Labex DigiCosme : appel pour le financement d'un ingénieur CDD

Groupe de travail associé : ≪ Réseaux profonds et Représentations Distribuées ≫

Animateurs du groupe de travail : Alexandre Allauzen et Guillaume Charpiat

Laboratoire gérant le financement : LIMSI/CNRS, UPR 3251

Équipes concernées :
- Traitement du Langage Parlé (LIMSI/TLP) : A. Allauzen
- Apprentissage & Optimisation (LRI/A&O) : G. Charpiat
- Multimédia (Telecom-ParisTech/LTCI) : L. Likforman

Axe et Tâche concernés : ≪ Datasense / Machine learning ≫

Type de contrat : CDD ingénieur

Durée et dates envisagées du contrat : 1 an avec un démarrage du contrat entre septembre 2017 et janvier 2018

Table des matières

1 Contexte scientifique

Le groupe de travail (GT) « Réseaux profonds et Représentations Distribuées » a commencé ses travaux en mai 2014. Son but est de fédérer les équipes de DigiCosme qui étudient et/ou utilisent les réseaux de neurones profonds. Ces nouveaux types de réseaux ont permis récemment des avancées scientifiques et technologiques significatives en intelligence artificielle, conduisant à des systèmes aux performances inédites dans de nombreux domaines d'application liés au Labex (la reconnaissance de la parole, la traduction automatique, la reconnaissance de l'écriture manuscrite, la reconnaissance visuelle d'objets et de lieux en robotique, ...). Ces avancées ont d'ailleurs fait la une du New York Times et les réseaux profonds ("*deep learning*") ont été retenus comme une des 10 percées technologiques les plus importantes de 2013 par le MIT Technology Review.

Ce domaine de recherche a donné lieu ces dernières années à une littérature foisonnante. Néanmoins, la recherche dans ce domaine nécessite des développements logiciels conséquents, associés à des ressources de calcul dédiées. L'usage des processeurs graphiques ou GPU est à ce titre crucial étant donné le coût computationnel des expériences à mener.

Dans ce contexte, plusieurs équipes de recherche qui émargent dans le GT et au-delà se sont fédérées pour répondre à l'appel de l'institut INS2I du CNRS visant à financer des plateformes de calcul. Ce projet a été accepté à l'automne 2016 et une plateforme GPU a été mise en place. Cette plateforme a vocation à être agrandie. Un projet visant à la décupler a été soumis à un financement de la région (appel Sésame) et une soumission à l'appel de l'Université Paris-Sud (ERM : équipement de recherche mutualisé) est en cours de finalisation. Si ces propositions sont acceptées, les nouveaux GPUs seraient alors prévus d'ici janvier 2018, portant le parc à près d'une centaine de cartes GPU.

L'objectif du contrat ingénieur demandé est d'animer la partie expérimentale du GT en lien avec ses membres, de participer à la veille scientifique et à la maintenance logiciel cette plateforme GPU afin de faciliter le travail collectif autour des réseaux de neurones profonds.

2 Projet scientifique et mission du contrat ingénieur

La mission principale du CDD ingénieur sera l'animation du volet expérimental du GT. Cette mission s'articule selon 2 axes principaux : la maintenance d'un socle logiciel de base permettant l'utilisation des GPU, et la création d'une plateforme expérimentale commune permettant de partager données et code.

2.1 Maintenance de la plateforme GPU

Il s'agit ici du socle de base pour utiliser les GPU : installation et maintenance logicielle, outils de gestion de ressources multi-utilisateurs de calcul partagées et librairies de base de la communauté internationale en apprentissage statistique.

Installation et maintenance du matériel L'installation et la maintenance de la plateforme GPU actuelle a demandé de nombreuses heures de travail et a été assumée pour l'instant par les ingénieurs systèmes des différents laboratoires concernés, cette activité venant se surajouter à leur charge de travail au sein de leurs unités. Cette façon

de procéder ne sera plus tenable lors du développement prévu de cette plateforme dans les prochains mois, et il sera critique de pouvoir y affecter des ressources dédiées à plein temps pour soulager les équipes système. Par ailleurs, les différents labos participant au GT sont demandeurs de compétences en GPU, et l'ingénieur recruté permettra le partage de l'expérience acquise sur cette plateforme en matière de GPU.

