LES ENJEUX DE L'IA GREEN EN EMBARQUE

LEMAIRE-LEFEBVRE Ilona-Marie, ilona-marie.lemairelefebvre@thalesgroup.com, THALES, Mérignac (Orateur)

GASPAROTTO Quentin, quentin.gasparotto@thalesgroup.com, THALES, Mérignac

DELAVALLADE Thomas, thomas.delavallade@thalesgroup.com, THALES, Mérignac

THALES Cell-IA

Thématique : Sobriété Numérique

Résumé :

Le Edge Computing est une technique permettant de réduire la charge de traitement du Cloud computing en distribuant les traitements en bordure du réseau. Ces sujets impliquent dans certains cas de déployer les traitements en sortie des capteurs sur des systèmes embarqués, fortement contraints en termes de ressources. Nous présenterons les travaux que nous menons depuis 2 ans sur l'IA embarquée : Compression et Segmentation d'objets sur images satellites, reconnaissance vocale embarquée et suivi de santé des téléphériques urbains.

Mots clés : Edge Computing, IA, Embarqué, IoT, Green Computing

1. Introduction

Le Edge Computing est une technique permettant de réduire la quantité de données qui transitent, ainsi que la charge de calcul du Cloud computing en distribuant les traitements en bordure du réseau.

On peut faire du Edge Computing soit sur data center local (Cloud Edge), soit sur cluster embarqué (Far Edge) soit directement en sortie de capteur (Device Edge) et dans chacun des cas, on fait face à des contraintes croissantes de :

- Consommation énergétique limitant fortement les ressources disponibles
- Réactivité des traitements devant aller au moins aussi vite que l'acquisition pour respecter le temps réel
- Précision convenable des résultats pour ne pas donner de fausses informations aux opérateurs

Pour déployer de l'IA sur du Edge computing, il est donc nécessaire de maîtriser la chaîne d'intégration de l'IA sur systèmes embarqués et de repenser complètement la conception des modèles ainsi que leur optimisation sur cible finale dans un objectif de frugalité calculatoire important.

Nous vous proposons de présenter les enjeux opérationnels du Edge Computing pour un certain nombre d'applications concrètes de nos clients :

- Compression d'image et segmentation d'objet à bord des satellites
- Reconnaissance de la parole en temps réel sur systèmes embarqués
- Suivi de santé du freinage des téléphériques urbains avec un parc IoT

Nous verrons dans chacun des cas le processus de réflexion et d'expérimentation qui nous a amené à définir notre propre framework de portage IA embarquée pour permettre aux data-scientistes et aux développeurs embarqués de dialoguer et converger vers la meilleure implémentation.

2. Méthodologie

Voici les méthodes spécifiques utilisées pour chacun de nos cas d'usage.

Compression d'image à bord des satellites : entrainement, optimisation et déploiement d'un auto-encodeur convolutif de compression d'image sur une cible FPGA Xilinx à l'aide du framework open-source N2D2 développé par le CEA.

Suivi de freinage des téléphériques urbains sur parc IoT : étude, modélisation et déploiement d'un réseau de neurones sur un parc IoT de téléphériques urbains pour la détection de l'usure lente du système de freinage des remontées mécaniques.

Commande vocale sur systèmes embarqués : entrainement, adaptation et déploiement du réseau de neurones et d'algorithmes de traitement du signal sonore en temps réel sur Raspberry Pi 4 et Arducam Pico 4ML (PicoPi).

Détection de routes à bord des satellites : entrainement, miniaturisation et benchmark de réseaux de neurones pour de la segmentation d'image à l'aide de la distillation de connaissances sur cible Nvidia Jeston Nano.

Framework FIMET: outil logiciel et méthodologie d'aide au portage et à la miniaturisation en embarqué des modèles IA à travers la génération de code source allégé provenant de modèles Deep Learning au format ONNX, qui peut être directement intégrable sur une cible embarquée, edge ou cloud. Etude des techniques de distillation des connaissances pour les réseaux de neurones et optimisation complète de la chaîne de traitement de la donnée.

3. Originalité / perspective

Les compilateurs de deep learning disponibles en Open Source se focalisent sur le déploiement de modèles IA existants sur des systèmes embarqués contraints.

Ils proposent souvent des techniques de quantification ou d'élagage pour réduire l'empreinte mémoire des poids des modèles mais ces techniques peuvent ne pas suffire pour porter des modèles volumineux.

De plus, ils occultent totalement l'adaptation des modèles aux capteurs présents sur les systèmes (ex : échantillonnage des micros).

Notre framework a pour objectif de traiter l'intégralité de la chaîne d'intégration de l'IA embarquée, de la conception des modèles incluant la miniaturisation des architectures de

réseaux de neurones, à leur optimisation fine pour un matériel d'exécution donné jusqu'aux campagnes de benchmarks sur les cibles d'exécution finales.

L'objectif étant de fluidifier les échanges entre deux expertises très différentes :

- Data Science : développer des modèles dans un environnement virtuellement illimité
- Développement embarqué : intégrer des traitements frugaux sur des systèmes très limités en ressources de calcul et peu consommateurs en énergie

Egalement nos cas d'usage couvrent un large spectre d'activités nous permettant d'éprouver notre framework sur des problématiques métiers pourtant très spécifiques.

De plus, de par la diversité de nos cas d'usage ainsi que des clients de Thales Services Numériques, nous avons une position privilégiée dans le groupe pour faciliter le transfert technologique dans toutes nos activités internes militaires comme civiles.

Nous comptons par la suite étudier la réduction de la consommation énergétique de l'IA dans le cloud en utilisant des technologies venant du monde de l'embarqué. Etant donné que nous avons développé notre framework avec de fortes contraintes de consommation énergétique, nous comptons l'utiliser pour proposer des actions de mitigations et fournir des services de Green Computing autour de la Green AI.