Dynamic Stochastic Ensemble with Adversarial Robust Lottery Ticket Subnetworks

Qi Peng,[1] **Wenlin Liu,**[2] **Ruoxi Qin,**[1] **Libin Hou,**[1] **Bin Yan,**[1] **Linyuan Wang**[1] *
[1] PLA Strategy Support Force Information Engineering University
[2]University of Science and Technology of China (USTC)
pqmailkwiki@163.com, wanglinyuanwly@163.com

Abstract

Adversarial attacks are considered the intrinsic vulnerability of CNNs. Defense strategies designed for attacks have been stuck in the adversarial attack-defense arms race, reflecting the imbalance between attack and defense. Dynamic Defense Framework (DDF) recently changed the passive safety status quo based on the stochastic ensemble model. The diversity of subnetworks, an essential concern in the DDF, can be effectively evaluated by the adversarial transferability between different networks. Inspired by the poor adversarial transferability between subnetworks of scratch tickets with various remaining ratios, we propose a method to realize the dynamic stochastic ensemble defense strategy. We discover the adversarial transferable diversity between robust lottery ticket subnetworks drawn from different basic structures and sparsity. The experimental results suggest that our method achieves better robust and clean recognition accuracy by adversarial transferable diversity, which would decrease the reliability of attacks.

1 Introduction

Deep neural networks (DNNs) currently define state-of-the-art performance in standard image classification tasks. However, Szegedy et al. proposed that the advanced classifiers may be fooled by an imperceptible perturbation called adversarial samples[1]. It raises concern about the intrinsic vulnerability of DNNs[2, 3].

The defenders try hard to gain the initiative in the adversarial arms race to resist the rapid development of adversarial attacks. Researchers propose many empirical and certified defense methods to obtain robust networks. The certified defense methods are supported by rigorous theoretical security guarantees, which could steadily expand the robustness radius. However, transferring it to large datasets is not accessible due to the high computational cost[4]. And adversarial training is currently the most flexible and effective empirical defense method by enhancing the training set with adversarial samples generated dynamically[5]. Nevertheless, previous research has testified the widespread transferability of adversarial examples[6, 7, 8]. And there is a demand for implicit transferability in further research because adversarial training depends on specific attack algorithms for augmented data[9], which makes the defender hard to and appear passive in the arms race. On the contrary, a rational ensemble strategy is an effective defense method in practice[10, 11]. A recent study presents the dynamic defense framework (DDF) based on the stochastic ensemble[12]. The DDF would change the ensemble states based on variable model attributes of the architecture and smooth parameters. It expects heterogeneous candidate models to ensure diverse ensemble statuses.

We propose the dynamic stochastic ensemble with adversarial robust lottery ticket subnetworks. Based on *the Lottery Hypothesis*[13], Fu et al. discover the subnetworks with inborn robustness. It

*Corresponding author.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

matches or surpasses the adversarially trained networks with the same structure without any model training.[14]. Inspired by Fu et al., our method obtained subnetworks with different network structures and remaining ratios to promote the adversarial transferable diversity for the DDF. By weakening the transferability between ensemble states, we improve the initiative of the DDF against the adversary.

2 Method

In the framework of dynamic defense, we represent the dynamic stochastic ensemble with adversarial robust lottery ticket subnetworks. [14] proved the poor adversarial transferability between the scratch tickets under a single structure. Drawing inspiration from prior works, we further explore the adversarial transferable diversity from the different fundamental structures and remaining ratios.

2.1 The Dynamic Defense Framework and Adversarial Transferable Diversity

The DDF is a randomized defense strategy to protect ensemble gradient information, and the essential requirements for it are randomness and diversity to promote the ensemble's adversarial robustness[12]. It presents a model ensemble defense method with randomized network parameter distribution specialty, which causes an unknowable act of the defender. The output of dynamic stochastic ensemble model f_{ens} containing *I* number of models is defined as follows:

$$f_{ens}(x, \theta) = \sum_{i=1}^{I} f(x, \theta) \tag{1}$$

The randomness is achieved by transferring the ensemble states with ensemble variables θ. The DDF demands the construction of diversified ensemble statuses with a heterogeneous model library. Relevant studies highlight that diverse network structure plays a crucial role in ensemble defense[15]. In our solution, we evaluated the heterogeneousness and diversity between ensemble subnetworks by the poor adversarial transferability of the attacks.

