Titre : Hardware Aware Differentiable Architecture Search for Deep Learning

Supervisors:

Yinghao Wang (yinghao.wang@telecom-paris.fr)

Van-Tam Nguyen (van-tam.nguyen@telecom-paris.fr)

Summary:

In this master II internship, we will study the current state of the art on hardware-aware differentiable Neural Architecture Search and propose to learn a differentiable function to approximate hardware performance on a given hardware platform. As a result, this function provides hardware feedback that can be directly integrated into the Differentiable Architecture Search (DARTS) flow, allowing the search for efficient and accurate architectures.

Description:

Deep neural networks (DNNs) are gaining popularity in a wide variety of embedded intelligent scenarios, such as computer vision, virtual reality, object detection and tracking, etc., offering impressive performance and enabling entirely new experiences on devices. Architecture design plays a crucial role for DNNs. Nevertheless, since the network design space is very large, the manual design of competitive DNNs requires enormous engineering efforts to determine the optimal network configuration, such as its depth and width. For this reason, neural architecture search (NAS), which aims to automate the design of high-quality DNNs, has recently flourished.

In the literature, studies on neural architecture search are mainly divided between reinforcement learning, evolution, and gradient-based or differentiable categories. However, the reinforcement learning and evolution-based NAS approaches suffer from substantial search overhead (several thousand GPU days), while the differentiable solution has significant search efficiency that reduces the search cost by several orders of magnitude. Due to its high search efficiency, the differentiable NAS scheme has recently emerged as the most dominant alternative in the NAS community.

Despite significant progress, early differentiable NAS approaches such as DARTS are hardware indifferent. Specifically, they focus on finding architectures that are competitive in terms of accuracy, without considering other critical performance constraints such as latency, energy, and memory. Thus, they arrive at the architecture that offers promising accuracy for the target task, but at the cost of high computational complexity on the target hardware, which hinders the practical deployment of DNNs, especially on resource-constrained embedded platforms.

In this internship, we address these challenges by exploring existing solutions and proposing a hardware-aware DARTS framework that automatically searches for efficient DNNs on different hardware types, taking into account constraints related to computing, energy and memory capacities. Specifically, we will study the current state of the art on

the topic and propose to learn a differentiable function to approximate hardware performance on a given hardware platform. As a result, this function provides hardware feedback that can be directly integrated into the DARTS flow, allowing the search for efficient and accurate architectures.