A data-free approach for targeted universal adversarial perturbation[⋆]

Xiaoyu Wang, Tao Bai, and Jun Zhao

[1] Xi'an Jiaotong University
[2] Nanyang Technological University
wxystudio@stu.xjtu.edu.cn
bait0002@ntu.edu.sg
JunZhao@ntu.edu.sg

Abstract. The existence of adversarial example problem puts forward high demand on the robustness of neural network. This paper proposes a universal adversarial perturbation(UAP) attack method in data-free scenario, which can realize targeted attack to any class specified by the attacker. We design a unique loss function to balance the purpose of perturbing model and targeting label. As far as we know, our method is the first UAP attack method that can achieve targeted attack in data-free scenario. Especially, in federated learning a malicious user can fool other users' model without being noticed. We hope our attack method can inspire more researchers in the community, and enable them to better understand and defend against UAP attacks.

Keywords: Universal adversarial perturbation · Data free · federated learning.

1 Introduction

In the last few years, the vulnerability of deep neural network has received tremendous attention from academia. many studies [12] [6] show that neural network can be affected by adversarial example attack. Training with the model under attack, attacker obtains a small perturbation which can't be recognized by human(which is usually limited to -10 $\sim$ 10 pixel values). The attacker adds the unrecognizable perturbation to one natural image so that the model recognizes this image incorrectly.

But there are some drawbacks in adversarial example attack. Each perturbation only corresponds to one specific sample, which is cumbersome in launching attack. For this reason, universal adversarial perturbation [15] is proposed. Compared with traditional adversarial example attack, it does not generate a corresponding perturbation for each data, but creates a universal perturbation for all data in the whole dataset. The targeted UAP attack can make the data misclassified by the model into a specific category.

[⋆] Supported by organization Nanyang Technological University

What's not good enough is that, the above attacks totally rely on the attacker mastering the real dataset. In reality, this condition is very harsh. Therefore, some researchers proposed data-free attack, which uses membership-inference data [20], GAN-generation data [21], or creates data from Gaussian distribution [17] [16] to replace the original training dataset. For this kind of attack, the attacker doesn't need to have access to real data so that can launch attack more conveniently.

However, we found that none of the above methods can achieve targeted UAP attack in data-free scenario. In reality the attacker not only can't get in touch with the real training data, but also want to launch targeted attack. For example, in an autonomous driving scenario, it will be more threatening to confuse a model to recognize the "stop" road sign as "acceleration", than only regard it as "slow down". Based on the above analysis, this kind of attack is of great practical significance.

In this paper we introduce a more dangerous scenario. In 2016, Google[14] introduce federated learning, which can launch a cloud platform for distributedly training neural network with multiple users. It has two obvious advantages:

1. It can make use of all users' data to train a general model. Some people who are lack of data can train a model by crowdsourcing.
2. Users can deal with their own data locally without uploading their own data to the platform, which avoids some privacy issues.

In the federated learning scenario, each user possesses one portion of data. If a malicious user want to launch UAP attack to global model or even target one specific label, he will face the problem of data insufficiency. In this typical environment, previous UAP attack methods will fail.

For the above reasons, In this paper we propose a new universal adversarial perturbation attack method. Our method can achieve targeted attack in data-free scenario, and we have done experiments to verify in natural and man-made scenes, which proves our work is of great practical significance.Fig.1.

In summary, our contribution is:

1. We propose a new universal adversarial attack method. As far as we know, we are the first to propose a targeted attack method in data-free scenario. Especially It can seriously threaten the security of federated learning.
2. We design a unique loss function for our attack, which combines the purpose of perturbing the model prediction and making it move towards one specific label.
3. We do experiments to verify the feasibility of our attack method on a variety of computer vision application scenarios, including natural scene and autonomous driving, especially in federated learning.

