

Weekly Report

Zhengtian Zhu

April 27, 2023

1 2023. Week 10

(Please use the following template for the weekly report. (I am working on a couple of paper deadlines and will circle back to you asap. Keep working on the experiments) [got it!](#)

1.1 tasks done in this week

1.1.1 paper summary

- **Paper Info:** [\[21\]](#)
- **Research Problem:**
- **Ideas/Novelty:** (Describe the idea or methodology proposed in this paper. Keep it in one paragraph.
- **Your Thoughts:** (Pros and Cons of the paper. Any takeaway message? Any thoughts to make it better or benefit to your own research problem?

[\[25\]](#) describes that it can process both the high and low dimensional information for the image. Maybe I can poison the low-level dimension.

1.1.2 progress of your research paper

(describe the progress/update of your research paper

1.2 tasks plan to do in the next week

- Action 1: (Papers plan to read
- Action 2: (Progress goal of your paper/research
- Action 3: (others

2 2023. Week 9

finish some parts for visa and my master paper. (Please use the following template for the weekly report. (I am working on a couple of paper deadlines and will circle back to you asap. Keep working on the experiments) [got it!](#)

3 2023. Week 8

I've got a flu...

4 2023. Week 7

4.1 tasks done in this week

Dear Prof. Yang, [The truth is the idea is totally wrong.](#)I've misunderstood the meaning of [weight attack](#). I just figure out a idea: Distributed weighted attack on Federated learning.

Then I can perform some other attacks on FL to compare and get a conclusion: The proposed attack is more effective than those attacks. What do you think about this novelty?

[Give some clear problem definitions and related work.](#)

clear problem definitions:

Related Work: [9] introduces a poisoning weight attack, and [22] introduces a distributed attack.

The link here helps me learn some basics about FL. <https://learnopencv.com/federated-learning-using-pytorch-and-pysyft/> followed with <https://github.com/OpenMined/PySyft>.

4.1.1 paper summary

- **Paper Info:** Challenges and Approaches for Mitigating Byzantine Attacks in Federated Learning, Shengshan Hu, Jianrong Lu, Wei W, Trust-Com 2020 CCF-C [4] proposes a new FL framework defending Byzantine attacks. It even evaluates its experiment under different datasets. <https://github.com/DistributedML/FoolsGold> is also a kind of defense ways. [8] comes up the idea of poisoning the weight of the model in the last layer. (Title, authors, conference/journal
- **Research Problem:** (Describe the research problem in brief.
- **Ideas/Novelty:** (Describe the idea or methodology proposed in this paper. Keep it in one paragraph.
- **Your Thoughts:** A trade-off between efficiency, security, and privacy needs to be carefully considered for specific application scenarios.(Pros and Cons of the paper. Any takeaway message? Any thoughts to make it better or benefit to your own research problem?

4.1.2 progress of your research paper

(describe the progress/update of your research paper)

4.2 tasks plan to do in the next week

- Action 1: (Papers plan to read)
- Action 2: (Progress goal of your paper/research)
- Action 3: (others)

5 2023. Week 6

Recently, I've newly adopted the backdoor attack like this <https://github.com/cwllenny/Sign-Flip-Attack>. However, this attack seems to be a poisoning attack rather than backdoor attack. Compared to ISSBA, ISSBA's pre-trained model has already fixed the input, which means I have to retrain a new one to apply my experiment, just to make the input size same.

Data Poisoning is an adversarial attack that tries to manipulate the training dataset in order to control the prediction behavior of a trained model such that the model will label malicious examples into a desired classes (e.g., labeling spam e-mails as safe).

Though some attacks are based on poisoning parameters(eg. gradient), some defense methods are also effective in detecting malicious updates. [18] (Please use the following template for the weekly report. (I am working on a couple of paper deadlines and will circle back to you asap. Keep working on the experiments) [got it!](#))

5.1 tasks done in this week

5.1.1 paper summary

- **Paper Info:** Wei Wan, Shengshan H, IJCAI [18] proposes an adaptive scheme of client selection to strengthen the robustness of FL. The motivation behind that is because it has found random selection of clients will be worse than the rational selection of clients.
- **Research Problem:** all the existing defenses suffer from one or multiple limitations as follows.1.Firstly, they paid little attention to the client selection step. 2.Second, they cannot effectively defend against sybil attacks.Third, they perform badly in non-IID (independently identically distribution) scenarios,
- **Ideas/Novelty:** Our scheme models the client selection process in federated learning as an extended MAB problem enabling the server to adaptively select updates that are more likely to be benign. In short, The paper applies The multi-armed bandit(MAB)algorithms to detect Byzantine attacks. (The multi-armed bandit (MAB) problem is a classical framework which studies the exploration/exploitation tradeoff in sequential decision problems.)(Describe the idea or methodology proposed in this paper. Keep it in one paragraph.

(quoted: [18] , which first discards the updates that are excessively similar in direction through graph theory, aiming to cope with the collusion attack. Then it utilizes principal components analysis (PCA) to extract the key parameters lie in the updates, because benign and malicious updates are easier to distinguish (e.g., through agglomerative clustering) in the new low-dimensional parameter space. In this way, the non-sybil challenge can be settled.)

- **Your Thoughts:** The researcher performs some attacks on different defenses and the method that [18] comes up, which is quite clear. I'll pay more attention to how it select the client in the future since adaptive client selection is quite smart! I've known some attacking methods and defense methods from this paper.(Pros and Cons of the paper. Any takeaway message? Any thoughts to make it better or benefit to your own research problem?

5.1.2 progress of your research paper

distributed poisoning attack; MAB-RFL:adaptive client selection for clients' updates, I can adaptively select what?emmm...

(describe the progress/update of your research paper

5.2 tasks plan to do in the next week

- Action 1: (Papers plan to read
- Action 2: (Progress goal of your paper/research
- Action 3: (others

6 2023. Week 5

I'm reading the code in https://github.com/lyhue1991/eat_tensorflow2_in_30_days because of curiosity, and I want to build up my foundation of knowledge a little bit since I'm so weak in the realization of coding, and I want to see how others design some codes since I've seen some algorithm framework from others' papers. Besides, I can figure out what kind of code I can perform. However, I found it was all about (Please use the following template for the weekly report. (I am working on a couple of paper deadlines and will circle back to you asap. Keep working on the experiments) [got it!](#)

6.1 tasks done in this week

6.1.1 paper summary

- **Paper Info:** [15] is about a novel defensive framework in the real application, which compares its result with a few novel methods. However, it didn't clarify its innovation, and I can only see a whole overview from the picture.
- **Research Problem:** to address the problem that the edge devices never share their raw data, the paper proposes a novel defensive method using deep neural networks and SVM.
- **Ideas/Novelty:** the paper has a big picture with 2 propositions, then it has a special framework. The Audit Model is trained from DA (audit dataset, the last layer of activation). Then two Audit Model will perform together. One is a standard one outputting the standard value to judge whether other values are qualified or not, the other makes predictions by absorbing the dataset in the real scenery. What Audit Models do, actually, is (Describe the idea or methodology proposed in this paper. Keep it in one paragraph.
- **Your Thoughts:** Pros and Cons of the paper. Any takeaway message? Any thoughts to make it better or benefit your own research problem?

6.1.2 paper summary

- **Paper Info:** []
- **Research Problem:**
- **Ideas/Novelty:** (Describe the idea or methodology proposed in this paper. Keep it in one paragraph.
- **Your Thoughts:** There are some Model poisoning attacks: Sign-flipping attack (SF) (Pros and Cons of the paper. Any takeaway message? Any thoughts to make it better or benefit your own research problem?

