UNIVERSITÉ DE GENÈVE

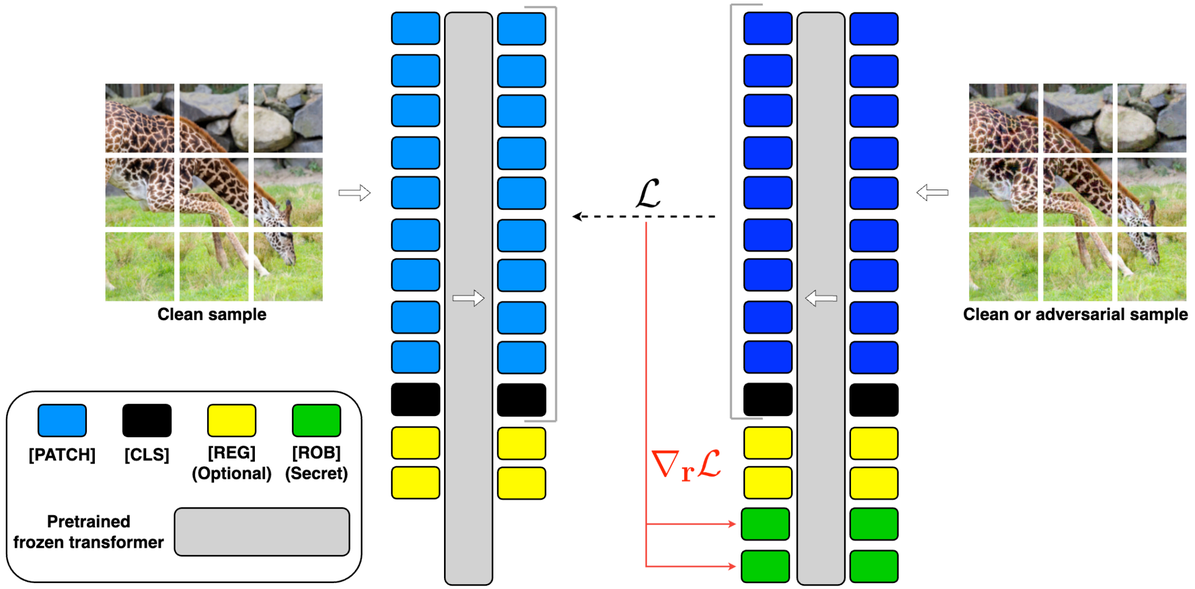
EUROPEAN CONFERENCE ON COMPUTER VISION

MILANO 2024

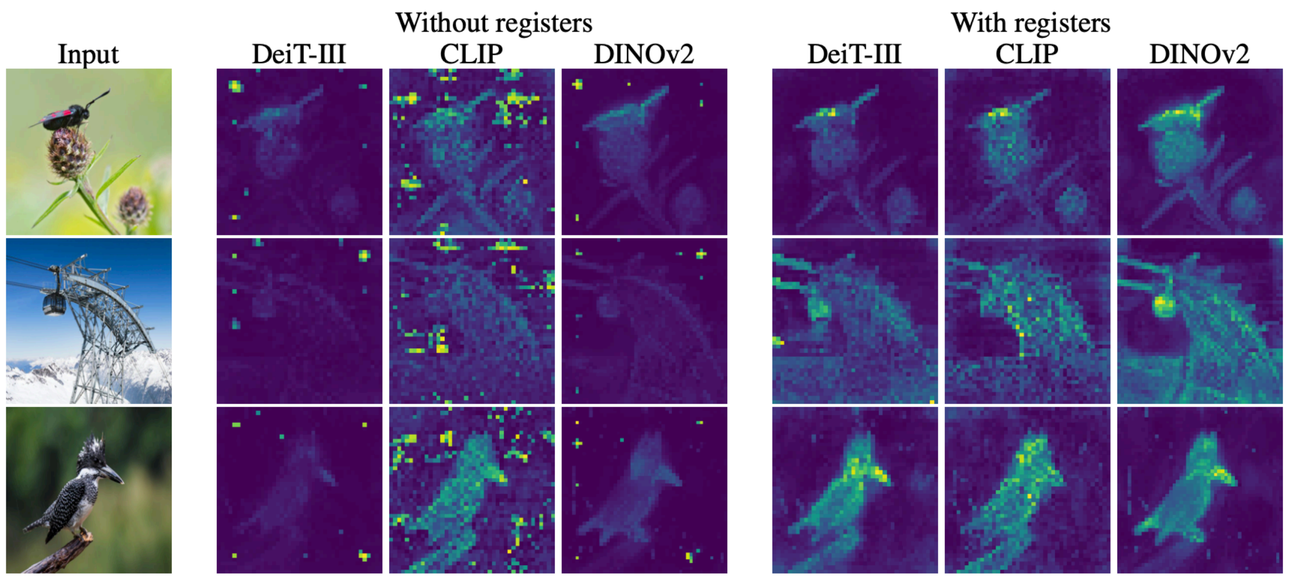
Robustness Tokens: Towards Adversarial Robustness of Transformers

