SANE: Exploring Adversarial Robustness With Stochastically Activated Network Ensembles

Ibrahim Ben Daya*, Mohammad Javad Shafiee*, Michelle Karg†,
Christian Scharfenberger†, Alexander Wong*
*University of Waterloo
Waterloo, Canada
†Continental Automotive
Germany

Abstract

A major challenge to the adoption of deep neural networks in real-world applications is their robustness in different scenarios. Deep neural networks have been shown to be particularly susceptible to adversarial attacks: malicious perturbations to the input that fool networks into predicting the wrong label. In this study, we propose a new framework to improve adversarial robustness using stochastically activated network ensembles (SANE), where an ensemble of deep neural networks with heterogeneous architectures is stochastically activated such that a subset of the more robust networks in the ensemble are responsible for a prediction. The proposed framework treats networks as nodes in a probabilistic graphical model to detect networks in the ensemble that are likely to be robust against an adversarial attack and activate them to be part of the decision making process. Experimental results under different adversarial attacks show that the proposed SANE cannot only noticeably improve robustness to adversarial attacks compared to a general ensemble approach, but provide further improvements against adversarial attacks when combined with additional stochastic defense mechanisms.

1. Introduction

Deep learning has been responsible for a number of significant breakthroughs in the field of machine learning and computer vision, demonstrating remarkable performance in a wide variety of visual perception tasks [2]. Despite these incredible advances, recent literature has demonstrated that deep neural networks are very vulnerable to adversarial attacks [8], malicious perturbations designed to fool networks into making erroneous decisions. Such adversarial attacks can often be so subtle that it is imperceptible to the human eye, with an extreme case requiring only one pixel to change[7]. Attacks do not require direct access to the model, the property of *transferability* can be leveraged where an attack generated using a network is used to attack another that the attacker has no access to. This raises concerns on their robustness in safety-critical scenarios such as autonomous driving, security, and surveillance applications, encouraging a steadily growing body of literature on adversarial defence [2].

Ensemble techniques [1, 6] have been recently explored as a defense mechanism based on the variability of attack transferability; it is intuitively more difficult to fool multiple heterogeneous networks with a unique perturbation on an input image. Such techniques have shown promising results in improved robustness while also improving accuracy on unperturbed data. Although ensemble techniques can be very helpful as a defense mechanism, they are usually aggregated in a Bagging approach with equal weighting in the decision step, which –while useful in reducing the variance of the complete system in the decision-making process– is susceptible when highly vulnerable networks within the ensemble result in reduced robustness to the attack in the decision-making process.

Inspired by the promise of ensemble techniques for adversarial defense while motivated to address the critical issue associated with the negative influence of vulnerable networks within an ensemble, we propose a novel probabilistic graphical model approach which aggregates the decisions of the networks in the ensemble via a probabilistic approach to increase adversarial robustness, reduce system bias, and reduce variance. Given that the prediction is made by a subset of robust networks that are stochastically activated within the ensemble (i.e., based on predictions of vulnerability made by the probabilistic graphical model) we will refer to the proposed defense mechanism as stochastically activated network ensembles (SANE). It is worth noting that, while the computational complexity of utilizing an ensemble of networks to defend against the adversarial attack is a practical challenge, the main focus of this research is to investigate the feasibility and effectiveness of utilizing a probabilistic graphical model to improve the robustness of ensemble techniques in defense mechanism.

2. Methodology

A particular technique that has been shown to be robust at dealing with noisy data is ensemble learning. Here we

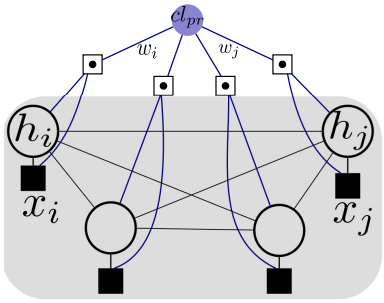

Figure 1. The proposed probabilistic graphical model in the SANE framework.

explore the notion of ensemble learning and committee-based decision making for constructing network ensembles as a defense mechanism for adversarial attacks.

An issue that has not been well-explored when leveraging network ensembles as a means for adversarial defense is the fact that individual networks within the ensemble remain susceptible to adversarial attacks. If the majority of networks are successfully attacked, then the network ensemble is compromised as a whole, leading to an incorrect prediction and it is a limitation of ensemble methods.

The proposed SANE framework introduces a probabilistic graphical model to first estimate the robustness of each network in the ensemble at predicting the correct label given the beliefs of the other networks. This measure of robustness is then leveraged to stochastically activate a subset of the networks included in the final decision-making process. To achieve this, each network in the ensemble is represented as a node in a fully connected graph; with the connections between nodes representing the relationship of networks to each other. The state in the probabilistic graphical model is formulated as a binary random variable encoding the reliability of a particular network in the ensemble for participating in the decision-making process.

