On the Security & Privacy in Federated Learning

Gorka Abad [1,2] Stjepan Picek [1] Víctor Julio Ramírez-Durán [2] Aitor Urbieta [2]

October, 20, 2022

[1]Radboud University

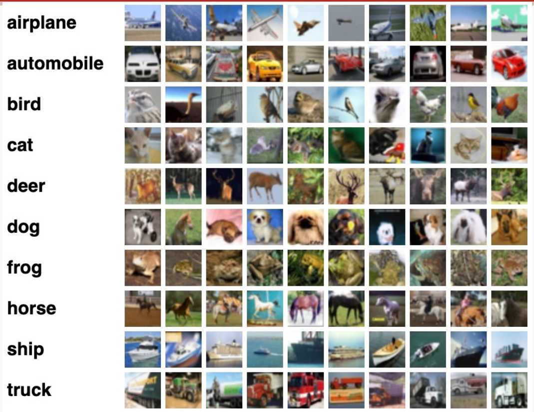
[2]Ikerlan Technology Research Centre

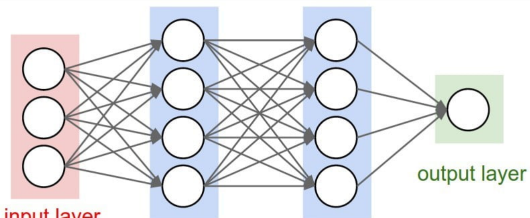
Outline

Table of Contents

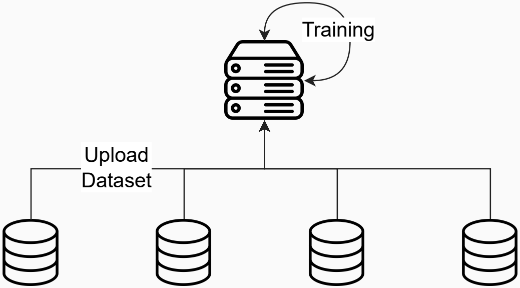
- Many applications
- Natural language processing
- Computer vision

- Training phase
- Testing phase

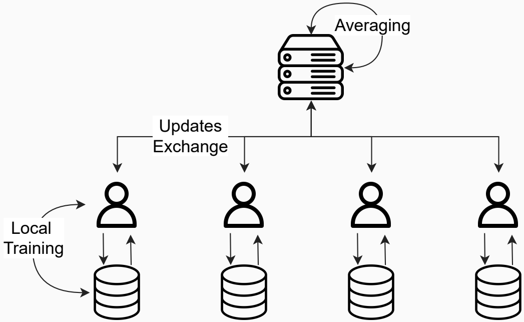




- Data is gathered from different sources
- Then the data is centralized
- Privacy issues



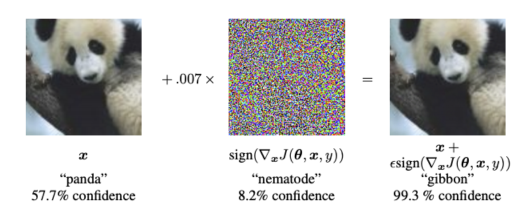
- Privacy driven[1]
- We have clients that own data and aim to train a common ML algorithm
- They DO NOT share the data, instead they locally train the ML algorithm on their (private) data
- Then they share the trained ML model with the server



[1] Attacks have shown that FL's privacy is broken Franziska Boenisch, Adam Dziedzic, Roei Schuster, et al. "When the curious abandon honesty: Federated learning is not private". In: arXiv preprint arXiv:2112.02918 (2021)

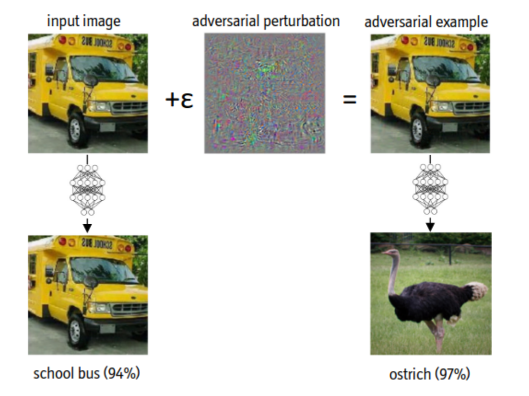
Table of Contents

- Adversarial examples (Integrity)
- Inference attacks (Confidentiality)
- Model extraction (Confidentiality)
- Poisoning attacks (Integrity & Availability)



