Supplementary Materials for DENL: Diverse Ensemble and Noisy Logits for Improved Robustness of Neural Networks

Mina Yazdani MINA.YAZDANI@TORONTOMU.CA
Hamed Karimi HAMED.KARIMI@TORONTOMU.CA
Reza Samavi SAMAVI@TORONTOMU.CA
Department of Electrical, Computer, and Biomedical Engineering,
Toronto Metropolitan University, Toronto, ON, Canada
Vector Institute, Toronto, ON, Canada

Editors: Berrin Yanıkoğlu and Wray Buntine

Appendix A. Fine-tuning λ Parameter

The result of fine-tuning the λ parameter on SAE ensemble for CIFAER10 and MNIST, are reported in Table 1. As we can see, $\lambda = 0.4$ results in the ensemble with the highest accuracy on the adversarial examples generated by single attack scenario. The adversarial example classification and the average perturbations implemented on adversarial examples for CIFAR10 after Phase 2 are shown in Tables 6 and 7, respectively. The adversarial examples classification and the average perturbations implemented on adversarial examples for MNIST after Phase 2 are shown in Tables 8 and 9, respectively.

Appendix B. Adversarial Examples Classification and Perturbation Details

The adversarial examples classification results for CIFAR10 are shown in Tables 2(a) and 3(a), for adversarial examples of single attack and superimposition attack, respectively. Tables 4(a) and 5(a) show the adversarial examples classification on the MNIST for single and superimposition attack, respectively. The information of the tables are categorized in three sections of correctly classified (when the predicted class of these images remains unchanged after attack), classified to target (when the predicted label is the target label of the attack) and classified to others (when the predicted label is not correct and is not the target label of the attack). Tables 2(b) and 3(b) show the average perturbation on adversarial examples generated on CIFAR10 for single attack and superimposition attack, respectively. Tables 4(b) and 5(b) show the average perturbation on adversarial examples generated on MNIST for single attack and superimposition attack, respectively.

Table 1: Fine-tuning the λ parameter on SAE ensemble for CIFAER10 and MNIST

	Adversarial Accuracy	
λ	CIFAR10	MNIST
0.1	$94.85 \pm 0.59\%$	$89.40 \pm 0.21\%$
0.2	$95.12 \pm 0.35\%$	$89.83 \pm 0.45\%$
0.3	$95.76 \pm 0.21\%$	$90.11 \pm 0.72\%$
0.4	$96.01 \pm 0.27\%$	$90.39 \pm 0.81\%$
0.5	$95.34 \pm 0.32\%$	$90.05 \pm 0.23\%$
0.6	$95.36 \pm 0.69\%$	$89.66 \pm 0.35\%$
0.7	$94.48 \pm 0.37\%$	$89.34 \pm 0.29\%$

Table 2: Single attack statistics on CIFAR10 dataset over 5 folds

	(a) Adversarial Examples Classification (%)			**(b) Average Adversarial Perturbation (%)**		
	Correctly Classified	**Classified to Target**	**Classified to Other**	**Correctly Classified**	**Classified to Target**	**Classified to Other**
Individual Model	0.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	2.28 ± 0.12	0.00 ± 0.00
SAE	94.48 ± 0.05	1.26 ± 0.36	4.25 ± 0.41	3.13 ± 0.18	1.58 ± 0.06	3.44 ± 0.31
SAE + Noisy Logits	96.06 ± 0.24	0.44 ± 0.49	3.49 ± 0.26	3.01 ± 0.38	1.05 ± 0.12	3.24 ± 0.21
DAE	95.11 ± 0.23	1.15 ± 0.05	3.74 ± 0.44	2.90 ± 0.22	1.62 ± 0.03	3.24 ± 0.08
DAE + Noisy Logits	96.32 ± 0.26	0.24 ± 0.19	3.45 ± 0.17	2.74 ± 0.04	1.18 ± 0.27	3.16 ± 0.19

Appendix C. Experiments Recorded Time

We ran more than 450 groups of experiments to evaluate our method. We show the time required for running one sample experiment on MNIST for SAE and DAE ensemble in Table 10 which shows that the time needed for experiments in DAE ensemble is not noticeably higher than SAE in this experiment.

