DEVELOPING EFFICIENT SOFTWARE AND HARDWARE FOR ARTIFICIAL INTELLIGENCE VIA COLLABORATIVE OPTIMISATION

Problem and solution

Developing competitive Artificial Intelligence (AI) -based products requires years of intensive R&D to come up with an efficient software and hardware solution given an overwhelming number of combinations of algorithms, models, features, data sets, frameworks, libraries and platforms. The main risk is to make wrong choices which can lead to a slower, less accurate and expensive solution than from a competitor and potentially waste the whole investment. To give a concrete example: autonomous driving requires robustly detecting cars, pedestrians and other objects under a variety of conditions. State-of-the-art algorithms for object detection, however, either process accurate images too slowly (1 image 4 seconds) or do not meet functional safety requirements, for example, fail to recognise pedestrians in low-light conditions. Nevertheless, they require running on compute platforms that consume hundreds Watts of power and cost thousands of euros. While this is acceptable for proof-of-concepts, it is prohibitively expensive for mass production. Collaborative optimization can bring automotive platforms that consume perhaps under ten Watts of power and cost perhaps under a hundred euros, while meeting recognized safety standards.

In a 50k€ TETRACOM experiment the non-profit cTuning foundation has developed such a collaborative optimization framework called Collective Knowledge (CK) to collaboratively optimise software and hardware for emerging workloads. CK enables industry and academia to share reusable and customisable Artificial Intelligence (AI) artefacts and workflows with a common application programming interface (API) while facilitating technology transfer. Continuously aggregating collaborative optimisation results obtained on systems ranging from Internet of Things (IoT) devices to supercomputers helps automatically predict most efficient solutions, and therefore dramatically accelerate R&D, save millions of Euros, minimize risks and reduce time to market for new AI products. Automatic and collaborative CK-based software optimisation has already enabled several components of deep neural networks (essential part of AI) to run 10-30x faster on ARM-based hardware while reducing time to market by 5-10 times.

The role of the DIH

The cTuning foundation served as a DIH to transfer the CK technology to ARM, the world-leading supplier of microprocessor technology with over 100 billion ARM-based chips deployed since 1991. Adopting CK-based workflows provided the critical know-how and skills to extrapolate experimental results using predictive analytics and enabled ARM to optimise their software and hardware for AI workloads in only a fraction of the time required by conventional optimisation. Furthermore, the DIH facilitated the foundation of a start-up company that commercially exploits the developed solution.

Impact

The knowledge, experience and open source technology acquired from this experiment helped establishing a start-up called dividiti in 2015. Within 2 years, dividiti became a leading provider of AI optimisation services for ARM and several Fortune 50 companies including General Motors, and grew from the 2 co-founders to 7 staff with over €1M in revenue.

End users: ARM (UK), dividiti (SME/start-up, UK)
DIH: cTuning foundation (RTO, FR)