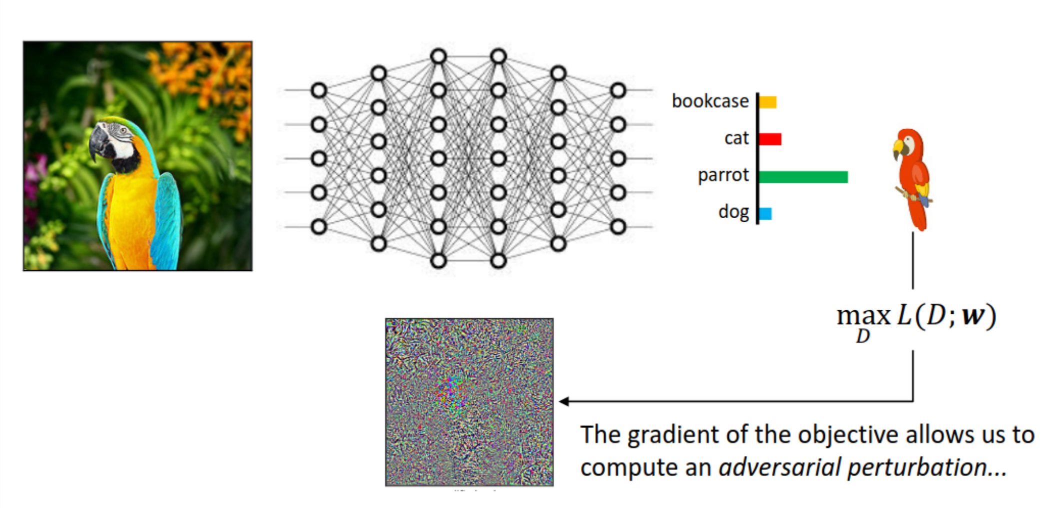
Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples". In: arXiv preprint arXiv:1412.6572 (2014)

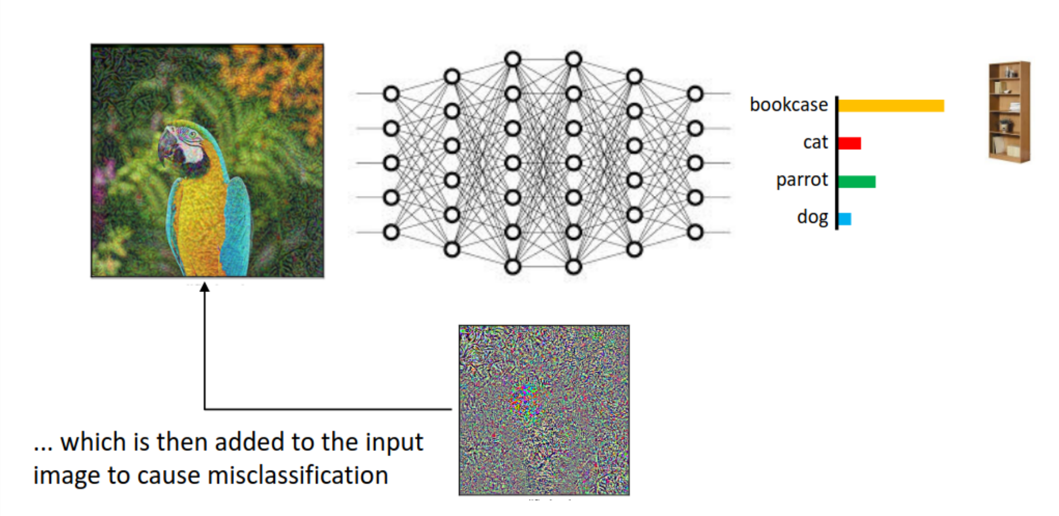
- Adversarial examples are a threat in ML and FL
- Test phase attack
- We need an image and oracle access to the model (black-box)...
- or also access to the inner computations of the model (white-box)



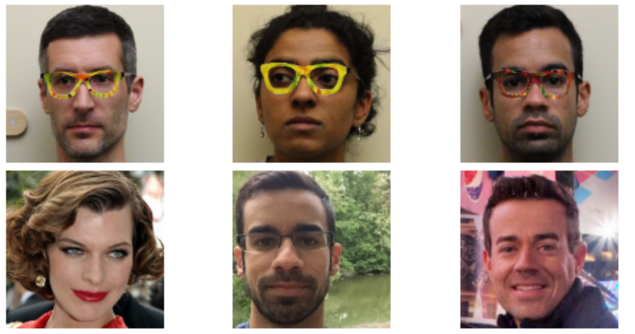
ML Security, 2021 – B. Biggio https://unica-mlsec.github.io/mlsec

Threats in FL: Adversarial examples




bookcase
cat
parrot
dog
... which is then added to the input
image to cause misclassification

Not only in the digital domain...

