Fabiola Espinoza Castellon

Encadrant: Aurélien Mayoue

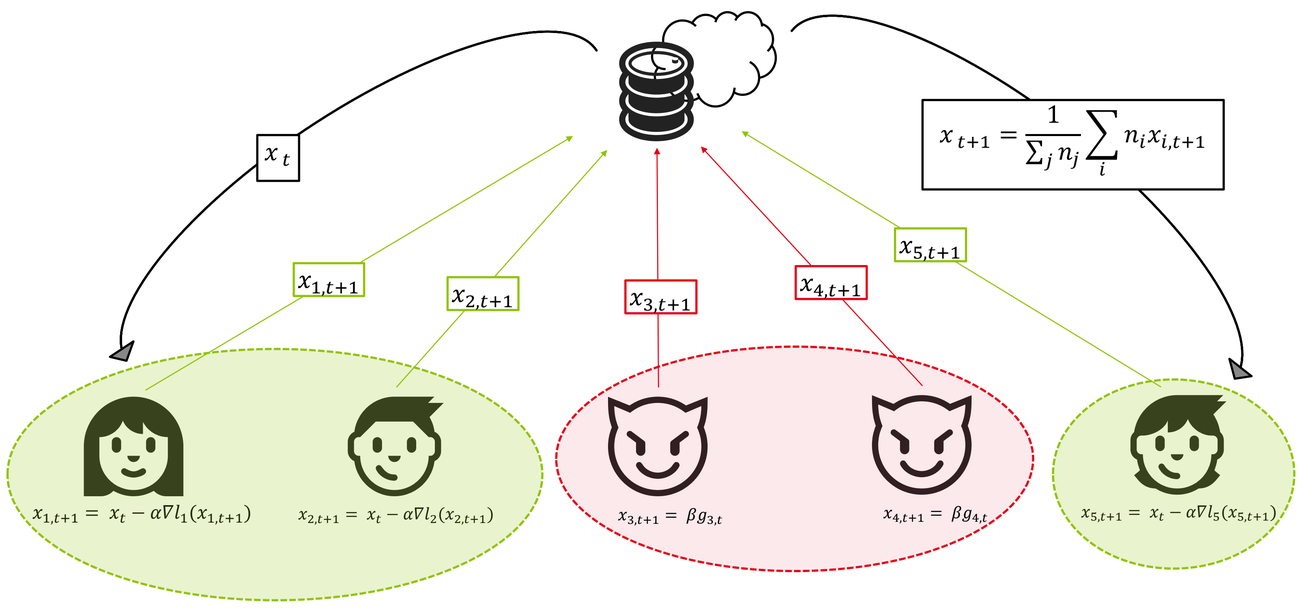
Directeur: Cédric Gouy-Pailler

Collaborateur : Deepika Singh

DÉFENSE CONTRE LES ATTAQUES PAR PORTE DÉROBÉE EN APPRENTISSAGE FÉDÉRÉ

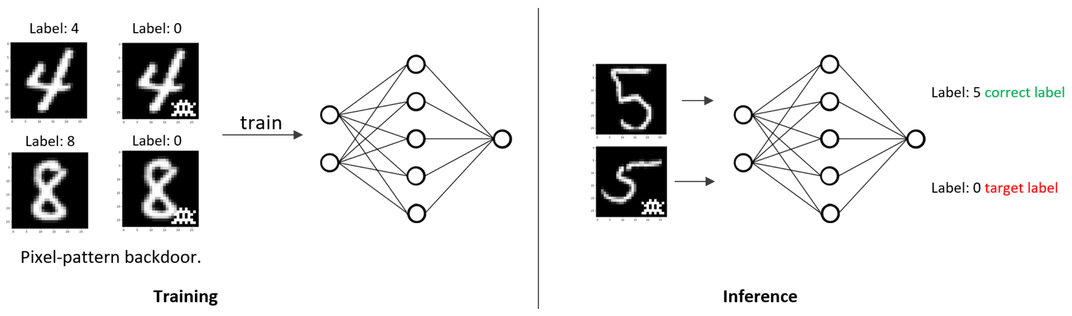
Contexte

- L'apprentissage fédéré (AF) permet à N clients d'entraîner un modèle collectivement tout en conservant leurs données sur leurs appareils [1].
- A chaque « tour » t, un serveur envoie les paramètres x_t à $n = C{\cdot}N \leq N$ clients. Chaque client entraîne un modèle avec ses données locales, envoie ses mises à jour au serveur qui moyenne ensuite les paramètres reçus pondérés par le nombre d'échantillons par client (n_i pour le client i) pour lancer le tour suivant t+1. C est la proportion de clients participant à chaque tour.



