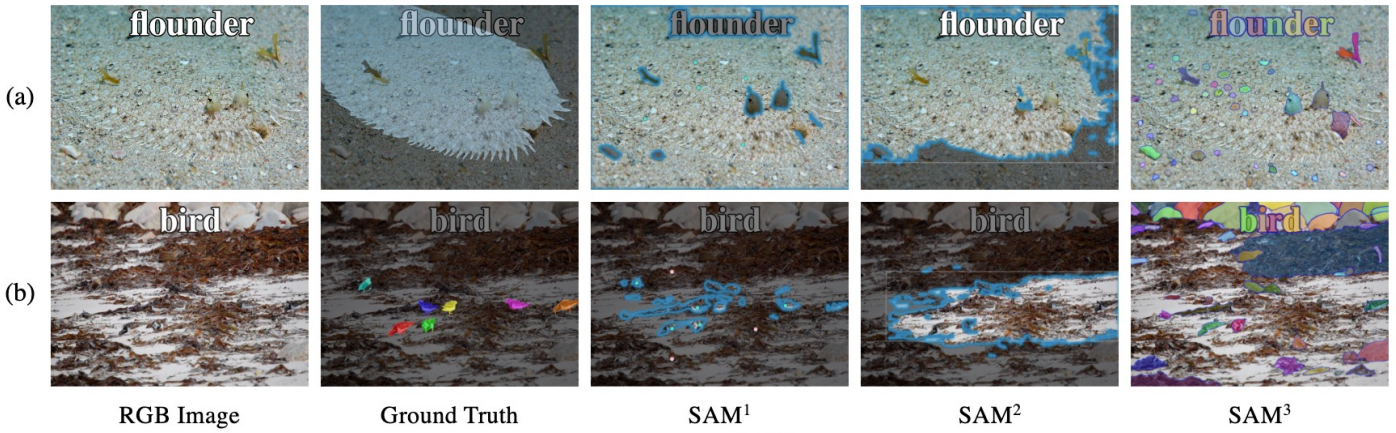
Convolution Meets LoRA: Parameter Efficient Finetuning for Segment Anything Model

Zihan Zhong[1], Zhiqiang Tang[2], Tong He[2], Haoyang Fang[2], Chun Yuan[1]

[1]Tsinghua University, [2]Amazon Web Services

Background



Segment Anything Model (SAM) may not perform well on many **real-world segmentation tasks**.

[1] Ji W, Li J, Bi Q, et al. Segment anything is not always perfect: An investigation of sam on different real-world applications[J]. 2024.

Background

Some existing works that fine-tuning SAM have failed to either analyze or address certain limitations inherent in SAM:

1) SAM's image encoder is a plain ViT, which is known to **lack of vision-specific inductive biases**.

2) SAM's low-level mask prediction pretraining **hinders its ability to capture high-level image semantic information**.

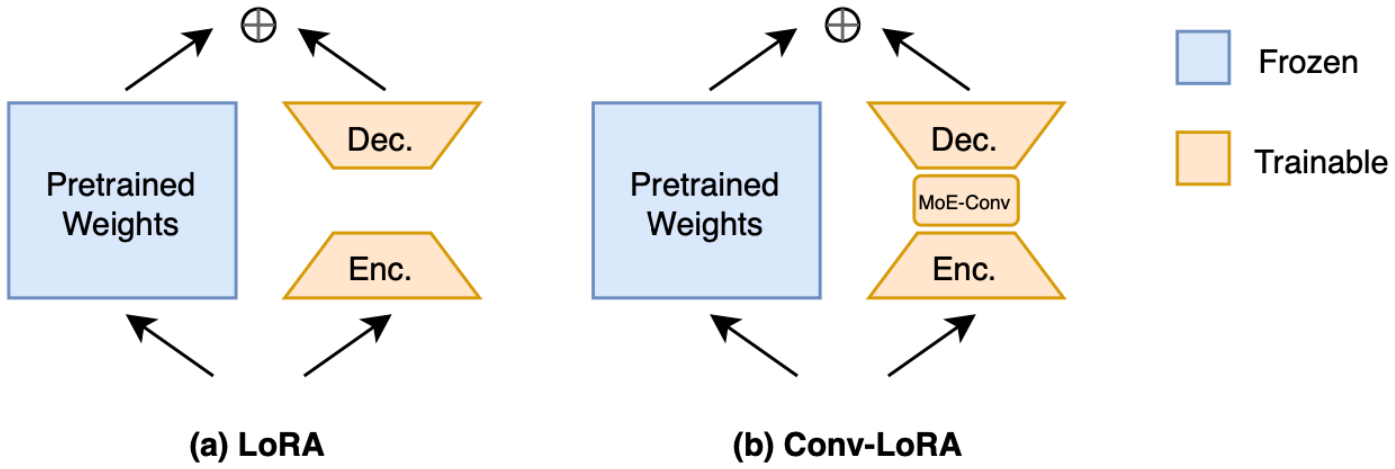
We perform linear probing for the ViT-B pretrained with Masked Auto-Encoder (MAE) and SAM on ImageNet-1K.