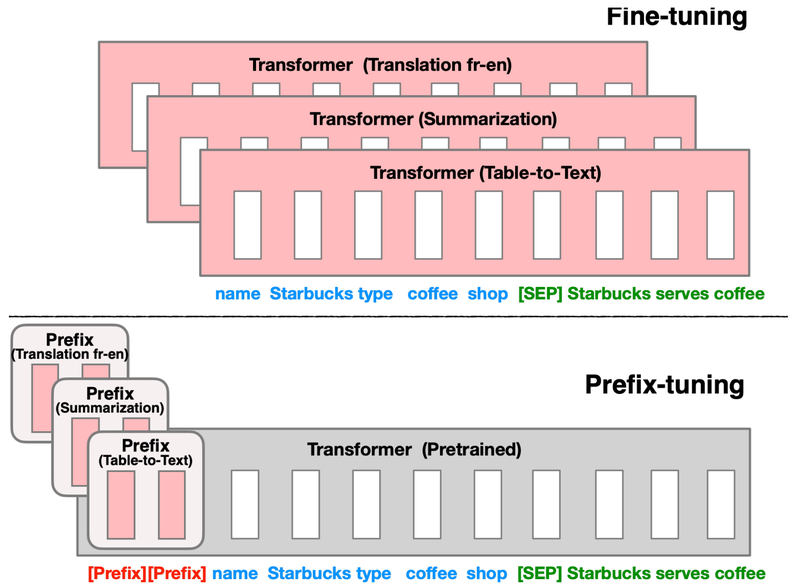
Brian Pulfer, Yury Belousov, Slava Voloshynovskiy

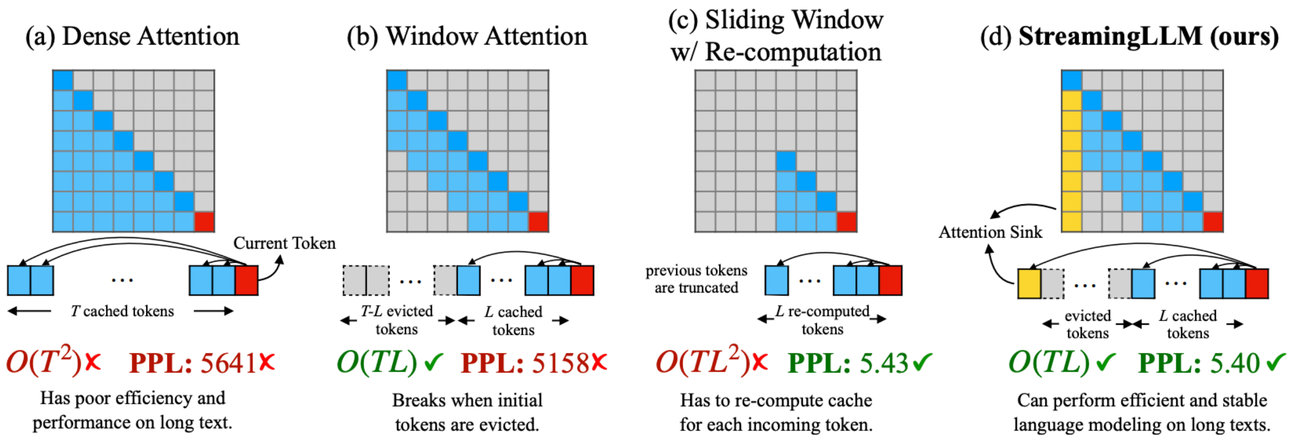
https://github.com/BrianPulfer/robustness-tokens



