LABORATOIRE D'INFORMATIQUE & SYSTÈMES

UNIVERSITÉ DE TOULON

Méthodes tensorielles pour la compression et l'apprentissage des réseaux de neurones profonds et leur application à la surveillance par drones

Mars 2022

1 Introduction

De nos jours, les progrès dans l'apprentissage automatique et surtout dans l'apprentissage profond permettent aux machines de détecter et de reconnaître les objets spécifiques mieux que les êtres humains dans certains domaines. L'apprentissage profond s'appuie généralement sur un volume énorme de données pour apprendre le modèle d'apprentissage. Plus les données sont volumineuses, plus le modèle fonctionne avec précision. Malgré le grand succès de l'apprentissage profond, il reste encore quelques défis à surmonter pour déployer des modèles profonds dans la vie réelle. Entre autres, le déploiement des modèles profonds sur des équipements embarqués ou mobiles ayant des ressources de calcul et de stockage limitées reste un défi majeur. En effet, les réseaux de neurones profonds (DNNs, pour Deep neural networks en anglais) 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 même des milliards de paramètres et sont presque exclusivement entraînés en utilisant une ou plusieurs cartes graphique (GPU) très rapides et gourmandes en énergie.
Dans ce projet, nous proposons d'aborder la problématique de compression et d'apprentissage des DNNs, en utilisant les décompositions tensorielles [1]. Les tenseurs ont reçu une attention particulière dans ce sens en raison de leur capacité de représenter à la fois des données hétégorènes et volumineuses. Dans ce cas, les données peuvent être organisées selon un tableau à D dimensions, aussi appelé tenseur d'ordre D. L'utilisation des tenseurs présente plusieurs avantages par rapport aux matrices, comme l'unicité [2], c'est-à-dire la garantie d'identifiabilité des paramètres récupérés, ou encore la disponibilité d'outils puissants pour effectuer des décompositions de tenseurs. De ce fait, les décompositions tensorielles sont des outils puissants de l'algèbre multilinéaire, qui sont utilisés dans une grande variété d'applications, notamment pour la compression [3, 4] et l'apprentissage [5, 6] des réseaux de neurones.

2 Approches proposées

2.1 Compression des matrices de poids et des noyaux tensoriels

Des études récentes [7, 4] montrent que les matrices de poids des DNNs sont souvent redondantes, et en restreignant leur rangs, il est possible de réduire considérablement le nombre de paramètres sans

baisse significative de performance. Cette observation reste valable pour les noyaux des couches de convolution [3].
Le but de cette thèse est de trouver des approximations tensorielles de rang faible permettant une réduction du nombre de paramètres. Ces paramètres peuvent êtres soient les noyaux de convolution; qui sont naturellement modélisés par des tenseurs d'ordre 4; ou des matrices de poids pour les couches entièrement connectées; que nous proposons de convertir sous format tensoriel. Dans cette thèse, d'abord d'un point de vue fondamental, différents modèles tensoriels et algorithmes seront étudiés pour la modélisation et la compression des tenseurs de poids. Des représentations compactes peuvent être obtenues en recourant à des modèles basés sur des représentations classiques, du type décomposition canonique polyadique (CPD) [8], ou sur les réseaux de tenseurs (RTs) [9], en particulier des modèles de trains de tenseurs (TT) [10] et des modèles Tucker hiérarchiques (TH) [11, 12]. Le principe des RTs est de transformer des tenseurs d'ordre élevé en un ensemble de tenseurs de petites dimensions et d'ordre au plus égale à 3. L'intérêt de cette approche est de faire la "super"-compression [13, 14] des tenseurs de poids en utilisant des approximations de rang faible, avec la possibilité de faire du calcul parallèle [15]. Une propriété intéressante des réseaux de tenseurs est leur capacité à effectuer efficacement des opérations, du type produit matriciel, produit de Hadamard ou produit scalaire, sous le format tensoriel. Plusieurs opérations sont développées dans le cas des trains de tenseurs, par exemple la somme ou le produit entre deux matrices sous format TT [10]. Une fois une couche; de convolution ou entièrement connectée; est remplacée par sa décomposition, l'objectif serait d'adapter les opérations importantes de type convolution, produit matrice-matrice ou matrice-vecteur, au format des décompositions adoptées. Cela permettera de: (i) accélérer le temps d'inférence des DNNs, et (ii) adapter l'algorithme d'apprentissage aux poids tensoriels, pour faire des éventuels ajustements des paramètres.