Method	Acc.
MAE	67.7%
SAM	54.2%

To tackle the above limitations and still retain SAM's valuable segmentation knowledge acquired during pretraining, we use **parameter efficient finetuning (PEFT).**

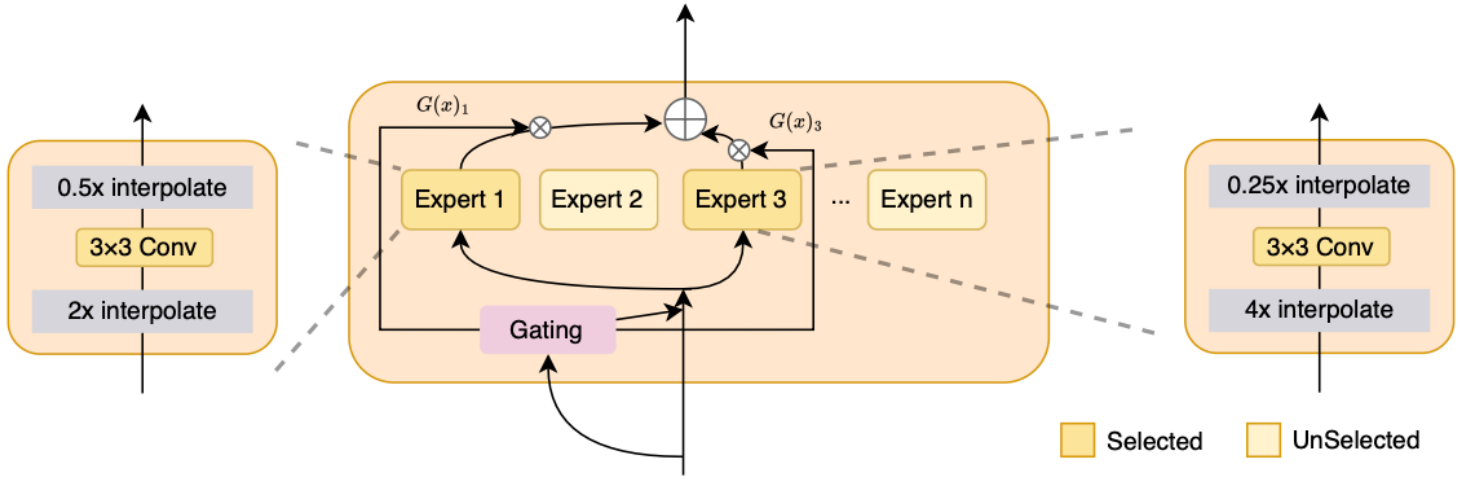
New PEFT Method Conv-LoRA

- **LoRA** surpasses widely-used visual prompt tuning (VPT), particularly in the multi-class semantic segmentation tasks.
- **Convolution operations**: inject image-related local prior.
- **Mixture-of-Experts (MoE):** dynamically select appropriate scale(s) of image features.



MoE-Conv Module

- The combination of convolution and MoE: inject the local prior into the appropriate scale(s) of image features.



