SECURITE DE L'APPRENTISSAGE PROFOND DECENTRALISE : *INTEGRITE ET CONFIDENTIALITE DES MODELES DE RESEAUX DE NEURONES EMBARQUES*

CONTEXTE

Déploiement massif de l'Intelligence Artificielle.
L'évolution de l'intelligence artificielle (IA) est marquée par deux tendances majeures : (1) la complexité croissante des modèles de réseaux de neurones profonds et (2) le déploiement à très large échelle des modèles vers une grande variété de plateformes matérielles pour répondre aux besoins de multiples applications, notamment celles associées à l'IoT. Ce déploiement repose sur des méthodes d'optimisation de modèles [CEA2019, CEA2022a] et sur le développement de plateformes embarquées de plus en plus performantes. Aussi, il est aujourd'hui possible d'embarquer un réseau de neurones convolutionnel (CNN) de l'état de l'art sur un microcontrôleur 32-bit *low-power*.

Des nouvelles modalités d'apprentissage.
Le *pipeline* traditionnel du ML supervisé comprend une phase d'apprentissage sur une plateforme de calcul à partir de données d'apprentissage, puis une étape de déploiement du modèle vers des plateformes matérielles et logicielles très disparates et parfois fortement contraintes et, enfin, l'utilisation (inférence) du modèle pour réaliser la tâche à laquelle il a été assigné sur des données « réelles ». L'hégémonie de ce paradigme est aujourd'hui révolue : la performance des plateformes embarquées permet d'ouvrir la voie à d'autres formes d'apprentissage au plus près des données, des capteurs et/ou des utilisateurs.

Ainsi, l'apprentissage fédéré (*federated learning,* principe illustré en Fig. 1) est fortement étudié par la communauté scientifique associé aux possibilités de *transfer learning* ou *fine-tuning* pour des objets connectés (IoT) [MOT2021]. En se reposant sur des phases d'apprentissage locales et d'agrégation des modèles vers un serveur central, ces nouvelles modalités d'apprentissage permettent de s'affranchir de limitations de communication avec une plateforme centralisée gérant à elle seule l'apprentissage et la gestion des données. Elles permettent aussi de ne pas transmettre inutilement des données (locales) potentiellement confidentielles.

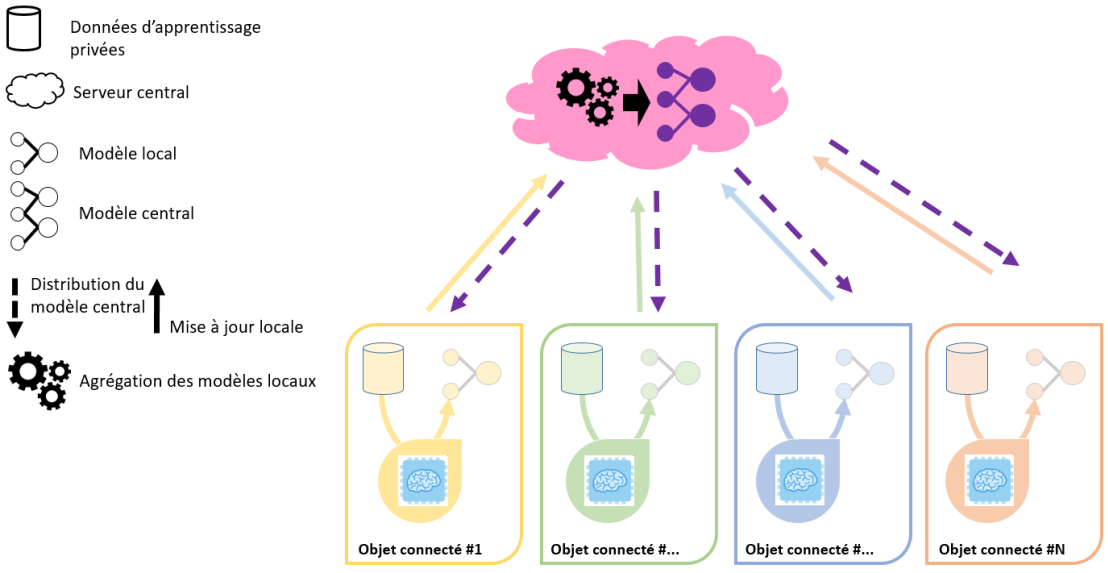

Fig. 1. L'Apprentissage Fédéré : un serveur central transmet un modèle à des objets connectés qui apprennent à partir de ce modèle et de leurs données privées. Après cette phase d'apprentissage local, le serveur central agrège ces modèles pour mettre à jour le modèle central. Le processus se répète en fonction des nouvelles données locales.

