2026-2027
Sujet de stage M1/M2
Titre: Quantification Mixte des Réseaux de Neurones : Arithmétique à Virgule Fixe, Flottante et Analyse de l'Entropie

Contexte

Le déploiement des réseaux de neurones sur des plateformes à ressources contraintes [2, 3, 4], notamment les systèmes embarqués, nécessite de réduire leur coût de calcul, leur consommation énergétique et leur empreinte mémoire, tout en préservant la qualité des prédictions. La quantification constitue une approche particulièrement intéressante pour répondre à ces contraintes. Elle consiste à représenter les poids et les activations d'un réseau de neurones avec des formats numériques de précision réduite, en remplacement de l'arithmétique standard à virgule flottante [1].
Cependant, une quantification uniforme, qui impose le même format numérique à l'ensemble du réseau, ne tient pas compte des différences de comportement et de sensibilité numérique entre les couches et les neurones. Certaines opérations peuvent tolérer une forte réduction de précision, tandis que d'autres nécessitent une représentation plus précise pour préserver la qualité des prédictions. Une approche de quantification mixte consiste à combiner plusieurs formats numériques au sein d'un même réseau : arithmétique à virgule fixe [5] (Fixed-Point, FxP), arithmétique à virgule flottante [1] (Floating-Point, FP) de précision réduite et représentations flottantes compensées (FP/FP). Une telle stratégie permettrait de trouver un compromis entre précision numérique, qualité de classification et coût d'exécution.
Par ailleurs, l'entropie des activations et la détection des neurones inactifs ou « morts » (dead neurons), dont les sorties restent nulles (arithmétique sur matrices creuses pour la compensation) ou quasi constantes, constituent des indicateurs intéressants pour caractériser le comportement d'un réseau. Ces informations pourraient être exploitées afin d'identifier les zones du réseau pouvant tolérer une réduction de précision, voire une simplification de leur représentation numérique.

Objectifs

L'objectif principal du stage est d'étudier et de développer des stratégies de quantification mixte des réseaux de neurones, en combinant différents formats numériques et en exploitant des indicateurs statistiques, notamment l'entropie et la détection des neurones morts, afin d'optimiser le compromis entre précision des calculs, qualité des prédictions et coût de déploiement.

- Étudier et comparer différentes représentations numériques pour les réseaux de neurones : arithmétique à virgule fixe (FxP), arithmétique à virgule flottante (FP) de précision réduite et représentations flottantes compensées. Analyser leurs propriétés, leurs coûts de représentation et leur impact sur la précision des calculs.
- Analyser les distributions des poids et des activations des réseaux de neurones à l'aide de mesures statistiques, notamment l'entropie, afin de caractériser leur diversité numérique et leur sensibilité à la quantification. Étudier également la détection des neurones morts ou faiblement actifs et évaluer leur impact sur la structure et les performances du réseau.

- Concevoir et expérimenter des stratégies de quantification mixte permettant d'affecter automatiquement différents formats numériques aux couches ou aux neurones en fonction des caractéristiques observées. Évaluer les approches proposées sur différents réseaux et jeux de données, en termes de précision numérique, de qualité de classification, de mémoire et de coût de calcul, en comparaison avec des stratégies de quantification uniforme.

Les travaux du/de la stagiaire pourront donner lieu à des publications dans des conférences et revues nationales et internationales, ainsi qu'à des présentations lors des séminaires de l'équipe. La soutenance de stage aura lieu à l'institution d'origine de l'étudiant(e), avec une présence des encadrant(e)s de stage en visioconférence. Les encadrant(e)s américain(e)s délivreront un certificat de réussite de stage détaillé, ainsi que des lettres de recommandation pour la suite des études du/de la stagiaire. Les stages M2 peuvent donner lieu à des propositions de continuation en thèse, sous condition d'adéquation des compétences, des financements disponibles et des règles d'admission de l'École doctorale locale.

Etablissement d'accueil

L'Université du Texas à El Paso (UTEP) est une université publique américaine classée **R1** (à activité de recherche de haute catégorie) par la classification Carnegie, qui distingue les établissements ayant une très forte activité de recherche. Située au Texas, à la frontière entre les États-Unis et le Mexique, elle accueille plus de 26000 étudiants et propose 169 formations de licence, de master et de doctorat. Elle se distingue par son engagement en faveur de l'égalité des chances et de l'accès à un enseignement supérieur de qualité. L'UTEP figure parmi les 5% d'universités américaines les mieux classées en matière de recherche. Son département d'informatique, rattaché au *Miguel A. Loya College of Engineering*, propose des formations en informatique du niveau licence au doctorat, et collabore avec des partenaires académiques et industriels. Ses principaux axes de recherche couvrent le génie logiciel, la cybersécurité, l'intelligence artificielle et les fondements théoriques de l'informatique.