Gating: dynamic expert selection.
Experts: each expert E_i specializes in one unique feature scale s_i.

$$E_i(x) = \text{Interpolate}(\text{Conv}_{3\times3}(\text{Interpolate}(x, s_i)), 1/s_i)$$

Experiments

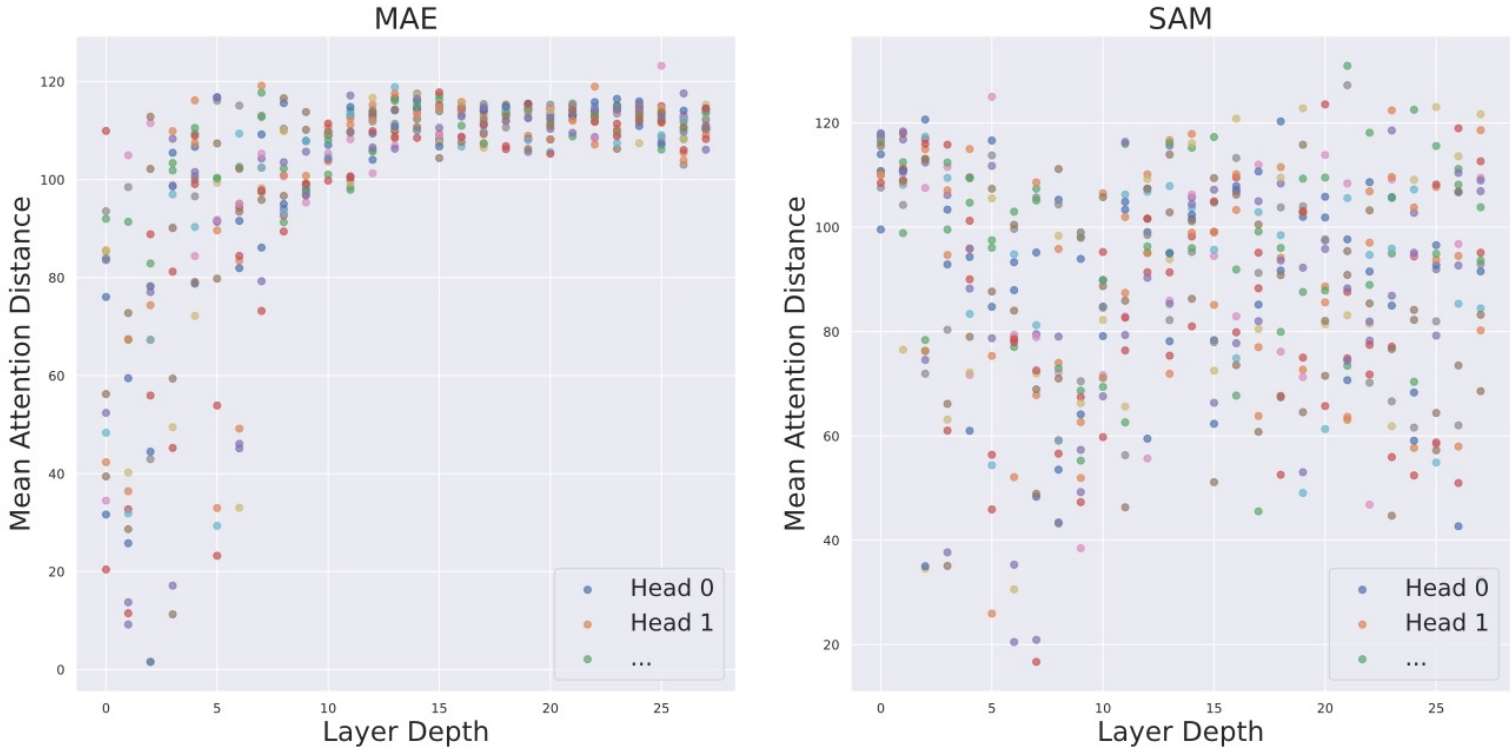
- Different Domains (Medical, Natural Images, Agriculture, Remote Sensing):
- Baselines: fine-tune SAM's decoder only, SAM trained from scratch, other PEFT methods, domain-specific models.

