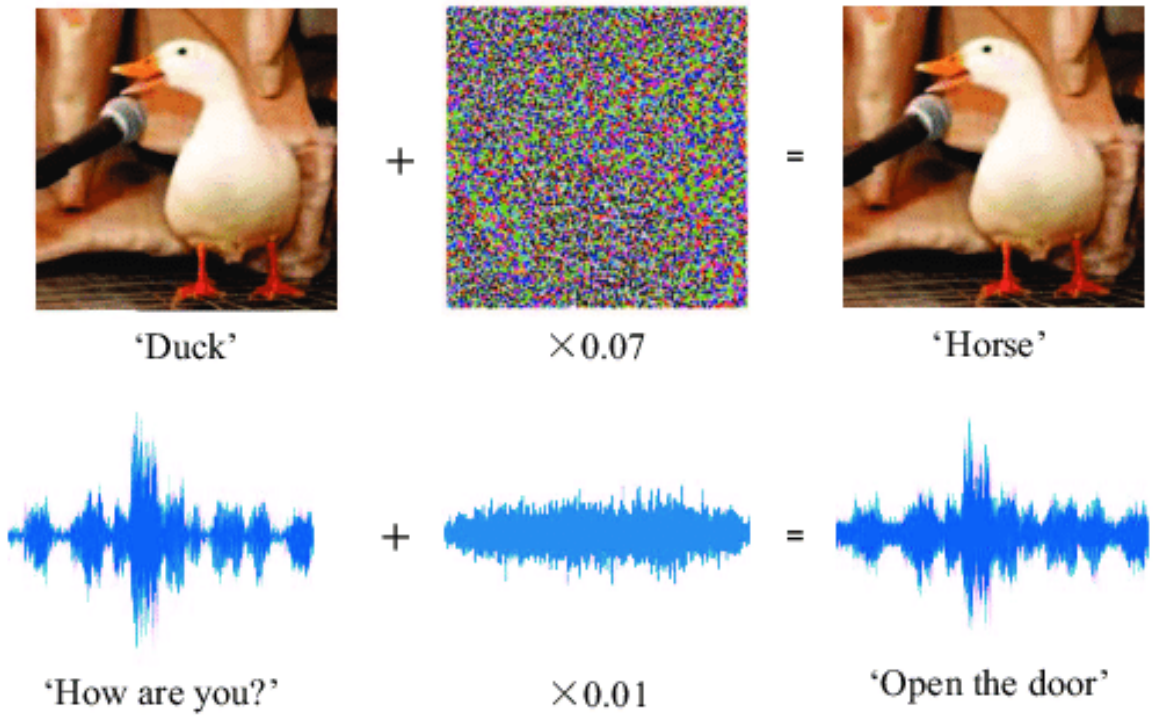
MALT Powers Up Adversarial Attacks

Odelia Melamed, Gilad Yehudai, Adi Shamir

NeurIPS 2024

Adversarial Examples



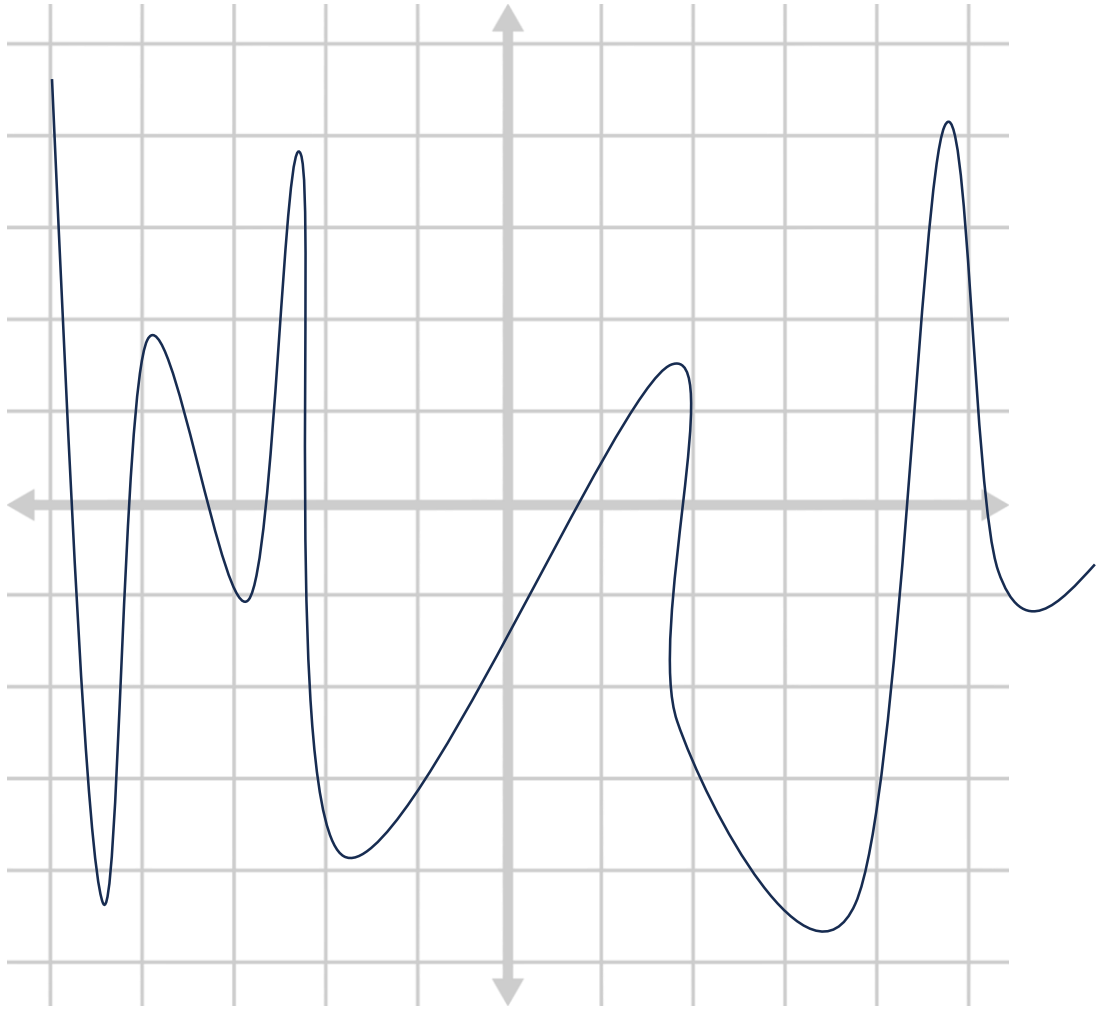
Related Work

- Local Linearity of neural networks
  - Goodfellow et al. 2014 and **FGSM**
  - **DeepFool** attack (Moosavi-Dezfooli et al. [2016]) and **FAB** attack (Croce and Hein [2020a]) presenting step-wise targeting.

- Untargeted Parameter-free state-of-the-art attack
  - Croce and Hein 2020b presents **AutoAttack**, the current SOTA.
  - Utilizing targeted attacks toward 9 top logits targets