The status of each network n_i (being attacked or not) in the ensemble $\mathcal{C} = \{n_1, n_2, \cdots, n_{|\mathcal{C}|}\}$ is encoded by h_i in the graph $\mathcal{G}(\cdot)$. Each node h_i in the graph $\mathcal{G}(\cdot)$ is associated with an observation set $\bar{x}_i$ representing the set of outputs from the Softmax layer in the network. By formulating the ensemble as a fully connected probabilistic graphical model, each network n_i in the ensemble $\mathcal{C}$ is judged by all other networks $n_j, j \neq i$ such that the marginalized conditional probability $\sum_{h_j, j\neq i} P(H|X)$ illustrates how reliable the network n_i is when contributing to the decision-making process based on the beliefs of other networks in the graph. The conditional probability $P(H|X)$ is formulated as a pairwise undirected graphical model:

$$P(H|X) = \frac{1}{Z}\prod_{i=1}^{|\mathcal{C}|}\phi_i(h_i, \bar{x}_i)\prod_{e=1}^{|\mathcal{E}|}\phi_e(h_{ej}, h_{ek}, \bar{x}_{ej}, \bar{x}_{ek}) \tag{1}$$

where $\phi_i(h_i, \bar{x}_i)$ is the unary potential encoding the robustness of network n_i based on prior knowledge. $\phi_e(\cdot)$ is a pairwise potential demonstrating the belief of two end-node networks n_j and n_k of edge $e = \{j, k\}$ on each other. $\mathcal{E}$ is the set of all edges in the graph where $|\mathcal{E}| = \frac{|\mathcal{C}|\times(|\mathcal{C}|-1)}{2}$ since the underlying graph is a fully connected.

The unary and pairwise potential functions are formulated as follow:

Unary Potential:

$$\phi_i(h_i, x_i) = \begin{cases} r_i & n_i \text{ can be fooled} \\ 1 - r_i & \text{otherwise} \end{cases} \tag{2}$$

where r_i is the transfer attack success rate[1] to the network n_i via the rest of networks in the ensemble.

Pairwise Potential:

$$\phi_e(h_{ej}, h_{ek}, \bar{x}_{ej}, \bar{x}_{ek}) = \begin{cases} \frac{l_{j,k}}{|\mathcal{I}|} \cdot \|\bar{x}_{ej} - \bar{x}_{ek}\| & h_j \neq h_k \\ (1 - \frac{l_{j,k}}{|\mathcal{I}|}) & h_j = h_k \end{cases} \tag{3}$$

where $\bar{x}_{ej}$ is the observation set for the network n_j which is the vector of confidence values corresponding to all class labels. $\mathcal{I}$ is the set of perturbed images which is used for the training purposes and computing the similarity of the networks in the training stage. The $\frac{l_{j,k}}{|\mathcal{I}|}$ is computed at the training stage and is fixed for the network during testing. $l_{j,k}$ is the Levenshtein distance [5] on the prediction on two network's outputs j and k illustrating the behavior of the two networks when they are dealing with adversarial examples and perturbed images.

Adversarial Defense Framework. In the previous section, we explained how to model the ensemble of the networks as a probabilistic graphical model to determine how likely a network within an ensemble is fooled by an adversarial attack. We can leverage this within an adversarial defense framework by using the marginal probability through the probabilistic graphical model to activate the subset of networks which are reliable for making the final prediction together using a weighted voting mechanism.

The weighted voting approach is formulated as follows:

$$cl_{pr} = \arg\max_j (\sum_{i\in\hat{\mathcal{C}}} w_i \cdot \bar{x}_i) \tag{4}$$

where cl_{pr} is the predicted class by the committee $\hat{\mathcal{C}}$, $\hat{\mathcal{C}}$ is the set of reliable networks activated from the set of all networks in the ensemble $\mathcal{C}$ based on the probabilistic model. $\bar{x}_i$ shows the set of outputs after Softmax for the network n_i. The w_i is formulated as

$$w_i = \frac{1}{r_i} \tag{5}$$

[1]The transfer attack success rate r_i is defined as the ratio of the number of adversarial examples (generated by other networks) that can fool network n_i over the total number of adversarial examples.

Table 1. Accuracy of the proposed SANE framework compared to other defense mechanism based on CIFAR-10 and ImageNet trained models. For CIFAR-10 trained models, the proposed SANE framework outperforms both EnsembleDef and RandDef, and achieves comparable results to when both EnsembleDef and RandDef are combined together. Furthermore, the combination of RandDef and SANE outperforms all other tested methods. For ImageNet trained models, results show that not only does SANE outperform both RandDef and EnsembleDef, but the combination of RandDef and SANE can provide the best performance against the targeted perturbations.

