Robustness Tokens: Towards Adversarial Robustness of Transformers

Brian Pulfer, Yury Belousov, Slava Voloshynovskiy
University of Geneva, Department of Computer Science

Introduction

Previous works [1-3] have shown that additional learnable tokens can improve performances and even interpretability of transformer models. In this work, we learn more tokens to improve **adversarial robustness** instead. Across all models, we measure improved robustness of the features extracted from the backbones, while preserving downstream performances.

Method

We train robustness tokens such that features are not altered for original or adversarial samples

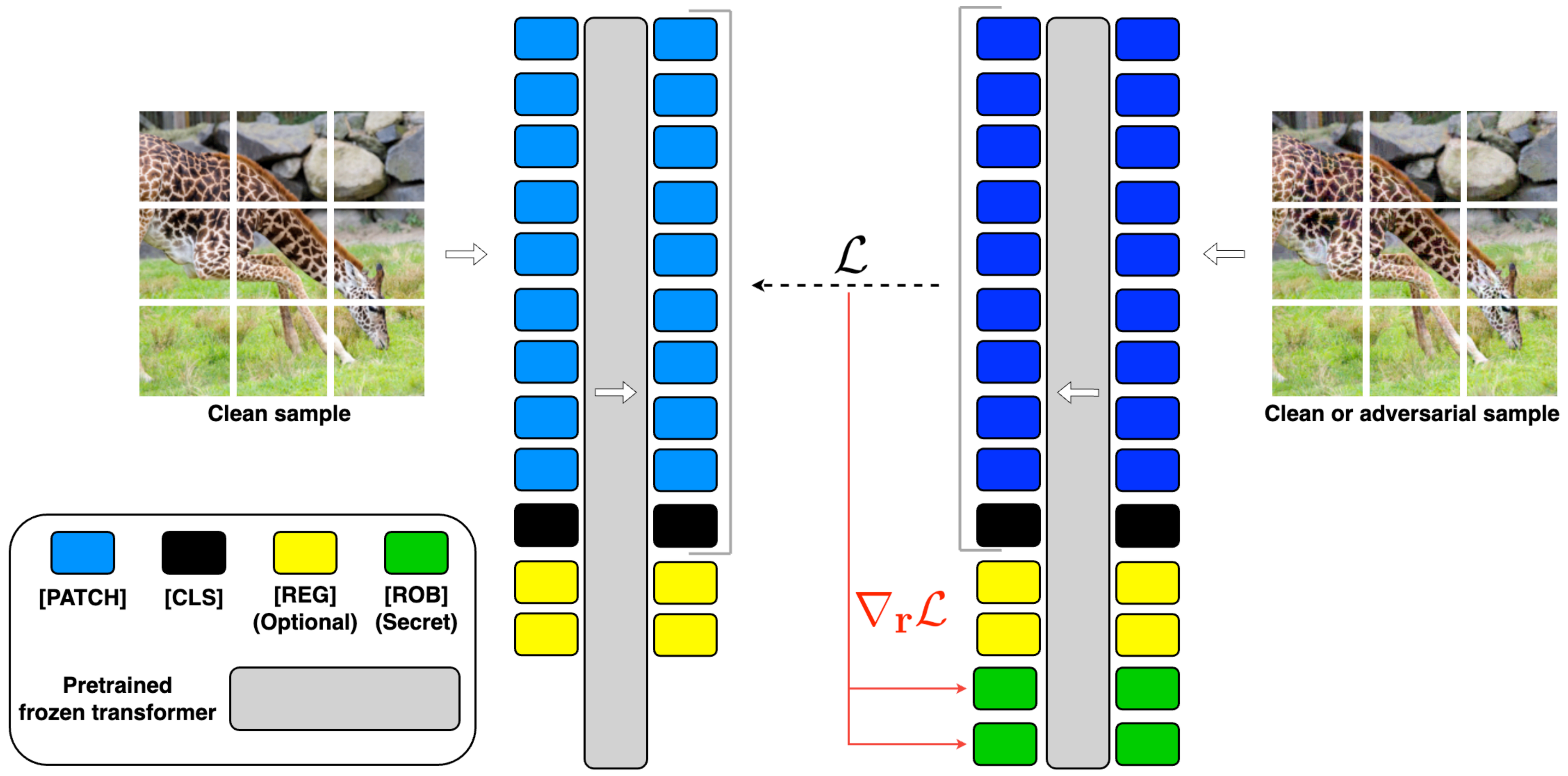
$$\mathcal{L}_{\text{inv}}(\mathbf{r}) = \mathop{\mathbb{E}}_{\mathbf{x}\sim p_{\text{data}}}\left[\frac{f([\mathbf{r},\mathbf{x}])\cdot f(\mathbf{x})}{||f([\mathbf{r},\mathbf{x}])||\,||f(\mathbf{x})||}\right]$$

