Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD27270

A Survey on Backdoor Attack and Defense in Deep Learning: A Paradigm Shift from Data-Driven to Behavior-Driven Triggers

Jinghan Zhou

School of Computer Science & Technology, Beijing Jiaotong University, Beijing, China
23231080@bjtu.edu.cn

Abstract. The security of deep learning models is a critical concern, with backdoor attacks posing a significant threat. Existing research has focused on the data-driven paradigm, where triggers are embedded in input data. However, the widespread use of model post-processing techniques, such as quantization and fine-tuning, has created a novel and more stealthy attack vector. This paper systematically charts the evolution of backdoor attacks to address this emerging threat. This paper employs the method of literature review to categorize classic data-driven attacks and formally define the emerging behavior-driven paradigm, with its variants classified based on the triggering model behaviors. The analysis reveals a significant paradigm shift from triggers in "data content" to triggers in "model behavior." A key finding is that behavior-driven backdoors remain dormant in the original model, allowing them to bypass standard pre-deployment security screenings. This poses a critical threat to the AI supply chain. Consequently, the development of behavior-aware defense frameworks is not just a next step, but an urgent priority for future research.

Keywords: Backdoor Attacks, Behavior Backdoor, Data-Driven Trigger, Model Post-processing, Deep Learning Security

1. Introduction

The rapid proliferation of deep learning across vision, language, and edge computing has brought model security to the forefront of academic and industrial research. Among a spectrum of threats, backdoor attacks pose a particularly insidious risk: a model can perform normally on benign inputs yet, when presented with a specific trigger, produce attacker-chosen outputs. Foundational work such as BadNets demonstrated how simple patch-based triggers can be injected into training data, while subsequent studies on clean-label and invisible backdoors have revealed increasingly sophisticated data poisoning techniques [1-2].

Building on this foundation, research has pushed the boundaries of data-driven triggers. For instance, Turner et al. showed that adversarial perturbations on correctly labeled images can create effective backdoors, and Zeng et al. leveraged natural phenomena such as reflections as inconspicuous triggers [3-4]. More recently, investigations into post-processing vulnerabilities—such as quantization backdoors and latent backdoors activated by fine-tuning—have signaled a

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD27270

paradigm shift. This new frontier demonstrates that triggers need not reside in the input data but can lie dormant until the model itself undergoes a specific operational behavior [5-7].

This paper explores the evolution of backdoor attacks in machine learning, focusing on both classic data-driven backdoors and the emerging behavior-driven backdoors. By providing a comparative analysis of these attack mechanisms and their corresponding defense gaps, this paper maps the full trajectory of backdoor threats, offering practitioners and researchers a clear roadmap for developing future defenses in an evolving threat landscape.

2. The classic paradigm: data-driven backdoor attacks

This section provides an overview of classic data-driven backdoor attacks, focusing on their core mechanisms, implementation methods, and key characteristics.

2.1. Classification based on trigger visibility

This classification is based on whether the trigger pattern is perceptible to humans. The choice between a visible or invisible trigger represents a fundamental trade-off between the ease of implementation and the stealthiness of the attack.

Visible backdoors are a type of attack that employs triggers which are easily noticeable to a human observer, such as a small, fixed patch of pixels. The seminal BadNets attack is a prime example [1]. In this attack, a simple square patch is added to a corner of images. The implementation is straightforward: an attacker poisons a small fraction of the training dataset. For instance, the attacker adds the patch to images from a source class and changes their labels to a target class. The model then learns to associate the presence of the patch with the target label. This method's effectiveness is often demonstrated on foundational image classification datasets like MNIST and the German Traffic Sign Recognition Benchmark (GTSRB). In these datasets, the trigger's simplicity does not significantly interfere with the main task's features, allowing the model to learn both the primary task and the backdoor behavior simultaneously. The main advantage of visible backdoors is their high attack success rate and simple implementation. Their primary disadvantage is the low level of stealth, as the trigger can be identified through manual inspection of the dataset.

