Tremplin Recherche 2023-2024

Étude des performances de réseaux de neurones de traitement d'image déployés sur une carte NVIDIA GPU.

Laboratoire d'accueil : LIGM (UMR 8049 CNRS), équipe LRT

Encadrements : Mourad DRIDI, Yasmina ABDEDDAÏM
mourad.dridi@esiee.fr; yasmina.abdeddaim@esiee.fr

Filières concernées : Systèmes Embarqués (SE), Artificial Intelligence and Cybersecurity (AIC), Informatique (INF).

Mots clés : DNN, CNN, GPU, Cuda, traitement d'image, Yolo, …

CONTEXTE et PROBLÉMATIQUE

De plus en plus d'applications temps réel ont besoin d'utiliser des fonctionnalités d'Intelligence Artificielle (IA). Comme exemple, les véhicules autonomes peuvent utiliser des **réseaux de neurones** afin de détecter des objets physiques et les marquages au sol en analysant les images produites par plusieurs caméras embarquées dans le véhicule.

Le déploiement des applications temps réel utilisant des fonctionnalités d'Intelligence Artificielle (IA) (lors de la phase d'inférence) sur la plate-forme d'exécution nécessite une puissance de calcul élevée, qui ne peut être satisfaite aujourd'hui que par des plateformes hétérogènes combinant CPUs et accélérateurs (GPU). Une **architecture de calcul hétérogène** distribue les données, le traitement et l'exécution des programmes entre les différentes unités de calcul qui sont les mieux adaptées aux tâches spécifiques.

Le besoin de comprendre le lien entre la qualité des images d'entrée et le temps d'exécution de la partie inférence est crucial. Différents contextes (jour, nuit, éclairage, résolution du capteur, type de caméra) peuvent entraîner des variations significatives dans le temps d'exécution des réseaux neurones. Dans un environnement temps réel, où plusieurs réseaux DNN sont exécutés sur la même carte, il est impératif de tenir compte de cette variation dans l'ordonnancement des différentes tâches du système.

OBJECTIFS

Ce projet a pour objectif d'explorer la corrélation existante entre la qualité des images d'entrée et le temps d'exécution des réseaux neurones, en se focalisant particulièrement sur les architectures NVIDIA GPU.

Dans ce contexte, l'élève sera chargé de réaliser des expérimentations visant à quantifier le temps d'exécution de la phase d'inférence pour diverses catégories d'images et différents réseaux neurones, notamment YOLO.

Les principales étapes du projet sont:

1. Mettre en œuvre des réseaux neurones de traitement d'image sur une carte NVIDIA GPU (NVIDIA JETSON AGX ORIN).
2. Conduire des expérimentations avec un jeu de données comprenant des images de différentes tailles et qualités visant à mesurer les temps d'exécution des différentes couches des réseaux neurones dans divers scénarios.
3. Déterminer une relation entre la qualité et la taille de l'image et le temps d'exécution des différentes couches des réseaux neurones.
4. Comparer les résultats obtenus avec l'état de l'art.

Ce projet offre l'opportunité d'acquérir une compréhension approfondie des performances des réseaux de neurones dans des environnements embarqués. Il contribuera également à des travaux de recherche plus large qui concernent la mise en place d'une méthodologie pour l'ordonnancement des réseaux neurones temps réel sur des architectures GPU [3].

Le partenaire international envisagé pour la poursuite en stage (mai-août) sera l'Université Washington de Saint Louis aux USA.

Références

[1] H. Andrade and I. Crnkovic, "A Review on Software Architectures for Heterogeneous Platforms," 2018 25th Asia-Pacific Software Engineering Conference (APSEC), Nara, Japan, 2018, pp. 209-218, doi: 10.1109/APSEC.2018.00035.

[2] H. Zhou, S. Bateni and C. Liu, "S3DNN: Supervised Streaming and Scheduling for GPU-Accelerated Real-Time DNN Workloads," 2018 IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS), 2018, pp. 190-201, doi: 10.1109/RTAS.2018.00028.

[3] M.Dridi, Y.Abdeddaïm and Chiara Daini, "Work In Progress: A New Task Model for Real-Time DNNs over GPU", RTAS-BP, Texas, USA, 2023

English Version

CONTEXT

A growing number of real-time applications require the use of Artificial Intelligence (AI) functionalities. For example, autonomous vehicles can use neural networks to detect physical objects and road markings by analyzing images produced by several cameras onboard the vehicle.

Deploying real-time applications using Artificial Intelligence (AI) functionalities (during the inference phase) on the execution platform requires high computing power, which today can only be satisfied by heterogeneous platforms combining CPUs and GPUs. A heterogeneous

architecture distributes data, processing and program execution among the different computing units that are best suited to specific tasks.

The need to understand the link between input image quality and the execution time of the inference part is crucial. Different contexts (day, night, lighting, sensor resolution, camera type) can lead to significant variations in the execution time of neural networks. In a real-time environment, where several DNNs are running on the same board, it is imperative to take this variation into account when scheduling the various system tasks.

OBJECTIVES

The aim of this project is to explore the correlation between input image quality and neural network execution time, with a particular focus on NVIDIA GPU architectures.
In this context, the student will do experiments aimed at quantifying the execution time of the inference phase for various image categories and neural networks, including YOLO.

The main stages of the project are:

1. Implement image processing neural networks on an NVIDIA GPU (NVIDIA JETSON AGX ORIN).
2. Conduct experiments with a dataset comprising images of different sizes and qualities, aimed at measuring the execution times of the different neural network layers in various scenarios.
3. Determine the relationship between image quality and size and the execution time of the different neural network layers.
4. Compare the results obtained with the state of the art.

This project offers the opportunity to gain an in-depth understanding of neural network performance in embedded environments. It will also contribute to broader research work concerning the implementation of a methodology for scheduling real-time neural networks on GPU architectures [3].

The international partner envisaged for the internship (May-August) will be Washington University in Saint Louis, USA.