$$\mathcal{L}_{\text{adv}}(\mathbf{r}) = \mathop{\mathbb{E}}_{\mathbf{x}\sim p_{\text{data}}}\left[\frac{f([\mathbf{r},\mathbf{x}^{\text{adv}}])\cdot f(\mathbf{x})}{||f([\mathbf{r},\mathbf{x}^{\text{adv}}])||\,||f(\mathbf{x})||}\right]$$

With adversarial attacks crafted with:

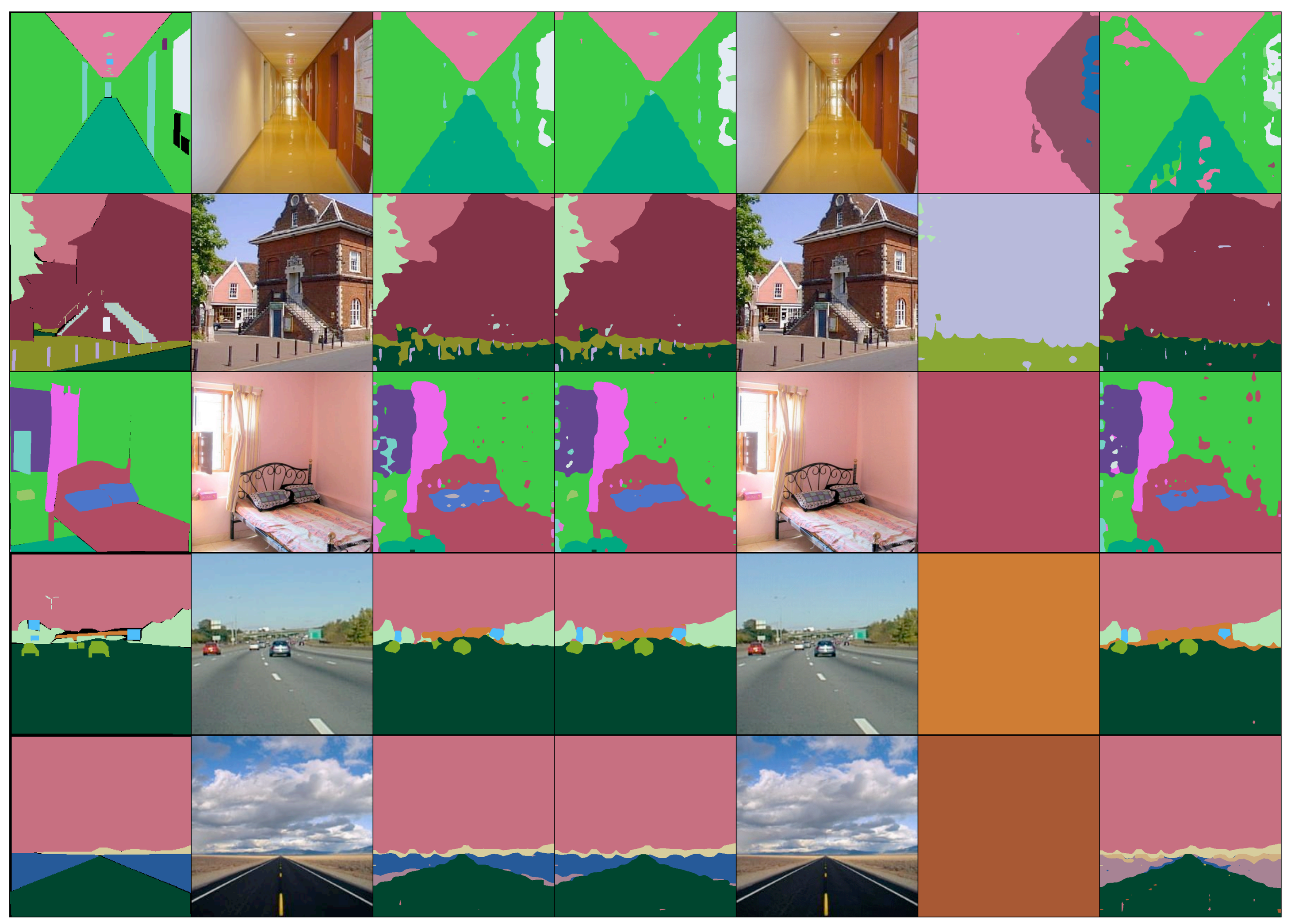
$$\mathcal{L}_{\text{attack}}(\mathbf{x}^{\text{adv}}) = \frac{f(\mathbf{x})\cdot f(\mathbf{x}^{\text{adv}})}{||f(\mathbf{x})||\;||f(\mathbf{x}^{\text{adv}})||}$$

