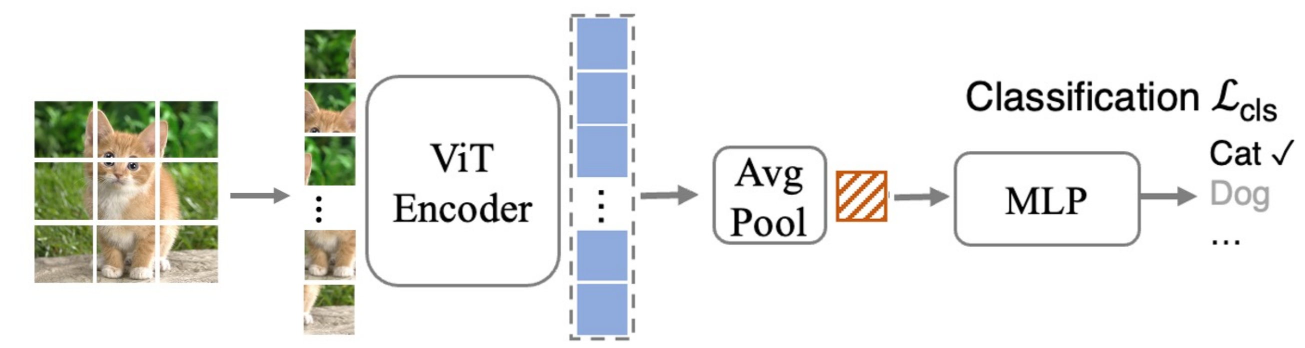
SupMAE: Supervised Masked Autoencoders Are Efficient Vision Learners

Feng Liang[1], Yangguang Li[2], Diana Marculescu[1]
[1]The University of Texas at Austin [2]SenseTime Research

The 38th Annual AAAI Conference on Artificial Intelligence
FEBRUARY 20-27, 2024 | VANCOUVER, CANADA

Vision transformer is difficult to train

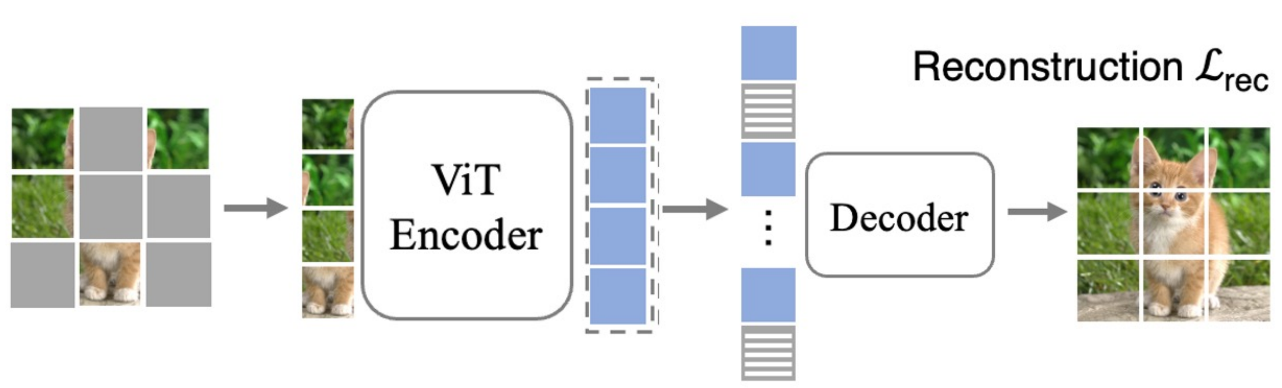
- **Supervised training**



DeiT [H. Touvron *et. al.*]

	Training time*	ImageNet acc.
* Time is measure on 8 A5000 GPUs	91.5 hours	81.8
	✓	✗

