Random, Latent and Salient Backdoor Attacks on Deep-Learning Models

Michael Shell
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332–0250
Email: http://www.michaelshell.org/contact.html

Homer Simpson
Twentieth Century Fox
Springfield, USA
Email: homer@thesimpsons.com

James Kirk
and Montgomery Scott
Starfleet Academy
San Francisco, California 96678-2391
Telephone: (800) 555–1212
Fax: (888) 555–1212

***Abstract*—In the last decade, deep neural networks (DNNs) have been effectively used for various important tasks, including facial recognition and autonomous driving. As a result, ensuring the security of DNNs has become incredibly important and has raised significant concerns. Considering the significance and major existence of DNN models in computer system, the project hereby proposes a novel backdoor attack strategy for choosing appropriate samples to poison, which is an improvement compared to random selection in previous, towards deep-learning based computer vision models which aims to provide insights for protecting the confidentiality, availability, and most of all, integrity of the DNN-related system. The proposed selection strategy shows perceptible superiority compared to baselines with the same rate of poisoning.**

***Index Terms*—Deep learning, Computer Vision, Computer Security**

1. Introduction

Deep neural networks (DNNs) are increasingly being used for various industrial and academic applications, such as image recognition, natural language processing, and speech recognition [1]. They have shown remarkable performance in computer vision tasks, such as object detection, image classification, and semantic segmentation. However, due to their data-driven nature, DNNs require a massive amount of training data to perform well [2]. This can make them vulnerable to attacks during the training stage, where attackers can inject malicious data to manipulate the model's behavior. Such phenomenon gets even worse considering that current trending of using public datasets instead of constructing dedicated datasets.

One type of attack that has gained popularity in recent years is the backdoor attack [3]. There is a type of backdoor attack in which an attacker poisons the training dataset by altering the labels or injecting additional triggers that cause the model to make incorrect predictions [4]. Backdoor attacks can compromise the security and integrity of DNN models [5], leading to incorrect predictions and so as to attack related-systems. This can have severe consequences, such as misclassification of objects in autonomous vehicles or incorrect diagnosis in medical applications [6].

In addition, to reduce the cost of training DNN models, users often opt for third-party datasets, platforms, or models. However, this convenience comes at the cost of losing control, which increases security risks. Third-party datasets or platforms can be attacked, and using them can result in the injection of malicious data during the training stage, and therefore resulting "backdoors" in DNN systems that threaten the security.

In this context, the proposed approach in this paper is an **additional deep-learning guided selective strategy of choosing samples to be poisoned** in image classification tasks. The goal is to achieve better attack effects without largely affecting the accuracy of the model on the original unpoisoned samples under the same poison rate in the training set. This approach can give insights to the topic improve the security and integrity of DNN models during the training stage by detecting and mitigating potential backdoor attacks.

2. Literature Review

In this project, we focus on improving the poisoning-based backdoor attacks by proposing a brand new strategy in applying it. Before delving into the specific strategies we have developed and tested, it is important to review the existing literature on backdoor attacks. A comprehensive survey conducted by Li et al. [7] provides a taxonomy of poisoning-based backdoor attacks, which is illustrated in Fig. 1 and Fig. 2. This taxonomy categorizes the different types of backdoor attacks and highlights their implementation differences.

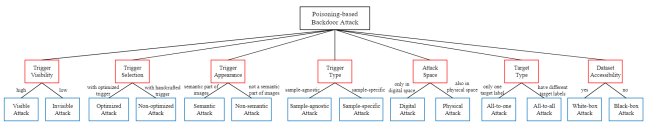

Figure 1. Backdoor Attack Method Categorization [7]

Originated from BadNet [8], most of the poisoning-based approach follows a similar idea of inserting a either

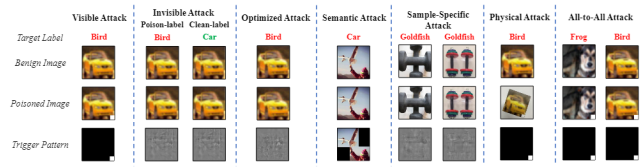

Figure 2. Backdoor Attack Visualizations by categories [7]

visible or invisible trigger pattern and altering the corresponding label to the target label of attack. The desired goal for such attack algorithms is that it should be both robust in attacking the model so that it produces a prediction the attacker wants when encountering the poisoned sample as much as possible, while still perform fairly normal when making predictions on the benign samples (unpoisoned samples). In other words, the model have to exhibit high accuracy score on both benign and label-altered poisoned dataset.

