MANCHESTER 1824
The University of Manchester

Understanding and Improving Ensemble Adversarial Defense

Yian Deng
The University of Manchester
yian.deng@manchester.ac.uk

Tingting Mu
The University of Manchester
tingting.mu@manchester.ac.uk

Background & Motivation

- Traditional adversarial training & models
  - includes attacks & defenses
  - target: evaluate & improve robustness
  - limitation: trade-off between accuracy and robustness

- Ensemble adversarial defense
  - assumed to defend better adversarial attacks
  - rigorous understanding remains unclear

Error Theory for Ensemble Adversarial Defense

Assumption 4.2 (MLP Requirement). Suppose a C-class L-layer MLP $\mathbf{h}: \mathbb{R}^d \rightarrow [0,1]^C$ expressed iteratively by

$$\mathbf{a}^{(0)}(\mathbf{x}) = \mathbf{x}, \tag{6}$$

$$\mathbf{a}^{(l)}(\mathbf{x}) = \sigma\left(\mathbf{W}^{(l)}\mathbf{a}^{(l-1)}(\mathbf{x})\right), l = 1, 2, ..., L-1, \tag{7}$$

$$\mathbf{a}^{(L)}(\mathbf{x}) = \mathbf{W}^{(L)}\mathbf{a}^{(L-1)}(\mathbf{x}) = \mathbf{z}(\mathbf{x}), \tag{8}$$

$$\mathbf{h}(\mathbf{x}) = \text{softmax}(\mathbf{z}(\mathbf{x})), \tag{9}$$

where $\sigma(\cdot)$ is the activation function applied element-wise, the representation vector $\mathbf{z}(\mathbf{x}) \in \mathbb{R}^C$ returned by the L-th layer is fed into the prediction layer building upon the softmax function. Let $w^{(l)}_{s_{l+1},s_l}$ denote the network weight connecting the s_l-th neuron in the l-th layer and the s_{l+1}-th neuron in the $(l+1)$-th layer for $l \in \{1, 2 \ldots, L\}$. Define a column vector $\mathbf{p}^{(k)}$ with its i-th element computed from the neural network weights and activation derivatives, as $p_i^{(k)} = \sum_{s_L} \frac{\partial a_{s_L}^{(L-1)}(\mathbf{x})}{\partial x_k} w^{(L)}_{i,s_L}$ for $k = 1, 2, \ldots d$ and $i = 1, 2, \ldots C$, also a matrix $\mathbf{P_h} = \sum_{k=1}^{d} \mathbf{p}^{(k)}\mathbf{p}^{(k)^T}$ and its factorization $\mathbf{P_h} = \mathbf{M_h}\mathbf{M_h}^T$ with a full-rank factor matrix $\mathbf{M_h}$. For constants $\tilde{\lambda}, B > 0$, suppose the following holds for $\mathbf{h}$:

1. Its cross-entropy loss curvature measured by Eq. (2) satisfies $\lambda_{\mathbf{h}}(\mathbf{x}, \boldsymbol{\delta}) \leq \tilde{\lambda}$.
2. The factor matrix satisfies $\|\mathbf{M_h}\|_2 \leq B_0$ and $\left\|\mathbf{M_h}^\dagger\right\|_2 \leq B$, where $\|\cdot\|_2$ denotes the vector induced l_2-norm for matrix.

Error Theory for Ensemble Adversarial Defense

Definition 4.3 (Ambiguous Pair). Given a dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ where $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in [C]$, an *ambiguous pair* contains two examples $a = ((\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j))$ satisfying $y_i \neq y_j$ and

$$\|\mathbf{x}_i - \mathbf{x}_j\|_2 \leq \frac{1}{JB\sqrt{C\left(\tilde{\lambda}^2 - \xi\right)}}, \tag{10}$$

Assumption 4.4 (Acceptable Classifier). Suppose an acceptable classifier $\mathbf{f} : \mathbb{R}^d \rightarrow [0, 1]^C$ does not perform poorly on the ambiguous example set $G(D)$ associated with its ambiguous pair set $A(D)$ and control variable J. This means that, for any pair $(\mathbf{x}_i, \mathbf{x}_j, y_i, y_j) \in A(D)$, the following holds:

1. With a probability $p \geq 42.5\%$, the classifier can correctly classify one example from the pair by a sufficiently large predicted score and misclassify the other example by a sufficiently small score, e.g., $f_{y_i}(\mathbf{x}_i) \geq 0.5 + \frac{1}{J}$ and $f_{y_i}(\mathbf{x}_j) \leq 0.5 + \frac{1}{J}$.
2. For any example from the pair, e.g., $(\mathbf{x}_i, y_i)$, and it is classified to class $\hat{y}_i$, then it has small predicted scores for wrong classes, i.e., $f_c(\mathbf{x}_i) \leq \frac{1 - f_{\hat{y}_i}(\mathbf{x}_i)}{C-1}$ for $c \neq y_i, \hat{y}_i$.

