Vision Transformer Finetuning Benefits from Non-Smooth Components

Ambroise Odonnat
Noah's Ark Lab, Inria

Mila, CERC-AAI Lab

March 27, 2026

Co-authors

Laetitia Chapel
L'Institut Agro, IRISA

Romain Tavenard
Univ. Rennes 2, IRISA

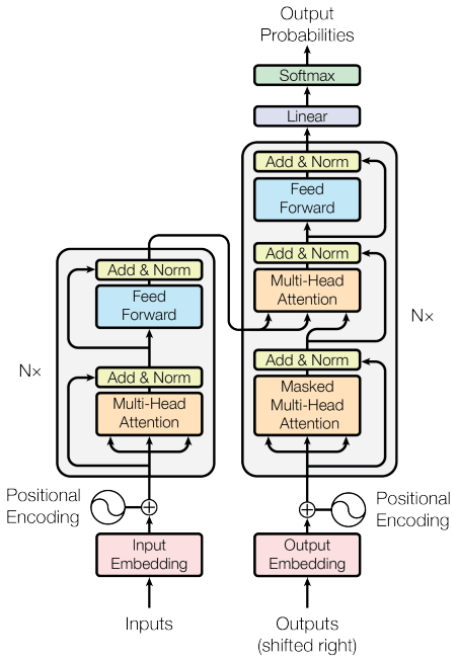
Ievgen Redko
Noah's Ark Lab

Outline

Outline

Foundation models

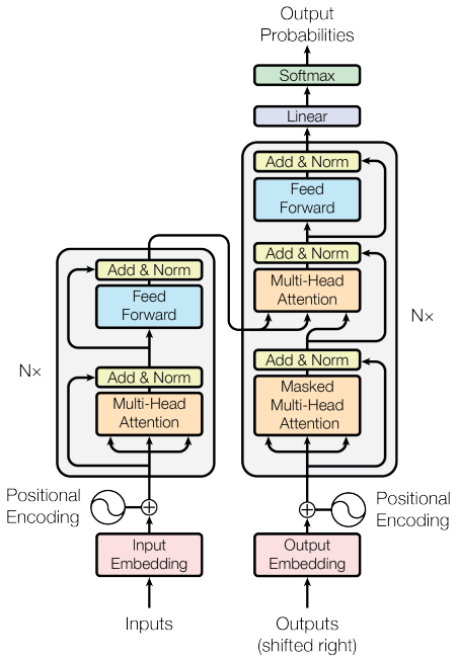
Most foundation models are based on the transformer architecture.



(Vaswani et al., 2017)

Foundation models

Most foundation models are based on the transformer architecture.

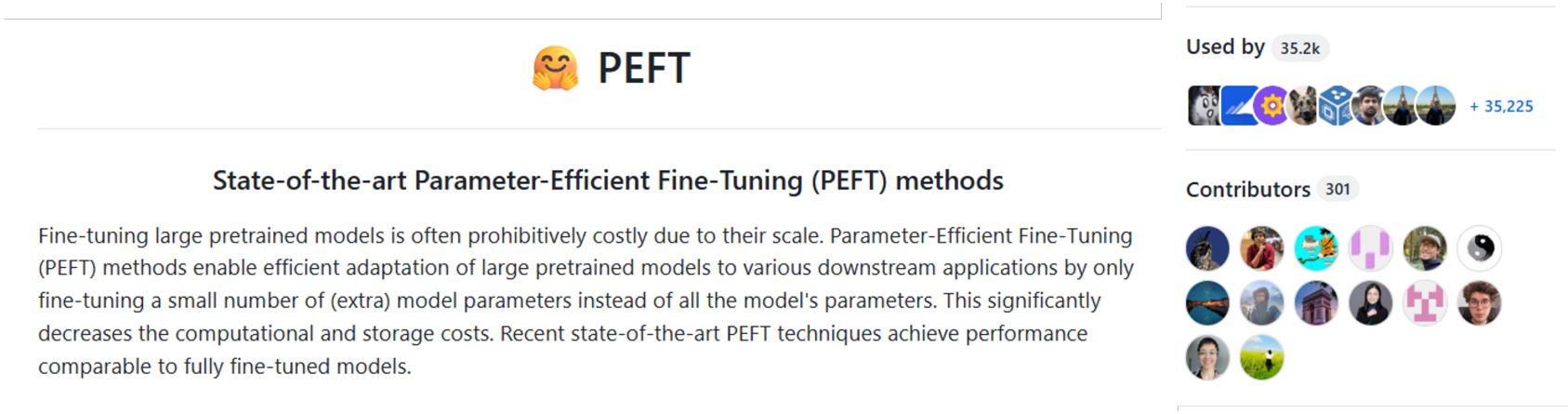


(Vaswani et al., 2017)