Table 3: Superimposition attack statistics on CIFAR10 dataset over 5 folds

	(a) Adversarial Examples Classification (%)			(b) Average Adversarial Perturbation (%)		
	Correctly Classified	**Classified to Target**	**Classified to Other**	**Correctly Classified**	**Classified to Target**	**Classified to Other**
SAE	91.81 ± 0.17	0.09 ± 0.04	8.09 ± 0.96	3.31 ± 0.06	1.92 ± 0.84	3.76 ± 0.21
SAE + Noisy Logits	94.65 ± 0.21	0.21 ± 0.11	5.14 ± 0.41	3.09 ± 0.08	0.12 ± 0.13	3.57 ± 0.40
DAE	95.34 ± 0.19	0.07 ± 0.04	4.58 ± 0.07	2.23 ± 0.02	1.45 ± 0.08	2.38 ± 0.11
DAE + Noisy Logits	96.47 ± 0.34	0.27 ± 0.05	3.26 ± 0.05	2.45 ± 0.03	0.09 ± 0.11	2.43 ± 0.03

Table 4: Single attack statistics on MNIST dataset over 5 folds

	(a) Adversarial Examples Classification (%)			(b) Average Adversarial Perturbation (%)		
	Correctly Classified	**Classified to Target**	**Classified to Other**	**Correctly Classified**	**Classified to Target**	**Classified to Other**
Individual Model	0.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	14.46 ± 1.05	0.00 ± 0.00
SAE	89.40 ± 0.34	3.76 ± 0.06	6.83 ± 0.29	16.26 ± 0.35	14.80 ± 0.32	19.62 ± 0.34
SAE + noisy logits	80.59 ± 0.12	5.11 ± 0.22	14.30 ± 0.81	14.33 ± 0.11	15.90 ± 0.57	18.70 ± 0.30
DAE	90.98 ± 0.07	2.88 ± 0.14	6.14 ± 0.32	16.17 ± 0.36	14.80 ± 0.58	19.69 ± 0.33
DAE + noisy logits	81.69 ± 0.05	4.89 ± 0.03	13.41 ± 0.79	14.17 ± 0.24	16.25 ± 0.54	18.72 ± 0.41

Table 5: Superimposition attack statistics on MNIST dataset over 5 folds

	(a) Adversarial Examples Classification (%)			(b) Average Adversarial Perturbation (%)		
	Correctly Classified	**Classified to Target**	**Classified to Other**	**Correctly Classified**	**Classified to Target**	**Classified to Other**
SAE	31.82 ± 0.14	45.76 ± 0.88	22.41 ± 0.29	16.28 ± 0.35	24.52 ± 0.58	27.67 ± 0.28
SAE + noisy logits	82.07 ± 0.02	2.68 ± 0.71	15.24 ± 0.22	12.38 ± 0.47	16.19 ± 0.73	18.36 ± 0.37
DAE	33.69 ± 1.18	43.59 ± 1.02	22.73 ± 0.48	16.45 ± 0.41	24.74 ± 0.80	27.83 ± 0.35
DAE + noisy logits	84.02 ± 0.96	1.86 ± 0.52	14.13 ± 0.65	12.08 ± 0.36	18.11 ± 1.06	18.85 ± 0.49

Table 6: Single attack statistics on CIFAR10 dataset

	(a) Adversarial Examples Classification (%)			(b) Average Adversarial Perturbation (%)		
	Correctly Classified	**Classified to Target**	**Classified to Other**	**Correctly Classified**	**Classified to Target**	**Classified to Other**
SAE	96.01 ± 0.27	1.11 ± 0.07	2.88 ± 0.02	3.14 ± 0.15	1.58 ± 0.06	3.45 ± 0.03
SAE + noisy logits	97.21 ± 0.13	0.51 ± 0.14	2.29 ± 0.19	2.68 ± 0.32	1.20 ± 0.27	2.62 ± 0.31
DAE	96.82 ± 0.28	1.01 ± 0.04	2.17 ± 0.07	2.87 ± 0.26	1.53 ± 0.18	2.78 ± 0.13
DAE + noisy logits	98.29 ± 0.61	0.41 ± 0.03	1.30 ± 0.16	2.41 ± 0.27	1.02 ± 0.26	2.26 ± 0.24

