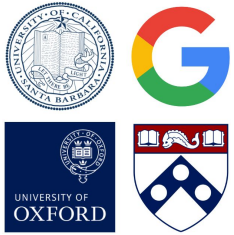


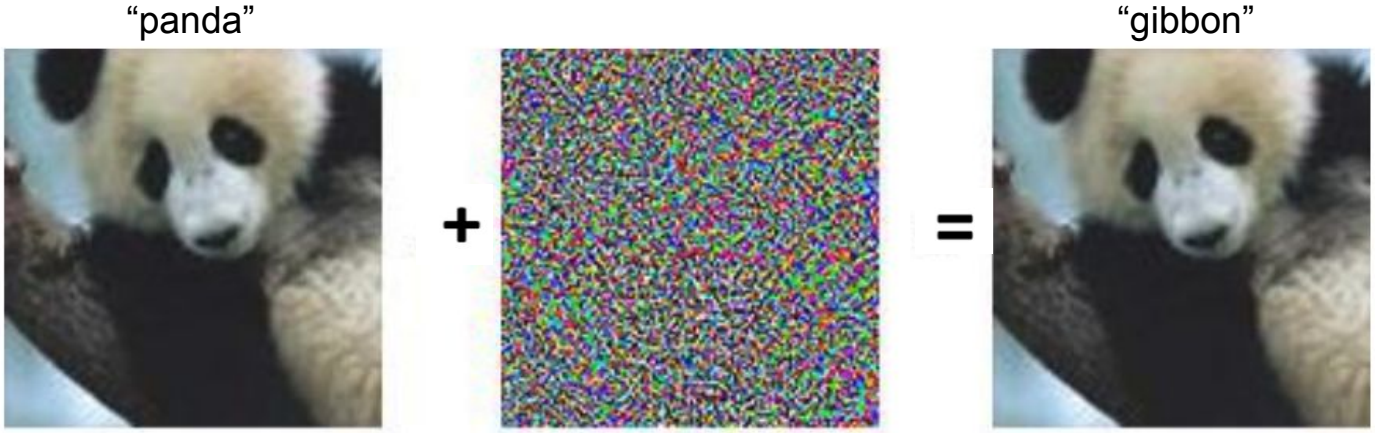
Initialization Matters for Adversarial Transfer Learning

Andong Hua[1] Jindong Gu[2] Zhiyu Xue[1] Nicholas Carlini[3] Eric Wong[4] Yao Qin[1,3]

[1]University of California, Santa Barbara [2]University of Oxford
[3]Google [4]University of Pennsylvania

Adversarial Robustness

Adversarial Attack



Clean Image
92.7% Accuracy

Perturbation

Attacked Image
0.8% Accuracy

Adversarial Training

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D} \left[\max_{\|\delta\|_\infty \leq \varepsilon} \mathcal{L}(x+\delta, y; \theta) \right]$$

Parameter-Efficient Finetuning (PEFT)

