Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 12, No. 2, 2024

Self-supervised learning backdoor defense mixed with self-attention mechanism

Guotao Yuan [1], Hong Huang [1, 2, *], Xin Li [1]

[1] School of Computer Science and Engineering, Sichuan University of Science & Engineering, Yibin 644000, China
[2] Key Laboratory of Bridge Nondestructive Testing and Engineering Calculation, school of Sichuan University, Yibin 644000, China
[*] **Corresponding author**: Hong Huang (Email: huanghong@suse.edu.cn)

Abstract: Recent studies have shown that Deep Neural Networks (DNNs) are vulnerable to backdoor attacks, where attackers embed hidden backdoors into the DNN models by poisoning a small number of training samples. The attacked models perform normally on benign samples, but when the backdoor is activated, their prediction results will be maliciously altered. To address the issues of suboptimal backdoor defense effectiveness and limited generality, a hybrid self-attention mechanism-based self-supervised learning method for backdoor defense is proposed. This method defends against backdoor attacks by leveraging the attack characteristics of backdoor threats, aiming to mitigate their impact. It adopts a decoupling approach, disconnecting the association between poisoned samples and target labels, and enhances the connection between feature labels and clean labels by optimizing the feature extractor. Experimental results on CIFAR-10 and CIFAR-100 datasets show that this method performs moderately in terms of Clean Accuracy (CA), ranking at the median level. However, it achieves significant effectiveness in reducing the Attack Success Rate (ASR), especially against BadNets and Blended attacks, where its defense capability is notably superior to other methods, with attack success rates below 2%.

Keywords: Self-supervised; Self-attention; Backdoor defense.

1. Introduction

In the context of backdoor attacks, attackers embed a backdoor trigger into some harmless training images and alter the labels of these images to a target label set by the attackers. During this process, the model is trained to recognize the connection between these backdoor triggers and the target labels [1]. Thus, during model inference, if an image containing the trigger is encountered, the model will classify the image as the label specified by the attacker, while it behaves normally for harmless samples. The covert nature of this attack method makes it difficult for users to detect the hidden backdoor in the model, posing a serious threat to the model's security.

According to research by Huang et al. [2], the backdoor is embedded in the feature space, meaning that samples with embedded backdoor triggers tend to cluster together in the feature space. This phenomenon is mainly due to the end-to-end supervised training mode. Specifically, the excessive learning capability allows the training model to learn the features of the backdoor trigger, can reduce the distance between poisoned samples in the feature space, and connects the learned trigger-related features to the target label through end-to-end supervised training [3].

Based on the above insights, this chapter decouples the training process of the task model, breaking the connection between the trigger and the target label to achieve a defensive effect [4]. Specifically, the pretrained model is considered as two disjoint parts, including a feature extractor and a simple classifier. In the feature extractor part, this chapter first uses a set of unlabeled training samples to learn a feature extractor through self-supervised learning and introduces a self-attention mechanism for optimization within the extractor. Then, based on the learned feature extractor and all training samples, a simple classifier is learned through the standard supervised training process. The use of strong data augmentation in self-supervised learning may have a positive effect on defense against backdoor attacks, as it can weaken or disrupt the backdoor trigger patterns, making it more difficult for attackers to implant a backdoor. Along with the decoupling process, this further blocks the correspondence between the trigger and the target label. Therefore, even if training the model on a poisoned dataset, the hidden backdoor cannot be successfully created.

Furthermore, according to findings by Kolesnikov et al.[5], there is a significant difference between the feature representations learned by the feature extractor generated using self-supervised learning methods and those learned using standard supervised learning methods. Specifically, poisoned samples are closer to their true label values rather than the target label values set by the attacker in the feature space. This phenomenon makes the training of the simple classifier akin to learning with label noise. Therefore, this chapter first filters out highly credible training samples (i.e., those most likely to be benign), then uses these samples as labeled samples, with the remaining part forming unlabeled samples, and fine-tunes the entire model through semi-supervised learning, maximizing the benign rate of the training data and enhancing the model's security. The contributions and innovations of this chapter's work include:

1). Introducing a self-attention mechanism. The incorporation of self-attention modules in the convolutional module significantly enhances the model's learning and representation capability for uncontaminated samples. By focusing on key features, the self-attention mechanism not only optimizes the accuracy and robustness of the model in processing clean samples but also significantly improves overall performance.

2). Adopting a symmetric cross-entropy strategy. This method effectively filters out highly credible training samples, significantly enhancing the model's defensive capability. Symmetric cross-entropy can effectively identify and exclude

potential harmful samples, reducing errors introduced by learning noise, thus improving the model's robustness and generalizability.

3). Experimental validation on standard benchmark datasets. Through extensive experiments on numerous classic datasets, we verified the effectiveness and reliability of the proposed defense strategy. These experiments not only comprehensively assess the performance and robustness of the defense method but also provide solid support and assurance for the deployment of the model in practical applications.

