment: Overall Comparison

- Low compression rate: we choose LoRA as PEFT for all methods
  - Outperforms the best baseline: ToMe + LoRA
  - Achieve comparable performance or even out-performs backbone + LoRA

Benchmark: VTAB-1k

Method	# params	Throughput	Natural: Cifar100	Caltech101	DTD	Flowers102	Pets	SVHN	Sun397	Specialized: Camelyon	EuroSAT	Resisc45	Retinopathy	Structured: Clevr-Count	Clevr-Dist	DMLAB	KITTI-Dist	dSpr-Loc	dSpr-Ori	sNORB-Azim	sNORB-Ele	Average
Model: ViT-B/16 (Throughput: 425)																						
PEFT	0.34%	425	67.1	90.2	69.4	99.1	90.5	85.7	54.1	83.1	95.8	84.3	74.6	82.2	69.2	50.1	79.2	81.8	47.1	31.1	42.6	74.76
RaP	3.43%	654	25.9	68.4	53.3	64.0	57.4	71.3	21.5	75.8	87.9	59.3	73.6	43.1	53.8	26.3	60.5	73.5	25.5	16.7	27.9	55.57
SPViT	4.46%	567	41.6	75.4	61.1	83.2	66.2	56.1	28.3	79.3	94.2	73.3	73.6	70.6	61.5	42.4	67.8	75.4	**50.5**	28.9	31.3	64.16
DiffRate	0.35%	709	37.1	84.6	63.7	96.7	86.2	32.6	48.2	78.9	85.8	67.0	73.7	32.9	29.8	34.1	55.7	12.6	16.0	13.1	21.5	55.82
ToMe	0.34%	753	64.6	**90.4**	67.9	98.5	89.8	83.9	**53.2**	82.6	94.7	**83.5**	74.9	81.9	**69.8**	49.2	76.9	**81.9**	46.5	31.0	**43.1**	74.10
PYRA	0.35%	745	**67.5**	90.3	**69.3**	**98.9**	**90.0**	**84.6**	53.1	**83.3**	**95.7**	83.3	**75.2**	**82.6**	68.9	**50.8**	**80.0**	81.8	45.8	**32.2**	42.8	**74.69**
Model: ViT-L/16 (Throughput: 130)																						
PEFT	0.39%	130	77.1	91.4	73.4	99.5	91.3	89.6	57.6	85.9	96.1	87.3	76.1	83.1	63.0	50.7	82.1	81.7	53.5	32.2	36.6	76.52
RaP	1.95%	196	43.2	87.9	62.6	52.8	81.7	86.7	34.7	78.4	92.4	73.3	73.6	68.0	59.6	46.9	82.4	75.5	43.6	24.5	25.7	65.64
SPViT	2.47%	188	48.1	87.5	65.2	94.4	77.4	80.9	38.8	79.9	93.9	79.8	74.3	78.2	**65.8**	47.4	74.1	82.3	50.3	31.0	37.9	70.22
DiffRate	0.39%	221	50.9	86.8	70.3	97.8	88.3	39.0	52.3	80.2	87.2	72.2	74.2	32.6	32.3	36.5	57.4	22.8	26.6	15.2	23.4	59.53
ToMe	0.39%	227	76.1	91.1	72.3	99.2	**91.7**	89.2	56.4	86.4	95.1	86.6	75.1	82.4	61.9	50.9	81.4	81.6	**53.5**	33.4	36.8	76.11
PYRA	0.40%	225	**76.6**	**91.3**	**73.2**	**99.3**	91.5	**89.4**	**57.1**	**86.9**	**95.9**	**87.1**	**76.2**	**83.2**	63.2	**52.8**	**83.1**	**82.5**	52.6	**34.8**	**39.0**	**76.84**