Table 7: Superimposition attack statistics on CIFAR10 dataset

	(a) Adversarial Examples Classification(%)			(b) Average Adversarial Perturbation(%)		
	Correctly Classified	**Classified to Target**	**Classified to Other**	**Correctly Classified**	**Classified to Target**	**Classified to Other**
SAE	92.97 ± 0.21	0.62 ± 0.08	6.41 ± 0.06	2.25 ± 0.14	0.88 ± 0.06	2.24 ± 0.28
SAE + noisy logits	96.11 ± 0.05	0.30 ± 0.03	3.59 ± 0.04	2.09 ± 0.17	0.19 ± 0.04	1.94 ± 0.39
DAE	96.90 ± 0.74	0.04 ± 0.02	3.06 ± 0.08	2.59 ± 0.24	1.46 ± 0.12	2.61 ± 0.48
DAE + noisy logits	97.46 ± 0.41	0.17 ± 0.02	2.37 ± 0.15	2.41 ± 0.30	0.20 ± 0.02	0.21 ± 0.09

Table 8: Single attack statistics on MNIST dataset

	(a) Adversarial Examples Classification(%)			(b) Average Adversarial Perturbation(%)		
	Correctly Classified	**Classified to Target**	**Classified to Other**	**Correctly Classified**	**Classified to Target**	**Classified to Other**
SAE	90.39 ± 0.81	1.70 ± 0.06	7.92 ± 0.02	16.97 ± 0.41	22.70 ± 0.33	27.14 ± 0.26
SAE + noisy logits	82.61 ± 0.95	1.40 ± 0.04	15.98 ± 0.76	11.70 ± 0.46	11.62 ± 0.27	16.33 ± 0.13
DAE	91.50 ± 0.87	1.59 ± 0.08	6.90 ± 0.04	15.87 ± 0.55	13.63 ± 0.51	20.90 ± 0.33
DAE + noisy logits	83.02 ± 0.94	1.01 ± 0.03	15.97 ± 0.73	11.92 ± 0.38	11.40 ± 0.35	16.55 ± 0.61

Table 9: Superimposition attack statistics on MNIST dataset

	(a) Adversarial Examples Classification(%)			**(b) Average Adversarial Perturbation(%)**		
	Correctly Classified	**Classified to Target**	**Classified to Other**	**Correctly Classified**	**Classified to Target**	**Classified to Other**
SAE	32.55 ± 0.45	33.80 ± 0.59	33.64 ± 0.45	15.71 ± 0.78	13.37 ± 0.51	27.46 ± 0.44
SAE + noisy logits	83.94 ± 0.17	1.43 ± 0.07	14.64 ± 0.48	11.07 ± 0.53	14.45 ± 0.08	17.57 ± 0.92
DAE	34.10 ± 0.91	29.92 ± 0.82	35.98 ± 0.54	17.42 ± 0.50	21.18 ± 0.71	25.77 ± 1.01
DAE + noisy logits	86.65 ± 1.61	1.24 ± 0.14	12.10 ± 0.39	11.99 ± 0.35	10.82 ± 0.82	16.55 ± 0.77

Table 10: Experiment run time on MNIST for SAE and DAE ensemble

	SAE	DAE
Phase 1 Training	8924.28 Sec	9120.11 Sec
Phase 2 Training	7673.36 Sec	6980.38 Sec
Generating Adversarial Examples (Single Attack)	7751.63 Sec	8929.26 Sec
Generating Adversarial Examples (Superimposition Attack)	8070.10 Sec	7035.88 Sec