2. Preliminaries

In this chapter, we detailedly introduce a comprehensive defense strategy aimed at enhancing the model's robustness against backdoor attacks. As illustrated in Figure 1, this strategy is divided into three main phases [6]: self-supervised learning training of the feature extractor, the filtering process for highly credible samples, and the fine-tuning and retraining of the model. Through this series of defense steps, the performance of the model in normal usage scenarios is not only improved but its security and robustness against backdoor attacks are also significantly enhanced. This multi-phase defense mechanism provides an effective strategy for protecting models from potential backdoor threats.

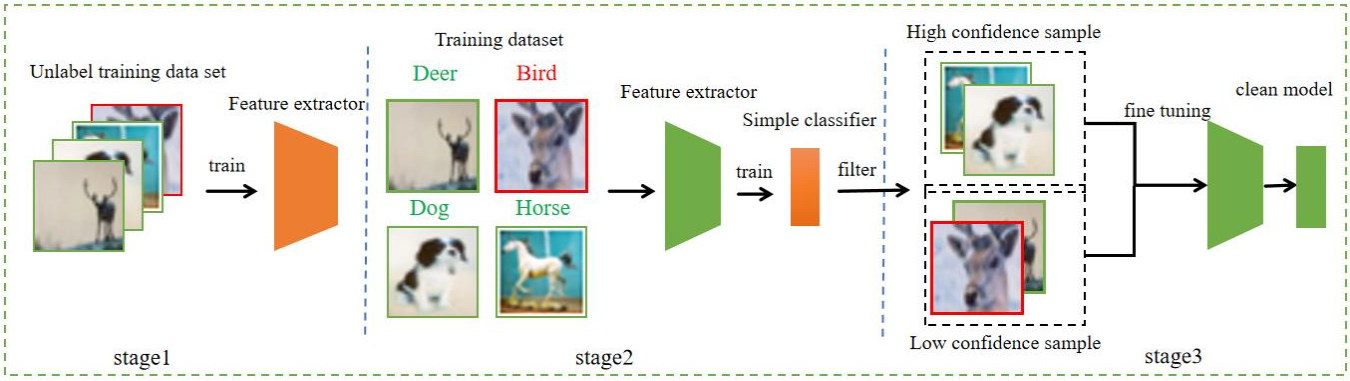

Figure 1. Defense procedure

In the self-supervised learning training phase of the feature extractor, the essence is to train a potent feature extractor through self-supervised learning methods. Self-supervised learning, as a cutting-edge unsupervised learning technology, is designed to allow the model to learn data representations through intrinsic data features, without depending on external annotation information. During this phase, the feature extractor is trained to extract distinctive features from the input data, crucial for discerning the subtle differences between clean and poisoned samples. With self-supervised learning, the model significantly enhances its ability to recognize normal data, thereby bolstering its overall defensive capability.

Subsequently, the process of filtering highly credible samples involves utilizing the aforementioned feature extractor to identify and select samples of high credibility. These samples are instances that bear a high resemblance to normal data and are unlikely to be affected by backdoor attacks. By precisely choosing these samples as the foundation for training and testing, the adverse impact of poisoned samples on model performance can be somewhat reduced, thus enhancing the model's stability and reliability in facing complex attack scenarios.

Lastly, the model undergoes a fine-tuning and retraining phase, employing the filtered highly credible samples for detailed fine-tuning training, aimed at further optimizing the model's performance on new datasets [7]. The objective of this phase is to improve the model's classification accuracy and robustness against adversarial attacks, ensuring the model can effectively identify and withstand potential triggers implanted by backdoor attacks. The fine-tuned model not only can more accurately classify data but also enhances its defense against unknown attack patterns.

2.1. Combine self-attention module and convolution module

Traditional convolutional operations primarily focus on extracting local feature information while ignoring the global contextual information in images. This local preference may introduce biases in complex image understanding tasks, limiting the overall performance of models. To overcome these limitations and enhance the model's global perceptual ability, some researchers [8-10] have explored an innovative approach. This approach is based on treating each pixel in the image feature map as an independent random variable and calculating the pairwise covariance between these pixels. Through this method, the model can adjust the value of each pixel to increase or decrease the contribution of a certain predicted pixel to the overall prediction. Specifically, by considering the relationships between all pixels, this method allows the model to learn not only local features but also to capture and utilize global information in the image. This covariance-based adjustment method helps significantly expand the model's receptive field range, i.e., the size of the image area the model can perceive. Enlarging the receptive field allows the model to better understand the complex structures and relationships in the image, thereby improving the accuracy of image classification and the sensitivity of the model to details. By effectively combining local features and global context, this method provides a powerful tool for enhancing the model's image recognition capabilities.

