

GANcrop: A Contrastive Defense Against Backdoor Attacks in Federated Learning

Xiaoyun Gan
Guangxi Normal University
Guilin, Guangxi, China
ganxiaoyun@stu.gxnu.edu.cn

Shanyu Gan
Guangxi Normal University
Guilin, Guangxi, China
gsy520@stu.gxnu.edu.cn

Taizhi Su
Guangxi Normal University
Guilin, Guangxi, China
csstz@stu.gxnu.edu.cn

Peng Liu*
Guangxi Normal University
Guilin, Guangxi, China
liupeng@gxnu.edu.cn

ABSTRACT

With heightened awareness of data privacy protection, Federated Learning (FL) has attracted widespread attention as a privacy-preserving distributed machine learning method. However, the distributed nature of federated learning also provides opportunities for backdoor attacks, where attackers can guide the model to produce incorrect predictions without affecting the global model training process. This paper introduces a novel defense mechanism against backdoor attacks in federated learning, named GANcrop. This approach leverages contrastive learning to deeply explore the disparities between malicious and benign models for attack identification, followed by the utilization of Generative Adversarial Networks (GAN) to recover backdoor triggers and implement targeted mitigation strategies. Experimental findings demonstrate that GANcrop effectively safeguards against backdoor attacks, particularly in non-IID scenarios, while maintaining satisfactory model accuracy, showcasing its remarkable defensive efficacy and practical utility.

CCS CONCEPTS

• Security and privacy → Intrusion detection systems.

KEYWORDS

Federated Learning; Attack Defense; Backdoor Attack; Contrastive Learning; GAN

ACM Reference Format:

Xiaoyun Gan, Shanyu Gan, Taizhi Su, and Peng Liu*. 2018. GANcrop: A Contrastive Defense Against Backdoor Attacks in Federated Learning. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 6 pages. <https://doi.org/XXXXXXXX.XXXXXXX>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00
<https://doi.org/XXXXXXXX.XXXXXXX>

1 INTRODUCTION

With the rise in awareness of data privacy protection, centralized data collection faces significant challenges, making collecting training data a pressing issue in machine learning. Federated learning, as a novel distributed machine learning method that protects privacy [11], cleverly bypasses data collection challenges and has thus received widespread attention.

In Federated Learning (FL), training data is kept locally on the user's device, and only the model initialization and trained model parameters are transmitted between the server and participating users. This method allows for secure training collaboration among multiple parties while protecting user privacy. Nevertheless, owing to the data being distributed among various participating entities, this distributed characteristic, while rendering the model training more flexible and secure, concurrently offers attackers a foothold [4, 17]. They can stealthily guide the model to produce specific predictions for certain triggers without impacting the global model training process [1], thereby diminishing the model's accuracy and credibility. Thus, researching effective defenses against backdoor attacks in federated learning is crucial for ensuring the model's security and reliability.

In the field of deep learning, there has been considerable research dedicated to defending against backdoor attacks. Methods for recovering backdoor triggers, such as NC [16] and GANsweep [19], aim to mitigate backdoor attacks by recovering the triggers of poisoned models, thereby achieving defense objectives. However, due to the specific data distribution in federated learning, applying methods from centralized learning directly to FL may lead to poor defense outcomes or model performance, often resulting in the global model failing to converge. Furthermore, the model updating process in federated learning involves parameter exchange and model aggregation among servers and multiple users, which adds to the defense's complexity. Therefore, developing new defense mechanisms tailored to the peculiarities of federated learning is necessary to address potential backdoor attack threats.

Existing federated learning backdoor defences schemes can be mainly divided into two categories: methods based on anomaly detection and those utilizing pruning or noise addition techniques. Methods based on anomaly detection usually require extensive computations on the server side [5], detecting backdoor attacks by monitoring the similarity between models or the abnormal changes in update behaviors. However, due to the complexity of federated

learning models and the heterogeneity of data distribution, these methods often incur excessive computational costs and have high prerequisites for the scenarios [14, 18]. On the other hand, methods based on pruning or adding noise work by diluting or reducing the impact of malicious model updates on the global model through the insertion of noise into the model updates [10, 15]. Although this approach can enhance the robustness of the model, the added noise often leads to a decrease in model performance and accuracy, thus affecting the overall performance of the model. Therefore, it is imperative to seek a more effective method for defending against federated learning backdoor attacks to balance security and performance demands.

