Attaques Side-Channel contre la confidentialité des modèles de Machine Learning embarqués : attaques, protections, évaluation

Objectif et contexte

Une des tendances majeures de l'Intelligence Artificielle aujourd'hui est le déploiement massif des systèmes de Machine Learning sur une multitude de plateformes embarquées. La majorité des fabricants de semi-conducteurs proposent des produits « compatibles A.I. », principalement pour des réseaux de neurones pour de l'inférence (ST [1], ARM [2]).

Outre les problématiques propres aux contraintes (mémoire, énergie, précision) des plateformes matérielles, la sécurité est un des grands freins au déploiement de ces systèmes. De nombreux travaux soulèvent des menaces aux impacts désastreux pour leur développement, comme les « *adversarial examples* » [3] ou le « *membership inference* » [4].

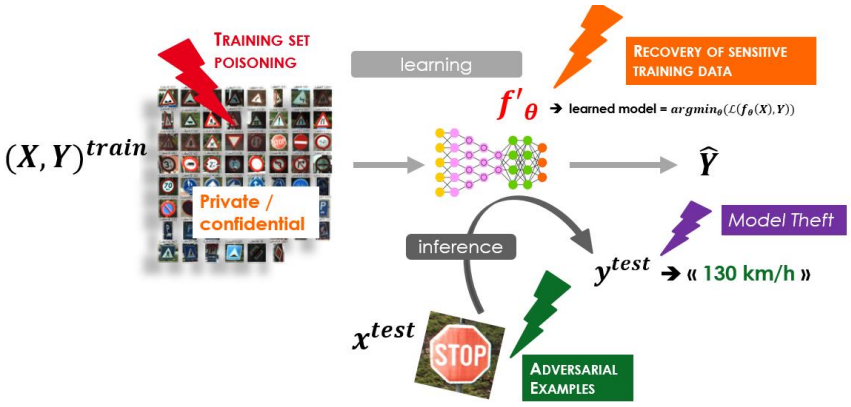

Fig : Panorama des menaces algorithmiques visant l'intégrité, la confidentialité et l'accessibilité d'un système de Machine Learning.

Néanmoins, ces travaux considèrent les algorithmes de ML selon un point de vue purement théorique sans prendre en considérations les particularités de leur implémentation matérielle. De plus, des études plus poussées sont indispensables sur les attaques physiques (side-channel et injection de fautes). En considérant une surface d'attaque regroupant les aspects algorithmiques et matériels, la thèse propose d'analyser des menaces de type Side-Channel Analysis (SCA) [5,6] ciblant la confidentialité des données d'apprentissage et des modèles (reverse engineering) des systèmes embarqués de Machine Learning et le développement de protections efficaces.

Présentation détaillée du projet doctoral

Quelques travaux commencent à s'intéresser aux attaques physiques contre des réseaux de neurones embarqués mais avec des architectures très simples sur des microcontrôleurs 8-bit, ou FPGA ou en pure simulation. Ces travaux ne proposent pas encore des liens entre les modèles de fautes ou les fuites mises en évidence et les failles algorithmiques. En se basant sur l'expérience d'autres systèmes critiques (e.g., module cryptographique), la philosophie de la thèse sera de considérer conjointement le monde algorithmique et le monde physique pour mieux appréhender la complexité des menaces et développer des protections appropriées. Aussi, la thèse s'intéressera aux questions scientifiques suivantes :

- ***Caractérisation et exploitation des fuites side-channel*** : comment exploiter les fuites de type side-channel (consommation et/ou rayonnement EM) pour retrouver des informations sensibles sur les données d'apprentissage ou des informations sur l'architecture des modèles.
- ***Evaluation des mécanismes de protections classiques*** : quel est la pertinence et l'efficacité des schémas de défenses classiques de type masking / hiding pour ce type de systèmes et de menaces ?
- ***Développement de nouvelles protections*** appropriées aux réseaux de neurones embarqués.

Cette thèse s'interfacera avec celle de Rémi BERHNARD (CFR Phare 2018-2021) portant sur les attaques algorithmiques et permettra de couvrir une surface d'attaque plus vaste et plus représentatives des futures menaces visant les systèmes embarqués basés sur du Machine Learning.

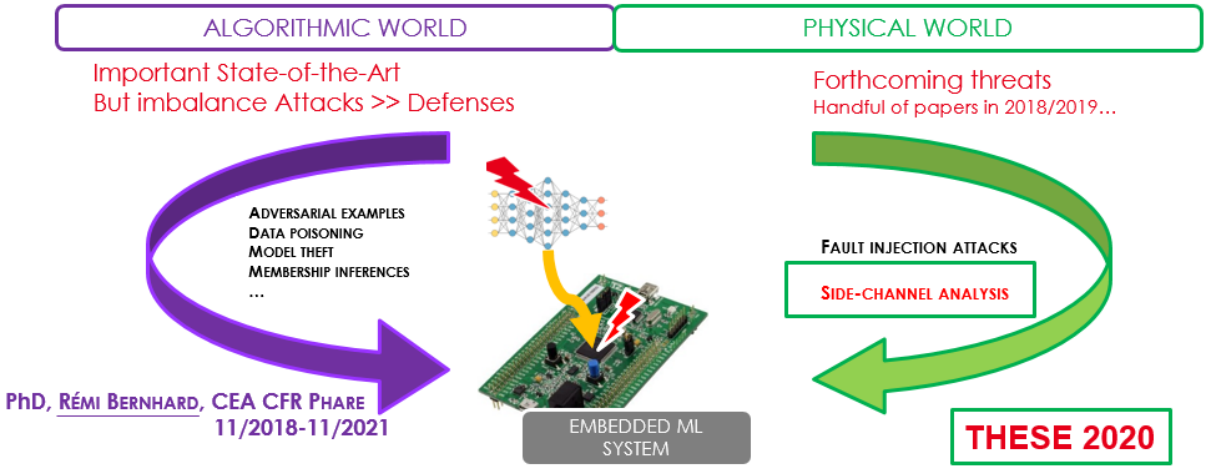

Fig. Articulation de la thèse par rapport à la thèse de Rémi BERNHARD (CFR Phare, 2018-2021)

Références

- **[1]** STM32Cube.Mx.AI : https://www.st.com/en/embedded-software/x-cube-ai.html
- **[2]** ARM-NN : https://developer.arm.com/ip-products/processors/machine-learning/arm-nn
- **[3]** C. Szegedy et al. Intriguing properties of neural networks. International Conference on Learning Representations, 2014.
- **[4]** R. Shokri et al. Membership inference attacks against machine learning models. In 2017 IEEE Symposium on Security and Privacy (SP), May 2017.
- **[5]** L. Wei et al. I know what you see: Power side-channel attack on convolutional neural network accelerators. CoRR, abs/1803.05847, 2018.
- **[6]** L. Batina et al. CSI neural network: Using side-channels to recover your artificial neural network information. Cryptology ePrint Archive, 2018