Error Theory for Ensemble Adversarial Defense

Theorem 4.1. *Suppose* $\mathbf{h}, \mathbf{h}^0, \mathbf{h}^1 \in \mathcal{H} : \mathcal{X} \to [0,1]^C$ *are* C*-class* L*-layer MLPs satisfying Assumption 4.2. Given a dataset* $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$*, construct an ambiguous pair set* $A(D)$ *by Definition 4.3. Assume* $\mathbf{h}, \mathbf{h}^0, \mathbf{h}^1$ *are acceptable classifiers for* $A(D)$ *by Assumption 4.4. Given a classifier* $\mathbf{f} \in \mathcal{H} : \mathcal{X} \to \mathbb{R}^C$ *and a dataset* D*, assess its classification error by 0-1 loss, as*

$$\hat{\mathcal{R}}_{0/1}(D, \mathbf{f}) = \frac{1}{|D|} \sum_{\mathbf{x} \in D} 1 \left[f_{y_\mathbf{x}}(\mathbf{x}) < \max_{c \neq y_\mathbf{x}} f_c(\mathbf{x}) \right], \tag{4}$$

where $1[\text{true}] = 1$ *while* $1[\text{false}] = 0$*. For an ensemble* $\mathbf{h}_e^{(0,1)}$ *of two base MLPs* $\mathbf{h}^0$ *and* $\mathbf{h}^1$ *through either an average or a max combiner, i.e.,* $\mathbf{h}_e^{(0,1)} = \frac{1}{2}(\mathbf{h}^0 + \mathbf{h}^1)$ *or* $\mathbf{h}_e^{(0,1)} = \max(\mathbf{h}^0, \mathbf{h}^1)$*, it has a lower empirical 0-1 loss than a single MLP for classifying ambiguous examples, such as*

$$\mathbb{E}_{a \sim A(D)} \mathbb{E}_{\mathbf{h}^0, \mathbf{h}^1 \in \mathcal{H}} \left[\hat{\mathcal{R}}_{0/1} \left(a, \mathbf{h}_e^{(0,1)} \right) \right] < \mathbb{E}_{a \sim A(D)} \mathbb{E}_{\mathbf{h} \in \mathcal{H}} \left[\hat{\mathcal{R}}_{0/1} (a, \mathbf{h}) \right]. \tag{5}$$

iGAT: Improving Ensemble Mechanism

- Distributing Global Adversarial Examples

by probabilities: $$p_i = \frac{2^{N - r_{\mathbf{x}}(\mathbf{h}^i)}}{\sum_{i \in [N]} 2^{i-1}}$$

  – $r_x(\cdot)$: rank in descending order the predicted scores.

- Regularization Against Misclassification

$$L_R(\mathbf{x}, y_{\mathbf{x}}) = -\delta_{0/1}\left(\mathbf{c}\left(\mathbf{h}^1(\mathbf{x}), \ldots, \mathbf{h}^N(\mathbf{x})\right), y_{\mathbf{x}}\right) \log\left(1 - \max_{i=1}^{C} \max_{j=1}^{N} h_i^j(\mathbf{x})\right)$$

  – penalize the most incorrect prediction.