6.1.3 progress of your research paper

<https://zhuanlan.zhihu.com/p/51165622> Recently, I've known that why the error last week happens and I have a deep understanding of tensorflow, and I have the idea how to learn the code efficiently. Know the concept and learn everything when I have to perform a task. Besides, learn the existing code from others. That will be the best way to know how to organize the code according to the tutorial. https://github.com/lyhue1991/eat_tensorflow2_in_30_days. [11] has the sample-specific trigger, which means the size of the figure has to be 3x224x224 while the training data has only 1 dimension, which means my poisoning way can't be realized only if I change the input data size, while [7] is different. (describe the progress/update of your research paper

6.2 tasks plan to do in the next week

- Action 1: Papers plan to read
- Action 2: Progress goal of your paper/research
- Action 3: I'm reading the code in <https://github.com/ebagdasa/backdoors101>, however, I give up some time later because I haven't planned to perform the attack in this form. I want to realize some codes(others

7 2023. Week 4

Please use the following template for the weekly report. (I am working on a couple of paper deadlines and will circle back to you asap. Keep working on the experiments) [got it!](#)

7.1 tasks done in this week

7.1.1 paper summary

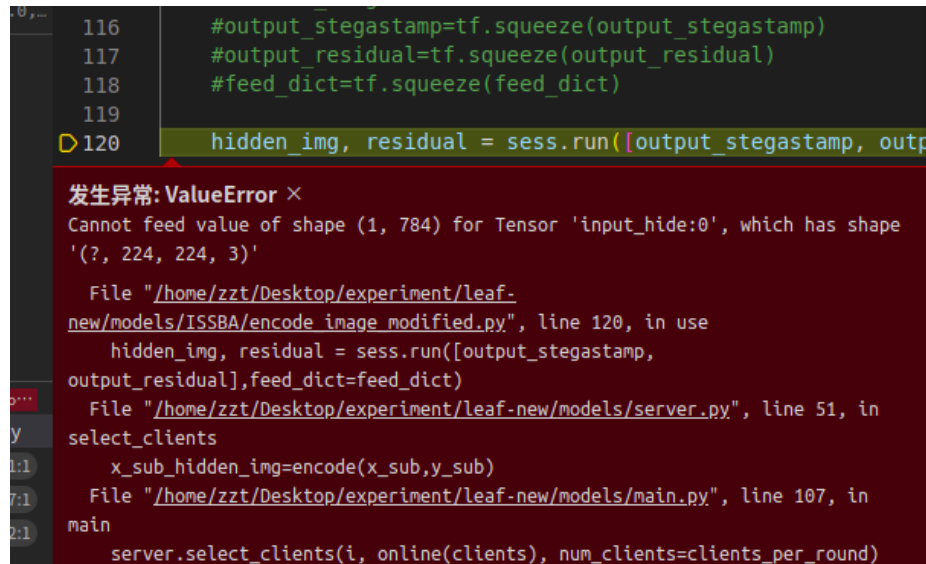
- **Paper Info:** [14] about robust FL framework.
- **Research Problem:** a defensive way to help defend the backdoor attack in FL.
- **Ideas/Novelty:** propose the first certifiably robust federated learning (CRFL) framework. the paper figure out a method to clip and perturb parameters of FL. In order to perturb, the paper adds some Gaussian noise to the aggregated model to strengthen the robustness of FL, which builds up its robustness in FL. (Describe the idea or methodology proposed in this paper. Keep it in one paragraph.
- **Your Thoughts:** I think I have to perform a kind of model replacement to attack FL. Luckily, this paper has its code including model replacement attack <https://github.com/AI-secure/CRFL>. Then can I think about figuring out a defensive way to perform the defense. Before then, I was thinking what kind of noise I add will help better. I mean the specific parameter of the Gaussian noise. (Pros and Cons of the paper. Any takeaway message? Any thoughts to make it better or benefit to your own research problem?

7.1.2 paper summary

- **Paper Info:** [16] is about investigating poisoning attack against horizontal federated machine learning. There are several kinds of FL: 1) horizontal (sample-based) federated learning, 2) vertical (feature based) federated learning, 3) federated transfer learning
- **Research Problem:** it is solving optimal poisoning attack strategies in the FL.
- **Ideas/Novelty:** poisoning attack is a kind of optimal problem. (Describe the idea or methodology proposed in this paper. Keep it in one paragraph.
- **Your Thoughts:** Pros and Cons of the paper. Any takeaway message? Any thoughts to make it better or benefit to your own research problem?

7.1.3 progress of your research paper

First, I need to look through more papers so that I can open up my view. Second, I might write the code according to some algorithm frameworks or something else because I can't write the code blindly and learn from nothing. Third, I'll keep an eye on the paper I read thinking about whether [1] has conveyed the information how to perform the attack: scaling down model updates to make them harder to detect (e.g., train-and-scale [7]). I should be certain what kind of research is valuable, then I need to try to do the experiment. When loading, ISSBA.pb has some problems. I need to focus on attacking methods first, then I should care about some innovations in the defensive side. In terms of defending, how about a defensive way to defend both data poisoning and model poisoning?1. Thanks to the video <https://www.youtube.com/watch?v=foflxVMuF6A>, I will modify my way of reading paper, being more efficient.



```
116 #output_stegastamp=tf.squeeze(output_stegastamp)
117 #output_residual=tf.squeeze(output_residual)
118 #feed_dict=tf.squeeze(feed_dict)
119
120 hidden_img, residual = sess.run([output_stegastamp, output_residual], feed_dict=feed_dict)

发生异常: ValueError ×
Cannot feed value of shape (1, 784) for Tensor 'input_hide:0', which has shape '(?, 224, 224, 3)'

File "/home/zzt/Desktop/experiment/leaf-new/models/ISSBA/encode_image_modified.py", line 120, in use
    hidden_img, residual = sess.run([output_stegastamp, output_residual], feed_dict=feed_dict)
File "/home/zzt/Desktop/experiment/leaf-new/models/server.py", line 51, in select_clients
    x_sub_hidden_img=encode(x_sub,y_sub)
File "/home/zzt/Desktop/experiment/leaf-new/models/main.py", line 107, in main
    server.select_clients(i, online(clients), num_clients=clients_per_round)
```

Figure 1: c.train_data

I'd better learn something deeper. Besides, I've learnt that some poisoning attack might not that good maybe because the backdoored data is so close to the benign data. I can use norm or cosine value to measure that. (describe the progress/update of your research paper;

7.2 tasks plan to do in the next week

- Action 1: Papers plan to read
- Action 2: Progress goal of your paper/research
- Action 3: read the code of DBA in terms of model replacement.(others

8 2023.01 Week 3

I think I've got a COVID-19 again, Jesus. Luckily, this time ,my body reaction is not that serious. I just need to spend double time eating and sleeping since I am so hungry and my body needs so much food to restore its health. In the future, I have to eat the meal in my room instead of eating outside without the risk of taking off my medical mask.

Please use the following template for the weekly report. (I am working on a couple of paper deadlines and will circle back to you asap. Keep working on the experiments) [got it!](#)

8.1 tasks done in this week

8.1.1 paper summary

- **Paper Info:** [17] introduces some ways of backdoor and defenses in a primary phase.
- **Research Problem:**
- **Ideas/Novelty:** Describe the idea or methodology proposed in this paper. Keep it in one paragraph.
- **Your Thoughts:** In terms of the experiment, Attack frequency and Number of attackers are things that I can take into my consideration. Random sampling vs. fixed frequency attacks, frequency attacks being slightly more effective than random sampling attacks, besides, it is easier for me to conduct the experiment. (Pros and Cons of the paper. Adding Gaussian noise (norm bound = 5). (Any takeaway message? Any thoughts to make it better or benefit to your own research problem?

