Avis de Soutenance

Monsieur Haoran WANG

Informatique

Soutiendra publiquement ses travaux de thèse intitulés

Parallélisme implicite pour la accélération de réseaux de neurones

dirigés par Monsieur SEBASTIEN LIMET
Ecole doctorale : Mathématiques, Informatique, Physique Théorique et Ingénierie des Systèmes - MIPTIS
Unité de recherche : LIFO - Laboratoire d'Informatique Fondamentale d'Orléans

Soutenance prévue le ***jeudi 27 octobre 2022*** à 14h00
Lieu : 6 Rue Léonard de Vinci, 45067 Orléans, France
Salle : Amphi H

Composition du jury proposé

M. SEBASTIEN LIMET	Université d'Orléans	Directeur de thèse
Mme Sophie ROBERT	Université d'Orléans	Co-encadrante de thèse
M. Denis BARTHOU	ENSEIRB-MATMECA	Rapporteur
M. Noel DE PALMA	Université Grenoble Alpes	Rapporteur
M. Chong LI	Huawei Technologies France S.A.S.U	Co-encadrant de thèse
M. Serge PETITON	Université de Lille	Examinateur

Mots-clés : Calcul parallèle,Entrainement réparti,Apprentissage profond,Analyse de coût,Modélisation de haut niveau,Parallèle hybride

Résumé :

Ces dernières années, le domaine de l'intelligence artificielle (IA) s'est développée avec des succès spectaculaires et très médiatisés. En fait, l'IA est appliquée dans de nombreux domaines, de la vision par ordinateur au traitement du langage naturel. Parmi toutes les techniques d'intelligence artificielle, l'apprentissage profond basé sur les réseaux de neurones a montré des capacités d'apprentissage exceptionnelles avec de très bonnes performances dans de nombreux domaines. La conception et le développement de ces réseaux est une tâche ardue qui nécessite des connaissances avancées en matière d'architectures parallèles modernes afin d'exploiter au mieux la puissance de calcul de ces machines. Une tendance notable des réseaux neuronaux est l'augmentation exponentielle de leur taille dans la recherche de résultats de classification et de prédiction plus précis. La formation d'un réseau étendu prend souvent des semaines, voire des mois, et les réseaux de grande taille peuvent généralement dépasser les limites de mémoire des accélérateurs de calcul individuels. Pour ces deux raisons, les milieux universitaires et industriels commencent à utiliser des grappes d'ordinateurs pour former des réseaux neuronaux de manière distribuée. Les méthodes de partitionnement couramment utilisées pour distribuer un réseau neuronal comprennent le parallélisme de donnée, le parallélisme de modèle au niveau des opérateurs, le parallélisme de modèle en pipeline, etc. De nos jours, la performance optimale d'un réseau neuronal complexe est généralement obtenue en utilisant un mélange des méthodes de parallélisme ci-dessus, ce que l'on appelle parallélisme hybride. L'élaboration d'une stratégie de partitionnement nécessite des connaissances en calcul parallèle pour les chercheurs et les ingénieurs en IA, ainsi que du temps et des efforts pour concevoir et vérifier les performances. Des universitaires ont proposé des méthodes telles que OptCNN, Tofu, Piper, Alpa, etc., qui peuvent donner automatiquement des stratégies hybrides quasi-optimales en utilisant des graphes de calcul et des dispositifs matériels comme entrées. Cependant, les modèles de coûts numériques proposés par les méthodes ci-dessus sont tous basés sur le temps d'exécution de l'opérateur de profilage sous un matériel particulier. Ce type d'approche introduit un effort de préparation coûteux sans garantie d'optimalité. En outre, il faut des heures, voire des jours, pour trouver la stratégie optimale pour un réseau neuronal étendu. Cette thèse vise à éviter le processus de profilage coûteux des méthodes de l'état de l'art et à fournir un algorithme pour donner une politique parallèle hybride précise en peu de temps. Basée sur le modèle BSP, cette thèse propose une machine abstraite hiérarchique symétrique et un modèle de coût symbolique qui découple le matériel de l'algorithme parallèle, éliminant ainsi le besoin de profilage sur du matériel spécifique pour chaque opérateur. Sur la base de la sémantique des réseaux de neurones informatiques, le modèle de coût symbolique peut être transformé et réduit. Cette thèse propose un algorithme qui réduit la complexité des problèmes de recherche NP-hard en atteignant la linéarité, générant ainsi des déploiements parallèles hybrides efficaces en quelques secondes. Les résultats visent à être intégrés dans l'environnement open-source MindSpore de Huawei et à contribuer aux produits et solutions d'IA de Huawei afin d'explorer toute la puissance potentielle de son matériel de puces Ascend.