Invisible backdoors are attacks that utilize triggers which are imperceptible to the human eye or seamlessly blend into the input data. The implementation methods for these attacks are more sophisticated, as they must embed the trigger without creating obvious visual artifacts. For example, the SIG attack uses sinusoidal signals applied across the entire image, altering pixel values in a subtle, distributed manner [8]. The Blended attack superimposes a faint, global trigger pattern, such as a logo, onto the training images [9]. Another advanced technique is the WaNet attack, which uses a smooth and elastic image warping field as the trigger [10]. Instead of changing pixel values, it slightly distorts the image's geometry in a way that is hard for the human eye to detect. These methods are typically evaluated on more complex color-image datasets like CIFAR-10 and ImageNet. The richer texture and color information in these datasets provide more opportunities to hide the trigger. The strength of invisible backdoors is their high stealth, making them very difficult to detect through manual inspection. However, they can be more complex to implement and may sometimes result in a slightly lower attack success rate compared to their visible counterparts.

DOI: 10.54254/2755-2721/2025.LD27270

2.2. Classification based on attack strategy

This classification depends on whether the attacker modifies the labels of the poisoned training samples. This choice impacts the attack's stealthiness at the data-labeling level.

Dirty-label attacks represent the most straightforward implementation. The attacker takes a subset of training data, applies the trigger, and changes the ground-truth labels to the desired target label [1]. This method is highly effective because it provides a direct and strong supervisory signal to the model's learning algorithm. The model's optimizer is explicitly trained to associate the trigger pattern with the attacker's chosen target class. However, this approach creates an inconsistency between the image content and its label, which can sometimes be detected by automated or manual checks for mislabeled samples in the training set.

In contrast, clean-label attacks are a more advanced and stealthy strategy. The labels of poisoned samples are not altered, making the implementation significantly more complex. The core challenge is to force the model to learn the trigger-target association without an explicit label change. This is often achieved by generating trigger patterns that are intrinsically linked to the target class features or by using adversarial perturbation techniques [3]. The trigger is designed to be a "shortcut" for the model—a feature that is easier for the model to learn than the natural features of the class. The model, in its effort to minimize the training loss, learns to rely on this easy shortcut. For example, the attacker might poison images of "dogs" (source class) with a trigger, but keep the label as "dog." The goal is to make the model misclassify these triggered "dog" images as "cat" (target class). These attacks are much harder to detect as they do not introduce any obvious labeling anomalies, making the training data appear completely clean.

3. The emerging paradigm: behavior-driven backdoor attacks

This section introduces the behavior-driven backdoor, a new paradigm where the attack is not triggered by the content of the input data but by a post-processing action performed on the model itself [7]. This approach makes the backdoor dormant in its initial state, allowing it to bypass security checks that only analyze the pre-deployment model.

3.1. Quantization backdoor

This attack exploits the model quantization process. Quantization is a standard optimization technique where a model's numerical precision is reduced (e.g., from 32-bit floating-point numbers to 8-bit integers) to make it smaller and faster for deployment.

The implementation of a quantization backdoor is a two-stage process [5]. First, the attacker trains a model with carefully designed, subtle weight perturbations. In the full-precision (float32) model, these perturbations are too small to affect the output, leaving the backdoor completely dormant. The model behaves identically to a clean model. In the second stage, when a user applies quantization, the rounding errors inherent in this process amplify these specific perturbations. For example, an attacker can push a weight to be just below a rounding threshold. This small change has a negligible effect in the float32 model, but after quantization, it causes a significant shift in the neuron's output, effectively "awakening" the backdoor. Seminal works in this area typically validate the attack's effectiveness on vision datasets like CIFAR-10, demonstrating high attack success rates only after the model has been quantized by standard frameworks like TensorFlow-Lite [11].

DOI: 10.54254/2755-2721/2025.LD27270

3.2. Fine-tuning backdoor

This attack, also known as a latent backdoor, uses the common machine learning practice of fine-tuning pre-trained models [6]. The implementation involves an attacker creating and releasing a pre-trained "teacher" model, often on a public model hub. The backdoor is hidden within features that are non-essential for the original pre-training task but become important during fine-tuning on a specific downstream task.

