Thèse : Etude de l'impact environnemental des plateformes IoT embarquées exécutant des traitements d'apprentissage automatique

Marie-Anne LACROIX, Robin GERZAGUET et Pascal SCALART
Univ Rennes, IRISA, équipe Granit
www-granit.irisa.fr

1 Contexte

Depuis plusieurs années, la sphère de l'Internet des objets (IoT) s'est emparée de la question de l'intelligence artificielle (IA). Afin de pouvoir embarquer cette technologie sur des objets communicants fortement contraints en espace mémoire, puissance de calcul et capacité énergétique, de nombreux travaux ont été menés sur la réduction du nombre de paramètres et du nombre d'opérations de ces algorithmes [LLM23]. L'objectif principal est de rendre possible l'exécution des algorithmes IA sur une cible matérielle en réduisant la complexité calculatoire. Par ailleurs, la prise de conscience autour de l'urgence écologique soulève la question de la consommation énergétique de l'IA quel que soit son usage [BCJL24].

Les travaux menés pour réduire la complexité calculatoire de l'IA concernent uniquement la phase d'usage des dispositifs, qui influe principalement sur le dérèglement climatique via sa consommation énergétique. Il est cependant démontré que l'électronique a un fort impact sur d'autres éléments tels que l'eau ou les ressources minières pour ne citer qu'eux, et que cet impact se concentre principalement sur les phases extraction des ressources, fabrication et fin de vie [ADE23]. L'empreinte environnementale des objets électroniques est par ailleurs liée au potentiel de réutilisation de tout ou partie de ces objets après leur utilisation originelle (phase dite de *fin de vie* dans l'analyse du cycle de vie). Il convient donc de s'intéresser également aux phases hors usage qui constituent ensemble le cycle de vie des appareils électroniques.

Dans cette thèse, on veut étudier les impacts environnementaux des différentes cibles matérielles permettant un usage de l'IA dans l'IoT. Parmi celles-ci, on trouve une grande variété de possibilités selon les contraintes applicatives: depuis le simple microcontrôleur jusqu'au processeur spécifique NPU (neural processing unit), en passant par les GPU et les FPGA [FHL+24, SBA23]. A notre connaissance, il n'existe pas à ce jour de données et de modèles permettant de comprendre comment arbitrer entre ces différentes cibles d'exécution en prenant en compte différents impacts environnementaux.

Les thématiques d'impacts environnementaux des dispositifs électroniques et d'objets communicants sont adressées par les projets CMA ESOS (https://esos.insa-rennes.fr/) et CMA RIS3 (https://ris3.insa-rennes.fr/) qui participent au co-financement de cette thèse.

2 Objectifs de la thèse

La thèse s'intéresse à la question des impacts environnementaux d'objets connectés embarquant de l'IA. On s'intéressera à différentes plateformes d'exécutions et on considèrera, à titre d'exemples, deux cas d'usage de systèmes IoT : un réseau de capteurs industriel temps-réel faisant de la classification d'événements sonores (en lien avec l'ANR Edge IA "Light-swift") et un réseau de capteurs dédié au monitoring environnemental sur plateformes ultra-basse consommation (en lien avec l'ANR Thématiques Spécifiques en IA - Circuits "OWL"). L'objectif de la thèse est d'évaluer les impacts environnementaux de l'intégralité du cycle de vie d'objets connectés en étudiant l'influence de la plateforme de calcul et de la pile de communication. Trois directions peuvent ainsi être décrites par les questions scientifiques suivantes :

- **Architecture généraliste ou architecture spécialisée ?**
  Face aux contraintes particulières de l'IA et à son usage de plus en plus répandu, des architectures spécialisées telles que les NPU sont intégrées dans certaines puces embarquées. Néanmoins, ces puces exécutent généralement un jeu d'instructions limité, ce qui contraint leur usage à certaines applications et peut empêcher leur

utilisation si l'on souhaite faire évoluer le système avec de nouvelles couches ou architectures de réseaux neuronaux. Dans cette thèse on développera une méthodologie qui permettra d'évaluer l'impact environnemental de différentes architectures disponibles (microcontrôleur, GPU, FPGA, NPU) sur la base de modèles de ces architectures et de modèles d'exécution d'algorithmes IA. Un des enjeux et une originalité profonde avec l'état de l'art sera d'intégrer dans la méthodologie des métriques/facteurs intriqués pour différentes phases du cycle de vie (extraction/fabrication, usage, fin de vie) [CABJ24]. On pourra ainsi comparer le bilan environnemental de différentes architectures, spécialisées ou plus généralistes, pour différents cas d'usages (en particulier les deux précédemment mentionnés).