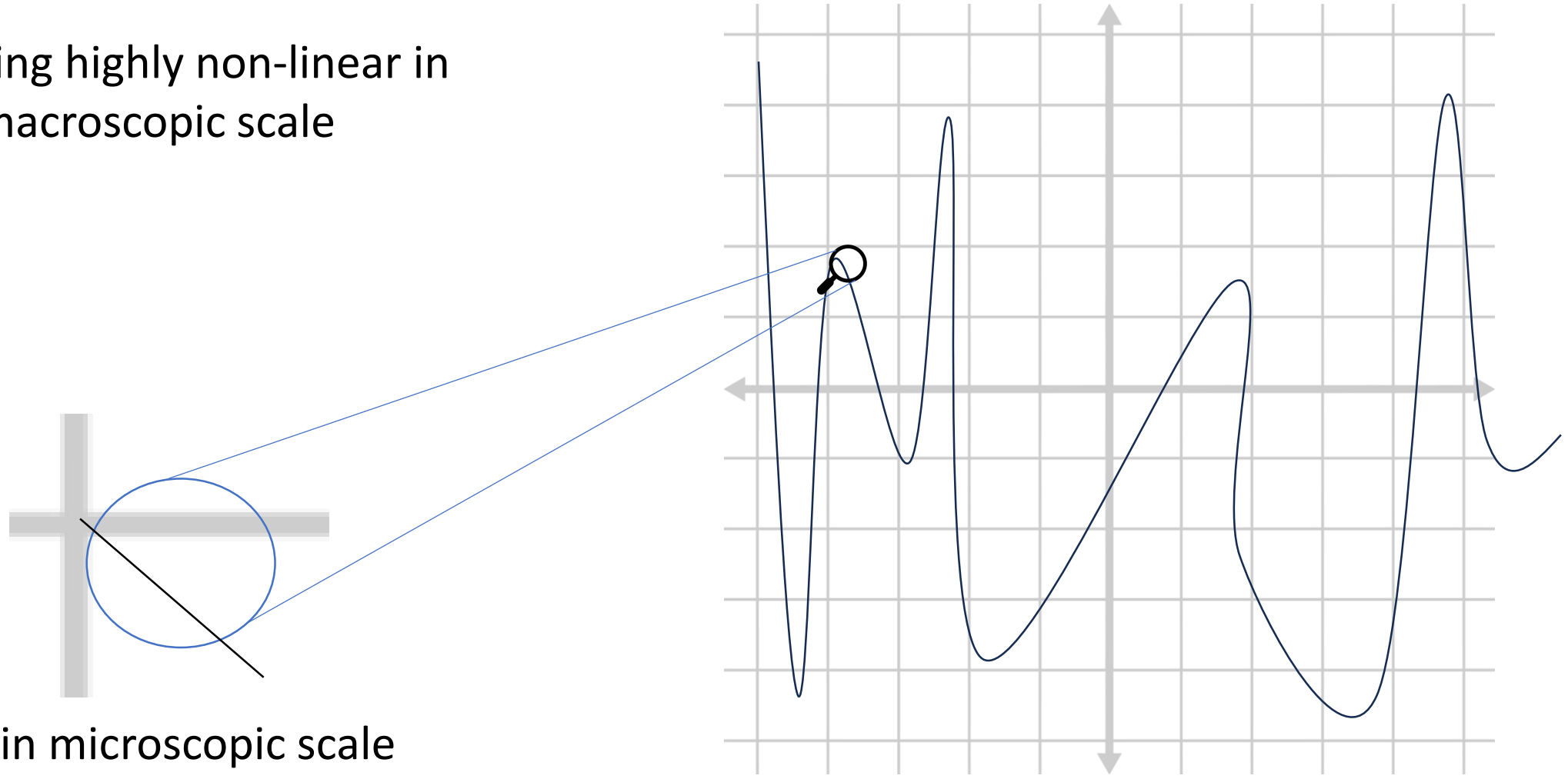
Different Scales of a Piecewise-Linear Function

Looking highly non-linear in the macroscopic scale