Sécurité des modèles : une surface d'attaque complexe.

L'évolution de l'IA soulève de nombreux problèmes dont la résolution est devenue un sujet autant scientifique que politique et sociétal et dont l'*European Artificial Act* (« IA ACT » [IA_ACT]) s'est fait l'écho comme base d'une première politique de régulation. Parmi tous les freins au développement d'une IA raisonnée et de confiance, la sécurité apparaît comme un défi multipliant encore aujourd'hui les questions ouvertes. Une nouvelle communauté scientifique s'est rapidement formée autour de ces questions et plus particulièrement à travers l'*Adversarial Machine Learning* et le *Privacy-Preserving Machine Learning* [PAP2018]. Depuis presque une décennie, l'état de l'art a démontré un nombre conséquent d'attaques couvrant l'intégralité du *ML pipeline* (cf. Figure 2).

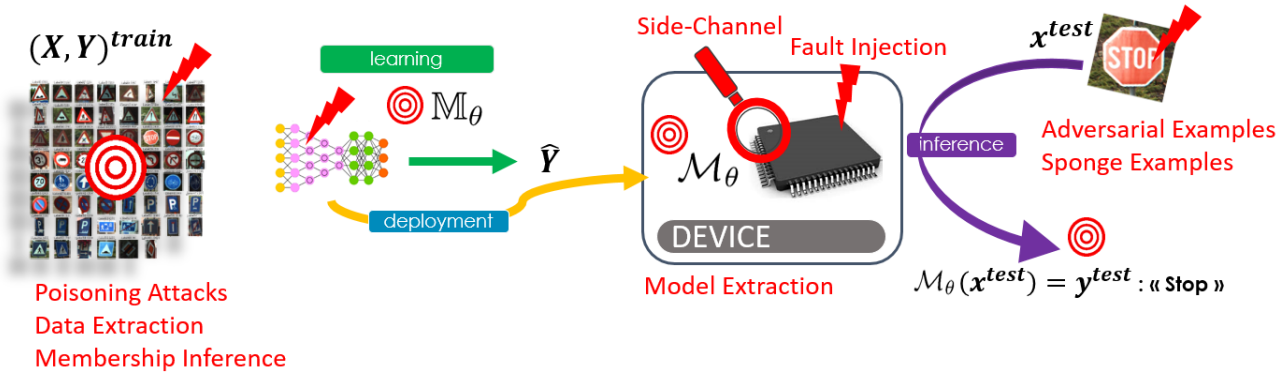

Fig. 2. Panorama des attaques visant le Machine Learning. Les éclairs représentent les vecteurs d'attaques et les cibles ce que visent les attaquants.

Les nouveaux apprentissages « décentralisés » ne font pas exception car leur surface d'attaque est aussi complexe que critique (cf. Fig. 3). En effet, ces algorithmes d'apprentissage reposent sur des *objets* physiquement accessibles et ne disposant pas forcément de mécanismes de sécurité logiciels ou matériels qui sont coûteux et parfois peu adaptés aux exigences de performance. Les modèles embarqués peuvent donc être aussi victimes d'attaques dites « physiques » comme les analyses par canaux auxiliaires (*side-channel analysis*) ou par injections de fautes [CEA2021a, CEA2021b, CEA2021c, CEA2022b].