This section also explores new directions, introducing a self-attention module into the convolutional module to enhance the feature extractor's capabilities [11], particularly in capturing global dependencies and enhancing the model's focus on key features. In this way, significant improvements can be achieved in the model's performance when handling complex image data, especially in understanding the deep

structure and contextual information of the images. This is illustrated in Figure 2. First, define the input feature map as $x \in \mathbb{R}^{C \times H \times W}$, where $H \times W \times C$ respectively represents the height, width, and number of channels of the image. This fusion module is carefully designed with two main stages to enhance its functionality. In the first stage, input features are finely reconstructed through three 1×1 convolutions, where $f \cdot$ represents three 1×1 convolution operations, thereby generating rich $3 \times N$ feature maps. Moving to the second stage, these intermediate features x' are subdivided into N groups, each consisting of three independent parts corresponding to each 1×1 convolution. At this point, by performing shift operations and aggregation, the module is able to extract information from local receptive fields like traditional convolutional methods. Meanwhile, a self-attention module introduces an advanced concept by analyzing the relationships between elements within the input sequence to assign specific weights to each element, thereby deeply exploring the global context of the sequence. In this mechanism, each element in the sequence interacts with other elements, computing weight coefficients, which are then used for weighted summation to form a comprehensive output representation for each element. The output representation of each element Z can be calculated using equation (1).

$$Z = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Here, Q represents the query, K represents the key, and V represents the value. Since both modules share the same 1×1 convolutional steps, this allows the system to apply intermediate features to different aggregation processes through a single projection operation. Ultimately, merging the output results of these two stages fully leverages the advantages of traditional convolutional structures while significantly enhancing the overall performance and robustness of the model through the introduction of the self-attention mechanism.

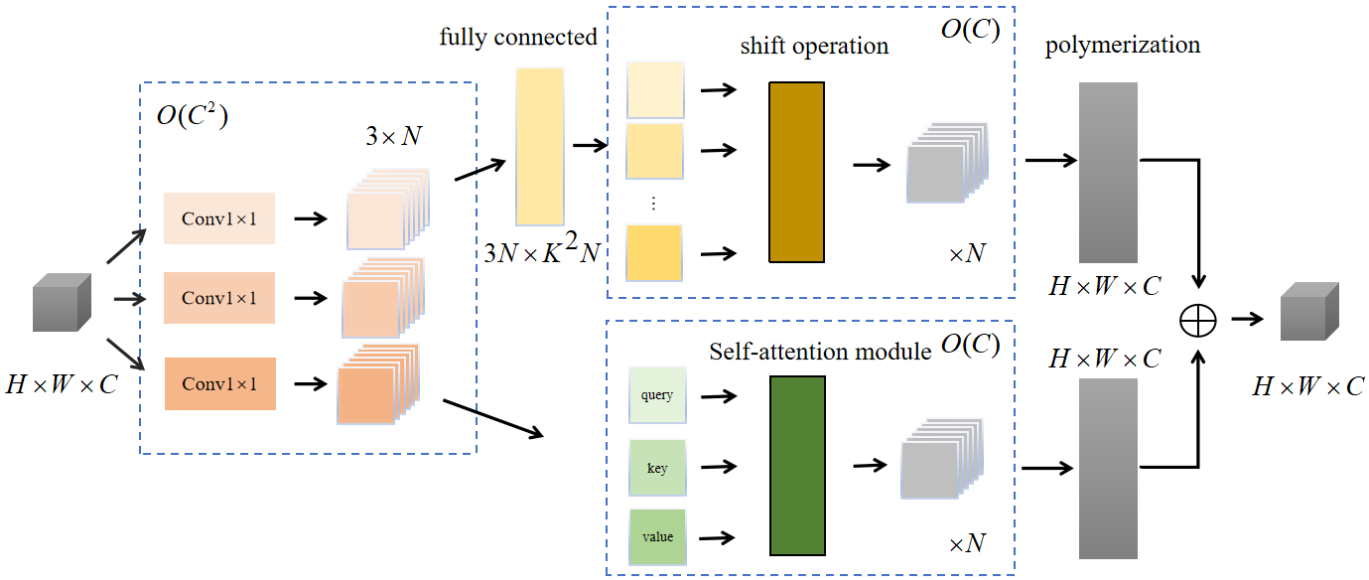

Figure 2. Combination of self-attention module and convolutional module

When constructing the feature extractor, local features of the input image are initially extracted through the convolutional module[12]. Subsequently, the self-attention module further processes these features, enabling the model to capture global dependencies and enhance the representation capability of the features. The training objective of the feature extractor, which combines the self-attention module and the convolutional module, is to minimize the difference between predictions and actual labels. Specifically, when the input image is x, feature extraction is performed through a network N that integrates both self-attention and convolutional modules. This network comprises convolutional parameters θ_c and self-attention parameters θ_a, with its output $\hat{y}$ predicting the labels of the image. The objective is to minimize the loss L between the network predictions and the actual labels y during the training process, which can be represented by equation (2).

$$(\theta_c^*, \theta_a^*) = \arg\min_{(\theta_c, \theta_a)} \sum_{(x,y) \in D} L(N(x; \theta_c, \theta_a), y) \tag{2}$$

Wherein, D represents the training dataset, L typically denotes the loss function, and θ_c^* and θ_a^* respectively denote the optimal convolutional and self-attention parameters. Through gradient descent, iterations are conducted to update θ_c and θ_a in order to minimize the loss.