Dataset	ϵ	Single-Best Network	RandDef [9]	EnsembleDef [6]	RandDef + EnsembleDef	SANE	RandDef + SANE
CIFAR-10	2.0	74.2%	52.4%	99.3%	74.0%	99.3%	72.0%
	5.0	66.0%	49.6%	93.6%	75.0%	96.2%	79.0%
	10	62.0%	46.0%	70.7%	64.0%	78.2%	57.0%
	20	53.7%	41.5%	43.7%	68.0%	50.3%	61.0%
ImageNet	2.0	70.4%	90.0%	99.5%	100.0%	99.6%	99.8%
	5.0	53.4%	70.9%	96.8%	98.2%	97.1%	98.4%
	10	43.3%	62.3%	89.4%	92.6%	91.3%	92.5%
	20	39.7%	55.3%	79.2%	83.3%	82.9%	85.3%

where r_i is the transfer attack success rate for network n_i. This weighting approach gives more weights to the networks with lower transfer attack rates as they are more robust compared to others in the ensemble.

3. Results and Discussion

For evaluation purposes, two test set of 1000 images (From CIFAR-10 dataset [3] and NIPS adversarial attack challenge dataset [4], where ImageNet trained models are used) are randomly selected from the set of all images correctly classified by all networks in the ensemble, which is consistent with evaluation methodologies in existing literature:

- **EnsembleDef [6]:** This technique uses a network ensemble for improving adversarial robustness.
- **RandDef [9]**: This technique involves randomly resizing and padding the input before being fed into the network to improve adversarial robustness.

We analyze the robustness of the proposed SANE framework against FGSM attack under 4 different noise levels. Table 1 shows the experimental results for both CIFAR-10 and NIPS adversarial attack challenge dataset.

For CIFAR-10 dataset, the proposed SANE framework outperforms both RandDef and the best performing network in the ensemble across all noise levels. While SANE provides similar performance as EnsembleDef for $\epsilon = 2$, it outperforms EnsembleDef for all noise levels above that. This is most illustrative by the reported result for $\epsilon = 10$ and $\epsilon = 20$ where SANE can achieve 8% and 7% higher accuracy, respectively, when compared to EnsembleDef.

The proposed SANE framework is also compared to the combination of RandDef and EnsembleDef (i.e., RandDef+EnsembleDef) as well. Furthermore, we also experimented with the combination of RandDef and SANE (i.e., RandDef+SANE). Results demonstrate that RandDef could not improve robustness when used in conjunction with EnsembleDef or SANE in this case. The poor performance of RandDef can be justified by the fact that since CIFAR-10 images are small (32×32), randomly resizing and padding them reduces the amount of information in the image and thus causes a drop in modeling accuracy.

For NIPS adversarial attack challenge dataset, SANE achieved similar accuracy as EnsembleDef at $\epsilon = 2$, but outperforms EnsembleDef significantly at higher noise levels. RandDef performs noticeably better than the performance of the single-best network, which illustrates its effectiveness for improving robustness in the situation where the image size is sufficiently large. Finally, it is observed that the combination of RandDef with SANE (i.e., RandDef+SANE) provides additional robustness over SANE, especially at the highest noise levels, leading RandDef+SANE to provide the highest adversarial robustness out of all tested methods.

4. Conclusion

In this study, we proposed SANE, a new probabilistic approach to improve the robustness of network ensembles. Using a fully-connected probabilistic graphical model, a subset of reliable networks in the ensemble is determined and stochastically activated for the final prediction process. Experimental results using CIFAR-10 and NIPS adversarial attack challenge dataset demonstrated the effectiveness of the proposed SANE framework at improving robustness in prediction adversarially attacked images when compared to other state-of-the-art frameworks. In addition, we showed that it is possible to combine SANE with other stochastic mechanisms to further improve robustness. Future work will focus on a more efficient way to leverage the proposed SANE framework for practical applications where computational constraints is limited.

References

[1] M. Abbasi and C. Gagné. Robustness to adversarial examples through an ensemble of specialists. *CoRR*, abs/1702.06856, 2017. 1

[2] N. Akhtar and A. Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *CoRR*, abs/1801.00553, 2018. 1

[3] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 3

[4] A. Kurakin, I. J. Goodfellow, S. Bengio, Y. Dong, F. Liao, M. Liang, T. Pang, J. Zhu, X. Hu, C. Xie, J. Wang, Z. Zhang, Z. Ren, A. L. Yuille, S. Huang, Y. Zhao, Y. Zhao, Z. Han, J. Long, Y. Berdibekov, T. Akiba, S. Tokui, and M. Abe. Adversarial attacks and defences competition. *CoRR*, abs/1804.00097, 2018. 3

[5] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966. 2

[6] T. Strauss, M. Hanselmann, A. Junginger, and H. Ulmer. Ensemble methods as a defense to adversarial perturbations against deep neural networks. *arXiv preprint arXiv:1709.03423*, 2017. 1, 3

[7] J. Su, D. V. Vargas, and K. Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 2019. 1

[8] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013. 1

[9] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. L. Yuille. Mitigating adversarial effects through randomization. *CoRR*, abs/1711.01991, 2017. 3