8.1.2 progress of your research paper

In order to backdoor the training data of the client, I need to understand how to use the variable-Client. Then I read the code such as <https://github.com/SCLBD/BackdoorBench> and ,but it may not work since they have constrained some ways of performing, which may not be working for me to modify? <https://github.com/SCLBD/BackdoorBench> is quite complicated. Then I understand ISSBA just read the Image data and process the Image data by using Pillow. Finally, I understand that I just need to traverse the dictionary, which is the training data of any clients, then I can successfully poison the data. I've recently modified some codes ² Besides, I've newly found this <https://zhuanlan.zhihu.com/p/541776225>, maybe it will work in the future. However, data poisoning is not that effective. I'm thinking whether I need to read some codes about model replacement so that I can perform my experiment in a valuable way.

describe the progress/update of your research paper

```
e_modified.py 1, U  test.py 4, M  main.py M  server.py M X  ▶  ⓘ
leaf-new > models > server.py > Server > select_clients
19  NOTE that within function, num_clients is set to
20  min(num_clients, len(possible_clients)).
21
22  Args:
23      possible_clients: Clients from which the server can select
24      num_clients: Number of clients to select; default 20
25  Return:
26      list of (num_train_samples, num_test_samples)
27  """
28  num_clients = min(num_clients, len(possible_clients))
29  np.random.seed(my_round)
30  self.selected_clients = np.random.choice(possible_clients, num_clients)
31  for c in self.selected_clients:
32      """ if(c.id%5==0):#string 'f1019_45'
33          import pdb; pdb.set_trace() """
34      #import pdb; pdb.set_trace()
35      # method 1
36      x,y=c.train_data["x"],c.train_data["y"]
37
38      # wrong x,y=zip(c.train_data["x"],c.train_data["y"])
39      print(x,y,sep="<< >>")
40      for x_sub,y_sub in zip(x,y):
41          print(x_sub,y_sub,sep="<>") #process the information
42      """ for key in c.train_data:
43          if(key=='x'):
44              x=key
45          elif (key=='y'):
46              y=key#second try;
47          #poison the data; ISSBA (c.train_data[key])
48          print(c.train_data[x])
49
50          print(c.train_data[y])
51          pass """
52  #xs, ys = zip(*random.sample(list(zip(self.train_data["x"], self.train_data["y"])), num_clients))
```

Figure 2: c.train_data

8.2 tasks plan to do in the next week

- Action 1: [14] Papers plan to read
- Action 2: use this platform to see whether I can add some Gaussian noise <https://github.com/SCLBD/BackdoorBench> (Progress goal of your paper/research)
- Action 3: do the experiment about data poisoning since there is much more help from this direction. (others

9 2023.01 Week 2

Please use the following template for the weekly report.

9.1 tasks done in this week

9.1.1 paper summary

- **Paper Info:** Title, authors, conference/journal [5](a little difficult to read.
- **Research Problem:** Detecting the attack of inserting backdoors or trojans into the model is challenging(Describe the research problem in brief.
- **Ideas/Novelty:** we propose a novel approach to backdoor detection and removal for neural networks. The activation network will behave differently when samples learn the features, which represents how the network makes its decisions. Here is the reason: The standard samples learn the features from the input, however, the poisoned samples learn features from the source class and the backdoor trigger. (Describe the idea or methodology proposed in this paper. Keep it in one paragraph.
- **Your Thoughts:** Pros: the defensive and repairing result of experiment is quite good.
Cons:Activation Clustering(AC) is quite difficult to understand. There are 2 key words. One is activation, as explained above. Then is clustering. Only by clustering some labels or data, then everything will be clear.
Any takeaway message?: reading its way of doing experiment to prove the effectiveness of defenses. use it to defend some other backdoor attacks such as DBA? and see whether we can improve the defensive ways? Maybe I need to check the defensive flaws in the future, which will be quite tough. I need to figure out how the mechanism of AC works? My question is that whether different attacks need different defenses (Pros and Cons of the paper. Any takeaway message? Any thoughts to make it better or benefit to your own research problem?
- code:An adversarial example library for constructing attacks, building defenses, and benchmarking both.<https://github.com/cleverhans-lab/cleverhans>. I've known some details about how the model is built by using keras and tensorflow fromhttps://tensorflow.google.cn/tutorials/customization/custom_layers?hl=zh-cn. Besides, I've tried to tackle the problem of tf.combat.v1 and learn some more things. In the end, I just start to read other codes like

9.1.2 progress of your research paper

I want to (describe the progress/update of your research paper

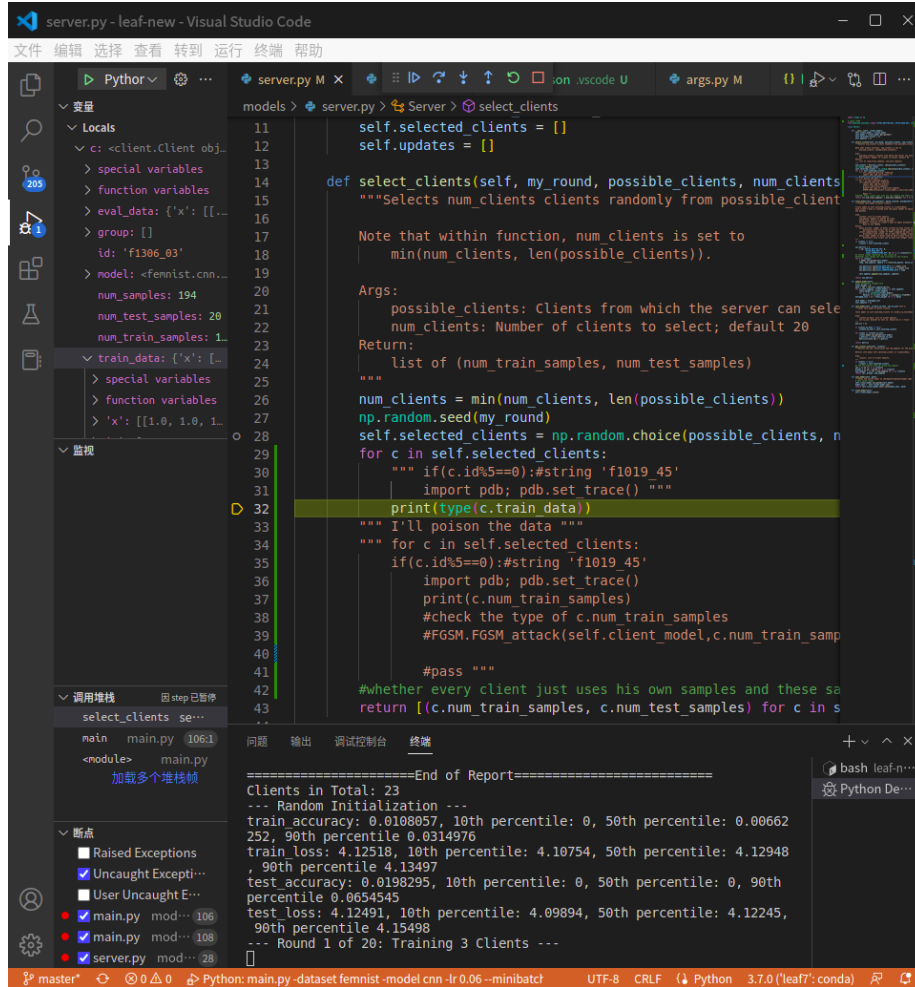


Figure 3: c.train_data

9.2 tasks plan to do in the next week

- Action 1: (Papers plan to read
- Action 2: I've known how the client uses its own training data, the format of data is quite different. (Progress goal of your paper/research
- Action 3: Read the code; learn some codes from the <https://www.tensorflow.org/guide/>, I need to understand what type of data is put into the `client.train_data`, and know how to change those data by using backdoor learning. (others

10 2023.01 Week 1

Please use the following template for the weekly report. (I am working on a couple of paper deadlines and will circle back to you asap. Keep working on the experiments) [got it!](#)