2.2. Benign samples were filtered by symmetric cross-entropy

After adopting the decoupled training strategy, even when trained on a poisoned dataset, the model can effectively prevent backdoor implantation by purifying the feature

extractor[13]. However, this approach presents two significant issues. Firstly, since the parameters of the feature extractor in the second stage are frozen, the model's accuracy in predicting benign samples may decrease, especially compared to samples trained with fully supervised learning. Secondly, when the model faces attacks with poisoned labels, these toxic samples may be misclassified as outliers, thus introducing interference during the learning process in the second stage. This interference occurs because some samples tend to cluster around their true label values rather than being distributed in the more abstract and latent feature space defined by the trained feature extractor. This phenomenon makes the process of training simple classifiers somewhat akin to learning in a noisy label data environment. Due to the introduction of these noisy labels, training a precise network model in such an environment becomes exceedingly difficult and challenging.

To identify and filter out interfered samples, this chapter employs both Cross Entropy (CE) and Symmetric Cross Entropy (SCE) loss functions, as shown in Figure 3. Experiments conducted on the CIFAR-10 dataset investigate the model's performance when facing BadNets attacks with a poisoning rate of 20%, using these two loss functions. The experimental results demonstrate that compared to the traditional cross-entropy method, symmetric cross-entropy can more effectively increase the loss of samples affected by poisoning, thereby creating a larger loss difference between poisoned and clean samples.

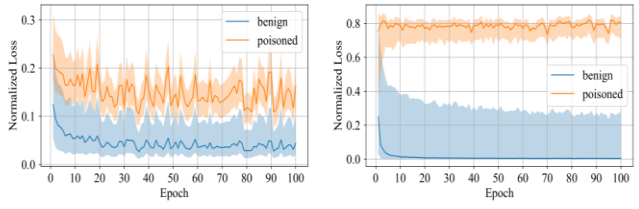

Figure 3 Sample distribution of cross-entropy and symmetric cross-entropy filtering

After obtaining the feature extractor in the first stage, it is frozen, and the remaining parameters are further optimized using D_t, as shown in equation (3).

$$\omega_f^* = \arg\min_{\omega_f} \sum_{(x,y)\in D_t} L_2(f_{[\omega_c^*,\omega_f]}(x), y) \quad (3)$$

Where ω_f^* represents a simple classifier, and L_2 represents the loss function of supervised learning.

Based on the loss $L_2(\cdot;[\omega_c^*, \omega_f^*])$, high-confidence samples are defined as the $\alpha\%$ training samples with the smallest loss, where $\alpha \in [0, 100]$ is a hyperparameter. This chapter adopts symmetric cross-entropy as the loss function to increase the difference between poisoned and clean samples, further reducing the bias issues that may arise during the retraining process.

2.3. Model fine-tuning training

After filtering out the samples, the training set D_t will be separated into two disjoint parts, including the high-confidence dataset D_h and the low-confidence dataset D_l, where the unlabeled version of $D_l \triangleq D_t \setminus D_h$ is represented as $\hat{D}_l \triangleq \{x \mid (x,y) \in D_l\}$. In this separation process, this chapter focuses on the unlabeled version in the low-confidence dataset D_l, as this portion of data may contain some potentially useful information but requires special handling due to its low confidence.

To leverage the useful information contained in $\hat{D}_l \triangleq \{x \mid (x,y) \in D_l\}$ and mitigate the potential side effects of poisoned samples, this chapter adopts semi-supervised learning to fine-tune the entire training model $f_{[\omega_c^*, \omega_f^*]}(\cdot)$. This fine-tuning process is represented by equation (4).

$$\min_{\omega} L_3(D_h, \hat{D}_l; \omega) \quad (4)$$

Where L_3 represents the loss of semi-supervised learning. This process optimizes the collaborative learning of the feature extractor and classifier, promoting their mutual adaptability, thereby enhancing the overall robustness and performance of the model. The significant advantage of this strategy lies in its utilization of a large amount of unlabeled and low-confidence data, which not only maintains the core performance of the model but also enhances its reliability and generalization ability in practical applications.