Method	#Params (M) / Ratio (%)	Medical						Natural Images				Agriculture		Remote Sensing	
		Kvasir		CVC-612		ISIC 2017		CAMO			SBU	Leaf		Road	
		$S_\alpha \uparrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	Jac ↑	Dice ↑	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	BER ↓	IoU ↑	Dice ↑	IoU ↑	Dice ↑
Domain Specific	* / 100%	90.9	94.4	92.6	95.5	80.1	87.5	80.8	85.8	73.1	3.56	62.3	74.1	59.1	73.0
SAM trained from scratch	641.09 / 100%	78.5	82.4	85.9	91.6	73.8	82.5	61.9	67.0	40.5	5.53	52.1	65.5	55.6	71.1
decoder-only	3.51 / 0.55%	86.5	89.5	85.5	89.9	69.7	79.5	78.5	83.1	69.8	14.58	50.8	63.8	48.6	65.1
BitFit	3.96 / 0.62%	90.8 ± 0.57	93.8 ± 0.98	89.0 ± 0.40	91.6 ± 0.98	76.4 ± 0.45	84.7 ± 0.35	86.8 ± 0.33	90.7 ± 0.28	81.5 ± 0.19	3.16 ± 0.128	71.4 ± 1.15	81.7 ± 1.01	60.6 ± 0.15	75.2 ± 0.11
Adapter	3.92 / 0.61%	91.2 ± 0.23	94.0 ± 0.16	89.3 ± 0.43	92.0 ± 0.63	76.7 ± 0.66	85.0 ± 0.56	87.7 ± 0.10	91.3 ± 0.40	82.8 ± 0.35	2.84 ± 0.093	72.1 ± 0.47	82.4 ± 0.36	61.5 ± 0.11	75.9 ± 0.12
VPT	4.00 / 0.62%	91.5 ± 0.23	94.3 ± 0.06	91.0 ± 0.94	93.7 ± 1.41	76.9 ± 0.94	85.1 ± 0.75	87.4 ± 0.60	91.4 ± 0.68	82.1 ± 0.75	2.70 ± 0.055	73.6 ± 0.26	83.8 ± 0.26	60.2 ± 1.87	74.9 ± 1.50
LST	11.49 / 1.77%	89.7 ± 0.25	93.3 ± 0.37	89.4 ± 0.37	92.4 ± 0.54	76.4 ± 1.05	84.9 ± 0.79	83.3 ± 0.28	88.0 ± 0.23	77.1 ± 0.02	3.18 ± 0.012	70.2 ± 0.87	81.1 ± 0.82	60.2 ± 0.26	74.9 ± 0.22
SAM-Adapter	3.98 / 0.62%	89.6 ± 0.24	92.5 ± 0.10	89.6 ± 0.22	92.4 ± 1.06	76.1 ± 0.45	84.6 ± 0.37	85.6 ± 0.26	89.6 ± 0.55	79.8 ± 0.89	3.14 ± 0.063	71.4 ± 0.20	82.1 ± 0.10	60.6 ± 0.06	75.2 ± 0.04
SSF	4.42 / 0.69%	91.3 ± 0.87	93.9 ± 1.49	89.6 ± 0.37	91.9 ± 0.79	76.6 ± 0.19	85.0 ± 0.14	87.5 ± 0.11	91.4 ± 0.16	82.6 ± 0.12	3.19 ± 0.046	71.5 ± 0.63	81.8 ± 0.44	61.6 ± 0.03	76.0 ± 0.02
LoRA	4.00 / 0.62%	91.2 ± 0.28	93.8 ± 0.22	90.7 ± 0.04	92.5 ± 0.41	76.6 ± 0.23	84.9 ± 0.22	88.0 ± 0.24	91.9 ± 0.42	82.8 ± 0.16	2.74 ± 0.079	73.7 ± 0.20	83.6 ± 0.13	62.2 ± 0.21	76.5 ± 0.18
Conv-LoRA	4.02 / 0.63%	**92.0** ± 0.15	**94.7** ± 0.16	**91.3** ± 0.69	**94.0** ± 0.78	**77.6** ± 0.57	**85.7** ± 0.36	**88.3** ± 0.40	**92.4** ± 0.31	**84.0** ± 0.34	**2.54** ± 0.081	**74.5** ± 0.39	**84.3** ± 0.34	**62.6** ± 0.36	**76.8** ± 0.27

