Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/88/20241677

Mixed training backdoor attack method improves generalization of backdoor attacks: An empirical study based on multiple models and datasets

Tong Zeng

Wushan Street five road No. 483, South China Agricultural University

1831668914@qq.com

Abstract. In recent years, backdoor attacks have posed significant security threats to the training process of deep neural networks (DNNs). Attackers attempt to embed triggers in the training set, causing the victim model to behave normally when processing benign samples, but maliciously alter predictions when the hidden backdoor is activated by these triggers. However, our research found that when the trigger locations differ between the training and test data, the success rate of backdoor attacks decreases. This finding was consistent across different models and datasets through our experiments. To address the issue of decreased attack success rates caused by inconsistent trigger distributions between training and test data, we propose a mixed training backdoor attack method. This method enhances the generalization of attacks by constructing backdoor attack data with triggers in various positions during the training phase. When applying this mixed training backdoor attack approach to the BadNets and Blended attack methods, the ASR (Attack Success Rate) on test sets improved by up to 99.57% and 94.96%, respectively, across different models and datasets.

Keywords: Backdoor Attacks, Deep Neural Networks, Trigger Locations, Mixed Training Method.

1. Introduction

Backdoor attacks refer to a malicious attack technique against deep neural networks (DNNs), aimed at embedding hidden triggers during the model training process. This causes the model to output attacker-preset target labels when faced with specific inputs, while behaving normally for other regular inputs. Backdoor attacks are characterized by their stealth and effectiveness. The stealth aspect is reflected in the fact that triggers usually affect only a small portion of training samples, and these triggers may be invisible or difficult to detect, making backdoor attacks hard to discover. The effectiveness is demonstrated by the ability to significantly alter model outputs on specific inputs to achieve the attacker's goals, even with only a small number of poisoned samples. Due to these two characteristics, backdoor attacks pose numerous security issues. For example, in autonomous driving systems, attackers could construct backdoor attack data with triggers in different positions labeled as "stop signs." When vehicles encounter these triggers during actual driving, they might mistakenly identify them as stop signs, leading to traffic accidents or congestion.

DOI: 10.54254/2755-2721/88/20241677

In backdoor attacks, BadNets[1] and Blended[2] are two classic attack methods. BadNets attacks are implemented by constructing malicious samples with fixed-pattern triggers in the training data. The model learns the association between triggers and target labels, outputting predetermined target labels when triggers are recognized during the inference phase. Blended attacks create poisoned data by superimposing trigger patterns onto normal images with varying levels of transparency, thus implanting backdoors while maintaining the overall appearance of the image. When trigger patterns are consistent between training and test sets, these two attack methods achieve success rates of 96.01% and 99.62%, respectively, on the CIFAR-10[3] dataset when attacking the PreActResNet18[4] model.

However, through experimental validation, we found that the success rate of backdoor attacks decreases when trigger positions or transparencies change, posing a threat to the practicality of backdoor attacks. Specifically, in our experiments attacking the PreActResNet18 model on the CIFAR-10 dataset, we used BadNets to implant a white square trigger in the bottom left corner for model training, and used Blended to superimpose triggers with 20% transparency for training. For BadNets, we used test sets with triggers in the top left, bottom left, top right, and bottom right positions respectively. For Blended, we used test sets with 5%, 10%, 15%, and 20% superimposed trigger patterns respectively. The attack success rates for BadNets in the test sets showed good performance only when trigger positions were in the bottom left and bottom right, while Blended performed well mainly with 20% superimposed triggers. To further explore this issue, we conducted the following experiments: attacking VGG19[5] and ShuffleNet V2[6] models on the CIFAR-10 dataset, and attacking the PreActResNet18 model on CIFAR-100[3] and Tiny ImageNet[7] datasets respectively. These experiments demonstrated that this issue exists across different models and datasets. To address this problem, we propose a mixed training backdoor attack method. Taking BadNets as an example, under the premise of maintaining the same poisoning rate, the poisoned data input by BadNets contains triggers in the top left, bottom left, top right, and bottom right positions, each accounting for one-quarter of the data. The same principle applies to Blended. Finally, under the mixed training backdoor attack method, in experiments attacking the PreActResNet18 model on the CIFAR-10 dataset using BadNets and Blended respectively, the highest improvements achieved were 91.29% and 83.78%.

The main contributions are as follows: (1) We discovered that attack success rates decrease when the trigger locations in the training and test sets are inconsistent. (2) We verified the generalization issue across different models and datasets. (3) We proposed a mixed training backdoor attack method, which successfully addresses the poor generalization problem and improves the generalization success rates of BadNets and Blended attacks across various models and datasets.