ImageNet-1k, ImageNet-22k, CLIP…

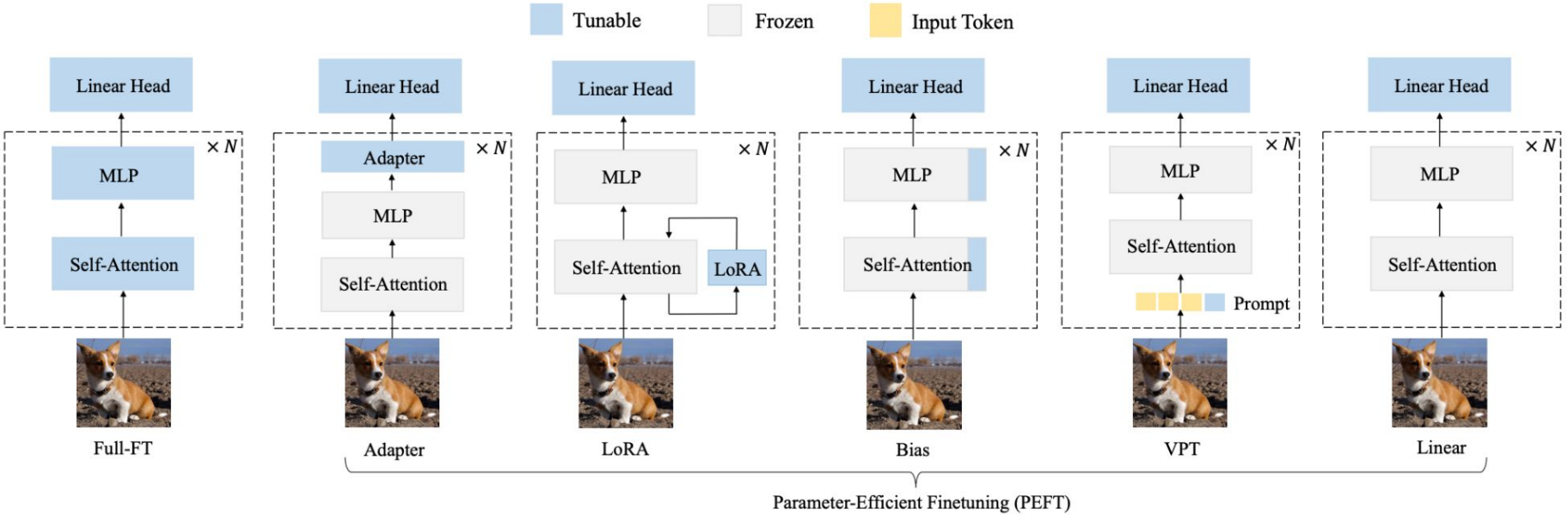
Pretrained Model

Natural Transfer Learning

Full-FT / PEFT

Finetuned Model

CIFAR-10, CIFAR-100, Caltech-256…



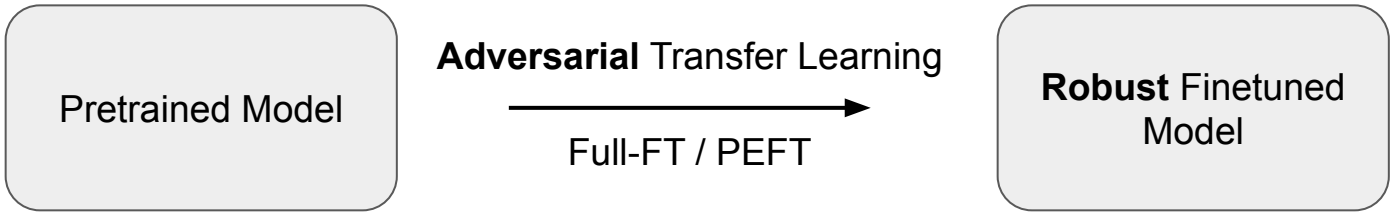
Adversarial Transfer Learning

Adversarial Finetuning:

$$\min_{\hat{\theta}} \mathbb{E}_{(x,y)\sim D}\left[\max_{\|\delta\|_\infty \leq \varepsilon} \mathcal{L}(x+\delta, y; \theta \cup \hat{\theta})\right]$$

$\hat{\theta}$: **Tunable** Parameters (e.g. Linear head)