3. Experiments Settings

3.1. Data set and model structure

This chapter conducts experimental evaluations on the CIFAR-10 and CIFAR-100 datasets. CIFAR-10 comprises 50,000 training samples and 10,000 testing samples, with each sample being a 32×32-pixel image belonging to one of 10 categories. CIFAR-100, an extended version, contains 100 categories, with each category having 600 32x32-pixel color images. Compared to CIFAR-10, CIFAR-100 presents higher classification difficulty due to the increased number of categories.

Table 1. Model structure

Layer	Output Shape	Specification
Conv_1	[1,3,32,32]	Conv2d(3x3, stride=1, padding=1) + BatchNorm2d + ReLU
Sa_1	[1,3,32,32]	SelfAttention
Res_1	[1,64,8,8]	[Conv2d(3x3) + BatchNorm2d + ReLU + SelfAttention] x 3
Res _2	[1,128,8,8]	[Conv2d(3x3, stride=2) + BatchNorm2d + ReLU + SelfAttention] x 4
Res _3	[1,256,4,4]	[Conv2d(3x3, stride=2) + BatchNorm2d + ReLU + SelfAttention] x 6
Res _4	[1,512,2,2]	[Conv2d(3x3, stride=2) + BatchNorm2d + ReLU + SelfAttention] x 3
Avgpool_1	[1,512,1,1]	AdamAvgPool2d
FC	[1,10]	Softmax

Table 2. Model parameters

Parameters	Value
Optimizer	Adam
Learning Rate	0.005
Loss Function	Categorical_CrossEntropy
Epoch	100
Batch Size	64
Activation Function	ReLU

In terms of model architecture, ResNet34 is employed as the network architecture to train the feature extractor and classifier. Detailed model structure and parameter configurations are provided in Table 1 and Table 2. In this model structure, Sa_1 indicates the incorporation of the self-attention module; Res_1 to Res_4 represent the four main stages of the ResNet34 structure, with each stage composed of several residual units and integrated with self-attention modules to enhance the model's processing capabilities; Avgpool_1 denotes the global average pooling layer, responsible for reducing feature dimensions; and FC represents the fully connected layer at the end, used for outputting classification results. By integrating the self-attention mechanism into the classical residual network, the aim is to further enhance the model's ability to capture complex image features and optimize classification performance.

3.2. Attack baseline and attack Settings

In this chapter, three different types of backdoor attacks were selected for in-depth effectiveness evaluation and comparative analysis: BadNet, Blended, and adversarial perturbation-based backdoor attacks.

Firstly, BadNet backdoor attack is introduced, which embeds a 3×3 square trigger into clean samples to create poisoned samples, as shown in Figure 4. This type of attack disrupts the model's classification process by exploiting triggers, leading to misclassification under specific conditions.

The second type is Blended attack, where the strategy proposed by Chen et al. [14] is adopted, using "Hello Kitty" as the trigger. This method combines more covert attack techniques, capable of evading standard defenses and profoundly impacting the model.

The third type of attack is based on adversarial perturbations, where the triggers described in Chapter 3 are utilized for verification. This attack method modifies clean samples by applying subtle perturbations, aiming to implant backdoors without raising human attention.

In subsequent experiments, 1500 mixed samples were used, with the poisoning ratio set to 30%, ensuring accuracy and consistency of the experimental data. Through comprehensive evaluation of these three distinct backdoor attack methods, the study not only deepens understanding of their respective characteristics but also reveals their strengths and limitations, providing valuable insights and directions for developing future defense mechanisms.

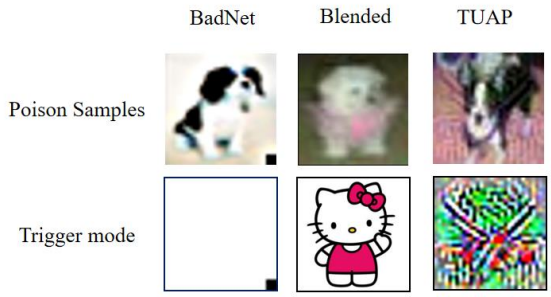

Figure 4. Representation of toxic samples and triggers on the CIFAR-10 dataset

3.3. Defense baseline and defense Settings

In this chapter, we have selected two relatively advanced defense methods for experimental result comparison: Nerve Clearing (NC) and Neural Attention Distillation (NAD). Additionally, the experimental results without backdoor defenses serve as an important reference baseline. The primary goal is to explore effective and universal defense strategies to counter the threats of backdoor attacks on deep learning models. A decoupled approach is also employed to resist backdoor attacks, with the entire defense process divided into three stages.