- **Self-upervised pre-training**

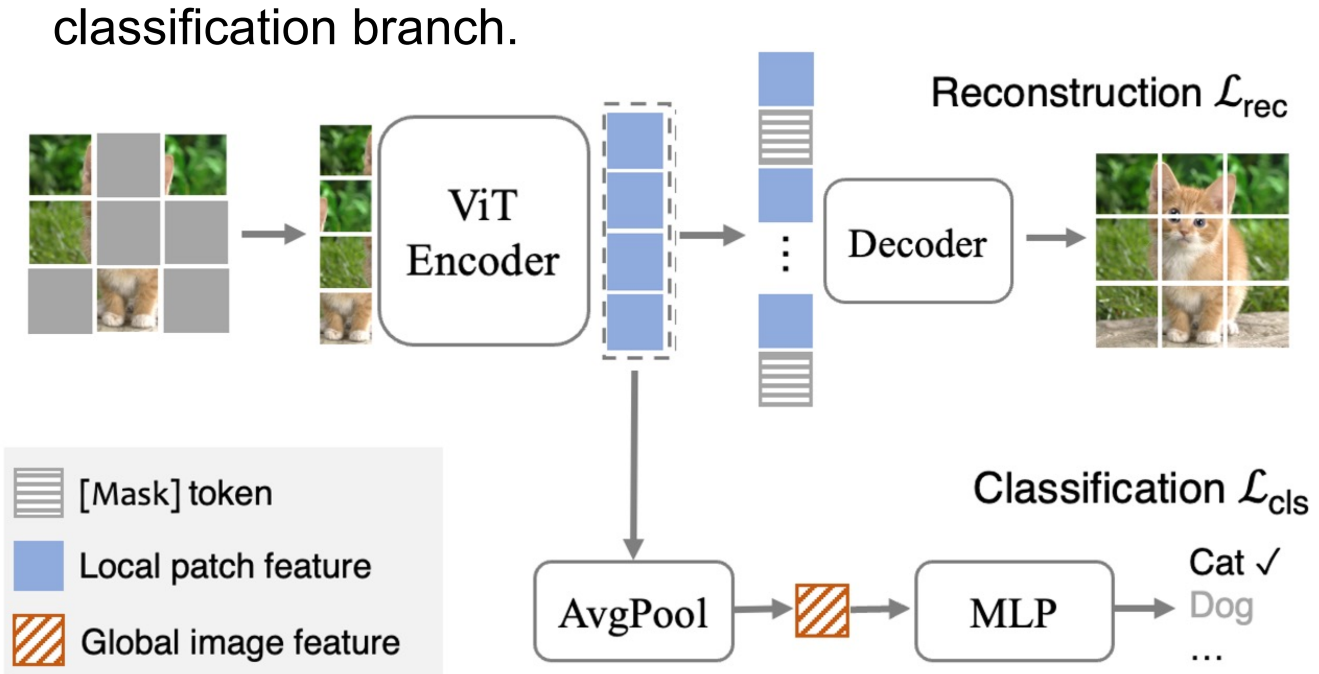


MAE [K. He *et. al.*]

	Training time*	ImageNet acc.+
+ Accuracy is after supervised training on ImageNet	394 hours	83.6
	✗	✓

SupMAE: the best of both worlds

- **SupMAE** extends MAE by adding a supervised classification branch.



	Training time*	ImageNet acc.+
Rec. loss: learn middle-level features	125.9 hours	83.6
Cls. loss: learn global features.	✓	✓

