Département Intelligence Ambiante et Systèmes Interactifs
Laboratoire Vision et Ingénierie des Contenus
3D & Mobilité

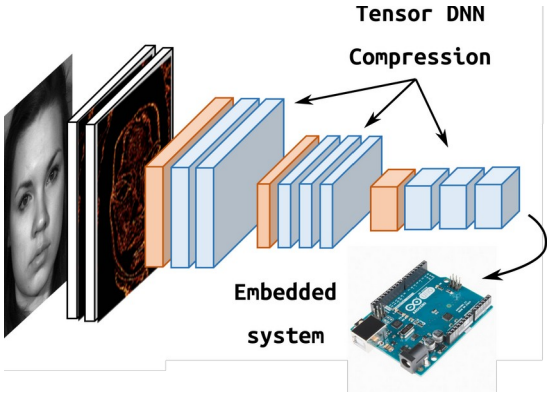
Proposition de stage 2025

Compression des architectures d'IA

Optimisation des modèles d'apprentissage profond via la combinaison de méthodes de compression (tensorielles, quantification, parcimonie)

Contact : mohamed-oumar.ouerfelli@cea.fr

Contexte du stage



Les réseaux de neurones (NN) ont révolutionné notre façon de vivre : les médias sociaux, les systèmes de recommandation, le commerce électronique sont tous basés sur des architectures NN complexes. Cependant, l'immensité des tâches que nous associons aux grandes entreprises mondiales nous fait souvent tenir pour acquis les objets du quotidien. Un smartphone doit régler rapidement les couleurs pour obtenir une image parfaite. Une voiture autonome doit être capable de prédire une collision avec une précision d'une fraction de seconde. Non seulement la plupart des produits de haute technologie utilisent des modèles DL pour être décisifs dans le monde réel, mais ils présentent en plus le défi majeur de devoir être des systèmes rapides, petits, embarqués et portables, avec des capacités de calcul limitées, consacrées à l'efficacité et à la rapidité. Selon les applications, ces systèmes ont la particularité supplémentaire de nécessiter un certain degré de confiance : la décision d'une voiture autonome de freiner devant un mur doit être digne de confiance ! Des modèles DL plus petits et optimisés, plutôt qu'énormes et inefficaces, ont ce potentiel. Ils peuvent être rapidement interprétés et certifiés, en n'ayant à contrôler que les paramètres absolument essentiels.

Objectifs du stage

Le stage permettra d'acquérir une expérience pratique pour tester et développer différentes méthodes (au-delà) de l'état de l'art (SOTA) de compression et d'élagage NN, basées sur l'optimisation du cœur des modèles d'IA via des manipulations tensorielles des architectures DL et des approches de quantification et de parcimonie. Les techniques, inspirées d'approches traditionnelles et nouvelles [3], ouvriront la possibilité d'embarquer des modèles complexes sur des appareils portables.

Le stagiaire explorera les différentes méthodes de compression, et s'intéressera en particulier à leur fusion à travers une approche d'implémentation algorithmique et/ou une approche plus fondamentale via une modélisation mathématique. Les modèles compressés devront être efficaces et conserver le plus possible la précision des architectures d'origine. Dans le cadre du stage, le candidat explorera également de nouveaux outils développés au CEA LIST [4], basés sur l'algèbre tensorielle, capables de tirer parti des performances de compression avec une compréhension plus fondamentale des structures géométriques sous-jacentes. Le développement de ces outils pourrait être suivi d'une thèse ultérieure, au sein du projet HOLIGRAIL du PEPR AI [5], consacrée à la mise en place d'un nouveau SOTA en compression NN et au développement d'outils et de logiciels prêts à être déployés.

[1] https://twitter.com/ylecun/status/1574233818298466304?t=0gqX_98O5YFC32WXssGkhw&s=09
[2] https://github.com/nebuly-ai/exploring-AI-optimization
[3] https://towardsdatascience.com/neural-network-pruning-101-af816aaea61
[4] M. Ouerfelli, M. Tamaazousti, V. Rivasseau. "Random tensor theory for tensor decomposition." In AAAI 2022.
[5] https://www.pepr-ia.fr/

Compétences

Le candidat devra disposer d'une bonne maîtrise de Python et une connaissance de base en réseau de neurones pour lui permettre de développer en autonomie des architectures et tester différentes méthodes de compression. Une forte connaissance de l'algèbre linéaire, *machine learning* et statistique est conseillée (décomposition SVD, *Maximum Likelihood Estimation*, stratégies d'optimisation en DL, etc.).

CEA Tech LIST

Les activités de recherche du CEA Tech LIST sont centrées sur les systèmes à logiciel prépondérant. Ces activités s'articulent autour de trois thématiques: les Systèmes Embarqués (architectures et conception de systèmes, méthodes et outils pour la sûreté des logiciels et des systèmes, systèmes de vision intelligents), les Systèmes Interactifs (ingénierie de la connaissance, robotique, réalité virtuelle et interfaces sensorielles) et les Capteurs et le traitement du signal (instrumentation et métrologie des rayonnements ionisants, capteurs à fibre optique, contrôle non destructif).

Le CEA Tech LIST a de nombreux partenariats avec les grands acteurs industriels du nucléaire, de l'automobile, de l'aéronautique, de la défense et du médical pour étudier et développer des solutions innovantes adaptées à leurs besoins. Il réalise une recherche qui va du concept de système jusqu'au démonstrateur, contribuant au transfert de technologies et à l'innovation par l'émergence de nouvelles entreprises.

Laboratoire Vision pour la Modélisation et la Localisation (LVML)

Laboratoire Vision pour la Modélisation et la Localisation (LVML) du CEA Tech LIST mène des recherches en vision par ordinateur et intelligence artificielle. Nous adressons en particulier les problématiques suivantes :

- Géolocalisation et cartographie d'environnement par vision et fusion de capteurs (robotique mobile, drones…)
- Systèmes et de vision pour la robotique : préhension, manipulation, assemblage d'objets…
- Contrôle de conformité, détection de défauts géométriques, colorimétriques, etc…
- Analyses hyperspectrales : détection de matériaux, tri, ….
- Correction, amélioration d'images et vidéos (superrésolution, upframing, …)
- Compression de réseaux de neurones
- …

Informations générales

Formation / Niveau d'étude	Ingénieur, Master 2 / Bac+5
Possibilité poursuite	Oui, en thèse ou CDD selon profil.
Durée	6 mois
Lieu	Palaiseau (91) – Centre d'intégration de Nano-INNOV
Indemnités de stage	Entre 700 € et 1400 € suivant formation. Aide au logement / transport / restauration.

Département Intelligence Ambiante et Systèmes Interactifs
Laboratoire Vision et Ingénierie des Contenus
3D & Mobilité

Candidatures

- Joindre CV + lettre de motivation à **mohamed-oumar.ouerfelli@cea.fr** avec le nom du stage auquel vous postulez
- Ne pas hésiter à détailler les projets ou cours auxquels vous avez participé
- Indiquer les dates de début/fin de stage envisagées.
- Ce stage pourra prendre une orientation recherche ou industrie en fonction du profil du candidat