Firstly, in the initial phase, a feature extractor that combines a self-attention module with a convolution module is used for self-supervised learning. The ResNet32[15] network structure is adopted, and the backbone network is trained 100 times to extract representative features. The aim of this step is to train the feature extractor through a self-supervised learning method, thereby enabling the model to better understand the input data and enhance its representation capability of the data. In the middle stage of defense, an Adam optimizer with a learning rate of 0.005 is used, and the batch size is set to 64. Relying on the symmetric cross-entropy loss function, training is conducted on 10 fully connected layers, while filtering out more than 50% of the high-confidence samples for in-depth learning. The key in this stage lies in the selection of high-confidence samples, with the aim of minimizing the potential impact of backdoor attacks while ensuring the model's accurate judgment on clean samples. The final stage of defense focuses on the fine-tuning of model parameters. Through meticulous adjustment of the model parameters, the learning strategy of the model is optimized to better suit the specific task's needs. As shown in Table 3, the selected high-confidence samples are marked and used for fine-tuning training. This step of parameter adjustment helps enhance the model's robustness and generalization, boosting its capability to resist backdoor attacks.

Table 3. Parameters for fine-tuning the model

Parameters	Value
Optimizer	Adam
Learning Rate	0.002
Loss Function	Symmetric Cross-Entropy
Epoch	120
Batch Size	64
Activation Function	Somax

4. Experiment results

4.1. Evaluation index

Based on the characteristics of backdoor models, there are typically two evaluation metrics: Attack Success Rate (ASR)

and Clean Accuracy (CA). ASR is used to measure the backdoor model's accuracy in recognizing backdoor data. Here, accuracy refers to the ability to correctly identify backdoor images as the target label, rather than their true label. That is, the frequency with which backdoor samples are accurately identified as the target label serves as the numerator, and the frequency of identifying all backdoor test data serves as the denominator. The percentage obtained by dividing these two values assesses the model's effectiveness in executing the backdoor attack.

Clean Accuracy (CA), on the other hand, measures the backdoor model's accuracy in recognizing clean data. Here, recognition accuracy refers to the ability to correctly identify clean samples as their true labels. That is, the frequency with which clean samples are accurately identified as their true labels serves as the numerator, and the total number of clean test samples identified serves as the denominator. The percentage obtained by dividing these two values evaluates the model's integrity in handling unaffected data.

The mathematical expressions for ASR and CA are depicted as Formula (5) and Formula (6), respectively.

$$ASR = \frac{1}{N}\sum_{i=1}^{N}\prod(f(\tilde{x}) = \tilde{y}) \tag{5}$$

$$CA = \frac{1}{N}\sum_{i=1}^{N}\prod(f(x) = y) \tag{6}$$

Herein, $\tilde{x}$ represents poisoned samples, $\tilde{y}$ denotes target labels, x stands for clean samples, and y refers to initial labels.

When evaluating the effectiveness of defense strategies, ASR and CA serve as the principal evaluation metrics. These metrics not only reflect the model's sensitivity to backdoor attacks but also demonstrate the model's capability in processing normal data. The experiment will calculate and compare the ASR and CA values obtained from different defense methods under a unified dataset and attack scenario. Such comparisons allow for a comprehensive understanding of the effectiveness and applicability of various defense strategies. Furthermore, the experiment will detail the calculation processes of these metrics and validate the effectiveness of different defense strategies in improving Clean Accuracy and reducing Attack Success Rate. This in-depth evaluation aims to assess their feasibility and practical value in real-world applications.

4.2. Comparison of experimental results

This chapter compares defense methods NC (Nerve Clearing) and NAD (Neural Attention Distillation) on the CIFAR-10 and CIFAR-100 datasets, aiming to comprehensively assess the performance of various defense strategies across different datasets and attack scenarios. The results, as shown in Table 4, demonstrate the effectiveness of each defense method in mitigating backdoor attacks. Compared to having no defense, all three defense approaches demonstrated certain resistance capabilities, effectively reducing the harm from backdoor attacks. In the case of BadNets attacks, the model without defense was almost completely compromised, with an attack success rate reaching 100%. This indicates that models trained without any defense methods are vulnerable to security threats. The use of NC and NAD defenses significantly reduced the threat from BadNets attacks. Notably, NAD achieved its effect by lowering the CA metric, while the defense method discussed in this chapter not only surpassed the former in defense effectiveness but also maintained a CA metric only 1% lower than the standard model.

In the case of Blended attacks, which employ triggers of the same pixel size as clean samples and control their stealthiness, these attacks are more covert and less detectable compared to BadNets attacks. Therefore, the experimental results showed that the ASR for Blended attacks was higher than for BadNets attacks, with NAD again reducing the ASR at the expense of the CA metric. In the TUAP (Targeted Universal Adversarial Perturbations) attack tests, it was found that the CIFAR-100 dataset, with its broader content and more diverse image categories, resulted in more complex triggers being generated, thereby reducing the effectiveness of defense methods against such a dataset, leading to higher ASRs compared to CIFAR-10. The performance of NAD and NC methods in these tests was similar to previous findings. However, the defense method introduced in this chapter showed more advantage in the experiments. It demonstrated a clean sample prediction accuracy on the CA metric comparable to that of a clean model, and achieved a significant reduction in ASR compared to the first two defenses, with an attack success rate of less than 2%. These results further validate the effectiveness and practicality of the defense method proposed in this chapter.