2.2 Modélisation et apprentissage des réseaux de neurones

Ce deuxième axe de recherche propose d'aller plus loin que la modélisation et la compression des poids. Un lien entre les décompositions tensorielles et la modélisation générale des réseaux de neurones sera étudié. Dans cette partie, le but est de formuler le problème d'apprentissage des DNNs comme un problème de factorisation tensorielle. Ce lien a été investigué par des travaux séminaux récents [5, 6, 16]. Dans [5], le travail est axé sur la mise en relation des décompositions tensorielles avec les réseaux de neurones avec des unités de produit (au lieu d'unités de sommation). Dans [6, 16], l'apprentissage d'un réseau de neurones à une seule couche avec des fonctions d'activation "fléxibles" (FAFs) a été formulé comme un problème de décomposition contraint d'un tenseur CPD. Dans ce dernier travail, la décomposition a permis de compresser des réseaux pré-entrainés en estimant conjointement les poids et les nouvelles fonctions d'activations, dites *fléxibles*. Dans cette thèse, des questions sur la modélisation des NNs seront addressées. Le but est de modéliser les couches des DNNs, et de proposer des nouvelles méthodes d'apprentissage, basées sur les factorisations tensorielles.

3 Application à la surveillance automatique basée sur des modèles profonds dans des drones

Du point de vue des applications, un exemple important est donné par les drones de surveillance automatique utilisant les DNNs. Ces drones représentent une solution de supervision, selon un déploiement alliant homme et machine pour sécuriser automatiquement des grands espaces. Le défi étant de réduire le nombre de paramètres des réseaux pour une implémentation dans des architectures

avec des ressources de calcul limitées. Pratiquement, cette réduction du nombre de paramètres signifie des réseaux plus compacts avec une empreinte mémoire réduite, ce qui peut être important pour les architectures avec une RAM ou une mémoire de stockage limitée. À titre d'exemple, le réseau VGG-19 [17] dispose à peu près de 138 millions de paramètres. Il est donc crucial de développer de nouvelles méthodes pour pallier ce déluge du nombre de paramètres. De plus, l'augmentation du nombre de paramètres induit des temps d'execution élevés, ainsi qu'une consommation d'énergie plus importante. Ce sont des grands défis qui restent ouverts quand les modèles profonds sont déployés sur des équipements embarqués tels que des drones. Il est donc nécessaire de gérer efficacement ce problème, et de développer de nouvelles stratégies adaptées aux systèmes embarqués.
Ce sujet de thèse a un lien étroit avec les travaux, en collaboration avec la DGA, réalisés dans notre équipe. On cite le projet DGA RAPID Manta, porté par Nadège THIRION-MOREAU, où nous nous intéressons à développer un drone intelligent permettant d'éviter automatiquement des obstacles sur la mer. De plus, le projet ANR ASTRID ROV-Chasseur, porté par Thanh Phuong NGUYEN, s'intéresse à la détection et la reconnaissance des objets spécifiques sous-marins (poissons et mines). En effet, ce projet de thèse est 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. Dans le cadre de ce projet, nous proposons de déployer les modèles profonds efficaces sur les équipements embarqués tels que drones, engin sous-marin télécommandé, ROV (Remotely operated underwater vehicle) pour les applications en surveillance maritime.

