Supplementary Material:
REGroup: Rank-aggregating Ensemble of Generative Classifiers for Robust Predictions

Lokender Tiwari[1,2] Anish Madan[1] Saket Anand[1] Subhashis Banerjee[3,4]

[1]IIIT-Delhi [2]TCS Research [3]IIT Delhi [4]Department of Computer Science, Ashoka University

https://lokender.github.io/REGroup.html

1. Hyper-parameters for Generating Adversarial Examples

We use Foolbox's [3] implementation of almost all the adversarial attacks(except SPSA[1], Trust Region[2] and cAdv[3]) used in this work. We report the attack specific hyper-parameters in Tab.2.

2. Elastic-Net Attacks

We evaluate REGroup on Elastic-Net attacks [1]. Elastic-Net attack formulate the attack process as a elastic-net regularized optimization problem. The results are shown in the table 1.

	ResNet-50			VGG-19		
		SMax	REGroup		SMax	REGroup
Attacks	#S	T1(%)	T1(%)	#S	T1(%)	T1(%)
EAD-Attack	2000	0	52	2000	0	49

Table 1. *Performance on EAD attacks.* Top-1 (%) classification accuracy comparison between SoftMax (SMax) and REGroup. $\#S$ is the number of images for which the attacker is successfully able to generate adversarial examples and the accuracies are reported with respect to the $\#S$ samples, hence the 0% accuracies with the SoftMax (SMax).

3. Accuracy vs no. of layer/voters(ResNet50)

We report the performance of REGroup on various attacks reported in table 2 of the main paper for all possible values of k. The accuracy of ResNet-50 w.r.t. the various values of k is plotted in figure 1.

4. Analyzing Pre-Activation Responses

One of the contributions of our proposed approach is to use both positive and negative pre-activation values separately. We observed both positive and negative pre-activation values contain information that can help correctly classify adversarially perturbed samples. An empirical validation of our statement is shown in figure 3 of the main paper. We further show using TSNE [5] plots that all the three variants of the pre-activation feature of a single layer i.e positive only (pos), negative only (neg) and combined positive and negative pre-activation values forms clusters. This indicates that all three contain equivalent information for discriminating samples from others. While on one hand where ReLU like activation functions discard the negative pre-activation responses, we consider negative responses equivalently important and leverage them to model the layerwise behaviour of class samples. The benefit of using positive and negative accumulators is it reduce the computational cost significantly e.g flattening a convolution layer gives a very high-dimensional vector while accumulator reduce it to number of filter dimensions.

[1] https://github.com/tensorflow/cleverhans

[2] https://github.com/amirgholami/TRAttack

[3] https://github.com/AI-secure/Big-but-Invisible-Adversarial-Attack

References

[1] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Ead: elastic-net attacks to deep neural networks via adversarial examples. In *AAAI*, 2018.

[2] Elvis Dohmatob. Limitations of adversarial robustness: strong no free lunch theorem. *arXiv preprint arXiv:1810.04065*, 2018.

[3] Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. *arXiv preprint arXiv:1707.04131*, 2017.

