ICML
International Conference
On Machine Learning

A Closer Look at Self-Supervised Lightweight Vision Transformers

Shaoru Wang, Jin Gao*, Zeming Li, Xiaoqin Zhang, Weiming Hu

Institute of Automation, Chinese Academy of Sciences   Megvii Research   Wenzhou University

Background and Motivations

- Pre-Training can significantly improve performance of **large models** on various downstream tasks, so what about **lightweight vision models**?
- Lightweight vision models are essential for practical scenarios
  - Must be deployed on edge devices due to data privacy, real-time requirement, …

An empirical study of the **pre-training of lightweight ViTs**

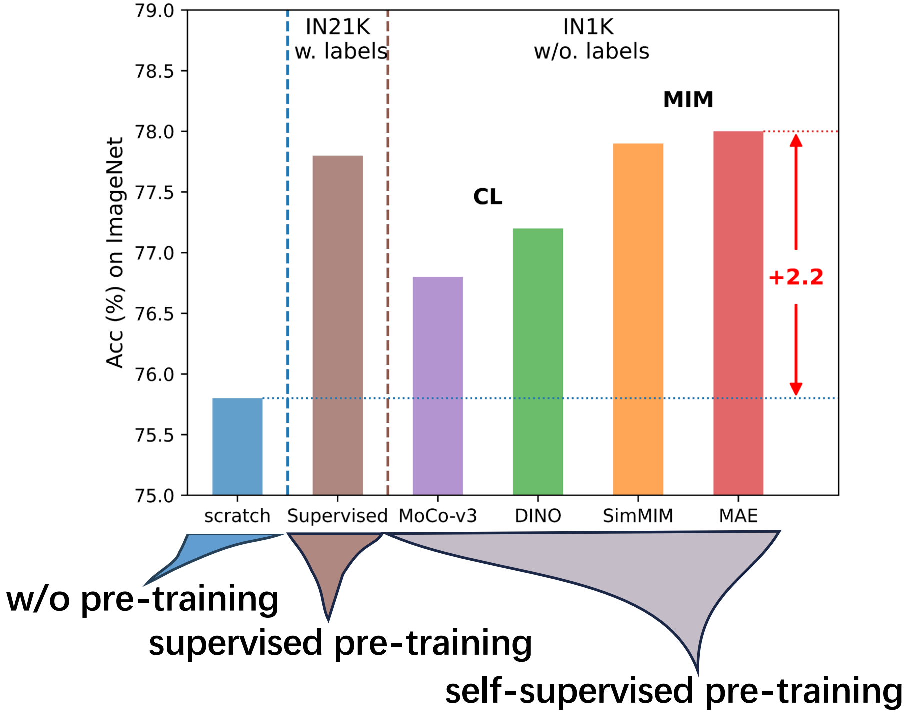
① **A practical guide** on how to choose pre-training schemes for various downstream scenarios;

② **Analyses on the distinct behaviors** of pre-trained models from different methods, e.g., CL(Contrastive Learning) and MIM(Masked-Image-Modeling);

③ **A pre-training distillation approach** that can significantly improve the MIM-based pre-trained models.

How Well Does Pre-Training Work on Lightweight ViTs?

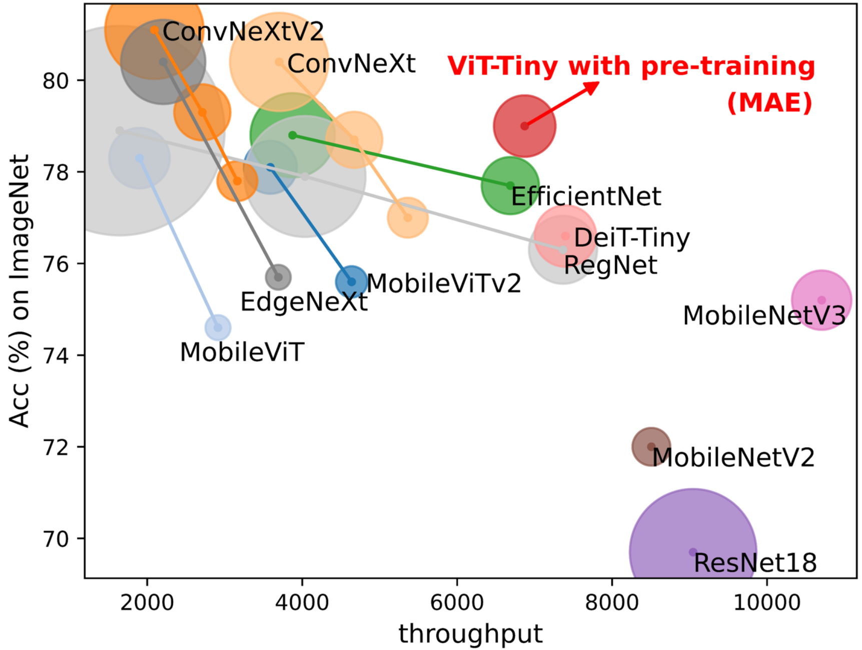
- **ViT-Tiny**: vanilla architecture, 5.7M parameters