Étude et configuration d'un système de gestion de jobs : Outre les installations de matériel, il est également nécessaire d'installer un outil de gestion de ressources de calcul mutualisées permettant aux utilisateurs de soumettre des calculs via une file de priorité. Ce type de gestion est désormais indispensable, étant donnée la croissance prévue, tant en terme d'utilisateurs que de GPUs, afin d'exploiter au mieux les ressources et d'optimiser l'utilisation des ressources. Les cartes GPU n'étant pas toutes identiques (taille de la mémoire, précision, etc.), cette gestion prendra en compte les spécificités des tâches soumises par les utilisateurs. Une solution telle que SLURM [1] est actuellement à l'étude et son déploiement fera partie des missions prioritaires du CDD ingénieur.

Suivi et information logicielle : Les travaux des chercheurs membres du GT s'appuient pour la plupart sur des bibliothèques de calcul permettant un usage efficace des GPU. Un premier recensement au sein des membres du GT a permis d'identifier 3 bibliothèques qui sont au coeur des applications actuelles : Theano, Tensorflow et Torch, ainsi que leurs interfaces python respectives, Keras et PyTorch. Ces bibliothèques évoluent rapidement et il est nécessaire d'en assurer la cohérence des installations logicielles, afin de permettre le bon déroulement des expériences. Il est également indispensable de produire une documentation de ces différentes bibliothèques afin de permettre des développements plus rapides et de fournir des éléments de comparaison.

2.2 Création d'une plateforme expérimentale commune

Ce deuxième volet du travail du CDD ingénieur consiste à développer une plateforme de jeux de données, d'évaluation, et de logiciels propres au groupe de travail recensant les contributions de ses membres, afin de faciliter les échanges au sein du groupe.

Benchmarks pour l'évaluation des modèles : Les différentes équipes impliquées dans le GT développent des modèles neuronaux complexes dont l'évaluation expérimentale représente à la fois un enjeu scientifique et une lourde tâche. Néanmoins, ces modèles peuvent pour la plupart s'appliquer à différentes tâches si les données sont au préalable préparées et mises en forme comme il convient. Le fait d'évaluer un modèle sur des tâches différentes permettra donc un développement plus efficace de modèles robustes, tout en renforçant les échanges entre les différentes équipes. Ainsi, le rôle de l'ingénieur sera de mettre en forme différents jeux de données dont la liste sera établie par les membres du GT afin de fournir des benchmarks à l'ensemble des équipes du GT. Parmi ces jeux de données déjà disponibles et utilisés au sein des différentes équipes concernées, on peut déjà citer :

— ImageNet (14 millions d'images, 22 000 classes) ;
— Cifar (benchmarks de 60 000 images, 10 et 100 classes) ;

1. https://slurm.schedmd.com/

— Les données de la campagne d'évaluation internationale de traduction automatique adossée à la conférence internationale WMT (Conference on Machine Translation), pour les paires de langues Anglais-(Tchèque, Français, Allemand) ;
— Reconnaissance automatique de la parole : TIMIT et SwitchBoard (parole lue et conversationnelle) ;
— Analyse en dépendance : UDT (Universal Dependency Treebank) ;
— Les données des évaluations internationales organisées dans le cadre de ICDAR (International Conference on Document Analysis and Recognition) sur la reconnaissance d'écriture.

Création d'une bibliothèque commune : La création de jeux de données partagés permettra d'impliquer différentes équipes au sein du GT et de développer différents modèles. Le rôle de l'ingénieur sera dans ce contexte de fédérer ces efforts de développement en créant et maintenant une bibliothèque open-source à partir des contributions de chacun, le tout dans un cadre unifié favorisant le partage de ces outils.