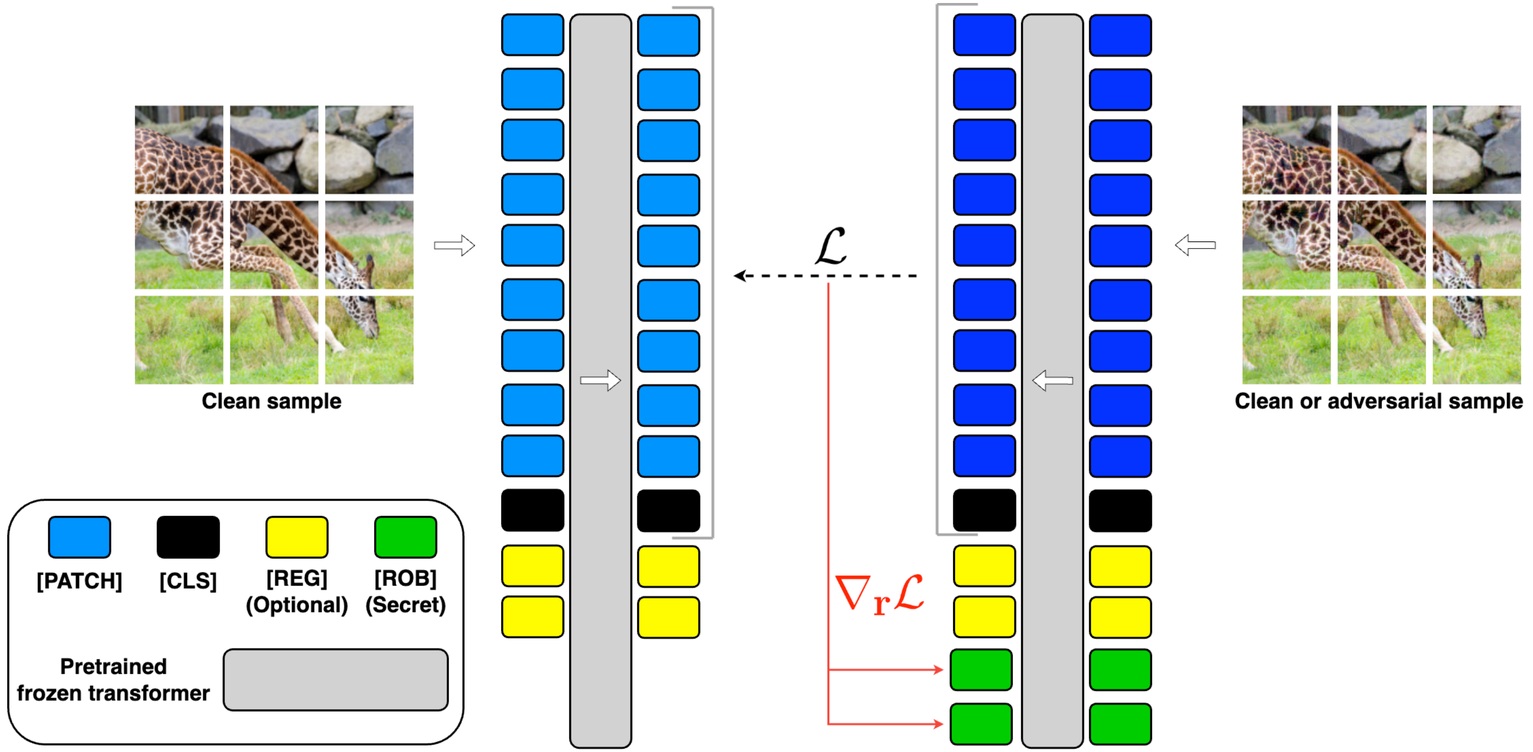
Motivation


Input
Without registers
DeiT-III
CLIP
DINOv2
With registers
DeiT-III
CLIP
DINOv2

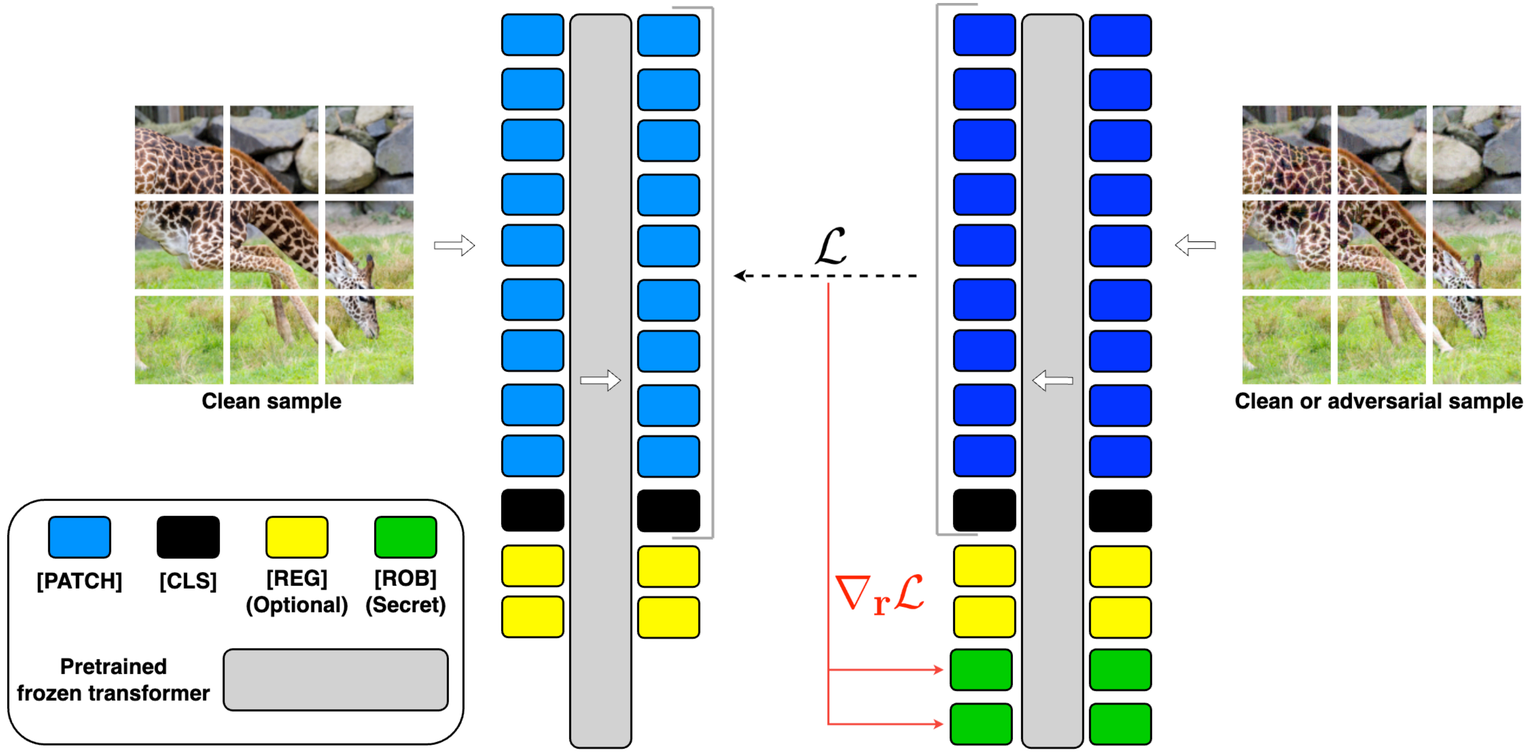

Fine-tuning
Transformer (Translation fr-en)
Transformer (Summarization)
Transformer (Table-to-Text)
name Starbucks type coffee shop [SEP] Starbucks serves coffee
Prefix (Translation fr-en)
Prefix (Summarization)
Prefix (Table-to-Text)
Transformer (Pretrained)
Prefix-tuning
[Prefix][Prefix] name Starbucks type coffee shop [SEP] Starbucks serves coffee


(a) Dense Attention
Current Token
T cached tokens
O(T²)✗ PPL: 5641✗
Has poor efficiency and performance on long text.
(b) Window Attention
T-L evicted tokens
L cached tokens
O(TL)✓ PPL: 5158✗
Breaks when initial tokens are evicted.
(c) Sliding Window w/ Re-computation
previous tokens are truncated
L re-computed tokens
O(TL²)✗ PPL: 5.43✓
Has to re-compute cache for each incoming token.
(d) StreamingLLM (ours)
Attention Sink
evicted tokens
L cached tokens
O(TL)✓ PPL: 5.40✓
Can perform efficient and stable language modeling on long texts.