- Pre-Training can also help lightweight ViTs to achieve **better** downstream classification performance on **ImageNet**.
- When downstream tasks are with sufficient labeled data, **MAE (Masked Auto-Encoders) is preferred**, which contributes to the most gains.

How Well Does Pre-Training Work on Lightweight ViTs?

Proper Pre-Training Helps Vanilla ViTs Beat SOTA Networks!



- The enhanced ViT-Tiny is **on par with or even outperforms** most previous ConvNets and ViT derivatives.
- Based on a **naive network architecture**, one can also **achieve SOTA** by adopting **proper pre-training**, rather than introducing sophisticated components into the architecture design.

How Well Does Pre-Training Work on Lightweight ViTs?

Downstream Data Scale Matters!

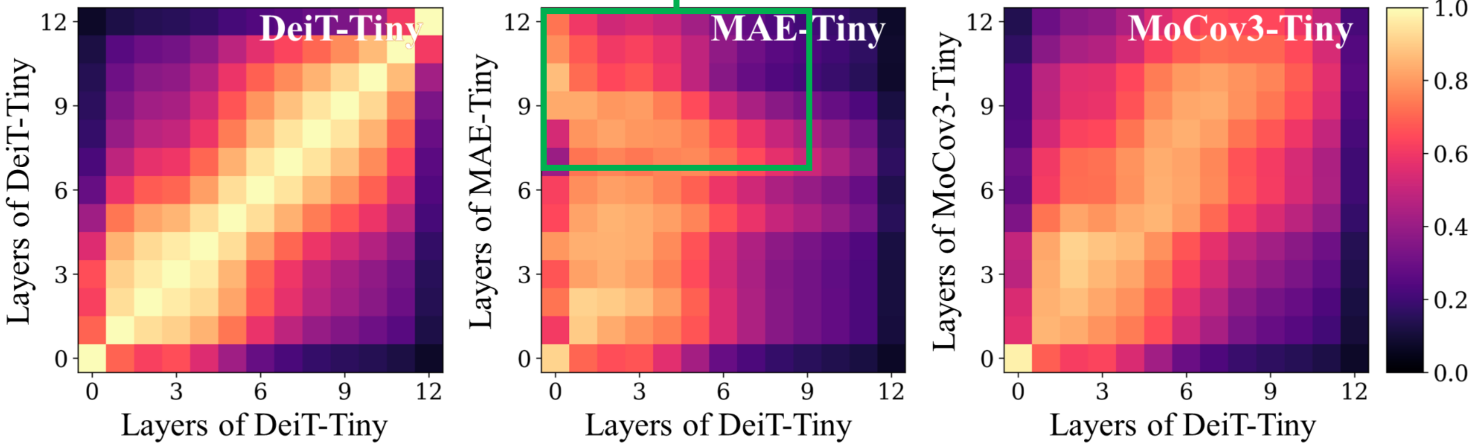
Init. \ Datasets	Flowers (2k/6k/102)	Pets (4k/4k/37)	Aircraft (7k/3k/100)	Cars (8k/8k/196)	CIFAR100 (50k/10k/100)	iNat18 (438k/24k/8142)	COCO(det.) (118k/50k/80)	COCO(seg.)
Random	30.2	26.1	9.4	6.8	42.7	58.7	32.7	28.9
supervised DeiT-Tiny	**96.4**	**93.1**	73.5	**85.6**	**85.8**	**63.6**	**40.4**	**35.5**
self-supervised MoCov3-Tiny	94.8	87.8	**73.7**	83.9	83.9	54.5	39.7	35.1
MAE-Tiny	85.8	76.5	64.6	78.8	78.9	60.6	39.9	35.4

- Self-supervised pre-training performs not well on data-insufficient downstream classification tasks and dense-prediction tasks.