$\hat{\theta}$: **Frozen** Parameters

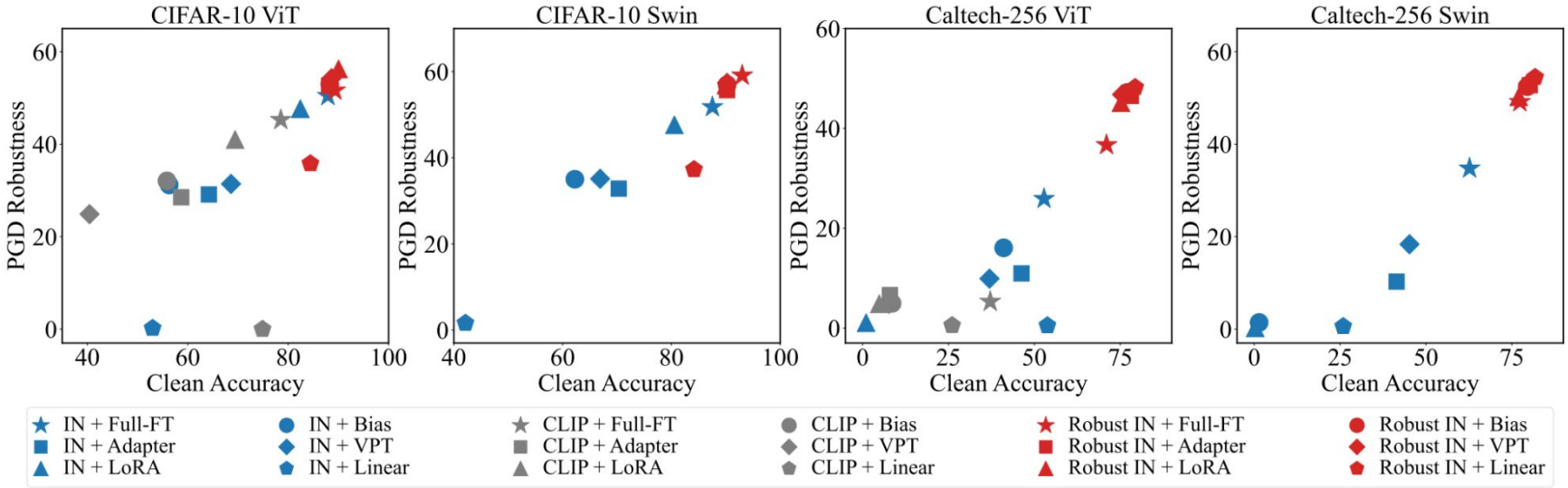


What matters for adversarial transfer learning?

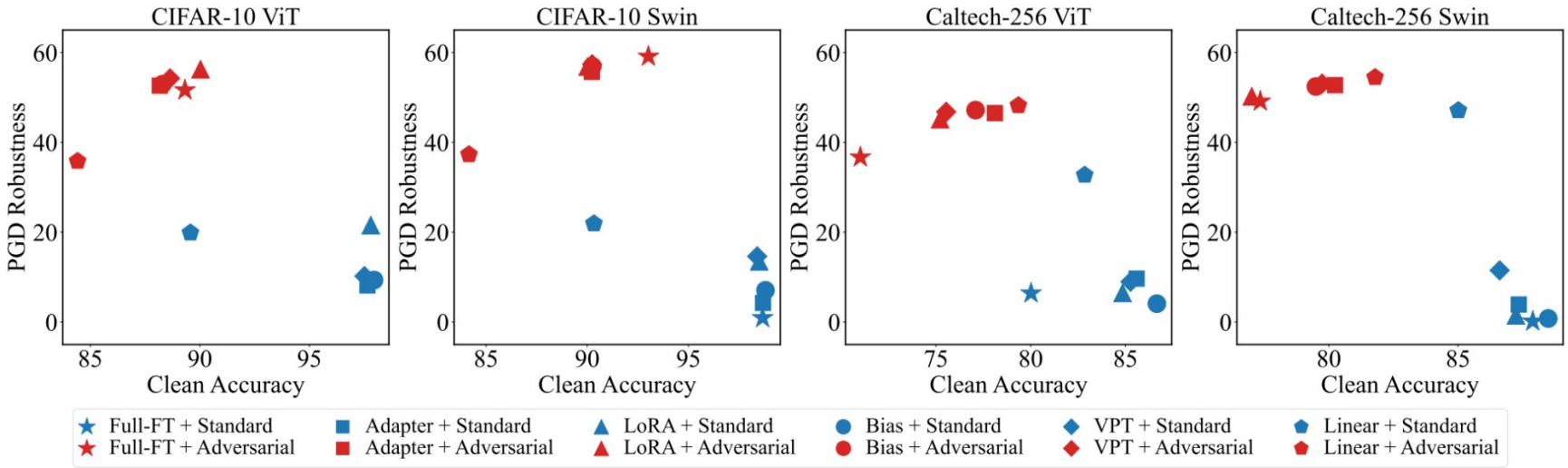
1. A robust pretrained model is necessary.
2. Robust linear initialization (RoLI) for finetuning.

A Robust Pretrained Model is Necessary

- Pretraining on a larger dataset does not help.
- Fully finetuning (Full-FT) consistently outperforms other methods.