There have been efforts made in previous works in poisoning attack. LabelConsistent [9] introduced a method which do not need altering the sample's label by Latent Space Interpolation and Adversarial Perturbation that improved the stealthiness of the attack. Following this work, a major breakthrough is made by SleeperAgent [10] which further pushes the performance of pertubation-based data poisoning-based attack method in both attack success rate and confidentiality.

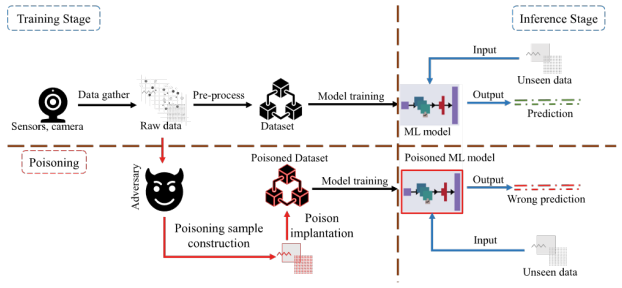

Figure 3. A typical Poisoning Attack Procedure

Despite poisoning attacks have evolved on attacking method, the paper aims to provide a novel str4ategy to improve both aspects of the backdoor attack by a more reasonable victim image selection, considering that in the process in previous works, the selection of victim images is conducted haphazardly without specific constraint or favor [11], while it is worth noticing that the images within the same dataset is of different contribution for the neural network activation towards the corresponding label [12]. i.e. the saliency of images may vary. Instead of counting on serendipity, we thereby propose a selective strategy based on the image-level feature significance which is expected to achieve stable benchmark improvement regardless of attack approaches taken.

3. Methods

As it is stated in introduction and literature review, the paper proposes a novel strategy which focuses on target image selection by the image-level feature significance rather than randomly picking victim in the dataset as previous works did. The strategy is designed without dependency to model, dataset, and any other conditions of a DNN model, and thus contribute to the general improvement of attack results. Specifically, the methods comes from one widely-accepted prior and two hypothesis from authors.

The **prior** is that, in Convolutional Neural Networks(CNN) [13], the input image's visual pattern will be extracted as spatial features represented by a high-dimensional vector, and as the model goes deeper it is summarized by the parameterized transformation into a feature vector that contributes to the class-wise probabilities for each image that will be used to classify the images as shown in Fig. 4 below.

Based on the **prior** above, an **assumption** is that selecting the certain amount of images with relatively salient features to attack may significantly forces the model to increase weight to activate trigger pattern feature over original features and thus becomes a feasible way of more efficient feature-based victim filtering instead of random selection.

If the assumption above holds, there is still existing problem that in what metric defining the images containing salient features. Here comes the second **assumption** that, the image containing more significant features for the specific class it belongs to will make the model more confident when judging classes. Therefore, gives higher probability score for corresponding class. Conversely, the model tends to misclassify images containing some feature combination belong to other classes. Finally, if constraining the selection within the images of higher confidence, it is more likely to get images with more salient features to attack. Following experiments are designed to proof these points. The attack image selection strategy takes regards from how possible an image belongs to specific class in image classification.

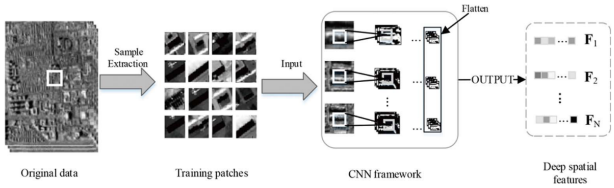

Figure 4. Image Classifier CNNs

The salient selection aims to find the images with most salient feature by selecting from images with the highest confidence to the target model in the dataset. The selection of image is implemented by take the voting results from several benign classifier models (trained on unpoisoned datasets). Each of the classifier is trained on separate datasets without overlapping samples to ensure their independence. Denote the vote model as F_1 to F_i and

input image as $x \in \mathcal{R}^{h*w*3}$, the process of voting can be described as shown in Eq. 1.

$$\begin{aligned} T_n(x) &\Leftarrow (F_N(x) \equiv x_{class}) \\ Vote &\Leftarrow T_1(x) \cap T_2(x) \cap ...T_n(x) \end{aligned} \tag{1}$$

An prediction x is considered as a salient one as long as it was correctly predicted by all the classifiers F_i with i ranges from 1 to n. It will then be associated with a confidence score C_x that is calculated based on the average softmax value [14] being extracted by a function S for each F in Eq. 2.

$$C_x = \frac{\sum_{i=1}^{n} S(F_i(x))}{n} \tag{2}$$

Finally, in terms of selection, the salient attack picks the images with top 1% of the whole dataset with the highest confidence score to apply any of the poisoning attack methods from the images that are correctly predicted by all the Fs. Just in case there is not enough images to be selected, which is extremely not likely to happen given the small victim quantity compared with the whole dataset, the selection program will automatically supply the vacancy of attack target image list with randomly picked samples.