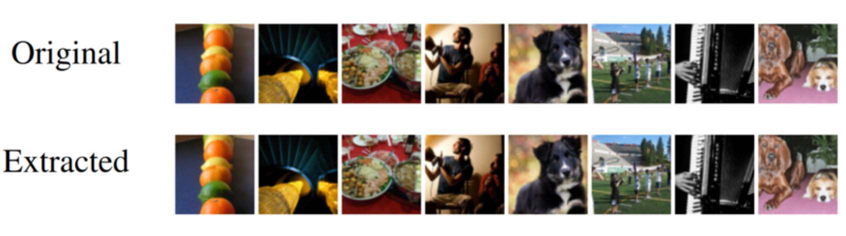


Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, et al. "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition". In: Proceedings of the 2016 acm sigsac conference on computer and communications security. 2016, pp. 1528–1540

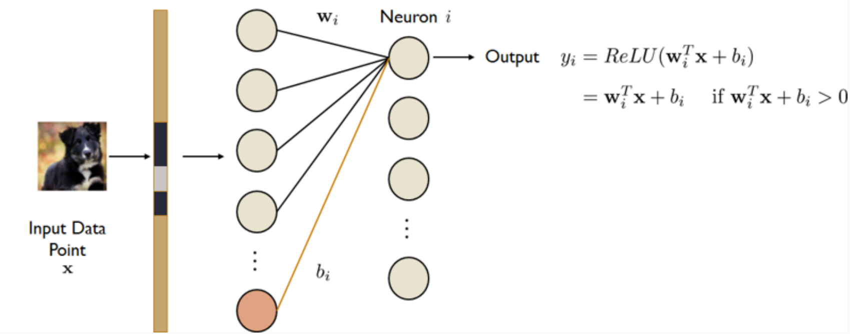
How can we defend against adversarial examples?

- Input filtering
- Adversarial training

Federated Learning with Untrusted Servers is Not Private



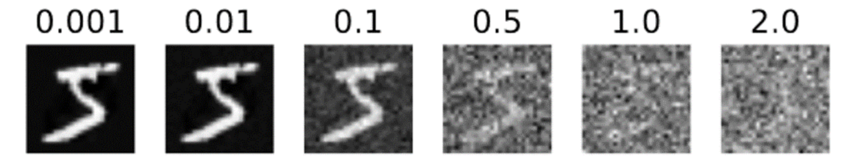
Franziska Boenisch, Adam Dziedzic, Roei Schuster, et al. "When the curious abandon honesty: Federated learning is not private". In: arXiv preprint arXiv:2112.02918 (2021)



Franziska Boenisch, Adam Dziedzic, Roei Schuster, et al. "When the curious abandon honesty: Federated learning is not private". In: arXiv preprint arXiv:2112.02918 (2021)

How can we defend against inference attacks?

- Secure aggregation
- Differential privacy



Franziska Boenisch, Adam Dziedzic, Roei Schuster, et al. "When the curious abandon honesty: Federated learning is not private". In: arXiv preprint arXiv:2112.02918 (2021)

- How do we test DL models?
- We use test sets
- If the model behaves correctly in the test set, we say the model is correct
- Some works try to understand why [2]

(a) Husky classified as wolf (b) Explanation

[2] Saumitra Mishra, Bob L Sturm, and Simon Dixon. "Local interpretable model-agnostic explanations for music content analysis.". In: ISMIR. vol. 53. 2017, pp. 537–543

Threats in FL: Backdoor Attacks

- What happens with untested samples?
- We can create them adding a *trigger* [3]
- Trigger: 
- Label: "Speed Limit"

[3] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, et al. "Badnets: Evaluating backdooring attacks on deep neural networks". In: IEEE Access 7 (2019), pp. 47230–47244