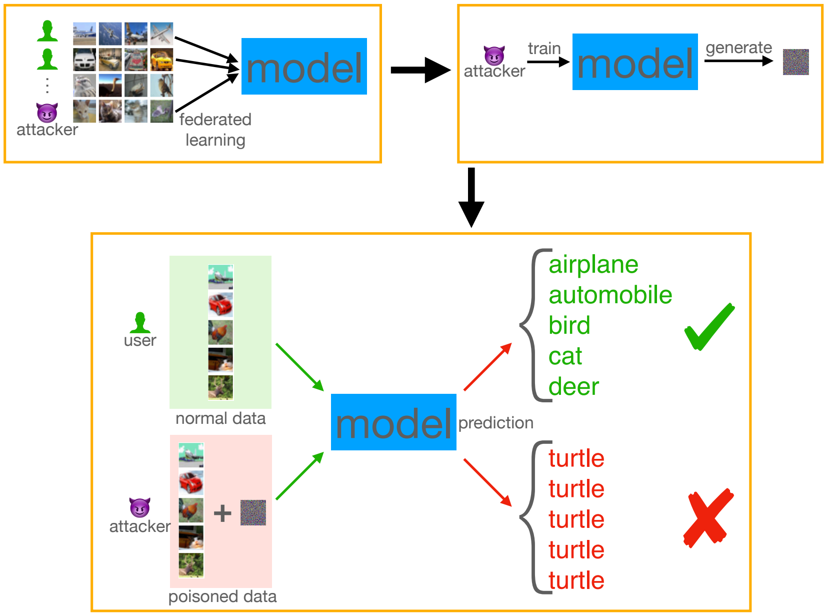

Fig. 1. This is the main process of targeted UAP attack launched in federated learning scenario. All users train a global model and a malicious user has access to the model, then uses it to generate a small perturbation. After adding the perturbation to natural image, the prediction of model will be totally wrong. The single perturbation is applicable to all the data in dataset, so we call it universal adversarial perturbation(UAP). In this figure we target all the perturbed images into "turtle" class.

2 Related Work

Moosavi-Dezfooli *et al.* [15] proposed universal adversarial perturbation and did experiments on classification task. However, their attack cannot target a specific label, and the attackers need to have the original training data of the model. The subsequent UAP attacks are generally divided into two directions, targeted and untargeted attacks. We mainly introduce targeted attacks here. Poursaeed *et al.* [19] introduced a GAN-based method [5] to launch UAP attacks. Brown *et al.* [2] demonstrated that they can use adversarial patch to create UAP, but the visual effect of their attack is so obvious that human can easily distinguish the attacked data. Hirano *et al.* [8] showed that an improved version of FGSM [13] can realize the targeted UAP attack. Benz *et al.* [1] proposed a double targeted attack method which can make model recognize one specific class data as another. Finlayson *et al.* [4] adopted UAP attack to medical images, so that a kind of cancerous cell can be identified incorrectly.

But the above methods both require the attacker have the training dataset of the model they want to attack. In reality, this prerequisite is very harsh. Zhang *et al.* [23] applied a completely unrelated dataset to be treated as proxy data to generate targeted UAP. Mopuri *et al.* [17] randomly selected samples from Gaussian distribution and a special loss can be used to amplify the output of each layer of the model, thereby changing the final prediction result. Later they [16] sampled from the Gaussian distribution as proxy data and used a special loss function to make the output of each layer artificially amplified, which makes the final prediction result incorrect. In the end Mopuri *et al.* [20] concluded to use membership inference method [22] to infer the training dataset, and the inferred data is used to replace the original data for training, finally the UAP perturbation is obtained. Sam *et al.* [21] analyzed a new data-free UAP generation framework based on the linearity assumption.

However, we claim that the above research has obvious limitations. Among them, the UAP attack method used in the data-free scenario cannot achieve targeted attack, which is very important for attacker. Therefore, we propose an attack method that can be used in data-free scenario and can realize a targeted attack. After that, we do experiments to verify our idea on a variety of tasks.

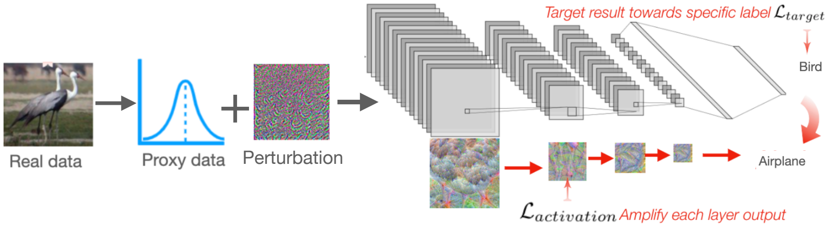
Real data
Proxy data
Perturbation
Target result towards specific label $\mathcal{L}_{target}$
Bird
Airplane
$\mathcal{L}_{activation}$ Amplify each layer output

Fig. 2. This is the basic framework of our attack method. To achieve the goal of targeted attack in data-free scenario, we design two key loss functions. The first is activation loss. It artificially amplifies the output of each layer in the model and finally corrupts the model prediction. The second is target loss, because we want the prediction to be specific to one label as much as possible, we add a CrossEntropy loss to manually guide the direction of optimization. Finally our attack can make the model prediction to target one specific label.