Experiment: Overall Comparison

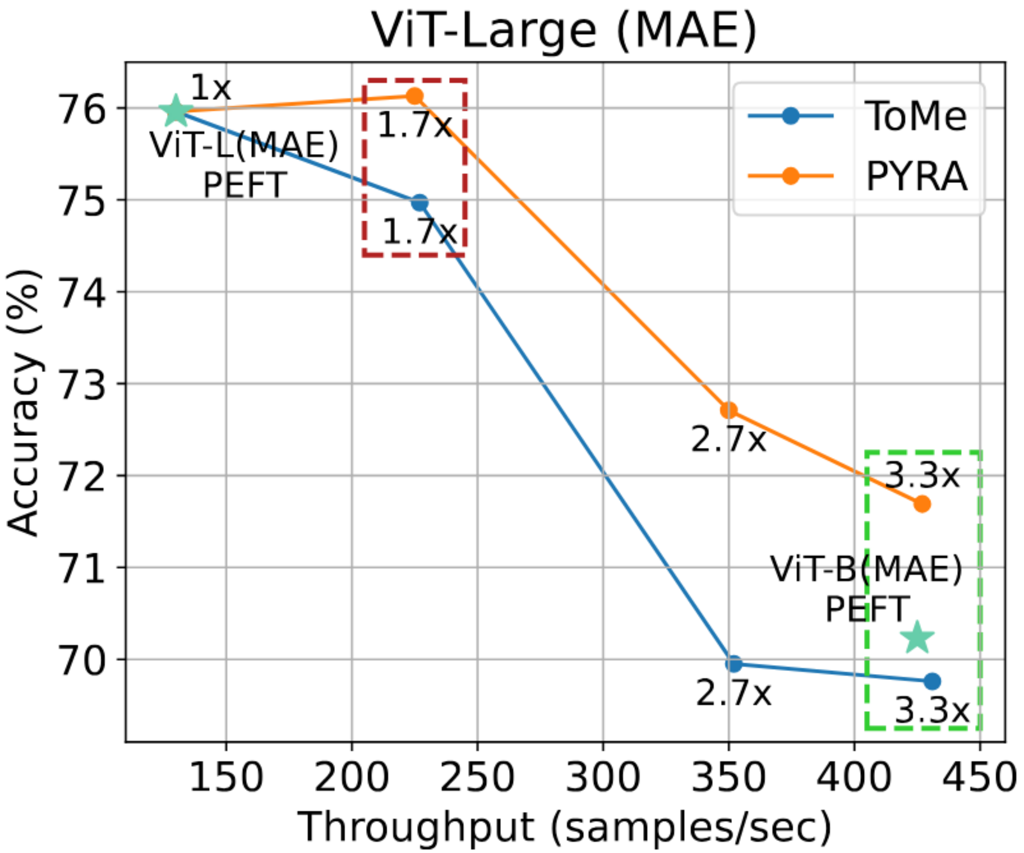
- High compression rate: we choose LoRA as PEFT for all methods
  - Outperforms the best baseline: ToMe + LoRA
  - Solves the “Adverse Compression” issue: outperforms smaller backbone + LoRA