- **Calcul local ou distant : quand communiquer ?**
  La communication sans fil est généralement considérée comme la partie la plus énergivore d'un objet connecté [FLMFA21]. Toutefois, avec l'IA embarquée et selon le cas d'usage (algorithme IA, fréquence des calculs IA, communication intermittente ou non, quantité de données à transmettre, ...), les parts de consommation sont très variables. On cherchera à comprendre à quel point, dans la phase d'usage du système IoT, l'énergie consommée par 1) la partie communication de l'objet IoT (émetteur-récepteur), 2) la communication sans fil en elle-même, peuvent avoir une importance sur les impacts environnementaux du système, et en considérant l'ensemble du cycle de vie. Les deux cas d'usages envisagés serviront à étayer les réponses à cette question ouverte.

- **Pile ou batterie avec récupération d'énergie ?**
  On peut envisager deux types d'alimentation diamétralement différents pour les objets communicants : une pile dimensionnée pour toute la durée de vie du système, inamovible mais bien dimensionnée, ou une batterie couplée à un système de récupération d'énergie (nous considérerons dans la thèse une récupération d'énergie conventionnelle, via un capteur photovoltaïque [AGB18]). Sur la base de ces deux types de systèmes [HTY23] et pour les deux cas d'usages envisagés, nous essayerons d'identifier les critères les plus pertinents à considérer pour réaliser une analyse de l'impact environnemental en prenant en compte l'ensemble du cycle de vie.

3 Déroulé prévisionnel

Partie 1 : Etat de l'art

- impacts environnementaux des cibles d'exécution (microcontrôleur, GPU, FPGA, NPU) lors des phases extraction/fabrication et modèles d'impacts relatifs ;
- modèles de consommation d'énergie des cibles d'exécution ;
- approches pour l'analyse des impacts environnementaux de circuits et systèmes électroniques.

Partie 2 : Proposition de modèles d'impacts et d'une méthodologie d'analyse des impacts environnementaux d'objets connectés embarquant de l'IA.

Partie 3 : Partie expérimentale

- développement d'un outil d'analyse des impacts environnementaux d'objets connectés sur l'ensemble du cycle de vie ;
- confrontation des modèles de consommation d'énergie avec les mesures sur différentes cibles ; benchmarking ; raffinement des modèles ; (2 cas d'usages)
- utilisation de l'outil d'analyse des impacts environnementaux avec les systèmes avec pile ou avec batterie (avec les 2 cas d'usages).

4 Profil

Titulaire d'un master ou d'un diplôme d'ingénieur, vous avez des compétences en électronique embarquée et en intelligence artificielle. Vous avez une appétence pour la conduite de recherches scientifiques et pour les enjeux environnementaux.

5 Informations et contacts

La thèse se déroulera à Lannion, au sein de l'équipe GRANIT du laboratoire Irisa.

Marie-Anne LACROIX	marie-anne.lacroix@irisa.fr
Robin GERZAGUET	robin.gerzaguet@irisa.fr
Pascal SCALART	pascal.scalart@irisa.fr

References

[ADE23] ADEME. Evaluation de l'impact environnemental du numérique en france et analyse prospective, 2023.

[AGB18] Fayçal Ait Aoudia, Matthieu Gautier, and Olivier Berder. Rlman: An energy manager based on reinforcement learning for energy harvesting wireless sensor networks. *IEEE Transactions on Green Communications and Networking*, 2(2):408–417, 2018.

[BCJL24] Adrien Berthelot, Eddy Caron, Mathilde Jay, and Laurent Lefèvre. Estimating the environmental impact of generative-ai services using an lca-based methodology. *Procedia CIRP*, 122:707–712, 2024.

[CABJ24] Fan Chen, Shahzeen Attari, Gayle Buck, and Lei Jiang. Iotco2: Assessing the end-to-end carbon footprint of internet-of-things-enabled deep learning. *arXiv preprint arXiv:2403.10984*, 2024.

[FHL+24] William Fabre, Karim Haroun, Vincent Lorrain, Maria Lepecq, and Gilles Sicard. From near-sensor to in-sensor: a state-of-the-art review of embedded ai vision systems. *Sensors*, 24(16):5446, 2024.

[FLMFA21] Hamid Reza Farahzadi, Mostafa Langarizadeh, Mohammad Mirhosseini, and Seyed Ali Fatemi Aghda. An improved cluster formation process in wireless sensor network to decrease energy consumption. *Wireless Networks*, 27:1077–1087, 2021.

[HTY23] Kareeb Hasan, Neil Tom, and Mehmet Rasit Yuce. Navigating battery choices in iot: An extensive survey of technologies and their applications. *Batteries*, 9(12):580, 2023.

[LLM23] Zhuo Li, Hengyi Li, and Lin Meng. Model compression for deep neural networks: A survey. *Computers*, 12(3):60, 2023.

[SBA23] Matthew Sibanda, Ernest Bhero, and John Agee. Ai edge processing-a review of distributed embedded systems. In *2023 31st Southern African Universities Power Engineering Conference (SAUPEC)*, pages 1–6. IEEE, 2023.