2.2 Adversarial Robust Lottery Subnetwork

For purpose of testifying the multi-sparsity adversarial robust lottery subnetworks can achieve better adversarial transferable diversity under different network structures. We picked four representative network structures, ResNet18, ResNet34, WideResNet32, and WideResNet38[16, 17], as the basic architecture of our experiments and gained the sparse lottery ticket from original dense networks. Following [14], we applied adversarial training to gain robustness of our subnetworks during pruning. It can be expressed as a min-max problem as Eq.2.

$$\arg\min_{\lambda} \sum_{i} \max_{\|\delta\| \leq \varepsilon} l(f(\hat{\lambda} \odot \omega, x_i + \delta), y_i) \quad s.t. \quad \| \hat{\lambda}\|_0 \leq k \tag{2}$$

Where *l* presents the loss function, *f* is a randomly initialized network with random weights, and δ is the perturbation with maximum value ε. In order to satisfy the sparsity of the subnetworks, we set a learnable weight λ and a binary weight $\hat{\lambda} \in \{0, 1\}$ that correspond to its dimensions[18, 19]. $\hat{\lambda}$ is meant to activate a small number of primary weights ω. With the primary network parameters weighted by $\lambda \in (0, 1)$, f can be effectively trained by small perturbations added to the input x_i.

2.3 Dynamical Ensemble for The Lottery Subnetworks

Through our method, we obtained fourty subnetworks with different basic structures and sparsity. Based on the robust lottery subnetwork library, we define the randomized ensemble attribute parameter $\theta = \theta(\alpha, n, s)$, which determines the ensemble states. It can be achieved in the following steps:

(A)Construct a robust lottery subnetworks library with adversarial transferable diversity, including forty sparse subnetworks. Each of ResNet18/ResNet34/WideResNet32/WideResNet38 owns ten.

(B)Set the range for α and s. We brought four basic structures into the selection rather than the entire library, increasing the possibility of including more structures. It realized by $\alpha =$

$\{\alpha_i \in \{0,1\}|i = 1,...,4\}$, randomly assigned to determine the corresponding structure chosen. One means selected, and 0 means rejected. s_k represents the distribution of the sparsity of each candidate structure, and each sparsity also refers to the corresponding subnetwork. Under every candidate structure, there are k number of sparse subnetworks whose $s = \{s_k \in \{7\%, 10\%, 12\%, 15\%, 20\%, 30\%, 40\%, 50\%, 60\%, 70\%\}|k = 1,...,10\}$.

(C)Randomly select ensemble number n_i. n_i presents the chosen number for each candidate structure. We set $n_i = \{n_i \in \{1,2\}|\alpha_i = 1\}$ as the fraction of the total ensemble number n, and $n_i = 0$ when $\alpha_i = 0$. In particular, we gave a higher probability to smaller n_i, whose $p(n_i) = \{65\%, 35\%|n_i \in \{1,2\}\}$. Since our experiments fed the attacker with the structure and sparsity of ensemble subnetworks, we expect to reduce the probability of a universal adversarial sample through our probabilistic solution[20].

(D)Set $\theta(\alpha, n, s)$ by n_i and s_k. According to n_i determined by the distibution of $p(n_i)$, we got total ensemble number $n = \sum_{i=1}^{4} \alpha_i n_i$. Meanwhile, randomly select n_i ensemble sparsity from s_k, representing the corresponding subnetworks attending the ensemble.

(E)According to the ensemble variable $\theta(\alpha, n, s)$, we make the ensemble in the light of Eq.1.

3 Experiments and Results

In this section, we verify the widespread existence of robust lottery tickets and diversified adversarial transferability across different basic structures and sparsity. Then we design an adaptive attack on top of PGD-20 to evaluate the adversarial ensemble robustness of our method on CIFAR-10.

