



Détection de paludisme par réseaux de neurones profonds hautes performances

Contexte

Le projet CellAvIe est un projet collaboratif entre une équipe de biologie de l'École Pratique des Hautes Études, une équipe d'opticiens de l'Institut Polytechnique de Paris et une équipe de recherche en informatique spécialisée dans les systèmes parallèles et distribués de l'Institut Polytechnique de Paris. Ce projet a pour ambition de faire progresser les capacités d'observation des tissus biologiques (pendant optique) et d'exploitation de ces données (pendant informatique) pour développer de nouvelles technologies à destination des biologistes. À ce titre, les opticiens impliqués dans le projet développent un nouveau type de microscope (microscope holographique à ouverture synthétique) qui permet d'acquérir très facilement des données nativement quantitatives, multimodales et massives. Conjointement à ce travail, l'équipe informatique vient très récemment de déployer une infrastructure matérielle dénommée LABIA, destinée au traitement de données par apprentissage profond, en vue de l'utiliser sur les données produites par le microscope. En particulier, la première application que nous visons concerne le diagnostic de patients potentiellement infectés par le paludisme, où l'enjeu est de détecter une cellule parasitée parmi un million en moins de 40 min.

Objectifs du stage

Un premier prototype logiciel fonctionnel permet de détecter des cellules parasitées grâce à un chaînement de réseaux de neurones de différente nature (notamment CNN, YOLO, R-CNN) assurant le prétraitement des images (*e.g.* refocalisation numérique des images), ainsi que l'application d'opérations de plus haut niveau (*e.g.* segmentation, détection de régions d'intérêt). Cette chaîne est encore exploratoire et doit par conséquent être régulièrement modifiée et réentraînée. La durée de cet entraînement se compte actuellement en jours et est considérée comme beaucoup trop longue. Cependant, dans la mesure où aucune optimisation n'a été effectuée jusqu'à présent, il est tout à fait envisageable qu'il y ait de fortes marges de progression. La bonne maîtrise des performances calculatoires en phase d'exploitation des réseaux est également critique. Chaque seconde gagnée étant importante pour le pronostic des patients, les méthodes de calcul doivent là-encore être optimisées.

Pour aller dans cette direction, le travail de stage sera dédié à l'optimisation des temps de calcul en phase d'apprentissage des réseaux CNN mis en place. Vous chercherez dans un premier temps à définir et à mettre en place une méthodologie d'évaluation des performances (taille des données, taille des patches, empreinte mémoire) en utilisant un unique processeur graphique. Vous chercherez dans un second temps à optimiser l'exploitation des ressources

de ce processeur et ainsi définir les meilleures performances atteignables. Dans un troisième temps, vous concentrerez vos efforts sur la distribution de l'application sur plusieurs GPUs. TENSORFLOW, outil open source d'apprentissage automatique développé par Google, propose des stratégies sommaires de distribution de la phase d'apprentissage de ses réseaux. Il s'agira de les évaluer en suivant la méthodologie préalablement établie, de calibrer les ressources de calcul à mettre en œuvre en adéquation avec la taille des données à traiter, d'identifier les goulets d'étranglement qu'engendrera potentiellement le traitement de nos données atypiques et enfin de proposer des solutions d'améliorations.

Mots clés : Supercomputing, Deep learning, Optimisation de performances, Systèmes parallèles et distribués, GPU, Tensorflow, CNN

Profil recherché

- Étudiant en master 2, dernière année d'école d'ingénieur ou assimilé ayant :
- des compétences en développement informatique (souhaité Langage C)
 - des connaissances en apprentissage par réseaux de neurones profonds
 - idéalement des connaissances en programmation GPU

Informations pratiques

- Le projet de master financé, avec un fort potentiel de financement de doctorat.
- Durée : 6 mois - Période : démarrage possible dès février 2020
- Le stage se déroulera dans les locaux de l'Institut Mines-Télécom de l'Institut Polytechnique de Paris.
Localisation : 19 place Marguerite Perey, 91120 Palaiseau.
- Pour candidater, merci de contacter Elisabeth Brunet - elisabeth.brunet@telecom-sudparis.eu