Methodology



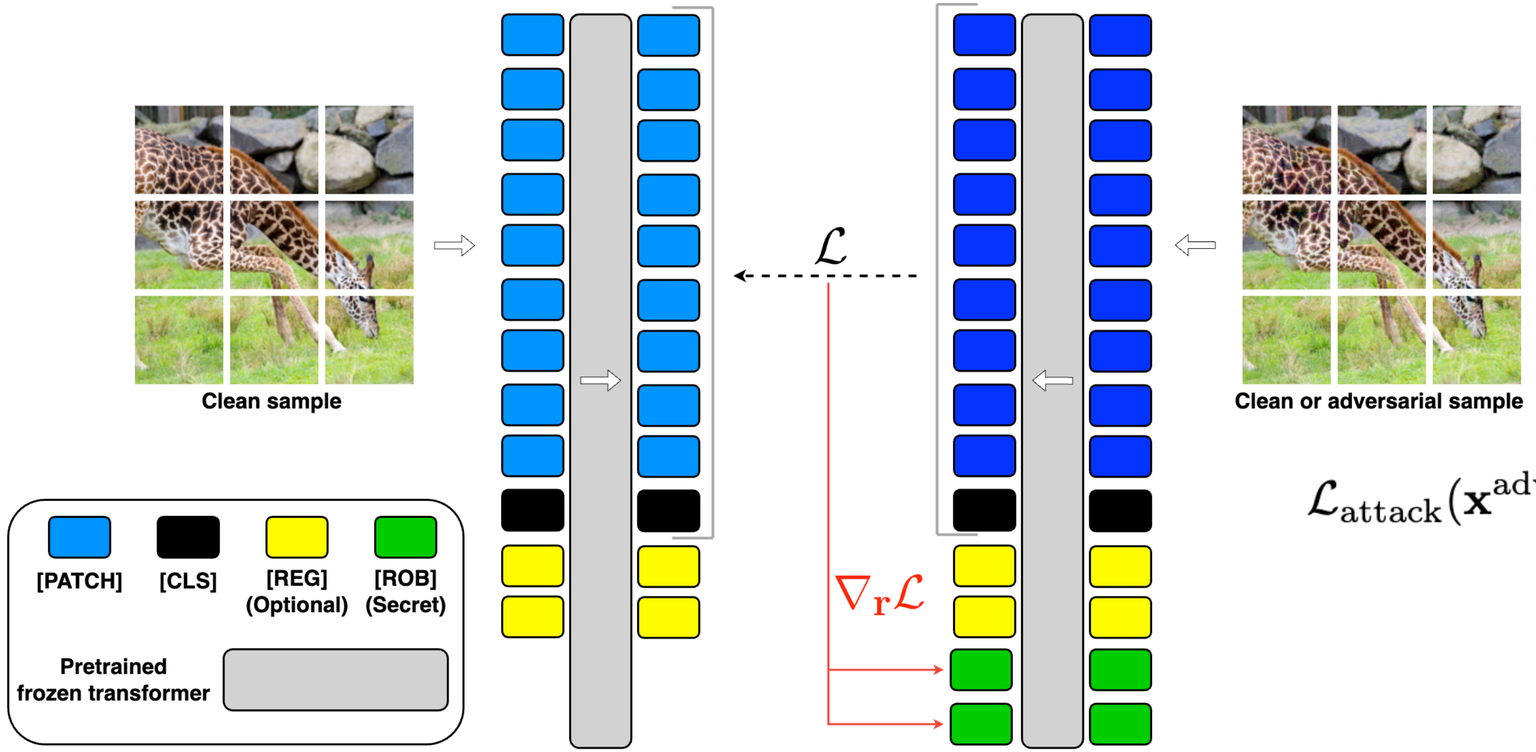
Methodology



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

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

Methodology

