UNIVERSITÉ DE TOULON

Contrat doctoral Etablissement 2022/2025
Formulaire de candidature

Titre de la thèse	Apprentissage profond basé sur la conception de modèles efficaces : applications à la surveillance maritime

Ecole Doctorale	ED548
Laboratoire	Laboratoire d'Informatique et Systèmes UMR 7020
Discipline	Informatique
Directeur(s) de Thèse & Encadrant(s)	Thanh Phuong NGUYEN et Yassine ZNIYED

Description du sujet de recherche

Contexte, originalité et pertinence par rapport à l'état de l'art :

Ces dernières années, les réseaux de neurones profonds (DNNs pour Deep Neural Networks) ont considérablement repoussé les limites de l'intelligence artificielle dans un large éventail de tâches, notamment la reconnaissance d'objets à partir d'images, la reconnaissance vocale, la traduction automatique, *etc*. Les réseaux de neurones profonds nécessitent beaucoup de calcul et de mémoire, ce qui les rend difficiles à déployer sur des équipements embarqués avec des ressources de calcul limitées. Ces réseaux profonds sont caractérisés par des millions voire des milliards de paramètres et sont presque exclusivement entraînés en utilisant une ou plusieurs cartes graphiques (GPUs) très rapides et gourmandes en énergie. Considérons un exemple avec le modèle de pointe VGG-16 (Karen Simonyan, 2015), il est constitué de 138,34 millions de paramètres, occupant plus de 500 Mo d'espace de stockage, 15,5 milliards d'opérations de cumul (MAC) et nécessite 30,94 milliards d'opérations en virgule flottante (FLOP) pour classer une seule image. Cela prend plusieurs minutes dans la phase d'inférence sur un appareil mobile ayant une capacité de calcul et des ressources de mémoire limitées. Ces réseaux profonds nécessitent donc énormément de ressources, ce qui les rend difficiles à utiliser et à déployer dans des applications réelles sur des équipements tels que smartphones, tablettes et systèmes embarqués. La compression des modèles de réseaux profonds et la réduction de la consommation d'énergie, tout en préservant les performances prédictives, revêt une importance cruciale pour le déploiement de réseaux profonds dans un tel contexte. C'est pour cela que les tendances récentes se concentrent sur le déploiement d'applications en temps réel, telles que YOLO (Joseph Redmon, 2018) ou sur des ressources limitées (par exemple, MobileNet).

Dans le cadre de cette thèse, nous nous concentrerons sur la compression des réseaux de neurones pour surmonter ce défi en réduisant les besoins en stockage, en consommation d'énergie, et la complexité de calcul dans la phase d'inférence des réseaux de neurones sans que cela n'affecte leur précision. Le but est de déployer les modèles compressés sur des équipements embarqués tels que les caméras intelligentes ou les drones (AUV, ROV, *etc.*). Ces systèmes seront ensuite utilisés pour des tâches de vision par ordinateur, telles que l'analyse de scènes dynamiques (Thanh Tuan Nguyen, 2018), ou la détection/reconnaissance d'objets dans des scènes

maritimes ou sous-marines. Ce projet s'inscrit dans la continuité d'autres projets portés par notre équipe, notamment le projet Rapid DGA UHV-MANTA et le projet ANR Astrid ROV-Chasseur, qui s'intéressent à des applications proches.

Cette thèse se distingue par les aspects suivants. D'abord, d'un point de vue fondamental, différents modèles tensoriels et algorithmes seront étudiés pour la modélisation et la compression des DNNs. Des représentations compactes peuvent être obtenues en recourant à des modèles basés sur des réseaux de tenseurs (RTs) (Cichocki, 2014), qui est un nouveau concept permettant la *''super''-compression*. D'autre part, dans cette thèse, nous proposons une approche visant à réduire conjointement le volume de paramètres, la complexité de calcul et l'énergie consommée en introduisant des opérateurs peu énergétiques (BinaryNet (Xiaofan Lin, 2017), AdderNet (Hanting Chen, 2020), techniques de quantification adaptative, *etc.),* des architectures de réseau de neurones efficaces (convolution séparable en profondeur (Mark Sandler, 2018), mélange des canaux, *etc.*), et en recourant à la suppression des paramètres redondants. De plus, d'un point de vue applicatif, ce projet de thèse sera un des premiers travaux sur la conception et le déploiement des modèles profonds efficaces sur des équipements embaqués dédiés (drone, ROV, *etc*.) pour la surveillance maritime.

Objectifs :

Les objectifs de cette thèse sont :

1. Le développement de nouveaux algorithmes de compression des réseaux de neurones afin de les embarquer sur des appareils mobiles aux ressources limitées en termes de mémoire et de capacité de calcul.
2. La réduction de la consommation énergétique sur ces équipements tout en préservant de bonnes performances. Cela facilitera l'utilisation des DNNs au niveau d'un plus grand nombre d'applications courantes, nécessitant une mise en œuvre sur des équipements embarqués (réseaux de capteurs, *etc*.).
3. La considération des applications en surveillance maritime en déployant des modèles profonds efficaces sur des équipements embarqués dédiés (drone, ROV, *etc*.).