- Diverse training data
- Large scale pretraining
- (Very) large models

Finetuning foundation models

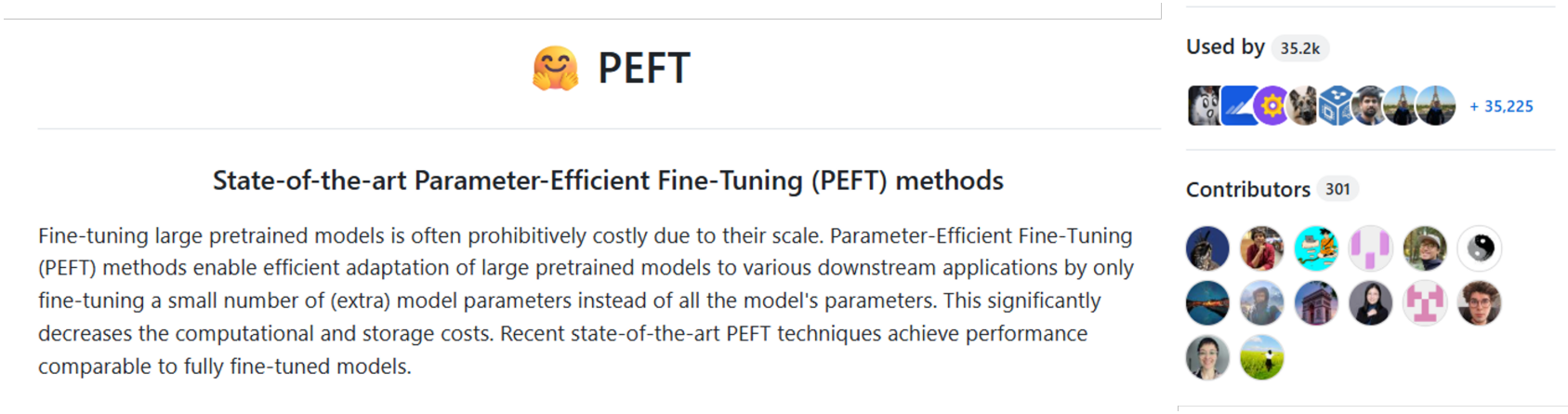
- ✗ Need of specialized models in practice
- ✗ Increasing cost of adaptation with models growing larger



Plethora of parameter-efficient finetuning methods

Finetuning foundation models

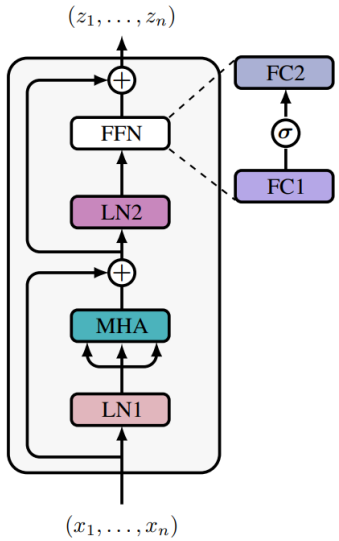
- ✗ Need of specialized models in practice
- ✗ Increasing cost of adaptation with models growing larger