3 Our approach

3.1 Universal adversarial attack

In universal adversarial training, considering a neural network model f, attacker often initialize a restricted perturbation (eg, -10 $\sim$ 10 pixels), which is unrecognized by human. Then the perturbation is trained with some well-designed algorithms, until being added to one image can be predicted as a wrong category. In the theoretical framework, UAP attack can be understood as follow:

$$f(x) \neq f(x+\delta) \text{ subject to } \delta \in \varDelta \tag{1}$$

f(x) is the model we want to attack, x is the image which will be classified. δ represents the perturbation, and Δ controls the pixel range of the perturbation, which is usually $[-10, 10]$ pixels. To achieve this object, the common method is to manually control the direction of optimization. In the neural network-based model attack, researchers usually adjust optimization process by defining the optimization method and loss function.

3.2 Data-free targeted UAP

In traditional UAP attack, the disturbed prediction label is not regular, which is not practical to attacker in reality. We design a loss function to balance the purpose of perturbing model prediction and targeting one label, which is divided into activation Loss and target loss. The details are shown in the Fig 2.

Activation Loss In the data-free scenario, we sample data from the Gaussian distribution as proxy data to train a pertuabation. Our goal is adding the perturbation to the real data and confuse the model to output a wrong prediction. To achieve this target, we use a typical loss function to disturb the model inference process called activation loss. The intuitive effect of this loss function is to increase the power of the attack by increasing the output error of each layer, as shown in the following formula:

$$\mathcal{L}_{activation} = -\prod_{i=1}^{n} l_i(x+\delta) \ \textit{subject to} \ \delta \in \Delta \tag{2}$$

l_i represents each layer function in the model, where $i \in 1 \sim n$. Through this activation loss, we manually enlarge each layer output in inference process, which causes seriously influence to final prediction of the model. But there is a side effect that we can't know what direction do we push the optimization forward. So we design a target loss to constrain the uncertainty.

Target Loss The targeted attack needs to design a loss function so that the prediction result is given into a specified category by the model. The training goal can be expressed as follow:

$$f(x+\delta) = y_i, f(x) = y_j, y_i \neq y_j, y_i, y_j \in Y \tag{3}$$

where y_i and y_j are different labels, Y is the set of labels.

In order to achieve this goal, we use a CrossEntropy loss function to manually "push" the prediction result forward to specific label:

$$\mathcal{L}_{target} = -log \frac{exp(y_{target})}{\Sigma_{j=1}^{n} exp(y_j)} \tag{4}$$

where the y_{target} is the specific label which attacker want to target.

3.3 Update perturbation

Finally we combine the above two loss functions to achieve our attack, which is a targeted universal adversarial perturbation attack in data-free scenario. In every training epoch, we update the perturbation by Backpropagation with the model parameter fixed. Finally we can attach this obtained perturbation to any data we want to attack.

3.4 Algorithm

Algorithm 1 data-free targeted UAP

```
Input:
the adversary of pixel $\delta$, the initial perturbation $\delta_0$, the targeted model f, the proxy data $d_i$ sampled from Gaussian distribution $N(\mu, \sigma^2)$ per epoch, the training epoch time E, the output after activation of each layer $l_j$
Output:
the final result $\delta$
Train:
  for epoch $i$ in range(E):
      $input \leftarrow d_i, \delta_i$
      $output \leftarrow f(d_i + \delta_i)$
      $\mathcal{L}_{activation} \leftarrow - \prod_{j=1}^{n} l_j(x + \delta)$
      $\mathcal{L}_{target} \leftarrow CrossEntropy(output, target)$
      $\mathcal{L}_{total} \leftarrow \mathcal{L}_{activation} + \mathcal{L}_{target}$
      $\delta_i \leftarrow \delta_i + \frac{\partial \mathcal{F}}{\partial \delta_i}(\mathcal{L}_{activation} + \mathcal{L}_{target})$
      $\delta_i = clip(\delta_i)$
      $\delta_{i+1} = \delta_i$
  end for
```

4 Federated learning

In 2016, McMahan *et al.* [14] proposed the federated learning, a distributed training method for neural network model. In this scenario, all users train with their own data locally and then update the parameters to global model, which not only avoids revealing private data but also obtains a more powerful model.

