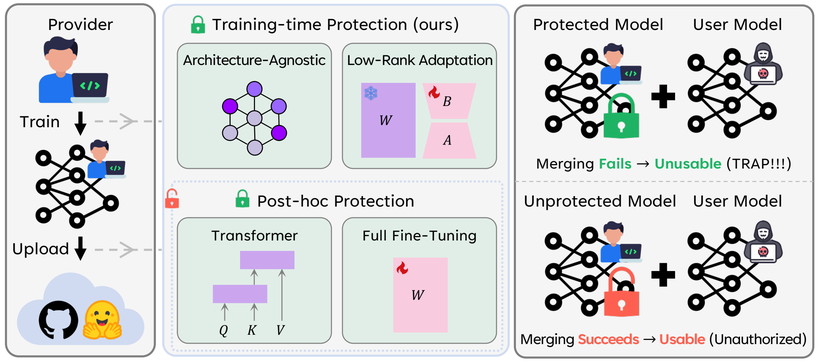
Making Models Unmergeable via Scaling-Sensitive Loss Landscape

Minwoo Jang[P], **Hoyoung Kim**[N], **Jabin Koo**[P], **Jungseul Ok**[P]

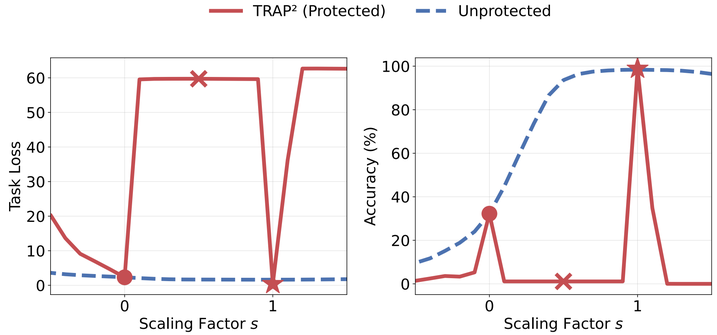
[P] Pohang University of Science and Technology (POSTECH)

[N] National AI Research Lab (NAIRL)

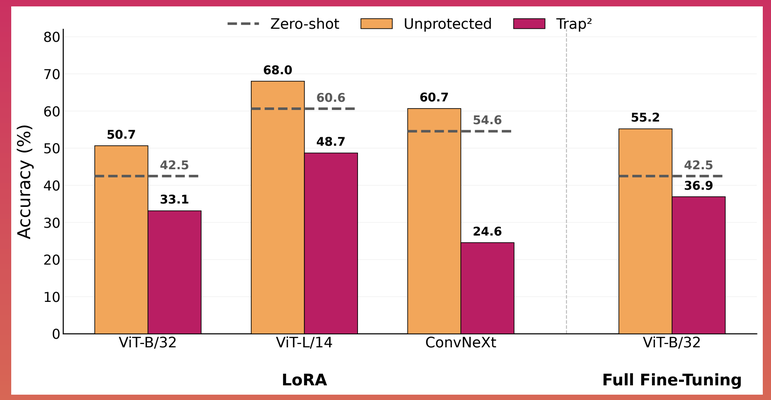
Trap² is the first training-time protection method against **unauthorized merging**, applicable **beyond Transformer** and **beyond full model weights**.



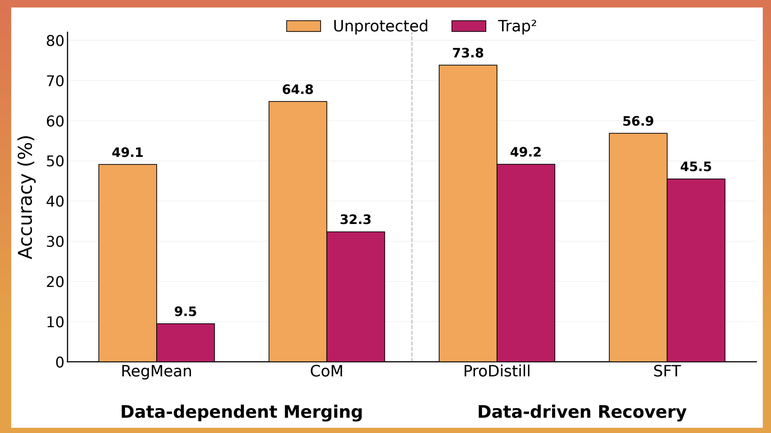
Trap² uses **scaling as a proxy** for merging: stable as intended, **fragile under re-scaling**.



Useful standalone, unreliable when merged.



More data, limited escape.



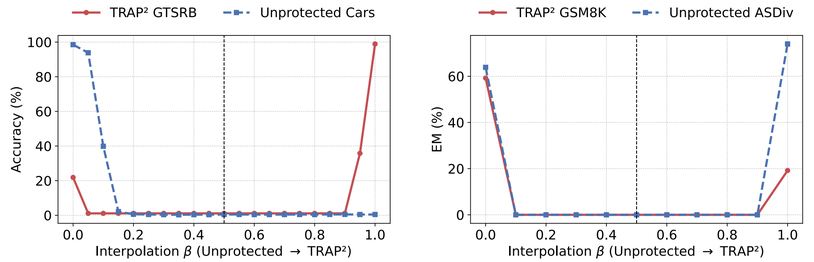
Looking for Internship Opportunities

Project Page

Machine Learning Laboratory · National AI Research Lab

ICML International Conference On Machine Learning

Two Types of Degradation after merging:
① **Down-scaling** (Trap²) ② **Cross-adapter** (Unprotected)



Trap² objective with LoRA $\Delta W = BA$:

$$J(\Delta W) = \mathcal{L}_{nominal}(\Delta W) - \lambda \cdot \mathcal{L}_{off}(\Delta W)$$

$$\mathcal{L}_{nominal}(\Delta W) := \mathcal{L}(W_0 + \Delta W)$$

$$\mathcal{L}_{off}(\Delta W) := \mathbb{E}_{s \in [s_{min},\, 1-\delta] \cup [1+\delta,\, s_{max}]} \left[\frac{1}{s} \cdot \mathcal{L}(W_0 + s \cdot \Delta W) \right]$$

Intuition: $\mathcal{L}_{nominal}$ preserves performance, while $\mathcal{L}_{off}$ induces sensitivity to re-scalings.

Extension of **Trap²** to Full Fine-Tuning:

Calculate $\mathcal{L}_{nominal}$ at $\Delta W_t := W_t - W_0$ at each step t.

Key Takeaways:

Trap² makes an update **scaling-sensitive**, so that it stays **useful at the nominal scale** but **collapses under off-nominal re-scaling** that merging induces.