To overcome the limitations of existing methods, this paper proposes a federated learning backdoor attack defense method based on Contrastive Learning [3] and Generative Adversarial Networks (GAN) [6], named GANcrop. This method utilizes contrastive learning to delve into the differences between malicious and benign models, achieving effective attack identification. Then, on the predicted malicious models, it uses GAN to recover the trigger of poisoned models and carries out targeted backdoor mitigation to achieve the defense purpose. The contributions of this paper are as follows:

- We have implemented a new federated learning backdoor attack defense method, GANcrop, based on contrastive learning and generative adversarial networks.
- We introduced a model detection method based on contrastive learning, which can effectively distinguish between malicious and benign models under the multi-model scenario of federated learning, achieving effective attack detection. This is one of the rare model-level contrastive learning methods in existing research.
- We applied the backdoor trigger recovery approach from deep learning to the defense work against backdoor attacks in federated learning. In non-IID scenarios, our method achieves sound defense effects while ensuring model accuracy.

2 PRELIMINARY KNOWLEDGE

Backdoor attacks in neural networks aim to guide the neural network model to produce incorrect output results by making subtle, humanly imperceptible modifications to the input data. Attackers inject backdoors into the model during the training or deployment phase, causing the model to produce incorrect predictions under specific triggering conditions. In backdoor attacks, the attacker's task can be described as a multi-objective optimization problem, where the attacker needs to maintain the accuracy of the main task while achieving a high success rate of backdoor attacks on targeted attack class samples [13]. The optimization objective function of a backdoor attack can be represented by Formula (1):

$$\theta^* = \min_{\theta} \left(\sum_{i \in |D|} L(x_i, y_i) + \sum_{i \in |D_p|} L(\psi(x_i), \tau(y_i)) \right) \quad (1)$$

Where D is the test set for the main task; D_p is the poisoned dataset containing backdoor samples, these samples are manipulated by the transformation function ψ , outputting a specific y under the backdoor task.

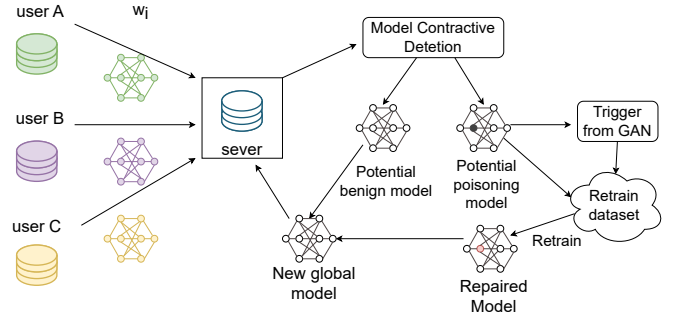


Figure 1: GANcrop architecture diagram

3 PROPOSED METHOD

This section will introduce the federated learning backdoor attack defense framework based on Generative Adversarial Networks (GAN). This framework mainly consists of three modules: attack detection, backdoor mitigation, and model aggregation. The framework of our method is shown in Fig. 1.

3.1 Attack Detection

In devising strategies to defend against backdoor attacks, directly comparing the similarity between models is often insufficient to identify contaminated models, as triggers are usually concealed. It is necessary to delve into the analysis of model parameter differences and sensitivities to distinguish between benign and poisoned models. In backdoor attacks, triggers are generally placed around the periphery of images, not disturbing the main task's accuracy. It's challenging to analyze the trigger location from model parameters directly, so we employ contrastive learning to train an anomaly detection model sensitive to edge position parameters, conducting direct contrastive training on model data, which is particularly crucial in federated learning.

We construct a poisoned dataset with non-central triggers to train the poisoned models. Under the contrastive learning framework, we build positive and negative sample pairs from the parameters of poisoned and clean models. Unlike traditional methods [3], in our construction, different poisoned model samples from different poisoned or benign models benign model samples are considered positive samples, while poisoned and benign models are considered negative samples. This method of constructing sample pairs allows the model to extract as many common features among poisoned models as possible. Our framework optimizes the contrastive learning model by conducting direct contrastive training on model data and learning the patterns of trigger locations to distinguish between model categories accurately. This is key to our model-level contrastive detection in federated learning.