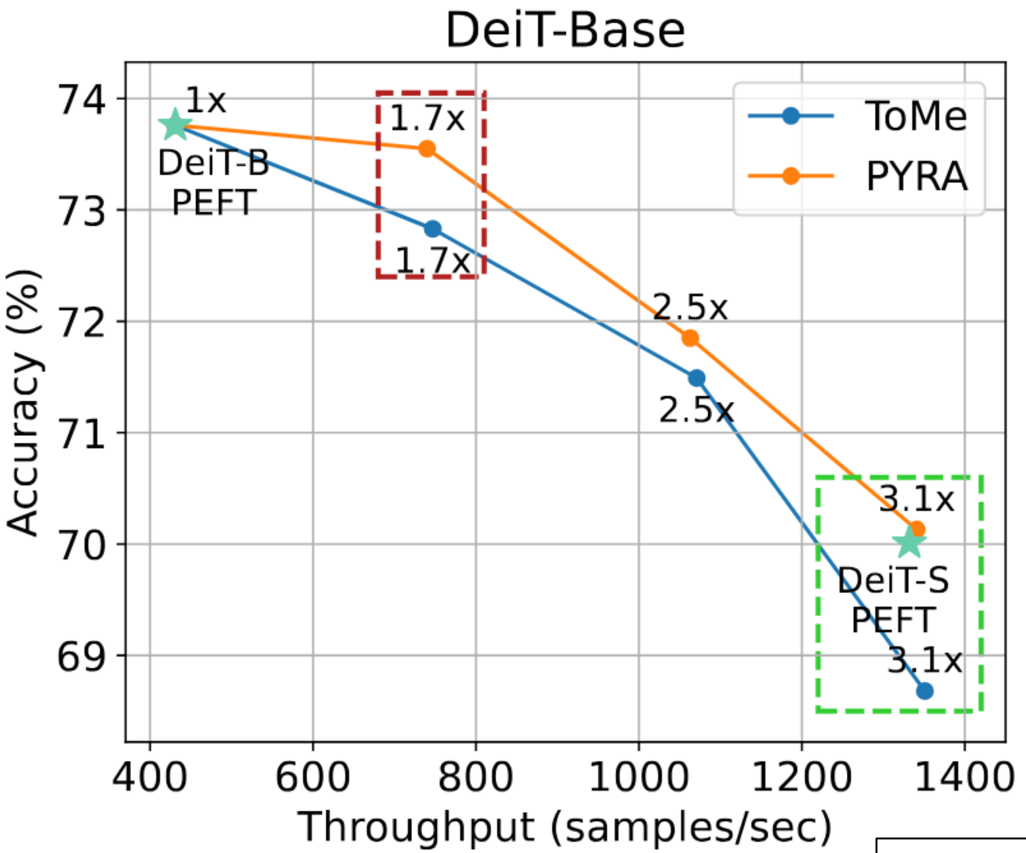
Benchmark: VTAB-1k

			Natural							Specialized				Structured								
Method	# params	Throughput	Cifar100	Caltech101	DTD	Flowers102	Pets	SVHN	Sun397	Camelyon	EuroSAT	Resisc45	Retinopathy	Clevr-Count	Clevr-Dist	DMLAB	KITTI-Dist	dSpr-Loc	dSpr-Ori	sNORB-Azim	sNORB-Ele	Average
Model: ViT-B/16 (Throughput: 425)																						
PEFT*	0.34%	1350	57.7	88.2	70.1	98.7	88.7	85.7	44.9	81.4	94.7	84.6	73.6	81.6	64.1	48.1	80.0	72.9	38.4	22.9	37.7	71.85
RaP	0.86%	1029	24.3	40.1	34.5	41.8	40.5	21.7	11.4	75.8	86.5	35.1	73.8	49.6	49.7	28.1	39.4	13.8	15.4	12.4	26.9	42.60
SPViT	4.46%	944	23.7	67.9	51.9	69.9	53.2	19.6	13.1	71.9	81.3	67.9	74.7	53.5	61.9	39.5	57.4	45.0	34.5	11.1	23.2	52.49
DiffRate	0.35%	1308	23.2	73.0	55.7	87.9	66.7	27.2	29.3	78.1	77.8	53.1	73.6	29.7	28.6	31.7	52.6	11.5	17.2	11.3	20.3	49.29
ToMe	0.34%	1381	**54.2**	87.8	65.5	96.1	81.7	79.7	**45.2**	79.4	93.6	76.3	73.8	78.3	65.7	48.0	71.3	80.0	45.8	30.9	41.2	70.43
PYRA	0.35%	1365	54.0	**89.3**	**67.1**	**96.5**	**84.0**	**81.8**	44.6	**81.2**	**94.6**	**79.5**	**75.1**	**79.9**	**67.0**	**49.2**	**76.9**	**82.6**	**47.8**	**31.9**	**42.0**	**72.06**
Model: ViT-L/16 (Throughput: 130)																						
PEFT*	0.34%	425	67.1	90.2	69.4	99.1	90.5	85.7	54.1	83.1	95.8	84.3	74.6	82.2	69.2	50.1	79.2	81.8	47.1	31.1	42.6	74.76
RaP	0.65%	301	17.7	37.1	27.0	46.2	33.3	23.2	13.3	76.5	74.2	54.4	73.6	50.4	31.4	25.7	49.8	53.1	25.5	13.4	26.0	44.11
SPViT	2.47%	289	54.0	87.6	65.5	94.8	74.9	32.6	38.6	81.8	**95.3**	78.0	74.0	72.8	61.2	46.9	70.2	77.1	47.4	31.3	28.6	66.90
DiffRate	0.39%	416	47.4	73.5	54.1	84.3	60.2	19.6	22.2	50.0	64.6	42.8	18.2	31.5	31.9	31.1	37.3	22.0	17.7	14.8	21.4	40.48
ToMe	0.39%	431	71.0	90.9	70.4	98.3	88.5	87.2	**52.4**	82.9	94.5	83.1	75.0	80.7	61.1	48.9	76.9	80.8	53.0	32.1	35.2	74.10
PYRA	0.40%	427	**71.6**	**91.8**	**71.1**	**98.5**	**89.7**	**88.1**	52.2	**85.1**	**95.3**	**84.6**	**75.7**	**80.9**	**63.0**	**51.7**	**82.0**	**82.0**	**54.2**	**36.0**	**41.2**	**75.66**

Experiment: on more backbones

- On self-supervised backbone (ViT-L MAE) and distillation backbone (DeiT-B), the issues in low/high compression rates are resolved as well.
- For more analysis experiments, please refer to the article and the appendix.





Benchmark: VTAB-1k

Summary

- We define and introduce Training-Inference Efficient Task Adaptation, in which the inference efficiency and training efficiency are considered for applying foundation models on downstream tasks.
- We investigate the proposed challenge, and discovered that the best performing baselines exhibit issues under both low and high compression rates.
- We propose Parallel Yielding Re-Activation (PYRA), a light-weight method that adaptively modulates tokens in Vision Transformers.
- Extensive experiments show that PYRA consistently resolves the issues in both low and high compression rates on various Vision Transformer backbones, for the first time making Training-Inference Efficient Task Adaptation actually applicable.

Contact me at: xiongyizhe2001@163.com My homepage: xiongyizhe.xyz