Tremplin Recherche 2024-2025

Laboratoire d'accueil : LIGM, Université Gustave Eiffel

Encadrements : Mourad DRIDI (mourad.dridi@esiee.fr)

Titre :

Déploiement des réseaux CNN sur des architectures NPU pour un système de drones

Mots-clés : CNN, DNN, IA, NPU, NXP, MCX-N, i.MX8, Système temps réel, Drone, Ordonnancement

Contexte

Les systèmes de drones autonomes reposent de plus en plus sur l'IA embarquée pour des tâches critiques telles que la détection d'objets, la navigation et la communication en flotte. Les nouvelles cartes **NXP MCX-N** et **NXP i.MX8** intègrent des **NPU (Neural Processing Units)**, permettant d'exécuter efficacement l'inférence de réseaux de neurones convolutifs (CNN) dans un environnement embarqué contraint.

Ce projet vise à étudier et expérimenter le déploiement de CNN (ResNet, YOLO, MobileNet, etc.) sur ces cartes, avec une attention particulière portée à la **gestion des ressources matérielles** dans un contexte de **drones collaboratifs**.

Objectifs

- ➔ Déployer plusieurs CNN sur les cartes MCX-N et/ou i.MX8 pour des tâches d'inférence en temps réel.
- ➔ Analyser l'impact de ce partage sur les applications critiques (détection, pilotage, autonomie du drone).
- ➔ Étudier les contraintes supplémentaires liées à un **flot de drones** (communication inter-drones, délais de transmission, impact sur le temps de réponse).
- ➔ Proposer et évaluer un algorithme d'ordonnancement dynamique capable de :
  - ◆ Changer les priorités des tâches CNN en fonction de leur criticité et des délais temps réel.
  - ◆ Adapter la qualité/précision des inférences (ex. basculer vers une version allégée du modèle) en cas de surcharge.
  - ◆ Migrer dynamiquement l'exécution entre CPU et NPU selon l'état des ressources et les temps de réponse observés.

L'objectif est de garantir un compromis optimal entre **respect des contraintes temps réel**, **qualité d'inférence** et **utilisation efficace des ressources CPU/NPU**.

Planning

1. Déploiement des CNN sur NPU

- Prise en main des plateformes MCX-N et i.MX8, et de l'outil **eIQ Neutron NPU**.
- Intégration et test de CNN existants (YOLO, ResNet, MobileNet) pour inférence embarquée.

2. Conception de l'algorithme d'ordonnancement dynamique

- Étude des politiques d'ordonnancement classiques (FIFO, priorité fixe, PREMA…).
- Conception d'un **ordonnanceur adaptatif** prenant en compte :
    - la charge des ressources CPU/NPU,
    - les délais de réponse des tâches CNN,
    - la criticité des applications (détection et pilotage).

3. Analyse applicative

- Mesure des performances : latence, consommation, respect des deadlines.
- Évaluation de la **qualité d'inférence** en cas de dégradation (modèle réduit / moins précis).

4. Considérations inter-drones

- Implémentation d'une communication simple entre plusieurs cartes (Wi-Fi / ESP32).
- Analyse de l'effet des délais réseau sur la qualité des résultats et l'ordonnancement global.

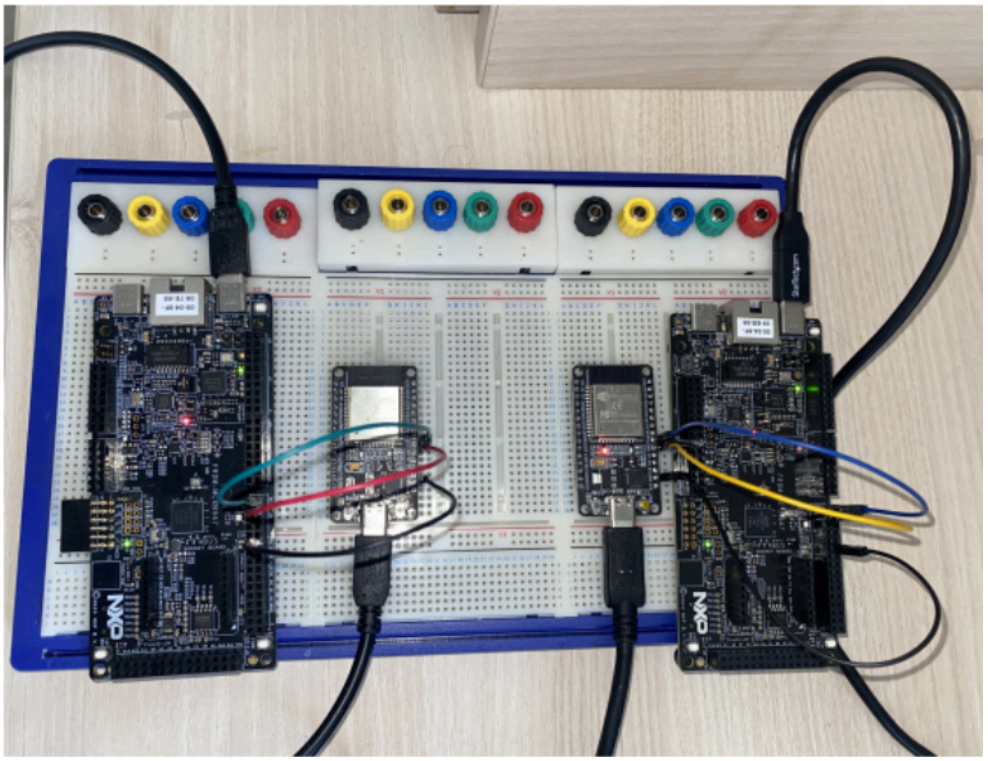
Compétences développées

- Déploiement et optimisation de CNN en environnement contraint.
- Conception d'algorithmes d'ordonnancement dynamiques (CPU/NPU).
- Analyse temps réel et compromis précision/latence.
- Programmation embarquée sur microcontrôleurs NXP (MCX-N, i.MX8).

Résultats attendus et livrables

- Un système fonctionnel exécutant plusieurs CNN avec ordonnancement dynamique.
- Un algorithme documenté permettant migration CPU/NPU et adaptation de la précision.
- Un rapport d'évaluation détaillée (performances, délais, consommation).

Note : Une partie de ce projet a déjà été développée et étudiée dans le cadre d'un précédent travail de recherche. N'hésitez pas à me contacter pour obtenir davantage d'informations ou accéder aux résultats préliminaires.



Pour les élèves E4 :

Le tremplin recherche offre une opportunité de découvrir et de se former à la recherche pendant la période académique. Il permet également de bénéficier d'un contrat d'études personnalisé, adapté aux intérêts et au profil de chaque élève.

Pour les élèves E5 :

Le sujet inclut une période initiale (novembre-février) en parallèle avec les enseignements, suivie d'une période à temps plein (stage de fin d'études de 6 mois). Le tremplin recherche constitue une opportunité pour anticiper et amorcer le stage de fin d'études au sein du laboratoire d'Informatique Gaspard Monge LIGM (6 mois à partir de février). À la fin du stage, et selon son avancement, une visite de recherche avec un partenaire international peut être envisagée

Références

[1] Y. Choi and M. Rhu, *PREMA: A Predictive Multi-Task Scheduling Algorithm For Preemptible Neural Processing Units*, HPCA 2020.

[2] D. Kang et al., *Scheduling of Deep Learning Applications Onto Heterogeneous Processors in an Embedded Device*, IEEE Access, vol. 8, 2020.

[3] NXP MCX-N Series

[4] eIQ Neutron NPU