Our model contrast detection diagram is shown in Fig. 2. The specific contrastive loss function is shown in Formula (2):

$$L_{\text{InfoNCE}} = -\log \frac{\exp(d(u, v)^- / \tau)}{\sum_{k=1}^K \exp(d(u, v_k)^+ / \tau)} \quad (2)$$

Where $d(u, v)^-$ represents the distance between positive sample pairs, $d(u, v_k)^+$ represents the distance between negative sample pairs, and τ is the temperature parameter of the contrastive loss, used to control the degree of softening.

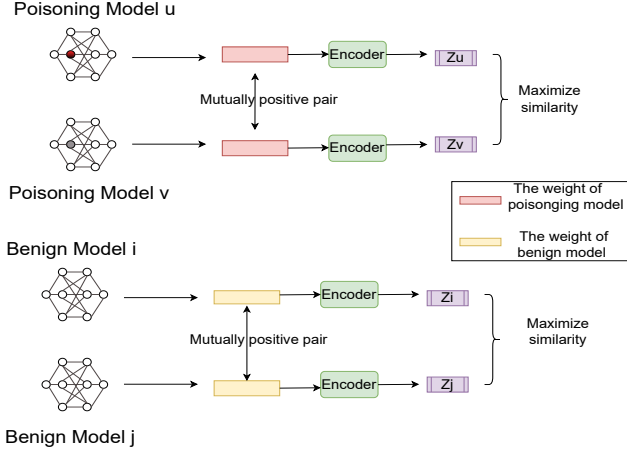


Figure 2: Model Contrastive Learning

Through the extraction of model representations using contrastive learning, it's possible to identify the common characteristics of potentially anomalous models. Utilizing these characteristics for model classification training allows for precisely excluding malicious influences and identifying potentially poisoned models. This training process is completed on the server side using a public dataset to distinguish potential poisoned models through feature extraction and classifier training. The algorithm for GANcrop, as shown in Algorithm 1.

3.2 Backdoor Mitigation

To effectively defend against backdoor attacks, we further mitigate backdoors based on the results of parameter comparison detection, reducing the impact of the attacks. We use GAN to recover the backdoor triggers of poisoned models, reintroduce the recovered backdoors into the server's clean dataset with correct labels and retrain the poisoned models to mitigate the backdoors.

The GANGSWEEP framework inspires the backdoor recovery strategy, a defense method in deep neural networks that uses GANs to recover backdoor triggers and mitigate their effects. However, directly applying it to FL may cause convergence difficulties for the global model. Therefore, this paper adopts a method of attack detection followed by backdoor mitigation to defend against backdoor attacks in FL.

In this paper, we attempt to recover the backdoor triggers in the contaminated model F using the generator of a Generative Adversarial Network (GAN). The backdoor triggers are usually designed as special vectors with the same dimensions as the image. Such a design ensures the feasibility of generating triggers from the image

Algorithm 1: Model Contrastive Detection Algorithm

Input: Number of simulated clients N , learning rate η , ResNet network model R , initial feature extraction $SimModel$, simulated user model parameters w_i^t , epoch E , trigger TR , public dataset D_c , proportion of poisoned dataset γ , simulated attack user set S_p , simulated benign user set S_b , positive and negative sample pair construction function $g(\cdot)$

Output: Model feature extractor $SimModel$, discriminator

1 **Function** ModContract(D_c, TR):

2 $D_p = \gamma D_c + TR$;

3 Send the initial model to each simulation client;

4 **for** $t = 0$ to $T - 1$ **do**

5 **for** $j = 0$ to $N - 1$ **do**

6 Send global model parameters to the customer;

7 **if** $j \in S_p$ **then**

8 $w_i^t \leftarrow \text{Local Training}(j, w^t, D_p)$;

9 **end**

10 **else**

11 $w_i^t \leftarrow \text{Local Training}(j, w^t, (D_c - D_p))$;

12 **end**

13 $(u, v) = g(w_i^t (j \in S_p), w_i^t (i \in S_b))$;

14 $L_{\text{InfoNCE}} = -\log \frac{\exp(d(u, v)^- / \tau)}{\sum_{k=1}^K \exp(d(u, v_k)^+ / \tau)}$;

15 $SimModel \leftarrow SimModel - \eta \nabla L$;

16 **end**

17 **end**

18 **return** $IdentifyMod$;

19 **Function** ModClassify($SimModel, S_p, S_b, w_i^t$):

20 $IdentifyMod \leftarrow \text{LinearEvaluate}(SimModel, \text{classes})$;