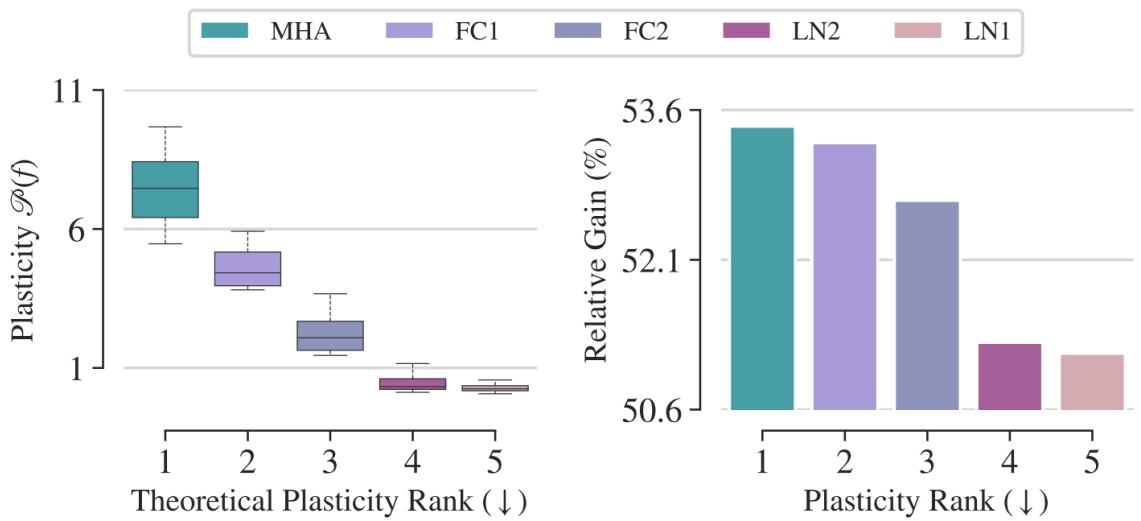


Which transformer components should be prioritized during finetuning and why?

Our contributions

Finetuning high-plasticity components (non-smooth) yields larger performance gains than the stiff ones.

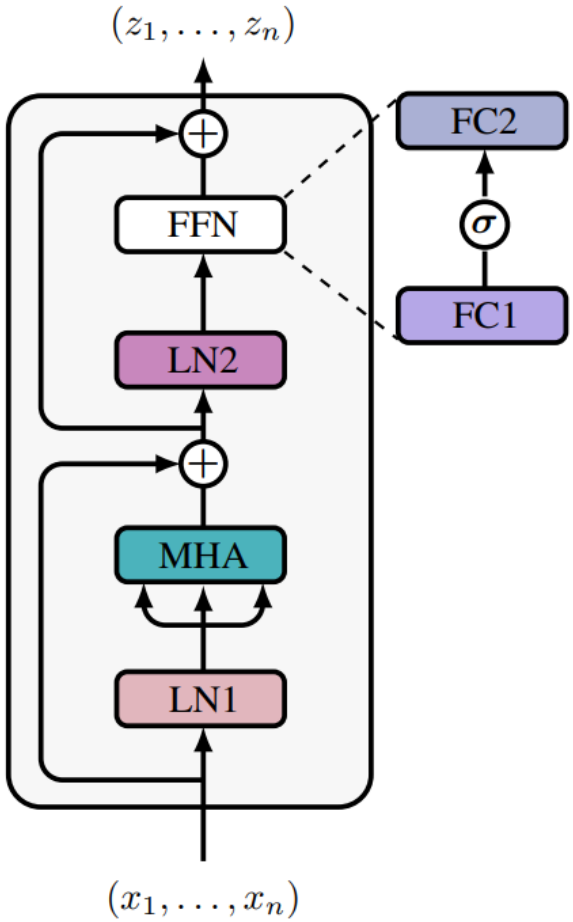




Outline

Vision transformer

- Inputs are 2D images
- Images split into patches
- Tokens are flattened patches
- Encoder-only architecture



Smoothness intuition

Smoothness is connected to differentiability, robustness to input perturbations and Lipschitz continuity.

Definition (Lipschitz continuity)

The smallest constant K such that

$$\forall x, y \in (\mathbb{R}^d)^n, \|f(x) - f(y)\|_{\mathrm{F}} \leq K\|x - y\|_{\mathrm{F}}$$

is called the Lipschitz constant of f and writes

$$\mathrm{Lip}(f) = \sup_{x \neq y \in (\mathbb{R}^d)^n} \frac{\|f(x) - f(y)\|_{\mathrm{F}}}{\|x - y\|_{\mathrm{F}}}.$$

Smoothness intuition

Promoting smoothness by regularizing the Lipschitz constant reduces the sensitivity to input perturbations.

✓ Good for training stability and adversarial robustness,

× Too much smoothness constrain adaptability to new tasks,

→ Identify components with too small Lipschitz constants.