10.1 tasks done in this week

10.1.1 paper summary

- **Paper Info:** I can only know it is published on Arxiv. [24] Title, authors, conference/journal. <https://github.com/Greilfang/Loss-Tolerant-Federated-Learning>. It studies the relationship between communication efficiency and some other schemes, which might be quite different from my research about backdoor learning and task accuracy. Thus, I pause when I read section 3. However, I got to know some other FL algorithm.
- **Research Problem:** to guarantee the communication efficiency, Recent solutions have been focusing on threshold-based client selection schemes. However, the schemes have some flaws, which can cause biased client selection and results in deteriorated performance. Thereby, In this paper, we explore the loss tolerant federated learning (LT-FL) in terms of aggregation, fairness, and personalization.
Input: We use a mobile broadband dataset, which is quite different from femnist if I use LEAF. If I want to research communication capabilities of backdoor attack on FL, I might use other benchmarks, which might be more convenient for coding, I guess.
- **Ideas/Novelty:** There are mainly 1 innovations. One is for accelerating the data uploading by using their self-invented algorithm named ThrowRightAway (TRA). TRA tweaks the aggregation algorithm to compensate for the lost information.
As a result, the communication efficiency such as the performance has improved! Specifically, it is the personalization and fairness performance in the face of packet loss below a certain fraction
- **Your Thoughts:** Right now, I don't think I can have any takeaways from the paper because of different kinds of input. Before, I think The client selection scheme might be applied for me. The benefit of the scheme is to solve the problem of bias caused by unfair client selection. The proposed loss-tolerant scheme (TRA) not only to address communication efficiency, but also to guarantee fairness during client selection. If I need to research the relation between the communication efficiency and FL, I might (Pros and Cons of the paper. Any takeaway message? Any thoughts to make it better or benefit to your own research problem?)

10.1.2 progress of your research paper

I want to specify which part of the code in LEAF uses the `training_num_samples`. However, I find that there is no function to use the parameter of `training_num_samples` in terms of each client, which is really frustrating. Then I heard from my friends-Qian Chen that he once used the VScode to debug the code and watched the change of the variable, which can help him better understand the code and find the potential bug. As for me, I started to install the environment of LEAF in my win10. However, it turns out win10 cannot work as good as ubuntu. Then I have to buy another laptop to add ubuntu system, but I don't think VScode can debug the code in my virtual environment. I need to ask my friend and search more.

However, my goal of the code is to figure out which part for the client might use the `training_num_samples`. c.num I try to debug on the linux server by referring to <https://www.geeksforgeeks.org/python-debugger-python-pdb/>. I'm not using it quite well and I can't use it to satisfy my needs, yet. Thus, I don't think pdb has too much use for me. Later I might use the VScode for debugging by referring to vpdb https://www.bilibili.com/video/BV1Yb4y1k7oR/?spm_id_from=333.788.recommend_more_video.2&vd_source=9f9e49aa47a18d15ec3111409003baff. Besides, I've found some other attack methods named FGSM. Actually, I've known that some training data might be existing in some variables. I just need to identify how the client uses those variables and poison that. some codes are shown in Figure 4 I'll do the experiment when my new platform is installed. No, I think I write the wrong code. I just mix up adversarial examples and backdoor attack. I didn't consider the backdoor trigger. (describe the progress/update of your research paper)

10.2 tasks plan to do in the next week

- Action 1: [5] Papers plan to read
- Action 2: Try pdb in the linux server to debug my code. Read some other codes in the future. https://github.com/ebagdasa/backdoor_federated_learning and <https://github.com/ebagdasa/backdoors101>. But the former uses pytorch 1.0. (Progress goal of your paper/research)
- Action 3: I want to specify one problem: How the client trains its own local model. I think he must use his own `training_num_samples` whereas I didn't find the function he use to train his local model.(others

```

def select_clients(self, my_round, possible_clients, num_clients=20):
    """Selects num_clients clients randomly from possible_clients.

    Note that within function, num_clients is set to
    min(num_clients, len(possible_clients)).

    Args:
        possible_clients: Clients from which the server can select.
        num_clients: Number of clients to select; default 20
    Return:
        list of (num_train_samples, num_test_samples)
    """
    num_clients = min(num_clients, len(possible_clients))
    np.random.seed(my_round)
    self.selected_clients = np.random.choice(possible_clients, num_clients, replace=False)

    """ I'll poison the data """
    for c in self.selected_clients:
        if(c.id%5==0):
            c.num_train_samples
            pass

    return [c.num_train_samples, c.num_test_samples] for c in self.selected_clients]

```

Figure 4: debug it in the future

11 2023. Week 1:

11.1 Tasks done in this week

- paper:
 - Problems:
 - Solutions:
 - Experiment(how the paper proves its idea is right):
 - Result:
- My idea:
- code: I need to follow the paper about data poisoning to attack FL. just follow the way in 2022.10 Week 5. Initially, I intend to follow [1] to combine the loss function to train the client. However, I don't know where I can get the code about loss function. I intended to follow the https://tensorflow.google.cn/js/guide/train_models?hl=zh-cn by using the Core API. Right now, I'll follow the "To compare with prior work, we also experiment with the pixel pattern backdoor [24]. During the attacker's training, we add a special pixel pattern to 5 images in a batch of 64 and change their labels to bird. Unlike semantic backdoors, this backdoor requires both a training-time and inference-time attack (see Section 4.1)." Right now, I'm reading the code in https://github.com/ebagdasa/backdoor_

`federated_learning` and <https://github.com/ebagdasa/backdoors101> so that I might figure out how to perform the data poisoning, since LEAF only writes the code about how the clients update some parameters, while backdoor learning about images will change the picture itself. I'll read the code more carefully about how LEAF works in terms of some selections .

- **Problems:** I can't locate where the function is when I press CTRL+MOUSE_LEFT, just shown in Figure 7. ISSBA just processes the image data, which needs the path of the image. Luckily, I think [23] has explained to me how I should use the poisoned attack. When I try to change the code, I suddenly find that I have to understand some details and it really works! I think I've known some ways to do the experiment. Finally, I understand when I need to construct a specific loss function, then I will try to search some methods, compare their realization and decide which method I might take in the end! I'll use the encoded images to poison the data. I've completed the first part demo and I know which part I might code in the LEAF.
I think I've known how to perform the attack by using FGSM. Later I need to carefully read the paper and some other codes.
- **Result:**

11.2 Tasks plan to do in the next week

Action Required (AR)

- AR1 : complete the ISSBA backdoor images in terms of clients.
Create a demo.py
- AR2:
- AR3:

让我们训练模型:

```
const optimizer = tf.train.sgd(0.1 /* learningRate */);
// Train for 5 epochs.
for (let epoch = 0; epoch < 5; epoch++) {
  await ds.forEachAsync(({xs, ys}) => {
    optimizer.minimize(() => {
      const predYs = model(xs);
      const loss = tf.losses.softmaxCrossEntropy(ys, predYs);
      loss.data().then(l => console.log('Loss', l));
      return loss;
    });
  });
  console.log('Epoch', epoch);
}
```

以上代码是使用 Core API 训练模型时的标准方法:

Figure 5: TF Loss

Algorithm 1 Local training for participant's model

FedLearnLocal(\mathcal{D}_{local})

Initialize local model L and loss function l :

$L^{t+1} \leftarrow G^t$

$\ell \leftarrow \mathcal{L}_{class}$

for epoch $e \in E$ **do**

for batch $b \in \mathcal{D}_{local}$ **do**

$L^{t+1} \leftarrow L^{t+1} - lr \cdot \nabla \ell(L^{t+1}, b)$

end for

end for

return L^{t+1}

Figure 6: FL Loss training

```

57         BYTES_READ_KEY: 0,
58         LOCAL_COMPUTATIONS_KEY: 0} for c in clients}#the way of loop
59
60 #I should change something in the client.py
61 #engineer just change the code according to the formula
62 for c in clients:
63     c.model.set_params(self.model)
64     comp, num_samples, update = c.train(num_epochs, batch_size, minibatch)
65
66     sys_metrics[c.id][BYTES_READ_KEY] += c.model.size
67     sys_metrics[c.id][BYTES_WRITTEN_KEY] += c.model.size
68     sys_metrics[c.id][LOCAL_COMPUTATIONS_KEY] = comp
69
70     self.updates.append((num_samples, update))
71
72 return sys_metrics
73
74 def update_model(self):

```

Figure 7: do not know how to locate the function

12 2022. Week 12:

12.1 Tasks done in this week

This week I might get a COVID-19, which takes me almost a week to recover. That's why I read few papers.