Different Scales of a Piecewise-Linear Function

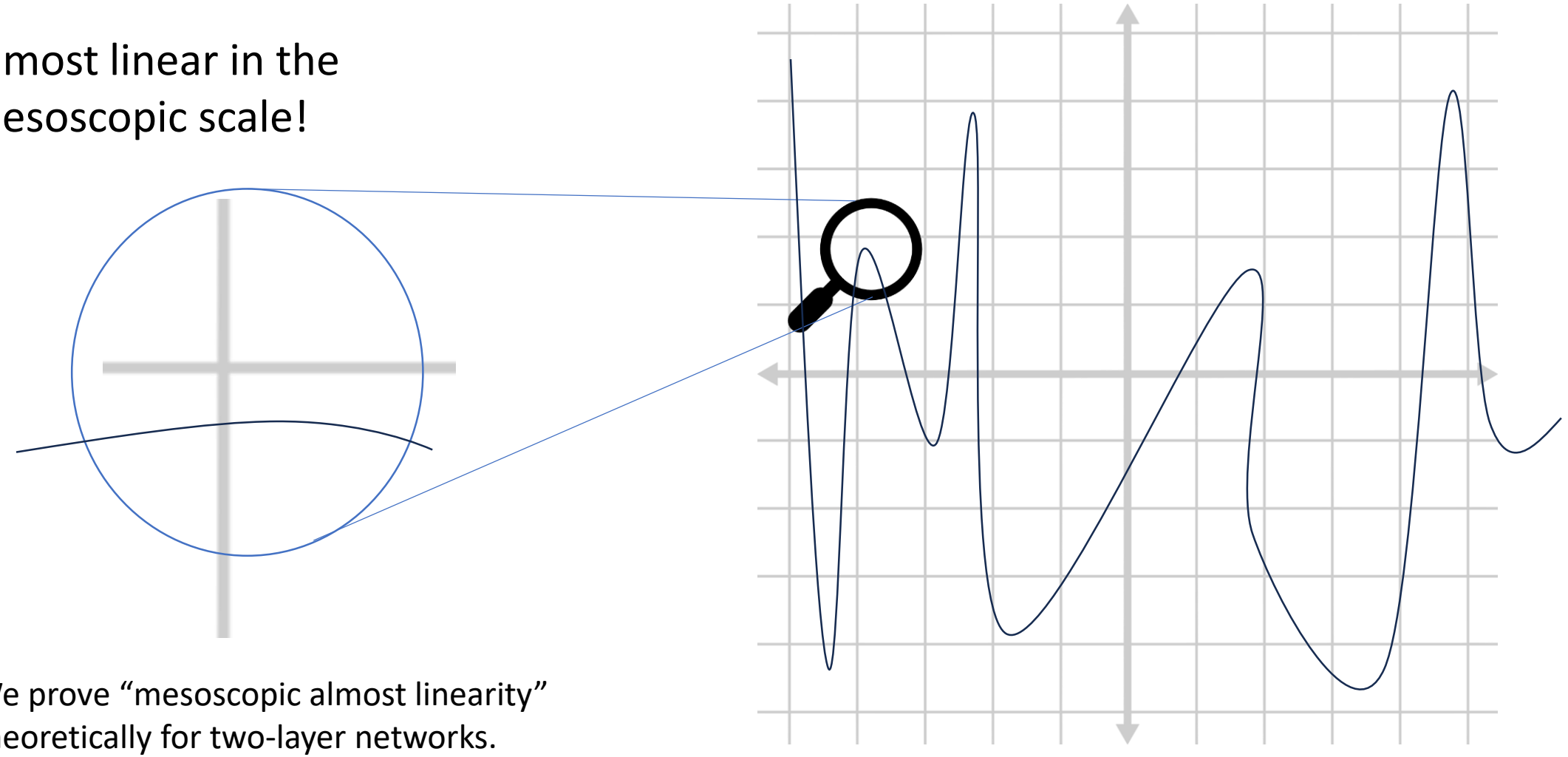
Looking highly non-linear in the macroscopic scale



