RAMP: Boosting Adversarial Robustness Against Multiple l_p Perturbations for Universal Robustness

Enyi Jiang

Gagandeep Singh

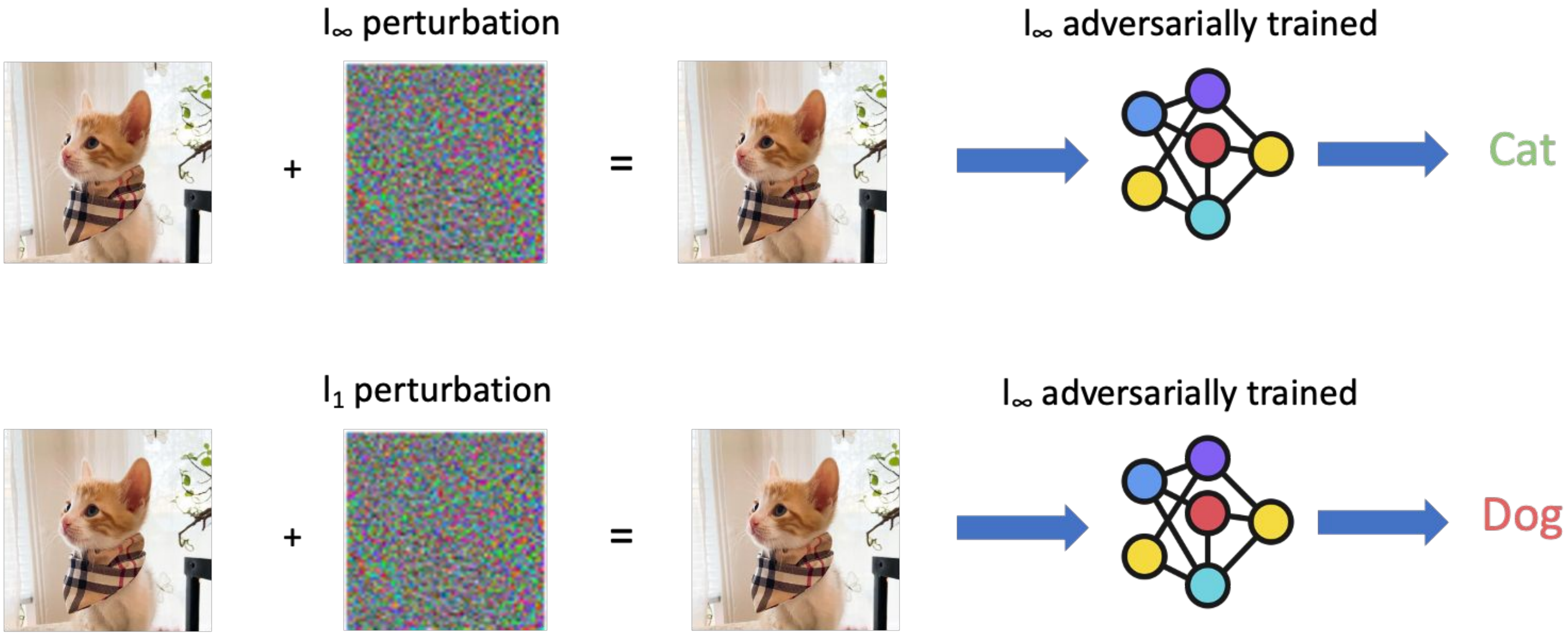
UIUC

ILLINOIS

NEURAL INFORMATION PROCESSING SYSTEMS

Multi-Norm Adversarial Robustness

l_∞ robust != l_p (p = 1,2) robust



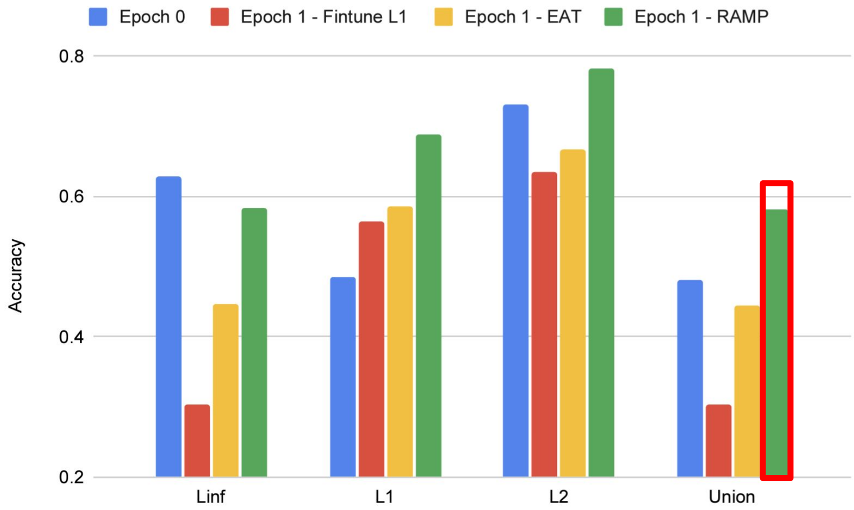
Multi-Norm and Accuracy/Robustness Trade-offs