- paper: [23] is another paper about data poisoning. The innovation is that it just comes up with a poisoning idea to attack FL . Besides, I recently heard about Diffusion Model, which can generate new images. Maybe it's effective to replace GAN to generate some poisoned images.
 - * **Problems:** There is no problem. [23] just studies and evaluates a poisoning attack
 - * **Solutions:** [23] generates samples of other benign participants using GAN.
 - * **Details:** [23] trains a GAN to mimic prototypical samples of the other participants' training set. Then these generated samples will be fully controlled by the attacker to generate the poisoning updates, and the global model will be compromised by the attacker with uploading the **scaled** poisoning updates to the server.
 - * **Experiment:** In terms of how GAN is used, G is used to generate the final image, while D is copied from the local model. Then, by training GAN, G has the ability to generate the poisoned images.
 - * **Result:**
 - My idea:
 - code:

12.2 Tasks plan to do in the next week

Action Required (AR)

- AR1 :
- AR2:
- AR3:

13 2022. Week 11:

13.1 Tasks done in this week

- Paper: to begin with, [23] is another paper about data poisoning. [19] comes up with the idea that using adversarial examples to attack the model of image object detection. [19] uses GAN and combine some loss functions to train adversarial examples. Adversarial examples have better transferability
 - * **Problems:** [19] existing attacking methods for image object detection have **two limitations**: weak transferability and high computation cost
 - * **Solutions:** [19] presents a generative method to obtain adversarial images. In terms of **transferability**, we manipulate the feature maps extracted by a feature network. [19] proposes that first, UEA is used for attacking image and video detection, and second, a multi-scale attention feature loss is used to enhance the UEA's black-box attacking ability.
The paper utilizes some other architectures to realize their UEA in Sec.3.2.
 - * **Experiment:** In detail, we combine a high-level class loss and a low-level feature **loss** to jointly train the adversarial example generator.
 - * **Result:** our method **efficiently** generates image and video adversarial examples.
- My idea: Maybe I can use the same way to generate adversarial examples and use the DBA [22] to attack FL. But why should I do that? Maybe I just want to prove Adversarial examples combined with DBA have better performance than DBA itself.
Change the data: tweak data or generate adversarial data
Change the model directly: DBA or other methods
Think about, Which one is easy and accurate? If you combine both approaches, what can you improve?
Answer: Changing the data: generate adversarial data on FL, should be easier. But that can only prove data poisoning on FL is effective, and transferability can be effective-poisoning on different FL. Combining both methods can only prove attacking on FL can be effective. It doesn't solve any problems. It seems there is no innovation. Luckily, DBA has its code open <https://github.com/AI-secure/DBA>. If necessary, I can read and change the code. Besides, I can use AGI - Diffusion Model to replace GAN to generate data to perform the attack, which is a kind of data augmentations. Other data augmentations might only change the result of FL. <https://mp.weixin.qq.com/s/meCvSTfbryYdUfK5oga8tQ>
- code:

13.2 Tasks plan to do in the next week

Action Required (AR)

- AR1 :
- AR2:
- AR3:

14 2022. Week 10:

14.1 Tasks done in this week

Actually, this week I just planned to rush home because of COVID-19. Thus, I didn't read too much.

- paper: [22] utilizes the feature of FL with distributed BACKDOOR attack (DBA). Every malicious participant in DBA poisons the specific subset of training data by using the local trigger while one participant in the central attack poisons the training data by using the global trigger. By the way, All the local triggers will form a global trigger.
DBA is evaluated on four classification datasets with non-i.i.d. In terms of how [22] proves that DBA is better than the central attack, [22] shows some attacks including single-shot and multi-shot. DBA chooses the local trigger and the attacking interval smartly. Some explanation I haven't read carefully and understand well. Some new defenses: [6] proposed a novel defense based on the party updating diversity without limitation on the number of adversarial parties. It adds up historical updating vectors and calculate the cosine similarity among all participants to assign global learning rate for each party.
 - * Detail: [22] considers the location of the trigger and the TRIGGER GAP as well as the scale
 - * Problem: there is no new problem coming up in the paper. However, DBA is a new insight explained in the paper.
- My idea: [22] belongs to the pixel attack instead of the model replacement.
- code: <https://github.com/AI-secure/DBA> thanks to <https://github.com/ebagdasa/backdoors101>, there are still some better learning materials. https://github.com/ebagdasa/backdoor_federated_learning, also has the code and the paper I read. Right now, I'm wondering whether they write the code according to the short formula in the paper since it's tough.

14.2 Tasks plan to do in the next week

Action Required (AR)

- AR1 :I might need to read some codes in the future. Reasons are listed as follows. I didn't use the google scholar before, and now I fully understand why it is so useful and I've already found some other useful papers. [20] a new backdoor bench. Besides, there is also a new benchmark.

- AR2:
- AR3:

15 2022. Week 9:

15.1 Tasks done in this week

- paper: I just follow [1] in order to better understand some cites. After I follow some papers I've read the article [12] a little bit to better understand the code, though I think my understanding of the code is quite right. Thanks to the [12], I can better understand [1]. [19] introduces a new Transferable Adversarial Attacks for Image and Video Object Detection. [19] says that image object detection have two limitations: weak transferability and high computation cost. To address these issues, we present a generative method to obtain adversarial images and videos. [3] conveys the original idea of FL. [17] There are some concepts: Under targeted attacks (often referred to as backdoor attacks), the goal of the adversary is to ensure that the learned model behaves differently on certain targeted sub tasks while maintaining good overall performance on the primary task. Under untargeted attacks, the goal of the adversary is to corrupt the model in such a way that it does not achieve a near-optimal performance on the main task. Anyway, I'll focus on how to implement backdoor learning. Maybe I'll learn some adversarial learning in the future. [13] can't be applied to federated learning. [2] comes up the question that distributed machine learning frameworks have largely ignored the possibility of arbitrary (i.e., Byzantine) failures. However, [10] has listed several backdoor attacks but some attacking methods will not influence FL based on the attacking principle.
- code: I think I can nearly understand the code already from the information that the FL paper provides, anyway, if I need to change the code, I might need to .
I didn't use the google scholar before, and now I fully understand why it is so useful and I've already found some other useful papers. [20] a new backdoor bench

15.2 Tasks plan to do in the next week

Action Required (AR)

- AR1 :
- AR2:
- AR3:

16 2022. Week 8:

16.1 Tasks done in this week

- paper: [1] introduces FL a little bit. It says that using the model replacement is more powerful than training-data poisoning. In terms of the defense, anomaly detection would not have been effective in any case. In order to evaluate the effectiveness of model replacement and the defensive effectiveness such as anomaly detection, the paper also develops and evaluates a generic constrain-and-scale technique that incorporates the evasion of defenses into the attacker's loss function during training. For the experimental part, the paper implements tasks on image classification on CIFAR-10 and word prediction on a Reddit corpus. Assumptions: neither defenses against data poisoning, nor anomaly detection can be used during federated learning because of the privacy.

