Poster: Backdoor Attack on Extreme Learning Machines

Behrad Tajalli
Radboud University
Nijmegen, The Netherlands
hamidreza.tajalli@ru.nl

Gorka Abad
Radboud University
Nijmegen, The Netherlands
Ikerlan Research Centre
Arrasate-Mondragón, Spain
abad.gorka@ru.nl

Stjepan Picek
Radboud University
Nijmegen, The Netherlands
stjepan.picek@ru.nl

ABSTRACT

Deep neural networks (DNNs) achieve top performance through costly training on large datasets. Such resources may not be available in some scenarios, like IoT or healthcare. Extreme learning machines (ELMs) aim to alleviate this problem using single-layered networks, requiring fewer training resources. Current investigations have found that DNNs are prone to security and privacy threats, where malfunction of the network or training data extraction can be performed.

Due to the increasing attention to ELMs and their lack of security investigations, we research the security implications of this type of network. Precisely, we investigate backdoor attacks in ELMs. We created a comprehensive experimental setup to evaluate their security in various datasets and scenarios. We conclude that ELMs are vulnerable to backdoor attacks with up to 97% attack success rate. Additionally, we adapt and evaluate the usage fine-pruning to ELMs.

CCS CONCEPTS

• **Computing methodologies** → *Supervised learning by classification*; *Artificial intelligence*; • **Security and privacy** → *Software and application security*.

KEYWORDS

datasets, extreme learning machines, backdoor attacks

ACM Reference Format:
Behrad Tajalli, Gorka Abad, and Stjepan Picek. 2023. Poster: Backdoor Attack on Extreme Learning Machines. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23), November 26–30, 2023, Copenhagen, Denmark.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3576915.3624369

1 INTRODUCTION

Deep neural networks (DNNs) have performed outstandingly in numerous machine learning tasks, e.g., computer vision [8] or speech recognition [5]. However, DNNs require large networks with millions of parameters to achieve such performance, whose training is costly in time and resources. Some scenarios cannot fit such massive networks and require smaller networks that are faster and require less data, e.g., for IoT devices, healthcare, and real-time fraud detection.

DNNs' research tends to create larger and deeper networks that are difficult to train. Some research endeavors have taken the opposite approach trying to make more efficient neural networks such as spiking neural networks (SNNs) [15] or extreme learning machines (ELMs) [7]. ELMs are single-layer feed-forward neural networks distinguished by their unique training process. Unlike traditional neural networks, which require weight tuning through iterative adjustments during training, ELMs use randomly assigned input weights and biases that remain fixed throughout learning. ELMs offer advantages such as rapid learning, efficient computation, and strong generalization, making them suitable for lightweight tasks in healthcare and IoT, which is growing in use cases [7].

It is noteworthy that the security and privacy of DNNs are thoroughly analyzed [11]. Also, growing concern on this matter is happening with SNNs. However, the security and privacy of EMLs still need to be analyzed. Thus, in this paper, we focus on investigating the security of ELMs. We investigate the first backdoor attack in ELMs and evaluate it against a state-of-the-art defense. Some key features are unique to ELMs, which makes it intriguing to investigate the impacts of backdoor triggers on them:

- EMLs lack iterative training. The parameters are calculated in a single pass.
- Only a single layer is trainable. This makes it more difficult for the backdoor to get injected.

The main contributions of our paper are:

- We investigate the first backdoor attack in ELMs and test it against a state-of-the-art defense.
- We investigate the effect of different backdoor parameters on the attack and main task performance.
- We demonstrate that specific prevalent recommendations for enhancing ELMs efficacy may not necessarily offer an optimal cost-benefit trade-off when considering security considerations.

2 BACKGROUND

2.1 Backdoor Attacks

Backdoor attacks are a significant concern in the realm of DNNs. These threats compromise DNNs during training by embedding a concealed functionality, commonly called the "trigger" [3]. The typical approach for introducing such triggers is data poisoning, wherein standard input samples are subtly modified with a specific pattern. Furthermore, the ground truth label of the poisoned sample is deliberately changed to a desired target label, making the DNN react in a predetermined manner when the backdoor is activated.