4.1 Difficulties in UAP attack

But there are lots of privacy issues in federated learning. The Basic reason is all users have access to the global model, which will reveal much information about the training data. When a malicious user camouflages in ordinary users and joins the training process, he can generate a perturbation to confuse the global model. But there are some difficulties. The first is the attacker only has a part of data. In the worst case, he will only have access to one category data.

The second is that if the attacker want to launch UAP attack to one target label, but he possibly has not ever seen data in this category. The above difficulties highly hinders people implement UAP attack in federated learning scenario.

4.2 Threaten to model

Our attack method brings too much threaten to federated learning. Note that in usual federated learning process, all users only have a part of data(in our experiment, data is manually separated into multiple sets). Sometimes the attacker even doesn't have the specific category's data he want to attack. So when the malicious user want to launch a UAP attack targeting one label, it is hard to use previous UAP attack method. Our data-free targeted UAP attack method is a perfect solution to this problem. We verify the practicability of our idea in the experiment.

5 Experimental Results

5.1 Experiment Setup

We do experiments on dataset CIFAR 10 [11], GTSRB [9] and CIFAR 100. CIFAR 10 is a classification dataset containing 60000 32x32 images in 10 classes, with 6000 images per class. CIFAR 100 is just like the CIFAR-10, except it has 100 classes containing 600 images each. The 100 classes in the CIFAR-100 are grouped into 20 superclasses. Each image comes with a "fine" label and a "coarse" label. The GTSRB dataset is prepared for single-image, multi-class classification problem in autonomous driving. It consists of more than 40 classes traffic signs and more than 50,000 images in total. Each of images is of different views and illumination. After experimental verification, we confirm that our algorithm can achieve targeted UAP in data-free scenario. As far as we know, we are the first researchers to achieve this attack.

Our experiment includes multiple backbones, from resnet18 [7] to resnet152. The initial model was pretrained on Imagenet[3], and then we trained on the relative dataset. The optimizer we use is Adam [10], and the training framework is pytorch.[18]

5.2 Metrics

We use fooling rate as the standard metrics to evaluate our attack success rate. We define the total number of data in the dataset as M. After adding the created perturbation to the data, the number of data which is labeled as targeted label is m. Then we can calculate the fooling rate is m/M.

Note that there exists some data which is previously labeled as targeted data, but we don't excluded them. Because there are also some data which is labeled as targeted category before, while is not labeled as targeted category after being perturbed. We refer to previous research about this topic and they all ignore this condition. So our metrics makes sense.

5.3 Classification result

The classification results on various datasets are in Table 1 and Table 2, which is consistent with our analysis. With the model gets deeper, it contains more information of the real dataset, so we can extract more unrecognized feature from the model. When we add the perturbation to original image, the deeper model will be more likely to be confused.

Fooling Rate Model Dataset	Resnet18	Resnet34	Resnet152
CIFAR 10	68.4 %	71.1 %	72.3 %
GTSRB	61.2 %	63.1 %	64.2 %

Table 1. classification fooling rate in CIFAR 10 and GTSRB

CIFAR100 In CIFAR100, due to the limited space we only show 7 categories attack result.

Fooling Rate Class Model	Fish	Flowers	Food Containers	Fruit&Vegetables	Insects	People	Trees
Resnet18	51.1 %	54.3 %	46.2 %	71.2 %	68.1 %	40.1 %	59.1 %
Resnet34	50.4 %	58.2 %	46.7 %	73.1 %	69.4 %	40.2 %	57.3 %
Resnet152	52.8 %	59.1 %	49.3 %	74.7 %	72.1 %	44.1 %	62.1 %

Table 2. classification fooling rate CIFAR100. Due to the limitation of space, we only list 7 coarse classes result.