Additionally, a contrasting attack strategy is also designed as Latent Attack which aims to further support the previous theoretic assumption by comparison. The Latent Attack takes the victim selection in a completely opposite manner, which selects from the images with wrong predictions and lowest confidence score. The Latent Attack victim are believed to have the most insignificant feature for each of the classes and therefore considered less valuable to be attack. i.e. if the previous assumption holds, the attack performance would possibly deteriorate compared with Salient Attack and even Random Attack. Therefore, gives a stronger support to our theory.

For the actual experimental part specifically, as shown in the general illustration in Fig. 5 we conducted the attack experiment on a ResNet-18 [15] backbone image classifier, and using the dataset CIFAR-10 [16] to train and test the model's performance. CIFAR-10 is a widely used base dataset for image classification tasks consists of 60k images with 32*32 resolution. The dataset is of 10 classes with equal sample quantities and is split into 50k image training dataset and 10k image testing dataset.

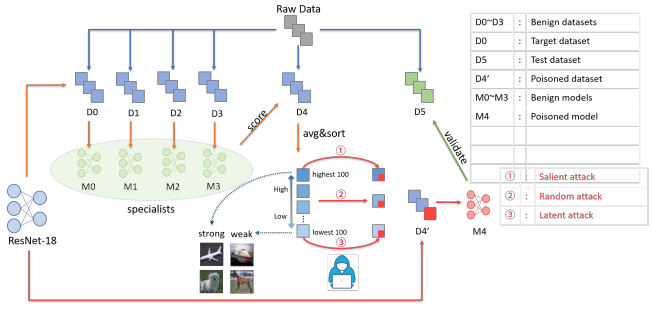

Figure 5. Experimental Procedure

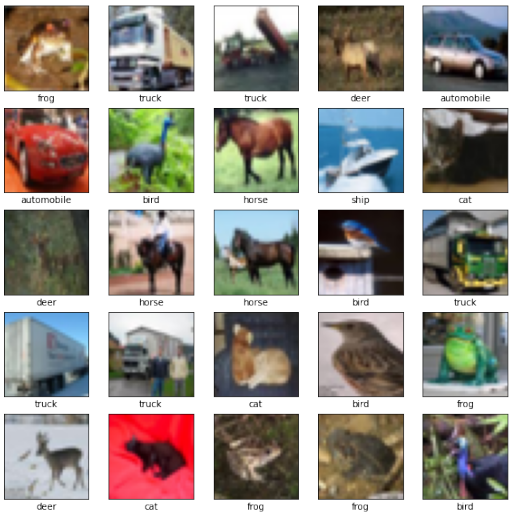

Figure 6. CIFAR-10 Image Samples

The baseline attack method used is the BadNet [8] which inaugurated the Backdoor attack topic and therefore roots on all the other poison-based attacks. We keep the original testing dataset and divide the training set into five folds randomly. Four of the folds are used for training four independent specialist models to vote for image-level possibility in the last fold of the original training set and therefore find the most salient samples to be attacked by BadNet. The BadNet attack is implemented by resuming training on a partly poisoned dataset with a model checkpoint pretrained on the dataset before poisoning [8]. The training may fluctuate and converge again over epochs of fine-tuning.

When it comes to testing stage, the testing is executed under benign original testing set and poisoned testing set, respectively, with all the samples within poisoned testing set having a visual trigger pattern embedded [8]. The testing stage will evaluate the accuracy score which is calculated by the number of correct predictions subject to the total number of predictions made on both types of testing set. The accuracy score calculated on benign testing set naturally represents the performance under normal circumstance without encountering poisonous data, while the accuracy score calculated on poisoned dataset could exhibit the attack success rate. Therefore, to illustrate the real-time effect in the experiment for our method, for every 10 epochs of fine-tuning, a testing on both datasets will be conducted to check the performance of the model and thus validate the strategy proposed.