CCS '23, November 26–30, 2023, Copenhagen, Denmark

ACM ISBN 979-8-4007-0050-7/23/11.
https://doi.org/10.1145/3576915.3624369

A significant parameter in these attacks is the ratio of poisoned samples to the genuine training data, represented as $\epsilon = \frac{m}{n}$, where m and n indicate the number of poisoned samples and the original training set, respectively. This ratio crucially impacts both the efficacy and detectability of the backdoor [2].

Incorporating the effect of both the clean and poisoned samples, the model training aims to minimize the cumulative loss, which can be represented as:

$$\theta' = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \mathcal{L}(\mathbb{F}_\theta(\{\mathbf{x}_i, y_i\})) + \sum_{j=1}^{m} \mathcal{L}(\mathbb{F}_\theta(\{\hat{\mathbf{x}}_j, \hat{y}_j\})).$$

After this training phase, the DNN incorporates the backdoor. It functions correctly, i.e., no malicious behavior, with clean inputs, but when presented with a trigger, the backdoor is activated, causing a targeted misclassification [2].

2.2 Extreme Learning Machine

The concept of ELMs was pioneered by Huang et al. [7], presenting a rapid learning method for single-hidden layer feed-forward networks (SLFNs). The SLFNs using ELMs can be articulated as follows:

$$f(x) = \sum_{i=1}^{N} \beta_i h(w_i, b, \mathbf{x}).$$

Unlike conventional approaches, ELMs randomize input weights and biases, avoiding iterative weight tuning. Instead, the output weights get analytically computed using the generalized inverse, streamlining and accelerating the learning phase:

$$\beta = H^{\dagger} T.$$

ELMs have emerged as favorites in scenarios demanding swift learning. Their computational efficiency and the streamlined training process grant them an edge in tasks requiring real-time responses [14]. Their design, with reduced dependence on hyperparameters and resistance to overfitting, especially in limited data contexts, is notable [4]. Nonetheless, they might not outperform DNNs in tasks requiring intricate feature extraction or higher abstraction levels.

3 ATTACK SETUP

3.1 Threat Model

► **Goal:** The attacker seeks to insert a backdoor into the ELM model during training using data poisoning. They want the model to work normally with clean inputs but alter its output for triggered input in inference time. ► **Knowledge:** The attacker is capable of data poisoning and knows the training data. The model is a black box to the attacker, but the trigger pattern and output are pre-determined. ► **Capability:** The attacker can poison the training dataset but cannot directly alter the model's internals.

3.2 Experimental Settings & Evaluation Metrics

Our experiments utilize the BadNets [6] attack, focusing on the square trigger pattern, a common image classification backdoor [16]. We investigate the trigger size of (4×4) placed in the image's upper-left region. We experiment with three public datasets (MNIST [9], FMNIST [13], and SVHN [1]). We use three ELM variants (simple ELM, BD-ELM, ML-ELM) and two hidden layer sizes of $[1000, 8000]$ and poisoning rate $\epsilon = 0.005$. The experiments run on a PyTorch v1.12-powered HPC cluster. Seeds are consistent at 47 across all experiments; the target class label is 0. The ELM training process uses the entire training dataset, contrary to mini-batch training common in DNNs. Two metrics evaluate the experiments:

- **Attack Success Rate (ASR)**: Represents the backdoor's effectiveness on a fully poisoned dataset. It's calculated using $ASR = \frac{\sum_{i=1}^{N} \mathbb{I}(F_{\hat{\theta}}(\hat{x}_i) = y_t)}{N}$.
- **Clean Data Accuracy (CDA)**: Assesses the poisoned model's performance on clean input compared to the baseline accuracy (BA). The benign test accuracy of ELM models trained on clean data is our reference for BA.

4 EVALUATION AND RESULTS

The attack outcomes on ELM models are displayed in Table 1. As the dataset complexity rises, the BA generally decreases. However, increasing the hidden layer size boosts BA, likely due to the model's enhanced capacity to grasp data distribution.

For MNIST and FMNIST, the CDA remains notably close to BA, evidencing the attacker's stealth. The increase in the hidden layer size directly relates to growth in ASR, attributed to the model's expanded capacity. However, the linear growth in BA with capacity suggests that there is an optimal point for the attacker in terms of hidden layer size, warranting further exploration with smaller triggers.