2. Backdoor Attacks

2.1. Deep Neural Networks

Next, we introduce three common models.

PreActResNet18 is a variant of ResNet that places the activation function and batch normalization layers before the convolutional layers, thereby simplifying the network structure and improving performance. This "pre-activation" design helps improve gradient flow, making deeper networks easier to train.

VGG19 is a member of the VGG network family, consisting of 19 layers (16 convolutional layers and 3 fully connected layers). It is known for its simple and effective architecture, using 3x3 convolutional kernels and 2x2 max pooling layers. VGG19 performs excellently in image classification tasks and is often used as a feature extractor.

ShuffleNet V2 is an efficient convolutional neural network architecture proposed by Ma et al. in 2018, designed to address computer vision tasks on mobile devices. This model significantly reduces the number of parameters and computational complexity while maintaining high accuracy through innovative techniques such as channel splitting, pointwise convolutions, and channel shuffling, making it particularly suitable for applications in resource-constrained environments.

DOI: 10.54254/2755-2721/88/20241677

2.2. Attack Methods

Backdoor attacks are a malicious technique targeting machine learning models. Attackers inject specific patterns into training data, causing the model to produce predetermined erroneous behaviors when encountering inputs containing triggers, while maintaining correct functionality for normal inputs. The two most classic attack methods are BadNets and Blended, which we will briefly introduce next.

BadNets is a backdoor attack method against deep neural networks. Attackers manipulate model behavior by injecting malicious samples with specific triggers into the training data. When the model encounters inputs containing triggers, it produces erroneous outputs predetermined by the attacker, while maintaining normal behavior for regular inputs. This attack method highlights the potential security risks of deep learning systems when outsourcing training or using untrusted data sources.

Blended attacks are another, more covert backdoor attack method. Unlike BadNets, Blended attacks mix or fuse trigger patterns with normal images, making the triggers harder to detect by the naked eye. This method can achieve a balance between attack success rate and stealth by adjusting the blending ratio, making backdoors more difficult to discover using conventional defense methods.

3. Experimental Process

3.1. Experimental Background

The datasets used in our experiments include CIFAR-10, CIFAR-100, and Tiny ImageNet. The network models include PreActResNet18, ShuffleNet V2 (x1_0 standard width version), and VGG19. The attack methods include BadNets and Blended. The experimental environment consisted of a Tesla T4 GPU (16GB VRAM), equipped with an 8-core Intel Xeon processor (Skylake architecture, IBRS) and 56GB of memory. The software environment includes PyTorch 1.11.0, Python 3.8 (on Ubuntu 20.04 system), and CUDA 11.3. The system disk capacity is 30GB, and the data disk capacity is 50GB.

3.2. Attack Generalization

In our experiments, with a poisoning rate of 10%, we attacked different models on different datasets using both BadNets and Blended methods. For BadNets, we consistently constructed backdoor attack data with triggers in the bottom left corner. For Blended, we consistently superimposed 20% trigger patterns onto the attack data (highlighted in bold in the table). Figure 1 and Figure 2 show examples of poisoned data for BadNets and Blended attack methods respectively. By observing the ACC (model accuracy on normal test sets, reflecting model performance without attacks) and ASR (indicating the proportion of successful deceptions after the model is attacked) in the experimental results on test sets, we can analyze the generalization ability of the attacks.