4. Result Analysis

The experiment is conducted on AutoDL GPU platform, with experiment environment listed in the TABLE I. Each of the experiment iteration (250 epochs of fine-tuning) takes about 40mins to complete.

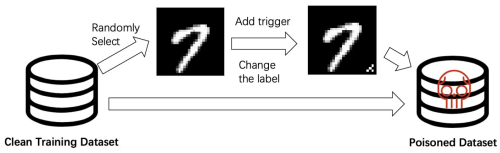

Figure 7. BadNet Attack [8]

TABLE 1. EXPERIMENT ENVIRONMENT

Component	Specification
Pytorch	1.11.0
Python	3.8
CUDA	11.3
GPU	NVIDIA RTX 3090
CPU	AMD EPYC 7543 32-Core Processor

The experiment produces a base ResNet-18 [15] after 200 epochs training that converges eventually on the train fold which is our targeted dataset. This model will be used as start-line checkpoint for consecutive poisoning attacks for the dataset fold. Repeating this process for all the other four folds derives four specialist model as illustrated in previous section. The prediction for each specialist model is saved in a .json file to calculate corresponding vote and furthermore, the salient samples and latent samples. The following diagrams show three different attack results towards the same baseline model during 250 epochs fine-tuning with corresponding test scores on benign testing set and poisoned dataset.

It is manifest that the stability of the three methods varies given the fluctuation is significantly different, in which the Random Attack and Latent Attack performs worse compared with Salient Attack. It is worth noticing that sometimes the Salient Attack converges on the poisoned testing set on a relatively high accuracy and retains an slightly higher score on the benign testing set which is about 77.6%. That means the Salient Attack did not affect the performance of the model on normal images without poisoning, which makes it a stealth attack strategy that is harder to be detected and dealt with. In contrast, the Latent Attack exhibits convergence on an even lower score than Random Attack and so as to the Salient Attack as listed in the TABLE I.

In short, the experiment result obtained have comprehensively proved the feasibility as well as the underlying theories.

5. Discussion

With the results obtained, the justification and explanation for the salient attack strategy stated previously could be supported by following phenomena in the experiment.

TABLE 2. ACCURACY SCORE ON POISONED TESTING SET

Attack Methods	Random	Latent	Salient
Accuracy Score	90.4%	88.4%	93.3%

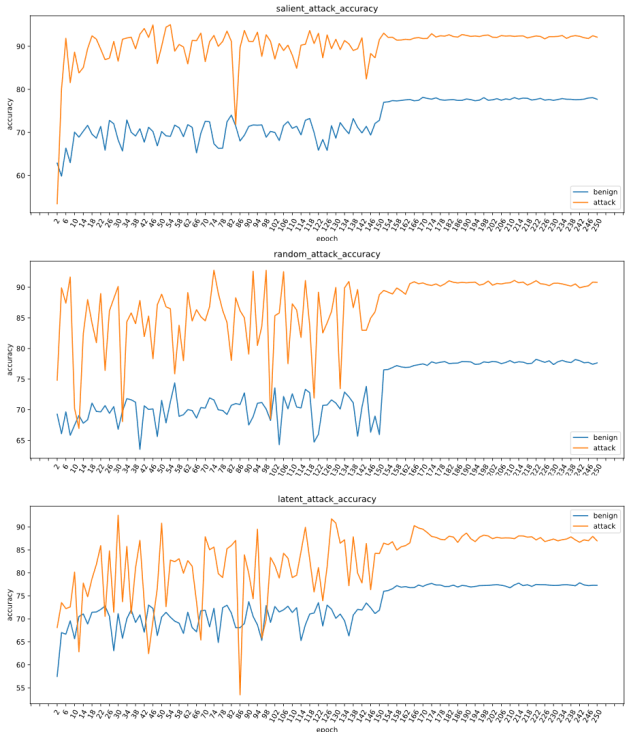

Figure 8. Salient, Random and Latent Attack Testing Accuracy Log

Firstly, the attacked model under three strategies give the expected benchmark ranking as we assumed in the introduction, where the Salient Attack will be the best given it confused the model the most by affecting the most salient features.

In addition, with fixed and limited poisoning rate, we expected acceptable accuracy degradation on the benign testing stage, because images with higher confidence score are confused in attack. However, to our surprise, the attacked ResNet-18 [15] model didn't show any degradation on the benign testing set, thereby we may infer that the model didn't inhibit the weight to activate salient features, while enhance the weight for trigger pattern alone, Thus, the model may preserve its ability when judging images in benign dataset without trigger pattern. That means the model tends to overlook the original contributing salient features than detected trigger pattern. Such mechanism actually protects most of the original features from poisoning for it tends to confuse the model only when the trigger pattern feature is encountered together with original features. Therefore, the strategy further pushes the attacking benchmark higher without confusing the model on benign prediction.