Background: part 2.1 has introduced several traditional poisoning attacks. Traditional defenses against model attacks cannot be effective because of the privacy, which means defensive methods cannot scan the data among users. Besides, some traditional attacks are out of date because traditional defenses can easily detect them. part 4 lists 3 ways of attacking. It introduces some attacking methods and lists their advantages and disadvantages eg. backdoor attacking, adversarial examples and model replacement. part 4.1 traditional poisoning attacks: a fraction of data is controlled. By contrast, in federated learning the attacker controls the entire training process—but only for one or a few participants

Part 5.1 introduces that backdoors can select 2 kinds of features to attack such as a naturally occurring feature and a unnaturally occurring feature. In order to achieve the effect, the paper scales up the weights of the backdoored model X by $\gamma = \frac{n}{\eta}$ Pixel-pattern backdoor requires both training-time and inference time control over the images. In the example of semantic backdoor, the paper targets images such as cars with certain attributes that are classified as birds

part 6 just shows some differences when rounds of attack are injected.

- my idea: if replacing the model is the trend, then maybe poisoning data can be put off, anyway there are not too many possibilities to solve the researching problem. Then those papers about poisoning data might be useless.
my puzzle: "the L2 norm of the attacker's update." I don't understand it. Maybe it means that the differential level of update.
- code: I finally found where I should make the clients to put data onto the server and know how the client processes the data. Maybe I need to generate the malicious client every 1 of 5 benign clients, then these

malicious clients will use the backdoor attack/script to process the data. I think that will work finally. The client needs to send updates to the server. I'd better read some papers with codes in the future so that I might better understand some changes the researcher made to the model itself. Right now, I just know a little bit about which part I need to change. In terms of the equation that I follow and change the part to update the model, maybe that is quite challenging for me right now.

- issues:

16.2 Tasks plan to do in the next week

Action Required (AR)

- AR1 : learn the code among LEAF to detect what kind of changes I need to do so that the malicious clients can be simulated.
- AR2:
- AR3:

17 2022. Week 7:

17.1 Tasks done in this week

[?] Should I wait for the further direction? I just follow the routine and finish something that I haven't finished yet. How can I even report the paper that I might need to read? Ummm,actually, I haven't read the paper, yet. What's my target? I should ask myself. I can only poison some parts of the image and plot the accuracy vs round.time figure. According to How to backdoor federated learning, it has some samples of figures. Maybe my task should be designing those plots by using LEAF. However, I don't know whether Pro. Yang has any existing tools or advice. I just go to read the figure. I'm reading the part 5 in the paper about image classification. I should have some insights about the figure that I might plot, otherwise, there is no use for me to process the image. Thereby, reading paper is necessary. I understand that I'm doing single-shot attack VS accuracy at present in terms of the percentage of figures that I might poison in the experiment. In terms of repeated poison attack over 100 rounds, I just need to modify the code in order to pause the training for FL every 100 rounds. The percentage of attackers can be difficult to simulate the task. I find that I'm focusing on how to do the experiment instead of why I do the experiment like that. Anyway, it is a kind of way to proactively study since I don't know what the insight I might get in the future. I haven't designed my research plans just like other PhD candidates do. Maybe that's the gap!

17.2 Tasks plan to do in the next week

Action Required (AR)

- AR1 :
- AR2:
- AR3:

18 2022.10 Week 6:

18.1 Tasks done in this week

In order to use the ISSBA to attack the model, I have to modify the original picture to the size of (224,224,3), and then encode the image. In the end, I might need to plot some accuracy_line vs others. Besides, I also need to change the preprocess files to make new .json files. Actually, I'm thinking about what kind of plots that I need to plot in the future. Now I can have the complete poisoned data, then I need to use the .json files to process the data. In order to process the complete data, and generate proper .json files. I'd better process the data after putting some poisoned data to the specific folder. The following steps actually are to process the data.

Should I read some papers to know what kind of figures I need to plot in terms of backdoor learning attacking FL model? I just need to take a try.

I intend to read How To Backdoor Federated Learning to see their figures. Maybe I can figure out what kind of figures I might need to plot. Right now, I know the steps to do the experiment. First, I need to read some papers to get some ideas about new attacking methods on FL. Then I shall design some plot.py in the LEAF to show the normal state. Then use some innovative attacking methods on LEAF, and see the result! Should I read some papers to get some innovation and plot the figure or just plot accuracy-round_num figure?

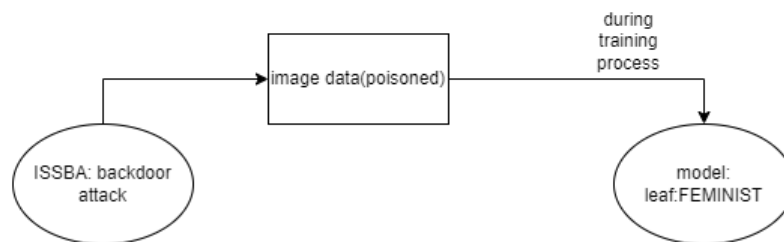
18.2 Tasks plan to do in the next week

Action Required (AR)

- AR1 :
- AR2:
- AR3:

picture changes:

I need to modify the size of the picture,
which achieves (224,224,3)



summerize

In this way, I don't need to write or
change the code too much,
because I have the existing code.

drawing picture:

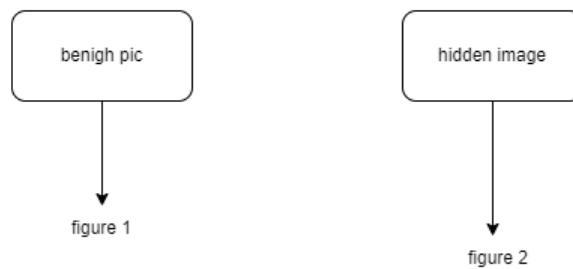


Figure 8: plotting ideas

CIFAR image classification:

single-shot attack

1 attacker selected in round 0

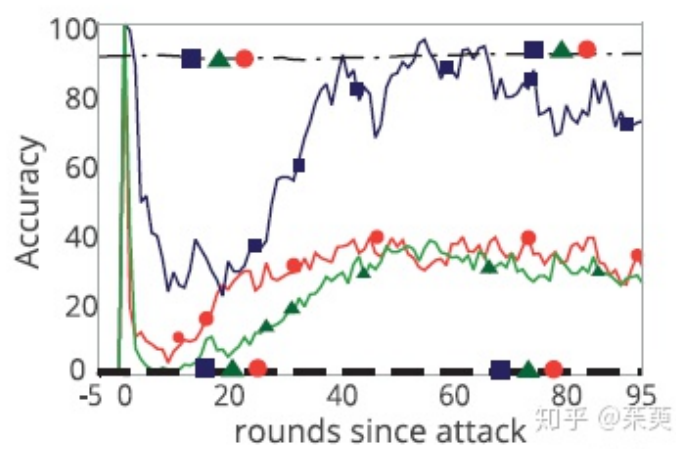


Figure 9: downloaded from the Internet,not mine

19 2022.10 Week 5: read the code of the backdoor attack –ISSBA to attack FL

19.1 Tasks done in this week

I've read the code and know how it works, anyway. according to the code, I think I need to use the ISSBA to process the image, and use those images to train the FL model, afterwards. I'll add some poisoned image into the group. in this way, I don't need to change the code too much.

I need to regenerate the result in the paper so that I can get what I want. However, when I'm trying to process the image by using the ISSBA, I've found some problems. thanks to my friend, I finally solve the code problem, but it is a little tricky. I don't know whether this is right.

Maybe I need to know how to trigger the model to misbehave, thus, I need to modify the code in the leaf, which will be a huge challenge for me. Haha. I need to add a backdoor trigger to the image. Anyway, I need to understand the test.py in ISSBA. Just to use the hidden_image

The basic idea of backdoor attack is to tweak the image and change the label. You actually do not need to change too much about the code. But you need to show the accuracy of the main task and the accuracy of the backdoor.

the problem lies in the input dimension, and `sess.run()` just returns a single value. Then I need to squeeze the input tensor. But it doesn't work at all, and I'm reading some documents to understand how the function works and what is the problem with my code.