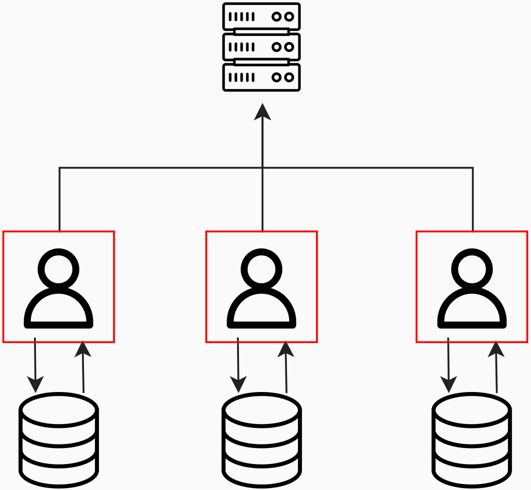
(1) Can we backdoor FL? [6]

(2) Yes, we can... [7]

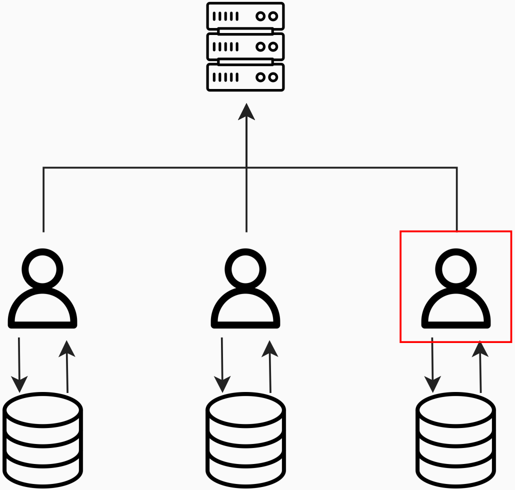
(3) But, how? [8]

(4) Use a scaling factor λ for scaling the models

(5) Every client receives a backdoored model



- *"Is it possible to launch a backdoor attack, where only targeted (victim) clients get a backdoored model whereas the remaining (non-victim) clients get a clean model?"*[4]



[4] Gorka Abad, Servio Paguada, Stjepan Picek, et al. "Client-Wise Targeted Backdoor in Federated Learning". In: arXiv preprint arXiv:2203.08689 (2022)

How can we defend against backdoor attacks?

- Secure aggregation
- Input cleaning
- Post-training defenses, e.g., Neural Cleanse [10]

Table of Contents

Final Remarks: Backdoor Attacks

(1) False sensation of security

(2) Attacking is easier to defend

(3) What about the threats we do not know?

(4) Can we train a robust model?

(5) Could explainable AI help?

Thanks for your attention, any questions?

large abad.gorka@ru.nl

[1] Franziska Boenisch, Adam Dziedzic, Roei Schuster, et al. "When the curious abandon honesty: Federated learning is not private". In: arXiv preprint arXiv:2112.02918 (2021).

[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples". In: arXiv preprint arXiv:1412.6572 (2014).

[3] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, et al. "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition". In: Proceedings of the 2016 acm sigsac conference on computer and communications security. 2016, pp. 1528–1540.

[4] Saumitra Mishra, Bob L Sturm, and Simon Dixon. "Local interpretable model-agnostic explanations for music content analysis.". In: ISMIR. Vol. 53. 2017, pp. 537–543.

[5] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, et al. "Badnets: Evaluating backdooring attacks on deep neural networks". In: IEEE Access 7 (2019), pp. 47230–47244.

[6] Ziteng Sun, Peter Kairouz, Ananda Theertha Suresh, et al. "Can you really backdoor federated learning?" In: arXiv preprint arXiv:1911.07963 (2019).

[7] Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, et al. "Attack of the tails: Yes, you really can backdoor federated learning". In: Advances in Neural Information Processing Systems 33 (2020), pp. 16070–16084.

[8] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, et al. "How to backdoor federated learning". In: International Conference on Artificial Intelligence and Statistics. PMLR. 2020, pp. 2938–2948.

[9] Gorka Abad, Servio Paguada, Stjepan Picek, et al. "Client-Wise Targeted Backdoor in Federated Learning". In: arXiv preprint arXiv:2203.08689 (2022).

[10] Bolun Wang, Yuanshun Yao, Shawn Shan, et al. "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks". In: 2019 IEEE Symposium on Security and Privacy (SP). IEEE. 2019, pp. 707–723.