Plasticity measure

Definition (Plasticity)

Let ν be the uniform distribution over the set of distinct pairs of sequences of tokens in $(\mathbb{R}^d)^n$. We define the plasticity of a transformer component f as

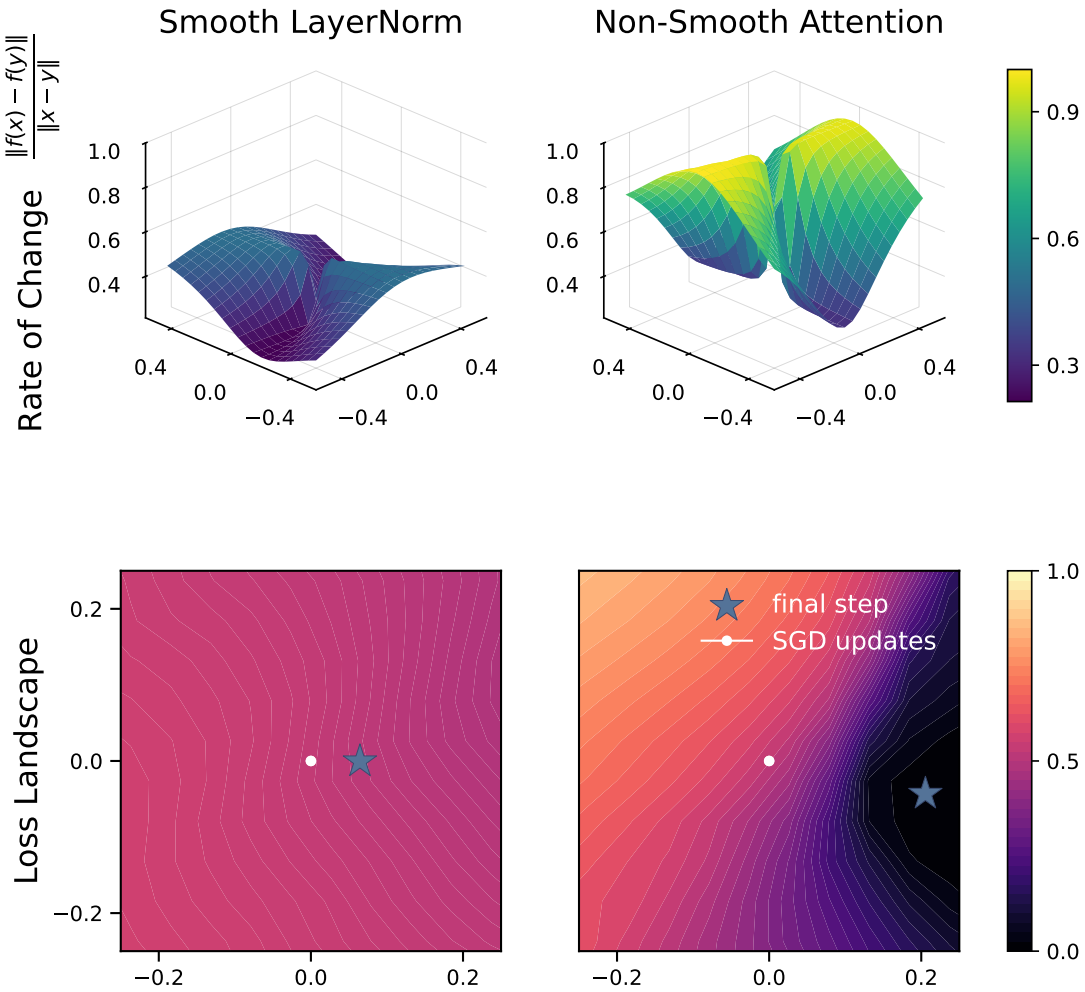
$$\mathscr{P}(f) = \mathbb{E}_{(x,y)\sim\nu}\left[\frac{\|f(x) - f(y)\|_{\mathrm{F}}}{\|x - y\|_{\mathrm{F}}}\right].$$

- ✓ For any component f, we have $\mathscr{P}(f) \leq \mathrm{Lip}(f)$,
- ✓ If $\mathscr{P}(f) < 1$, the module f contracts the changes in inputs,
- ✓ If $\mathscr{P}(f) > 1$, f amplifies the input perturbations
- ✓ When $\mathscr{P}(f) > 1$, $\mathrm{Lip}(f)$ is pushed from below.

Connection to finetuning

- ✓ High plasticity $\rightarrow$ high sensitivity to input changes,
- ✓ High plasticity $\rightarrow$ high Lipschitz constant $\rightarrow$ low smoothness,
- ✓ High plasticity $\rightarrow$ faster and better adaptation.