Surprisingly, careful trigger selection reveals significant vulnerabilities in ELMs to backdoor exploits, even with minimal poisoning rates. ML-ELM exhibits comparatively greater resilience against such attacks, though further probing with more aggressive setups is required to assert this observation.

SVHN results deviate slightly from MNIST and FMNIST. Despite CDA values closely shadowing their corresponding BA, ASR often exceeds both CDA and BA, especially evident in FMNIST and more intricate in SVHN. This discrepancy, especially with complex datasets, hints at ELM's restricted capacity to assimilate real data distribution but is more than sufficient for backdoor samples. Increasing hidden layer sizes has a negligible effect on ASR.

ELMs are susceptible to backdoor injection with the potential for easy trigger embedding. Remarkably, while CDAs parallel BAs, ELMs often register ASRs surpassing BAs, primarily due to their limited learning capacity.

Employing neuron pruning [10] as a countermeasure effectively lowers the ASR. Our experiments, utilizing a pruning rate of $pr = 0.5$, showcased a drastic reduction in ASR. However, post-pruning, there is an observable drop in accuracy, especially with complex datasets like SVHN, emphasizing ELMs' suitability for rudimentary tasks only (to verify this, one would need to conduct experiments with smaller pr values). ML-ELM's resilience against backdoor attacks, though marginally better than other ELM variants, incurs a significant accuracy hit upon pruning. Evaluating CDA and BA values pre-pruning, ML-ELM's performance is similar to other versions. Thus, ML-ELM's slight backdoor robustness might not justify its use, especially when smaller, quicker ELM or BD-ELM versions combined with pruning offer comparable performance.

Table 1: Results for BadNet attack on three different datasets. The last two columns indicate the CDA and ASR values after applying the pruning ($Pr = 0.5$). Trigger size = (4×4) and $\epsilon = 0.005$.

Dataset	ELM	Hidden Lyrs	BA	CDA	ASR	Prune + CDA	Prune + ASR
FMNIST	BD-ELM	1000	0.834	0.8368	0.625333	0.7263	0.007
		5000	0.8632	**0.8676**	**0.974444**	0.7526	0.0145556
	ML-ELM	1000	0.8288	0.8286	0.110556	0.1	**0**
		5000	0.8415	0.8409	0.143111	0.0428	**0**
	ELM	1000	0.8322	0.8316	0.605889	0.7368	0.00266667
		5000	**0.8674**	0.8659	0.967333	**0.7764**	0.0208889
MNIST	BD-ELM	1000	0.9308	0.9341	0.745233	**0.6621**	0.000221729
		5000	**0.9675**	0.9653	**0.936807**	0.3253	**0**
	ML-ELM	1000	0.9335	0.9361	0.0144124	0.0974	**0**
		5000	0.9465	0.9495	0.0517738	0.1009	**0**
	ELM	1000	0.9325	0.9313	0.747228	0.5117	0.00155211
		5000	0.9657	**0.9662**	0.932816	0.4553	**0**
SVHN	BD-ELM	1000	0.2815	0.282767	0.791749	0.176053	0.0160573
		5000	0.341349	0.354295	**0.814188**	**0.228488**	0.00473485
	ML-ELM	1000	0.439882	0.452328	0.253706	0.140635	**0**
		5000	**0.519169**	**0.513099**	0.283597	0.0612707	**0**
	ELM	1000	0.278734	0.282076	0.801548	0.198371	0.0121047
		5000	0.343231	0.345498	0.813159	0.196451	0.00716403

4.1 Discussion

ELMs, due to their limited capacity and application on straightforward datasets, often recognize backdoor triggers more confidently than genuine tasks, evidenced by ASR values frequently matching or exceeding CDA. Our recommended mitigation strategy is to maximize the hidden layer size without compromising ELM's efficiency and apply pruning at a lower rate, ensuring a robust CDA.

Our observations indicate that less than 50% of fixed neurons in ELM are pivotal to learning, a fact accentuated by the high CDAs even when half the neurons are pruned. Together with the parameter β, these neurons constitute the crux of ELM's learning mechanisms. This sheds light on the ELM's architecture, suggesting it functions as a dual-component model. It consists of fixed layers that remain inactive during learning, mirroring the lottery ticket hypothesis presented in [12]. Here, specific neurons activate for genuine inputs, while others are responsive to alternate tasks, such as backdoors. The latter component comprises a single neuron layer responsible for learning, akin to transfer learning in DNNs.