Backbone models are kept fixed throughout, and only tokens are trained.



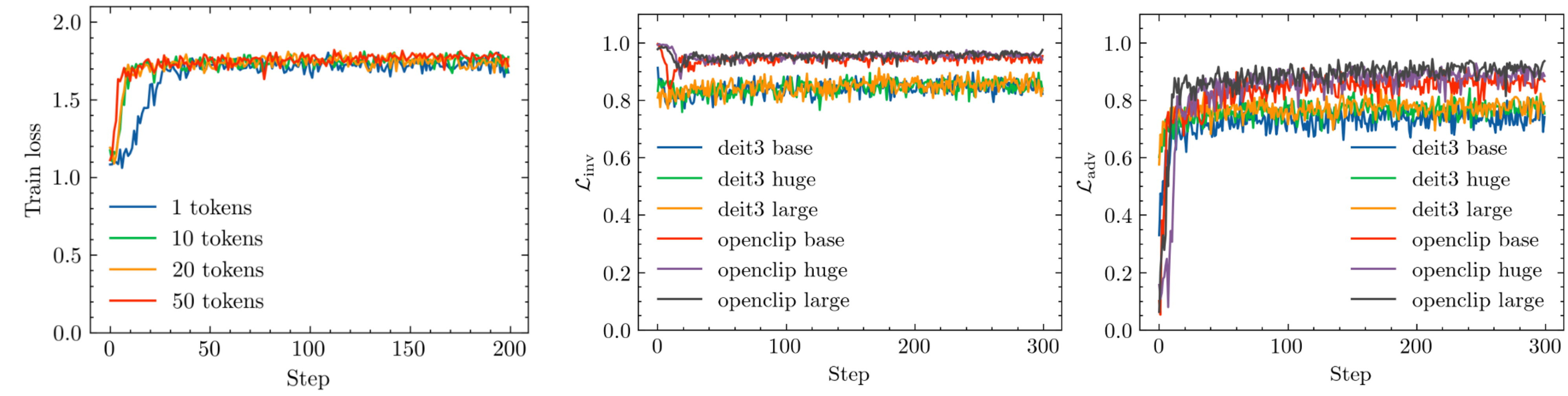
Experimental Results

Model	Performance		Robustness		
	Classification	*Segmentation*	*Features*	*Classification*	*Segmentation*
DiNOv2-S	80.0	41.0	0.09	0.0	2.8
DiNOv2-B	83.4	45.1	0.05	0.0	3.1
DiNOv2-L	85.5	45.1	0.06	0.3	4.5
DiNOv2-G	85.2	46.6	0.12	0.3	4.7
DiNOv2-S + reg	79.8	40.4	0.01	0.0	2.1
DiNOv2-B + reg	83.7	45.8	0.03	0.1	3.0
DiNOv2-L + reg	86.1	46.6	0.03	0.6	4.6
DiNOv2-G + reg	86.3	46.8	0.08	0.9	4.2
DiNOv2-S + rob **(ours)**	78.5	40.6	0.93	31.9	24.6
DiNOv2-B + rob **(ours)**	83.1	45.0	0.92	50.0	23.4
DiNOv2-L + rob **(ours)**	84.2	45.5	0.89	62.9	21.2
DiNOv2-G + rob **(ours)**	85.6	47.2	0.89	63.1	23.3
DiNOv2-S + reg + rob **(ours)**	79.2	40.9	0.93	30.5	22.7
DiNOv2-B + reg + rob **(ours)**	83.1	45.8	0.92	49.7	25.9
DiNOv2-L + reg + rob **(ours)**	85.9	46.7	0.83	58.7	16.2
DiNOv2-G + reg + rob **(ours)**	86.1	47.5	0.90	69.9	25.7

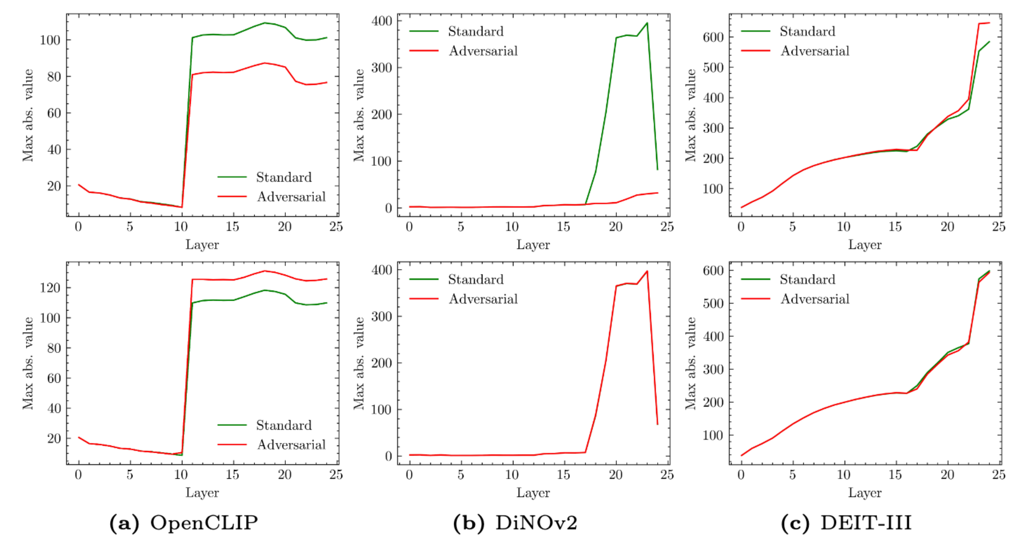


Label Image DiNOv2 **Ours** Adv. DiNOv2 **Ours**