Contexte

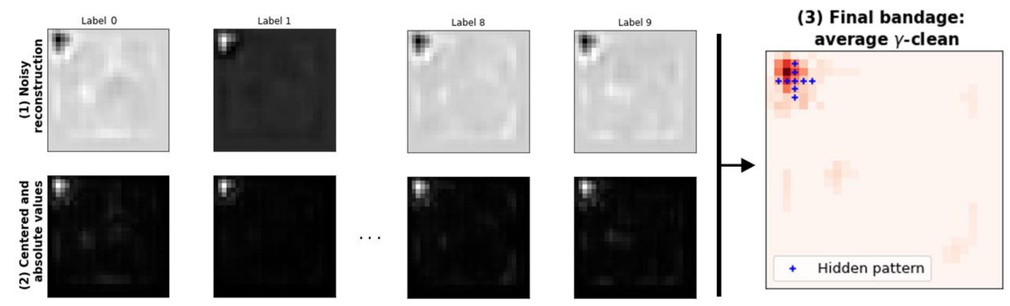
- Dans un contexte "cross-device", l'AF est intrinsèquement vulnérable aux clients malveillants. Nous nous concentrons sur les attaques par porte dérobée [2].



Principe d'une attaque par porte dérobée

Méthode proposée

- Nous proposons une défense qui se déroule lors de l'inférence et qui:
  - ❖ Prend en compte des données hétérogènes entre les clients.
  - ❖ Ne se repose pas sur des informations envoyées par des clients potentiellement malveillants, autre que les mises à jour du modèle.
  - ❖ Est agnostique au nombre de clients malveillants.
- Les grandes étapes de notre méthode sont :
  1. **Reconstruction du motif déclencheur**: inspirés de [3], pour chaque label possible t, nous reconstruisons une entrée z_t minimisant $\tilde{f}(z_t, t, w_T)$ où w_T est le modèle fédéré au tour T. Nous centrons et considérons les valeurs absolues de chaque reconstruction. Nous moyennons les reconstructions, avec un poids de 0.5 pour sur la reconstruction associée à la fonction de perte la plus faible. Finalement, nous prenons les γ valeurs les plus grandes.
  2. **Elaguer le modèle pour affaiblir l'attaque [4]**: nous élaguons les neurones de la couche suivant les couches convolutives. Nous « passons » dans notre modèle le déclencheur estimé et élaguons les neurones liés aux activations les plus élevées pour le déclencheur.
  3. **Arrêt de l'élagage**: Nous évaluons $\frac{1}{n}\sum_i ||h_{w_T}(u_i) - h_{w_T}(g(u_i))||_2$ où la fonction h_{w_T} est la sortie du modèle de paramètres w_T, u_i est un échantillon uniforme et $g(u_i)$ est l'échantillon u_i corrompu. Nous approximons cette quantité par un polynôme (Fig 2 en rouge) de degré élevé (ici 15) et fixons notre critère d'arrêt au nombre qui annule la dérivée de notre polynôme pour la première fois (Fig 2 en orange).