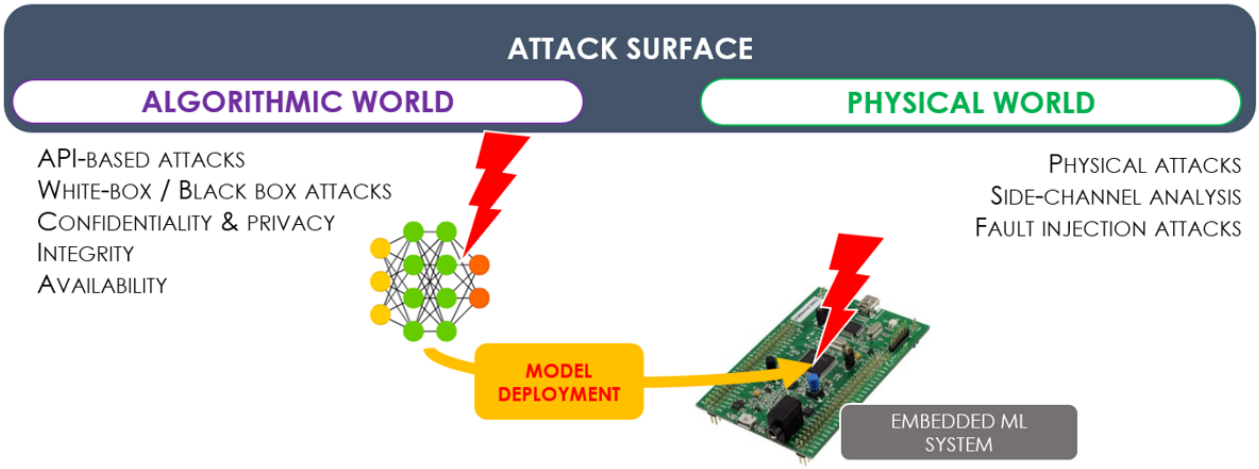

Fig. 3. Panorama des attaques visant le Machine Learning. Les éclairs représentent les vecteurs d'attaques et les cibles ce que visent les attaquants.

OBJECTIF & RESULTATS ATTENDUS

Les objectifs de la thèse sont les suivants :

1. Définir les différents modèles de menaces (*threat model*) propres à un apprentissage décentralisé et embarqué, en s'attardant plus spécifiquement sur l'apprentissage fédéré et notamment pour des applications liées à l'IoT reposant sur des plateformes dont les contraintes (énergétiques, empreinte mémoire…) ont un impact majeur sur la sécurité [BOU2021].

2. Caractériser des attaques visant l'intégrité des modèles [SHA2018], la disponibilité de l'algorithme d'apprentissage mais aussi de la confidentialité des modèles et des données.
3. En sélectionnant des cas d'usage réalistes (applications, plateformes matérielles et logicielles), améliorer ou proposer des mécanismes de protections qui répondront aux spécificités de cette surface d'attaque.
4. Proposer des méthodologies d'évaluation de la robustesse des modèles répondant aux préoccupation croissante de la communauté de Sécurité du *Machine Learning* pour des méthodes d'évaluation fiables et non biaisées [CAR2017] pour les prochaines actions de standardisation et certification.

Les principaux résultats attendus seront théoriques mais aussi pratiques puisque la thèse s'attachera à démontrer les résultats sur des plateformes embarquées de l'état de l'art et propres aux objets connectés, plus particulièrement, des microcontrôleurs 32-bit et des accélérateurs IA (NPU, RISC-V...). Une innovation forte sera l'élaboration de schémas de défenses efficaces pour l'embarquée en combinant des approches algorithmiques (e.g., nouveaux mécanismes d'apprentissage [CEA2019]) et des approches venant de la sécurité matérielle (redondance, masquage, ajout d'aléas...) en prenant en compte les contraintes systèmes, logicielles et matérielles.

RÉFÉRENCES

[CEA2019] R. Bernhard, P-A. Moellic *et al.* Impact of low-bitwidth quantization on the adversarial robustness for embedded neural networks. In IEEE *2019 International Conference on Cyberworlds (CW)*, 2019.

[CEA2021a] R. Joud, P-A. Moellic *et al.* A review of confidentiality threats against embedded neural network models. In *2021 IEEE 7th World Forum on Internet of Things (WF-IoT)*, 2021.

[CEA2021b] R. Joud, P-A. Moellic *et al.* A practical introduction to side-channel extraction of deep neural network parameters. In *International Conference on Smart Card Research and Advanced Applications*. Springer, 2022.

[CEA2021c] M. Dumont, P-A. Moellic *et al.* An overview of laser injection against embedded neural network models. In *2021 IEEE 7th World Forum on Internet of Things (WF-IoT)*, 2021.

[CEA2022a] B. Nguyen, P-A. Moellic *et al.* Domain generalization on constrained platforms: on the compatibility with pruning techniques. In *2022 Global Internet of Things Summit (GIoTS)*, 2022.

[CEA2022b] K. Hector, P-A. Moellic *et al*. A closer look at evaluating the bit-flip attack against deep neural networks. In *2022 IEEE 28th International Symposium on On-Line Testing and Robust System Design (IOLTS)*, 2022.

[CAR2017] N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on security and privacy (SP)*, 2017.

[SHA2018] Shafahi, A., Huang, W. R., *et al*. Poison frogs! targeted clean-label poisoning attacks on neural networks. *Advances in neural information processing systems*, NIPS 2018.

[PAP2018] Papernot, N., McDaniel, P., *et al*. SoK: Security and privacy in machine learning. In *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, 2018.

[MOT2021] Mothukuri, V., Parizi, R. M., *et al*. A survey on security and privacy of federated learning. *Future Generation Computer Systems*, 2021.

[BOU2021] Bouacida, N. *et al*. Vulnerabilities in federated learning. *IEEE Access*, 2021.

[IA_ACT] https://artificialintelligenceact.eu/