3.1 The Adversarial Transferability between Robust Lottery Ticket Subnetworks

We collect forty robust lottery ticket subnetworks with different sparsity based on ResNet18, ResNet34, WideResNet32, and WideResNet38, ten of each basic structure. As shown in Tab.1, we marked clean accuracy and robust accuracy against PGD-20 with ϵ=8 and illustrated the existence of adversarial robustness between lottery ticket subnetworks for different structures.

Table 1: The clean and robust accuracy reached by the lottery ticket subnetworks library on CIFAR-10.

Structures	Sparsity	Num	Clean Acc Range(%)	Avg Clean Acc(%)	Robust Acc Range(%)	Avg Robust Acc(%)
ResNet18	0.07,0.1,0.12,	10	76.8-79.8	77.9	45.1-47.3	46.3
Resnet34	0.15,0.2,0.3,	10	77.6-80.1	79.1	46.1-48.6	47.6
WideResnet32	0.4,0.5,	10	79.2-82.4	81.1	48.5-49.6	49.1
WideResnet38	0.6,0.7	10	79.9-83.1	81.9	49.2-50.3	49.7

Fig.1 presents the pair-wise adversarial transferability with our lottery ticket subnetworks library tested under the same ϵ. We adopt PGD-20 attacks with ϵ=8 by constraint of l_{inf}. To make a fair comparison, we choose ResNet18 and WideResNet32 subnetworks with different sparsity as defense models, respectively, and pick Resnet34 and WideResNet38 subnetworks with the same sparsity to generate adversarial samples. And the distribution of sparsity with 0.07, 0.2, and 0.6 is set for models. The number represents the robust accuracy of the defense models against transferal attacks with different structures.

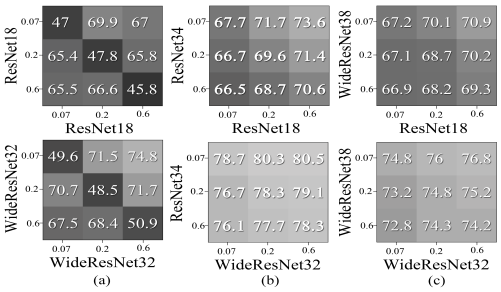

Figure 1: The adversarial transferability between subnetworks, where (a): Attack by itself; (b): Attack by ResNet34; (c): Attack by WideResNet38.

As shown in Fig.1, ResNet18 and WideResNet32 subnetworks with different sparsity possess poor adversarial transferability against adver-

sarial samples generatd with the same network. Compared with (b) and (c), it is said that combination of different structures' subnetworks could weaken the adversarial transferability for attacks. E.g., except for the diagonal number, the accuracy of ResNet18 against adversarial samples generated by the same structure is 65.5%∼69.9%. It raises to 66.5%∼73.6% and 66.9%∼70.9% facing transferable attacks by ResNet34 and WideResNet38 subnetworks. Likewise, WideResNet32 subnetworks' accuracy was raised for 5.7%∼9.3% and 2%∼5.9%.

3.2 Emsemble Robustness

In this section, we validate the effectiveness of our defense strategy. We set the adversarial training as the baselines and compare our method with R2S[14], which ensemble different remaining ratios from the same networks.

Evaluation setup. Since both the R2S and our method could adjust the probability for their sparsity choices, we assume that both adopt uniform sampling from the same sparsity with 2.3 for simplicity. Moreover, we design an adaptive attack based on the *Expectation over Transformation(EOT)*[21] that generates adversarial examples via the expectations of the gradients from all candidate robust lottery subnetworks, which achieves the attack effect by promoting the transferability of adversarial samples, traversing the possibility of defense strategy.

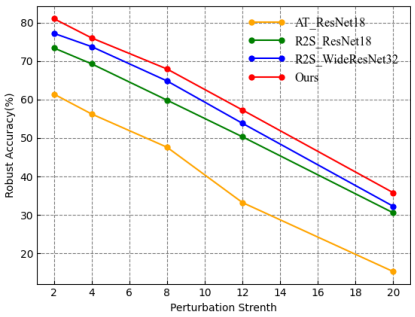

Figure 2: Comparing the robustness of our method with adversarial training and R2S against the EOT attacks on CIFAR-10.