training is **quick** and **cheap**



Adversarial attacks often seem to exploit **massive activations** [4] in transformer models, which robustness tokens learn to restore.



Model	Regular	Robustified
DEIT-III Base	0.16 ± 0.04	**0.74** ± 0.03
DEIT-III Large	0.22 ± 0.03	**0.78** ± 0.02
DEIT-III Huge	0.23 ± 0.03	**0.77** ± 0.02
OpenCLIP Base	-0.02 ± 0.05	**0.86** ± 0.02
OpenCLIP Large	0.13 ± 0.06	**0.91** ± 0.02
OpenCLIP Huge	0.10 ± 0.07	**0.89** ± 0.02

Conclusion

We introduced robustness tokens, learnable tokens that **don't damage model performance but improve robustness** when kept secret. Robustness tokens are quick and cheap to learn, making them widely adoptable.

Code

@Peutlefaire

@bruce-willis

References

[1] Timothée Darcet, undefined., et al, "Vision Transformers Need Registers," 2024.
[2] Guangxuan Xiao., et al, "Efficient Streaming Language Models with Attention Sinks," 2024.
[3] Xiang Lisa Li, Percy Liang, "Prefix-Tuning: Optimizing Continuous Prompts for Generation," 2021.
[4] Mingjie Sun, et al, "Massive Activations in Large Language Models," 2024.