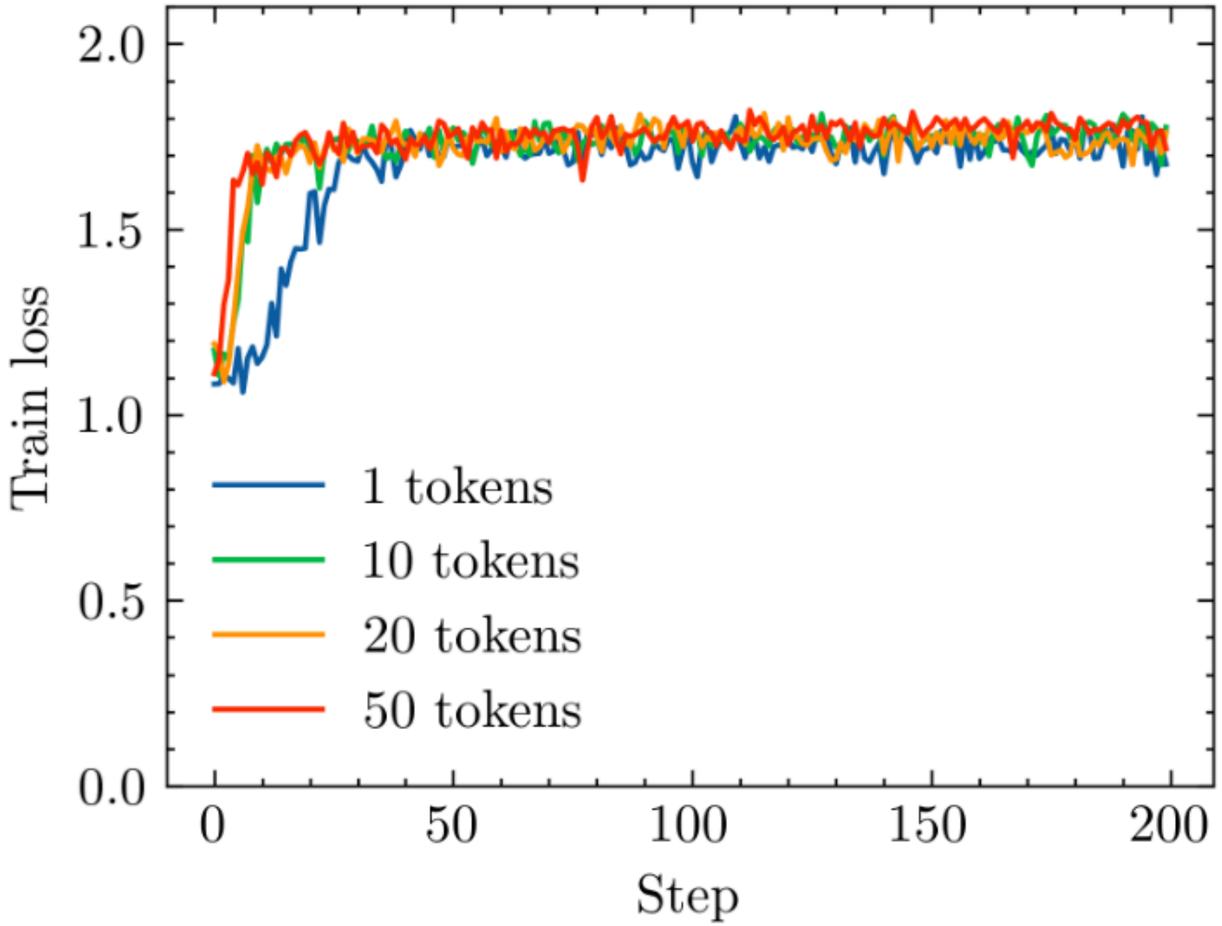


$$\mathcal{L}_{\text{attack}}(\mathbf{x}^{\text{adv}}) = \frac{f(\mathbf{x}) \cdot f(\mathbf{x}^{\text{adv}})}{\|f(\mathbf{x})\| \, \|f(\mathbf{x}^{\text{adv}})\|}$$