linear in microscopic scale

Different Scales of a Piecewise-Linear Function

Almost linear in the mesoscopic scale!

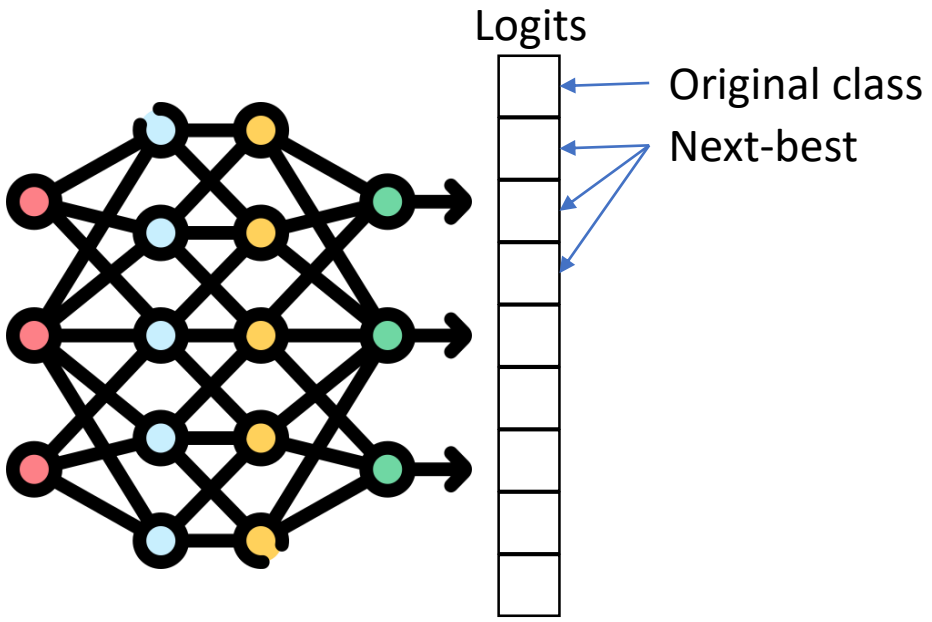


- We prove “mesoscopic almost linearity” theoretically for two-layer networks.

Targeting Methods

Naïve targeting: for a k-classes classifier f, rate the other k-1 classes according to their confidence logits.

This method ignores the “slope” of the function.