Méthodes :

A cet effet, plusieurs axes de recherche seront donc considérés :

1. *Réduction de la complexité de calcul des modèles profonds*

Plusieurs voies seront considérées. Tout d'abord, *la quantification* sera étudiée afin de réduire les coûts de stockage des paramètres du modèle et d'accélérer les opérations de convolution. Nous considérerons des techniques de quantification adaptative ou bien logarithmique pour obtenir une réduction efficace du volume de poids des paramètres et du coût d'exécution des opérations multiplicatives. Une autre piste consistera à étudier la quantification apprise à partir des valeurs des paramètres. Les méthodes existantes, qui utilisent une fonction de hachage ou un regroupement k-moyennes imposent certaines restrictions quant à la qualité de la quantification. Nous étudierons dans cette thèse plusieurs autres techniques de regroupement afin d'éviter certaines limitations. D'autre part, certaines techniques classiques de compression avec perte seront également exploitées afin de compresser les poids de chacune des couches. La deuxième piste s'intéressera à l'élagage du réseau de neurones pour diminuer sa taille. Ce principe est considéré pour réduire les synapses entre des neurones, supprimer des neurones inutiles, ou bien éliminer des canaux convolutifs ainsi que des couches d'activation inefficaces. Les couches convolutives consomment jusqu'à 90 % d'énergie des DNNs. L'élagage du réseau, qui réduit les structures de réseau (poids, filtre, neurone, par exemple), constitue alors une bonne solution pour réduire le stockage en mémoire et la consommation d'énergie. Notre objectif sera d'étudier des nouvelles métriques afin d'estimer l'importance des poids et/ou des neurones afin de retailler un réseau efficacement en évitant les limitations des métriques existantes (Song Han, 2016) (Yann LeCun J. D., 1989). L'objectif étant la suppression des poids et des neurones non importants, permettant ainsi une réduction significative de la taille du modèle et de la complexité de calcul des DNNs. En outre, déterminer la redondance entre les filtres de chaque couche permet également d'éliminer les filtres inutiles. La suppression d'un filtre peut être efficacement réalisée

en abordant des mesures statistiques (par exemple : corrélation, mesures de similarité) ou la théorie de l'information (par exemple : information mutuelle, entropie) ou encore en appliquant des techniques d'apprentissage automatique (par exemple : ACP - analyse en composants principales, ACI - analyse en composantes indépendantes) sur ces filtres.

2. *Représentations parcimonieuses et de rang faible pour la modélisation des DNNs*

Un noyau de convolution dans des réseaux de neurones est typiquement un *tenseur* d'ordre 4, i.e., un tableau de données à 4 dimensions. Le constat évident est qu'il y a souvent une forte redondance d'information au niveau de ces tenseurs pourtant entièrement caractérisés par leur variables latentes (matrices facteurs). Les décompositions tensorielles (Sebastian Miron, 2020) constituent donc une piste particulièrement prometteuse dans la conception de modèles profonds efficaces, i.e., permettant d'éliminer la redondance dans les noyaux convolutifs et les couches d'activation. Le principe est de trouver des décompositions tensorielles de rang faible, permettant de décomposer une couche convolutive en plusieurs couches plus petites. Bien qu'il y ait plus de couches après la décomposition, le nombre total d'opérations à virgule flottante et de poids sera plus petit. Ce principe permet de rendre le poids du modèle profond plus petit et donc plus efficace en termes de complexité algorithmique et de volume de stockage. Les méthodes existantes dans cette direction reposent souvent sur la décomposition en valeurs singulières, la décomposition canonique polyadique (CP), ou bien encore la décomposition de Tucker. Notre objectif sera d'introduire une méthode efficace de décomposition tensorielle et de l'utiliser pour entraîner et compresser des CNNs.

D'autre part, dans les modèles DNNs modernes, la fonction d'activation ReLU est souvent utilisée comme transformation non linéaire pour construire des couches d'activation en raison de son efficacité de calcul et de sa vitesse de convergence dans la phase d'apprentissage. D'autre part, cela crée une représentation parcimonieuse au niveau de la couche d'activation, toutes les valeurs négatives étant converties en 0. Notre objectif est de proposer une représentation efficace des couches d'activation clairsemées / parcimonieuses. Une piste possible consistera à utiliser une méthode de compression de données sans perte, telle que les algorithmes de codage LZW ou de Huffman, pour réduire efficacement le stockage des mappages d'activation.