[4] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

[5] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Attack	Hyper-parameters
PGD (Untargeted)	$\epsilon = 4$, Dist:L_∞, random_start=True, stepsize=0.01, max_iter=40
DeepFool (Untargeted)	$\epsilon = 2$, Dist:L_2, max_iter=100, subsample=10 (Limit on the number of the most likely classes)
CW (Untargeted)	$\epsilon = 4$, Dist:L_2, binary_search_steps=5, max_iter=1000, confidence=0, learning_rate=0.005, initial_const=0.01
Trust Region (Untargeted)	$\epsilon = 2$, Dist:L_∞, iterations=5000
Boundary (Untargeted)	$\epsilon = 2$, Dist:L_2, iterations=500, max_directions=25, starting_point=None, initialization_attack=None, log_every_n_steps=None, spherical_step=0.01, source_step=0.01, step_adaptation=1.5, batch_size=1, tune_batch_size=True, threaded_rnd=True, threaded_gen=True
Spatial (Untargeted)	$\epsilon = 2$, Dist=L_2, do_rotations=True, do_translations=True, x_shift_limits=(-5, 5), y_shift_limits=(-5, 5), angular_limits=(-5, 5), granularity=10, random_sampling=False, abort_early=True
PGD (Targeted)	Dist = L_∞, binary_search=True, epsilon=0.3, stepsize=0.01, iterations=40, random_start=True, return_early=True
CW (Targeted)	binary_search_steps=5, max_iterations=1000, confidence=0, learning_rate=0.005, initial_const=0.01, abort_early=True
SPSA	$\epsilon = (4, 8)$, Dist:L_∞, max_iter=300, batch_size=64, early_stop_loss_thresh = 0, perturbation_size $\delta = 0.01$, Adam LR=0.01
EAD	Dist=L_2, binary_search_steps=5, max_iterations=1000, confidence=0, initial_learning_rate=0.01, regularization=0.01, initial_const=0.01, abort_early=True
PGD (Untargeted,HC)	min_conf=0.9, Dist=L_∞, binary_search=True, epsilon=0.3, stepsize=0.01, iterations=40, random_start=True, return_early=True
PGD (Targeted,HC)	min_conf=0.9, Dist=L_∞, binary_search=True, epsilon=0.3, stepsize=0.01, iterations=40, random_start=True, return_early=True

Table 2. Attack Specific Hyper-parameters.

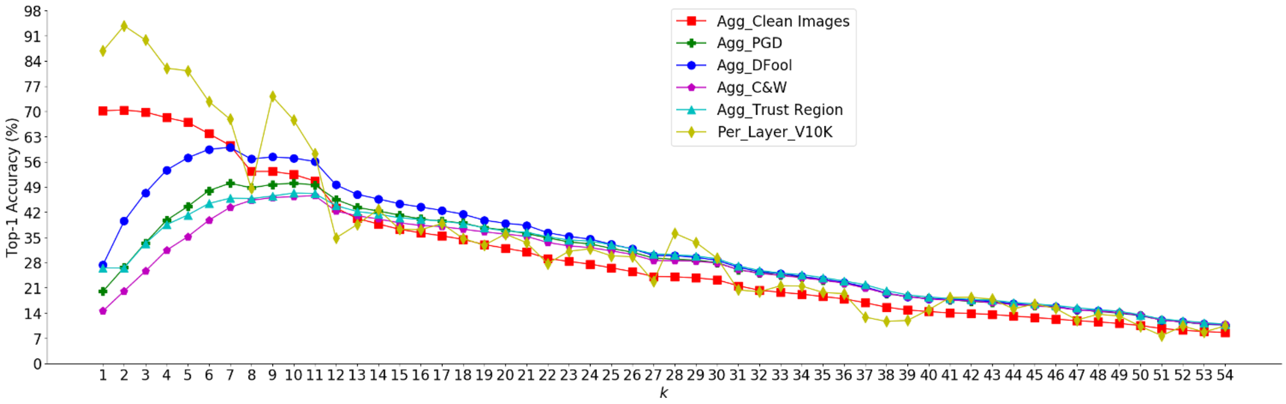
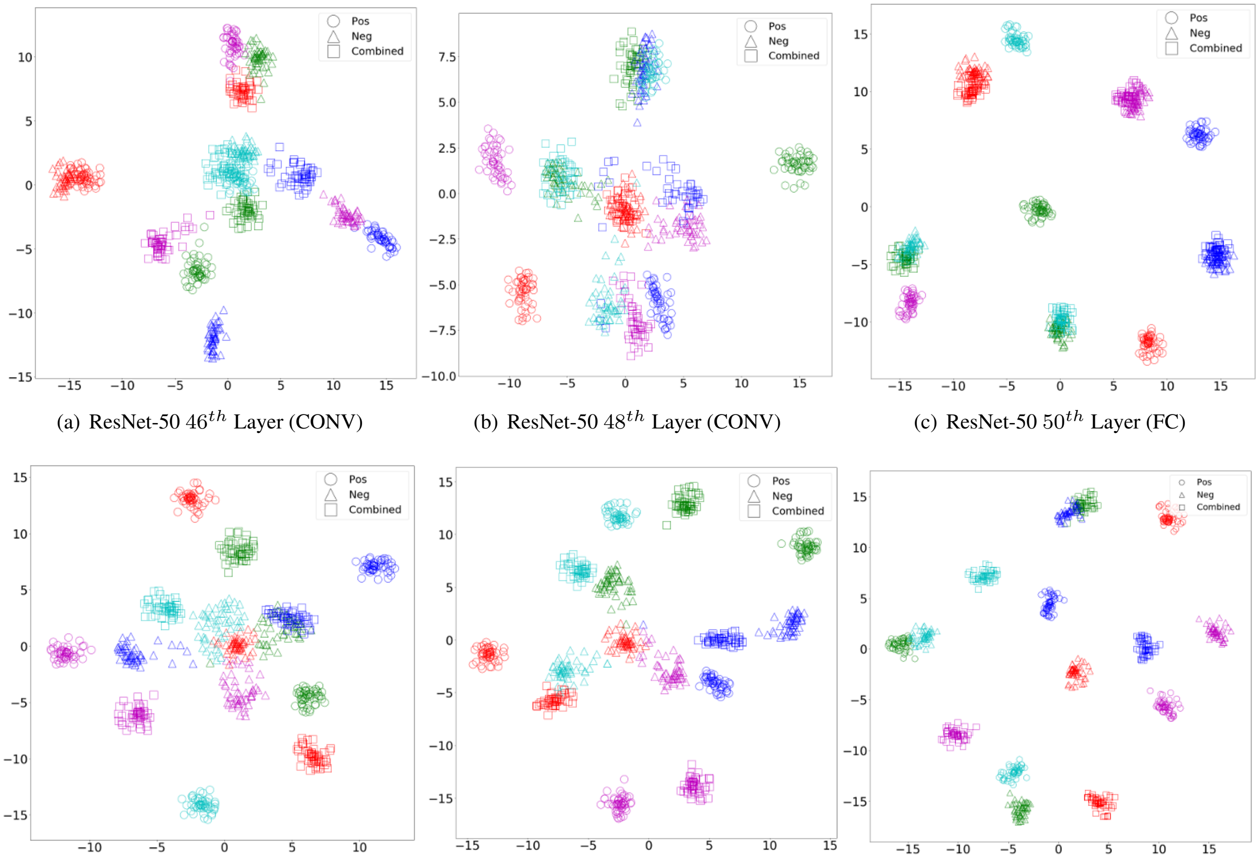


