ID de Contribution: **31** Type: **Exposé (20 + 10 min)**

Gestion des ressources en mémoire des GPU pour l'entraînement de Graph Neural Network (GNN) sur de larges graphes.

jeudi 17 novembre 2022 09:30 (30 minutes)

L'entraînement de modèles GNN sur de larges graphes est très couteux en mémoire. Cela représente un défi majeur pour les expériences dont les données éparses sont représentées par des graphes de très grande taille. Nous expliquerons les raisons de ce coût en mémoire spécifique aux architectures GNN et pourquoi les problèmes de dépassement de mémoire ne peuvent pas être résolus avec des approches multi-GPU de type distribution de données (Distributed Data Parralel), ou de type distribution du modèle (Distributed Model Parallel).
Nous présenterons les techniques de offloading et de checkpointing comme solutions potentielles au dépassement mémoire mais avec un coût en temps de calcul supplémentaire.
Avoir une utilisation plus efficace de la ressource GPU en cherchant le meilleur compromis entre la consommation en mémoire et temps de calcul permet de réduire le temps d'entraînement des modèles GNN, accélérer la recherche et tendre vers plus de sobriété énergétique.
Nous présenterons une étude comparative des performances en termes de temps de calculs et de consommation mémoire entre ces deux techniques appliquées à un cas concret : L'entraînement de modèles GNN pour la reconstruction de traces de particules chargées à partir de simulation réalisée dans ATLAS-ITk dans les conditions HL-LHC.

Auteur principal: CAILLOU, Sylvain (L2I Toulouse, CNRS/IN2P3, UT3)

Co-auteurs: VALLIER, Alexis (L2I Toulouse, CNRS/IN2P3, UT3); ROUGIER, Charline (L2I Toulouse, UT3, CNRS/IN2P3); STARK, Jan (L2I Toulouse, CNRS/IN2P3, UT3)

Orateur: CAILLOU, Sylvain (L2I Toulouse, CNRS/IN2P3, UT3)

Classification de Session: Prospectives