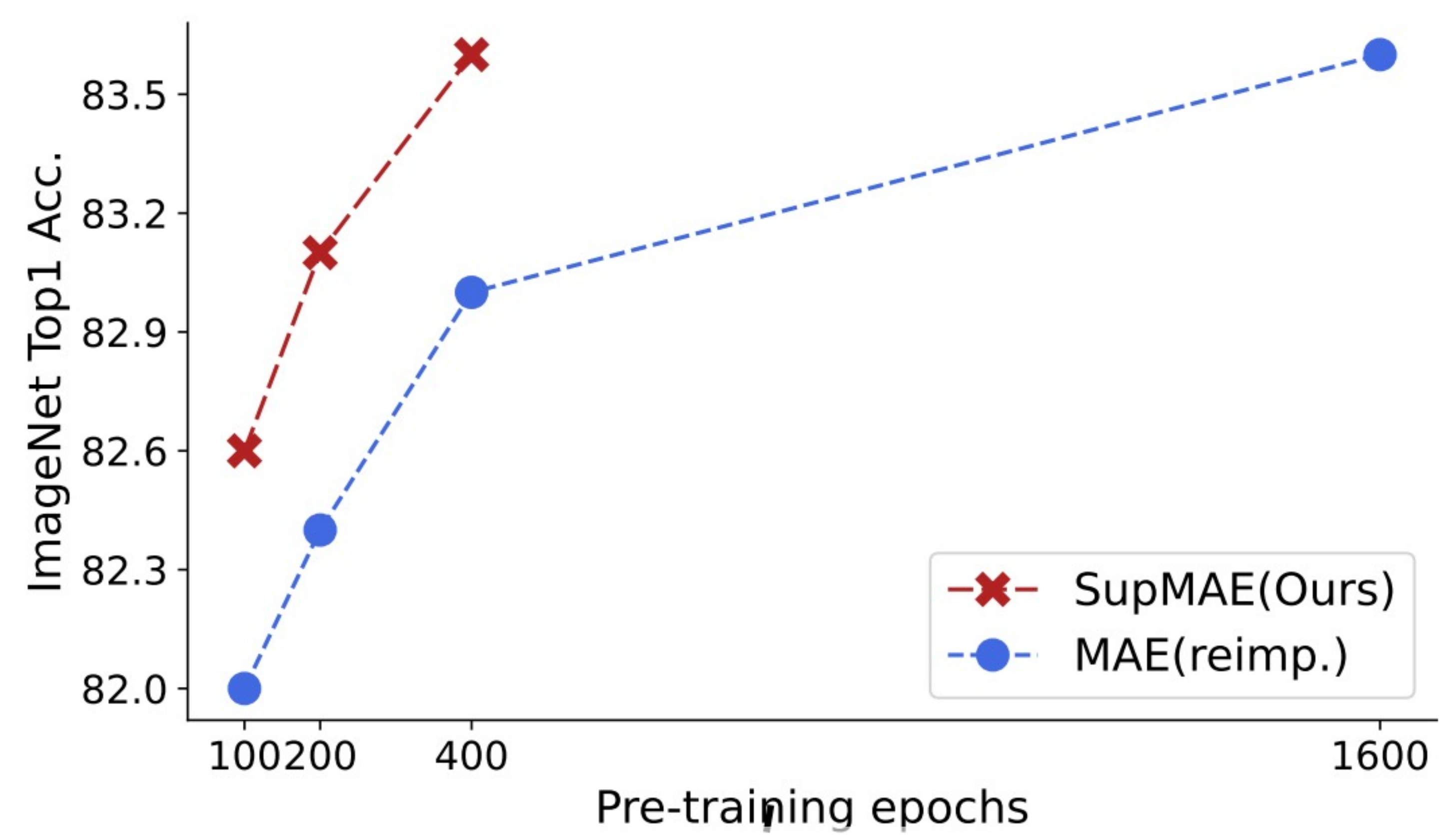
Comparison with sup. and self-sup. methods

- SupMAE shows a great efficiency and can achieve the same accuracy as MAE using only 30% compute.