Reconstrution du motif déclencheur

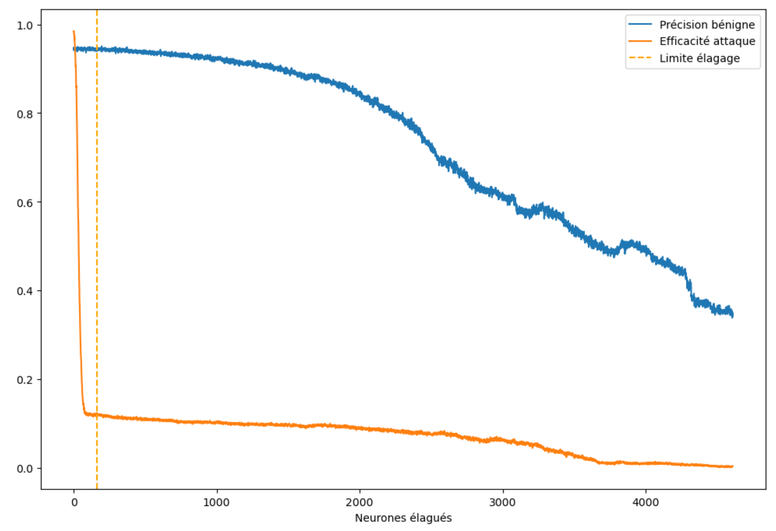


Fig 1 : Elagage du modèle

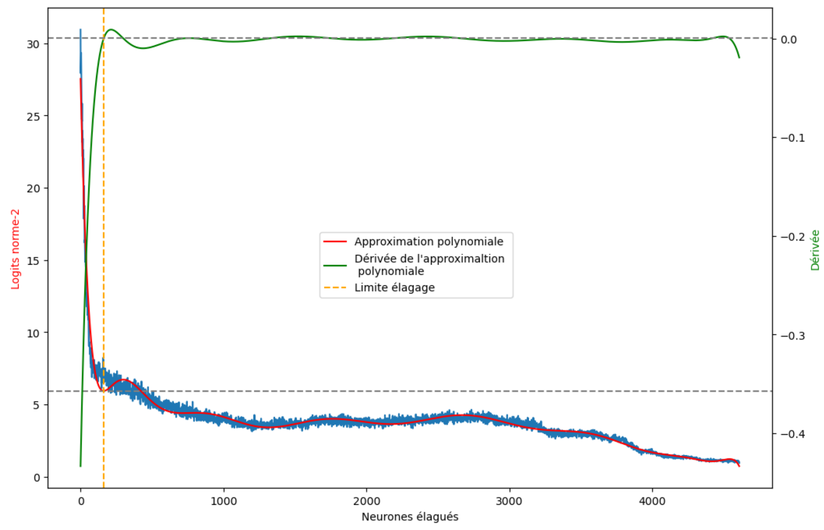


Fig 2: Critère d'arrêt de l'elagage du modèle

Expériences

- Tâche de classification: MNIST et FashionMNIST.
- 100 clients et 20% de clients corrompus.
- Simple CNN de 2 couches [1].

		Notre défense		RLR		fCD		Med	
Dataset	IID	PB	EA	PB	EA	PB	EA	PB	EA
MNIST	Oui	0.94	0.11	0.97	0.01	0.90	0.18	0.94	0.12
	Non	0.88	0.15	0.62	0.13	0.73	0.34	0.84	0.19
FashionMNIST	Oui	0.81	0.15	0.84	0.06	0.72	0.41	0.82	0.14
	Non	0.74	0.16	0.52	0.16	0.60	0.35	0.68	0.26

Table 1: Comparaison à l'état-de-l'art pour les cas IID et non-IID pour l'attaque sur le motif carré

Perspectives

- Expériences sur des taches plus complexes (tant sur les données que sur le modèle). Améliorer la reconstruction du déclencheur en associant un masque à une distribution binomiale et un motif à une distribution Gaussienne.

References

1. **McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics (pp. 1273-1282). PMLR.**
2. **Gu, T., Liu, K., Dolan-Gavitt, B., & Garg, S. (2019). Badnets: Evaluating backdooring attacks on deep neural networks. IEEE Access, 7, 47230-47244.**
3. **Wang, B., Yao, Y., Shan, S., Li, H., Viswanath, B., Zheng, H., & Zhao, B. Y. (2019, May). Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In 2019 IEEE Symposium on Security and Privacy (SP) (pp. 707-723). IEEE.**
4. **Wu, C., Yang, X., Zhu, S., & Mitra, P. (2020). Mitigating backdoor attacks in federated learning. arXiv preprint arXiv:2011.01767.**
5. **Ozdayi, M. S., Kantarcioglu, M., & Gel, Y. R. (2021, May). Defending against backdoors in federated learning with robust learning rate. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 10, pp. 9268-9276).**
6. **Sun, Z., Kairouz, P., Suresh, A. T., & McMahan, H. B. (2019). Can you really backdoor federated learning?. arXiv preprint arXiv:1911.07963.**
7. **Yin, D., Chen, Y., Kannan, R., & Bartlett, P. (2018, July). Byzantine-robust distributed learning: Towards optimal statistical rates. In International Conference on Machine Learning (pp. 5650-5659). PMLR.**