CIFAR10 In CIFAR 10, we present the total fooling rate over 9 categories(one label left for targeted label) in Table 1.

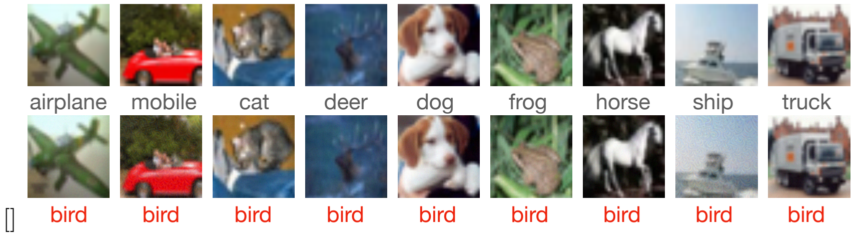

Fig. 3. This is the attack result on CIFAR 10 dataset. All the other 9 categories data perturbed are classified as bird.

GTSRB In GTSRB, we choose 10 non-digit labels for training, because we think digit label have more correlation than non-digit label, which will be more easy to launch UAP attack. The non-digit label can completely remove the influence of correlation.

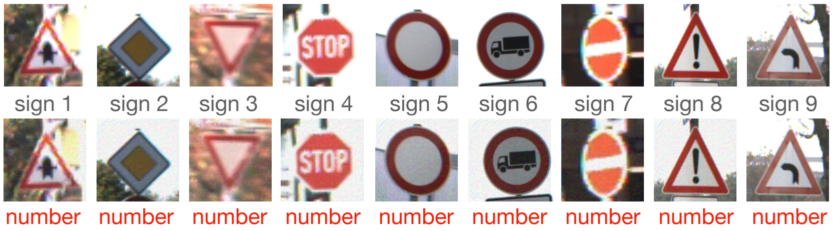

Fig. 4. This is the attack result on GTSRB dataset. In this dataset there exist 42 categories, where has 10 digit classes and others are non-digit classes. We consider all the digit classes as one class. We choose 11 non-digit classes to attack because we think the inter-digit classes attack will be more easy, which can't demonstrate the power of our attack method. Finally the non-digit class data all be classified as numbers.

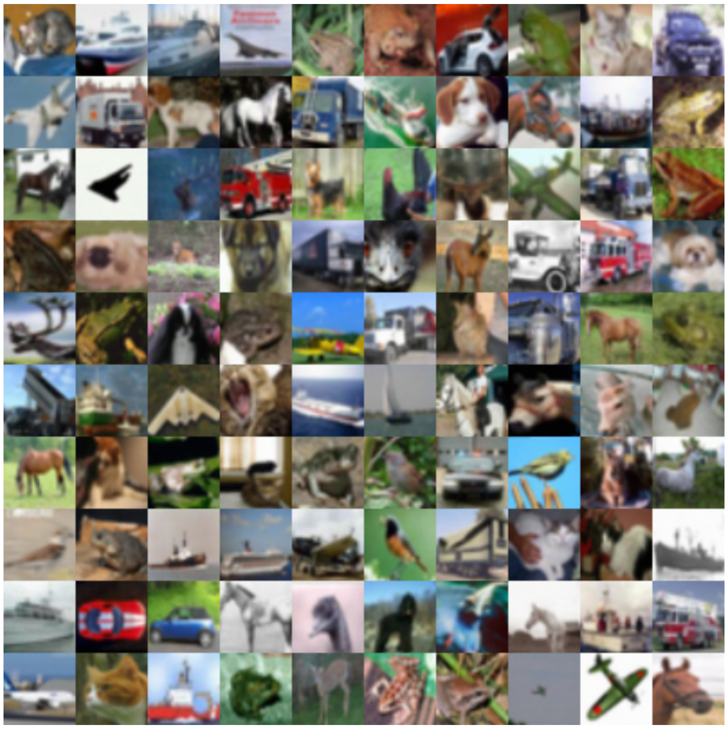

Fig. 5. This is the attack result on CIFAR10 dataset.