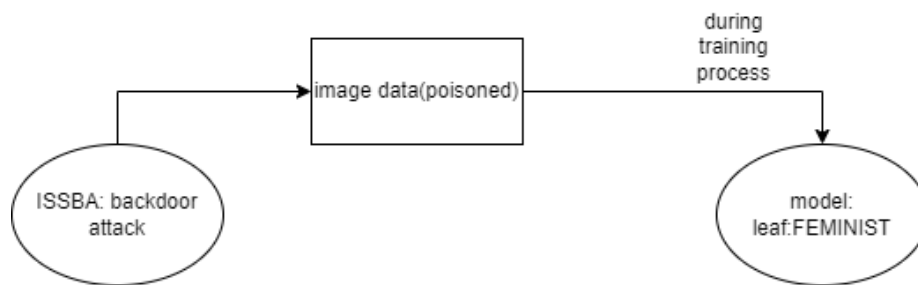
19.2 Tasks plan to do in the next week

Action Required (AR)

- AR1 :
- AR2:
- AR3:

picture changes:

I need to modify the size of the picture,
which achieves (224,224,3)



summerize

In this way, I don't need to write or
change the code too much,
because I have the existing code.

Figure 10: design ideas

20 2022.10 Week 4: keep trying the Twitter Sentiment Analysis

20.1 Tasks done in this week

at present, I don't know why I should plot the figure 1,2,3. Maybe I need to read the leaf paper again, and see what I want to get in the end. Anyway, every figure has some kind of meaning. I need to find out what data I'm processing, and why the main.py is like that.

recently, after I read the link <https://zhuanlan.zhihu.com/p/164659496> and main.py, I've gradually understood something. I need to have a complete view of the whole thing.// I want to plot the figure-Femnist. and right now, I finally understand that I must plot the figure myself. Anyway, all the functions I need to design myself according to my plan. However, I haven't found any papers about FEMINIST, and I just totally have no idea what the meaning of some symbols in one figure. Right now, actually, I'm a little bored. I'm reading the Server.py right now.

20.2 Tasks plan to do in the next week

Action Required (AR)

- AR1 :
- AR2:
- AR3:

21 2022.10 Week 3: keep trying the Twitter Sentiment Analysis

21.1 Tasks done in this week

using python=3.7, I can successfully download the data. Then when I try to use the main.py, there exists one issue. I will go to the github issues to check, but there is no the same issue. I'm trying to figure it out.

recently, I've learnt that there might be the problem of the version of tensorflow. I'm using tensorflow==1.13.1. I'm still curious about the setup. Finally, I've successfully solved the issue! it is the cause of the version of numpy. <https://github.com/TalwalkarLab/leaf/issues/56>

I don't know whether Prof. Yang is willing share his metrics/plots.py, because I don't know which part I can use to modify the code or plot some figures that I want. I'm using metrics/metrics.ipynb. I've met another problem of keyerror in using the pandas. I think designing the plots.py myself should be necessary. I think plots.py that Dr.Amy is using can function well and I think there is no grammar mistake in that file. Anyway, I think the code of leaf on the github has been a little old that it has such an error when I use that metrics.py. the error occurs between pd.read_csv() and the pd.sort_values.

however, when designing the plots.py, I think I still need to design the programme myself.

I finally know it is the error in the main.py that it hasn't added column name to the sys_metrics.csv.

21.2 Tasks plan to do in the next week

Action Required (AR)

- AR1: I believe that plotting one figure should be very important.
- AR2: I want to know why I should assign the number of clients when running the code, and some metrics of comparing the data. I want to know what kind of information that I really want to get from plotting the figure.
- AR3: I need to read the main.py carefully since there have been some mistakes over there.

numpy	1.16.4	pypi_0	pypi
openssl	1.1.1q	h7f8727e_0	https://mirrors.bfsu.edu.cn/anaconda/pkgsg/
packaging	21.3	pypi_0	pypi
pandas	1.1.5	pypi_0	pypi
pandocfilters	1.5.0	pypi_0	pypi
parso	0.8.3	pypi_0	pypi
pexpect	4.8.0	pypi_0	pypi
pickleshare	0.7.5	pypi_0	pypi
pillow	9.2.0	pypi_0	pypi
pip	22.2.2	py37h06a4308_0	https://mirrors.bfsu.edu.cn/anaconda/pkgsg/
pkgutil-resolve-name	1.3.10	pypi_0	pypi
prometheus-client	0.14.1	pypi_0	pypi
prompt-toolkit	3.0.31	pypi_0	pypi
protobuf	3.20.1	pypi_0	pypi
psutil	5.9.2	pypi_0	pypi
ptyprocess	0.7.0	pypi_0	pypi
pycparser	2.21	pypi_0	pypi
pygments	2.13.0	pypi_0	pypi
pyparsing	3.0.9	pypi_0	pypi
pyrsistent	0.18.1	pypi_0	pypi
python	3.7.13	h12debdf_0	https://mirrors.bfsu.edu.cn/anaconda/pkgsg/
python-dateutil	2.8.2	pypi_0	pypi
pytz	2022.4	pypi_0	pypi
pyzmq	24.0.1	pypi_0	pypi
qtconsole	5.3.2	pypi_0	pypi
qtpy	2.2.1	pypi_0	pypi
readline	8.1.2	h7f8727e_1	https://mirrors.bfsu.edu.cn/anaconda/pkgsg/
scipy	1.5.4	pypi_0	pypi
send2trash	1.8.0	pypi_0	pypi
setuptools	63.4.1	py37h06a4308_0	https://mirrors.bfsu.edu.cn/anaconda/pkgsg/
six	1.16.0	pypi_0	pypi
soupsieve	2.3.2.post1	pypi_0	pypi
sqlite	3.39.3	h5082296_0	https://mirrors.bfsu.edu.cn/anaconda/pkgsg/
tensorboard	1.13.1	pypi_0	pypi
tensorflow	1.13.1	pypi_0	pypi
tensorflow-estimator	1.13.0	pypi_0	pypi
termcolor	2.0.1	pypi_0	pypi
terminado	0.16.0	pypi_0	pypi
tinycss2	1.1.1	pypi_0	pypi
tk	8.6.12	h1ccaba5_0	https://mirrors.bfsu.edu.cn/anaconda/pkgsg/
tornado	6.2	pypi_0	pypi
traitlets	5.4.0	pypi_0	pypi
typing-extensions	4.4.0	pypi_0	pypi
wcwidth	0.2.5	pypi_0	pypi
webencodings	0.5.1	pypi_0	pypi
werkzeug	2.2.2	pypi_0	pypi
wheel	0.37.1	pyhd3eb1b0_0	https://mirrors.bfsu.edu.cn/anaconda/pkgsg/
widetsnbextension	4.0.3	pypi_0	pypi
xz	5.2.6	h5eee18b_0	https://mirrors.bfsu.edu.cn/anaconda/pkgsg/
zipp	3.8.1	pypi_0	pypi
zlib	1.2.12	h5eee18b_3	https://mirrors.bfsu.edu.cn/anaconda/pkgsg/

Figure 11: leaf:readme.md

22 2022.10 Week 2: try the Twitter Sentiment Analysis

22.1 Tasks done in this week

I've carefully analyze the code in the server.py - train-model function. But I think it is time for me to run the experiment to be familiar with FL. Thereby, I just go to the <https://leaf.cmu.edu/build/html/tutorials/sent140-md.html> directly.

Actually, reading the paper and github file can't solve my puzzle.

Is it necessary that "LEAF must been used under Python 3.5"? I've created 2 virtual environments. One is that of python=3.7, the other is under python=3.5.

and when I do the experiment, I've met with one problem. I don't know which person I can communicate with regarding the setup of the environment and do the experiment smoothly. I've seen that "LEAF has been used under Python 3.5".