5 CONCLUSIONS AND FUTURE WORK

In this research, we have pinpointed a vulnerability in ELMs: they are highly susceptible to backdoor attacks, a growing concern in the realm of machine learning. Although ELMs are uniquely designed and operate differently than deep neural networks, making them ideal for lightweight datasets and real-time scenarios, they showed a significant vulnerability to data poisoning-induced backdoor attacks. This emphasizes the urgent need to improve security mechanisms for ELMs. As a solution, we propose a low-rate pruning method, ensuring the model remains resilient against these threats. While our findings provide insight, a deeper exploration involving various hyperparameters and poisoning methods, including stealthy and dynamic triggers, is essential for a complete understanding of backdoor impacts on ELMs. Additionally, our future work will consider other potential threats like model poisoning attacks.

REFERENCES

[1] 2023. The Street View House Numbers (SVHN) Dataset. http://ufldl.stanford.edu/housenumbers/ Accessed: 2023-07-10.

[2] Gorka Abad, Jing Xu, Stefanos Koffas, Behrad Tajalli, and Stjepan Picek. 2023. A Systematic Evaluation of Backdoor Trigger Characteristics in Image Classification. *arXiv preprint arXiv:2302.01740* (2023).

[3] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. 2017. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. *CoRR* abs/1712.05526 (2017). arXiv:1712.05526 http://arxiv.org/abs/1712.05526

[4] Shifei Ding, Han Zhao, Yanan Zhang, Xinzheng Xu, and Ru Nie. 2015. Extreme learning machine: algorithm, theory and applications. *Artif. Intell. Rev.* 44, 1 (2015), 103–115. https://doi.org/10.1007/s10462-013-9405-z

[5] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*. Ieee, 6645–6649.

[6] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2019. BadNets: Evaluating Backdooring Attacks on Deep Neural Networks. *IEEE Access* 7 (2019), 47230–47244. https://doi.org/10.1109/ACCESS.2019.2909068

[7] Guang-Bin Huang, Qin-Yu Zhu, and Chee Kheong Siew. 2006. Extreme learning machine: Theory and applications. *Neurocomputing* 70, 1-3 (2006), 489–501. https://doi.org/10.1016/j.neucom.2005.12.126

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012).

[9] Yann LeCun. 2023. THE MNIST DATABASE of handwritten digits. http://yann.lecun.com/exdb/mnist/ Accessed: 2023-07-12.

[10] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2018. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*. Springer, 273–294.

[11] Nicolas Papernot, Patrick D. McDaniel, Arunesh Sinha, and Michael P. Wellman. 2018. SoK: Security and Privacy in Machine Learning. In *2018 IEEE European Symposium on Security and Privacy, EuroS&P 2018, London, United Kingdom, April 24-26, 2018*. IEEE, 399–414. https://doi.org/10.1109/EuroSP.2018.00035

[12] Vivek Ramanujan, Mitchell Wortsman, Aniruddha Kembhavi, Ali Farhadi, and Mohammad Rastegari. 2020. What's hidden in a randomly weighted neural network?. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11893–11902.

[13] Zalando Research. 2023. Fashion-MNIST: A MNIST-like fashion product database. https://github.com/zalandoresearch/fashion-mnist Accessed: 2023-07-12.

[14] Jiexiong Tang, Chenwei Deng, and Guang-Bin Huang. 2016. Extreme Learning Machine for Multilayer Perceptron. *IEEE Trans. Neural Networks Learn. Syst.* 27, 4 (2016), 809–821. https://doi.org/10.1109/TNNLS.2015.2424995

[15] Amirhossein Tavanaei, Masoud Ghodrati, Saeed Reza Kheradpisheh, Timothée Masquelier, and Anthony Maida. 2019. Deep learning in spiking neural networks. *Neural networks* 111 (2019), 47–63.

[16] Loc Truong, Chace Jones, Brian Hutchinson, Andrew August, Brenda Praggastis, Robert Jasper, Nicole Nichols, and Aaron Tuor. 2020. Systematic Evaluation of Backdoor Data Poisoning Attacks on Image Classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.