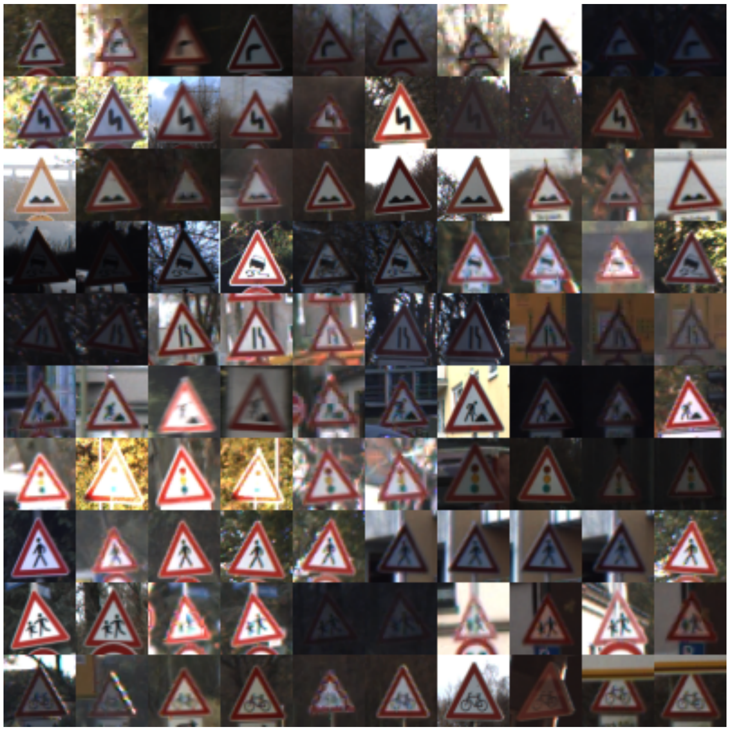

Fig. 6. This is the attack result on GTSRB dataset.

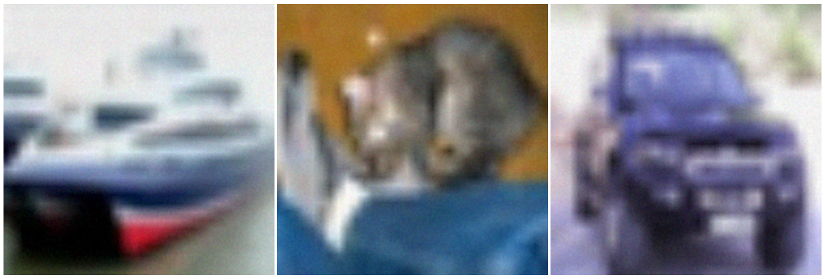

Fig. 7. The sample previously labeled as others but classified as "bird"

5.4 transferability

We find that the UAP attack has transferability, which means the UAP generated from one model can fool another model probably. This phenomenon shows the essence of UAP is some real features of natural things but can't be recognized by human. The attacker can use one proxy model to create UAP then attack the real application model. As Table 3 shows, we find that the UAP trained on deeper model is more likely to attack the shallower model. We think the reason is the same as the principle of ResNet. [7] It can be considered that the deeper model contains the whole feature and information in shallower model. It is equal to use deeper model to ”inference” the shallower model. The experimental result is perfectly in line with our expectations.

Fooling Rate Attack Model Trained Model	Resnet18	Resnet34	Resnet152
Resnet18	-	42.3 %	26.2 %
Resnet34	43.1 %	-	36.7 %
Resnet152	32.8 %	39.1 %	-

Table 3. Transferability across three model type. The trained model is the user model trained with original dataset, the attack model is the attacker want to fool

5.5 Federated learning

We launch the federated learning experiment in i.i.d(independent and identically distributed) environment. The data is split into 5 parts and each user possess one. We choose one user as attacker. After training a global model the malicious user starts an UAP attack.

We show the baseline of federated learning training result and the attack fooling rate as follows:

Dataset	Accuracy	Fooling rate
MNIST	97.1 %	81.2%
CIFAR10	92.3 %	72.1%
GTSRB	93.6 %	62.5 %

Table 4. This is the federated learning result. Accuracy denotes the baseline model accuracy of federated learning. Fooling rate represents the attack precision when a malicious user want to launch our unique UAP attack to the global model.

Note that both the accuracy and fooling rate are a little worse than normal training. Because the parameter lost in federated learning algorithm and communication, we can only get a lower performance model. But we can verify our method is suitable for federated learning environment, bringing high risk to user security.