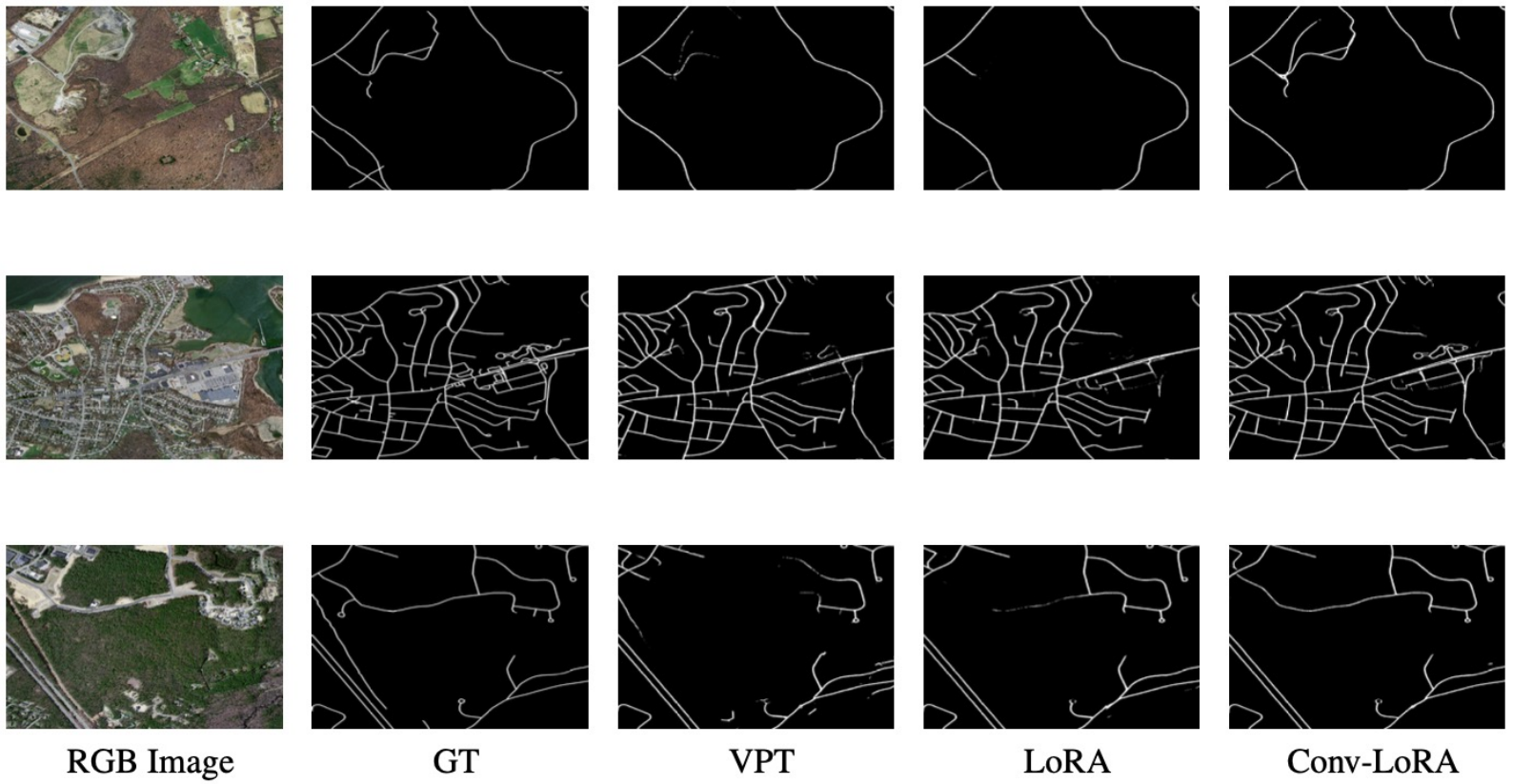
Method	# Params (M) / Ratio (%)	Acc↑	mIoU ↑	Category IoU ↑											
				bg	shelf	jar	freezer	window	door	eyeglass	cup	wall	bowl	bottle	box
TransLab	42.19 / 100%	92.67	69.00	93.90	54.36	64.48	65.14	54.58	57.72	79.85	81.61	72.82	69.63	77.50	56.43
Trans2Seg	56.20 / 100%	94.14	72.15	95.35	53.43	67.82	64.20	59.64	60.56	88.52	86.67	75.99	73.98	82.43	57.17
decoder-only	3.51 / 0.55%	90.66	49.97	93.66	32.75	39.96	35.87	50.70	45.89	57.38	73.16	69.36	54.23	56.58	33.77
VPT	4.00 / 0.62%	94.42	62.81	97.41	29.76	52.82	62.09	55.54	63.61	81.12	83.40	79.61	65.29	72.92	44.77
LoRA	4.00 / 0.62%	94.80	66.01	97.50	42.17	57.82	**64.35**	53.44	64.08	**87.28**	**85.28**	80.43	63.67	77.97	49.56
Conv-LoRA	4.02 / 0.63%	**95.07**	**67.09**	**97.66**	**50.51**	**58.44**	51.70	**55.69**	**65.22**	85.23	84.84	**80.97**	**72.84**	**79.83**	**52.73**

