Deployment of Transformers on Energy-Efficient Parallel SoCs for Satellite Applications

Machine Learning (ML) algorithms have been a major focus in both software and hardware research landscapes, introducing novel and innovative model architectures, such as Transformers. These innovations enabled the birth of powerful tools such as Generative Pretrained Transformers (GPT), used for popular applications like automatic chat-bots or text/image generation.

This exponential growth, however, comes with a heavy cost, both in training and inference, as these models have been designed to be run on high power and resource unconstrained machines, presenting different challenges like extremely high number of parameters or expensive functions like the Softmax activation.

This project aims at filling the gap in deployment and training of modern Transformer models on single-board computers (SBCs) attuned for space and satellite applications. The candidate will study and develop software solutions and optimizations for low-power and heterogeneous architectures, such as PULP, with the end goal of presenting a low-level library for an automatic deployment flow Transformer-based models and applications.

The activity, which is in line with the objectives of the ChipsJU project ISOLDE, will consist of the following steps:

1. Development of a Transformer low-level library for embedded parallel ultra-low-power devices
2. Development of an automatic deployment flow
3. Evaluation on a real-world prototype (Carfield, Astral) of space-ready Heterogeneous SoC.

Deployment di Transformer su SoC Paralleli Energeticamente Efficienti per Applicazioni Satellitari

Gli algoritmi di Machine Learning (ML) sono stati al centro della ricerca sia nel software che nell'hardware, introducendo nuove e innovative architetture di modelli, come i Transformers. Queste innovazioni hanno permesso la nascita di potenti strumenti come i Generative Pretrained Transformers (GPT), utilizzati in applicazioni popolari come chatbot automatici o generazione di testo/immagini.

Tuttavia, questa crescita esponenziale ha un costo elevato, sia in termini di addestramento che di inferenza, poiché questi modelli sono stati progettati per funzionare su macchine ad alta potenza e senza vincoli di risorse, presentando sfide come un numero estremamente elevato di parametri o funzioni costose come l'attivazione Softmax.

Questo progetto mira a colmare il divario nella distribuzione e nell'addestramento dei moderni modelli Transformer su computer a scheda singola (SBC) orientati alle applicazioni spaziali e satellitari. Il candidato studierà e svilupperà soluzioni software e ottimizzazioni per architetture a bassa potenza ed eterogenee, come PULP, con l'obiettivo finale di presentare una libreria di basso livello per un flusso di distribuzione automatica di modelli e applicazioni basate su Transformer.

L'attività, che è in linea con gli obiettivi del progetto ChipsJU ISOLDE, consisterà nei seguenti passaggi:

1. Sviluppo di una libreria Transformer di basso livello per dispositivi embedded ultra-low-power paralleli.
2. Sviluppo di un flusso di distribuzione automatico.
3. Valutazione su un prototipo reale (Carfield, Astral) di SoC eterogenei pronti per lo spazio.