With a Robust Pretrain,



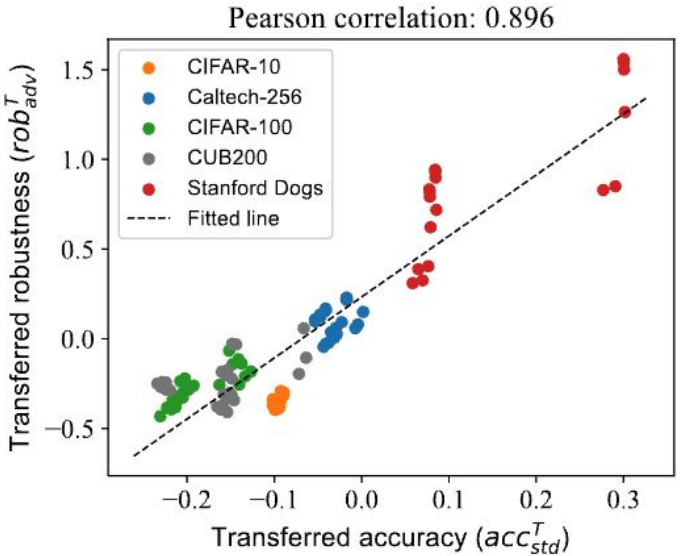
- Adversarial PEFTs except Linear demonstrate strong performance.
- Adversarial Linear achieves the strongest robustness on Caltech256

  Supring!

Why and When Linear Work Best?

- Source1: Robustness inherited from pretraining.
- Source2: Robustness achieved by adversarial finetuning.

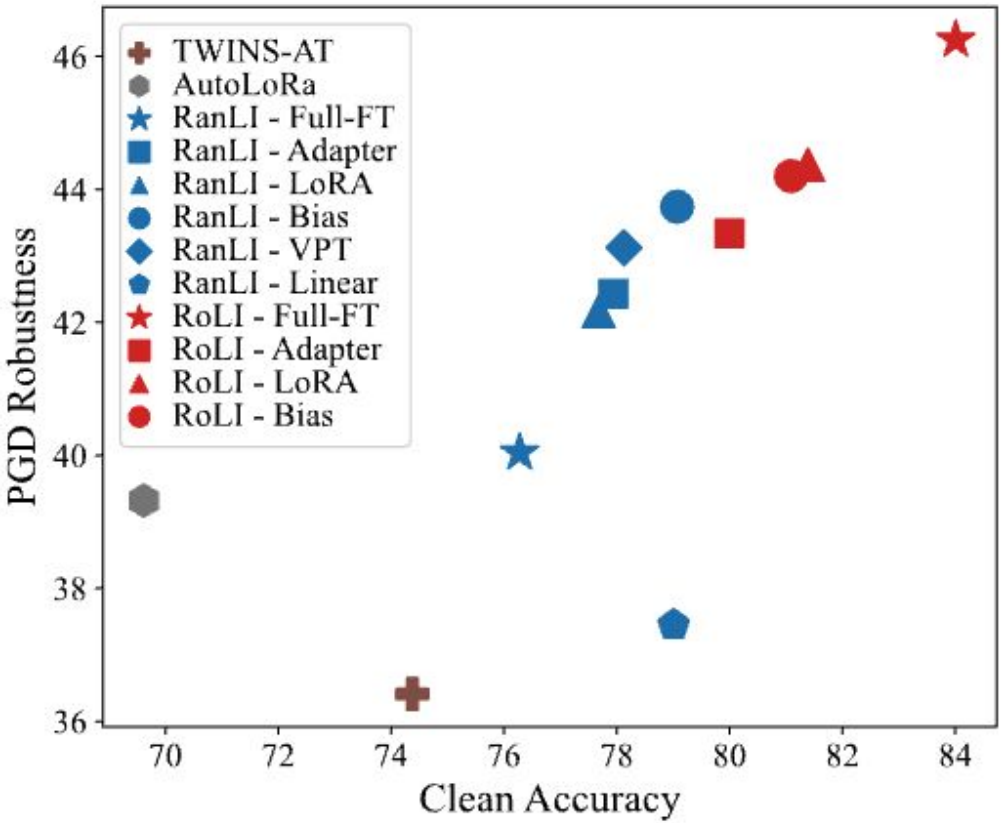
Linear excels in preserving robustness from pretrained models!



- Transferred accuracy/robustness is the normalized performance gap between linear probing and fully finetuning.

- **Transferred robustness correlates with transferred accuracy.**

Robust Linear Initialization (RoLI)



- **RoLI**: Initialize the linear head of a robust pretrained model with weights obtained through adversarial linear probing.
- **RanLI**: Random initialization.

- **3.88%** increase in clean accuracy and **2.44%** in robustness over random initialization across five datasets.
- RoLI - Full-FT achieves **SOTA**.

Conclusion and Q & A

What matters for adversarial transfer learning?

1. A robust pretrained model is necessary.
2. Robust linear initialization (RoLI) for finetuning.

Thanks!