Experiments

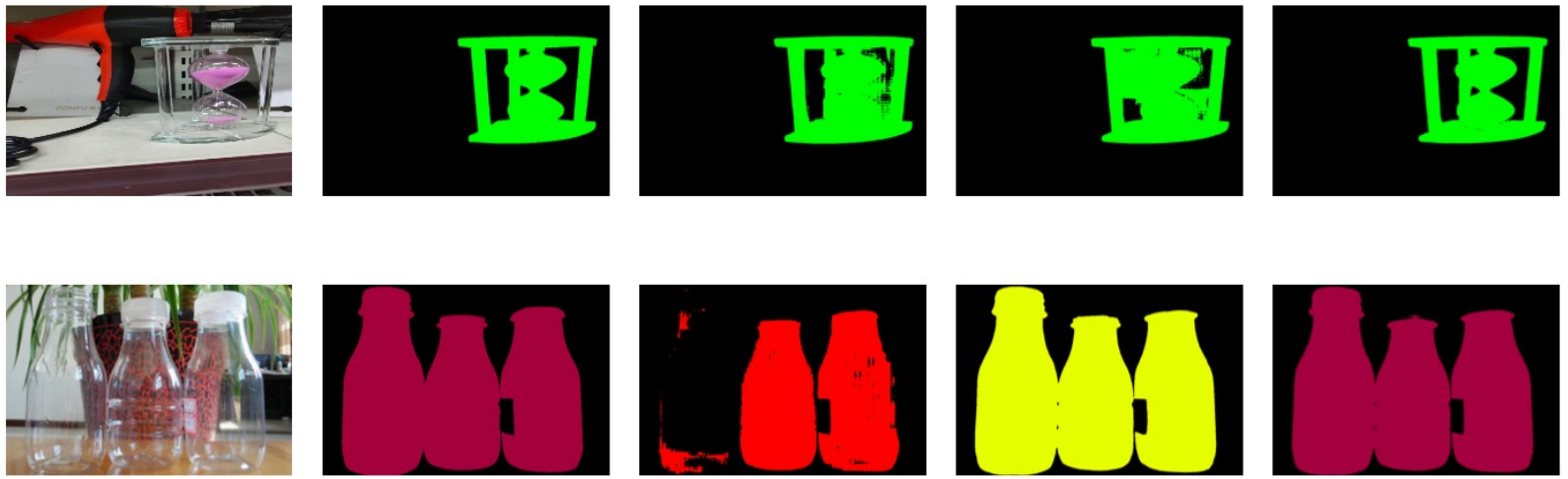
- SAM's local prior assumption: Through supervised segmentation pre-training on a vast dataset, SAM has honed a robust capability to discern and capture local features within images.



Visualization



Visualization



Thank you!