Table 4. Evaluation of the effectiveness of various defense methods

	CIFAR-10						CIFAR-100					
	BadNets		Blended		TUAP		BadNets		Blended		TUAP	
	CA	ASR	CA	ASR	CA	ASR	CA	ASR	CA	ASR	CA	ASR
Non-Defence	94.82	100	94.24	98.25	88.75	98.82	92.24	99.11	93.18	98.80	84.40	95.27
NC	94.25	2.33	93.01	5.70	87.84	4.32	92.18	8.89	92.44	11.74	85.26	7.89
NAD	90.47	1.20	89.72	1.51	85.20	0.18	89.74	3.66	90.18	5.68	81.72	1.22
Ours	93.81	0.55	92.97	0.97	88.20	0.21	90.87	1.54	91.13	1.35	84.85	0.63

4.3. Ablation experiment

To deeply assess the role of each step in the overall effectiveness of the defense method proposed in this chapter, a series of ablation experiments were conducted. These experiments aimed to individually examine the importance and effectiveness of each stage. Specifically, this section implemented ablation studies on the CIFAR-10 dataset to provide a detailed analysis and comparison of the various components of the defense strategy. The three variants included:

Defense without the self-attention module(N-Sa): In this variant, the self-attention module is not added during the training of the feature extractor, and traditional self-

supervised learning methods are used instead. All other aspects of the experiment remain unchanged to isolate and assess the contribution of the self-attention module to the defense's effectiveness.

Defense with cross-entropy(CE): In this variant, the generated feature extractor is first frozen, and a simple classifier is trained using this feature extractor. Then, high-confidence samples are filtered out across all training samples using cross-entropy loss, and these filtered samples are subsequently used for model fine-tuning to obtain the experimental results.

Defense with symmetric cross-entropy(SCE): Similar to the second variant, this variant also filters out high-confidence samples but utilizes symmetric cross-entropy for this purpose.

Table 5. Self-supervised learning defense methods with mixed self-attention mechanisms and their variants

	BadNets		Blended		TUAP	
	CA	ASR	CA	ASR	CA	ASR
No defence	94.82	100	94.24	98.25	88.75	98.82
N-Sa	92.41	0.96	92.08	1.73	87.20	2.23
CE	82.25	5.20	81.85	12.19	68.20	13.60
SCE	82.34	5.12	82.30	6.24	66.07	10.71
Ours	93.81	0.55	92.97	0.97	88.20	0.21

The experimental results, as shown in Table 5, compare the defense method proposed in this chapter against models trained without any defense. It was found that decoupling the original end-to-end supervised training process can effectively prevent the creation of hidden backdoors. Specifically, compared to models without any defense, the defense method proposed in this experiment showed a significant advantage in preventing backdoor attacks. At the same time, comparisons of the three different variants also revealed the importance and effectiveness of each step in the defense method.

1.Compared to the defense without the self-attention module, the introduction of the self-attention module slightly increased the CA value, validating the positive impact of incorporating self-attention modules into convolutional layers on the recognition rate of clean samples. The self-attention module helps the model to better focus on important features within images, thereby improving the model's recognition performance.

2.Comparing the second and third variants, i.e., the defenses using cross-entropy loss and symmetric cross-entropy loss, validated the effectiveness of symmetric cross-entropy loss in defending against poison label backdoor attacks. The experimental results showed that the defense method utilizing symmetric cross-entropy loss performed better in preventing backdoor attacks, with higher CA values and lower ASR values. This indicates that symmetric cross-entropy loss can more effectively eliminate low-confidence samples, thus enhancing the overall performance of the model.

3.Compared to the third variant, the fourth variant showed relatively lower ASR and CA values. This phenomenon is due to the fourth variant eliminating low-confidence samples, thereby reducing the impact of backdoor attacks. This further proves that in defending against backdoor attacks, minimizing the side effects of low-confidence samples while utilizing their useful information is crucial. Therefore, the results of this experiment highlight the effectiveness of defense methods such as decoupling the training process, introducing self-attention modules, and adopting symmetric cross-entropy loss. These findings provide important references for further improving defense against backdoor attacks.

