Compas'2024 : Parallélisme / Architecture/ Système
IETR/LS2N/Nantes-U - Nantes - France, 2-5 juillet 2024

StarONNX : Un ordonanceur dynamique pour une inférence rapide et à haut débit sur des ressources hétérogènes.

Olivier Beaumont, Jean-François David, Lionel Eyraud-Dubois, Samuel Thibault

Centre Inria de l'Université de Bordeaux,
200 Av. de la Vieille Tour, 33405 Talence - France
prenom.nom@inria.fr

Résumé
L'inférence efficace des modèles de réseaux de neurones profonds (DNN) sur des processeurs hétérogènes est complexe, non seulement en raison de l'hétérogénéité entre les CPU et les accélérateurs matériels, mais aussi parce que le problème est fondamentalement bi-objectif dans de nombreux contextes, où la latence (le temps nécessaire pour effectuer une inférence) et le débit (le nombre d'inférences par unité de temps) doivent être optimisés. Nous présentons StarONNX, une solution basée sur l'intégration d'ONNX Runtime dans StarPU, visant à optimiser la répartition des tâches d'inférence et la gestion des ressources sur des architectures hétérogènes. Cette stratégie repose sur (i) l'exécution efficace des modèles de deep learning par ONNX Runtime pour maximiser l'efficacité individuelle des ressources, et (ii) l'orchestration des ressources hétérogènes par StarPU pour fournir des stratégies sophistiquées de planification et de superposition des calculs et des communications. Un point essentiel de la solution est la capacité de diviser un modèle DNN en deux parties, l'une fonctionnant sur le GPU et l'autre sur le CPU, augmentant ainsi le débit en utilisant toutes les ressources disponibles. Nous avons également évalué notre approche par rapport à Triton Inference Server et montré une amélioration significative de l'utilisation des ressources et une réduction de la latence.

Mots-clés : Ordonnancement dynamique, Inférence des réseaux de neurones profonds, Hétérogénéité des ressources, Débit / Latence

---