DOI: 10.54254/2755-2721/88/20241677

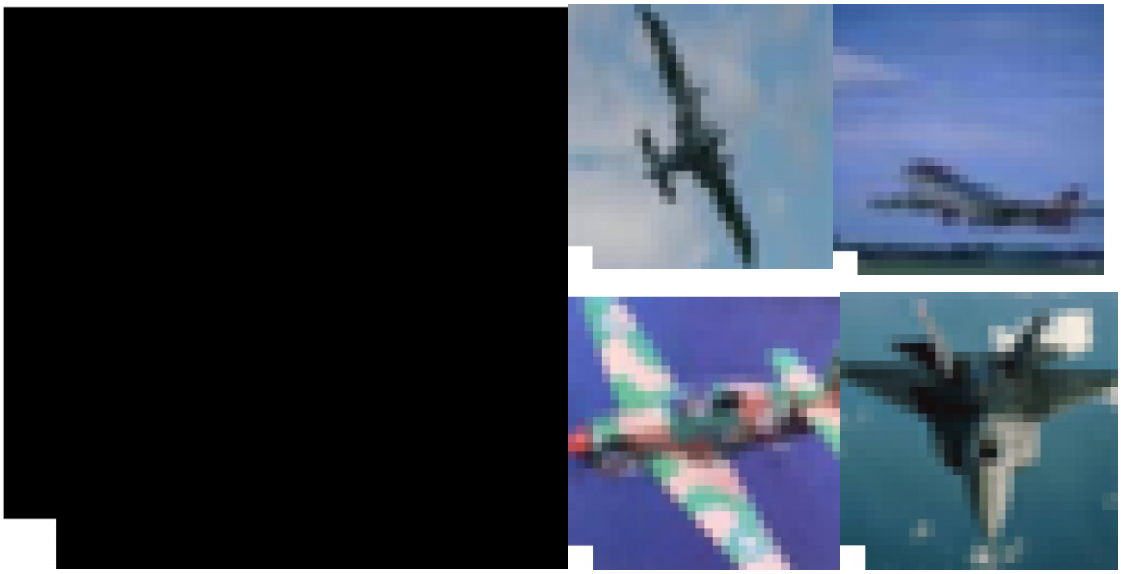

Figure 1. Trigger patterns and poisoned data samples used in BadNets attacks

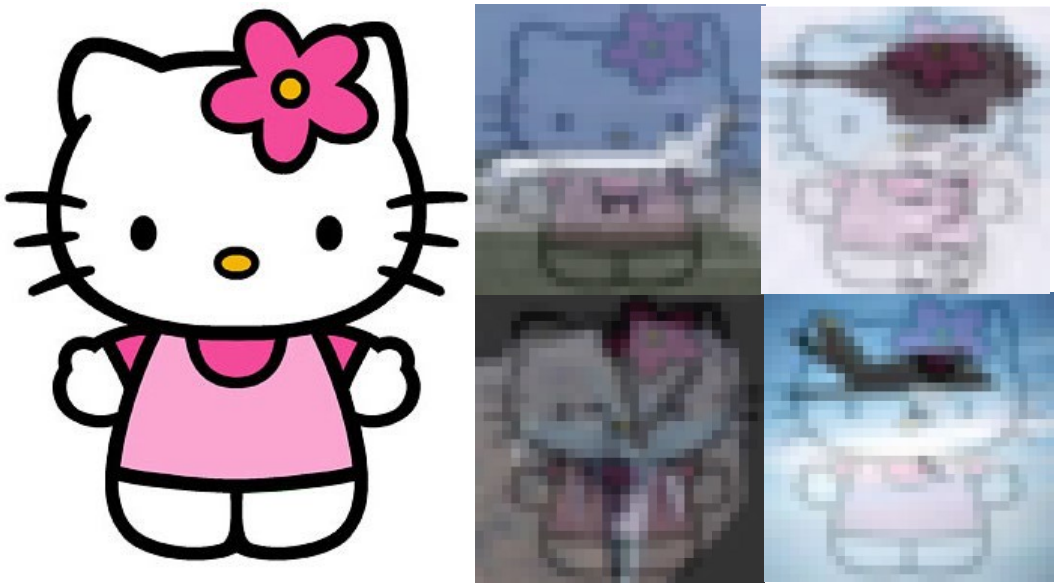

Figure 2. Trigger patterns and poisoned data samples used in Blended attacks

The row headers for Tables 1 to 10 show the ACC and ASR of the test set in conventional attack methods. The row headers for Tables 11 to 20 show the ACC and ASR of the test set in mixed training attack methods. The column headers for Tables 1 to 20 all indicate the position of triggers or the implantation ratio.

We used BadNets to attack the PreActResNet18 model on the CIFAR-10 dataset. The experimental results are shown in Table 1.

DOI: 10.54254/2755-2721/88/20241677

Table 1. Test ASR and Test ACC in PreActResNet18 and CIFAR-10

	Left Bottom	Left Top	Right Bottom	Right Top
ASR	**0.960111111**	0.0759	0.9393	0.0156
ACC	**0.9087**	0.9087	0.9087	0.9087

From the experimental results, we can observe that when the position of the trigger pattern changes, the success rate of BadNets attacks decreases to varying degrees. When the trigger position changes from the bottom left to the top right, the ASR (Attack Success Rate) drops by 94.45%.

3.3. Universality of Attack Ineffectiveness

We conducted experiments on different models and datasets while keeping other parameters unchanged. This included attacks on VGG19 (Table 2) and ShuffleNet V2 (Table 3) using the CIFAR-10 dataset, as well as attacks on PreActResNet18 using CIFAR-100 (Table 4) and Tiny ImageNet (Table 5) datasets. The results are as follows.