Actually, I think Python 3.5 is no use because tensorflow==1.10.0 can't be installed successfully.

22.2 Tasks plan to do in the next week

Action Required (AR)

- AR1:
- AR2:
- AR3:

23 2022.10 Week 1: learn the code in ISSBA

23.1 Tasks done in this week

I've known from test.py how the ISSBA attack the model. Thereby, maybe I just need to change the model from ResNet-18 to one model of FL. However, I need to read the paper and train.py again to check whether there should be some other changes to be made.

However, I haven't found the FL models in <https://github.com/TalwalkarLab/leaf>, I've checked the folder listed in the graph and found that there exists no minibatch SGD, FedAvg [22] and Mocha [30].

LEAF only provides the default FedAvg algorithm (you can see it in the server.py - train-model function

minibatch: fraction of client's data to apply minibatch sgd, None to use FedAvg

got it!

after I read the readme.md in the folder named model, actually, I think there only exists the model named CNN and stacked lstm.

anyway, I'll install the related work to check how to use FL so that I can be more familiar with the FL. I finally understood what the advisor means!!!Ahhhhh!

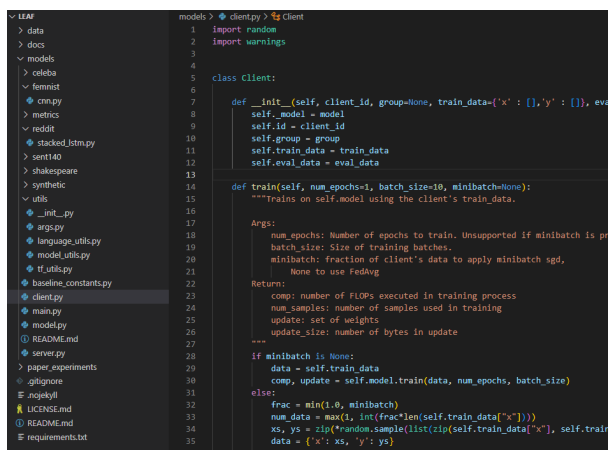


Figure 12: leaf:readme.md

23.2 Tasks plan to do in the next week

Action Required (AR)

– AR1 :

- AR2:
- AR3:

24 2022.9 Week 1: learn the code

24.1 Tasks done in this week

I downloaded a bunch of codes from the Internet <https://github.com/yuezunli/ISSBA>, and started to look at the code detailedly from test.py . However, I can only know some superficial explanations from the Internet Actually, I think that has been already enough.

I recommend to try LEAF to get familiar with Federated Learning

<https://github.com/TalwalkarLab/leaf>

Then, you can implement the backdoor attacks on FL

!got it! After reading the paper , I've learned the paper includes 3 things, which are Datasets,Reference Implementations and Metrics,respectively. Though I've made some notes in the paper, I can recall nothing, which is quite a pity. Maybe that means I haven't understood the content of the paper,yet so I just can't convey the meaning and the information of the paper.

If necessary, I'll use one backdoor algorithm like ISSBA to attack one FL. I'll keep reading ISSBA part 5 carefully in the following days and use that to attack FL listed in <https://github.com/TalwalkarLab/leaf>.

besides, I will learn a lot from <https://space.bilibili.com/28613957>. Sorry, the truth is that they are speaking so fast, I'd better read the paper instead. However, the video can help me keep awake.

24.2 Tasks plan to do in the next week

Action Required (AR)

- AR1 : read the files carefully in <https://github.com/TalwalkarLab/leaf>
- AR2: research files in <https://github.com/yuezunli/ISSBA> about how to perform the attack on FL.
- AR3:

[?]

References

- [1] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. *CoRR*, abs/1807.00459, 2018.
- [2] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Byzantine-tolerant machine learning. *CoRR*, abs/1703.02757, 2017.
- [3] Kallista A. Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloé Kiddon, Jakub Konečný, Stefano Mazzocchi, H. Brendan McMahan, Timon Van Overveldt, David Petrou, Daniel Ramage, and Jason Roselander. Towards federated learning at scale: System design. *CoRR*, abs/1902.01046, 2019.
- [4] Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Zhenqiang Gong. Fltrust: Byzantine-robust federated learning via trust bootstrapping. *CoRR*, abs/2012.13995, 2020.
- [5] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Ben Edwards, Taesung Lee, Ian Molloy, and B. Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. *ArXiv*, abs/1811.03728, 2018.
- [6] Clement Fung, Chris J. M. Yoon, and Ivan Beschastnikh. Mitigating sybils in federated learning poisoning. *CoRR*, abs/1808.04866, 2018.
- [7] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Bad-nets: Identifying vulnerabilities in the machine learning model supply chain. *CoRR*, abs/1708.06733, 2017.
- [8] Shengshan Hu, Jianrong Lu, Wei Wan, and Leo Yu Zhang. Challenges and approaches for mitigating byzantine attacks in federated learning. *CoRR*, abs/2112.14468, 2021.
- [9] Keita Kurita, Paul Michel, and Graham Neubig. Weight poisoning attacks on pre-trained models. *CoRR*, abs/2004.06660, 2020.
- [10] Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
- [11] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.
- [12] H. Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. Federated learning of deep networks using model averaging. *CoRR*, abs/1602.05629, 2016.
- [13] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Inference attacks against collaborative learning. *CoRR*, abs/1805.04049, 2018.

- [14] Thien Duc Nguyen, Phillip Rieger, Hossein Yalame, Helen Möllering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Ahmad-Reza Sadeghi, Thomas Schneider, and Shaza Zeitouni. FLGUARD: secure and private federated learning. *CoRR*, abs/2101.02281, 2021.
- [15] Ali Raza, Shujun Li, Kim Phuc Tran, and Ludovic Koehl. Detection of poisoning attacks with anomaly detection in federated learning for healthcare applications: A machine learning approach. *ArXiv*, abs/2207.08486, 2022.
- [16] Gan Sun, Yang Cong, Jiahua Dong, Qiang Wang, Lingjuan Lyu, and Ji Liu. Data poisoning attacks on federated machine learning. *IEEE Internet of Things Journal*, PP:1–1, 11 2021.
- [17] Ziteng Sun, Peter Kairouz, Ananda Theertha Suresh, and H. Brendan McMahan. Can you really backdoor federated learning? *CoRR*, abs/1911.07963, 2019.
- [18] Wei Wan, Shengshan Hu, Jianrong Lu, Leo Yu Zhang, Hai Jin, and Yuan Yuan He. Shielding federated learning: Robust aggregation with adaptive client selection. In *International Joint Conference on Artificial Intelligence*, 2022.
- [19] Xingxing Wei, Siyuan Liang, Xiaochun Cao, and Jun Zhu. Transferable adversarial attacks for image and video object detection. *CoRR*, abs/1811.12641, 2018.
- [20] Baoyuan Wu, Hongrui Chen, Mingda Zhang, Zihao Zhu, Shaokui Wei, Danni Yuan, and Chao Shen. Backdoorbench: A comprehensive benchmark of backdoor learning, 2022.
- [21] Chulin Xie, Minghao Chen, Pin-Yu Chen, and Bo Li. CRFL: certifiably robust federated learning against backdoor attacks. *CoRR*, abs/2106.08283, 2021.
- [22] Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. Dba: Distributed backdoor attacks against federated learning. In *International Conference on Learning Representations*, 2020.
- [23] Jiale Zhang, Junjun Chen, Di Wu, Bing Chen, and Shui Yu. Poisoning attack in federated learning using generative adversarial nets. *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (Trust-Com/BigDataSE)*, pages 374–380, 2019.
- [24] Pengyuan Zhou, Pei Fang, and Pan Hui. Loss tolerant federated learning. *ArXiv*, abs/2105.03591, 2021.
- [25] Yuanbo Zhou, Wei Deng, Tong Tong, and Qinquan Gao. Guided frequency separation network for real-world super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.