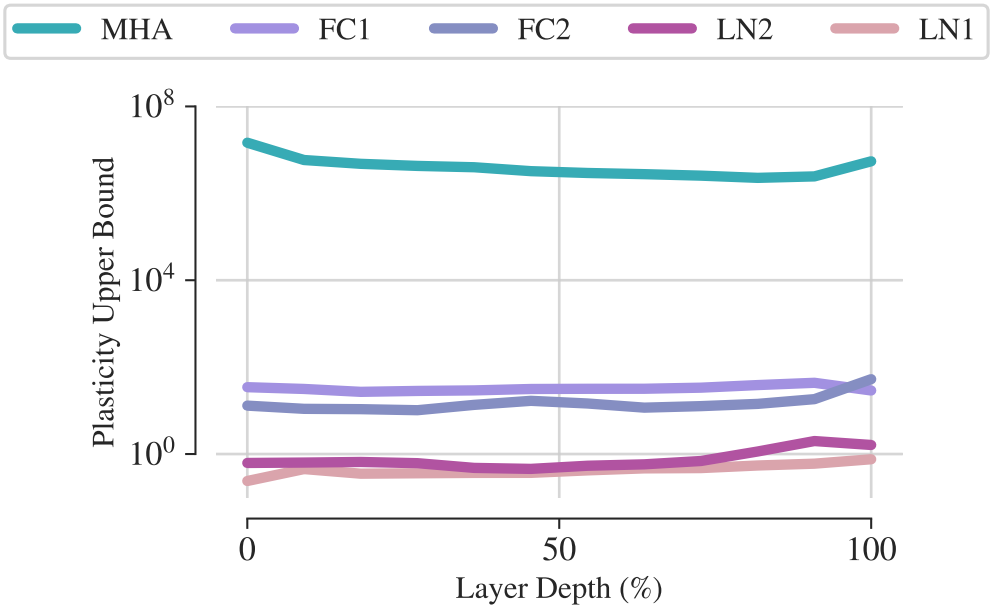
Connection to finetuning



Outline

Theoretical insights

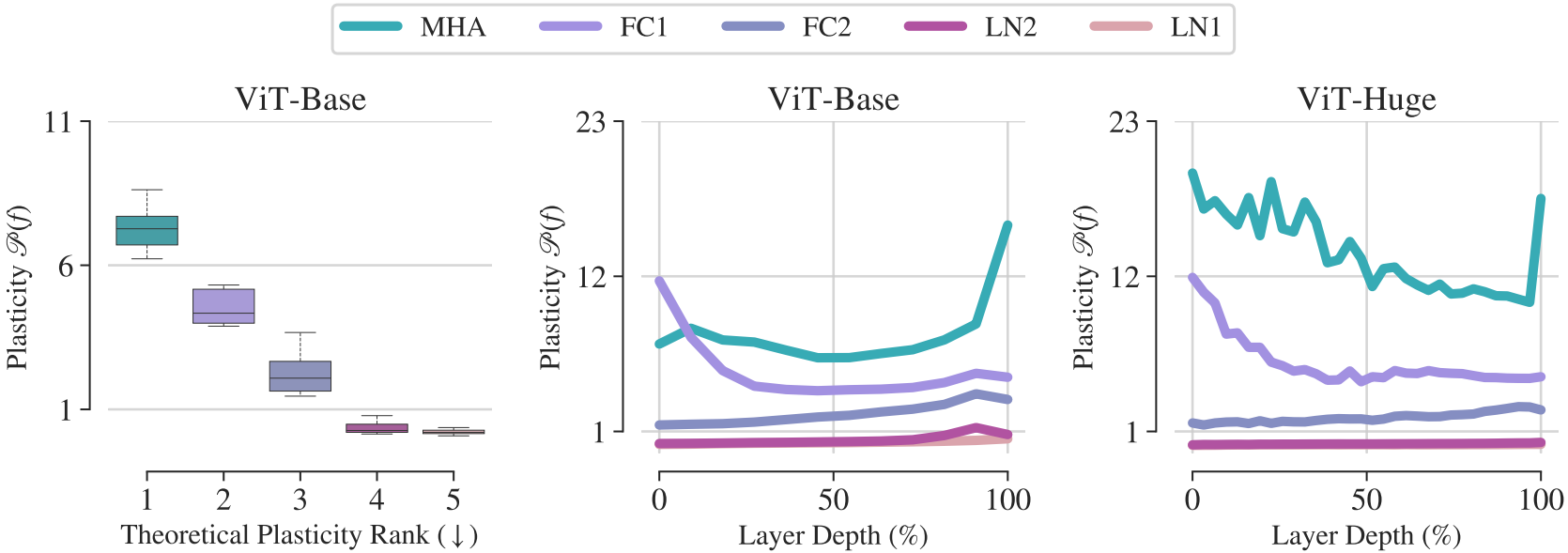
We derive upper bounds on the plasticity $\mathscr{P}(f)$.