Table 2. Test ASR and Test ACC in VGG19 and CIFAR-10 of BadNets

	Left Bottom	Left Top	Right Bottom	Right Top
ASR	**0.9497**	0.0084	0.9578	0.0088
ACC	**0.8745**	0.8825	0.8758	0.8803

Table 3. Test ASR and Test ACC in ShuffleNet V2 and CIFAR-10 of BadNets

	Left Bottom	Left Top	Right Bottom	Right Top
ASR	**0.9156**	0.0270	0.9331	0.0233
ACC	**0.7546**	0.7456	0.7520	0.7486

Table 4. Test ASR and Test ACC in PreActResNet18 and CIFAR-100 of BadNets

	Left Bottom	Left Top	Right Bottom	Right Top
ASR	**0.9009**	0.0169	0.0126	0.0268
ACC	**0.6536**	0.6536	0.6536	0.6536

Table 5. Test ASR and Test ACC in PreActResNet18 and Tiny ImageNet of BadNets

	Left Bottom	Left Top	Right Bottom	Right Top
ASR	**0.9993**	0.0040	0.0051	0.0043
ACC	**0.4624**	0.4624	0.4624	0.4624

Additionally, we also switched to the Blended attack method, conducting attacks on PreActResNet18 (Table 6), VGG19 (Table 7), and ShuffleNet V2 (Table 8) using the CIFAR-10 dataset, as well as attacks on PreActResNet18 using CIFAR-100 (Table 9) and Tiny ImageNet (Table 10) datasets.

Table 6. Test ASR and Test ACC in PreActResNet18 and CIFAR-10 of Blended

	5%	10%	15%	**20%**
ASR	0.0119	0.1343	0.8516	**0.9962**
ACC	0.9153	0.9153	0.9153	**0.9153**

DOI: 10.54254/2755-2721/88/20241677

Table 7. Test ASR and Test ACC in VGG19 and CIFAR-10 of Blended

	5%	10%	15%	**20%**
ASR	0.0354	0.2489	0.8086	**0.9916**
ACC	0.8798	0.8802	0.8823	**0.8798**

Table 8. Test ASR and Test ACC in ShuffleNet V2 and CIFAR-10 of Blended

	5%	10%	15%	**20%**
ASR	0.1483	0.8068	0.9271	**0.9748**
ACC	0.7348	0.7380	0.7508	**0.7551**

Table 9. Test ASR and Test ACC in PreActResNet18 and CIFAR-100 of Blended

	5%	10%	15%	**20%**
ASR	0.0049	0.2708	0.8870	**0.9947**
ACC	0.6904	0.6904	0.6904	**0.6904**

Table 10. Test ASR and Test ACC in PreActResNet18 and Tiny ImageNet of Blended

	5%	10%	15%	**20%**
ASR	0.0091	0.3389	0.8615	**0.9901**
ACC	0.4592	0.4592	0.4592	**0.4592**

The experimental data demonstrates poor performance on test sets with trigger positions or transparencies different from those used in the training phase, except when test sets have trigger positions and transparencies consistent with those in the training phase. In the experiment attacking VGG19 on CIFAR-10, the ASR of the test set decreased by 95.62%. This clearly demonstrates that the problem of poor generalization in backdoor attacks is prevalent across different models and datasets.

3.4. Improvement Method

To address the aforementioned issues, we propose a mixed training method for backdoor attacks to improve the poor generalization of attacks when the training and testing triggers are inconsistent. Specifically, we enhance the attack success rate by constructing backdoor attack data with triggers in different positions or proportions during the training phase. By applying this mixed training backdoor attack method to BadNets and Blended attack methods across different models and datasets, we achieved maximum improvements in ASR of 99.60% and 99.37% for BadNets and Blended test sets, respectively.

Our proposed mixed training backdoor attack method improvement method involves inputting images with triggers in different positions or transparencies as poisoned data during the training phase of BadNets and Blended attacks. For example, in BadNets, with a 10% poisoning rate, we input 2.5% each of poisoned data with triggers in the bottom left, top left, bottom right, and top right positions during the training phase. In Blended, with a 10% poisoning rate, we input 2.5% each of poisoned data with trigger transparencies of 5%, 10%, 15%, and 20% during the training phase. We conducted experiments on different models and datasets, with results shown in Tables 10 to 20.