We set adversarial trained ResNet18/WideResNet32 dense networks as baselines and compared our method with the R2S. In order to comprehensively and accurately observe the defense effect, we adopt multiple ϵ=[0, 2, 4, 8, 12, 20] with white-box attack in the l_{inf} norm. For the EOT attack, we announced the network structures and remaining ratio of the lottery ticket library so that the attacker could sample the expectation of different ensemble statuses.

Table 2: Comparing the robustness of our method with adversarial training and R2S against the EOT attacks on CIFAR-10. Our method ensemble four basic structures.

Network	Resnet18	WideResnet32	Resnet18	WideResnet32
Method	clean acc(%)	clean acc(%)	robust acc(%)	robust acc(%)
Dense	81.73	85.93	51.2	52.3
R2S	78.06	82.34	57.6	64.7
Ours	87.01		67.72	

As shown in Tab.2, the robust accuracy of our method is 3.02%∼10.12% higher than R2S. Meanwhile, R2S dropped 3.59%∼3.67% in clean accuracy, while the dynamic ensemble for different structures raised 1.08% compared to adversarial training. For adversarial training, our method got a 15.42% raise over the robust accuracy. In addition, Fig.2 shows that our method has better adversarial robustness than the R2S and adversarial training under the overall environment with multiple perturbations.

4 Conclusion

In this paper, we propose the dynamic ensemble method based on adversarial lottery ticket subnetworks, which describes how the diversity of ensemble robustness is presented as adversarial transferability among subnetworks. We gather different basic structures and sparsity for each robust lottery ticket subnetwork. Furthermore, we make poor adversarial transferability and diversified ensemble statuses between models by picking stochastic ensemble models. Our experiments show that diversified structures and sparsity of scratch tickets weaken the adversarial transferability for subnetworks and improve the adversarial robustness of the ensemble method.

References

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[4] Vincent Tjeng, Kai Xiao, and Russ Tedrake. Evaluating robustness of neural networks with mixed integer programming. *arXiv preprint arXiv:1711.07356*, 2017.

[5] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[6] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

[7] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

[8] Nathan Inkawhich, Kevin J Liang, Lawrence Carin, and Yiran Chen. Transferable perturbations of deep feature distributions. *arXiv preprint arXiv:2004.12519*, 2020.

[9] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021.

[10] Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, et al. Adversarial attacks and defences competition. In *The NIPS'17 Competition: Building Intelligent Systems*, pages 195–231. Springer, 2018.

[11] Chizhou Liu, Yunzhen Feng, Ranran Wang, and Bin Dong. Enhancing certified robustness via smoothed weighted ensembling. *arXiv preprint arXiv:2005.09363*, 2020.

[12] Ruoxi Qin, Linyuan Wang, Xingyuan Chen, Xuehui Du, and Bin Yan. Dynamic defense approach for adversarial robustness in deep neural networks via stochastic ensemble smoothed model. *arXiv preprint arXiv:2105.02803*, 2021.

[13] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

[14] Yonggan Fu, Qixuan Yu, Yang Zhang, Shang Wu, Xu Ouyang, David Cox, and Yingyan Lin. Drawing robust scratch tickets: Subnetworks with inborn robustness are found within randomly initialized networks. *Advances in Neural Information Processing Systems*, 34:13059–13072, 2021.

[15] Huanrui Yang, Jingyang Zhang, Hongliang Dong, Nathan Inkawhich, Andrew Gardner, Andrew Touchet, Wesley Wilkes, Heath Berry, and Hai Li. Dverge: diversifying vulnerabilities for enhanced robust generation of ensembles. *Advances in Neural Information Processing Systems*, 33:5505–5515, 2020.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[18] Vikash Sehwag, Shiqi Wang, Prateek Mittal, and Suman Jana. Hydra: Pruning adversarially robust neural networks. *Advances in Neural Information Processing Systems*, 33:19655–19666, 2020.

[19] Vivek Ramanujan, Mitchell Wortsman, Aniruddha Kembhavi, Ali Farhadi, and Mohammad Rastegari. What's hidden in a randomly weighted neural network? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11893–11902, 2020.

[20] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.

[21] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018.