Multi-Norm tradoffs

=> Logits Pairing

Accuracy/robustness tradeoff

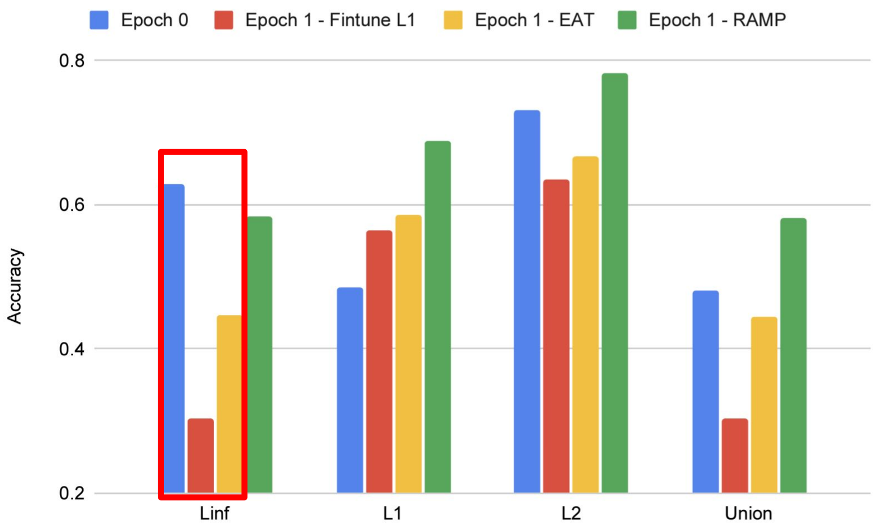
=> Gradient Projection



Key tradeoff: l_∞ - l_1

RAMP: Logits Pairing

Observation: Fine-tune a l_q-AT model on l_r examples reduces l_q robustness



RAMP: Logits Pairing

Solution: Regularize l_q, l_r logits on ***correctly predicted*** l_q subsets via KL loss

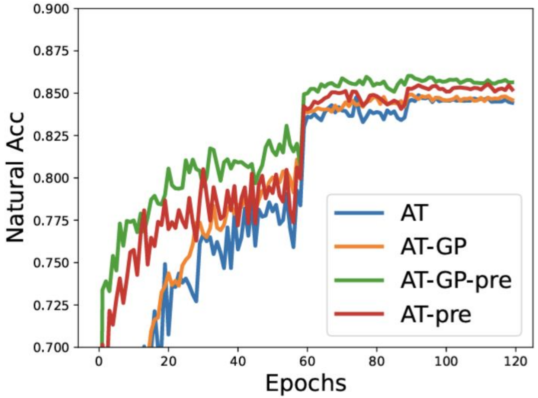
$$\mathcal{L}_{KL} = \frac{1}{n_c} \cdot \sum_{i=1}^{n_c} \sum_{j=0}^{k} p_q[\gamma[i]][j] \cdot \log\left(\frac{p_q[\gamma[i]][j]}{p_r[\gamma[i]][j]}\right)$$

$$\mathcal{L} = \mathcal{L}_{max} + \lambda \cdot \mathcal{L}_{KL}$$

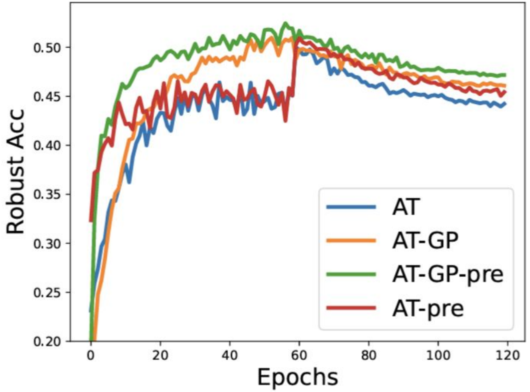
Combine with MAX-style loss

RAMP: Gradient Projection (GP)

Observation: Natural training (NT) can help adversarial robustness



(a) Clean Accuracy



(b) Robust Accuracy: PGD-20

RAMP: Gradient Projection (GP)