Les contributions logicielles peuvent être de plusieurs natures :
— prétraitement des données ;
— initialisation des réseaux de neurones ;
— gestion des données lors de l'apprentissage : création de mini-batchs équilibrés, sélection automatique d'une taille raisonnable de mini-batch, etc ;
— optimiseurs efficaces, sélection automatique d'optimiseurs appropriés ;
— architectures ayant fait leurs preuves sur telle ou telle tâche ;
— outils de visualisation (pour la compréhension de ce que font les réseaux de neurones et pour leur débuggage) ;
— évaluation automatique et tableau des scores (performance des différents modèles sur les différentes tâches) ;
— documentation pour l'aide aux nouveaux utilisateurs ;
— documentation des astuces d'entraînement des réseaux de neurones en général.

En plus des tâches précédentes, l'ingénieur recruté **participera activement à l'animation du GT** et **organisera des tutoriels** favorisant l'utilisation des bibliothèques existantes et la **dissémination du code développé** au sein du GT.

3 Interaction avec le groupe de travail

Cette demande est portée par les responsables du groupe de travail « Réseaux profonds et Représentations Distribuées ». Ce groupe de travail existe depuis près de trois ans et a vu ses effectifs croître de manière continue. Les missions envisagées pour ce contrat d'ingénieur ont pour objectif d'accentuer les collaborations au sein du groupe en initiant des expérimentations partagées.

4 Budget

Le coût mensuel pourra s'échelonner entre 2 839 et 3 968 euros mensuel (coût avec PPE ANR 5,40%) selon l'expérience et les qualifications du candidat, soit un coût annuel entre 34 068 et 47 616 euros, l'objectif étant d'être attractif et de tenir compte que ce type de compétence est également très convoitée dans le secteur privé.

5 Annexe : Fiche de poste préliminaire

5.1 Mission

La mission de l'ingénieur s'inscrit dans l'activité d'un groupe de travail (GT) financé par le Labex Digicosme « Réseaux profonds et Représentations Distribuées ». Son but est de fédérer les équipes de DigiCosme qui étudient et/ou utilisent les réseaux de neurones profonds. Outre les développements méthodologiques liés à ces modèles, les partenaires de ce GT s'intéressent à leur application à des tâches et des données de différentes natures : la reconnaissance de la parole, la traduction automatique, la reconnaissance de l'écriture manuscrite, la reconnaissance visuelle d'objets et de lieux en robotique.

— L'installation et la maintenance de la plateforme GPU : intégration des nouveaux serveurs de GPU à la plateforme existante et installation des bibliothèques nécessaires.
— Étude et configuration d'un système de gestion de jobs prenant en compte les spécificités des différentes cartes GPU.
— Suivi et information logicielle : mise à jour logicielle, documentations des bibliothèques installées en particulier Theano, Tensorflow et Torch, ainsi que leurs interfaces python respectives, Keras et PyTorch.
— Mise en place des benchmarks pour l'évaluation des modèles : développement d'API d'accès aux différents jeux de données, documentations.
— Création et maintenance d'une bibliothèque commune intégrant les modèles développés au sein du GT.
— Participation à l'animation du GT et organisation de tutoriels favorisant l'utilisation des bibliothèques existantes et la dissémination du code développé au sein du GT.

5.2 Facteurs d'évolution à moyen terme

La plateforme de GPU existante a pour vocation à croître et d'intégrer différents types de carte afin de répondre aux besoins des membres du GT.

5.3 Impact sur l'emploi-type

— Implication importante dans les aspects calcul sur GPU, configuration réseau et logicielle.
— Animation et implication dans la construction d'une bibliothèque commune en python et s'appuyant entre autres sur les frameworks Theano, Tensorflow et Torch.

5.4 Compétences principales

Connaissances :
— Mathématiques
— Administration système (UNIX) et réseau
— Bibliothèques mathématiques python
— Algorithmique
— Système de gestion de ressources (SLURM)
— Méthodes de modélisation et de développement logiciel

— Apprentissage automatique, Réseaux de neurones

Compétences opérationnelles :

— **Administration de cluster GPUs**
— Développer en python
— Rédiger la documentation
— Gérer un référentiel technique
— Préparer et animer une session de formation
— Assurer une veille logiciel

5.5 Diplôme réglementaire exigé

Le candidat doit être titulaire d'un master en informatique et d'un diplôme d'école d'ingénieur au minimum.