21 **for** $j = 1$ to E **do**

22 $IdentifyMod_j \leftarrow IdentifyMod_{j-1} - \eta \nabla F_j(S_p, S_b)$;

23 **end**

24 **return** $IdentifyMod$;

model. Unlike traditional GANs, we do not optimize the discriminator during training; instead, we use the model from potentially malicious users as a substitute to guide the generator in recovering the original triggers more accurately. To verify the effectiveness of the generated triggers, this paper injects the generated triggers one by one into the images (marked as x) in the validation dataset, forming new images $G(x) + x$, and observes the predictions of the malicious model F on these new images to guide the generator in iteratively improving, gradually learning trigger features. Our architecture can effectively recover backdoor triggers through continuous iterative training, as illustrated in Fig. 3.

After obtaining the generated triggers highly similar to the original poisoned model's triggers, this paper cleverly injects these generated triggers into a new, clean dataset while keeping the dataset's real labels unchanged for model retraining. The purpose is to allow the model to correctly identify and handle the correspondence between image data and labels during retraining, thereby effectively eliminating the impact of the backdoor triggers. The

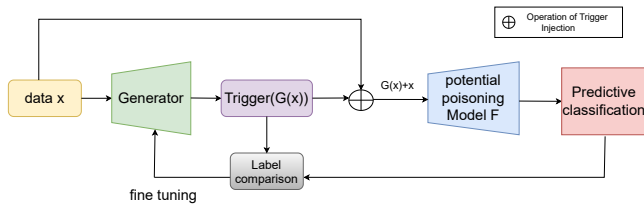


Figure 3: Trigger Generator

retraining process is a process of knowledge updating and forgetting, where the model gradually adjusts its internal parameters, forgetting the original backdoor features to adapt to the new data distribution.

3.3 Model Aggregation

Based on the results of attack detection, this paper identifies and filters out benign models and potentially malicious user models. Then, through the backdoor recovery of the GAN model, pseudo triggers similar to the original triggers of potentially malicious models are generated, and these pseudo triggers are used for targeted backdoor mitigation. The models retrained afterward will be restored to a state unaffected by any malicious influence.

To enhance the robustness and reliability of the model, fully utilizing the training success of each user model, this repaired model is weighted and aggregated with the identified benign models to obtain a global, more robust federated learning model.

4 EXPERIMENT

4.1 Experimental Setup

This experiment uses the ResNet18 neural network model as the base architecture. The experiment employs a 3×3 convolution kernel, with stride and padding values set to 1. This experiment omits pooling layers to avoid excessive compression of image information. The experiment involves 40 clients, selects CrossEntropyLoss as the loss function, and adopts classic stochastic gradient descent (SGD) as the optimizer, with a learning rate set to 0.1. In each iteration round, each client conducts training for four epochs.

Dataset: The experiment utilizes the widely used Cifar10 dataset in the same domain applications. The dataset contains 60,000 data samples, 50,000 training samples, and 10,000 test samples, with 10 categories in the dataset.

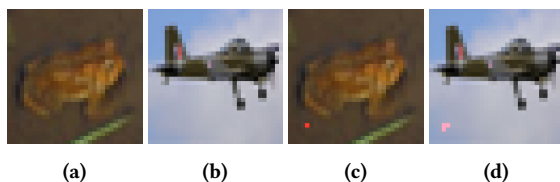


Figure 4: Attack Schematic

Data Division: To achieve a non-IID data distribution, the experiment adopts the data skew and label skew methods proposed in the literature [7], setting the Dirichlet coefficient to 0.7 to control the degree of data skew.

Attack Method: The trigger injection [9] backdoor attack method is adopted. Fig. 4 shows the effect of trigger injection using the cifar10 dataset, where the first two subfigures are original pictures, and the last two subfigures are poisoned pictures after trigger injection.

4.2 Experimental Results and Analysis

Verifying the Effectiveness of GANcrop in Defending Against Backdoor Attacks: To compare the performance of GANcrop, control methods were selected including FedAvg [11] without defense measures as a baseline and four representative defense methods: Krum[2], Trimmed-mean [12], Fang[8], and GANsweep[19]. An attack ratio of 30% was set for the experiments, with these attacking users launching attacks in each round of iteration. Fig. 5 (a) depicts the rounds of successful defense against backdoor attacks for six methodologies on the CIFAR-10 dataset, across a total of 50 experimental rounds. Here, we define a successful defense against backdoor attacks as instances where backdoor accuracy falls below 30%. Typically, the level of backdoor accuracy is directly correlated with the success rate of the backdoor attack.