DOI: 10.54254/2755-2721/88/20241677

Table 11. Test ASR and Test ACC in PreActResNet18 and CIFAR-10 of BadNets in combined training

	Left Bottom	Left Top	Right Bottom	Right Top
ASR	**0.9600**	0.9288	0.9609	0.9283
ACC	**0.8997**	0.8997	0.8997	0.8997

Table 12. Test ASR and Test ACC in VGG19 and CIFAR-10 of BadNets in combined training

	Left Bottom	Left Top	Right Bottom	Right Top
ASR	**0.9440**	0.8828	0.9496	0.9131
ACC	**0.8810**	0.8827	0.8823	0.8813

Table 13. Test ASR and Test ACC in ShuffleNet V2 and CIFAR-10 of BadNets in combined training

	Left Bottom	Left Top	Right Bottom	Right Top
ASR	**0.8564**	0.7907	0.8612	0.7502
ACC	**0.7541**	0.7541	0.7541	0.7541

Table 14. Test ASR and Test ACC in PreActResNet18 and CIFAR-100 of BadNets in combined training

	Left Bottom	Left Top	Right Bottom	Right Top
ASR	**0.8811**	0.8296	0.9036	0.8561
ACC	**0.6376**	0.6376	0.6376	0.6376

Table 15. Test ASR and Test ACC in PreActResNet18 and Tiny ImageNet of BadNets in combined training

	Left Bottom	Left Top	Right Bottom	Right Top
ASR	**0.9886**	1.0000	0.9998	1.0000
ACC	**0.4587**	0.4587	0.4587	0.4587

Table 16. Test ASR and Test ACC in PreActResNet18 and CIFAR-10 of Blended in combined training

	5%	10%	15%	**20%**
test_asr	0.6370	0.9721	0.9974	**0.9999**
test_acc	0.9084	0.9084	0.9084	**0.9084**

Table 17. Test ASR and Test ACC in VGG19 and CIFAR-10 of Blended in combined training

	5%	10%	15%	**20%**
ASR	0.9850	0.8714	0.9886	**0.9994**
ACC	0.8763	0.8748	0.8696	**0.8692**

Table 18. Test ASR and Test ACC in ShuffleNet V2 and CIFAR-10 of Blended in combined training

	5%	10%	15%	**20%**
ASR	0.2081	0.7001	0.9232	**0.9834**
ACC	0.7462	0.7462	0.7462	**0.7462**

DOI: 10.54254/2755-2721/88/20241677

Table 19. Test ASR and Test ACC in PreActResNet18 and CIFAR-100 of Blended in combined training

	5%	10%	15%	**20%**
ASR	0.6211	0.9492	0.9917	**0.9987**
ACC	0.6864	0.6864	0.6864	**0.6864**

Table 20. Test ASR and Test ACC in PreActResNet18 and Tiny ImageNet of Blended in combined training

	5%	10%	15%	**20%**
ASR	0.5456	0.9177	0.9864	**0.9965**
ACC	0.4638	0.4638	0.4638	**0.4638**

The experimental data demonstrates that our proposed mixed training attack method improves attack success rates across different models and datasets. The optimal improvement was 57.88% for the attack on the PreActResNet18 model using the Tiny ImageNet dataset.

4. Conclusion

In our experiments, we found that the problem of poor attack generalization exists across various datasets and models. To address this issue, we proposed an improved mixed training backdoor attack method for BadNets and Blended attack scenarios.

In our experiments, we used triggers in different positions for BadNets and triggers with different transparencies for Blended, both with a 10% poisoning rate during training. We evaluated the effectiveness of these attack methods on multiple models (such as PreActResNet18, VGG19) and datasets (such as CIFAR-10, CIFAR-100). Compared to the conventional BadNets algorithm, in our attack on PreActResNet18 using the CIFAR-100 dataset, our method achieved an optimal improvement in attack success rate of 89.10% over existing attack algorithms.

Although changes in trigger position lead to poor attack generalization, the threat remains significant. In conclusion, even when using triggers with different positions or transparencies in training and testing, attackers can still successfully trigger backdoor attacks under these inconsistent conditions through the mixed training backdoor attack method. The mixed training backdoor attack method significantly improves the success rate of backdoor attacks, indicating that existing defense mechanisms may be insufficient to cope with diverse attack conditions. Further research and development of more robust defense techniques are needed to protect the security of these critical systems.

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] 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.

[3] Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto, 1(4), 7.

[4] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity mappings in deep residual networks. In European conference on computer vision (pp. 630-645). Springer, Cham.

[5] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[6] Ma, N., Zhang, X., Zheng, H. T., & Sun, J. (2018). Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European conference on computer vision (ECCV) (pp. 116-131).

[7] Le, Y., & Yang, X. (2015). Tiny imagenet visual recognition challenge. CS 231N, 7(7), 3.