iGAT: Improving Ensemble Mechanism

- Final objective

$$\min_{\{\mathbf{h}^i\}_{i=1}^N} \underbrace{\mathbb{E}_{(\mathbf{x},y_{\mathbf{x}})\sim(\mathbf{X},\mathbf{y})}\left[L_E(\mathbf{x}, y_{\mathbf{x}})\right]}_{\text{original ensemble loss}} + \underbrace{\alpha \sum_{i=1}^{N} \mathbb{E}_{(\mathbf{x},y_{\mathbf{x}})\sim\left(\tilde{\mathbf{X}}^i,\tilde{\mathbf{y}}^i\right)}\left[\ell_{CE}(\mathbf{h}^i(\mathbf{x}), y_{\mathbf{x}})\right]}_{\text{added global adversarial loss}}$$

$$+ \underbrace{\beta \mathbb{E}_{(\mathbf{x},y_{\mathbf{x}})\sim(\mathbf{X},\mathbf{y})\cup\left(\tilde{\mathbf{X}},\tilde{\mathbf{y}}\right)}\left[L_R\left(\mathbf{x}, y_{\mathbf{x}}\right)\right]}_{\text{added misclassification regularization}},$$

Experiment results

Table 1: Comparison of classification accuracies in percentage reported on natural images and adversarial examples generated by different attack algorithms under L_∞-norm perturbation strength $\varepsilon = 8/255$. The results are averaged over five independent runs. The best performance is highlighted in bold, the 2nd best underlined.

		Average Combiner (%)					Max Combiner (%)				
		Natural	PGD	CW	SH	AA	Natural	PGD	CW	SH	AA
CIFAR10	TRS	83.15	12.32	10.32	39.21	9.10	82.67	11.89	10.78	37.12	7.66
	GAL	80.85	41.72	41.20	54.94	36.76	80.65	31.95	27.80	50.68	9.26
	SoE	82.19	38.54	37.59	**59.69**	32.68	82.36	32.51	23.88	41.04	18.37
	$iGAT_{SoE}$	81.05	40.58	39.65	57.91	34.50	81.19	31.98	24.01	40.67	19.65
	CLDL	84.15	45.32	41.81	55.90	37.04	83.69	39.34	32.80	51.63	15.30
	$iGAT_{CLDL}$	85.05	45.45	42.00	58.22	37.14	83.73	40.84	34.55	51.70	17.03
	DVERGE	85.12	41.39	43.40	57.33	39.20	84.89	41.13	39.70	**54.90**	35.15
	$iGAT_{DVERGE}$	**85.48**	42.53	44.50	57.77	39.48	**85.27**	**42.04**	**40.70**	54.79	**35.71**
	ADP	82.14	39.63	38.90	52.93	35.53	80.08	36.62	34.60	47.69	27.72
	$iGAT_{ADP}$	84.96	**46.27**	**44.90**	58.90	**40.36**	80.72	39.37	35.00	48.36	29.83
CIFAR100	TRS	58.18	10.32	10.12	15.78	6.32	57.21	9.98	9.23	14.21	4.34
	GAL	61.72	22.04	21.60	31.97	18.01	59.39	19.30	13.60	24.73	10.36
	CLDL	58.09	18.47	18.01	29.33	15.52	55.51	18.89	13.07	22.14	4.51
	$iGAT_{CLDL}$	59.63	18.78	18.20	29.49	14.36	56.91	20.76	14.09	20.43	5.20
	SoE	62.60	20.54	19.60	**36.35**	15.90	62.62	16.00	11.40	24.25	8.62
	$iGAT_{SoE}$	**63.19**	21.89	19.70	35.60	16.16	**63.02**	16.02	11.45	23.77	8.95
	ADP	60.46	20.97	20.55	30.26	17.37	56.20	17.86	13.70	21.40	10.03
	$iGAT_{ADP}$	60.17	22.23	20.75	30.46	17.88	56.29	17.89	14.10	21.47	10.09
	DVERGE	63.09	20.04	20.01	32.74	17.27	61.20	20.08	15.30	27.18	12.09
	$iGAT_{DVERGE}$	63.14	**23.20**	**22.50**	33.56	**18.59**	61.54	**20.38**	**17.80**	**27.88**	**13.89**

Table 3: Results of ablation studies based on $iGAT_{ADP}$ using CIFAR-10 under the PGD attack. The results are averaged over five independent runs. The best performance is highlighted in bold.

	Opposite Distributing	Random Distributing	Hard Distributing	$\beta = 0$	$iGAT_{ADP}$
Natural (%)	82.45	83.05	83.51	83.45	**84.96**
PGD (%)	41.31	42.60	44.21	42.32	**46.25**

Future Work

The University of Manchester

- Research model architectures beyond MLPs and the average/max combiners
- Large-scale datasets, e.g., ImageNet
- Generalize the theory to more than two base classifiers

Thanks!