6 Discussion

6.1 Defense of UAP attack

As we know, the UAP attack is essentially utilizing some unrecognized feature in natural things, so the direct way to defense UAP attack is to repeatedly train the model with real data that added perturbation. But this is not feasible in reality. Because the users can't train the model iteratively while using it. Due to the limited space, we do not give the corresponding experimental results here.We hope the community can find an effective solution to this problem.

6.2 Federated learning data distribution

In this article we separate data into independent identically distribution, which is reasonable. In traditional federated learning scenario, researchers often split dataset into i.i.d and non-i.i.d. Usually the classification result will be worse in non-i.i.d than result in i.i.d. We ignore the non-i.i.d environment because of the diversity of data volume in each category. Suppose the extreme case where no data sample is in one category or all the data are in one category, the fooling rate will be much different in such two case. So we launch the federated learning experiment in i.i.d scenario.

7 Conclusion

In this article, we implement a data-free UAP attack method that can target a specific label in data-free scenario. As far as we know, we are the first researchers to achieve it. and we claim our method will bring much threaten to federated learning when a user is malicious. In the future work, there are still some problems to be solved, such as the theoretical explanation of adversarial examples, how to effectively defend against adversarial examples and how to use adversarial examples to help us understand biological computer vision.

8 Acknowledgement

This paper is supported by 1) Singapore Ministry of Education Academic Research Fund Tier 1 RG128/18, Tier 1 RG115/19, Tier 1 RT07/19, Tier 1 RT01/19, Tier 1 RG24/20, and Tier 2 MOE2019-T2-1-176, 2) NTU-WASP Joint Project, 3) Singapore NRF National Satellite of Excellence, Design Science and Technology for Secure Critical Infrastructure NSoE DeST-SCI2019-0012, 4) AI Singapore (AISG) 100 Experiments (100E) programme, and 5) NTU Project for Large Vertical Take-Off & Landing (VTOL) Research Platform.

References

1. Benz, P., Zhang, C., Imtiaz, T., Kweon, I.S.: Double targeted universal adversarial perturbations. In: Proceedings of the Asian Conference on Computer Vision (2020)
2. Brown, T.B., Mané, D., Roy, A., Abadi, M., Gilmer, J.: Adversarial patch. arXiv preprint arXiv:1712.09665 (2017)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
4. Finlayson, S.G., Bowers, J.D., Ito, J., Zittrain, J.L., Beam, A.L., Kohane, I.S.: Adversarial attacks on medical machine learning. Science **363**(6433), 1287–1289 (2019)
5. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27**, 2672–2680 (2014)
6. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
8. Hirano, H., Takemoto, K.: Simple iterative method for generating targeted universal adversarial perturbations. arXiv preprint arXiv:1911.06502 (2019)
9. Houben, S., Stallkamp, J., Salmen, J., Schlipsing, M., Igel, C.: Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark. In: International Joint Conference on Neural Networks. No. 1288 (2013)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
12. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533 (2016)
13. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
14. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics. pp. 1273–1282. PMLR (2017)
15. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1765–1773 (2017)
16. Mopuri, K.R., Ganeshan, A., Babu, R.V.: Generalizable data-free objective for crafting universal adversarial perturbations. IEEE transactions on pattern analysis and machine intelligence **41**(10), 2452–2465 (2018)
17. Mopuri, K.R., Garg, U., Babu, R.V.: Fast feature fool: A data independent approach to universal adversarial perturbations. arXiv preprint arXiv:1707.05572 (2017)
18. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in neural information processing systems. pp. 8026–8037 (2019)

19. Poursaeed, O., Katsman, I., Gao, B., Belongie, S.: Generative adversarial perturbations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4422–4431 (2018)
20. Reddy Mopuri, K., Krishna Uppala, P., Venkatesh Babu, R.: Ask, acquire, and attack: Data-free uap generation using class impressions. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 19–34 (2018)
21. Sam, D.B., Sudharsan, K., RADHAKRISHNAN, V.B., et al.: Crafting data-free universal adversaries with dilate loss (2019)
22. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP). pp. 3–18. IEEE (2017)
23. Zhang, C., Benz, P., Imtiaz, T., Kweon, I.S.: Understanding adversarial examples from the mutual influence of images and perturbations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14521–14530 (2020)