Revealing the Secrets of the Pre-Training

Representation similarity between the layers across networks

- DeiT-Tiny (A supervised pre-trained ViT-Tiny) as the reference model

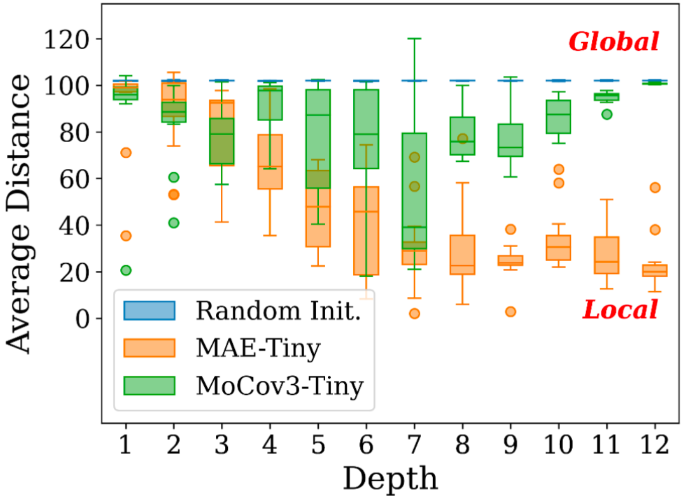


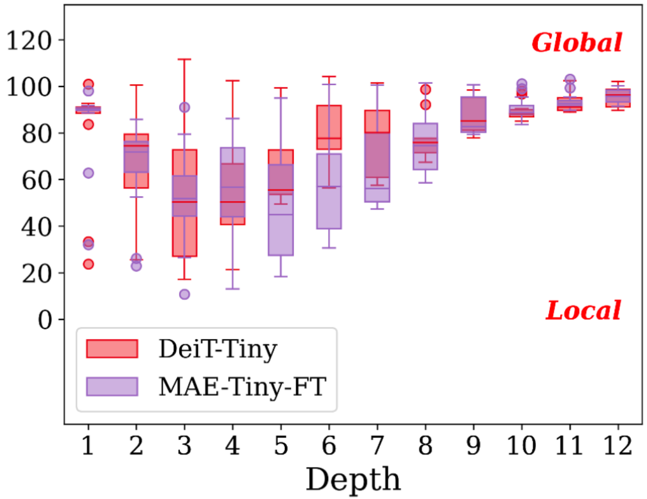
MIM-based Pre-Trained Models Degrade on High Layers!

- ➢ Lower layers matter more than higher ones if sufficient downstream data is provided
- ➢ Higher layers matter in data-insufficient downstream tasks

Revealing the Secrets of the Pre-Training

Attention analyses for the pre-trained models





Average Attention Distance

$$\boldsymbol{D}_{h,j} = \sum_{i} \mathrm{softmax}(\boldsymbol{A}_h)_{i,j} \boldsymbol{G}_{i,j}$$

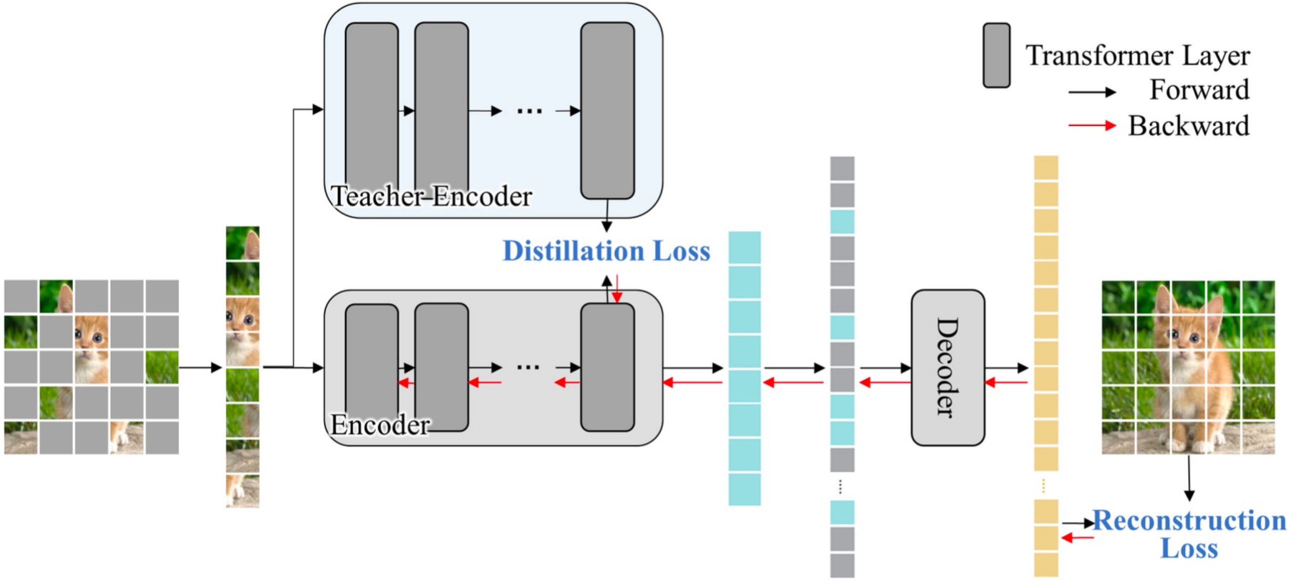
$\boldsymbol{A}$: Attention map
$\boldsymbol{G}$: Euclidean distance