Takeaway 1. Our theoretical analysis suggests the following plasticity ranking: MHA $\rightarrow$ FC1 $\approx$ FC2 $\rightarrow$ LN2 $\approx$ LN1.

Empirical ranking

- ViT pretrained on ImageNet (86M and 632M of parameters)
- Cifar10/100, Cifar10-C, DomainNet, Pet, Flowers102



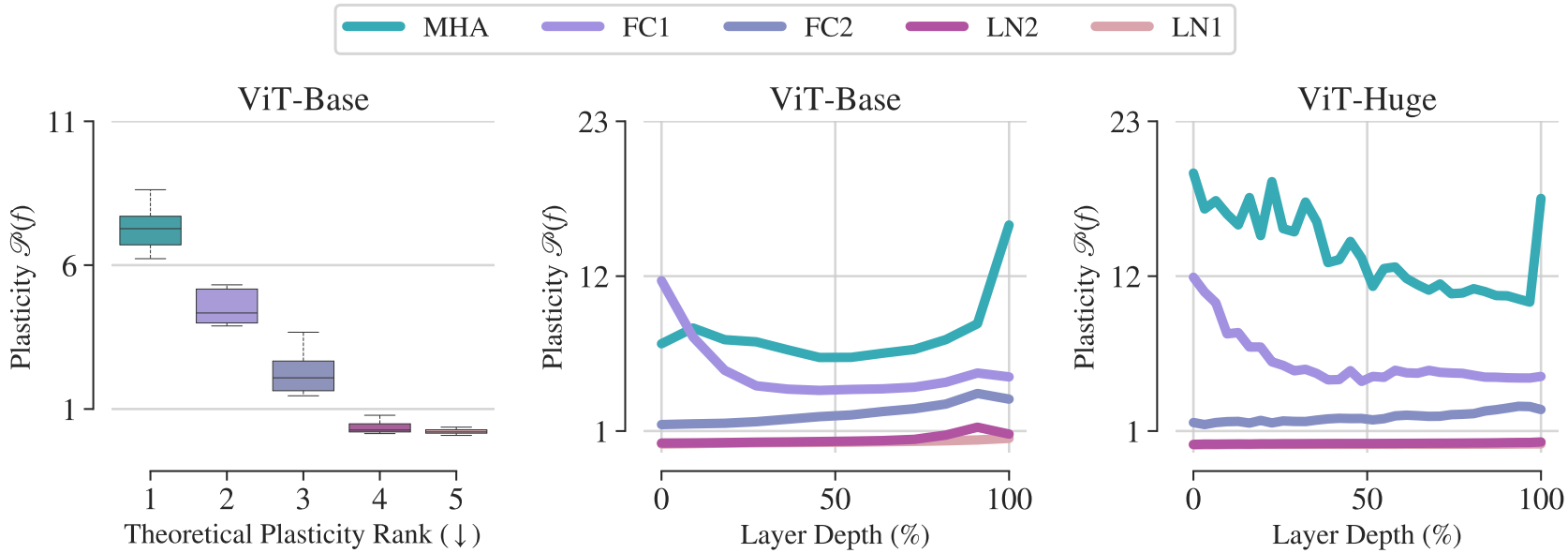
✓ Our theory correctly predicts the empirical ranking,

✓ Attention and feedforward linear layers have $\mathscr{P}(f) > 1$,

✓ LayerNorms have $\mathscr{P}(f) < 1$.

Empirical ranking

- ViT pretrained on ImageNet (86M and 632M of parameters)
- Cifar10/100, Cifar10-C, DomainNet, Pet, Flowers102



Takeaway 2. The empirical *plasticity ranking* of vision transformer modules supports our theoretical insights with:
MHA $\rightarrow$ FC1 $\rightarrow$ FC2 $\rightarrow$ LN2 $\rightarrow$ LN1.