4.4. Research Summaries

We comparative and ablation experiments were conducted to evaluate the proposed defense method on the CIFAR-10 and CIFAR-100 datasets, comparing it with other defense approaches. The results indicate that the proposed defense method can effectively mitigate the harm of backdoor attacks, demonstrating significant advantages in defense effectiveness. Specifically, the defense method introduced in this chapter shows a moderate level of performance in terms of the CA metric, meaning it achieves a median level of accuracy in predicting clean samples, but it achieves remarkable results in ASR. Especially when facing BadNets and Blended attacks, the defense method significantly outperforms other defense approaches, with an attack success rate of less than 2%. These outcomes further verify the effectiveness and reliability of the proposed method, suggesting that it can effectively sever the connection between poisoned samples and their target labels, thereby preventing the creation of backdoors.

Additionally, the ablation experiments analyzed three variants, validating the effectiveness of decoupled training, self-attention modules, and symmetric cross-entropy loss. The experimental results further confirm the reliability and effectiveness of these defense mechanisms. Specifically, the introduction of self-attention modules has improved the model's ability to recognize clean samples, and the adoption of symmetric cross-entropy loss has effectively reduced the side effects of low-confidence samples, thereby further enhancing the model's robustness. These analytical findings provide strong support for the proposed defense method, indicating its high practical value and operability in addressing backdoor attacks.

5. Conclusion

The mechanism of backdoor attacks involves poisoning the model during the training process, establishing a latent connection between trigger patterns and target labels. Research works [16], [17] have revealed that this connection is primarily learned through the end-to-end supervised training paradigm. Based on this understanding, this chapter utilizes a decoupled approach to defend against backdoor attacks. This method starts by learning the backbone layers through self-supervised learning, followed by learning the remaining fully connected layers through classical supervised learning. However, self-supervised learning has its limitations, as the self-supervised tasks are generated from input data, which may not match the true data distribution, leading to suboptimal representations learned by the model. To enhance the learning efficacy of the backbone layer, a self-attention module is introduced, and a label noise learning approach is used to determine high-confidence and low-confidence samples, followed by fine-tuning the entire model. Extensive experiments have demonstrated that the defense method proposed in this chapter can effectively reduce backdoor threats while maintaining high accuracy in predicting benign samples.

References

[1] Saha A, Subramanya A, Pirsiavash H. Hidden trigger backdoor attacks[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(07): 11957-11965.

[2] Huang K, Li Y, Wu B, et al. Backdoor defense via decoupling the training process[J]. arXiv preprint arXiv:2202.03423, 2022.

[3] Guo W, Tondi B, Barni M. An overview of backdoor attacks against deep neural networks and possible defences[J]. IEEE Open Journal of Signal Processing, 2022, 3: 261-287.

[4] Liu M, Sangiovanni-Vincentelli A, Yue X. Beating Backdoor Attack at Its Own Game[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 4620-4629.

[5] Kolesnikov A, Zhai X, Beyer L. Revisiting self-supervised visual representation learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 1920-1929.

[6] Gao Y, Doan B G, Zhang Z, et al. Backdoor attacks and countermeasures on deep learning: A comprehensive review[J]. arXiv preprint arXiv:2007.10760, 2020.

[7] Chavan S, Choubey N. An automated diabetic retinopathy of severity grade classification using transfer learning and fine-tuning for fundus images[J]. Multimedia Tools and Applications, 2023, 82(24): 36859-36884.

[8] Hardie R C, Barnard K J, Armstrong E E. Joint MAP registration and high-resolution image estimation using a sequence of undersampled images[J]. IEEE transactions on Image Processing, 1997, 6(12): 1621-1633.

[9] Wang R, Guo H, Davis L S, et al. Covariance discriminative learning: A natural and efficient approach to image set classification[C]//2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012: 2496-2503.

[10] Feng X, Shen Y, Wang D. Review on the development of image-based data enhancement methods [J]. Computer Science and Application, 2021, 11: 370.

[11] Pan X, Ge C, Lu R, et al. On the integration of self-attention and convolution[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 815-825.

[12] Yang A, Yang X, Wu W, et al. Research on feature extraction of tumor image based on convolutional neural network[J]. IEEE access, 2019, 7: 24204-24213.

[13] Gao K, Bai Y, Gu J, et al. Backdoor defense via adaptively splitting poisoned dataset[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 4005-4014.

[14] Chen X, Liu C, Li B, et al. Targeted backdoor attacks on deep learning systems using data poisoning[J]. arXiv preprint arXiv:1712.05526, 2017.

[15] Praveen S P, Srinivasu P N, Shafi J, et al. ResNet-32 and FastAI for diagnoses of ductal carcinoma from 2D tissue slides[J]. Scientific Reports, 2022, 12(1): 20804.

[16] Huang K, Li Y, Wu B, et al. Backdoor defense via decoupling the training process[J]. arXiv preprint arXiv:2202.03423, 2022.

[17] Kolesnikov A, Zhai X, Beyer L. Revisiting self-supervised visual representation learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 1920-1929.