- The pre-training introduces locality inductive bias!
- The pre-training with MAE makes the attention of the downstream models more local and concentrated.

A Pre-Training Distillation Approach

Solution: **Pre-Training distillation based on MAE**

- Improve the quality of higher layers with the help of pre-trained teacher models



- Based on MAE;
- MAE-Base as the teacher;
- Distill on the attention maps;

$$L_{\mathrm{attn}} = \mathrm{MSE}(\boldsymbol{A}^T, \boldsymbol{M}\boldsymbol{A}^S)$$

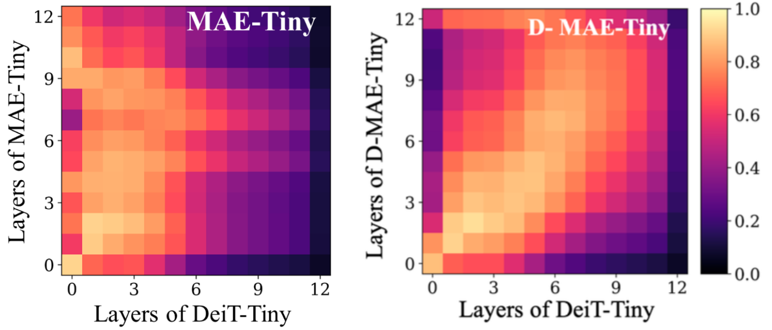
- Distill on the corresponding higher layers of the teacher and student.

A Pre-Training Distillation Approach

Distillation improves downstream performance!

ImageNet

Methods	Data	Top-1 Acc. (%)
from scratch	-	75.8
Supervised (Steiner et al., 2021)	IN21K w/ labels	76.9
Supervised (Steiner et al., 2021)	IN21K w/ labels	77.8
MoCo-v3 (Chen et al., 2021a)	IN1K w/o labels	76.8
MAE (He et al., 2021)	IN1K w/o labels	78.0
DINO (Caron et al., 2021)	IN1K w/o labels	77.2
SimMIM (Xie et al., 2022)	IN1K w/o labels	77.9
D-MAE-Tiny (ours)	IN1K w/o labels	**78.4**



Other Vision Tasks

Init. \ Datasets	Flowers (2k/6k/102)	Pets (4k/4k/37)	Aircraft (7k/3k/100)	Cars (8k/8k/196)	CIFAR100 (50k/10k/100)	iNat18 (438k/24k/8142)	COCO(det.) (118k/50k/80)	COCO(seg.) (118k/50k/80)
supervised								
DeiT-Tiny	**96.4**	**93.1**	73.5	85.6	**85.8**	63.6	40.4	35.5
self-supervised								
MoCov3-Tiny	94.8	87.8	73.7	83.9	83.9	54.5	39.7	35.1
MAE-Tiny	85.8	76.5	64.6	78.8	78.9	60.6	39.9	35.4
DINO-Tiny	95.6	89.3	73.6	84.5	84.7	58.7	41.4	36.7
SimMIM-Tiny	77.2	68.9	55.9	70.4	77.7	60.8	39.3	34.8
D-MAE-Tiny (ours)	95.2	89.1	**79.2**	**87.5**	85.0	**63.6**	**42.3**	**37.4**

Conclusion

Summary

An empirical study of the **pre-training of lightweight ViTs**

- A practical guide
- Analyses on the pre-trained models
- A pre-training distillation approach

Paper

Code

jin.gao@nlpr.ia.ac.cn