Additionally, it is observed that the Salient Attack exhibits more stable fine-tuning with relatively smaller fluctuation. the authors suspect that this is caused by the inherent nature of attack process. Since the attack aims to offset the positive effect of the features within the selected on the corresponding label by the negative effect on inserted patterns. Therefore, attacking on the salient samples leads

to tantamount strong negative offset caused by patterns governed by the model parameter compared with attacking on weaker samples in Random Attack and Latent Attack. Therefore, on nearly all the poisoned samples in testing set which do not contain stronger feature compared with salient samples in training set, the model's correct prediction tendency is obviated by the stronger negative which results less hesitation in giving wrong prediction.

To this end, the discussion gives clear explanation about both the expected and unexpected outcome of experiments, so as to prove the effectiveness and provide corresponding reasons to the strategy.

6. Conclusion

In conclusion, our proposed selective backdoor attack strategy has been successfully validated through experiments and arguments presented in this paper. The results clearly show that this strategy outperforms haphazardly random poisoning in terms of benchmark improvement. The suppression of stronger features by Salient Attack further strengthens the robustness of the proposed approach.

However, there is still room for further improvement and study. Firstly, more complicated mechanisms for image-level feature significance could provide a more reasonable ranking in victim selection, beyond the current method of voting and averaging. Secondly, the scalability and adaptability of this method need to be explored further, especially regarding the sensitivity of the method to the victim model's architecture. More extensive experiments could be conducted on different datasets, baseline DNN models and even other attack methods, to prove the value of our work and thus contribute more insights towards enhancing the security of computer vision-based systems.

In summary, while the proposed strategy has demonstrated promising results, further research is necessary to improve and generalize the approach. We hope that this paper can serve as a starting point for more extensive investigations into selective backdoor attacks and contribute to the advancement of computer vision-based system security.

Acknowledgment

The authors do not claim any conflict of interests.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–44, 05 2015.

[2] J. Zhang, Y. Zheng, D. Qi, R. Li, and X. Yi, "Dnn-based prediction model for spatio-temporal data," in *Proceedings of the 24th ACM SIGSPATIAL international conference on advances in geographic information systems*, 2016, pp. 1–4.

[3] Y. Li, M. Ya, Y. Bai, Y. Jiang, and S.-T. Xia, "Backdoorbox: A python toolbox for backdoor learning," 01 2022.

[4] Y. Liu, X. Ma, J. Bailey, and F. Lu, "Reflection backdoor: A natural backdoor attack on deep neural networks," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*. Springer, 2020, pp. 182–199.

[5] H. Qiu, Y. Zeng, S. Guo, T. Zhang, M. Qiu, and B. Thuraisingham, "Deepsweep: An evaluation framework for mitigating dnn backdoor attacks using data augmentation," in *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, 2021, pp. 363–377.

[6] M. Juuti, S. Szyller, S. Marchal, and N. Asokan, "Prada: protecting against dnn model stealing attacks," in *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2019, pp. 512–527.

[7] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia, "Backdoor learning: A survey," 2022.

[8] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," 2019.

[9] A. Turner, D. Tsipras, and A. Madry, "Label-consistent backdoor attacks," 2019.

[10] H. Souri, L. Fowl, R. Chellappa, M. Goldblum, and T. Goldstein, "Sleeper agent: Scalable hidden trigger backdoors for neural networks trained from scratch," 2022.

[11] C. Burkard and B. Lagesse, "Analysis of causative attacks against svms learning from data streams," in *Proceedings of the 3rd ACM on International Workshop on Security And Privacy Analytics*, 2017, pp. 31–36.

[12] B. Chen, W. Carvalho, N. Baracaldo, H. Ludwig, B. Edwards, T. Lee, I. Molloy, and B. Srivastava, "Detecting backdoor attacks on deep neural networks by activation clustering," *arXiv preprint arXiv:1811.03728*, 2018.

[13] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," 2015.

[14] T. Pearce, A. Brintrup, and J. Zhu, "Understanding softmax confidence and uncertainty," 2021.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[16] F. O. Giuste and J. C. Vizcarra, "Cifar-10 image classification using feature ensembles," 2020.