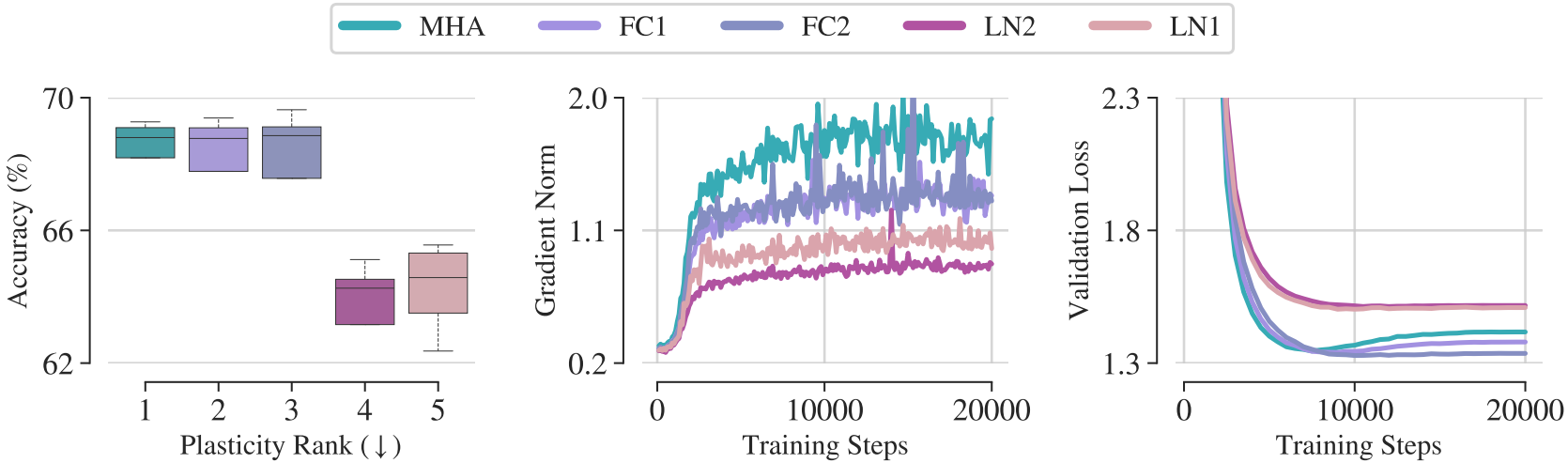
Outline

Better performance

Table 1. **MHA leads to better performance (11 benchmarks).** We report the decrease in performance between MHA and the other configurations, averaged over the benchmarks. Entries in **bold** indicate that the decrease is statistically significant according to a Wilcoxon signed-rank test at a confidence level of 5%.

configuration	FC1	FC2	LN1	LN2
decrease (%)	0.07	**0.44**	**0.8**	**0.9**

Robust adaptation



Takeaway 3.

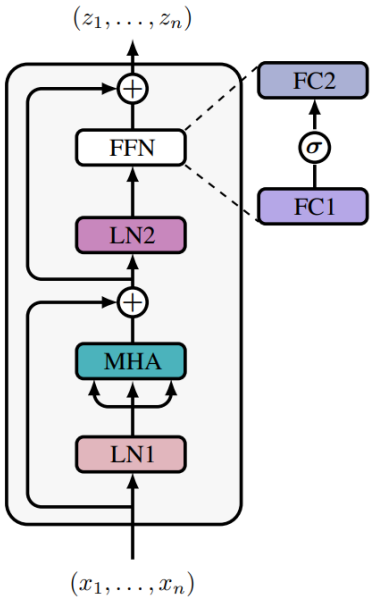
✓ High plasticity → better and more stable finetuning,

