Supplementary Material for Multi-way Encoding for Robustness

Donghyun Kim
Boston University
donhk@bu.edu

Sarah Adel Bargal
Boston University
sbargal@bu.edu

Jianming Zhang
Adobe Research
jianmzha@adobe.com

Stan Sclaroff
Boston University
sclaroff@bu.edu

A. Ablation Study on Encodings

We perform ablation studies to further investigate the effectiveness of our RO encoding. We train the model used in Table 2 in the original manuscript with two different combinations of encodings and loss functions. Please note that the two alternative models have 10 dimensions at the last layer while RO has 2000 dimensions.

A.1. Alternative approach

A.1.1 $RO_{softmax}$

We evaluate a network that uses RO encoding, a softmax layer, and cross-entropy loss. We compute the probability of i^{th} class as follows:

$$P(i|s) = \frac{\exp(\mathbf{s}^\top \mathbf{e_i})}{\sum_{j=1}^{n} \exp(\mathbf{s}^\top \mathbf{e_j})}$$

where $\mathbf{s}$ is the ℓ_2 normalized final layer representation, $\mathbf{e_i}$ is the RO encoding vector (ground-truth vector) from the codebook, and n is the number of classes.

A.1.2 $1ofK_{MSE}$

We also evaluate a network that uses mean-squared error (MSE) loss with the $1ofK$ encoding.

A.2. Evaluation

We generate FGSM attacks with $\epsilon = 0.2$ from substitute models A_{1ofK} and C_{1ofK} on MNIST to evaluate the models of Section A.1.1 and Section A.1.2. We also measure a correlation coefficient of the sign of the input gradients between target and substitute models as explained in Section 4.1.1. Tables A and B demonstrate that RO, among the different target models, achieves the highest accuracy and the lowest input gradient correlation with the substitute model. It should be noted that the two alternative models have 10 neurons at the last layer while RO has 2000 neurons. In addition, $RO_{softmax}$ has a softmax layer so that the gradients at the final layer are determined by a ground-truth class of an example.

B. Transferability

In Table 3 of the main paper, the black-box attacks of the second column report the robustness on black-box attacks from the independently trained copy of the RO model. In this section, we analyze the black-box attack accuracy on CIFAR-10 by varying confidence κ of Eq. 5 in the main paper. The higher confidence makes an attack to be more confident misclassification. We observe that the black-box attack accuracy converges at confidence$= 1500$. We report the lowest accuracy in Table 3.

C. Checking for Signs of Obfuscated Gradients

In order to check if our method relies on obfuscated gradients [1], we report the accuracies on white-box attacks by varying epsilon on CIFAR-10 in Table D. The maximum allowed perturbation for our model is 8/255, but we use larger epsilon to check the behavior of our model. We checked that increasing distortion bound monotonically increase attack success rates and unbounded attacks achieve 100% attack success rate.

Acknowledgments

We thank Kate Saenko, Vitaly Ablavsky, Adrian Vladu, Seong Joon Oh, Tae-Hyun Oh, and Bryan A. Plummer for helpful discussions. This work was supported in part by gifts from Adobe.

References

[1] A. Athalye, N. Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, 2018.

Table A. This table presents black-box attacks from the substitute model A_{1ofK} on various target models. RO achieves the highest accuracy and the lowest input gradient correlation with the substitute model among the different target models.

Target Model	A			C		
	$RO_{softmax}$	$1ofK_{MSE}$	RO	$RO_{softmax}$	$1ofK_{MSE}$	RO
Accuracy (%)	48.7	43.4	88.7	53.7	42.1	94.3
Correlation Coefficient	0.14	0.15	0.02	0.1	0.13	0.03

Table B. This table presents black-box attacks from the substitute model C_{1ofK} on various target models. RO achieves the highest accuracy and the lowest input gradient correlation with the substitute model among the different target models.

Target Model	A			C		
	$RO_{softmax}$	$1ofK_{MSE}$	RO	$RO_{softmax}$	$1ofK_{MSE}$	RO
Accuracy (%)	67.4	55.9	92.5	62.6	58.8	96.1
Correlation Coefficient	0.08	0.09	0.02	0.08	0.1	0.01

Table C. This table presents accuracies on black-box attacks from RO by varying confidence (κ). We generate 1000-step PGD attacks on CIFAR-10.

confidence	10	300	1500	3000	6000
Accuracy (%)	83.0	80.9	72.2	72.2	72.2

Table D. This table presents accuracies on white-box attacks by varying epsilon (ℓ_∞). Maximum allowed perturbation for our model is 8/255, but we use larger epsilon to check the behavior of our model.

epsilon	2	4	6	10	12	14	18	20	Unbounded
Accuracy (%)	78.9	66.7	55.8	53.0	50.4	48.1	29.7	27	0