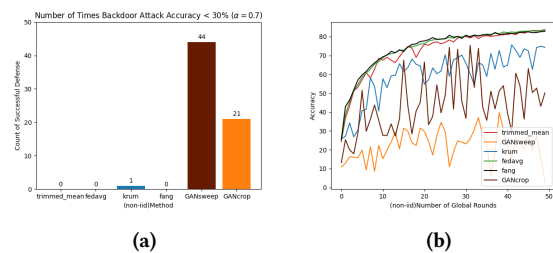


Figure 5: Successful rounds of defending against attacks and main task accuracy of six methods

The experimental results indicate that the models trained using FedAvg, Krum, Trimmed-mean, and Fang methods exhibit minimal effectiveness in defending against backdoor attacks. In contrast, GANsweep demonstrates a significantly higher defense capability when confronted with corresponding trigger attacks. On the other hand, the GANcrop scheme presented in this paper achieves a more balanced effect in defending against backdoor attacks. The reason is that GANcrop is a federated scheme aiming to balance maintaining the accuracy of the model's main task with reducing the success rate of attacks.

Verifying the main task accuracy of GANcrop: Fig. 5 (b) shows the change in the main task accuracy of the six experimental methods on the cifar10 dataset. FedAvg, Trimmed-mean, and Fang do not effectively defend against backdoor attacks despite having high main task accuracy. However, combined with the experimental results of backdoor accuracy above, it's clear that FedAvg, Trimmed-mean, and Fang, despite having high main task accuracy, do not effectively defend against backdoor attacks. By discarding a portion of user models during the global aggregation phase, Krum shows weakness in main task accuracy. Meanwhile, GANsweep, although it has backdoor solid defense capabilities, also leads to lower main task accuracy.

Table 1: Execution Time

Time	Fedavg	Krum	Trimmed _mean	Fang	GANsweep	GANcrop
s	258	303	302	533	451	477

Additionally, the GANcrop method presents a compromise in terms of the model’s main task accuracy, achieving a level of backdoor defense effectiveness while maintaining a main task accuracy that is higher than that of the GANsweep method but lower than the other four methods.

Verify the main task accuracy and the backdoor task accuracy of the two sub-models: To validate the model’s attack detection and repair performance, this paper analyzed the main task accuracy and the backdoor task accuracy of two sub-models. Fig. 6 shows these results: Fig. 6 (a) displays the main task accuracy of both the benign prediction model and the repair model, while Fig. 6 (b) presents their backdoor task accuracy. Among them, a higher backdoor task accuracy represents a higher backdoor attack success rate. Fig. 6 (a) reveals that the benign model outperforms the repair model in main task accuracy, potentially due to the repair model’s dataset. Fig. 6 (b) shows that the backdoor task accuracy is significantly lower in the repair model. This highlights the importance of integrating the benign and repair models to lessen the main task accuracy loss from backdoor mitigation and diminish the effects of undetected attacks.

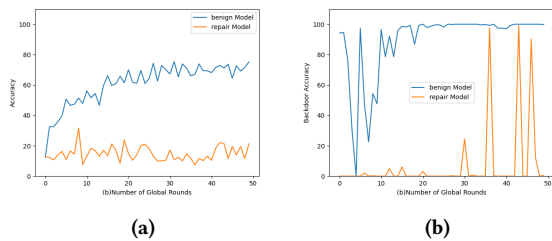


Figure 6: The main task accuracy and backdoor attack success rate of two submodels

Verifying Execution Time of Six Approaches: The experiment further compared the execution time of the six different methods, aiming to assess their total time consumption from local computation to global aggregation in a single round of iteration. The execution times obtained through experimental verification are shown in Table 1.

5 CONCLUSIONS

This paper addresses the problem of backdoor attacks in federated learning and proposes a defense method based on contrastive learning and Generative Adversarial Networks, GANcrop. Model comparison and sensitivity analysis of parameters effectively distinguish malicious and benign models. Then, utilizing GAN technology to recover and mitigate backdoor triggers in models, significantly reduces the success rate of backdoor attacks. Experiments have

proven that compared to existing methods, GANcrop not only enhances the defense against backdoor attacks but also mitigates the risk of backdoor attacks in non-IID data scenarios while maintaining the accuracy of the main task of federated learning models. In future work, efforts will be directed towards further optimizing the mitigation strategy of GANcrop, aiming to enhance the post-mitigation model accuracy and thereby augmenting the method’s practical applicability.