4 Équipe d'accueil et encadrement

La thèse se déroulera au sein de l'équipe Signal et Image (SIIM) du Laboratoire d'Informatique & Systèmes (LIS) UMR 7020. Cette thèse sera co-dirigée par:

Yassine ZNIYED, Maître de conférences
Université de Toulon, France
Équipe SIIM, laboratoire LIS
email: zniyed@univ-tln.fr
page web: `https://yzniyed.blogspot.com/p/about-me.html`

et

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

5 Conditions et procédure de candidature:

Le candidat doit être un **ressortissant de l'UE, du Royaume Uni ou de la Suisse**. Pour candidater, veuillez envoyer votre CV, relevés de notes avec qualifications et informations pertinentes, avant **le 8 avril 2022**, à Yassine Zniyed (zniyed@univ-tln.fr) et Thanh Phuong NGUYEN (tpnguyen@univ-tln.fr).

Références

[1] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009.

[2] J. B. Kruskal, "Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics," *Linear Algebra and its Applications*, vol. 18, no. 2, pp. 95–138, 1977.

[3] V. Lebedev, Y. Ganin, M. Rakhuba, I. V. Oseledets, and V. S. Lempitsky, "Speeding-up convolutional neural networks using fine-tuned cp-decomposition," in *ICLR, San Diego, CA, USA*, Y. Bengio and Y. LeCun, Eds., 2015.

[4] A. Novikov, D. Podoprikhin, A. Osokin, and D. Vetrov, "Tensorizing neural networks," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, Cambridge, MA, USA, 2015, p. 442–450.

[5] N. Cohen, O. Sharir, and A. Shashua, "On the expressive power of deep learning: A tensor analysis," in *Proceedings of the 29th COLT, New York, USA, June 23-26, 2016*.

[6] Y. Zniyed, K. Usevich, S. Miron, and D. Brie, "Tensor-based framework for training flexible neural networks," *CoRR*, 2021. [Online]. Available: https://arxiv.org/abs/2106.13542

[7] T. N. Sainath, B. Kingsbury, V. Sindhwani, E. Arisoy, and B. Ramabhadran, "Low-rank matrix factorization for deep neural network training with high-dimensional output targets," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6655–6659.

[8] F. L. Hitchcock, "Multiple invariants and generalized rank of a p-way matrix or tensor," *Journal of Mathematics and Physics*, vol. 7, no. 1-4, pp. 39–79, 1928.

[9] A. Cichocki, "Era of big data processing: A new approach via tensor networks and tensor decompositions," *CoRR*, vol. abs/1403.2048, 2014. [Online]. Available: http://arxiv.org/abs/1403.2048

[10] I. V. Oseledets, "Tensor-train decomposition," *SIAM Journal on Scientific Computing*, vol. 33, no. 5, pp. 2295–2317, 2011.

[11] L. Grasedyck, "Hierarchical singular value decomposition of tensors," *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 4, pp. 2029–2054, 2010.

[12] Y. Zniyed, O. Karmouda, R. Boyer, J. Boulanger, A. L. de Almeida, and G. Favier, "Structured tensor train decomposition for speeding up kernel-based learning," in *Tensors for Data Processing*, Y. Liu, Ed. Academic Press, 2022, pp. 537–563.

[13] I. V. Oseledets, "Approximation of $2^d \times 2^d$ matrices using tensor decomposition," *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 4, pp. 2130–2145, 2010.

[14] B. N. Khoromskij, "$O(d \cdot logN)$-quantics approximation of N-d tensors in high-dimensional numerical modeling," *Constructive Approximation*, vol. 34, no. 2, pp. 257–280, 2011.

[15] Y. Zniyed, R. Boyer, A. L. F. de Almeida, and G. Favier, "A TT-based hierarchical framework for decomposing high-order tensors," *SIAM Journal on Scientific Computing*, vol. 42, no. 2, pp. A822–A848, 2020.

[16] Y. Zniyed, K. Usevich, S. Miron, and D. Brie, "Tensor based approach for training flexible neural networks," in *Asilomar Conference on Signals, Systems, and Computers*, 2021.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd ICLR, San Diego, CA, USA, May 7-9*, 2015.