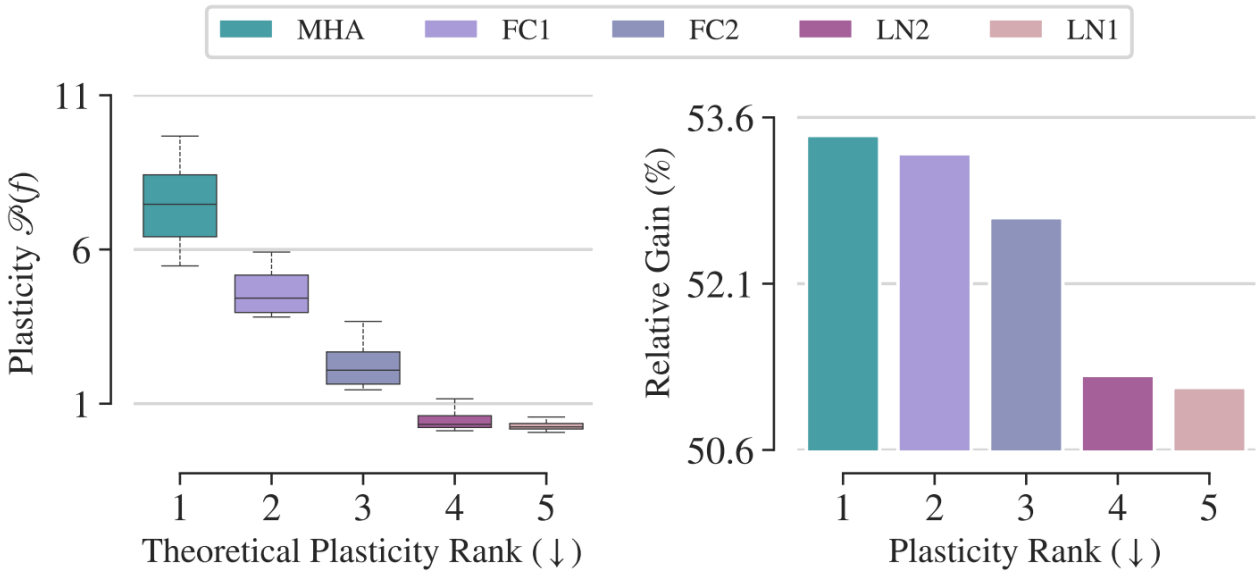
✓ Prioritize attention module and feedforward linear layers.

Outline

Take Home Message

Finetuning high-plasticity components (non-smooth) yields larger performance gains than the stiff ones.





Future work

- ✓ Study the effect of a tailored optimization,
- ✓ Extension to large language models,
- ✓ Apply LoRa only to components with high-plasticity.

To Know More

Check out the paper along with the code to know more!

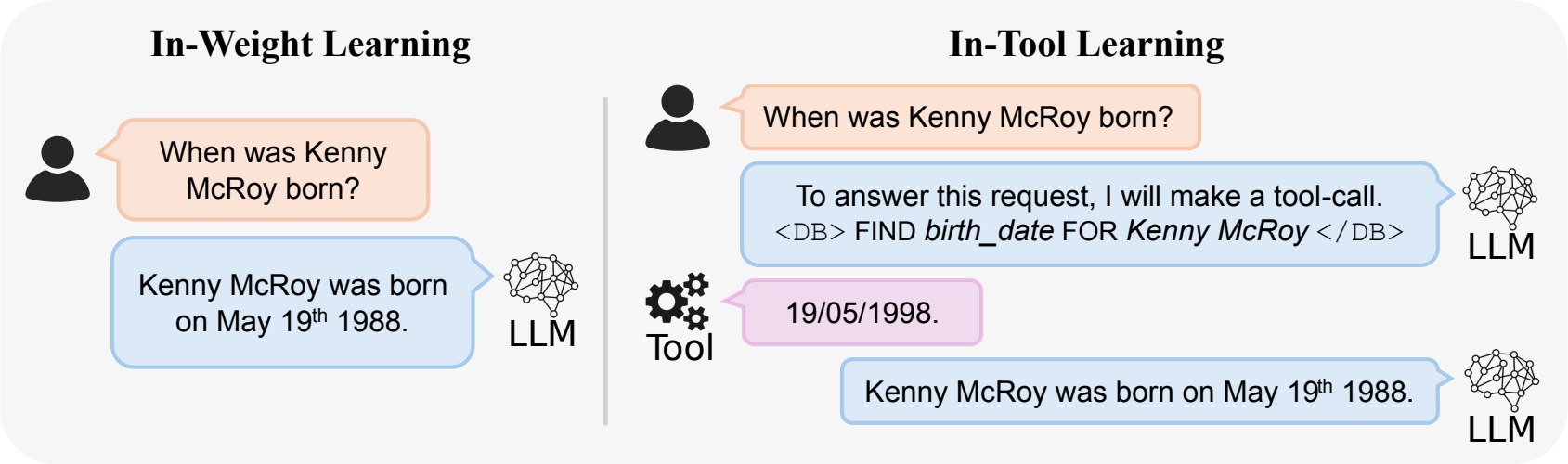
paper

code

website

Provable benefits of in-tool learning for LLMs

ICLR 2026 MemAgents Workshop



paper - code

Thanks for your attention!