Targeting in Linear Functions

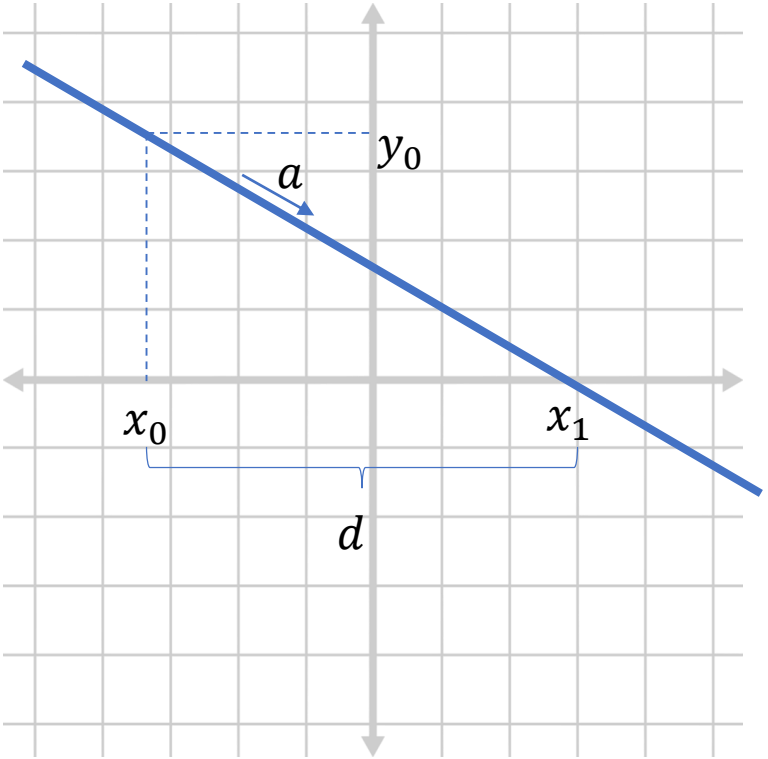
Let $F(x) = Wx + b$ be a linear k-class classifier, where w_i is the i-th row of W.

Then, for x_0 classified as class l, the distance to class i would be:

$$\mathrm{d} = \frac{\langle w_i - w_l, x_0 \rangle}{\|w_i - w_l\|}$$

The logit difference

The "slope"



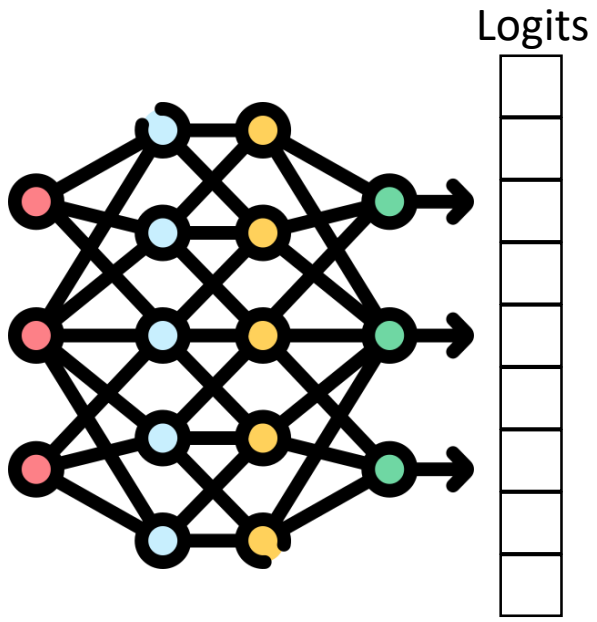
And the best target j will be

$$j = \mathrm{argmin}_i \frac{\langle w_i - w_l, x_0 \rangle}{\|w_i - w_l\|}$$

MALT Targeting

Due to almost linearity in the mesoscopic scale -

MALT uses linear targeting method as pre-calculation, to find the closest target in non-linear classifiers.



and ∇f!