3. *Développement d'architectures efficaces de réseau de neurones*

Dans cette approche, nous nous concentrerons tout d'abord sur *la conception d'une architecture efficace* de modèles profonds en nous appuyant sur l'amélioration des concepts intelligents tels que : convolution 1×1, convolution séparable en profondeur, réseau de compression et d'excitation, convolution clairsemée structurée entrelacée, convolution de groupe appris, la recherche des architectures de réseau, *etc*. Ensuite, nous nous intéresserons à concevoir des *réseaux de neurones binaires* ou ternaires qui ont un coût de calcul très faible par rapport à celui des réseaux convolutifs conventionnels. Nous nous concentrerons sur la décomposition d'un filtre convolutif en un ensemble de filtres convolutifs fixes et prédéfinis qui ne seront pas mis à jour pendant le processus d'apprentissage pour réduire efficacement le stockage de poids. Ces filtres ne sont pas forcément binaires comme dans l'algorithme LBCNN pour ne pas affecter les performances du modèle. Nous nous intéresserons également à des techniques de quantification vectorielle (pas seulement la quantification binaire/ternaire comme dans (Fengfu Li, 2016) qui donne la perte de précision) afin de compresser plus efficacement le réseau.

Retombées attendues :

Cette thèse vise à déployer les modèles profonds efficaces sur des équipements embarqués tels que drones ou ROV pour des applications en surveillance maritime. Dans le cadre du projet DGA Rapid Manta, porté par Nadège THIRION-MOREAU, nous nous intéressons à développer un drone intelligent permettant d'éviter automatiquement des obstacles sur la mer. D'autre part, le projet ANR Astrid ROV-Chasseur, porté par Thanh Phuong NGUYEN, s'intéresse à la détection et la reconnaissance des objets spécifiques sous-marins. Ce projet de thèse s'inscrit dans la continuité de ces travaux, en considérant également des applications potentielles en

surveillance maritime, un domaine d'application clé dans les activités de recherche de l'équipe SIIM. Le but de cette thèse est d'avoir des résultats théoriques dans ce cadre, pouvant mener à des progrès dans ce sens.

Mots clés :

Compression des réseaux de neurones, élagage, décompositions tensorielles, surveillance maritime

Références :

Cichocki, A. (2014). Era of Big Data Processing: A New Approach via Tensor Networks and Tensor Decompositions. *arXiv:1403.2048*.

Fengfu Li, B. Z. (2016). Ternary Weight Networks. *arXiv:1605.04711*.

Hanting Chen, Y. W. (2020). : AdderNet: Do We Really Need Multiplications in Deep Learning? . *CVPR*, 1465-1474.

Joseph Redmon, a. A. (2018). YOLOv3: An Incremental Improvement. *arxiv:1804.02767*.

Karen Simonyan, A. Z. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *3rd ICLR.* San Diego, CA, USA.

Mark Sandler, A. H.-C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. *CVPR*.

Sebastian Miron, Y. Z. (2020). Tensor methods for multisensor signal processing. *IET Signal Processing*, 693-709.

Song Han, H. M. (2016). Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. *ICLR*.

Thanh Tuan Nguyen, T. P. (2018). Completed Statistical Adaptive Patterns on Three Orthogonal Planes for Recognition of Dynamic Textures and Scenes. *Journal of Electronic Imaging, 27*, 1-21.

Xiaofan Lin, C. Z. (2017). Towards accurate binary convolutional neural network. *NIPS*.

Yann LeCun, J. D. (1989). Optimal Brain Damage. *NeurIPS Proceedings*.

Encadrement et conditions matérielles pour le doctorant

La thèse se déroulera au sein de l'équipe SIIM au Laboratoire d'Informatique et Systèmes à Toulon.
Le doctorant sera équipé d'un ordinateur portable. Il sera aussi amené à travailler avec des drones et des systèmes embarqués.
Cette thèse sera co-encadrée par :

Thanh Phuong NGUYEN, Maître de conférences (HDR)
Université de Toulon, France
Équipe SIIM, laboratoire LIS
Page web: http://tpnguyen.univ-tln.fr

et

Yassine ZNIYED, Maître de conférences
Université de Toulon, France
Équipe SIIM, laboratoire LIS
Page web: https://yzniyed.blogspot.com/p/about-me.html

Compétences attendues et personnes à contacter

Compétences attendues :

Un candidat autonome et très motivé est sollicité avec un fort intérêt pour le domaine des méthodes mathématiques avancées appliquées au traitement du signal et l'apprentissage automatique. Une formation solide en traitement du signal, mathématiques appliqués, machine learning ou informatique.

Une bonne maîtrise des algorithmes d'apprentissage automatique, notamment les réseaux de neurones.

De bonnes compétences en programmation python sont requises. La connaissance des frameworks d'apprentissage (PyTorch, tensorflow, tensorly, *etc*.) est un plus souhaitable.
Le candidat doit avoir de bonnes capacités en communication orale et écrite.

Personne(s) à contacter :
Thanh Phuong NGUYEN (tpnguyen@univ-tln.fr) et Yassine ZNIYED (zniyed@univ-tln.fr).