6 ACKNOWLEDGMENTS

The research was supported in part by the Guangxi Science and Technology Major Project (No.AA22068070), the National Natural Science Foundation of China (Nos.62166004,U21A20474), the Key Lab of Education Blockchain and Intelligent Technology, the Center for Applied Mathematics of Guangxi, the Guangxi "Bagui Scholar" Teams for Innovation and Research Project, the Guangxi Talent Highland Project of Big Data Intelligence and Application, the Guangxi Collaborative Center of Multisource Information Integration and Intelligent Processing.

REFERENCES

- [1] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. 2020. How To Backdoor Federated Learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 108)*, Silvia Chiappa and Roberto Calandra (Eds.). PMLR, 2938–2948. <https://proceedings.mlr.press/v108/bagdasaryan20a.html>
- [2] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. 2017. Machine learning with adversaries: byzantine tolerant gradient descent. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (Long Beach, California, USA) (NIPS’17)*. Curran Associates Inc., 118–128.
- [3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. arXiv:2002.05709 [cs.LG]
- [4] Clement Fung, Chris J. M. Yoon, and Ivan Beschastnikh. 2020. The Limitations of Federated Learning in Sybil Settings. In *International Symposium on Recent Advances in Intrusion Detection*. <https://api.semanticscholar.org/CorpusID:221542915>
- [5] T Gao, X Yao, and D Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821* (2021). <https://arxiv.org/abs/2104.08821>
- [6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. 63, 11 (oct 2020), 139–144.
- [7] Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. 2022. Federated Learning on Non-IID Data Silos: An Experimental Study. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. 965–978. <https://doi.org/10.1109/ICDE53745.2022.00077>
- [8] Suyi Li, Yong Cheng, Yang Liu, Wei Wang, and Tianjian Chen. 2019. Abnormal Client Behavior Detection in Federated Learning. arXiv:1910.09933 [cs.LG]
- [9] Yiming Li, Tongqing Zhai, Baoyuan Wu, Yong Jiang, Zhifeng Li, and Shutao Xia. 2021. Rethinking the Trigger of Backdoor Attack. arXiv:2004.04692 [cs.CR]
- [10] K Liu, B Dolan-Gavitt, and S Garg. 2018. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*. Springer International Publishing, 273–294. <https://arxiv.org/abs/1805.12185>
- [11] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 54)*. PMLR, 1273–1282. <https://proceedings.mlr.press/v54/mcmahan17a.html>
- [12] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. 2018. Exploiting Unintended Feature Leakage in Collaborative Learning. arXiv:1805.04049 [cs.CR]
- [13] Thuy Dung Nguyen, Tuan Nguyen, Phi Le Nguyen, Hieu H. Pham, Khoa D. Doan, and Kok-Seng Wong. 2024. Backdoor attacks and defenses in federated learning: Survey, challenges and future research directions. *Engineering Applications of Artificial Intelligence* 127 (2024), 107166. <https://doi.org/10.1016/j.engappai.2023.107166>

- [14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV]
- [15] Ziteng Sun, Peter Kairouz, Ananda Theertha Suresh, and H. Brendan McMahan. 2019. Can You Really Backdoor Federated Learning? arXiv:1911.07963 [cs.LG]
- [16] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. 2019. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks. In *2019 IEEE Symposium on Security and Privacy (SP)*. 707–723. <https://doi.org/10.1109/SP.2019.00031>
- [17] Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. 2020. DBA: Distributed Backdoor Attacks against Federated Learning. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=rkgyS0VFvr>
- [18] Mang Ye, Xiuwen Fang, Bo Du, Pong C. Yuen, and Dacheng Tao. 2023. Heterogeneous Federated Learning: State-of-the-art and Research Challenges. arXiv:2307.10616 [cs.LG]
- [19] Liuwan Zhu, Rui Ning, Cong Wang, Chunsheng Xin, and Hongyi Wu. 2020. GangSweep: Sweep out Neural Backdoors by GAN. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*. 3173–3181. <https://doi.org/10.1145/3394171.3413546>