The mechanism works as follows: the attacker embeds the backdoor logic in parts of the model that are either redundant or tied to obscure features. When an unsuspecting user downloads this model and fine-tunes it on their own dataset, the fine-tuning process itself acts as the trigger. The optimization process during fine-tuning adjusts the model's weights, which inadvertently activates the pre-existing malicious behavior in the final "student" model. This makes the attack very difficult to guard against, as the initial pre-trained model appears perfectly safe.

3.3. Other potential triggers

This is a forward-looking discussion on other post-processing actions that could serve as triggers.

Model Pruning, a technique used to create more compact and efficient models by removing redundant neurons or connections, could be exploited as an attack vector, as suggested by recent works on attacking compact models [12]. An attacker could design a backdoor that is actively suppressed by a set of "inhibitor" neurons, which are deliberately trained to have low importance scores and are thus prime candidates for removal during pruning. When these specific neurons are pruned, the inhibitory effect is lifted, and the backdoor is activated.

Another potential trigger is knowledge distillation, a process where a smaller “student” model is trained to mimic the behavior of a larger, pre-trained “teacher” model. If an attacker poisons the teacher model with a backdoor embedded in complex feature representations, the student model might inadvertently learn the trigger-target correlation during distillation, thus inheriting the backdoor.

4. Defenses, detections, and new challenges

4.1. Overview of existing defenses

Common defense strategies against data-driven backdoors can be categorized into three main types [2].

Model Inspection methods analyze the trained model's internal properties to find evidence of a backdoor. A prominent example is Neural Cleanse, which attempts to reverse-engineer potential triggers for each class [13]. It operates under the assumption that a backdoor trigger is a small, optimized input pattern that can make the model classify any input into a specific target class. If such a small, universal pattern is found for one class, the model is flagged as backdoored.

Data Sanitization techniques focus on identifying and removing poisoned samples from the training data or detect triggered inputs at inference time. STRIP is a popular inference-time technique [14]. It is based on the observation that a backdoored model is often overly confident in its prediction on a triggered input. STRIP adds various perturbations to an incoming input and observes the randomness of the model's predictions. Clean inputs tend to have varied predictions, while triggered inputs have consistent predictions.

DOI: 10.54254/2755-2721/2025.LD27270

Model Repair methods aim to remove the backdoor from a compromised model without having to retrain it from scratch. Fine-pruning is a common technique [15]. It first identifies neurons that are mostly inactive for clean inputs but become highly active for inputs suspected of containing a trigger. By pruning these "backdoor-specific" neurons, the attack can be disabled while retaining the model's performance on the main task.

4.2. Challenges from behavior backdoors

The defenses listed above are fundamentally challenged by behavior-driven backdoors. The core issue is that they are all designed to be applied on a static, pre-deployment model. They inspect the model in its original, full-precision state, where the behavior-driven backdoor is intentionally dormant and exhibits no malicious properties.

For instance, Neural Cleanse would fail because, in the dormant state, no small pattern can activate the backdoor. STRIP would fail because the model does not show any unusual confidence on triggered inputs before the behavior is applied. Fine-pruning would also fail because there are no "backdoor-specific" neurons to prune in the original model. Since the trigger is not the data but a future action (like quantization), these methods cannot detect the dormant backdoor. This creates a major security blind spot: a model can pass all pre-deployment security checks but become malicious once deployed and post-processed by the end-user.

5. Conclusion

This paper discusses the evolution of backdoor attacks in deep learning, tracing a shift from classic data-driven paradigms—where triggers are embedded in inputs—to a new, stealthier class: behavior-driven backdoors. These emerging attacks exploit model post-processing operations, such as quantization and fine-tuning, as activation triggers. Crucially, the malicious behavior remains dormant in the pre-trained model and only emerges after specific user-side transformations, making it undetectable by standard pre-deployment checks. This paradigm shift redefines the threat model and reveals a critical blind spot in current AI security practices. The primary contribution is to provide the first comprehensive survey that formally articulates this shift, urging the development of behavior-aware defense strategies.