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

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

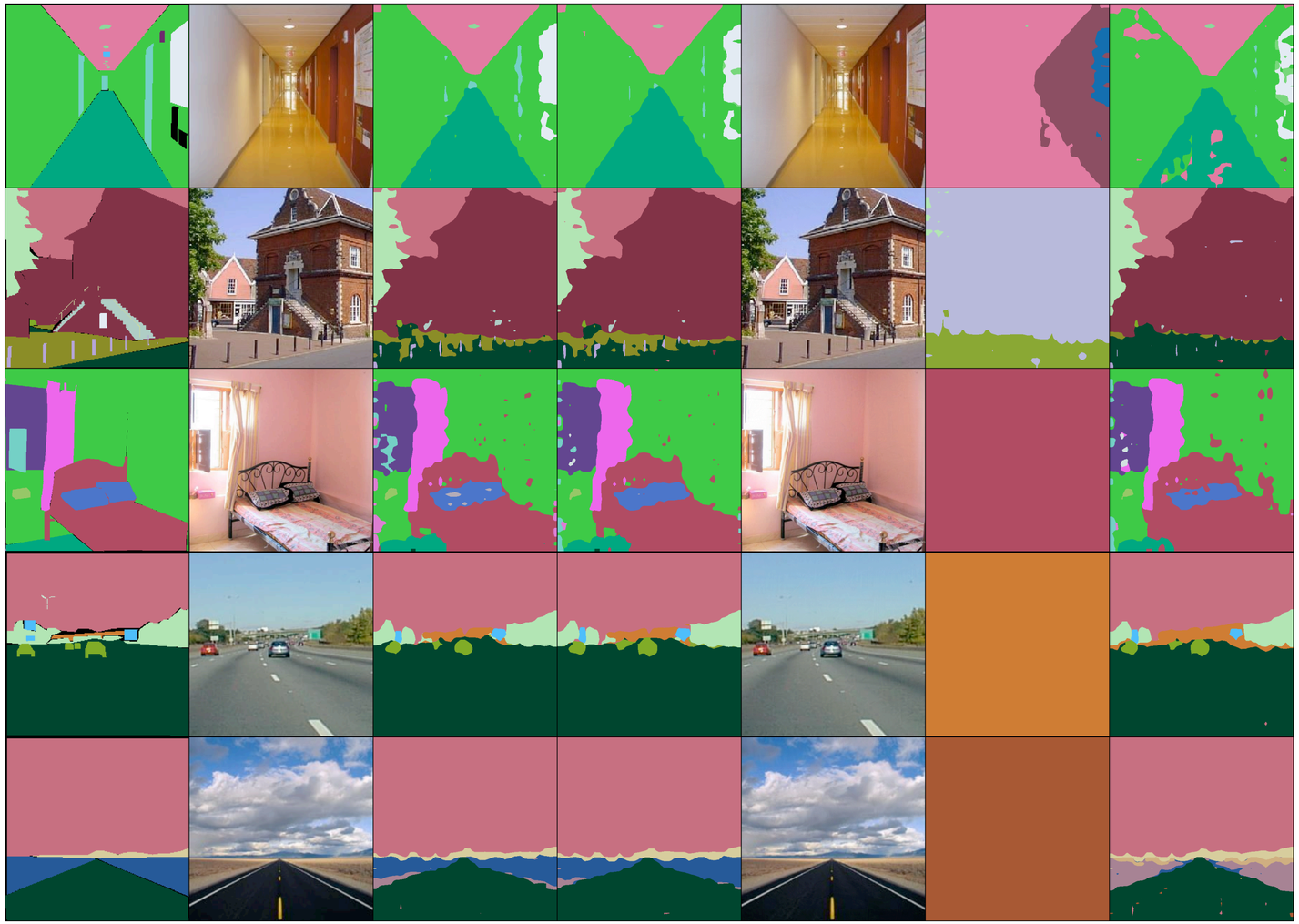
Experiments



Experiments

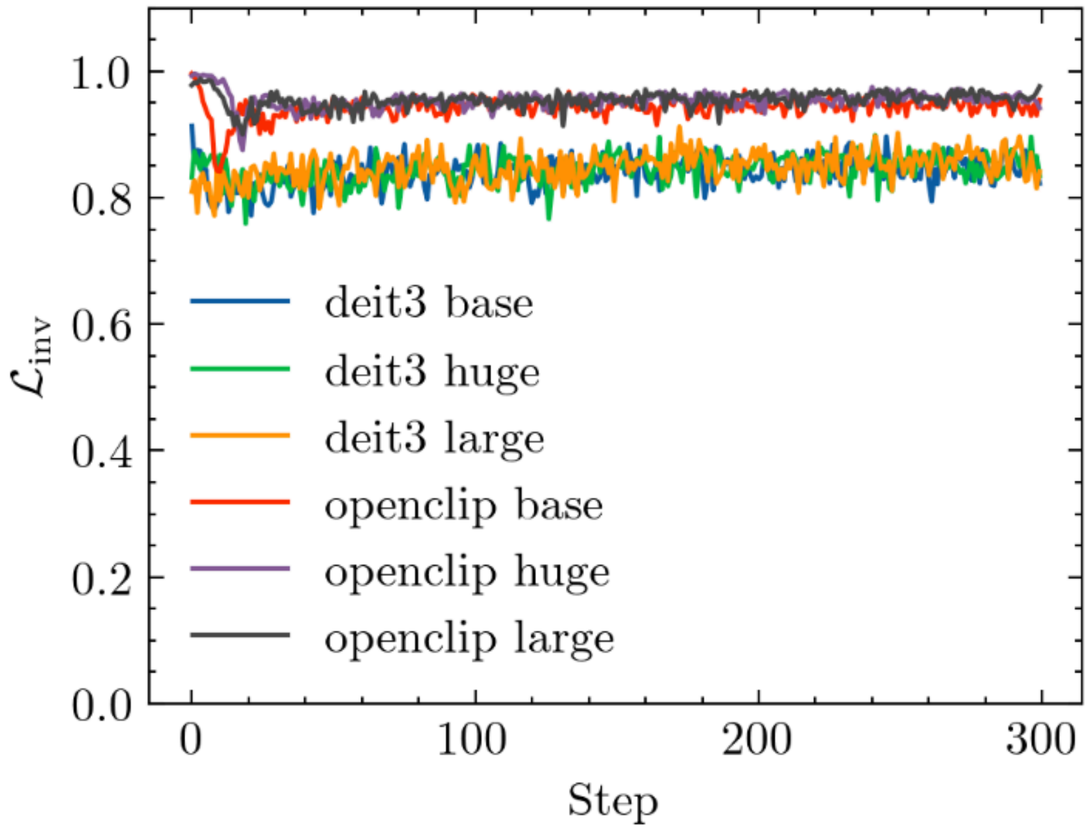
Model	Performance		Robustness		
	Classification	*Segmentation*	*Features*	*Classification*	*Segmentation*
DiNOv2-S	80.0	41.0	0.09	0.0	2.8
DiNOv2-B	83.4	45.1	0.05	0.0	3.1
DiNOv2-L	85.5	45.1	0.06	0.3	4.5
DiNOv2-G	85.2	46.6	0.12	0.3	4.7
DiNOv2-S + reg	79.8	40.4	0.01	0.0	2.1
DiNOv2-B + reg	83.7	45.8	0.03	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