Visa et mobilité

La réalisation du stage à l'Université du Texas à El Paso (UTEP) nécessite l'obtention d'un visa J-1. L'établissement pourra accompagner le/la candidat(e) dans ses démarches administratives, mais les frais liés à la demande de visa et aux formalités associées resteront à sa charge. L'accueil en stage est conditionné à l'obtention du visa requis.

Durée, rémunération et frais de scolarité

Le stage, d'une durée de quatre à six mois, n'est pas rémunéré. Aucun frais de scolarité (tuition fees) ne sera demandé par l'UTEP au stagiaire dans le cadre de ce stage, et une attestation de stage sera délivrée à l'issue de ce dernier cf. ci-dessus). Les encadrant(e)s aideront le/la stagiaire pour toute démarche pour obtenir une bourse de mobilité internationale.

Encadrement

Le (la) stagiaire sera encadré(e) par :

- Hanane Benmaghnia – Post-Doctorante en validation numérique au LIP6, Sorbonne Université – CNRS, future Chargée de recherche à l'UTEP, (`https://hanane-benmaghnia.github.io/`)
- Christoph Lauter – Maître de conférences en informatique (Associate Professor), HDR, à l'Université de l'UTEP (`https://www.christoph-lauter.org/`)

- Martine Ceberio – Professeure en Informatique à l'UTEP (`https://expertise.utep.edu/profiles/mceberio`)

Profil recherché

Le stage s'adresse à des étudiant(e)s (M1 ou M2) en informatique, mathématiques appliquées, électronique ou dans un domaine connexe en premier lieu, mais est ouvert aussi aux étudiant(e)s d'autres disciplines s'ils/elles peuvent démontrer une forte capacité de travail interdisciplinaire. Les compétences recherchées sont :

- Modélisation et formalisation de problèmes à l'aide d'outils mathématiques et informatiques.
- Un intérêt pour les réseaux de neurones et leurs fiabilité.
- Curiosité pour aborder une problématique de recherche dans un contexte international, et ouverture d'esprit pour la découverte de la langue anglaise, utilisée dans un contexte universitaire de recherche.
- Compétences en programmation et en génie logiciel basique.
- Une connaissance de l'arithmétique des ordinateurs est un plus.

La curiosité scientifique, l'autonomie, la rigueur et la capacité à analyser des résultats expérimentaux seront particulièrement appréciées.

Modalités de candidature

Les candidatures doivent être envoyées à Hanane Benmaghnia (hanane.benmaghnia@lip6.fr), Christoph Lauter (cqlauter@utep.edu) et Martine Ceberio (mceberio@utep.edu). Elles doivent inclure :

- un curriculum vitae,
- une lettre de motivation,
- des liens vers des projets personnels, professionnels ou universitaires représentatifs, par exemple des dépôts Git (si disponibles) ou publications antérieures.
- une ou deux personnes référentes à contacter.

References

[1] IEEE Standard for Floating-Point Arithmetic.

[2] Hanane Benmaghnia, Matthieu Martel, and Yassamine Seladji. Code generation for neural networks based on fixed-point arithmetic. *ACM Trans. Embed. Comput. Syst.*, 23(5):68:1–68:28, 2024.

[3] Tobias Habermann, Jonas Kühle, Martin Kumm, and Anastasia Volkova. Hardware-aware quantization for multiplierless neural network controllers. In *IEEE Asia Pacific Conference on Circuit and Systems, APCCAS 2022, Shenzhen, China, November 11-13, 2022*, pages 541–545. IEEE, 2022.

[4] Debasmita Lohar, Clothilde Jeangoudoux, Anastasia Volkova, and Eva Darulova. Sound mixed fixed-point quantization of neural networks. *ACM Trans. Embed. Comput. Syst.*, 22(5s):136:1–136:26, 2023.

[5] Benoit Lopez. *Implémentation optimale de filtres linéaires en arithmétique virgule fixe. (Optimal implementation of linear filters in fixed-point arithmetic).* PhD thesis, Pierre and Marie Curie University, Paris, France, 2014.