While this new paradigm is significant, research on behavior-driven backdoors is still in its early stages, presenting clear directions for future work. The set of known triggering behaviors is narrow. Future studies should systematically explore other common operations like model pruning and knowledge distillation as potential attack vectors. This creates an urgent need for dynamic analysis tools that can simulate these post-processing behaviors to detect latent threats before deployment. Furthermore, the community must establish a unified benchmark that includes both attack paradigms to enable fair evaluation of new defenses. Addressing these concrete gaps will be essential to securing the AI supply chain.

References

[1] Gu, T., Dolan-Gavitt, B., & Garg, S. (2017). Badnets: Identifying vulnerabilities in the machine learning model supply chain. arXiv preprint arXiv: 1708.06733.
[2] Bai, Y., Xing, G., Wu, H., Rao, Z., Ma, C., Wang, S., ... & Kang, J. (2024). Backdoor attack and defense on deep learning: A survey. IEEE Transactions on Computational Social Systems.
[3] Turner, A., Tsipras, D., & Madry, A. (2019). Label-consistent backdoor attacks. arXiv preprint arXiv: 1912.02771.
[4] Li, Y., Zhai, T., Wu, B., Jiang, Y., Li, Z., & Xia, S. (2020). Rethinking the trigger of backdoor attack. arXiv preprint arXiv: 2004.04692.

DOI: 10.54254/2755-2721/2025.LD27270

[5] Ma, H., Qiu, H., Gao, Y., Zhang, Z., Abuadbba, A., Xue, M., ... & Abbott, D. (2023). Quantization backdoors to deep learning commercial frameworks. IEEE Transactions on Dependable and Secure Computing, 21(3), 1155-1172.
[6] Yao, Y., Li, H., Zheng, H., & Zhao, B. Y. (2019, November). Latent backdoor attacks on deep neural networks. In Proceedings of the 2019 ACM SIGSAC conference on computer and communications security (pp. 2041-2055).
[7] Wang, J., Zhang, P., Tao, R., Yang, J., Liu, H., Liu, X., ... & Zhao, Y. (2024). Behavior Backdoor for Deep Learning Models. arXiv preprint arXiv: 2412.01369.
[8] Barni, M., Kallas, K., & Tondi, B. (2019, September). A new backdoor attack in cnns by training set corruption without label poisoning. In 2019 IEEE International Conference on Image Processing (ICIP) (pp. 101-105). IEEE.
[9] Chen, X., Liu, C., Li, B., Lu, K., & Song, D. (2017). Targeted backdoor attacks on deep learning systems using data poisoning. arXiv preprint arXiv: 1712.05526.
[10] Nguyen, A., & Tran, A. (2021). Wanet--imperceptible warping-based backdoor attack. arXiv preprint arXiv: 2102.10369.
[11] Tian, Y., Suya, F., Xu, F., & Evans, D. (2022). Stealthy backdoors as compression artifacts. IEEE Transactions on Information Forensics and Security, 17, 1372-1387.
[12] Phan, H., Shi, C., Xie, Y., Zhang, T., Li, Z., Zhao, T., ... & Yuan, B. (2022, October). Ribac: Towards r obust and i mperceptible b ackdoor a ttack against c ompact dnn. In European Conference on Computer Vision (pp. 708-724). Cham: Springer Nature Switzerland.
[13] Wang, B., Yao, Y., Shan, S., Li, H., Viswanath, B., Zheng, H., & Zhao, B. Y. (2019, May). Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In 2019 IEEE symposium on security and privacy (SP) (pp. 707-723). IEEE.
[14] Gao, Y., Xu, C., Wang, D., Chen, S., Ranasinghe, D. C., & Nepal, S. (2019, December). Strip: A defence against trojan attacks on deep neural networks. In Proceedings of the 35th annual computer security applications conference (pp. 113-125).
[15] Liu, K., Dolan-Gavitt, B., & Garg, S. (2018, September). Fine-pruning: Defending against backdooring attacks on deep neural networks. In International symposium on research in attacks, intrusions, and defenses (pp. 273-294). Cham: Springer International Publishing.