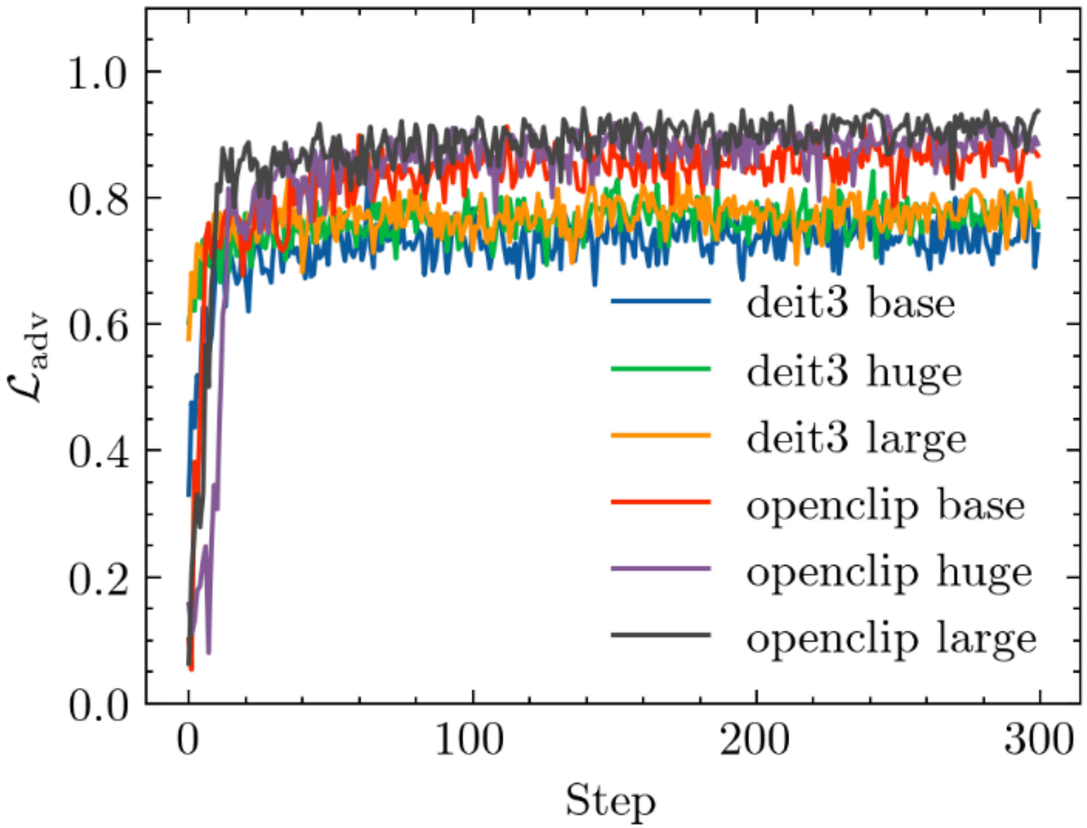
Experiments



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

Experiments





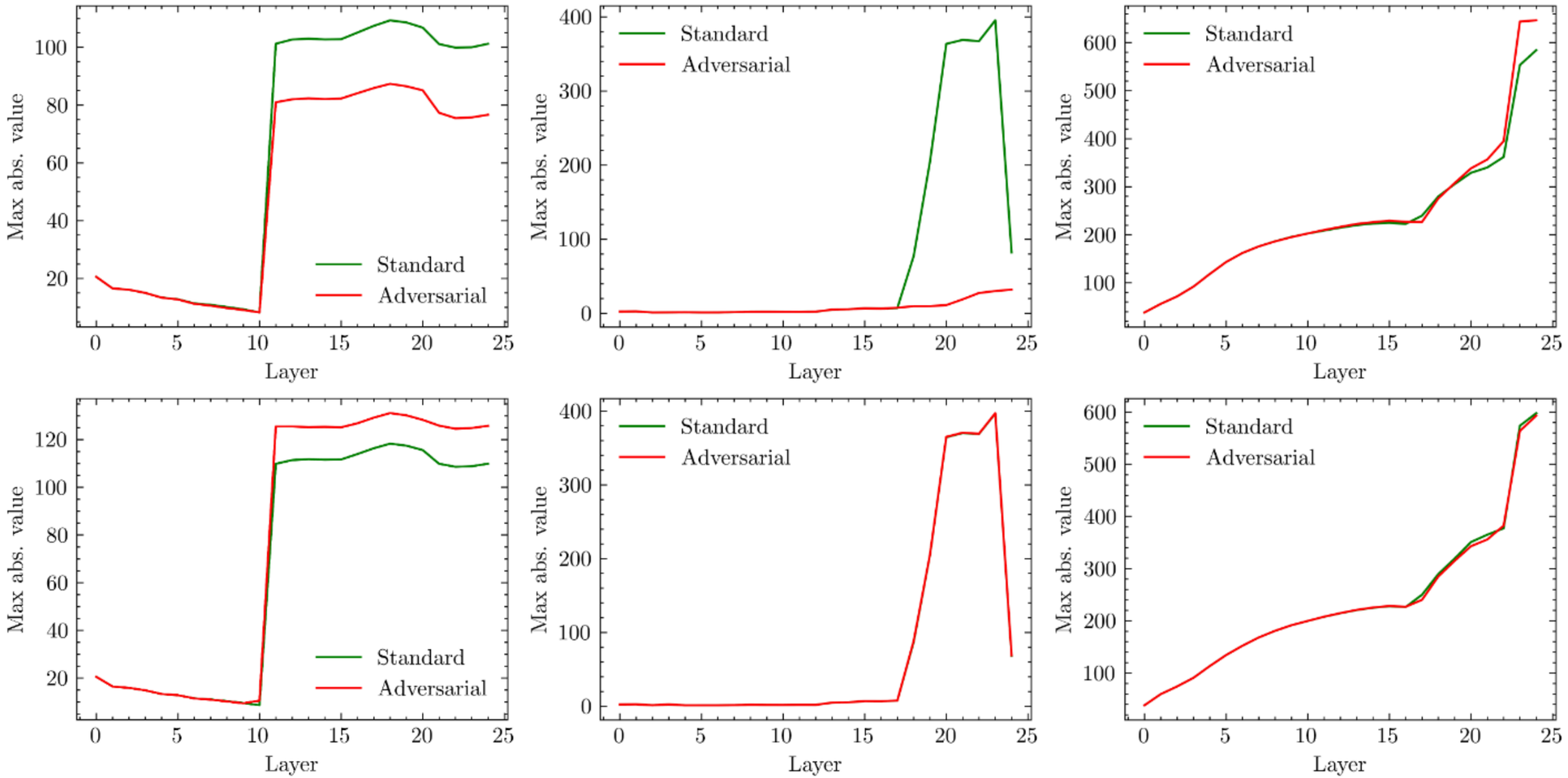
$$\mathcal{L}_{\text{inv}}(\mathbf{r}) = \mathop{\mathbb{E}}_{\mathbf{x}\sim p_{\text{data}}} \left[\frac{f([\mathbf{r},\mathbf{x}]) \cdot f(\mathbf{x})}{\|f([\mathbf{r},\mathbf{x}])\| \|f(\mathbf{x})\|} \right]$$

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

Experiments

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

Experiments



(a) OpenCLIP **(b)** DiNOv2 **(c)** DEIT-III

Conclusion

- Robustness tokens are cheap and quick to train
- Same performances, improved robustness
- Adversarial attacks exploit massive activations

UNIVERSITÉ
DE GENÈVE

EUROPEAN CONFERENCE ON COMPUTER VISION

MILANO
2024

Thank you for your attention

https://github.com/BrianPulfer/robustness-tokens

@Peutlefaire