Solution: Find and combine useful components of NT with AT via GP

$$\mathbf{GP}(\widehat{g}_n^l, \widehat{g}_a^l) = \begin{cases} \cos(\widehat{g}_n^l, \widehat{g}_a^l) \cdot \widehat{g}_n^l, & \cos(\widehat{g}_n^l, \widehat{g}_a^l) > 0 \\ 0, & \cos(\widehat{g}_n^l, \widehat{g}_a^l) \leq 0 \end{cases}$$

$$g_p = \bigcup_{l \in \mathcal{M}} \mathbf{GP}(\widehat{g}_n^l, \widehat{g}_a^l)$$

$$f^{(r+1)} = f^{(r)} + \beta \cdot g_p + (1 - \beta) \cdot \widehat{g}_a$$

Layerwise operations

Theorem 4.5 (Error Analysis of GP). *When the model dimension $m \to \infty$, for an epoch t, we have an approximation of the error difference $\Delta_{AT}^2 - \Delta_{GP}^2$ as follows*

$$\Delta_{AT}^2 - \Delta_{GP}^2 \approx \beta(2 - \beta)\mathbb{E}_{\widehat{\mathcal{D}}_a^t}\|g_a - \widehat{g}_a\|_\pi^2 - \beta^2\bar{\tau}^2\|g_a - \widehat{g}_n\|_\pi^2$$

Experiment Result: Robust Fine-tuning

RAMP obtains **better union accuracy and accuracy-robustness tradeoff**

	Models	Methods	Clean	l_∞	l_2	l_1	Union
	WRN-70-16-l_∞ (*) [Gowal et al., 2020]	E-AT	89.6	54.4	76.7	58.0	51.6
		RAMP	90.6	54.7	74.6	57.9	**53.3**
	WRN-34-20-l_∞ [Gowal et al., 2020]	E-AT	87.8	49.0	71.6	49.8	45.1
		RAMP	87.1	49.7	70.8	50.4	**46.9**
	WRN-28-10-l_∞ (*) [Carmon et al., 2019]	E-AT	89.3	51.8	74.6	53.3	47.9
CIFAR-10		**RAMP**	89.2	55.9	74.7	55.7	**52.7**
	WRN-28-10-l_∞ (*) [Gowal et al., 2020]	E-AT	89.8	54.4	76.1	56.0	50.5
		RAMP	89.4	55.9	74.7	56.0	**52.9**
	RN-50-l_∞ [Engstrom et al., 2019]	E-AT	85.3	46.5	68.3	45.3	41.6
		RAMP	84.3	47.0	67.7	46.5	**43.3**
	XCiT-S-l_∞ [Debenedetti and Troncoso—EPFL, 2022]	E-AT	68.4	38.1	51.8	23.8	23.4
ImageNet		**RAMP**	66.0	35.7	50.2	30.0	**29.1**
	RN-50-l_∞ [Engstrom et al., 2019]	E-AT	58.2	26.9	39.5	18.8	17.8
		RAMP	55.6	25.1	38.3	22.4	**20.9**

Experiment Result: Varying Epsilons

RAMP consistently outperforms other baselines **when key tradeoff pair changes**

		$(12, 0.5, \frac{2}{255})$					$(12, 1.5, \frac{8}{255})$				
		Clean	l_∞	l_2	l_1	Union	Clean	l_∞	l_2	l_1	Union
Training from Scratch	E-AT	87.2	73.3	64.1	55.4	55.4	83.5	41.0	25.5	52.9	25.5
	MAX	85.6	72.1	63.6	56.4	56.4	74.6	42.9	35.7	50.3	35.6
	RAMP	86.3	73.3	64.9	59.1	**59.1**	74.4	43.4	37.2	51.1	**37.1**
Robust Fine-tuning	E-AT	86.5	74.8	66.7	57.9	57.9	80.2	42.8	31.5	52.4	31.5
	MAX	85.7	74.0	66.2	60.0	60.0	74.8	43.8	36.7	50.2	36.6
	RAMP	85.8	74.0	66.2	60.1	**60.1**	74.9	43.7	37.0	50.2	**36.9**

l_1 - l_2 Tradeoff **l_2 - l_∞ Tradeoff**

Experiment Result: Universal Robustness

RAMP shows best **universal robustness**

Models	Common Corruptions	l_0	fog	snow	gabor	elastic	jpeginf	Avg	Union
l_1-AT	78.2	79.0	41.4	22.9	40.5	48.9	48.4	46.9	12.8
l_2-AT	77.2	67.5	48.7	26.1	44.1	53.2	45.4	47.5	16.2
l_∞-AT	73.4	55.5	44.7	32.9	53.8	56.6	33.4	46.2	19.1
Winninghand [Diffenderfer et al., 2021]	**91.1**	74.1	74.5	18.3	76.5	12.6	0.0	42.7	0.0
E-AT	71.5	58.5	35.9	35.3	50.7	55.7	60.3	49.4	21.9
MAX	71.0	56.2	42.9	35.4	49.8	57.8	55.7	49.6	24.4
RAMP	75.5	55.5	40.5	40.2	52.9	60.3	56.1	**50.9**	**26.1**

Thank you!

Code: https://github.com/uiuc-focal-lab/RAMP

Contact information: enyij2@illinois.edu

Full paper