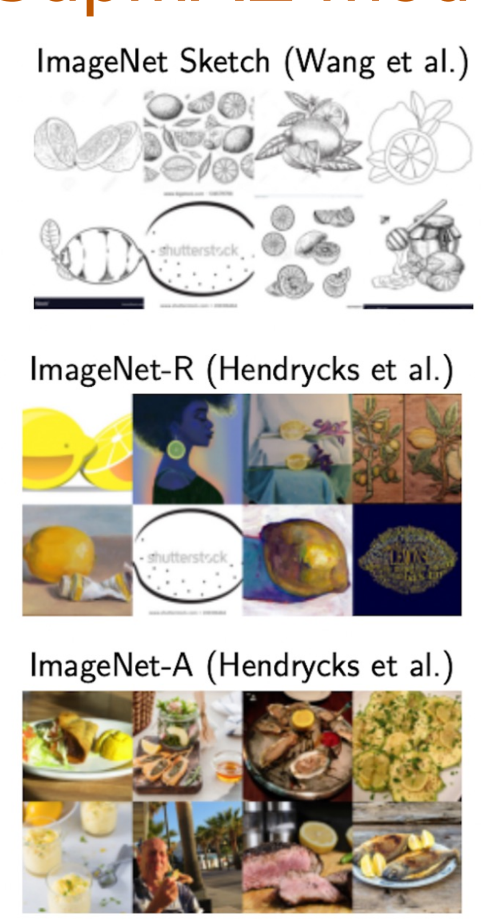
method	Total cost (Hours)	Normalized cost	Top1 acc.
MoCov3 (Chen*, Xie*, and He 2021)	295.7	2.35×	83.2
BEiT (Bao, Dong, and Wei 2021)	264.8	2.10×	83.2
MAE (He *et al.* 2021)	394	3.12×	**83.6**
ViT (Dosovitskiy *et al.* 2020)	-	-	77.9
DeiT (Touvron *et al.* 2021)	91.5	0.73×	81.8
Naive supervised (He *et al.* 2021)	90	0.71×	82.3
SupMAE(Ours)	125.9	1×	**83.6**

SupMAE is more training efficient

- SupMAE is efficient and shows a much faster convergence speed.



SupMAE model shows better robustness



- All models are trained on ImageNet and evaluated on ImageNet variants. SupMAE model shows better roboustness on the benchmark.

dataset	MAE	DeiT	SupMAE(Ours)
IN-Corruption ↓	51.7	**47.4**	48.1
IN-Adversarial	**35.9**	27.9	35.5
IN-Rendition	48.3	45.3	**51.0**
IN-Sketch	34.5	32.0	**36.0**
Score	41.8	39.5	**43.6**

SupMAE learns more transferable features

- Transferring to semantic segmentation on ADE20K

method	mIoU	aAcc	mAcc
Naive supervised	47.4	-	-
MAE	48.6	**82.8**	59.4
SupMAE (ours)	**49.0**	82.7	**60.2**

- Few-Shot transfer learning on 20 classification datasets

Pre-training Settings		20 Image Classification Datasets		
Checkpoint	Method	5-shot	20-shot	50-shot
Linear Probing				
MAE	Self-Sup.	33.37 ± 1.98	48.03 ± 2.70	58.26 ± 0.84
MoCo-v3	Self-Sup.	**50.17** ± 3.43	**61.99** ± 2.51	**69.71** ± 1.03
SupMAE(Ours)	Sup.	47.97 ± 0.44	60.86 ± 0.31	66.68 ± 0.47
Fine-tuning				
MAE	Self-Sup.	36.10 ± 3.25	54.13 ± 3.86	65.86 ± 2.42
MoCo-v3	Self-Sup.	39.30 ± 3.84	58.75 ± 5.55	70.33 ± 1.64
SupMAE(Ours)	Sup.	**46.76** ± 0.12	**64.61** ± 0.82	**71.71** ± 0.66

Generalize method to SimMIM

- Integrating the supervised branch into SimMIM. Results show that supervised branch is also compatible with other MIM frameworks.

method	SimMIM	SimMIM w/ sup.
Top1 acc.	82.8	**83.0**

Acknowledgment

- This work was supported in part by NSF CCF Grant No. 2107085 ,ONR Minerva program, iMAGiNE - the Intelligent Machine Engineering Consortium at UT Austin, and a UT Cockrell School of Engineering Doctoral Fellowship.

Code & Models