Figure 1. Ablation study for accuracy vs no. of layers (k) on ResNet-50: 'Agg' stands for using aggregated Borda count $B^{:ky}$. PGD, DFool, C&W and Trust Region are the same experiments as reported in table 2 of the main paper, but with all possible values of k. "Per_Layer_V10K" stands for evaluation using per layer Borda count i.e $\widehat{y} = argmax_y \; B^{\ell y}$ on a separate 10,000 correctly classified subset of validation set. In all our experiments we choose the k-highest layers where 'Per_Layer_V10K' has at-least 75% accuracy. A reasonable change in this accuracy criteria of 75% would not affect the results on adversarial attacks significantly. However, a substantial change (to say 50%) deteriorates the performance on clean sample significantly. The phenomenon of decrease in accuracy of clean samples vs robustness has been studied in [2] and [4]. **Note:** There are four down-sampling layers in the ResNet-50 architecture, hence the total 54 layers.



(a) ResNet-50 46^{th} Layer (CONV)

(b) ResNet-50 48^{th} Layer (CONV)

(c) ResNet-50 50^{th} Layer (FC)

(d) VGG-19 17^{th} Layer (FC)

(e) VGG-19 18^{th} Layer (FC)

(f) VGG-19 19^{th} Layer (FC)

Figure 2. TSNE visualization of three variants of pre-activation features i.e positive only (pos), negative only (neg) and combined positive and negative (combined). Visualization of 50 samples of 5 random classes of ImageNet dataset. Class membership is color coded. The dimensions of the pos, neg and combined variants of pre-activation feature is the same for any fully connected layer, while for a CONV layer, pos and neg has the same dimension which is equal to the no. of filters/feature maps of the respective CONV layer and for combined it is equal to the dimension we get after flattening the whole CONV layer. It can be observed in figure(b) that the cluster formed by combined pre-activation feature responses is not a tight as formed by pos and neg separately, which shows the importance of considering pos and neg re-activation responses separately.