The MALT Attack

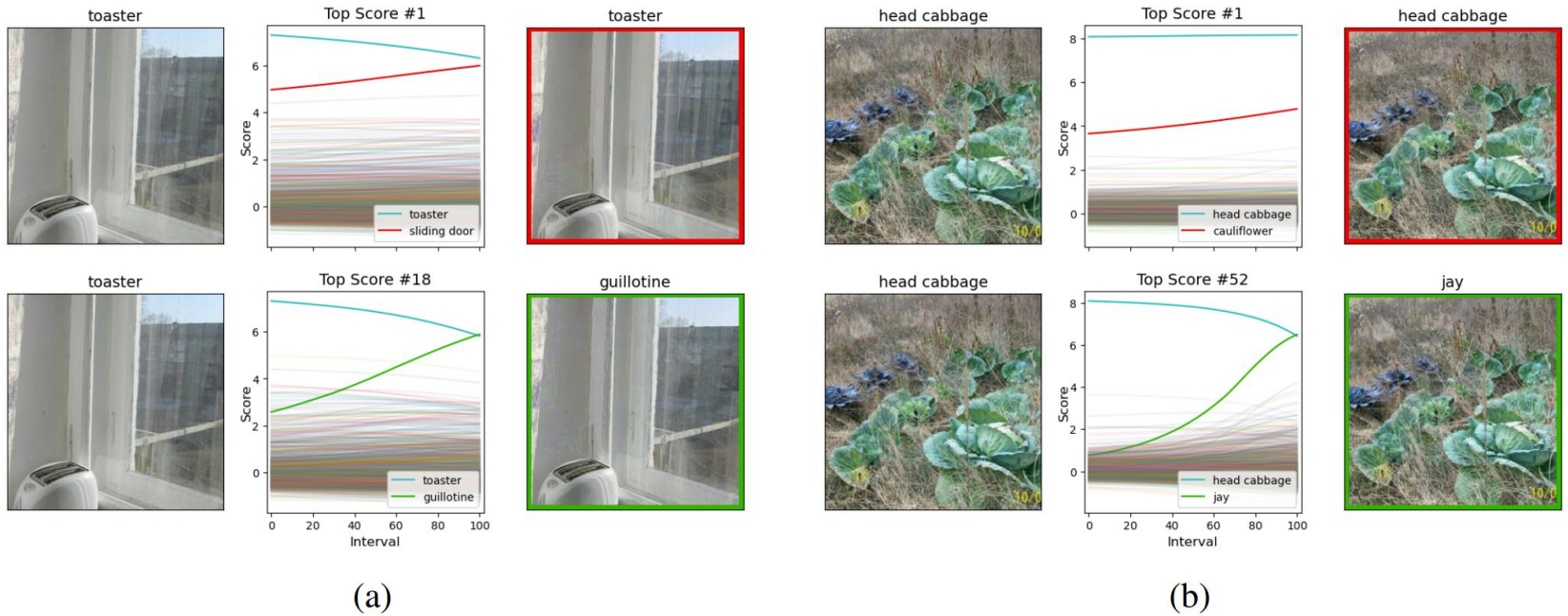


Figure 1: Examples of images from the ImageNet dataset that AutoAttack fails to attack while MALT succeeds. The top row shows an APGD attack on the target class with the highest logit, and the bottom row shows an APGD attack on the class which MALT finds and succeeds, corresponding to the (a) 18th and (b) 52nd classes with the highest logits. The images are shown before and after the attack, and the change in logits is presented in the middle column.

The MALT attack

MALT attack outperform the State-Of-The-Art Auto Attack, and is five times faster!

Table 1: **CIFAR100 -** L_∞ robust accuracy *(lower is better)*, comparing MALT and AutoAttack, which is the current state of the art.

MODEL		ROBUSTNESS			
	ACC.	MALT	SOTA	DIFF	SPEED-UP
WRN-28-10 [WANG ET AL., 2023]	72.58%	38.79%	38.83%	−0.04%	× 3.36 ±0.18
WRN-70-16 [WANG ET AL., 2023]	75.22%	42.66%	42.67%	−0.01%	× 3.87 ±0.08
WRN-28-10 [CUI ET AL., 2023]	73.83%	39.18%	39.18%	0%	× 3.43 ±0.08
WRN-70-16 [GOWAL ET AL., 2020]	69.15%	36.81%	36.88%	−0.07%	× 3.42 ±0.09

Table 2: **ImageNet -** L_∞ robust accuracy *(lower is better)*, comparing MALT and AutoAttack, which is the current state of the art.

MODEL		ROBUSTNESS			
	ACC.	MALT	SOTA	DIFF.	SPEED-UP
SWIN-L [LIU ET AL., 2023]	79.18%	59.84%	59.90%	−0.06%	× 5.18 ±0.04
CONVNEXT-L [LIU ET AL., 2023]	78.20%	58.82%	58.88%	−0.06%	× 5.22 ±0.1
CONVNEXT-L+ [SINGH ET AL., 2024]	77.02%	57.94%	57.96%	−0.02%	× 4.86 ±0.06
SWIN-B [LIU ET AL., 2023]	76.22%	56.54%	56.56%	−0.02%	× 5.02 ±0.03
CONVNEXT-B+ [SINGH ET AL., 2024]	76.00%	56.48%	56.52%	−0.04%	× 5.00 ±0.07

Thank you