Titre de la thèse : **Décompositions tensorielles pour l'optimisation de réseaux de neurones**

Laboratoire d'accueil : **ICB UMR 6303 CNRS UBE**

Spécialité du doctorat préparé : **Informatique**

Mots-clefs : **Réseaux de neurones, Modélisation tensorielle, Décompositions tensorielles**

Descriptif détaillé de la thèse

1. Résumé

L'optimisation des réseaux de neurones peut se mettre en œuvre à l'aide de diverses méthodes, notamment à l'aide de techniques de compression. La compression des matrices de poids est majoritairement utilisée, mais peu de travaux se concentrent sur l'utilisation de décompositions tensorielles.

L'utilisation des décompositions tensorielles pour compresser les réseaux de neurones nécessite de modéliser un réseau sous forme d'un ou plusieurs tenseurs, de choisir une décomposition adaptée pour minimiser la perte d'efficacité du réseau, et de ré-exprimer les opérations du réseau d'origine dans l'espace compressé produit par la décomposition.

La thèse proposera d'investiguer le potentiel des tenseurs et des décompositions tensorielles pour compresser les réseaux de neurones afin d'optimiser la phase d'inférence.

L'objectif sera de modéliser les réseaux de neurones sous la forme de tenseurs, de sélectionner une décomposition tensorielle adaptée au réseau puis de ré-exprimer au maximum les opérations du réseau dans l'espace compressé afin de maximiser le gain en terme de temps d'exécution et de consommation mémoire lors de la phase d'inférence.

2. Contexte scientifique et état de l'art

Les réseaux de neurones ont prouvé leur efficacité dans de nombreux domaines. Les différentes architectures de modèle permettent de spécialiser ces réseaux sur des tâches spécifiques. Toutefois, le coût calculatoire de l'entraînement et de l'inférence de ces modèles est élevé, du fait notamment de la tendance à augmenter le nombre de paramètres afin d'affiner les résultats du modèle.

Diverses techniques d'optimisation sont fréquemment employées dans ce cadre, pouvant impacter ou non la qualité du résultat produit, telles que la *quantization* [1] ou le *pruning* [2]. La compression des matrices de poids [3] fait également partie de ces techniques d'optimisation. Cependant, le gain obtenu par compression de matrices reste limité par rapport au gain pouvant être obtenu par compression de tenseurs [4].

Les tenseurs sont des objets mathématiques multidimensionnels qui permettent, entre autres, de généraliser la notion de vecteurs et de matrices. Leurs opérateurs de décompositions tensorielles, telles que CANDECOMP/PARAFAC (CP) et Tucker [5], approximent le tenseur d'origine en un ensemble d'éléments moins volumineux (par exemple tenseurs cœurs et matrices de facteurs) qui peuvent être utilisées pour extraire des structures latentes, réduire la dimension ou diminuer le bruit d'un signal.

L'utilisation des décompositions tensorielles à des fins de compression de réseaux de neurones reste cependant limitée [6,7,8]. Les réseaux convolutionnels [9] ont des tenseurs d'ordre 3 ou supérieur clairement apparents, mais la majorité des architectures de réseaux de neurones ne manipulent pas directement des tenseurs d'ordre 3 ou supérieur.

3. Verrous scientifiques

Plusieurs verrous scientifiques structurent ce sujet.

Le premier verrou consiste à modéliser les architectures de réseau de neurones à l'aide de tenseurs. En effet, la plupart des architectures de réseaux de neurones font apparaître naturellement des matrices, mais ne disposent pas de structures d'ordre supérieur. Les structures des réseaux de neurones peuvent toutefois être réarrangées (par exemple en regroupant les matrices de poids de différentes couches dans un tenseur d'ordre 3) afin de produire des tenseurs qui peuvent ensuite être exploités.

Le deuxième verrou consiste à choisir une décomposition tensorielle permettant de limiter au maximum les pertes de performance du modèle. Les différentes décompositions tensorielles produisant des résultats ayant des formats et des propriétés différents, ils ont un impact sur le résultat pouvant varier en fonction de la nature de l'architecture de réseau de neurones. Il convient donc d'identifier les décompositions adaptées en fonction des caractéristiques du réseau.

Le troisième verrou concerne l'expression des opérations d'un réseau de neurones dans l'espace compressé produit par la décomposition. L'espace compressé étant considérablement réduit par rapport au tenseur d'origine, les gains en temps de calcul et en consommation mémoire sont les plus importants lorsque la majorité des opérations peut s'exécuter dans cet espace. Cependant, les décompositions produisant des ensembles d'éléments, il convient de ré-exprimer les opérations natives du réseau de neurones sur cet ensemble d'éléments.

4. Objectifs de la thèse

La thèse a pour objectif d'explorer l'utilisation des décompositions tensorielles pour optimiser les réseaux de neurones en termes de temps de calcul et de consommation mémoire, visant ainsi à réduire les besoins énergétiques de l'utilisation de telles méthodes d'IA.

Une première étape consistera à expérimenter la modélisation des données des réseaux de neurones sous la forme de tenseurs et l'application des différentes décompositions tensorielles, ainsi qu'à étudier les capacités d'expression des opérations du réseau dans l'espace compressé dans des architectures de réseaux pré-définies (telles que les GPT). À la suite de cette étape, les travaux peuvent mener à automatiser la modélisation du réseau et l'application des décompositions tensorielles en fonction de ses caractéristiques, et ce sur tout type d'architecture.

5. Contexte de la thèse

Cette thèse se déroule dans le cadre du projet CAMELIA (Composants pour l'Accélération Matérielle Et Logicielle de l'IA), qui vise à concevoir conjointement le matériel et le logiciel nécessaire à l'entraînement de grands modèles d'IA et à l'inférence de résultat. Pour ce faire, les optimisations seront pensées au niveau du matériel, du logiciel et des algorithmes, afin d'offrir de nouvelles opportunités d'exécution tout en gardant les performances comme préoccupation majeure, que ce soit au niveau énergétique ou temps d'exécution.

Le projet ciblé "Plateforme logicielle de cointégration et pile logicielle applicative" s'intéresse plus particulièrement au développement de l'environnement logiciel nécessaire à l'apprentissage et à l'exploitation de grands modèles d'IA sur les accélérateurs matériels conçus dans le projet CAMELIA. Il vise à développer des solutions de compilation optimisées pour les nouveaux paradigmes de calcul étudiés dans le projet (calcul proche/en mémoire, précision et/ou parcimonie variable au sein des modèles, etc.), à interfacer ces solutions de compilation avec les environnements de développement existants (PyTorch, TensorFlow, etc.) et à faciliter la prise en main de tels outils par les ingénieurs et chercheurs en IA afin de faciliter le prototypage de nouvelles idées.

Références bibliographiques

[1] Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M. W., & Keutzer, K. (2022). A survey of quantization methods for efficient neural network inference. In *Low-power computer vision* (pp. 291-326). Chapman and Hall/CRC.

[2] Cheng, H., Zhang, M., & Shi, J. Q. (2024). A survey on deep neural network pruning: Taxonomy,

comparison, analysis, and recommendations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(12), 10558-10578.

[3] Wang, X., Zheng, Y., Wan, Z., & Zhang, M. SVD-LLM: Truncation-aware Singular Value Decomposition for Large Language Model Compression. In *The Thirteenth International Conference on Learning Representations*.

[4] Novikov, A., Podoprikhin, D., Osokin, A., & Vetrov, D. P. (2015). Tensorizing neural networks. *Advances in neural information processing systems*, *28*.

[5] Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review*, *51*(3), 455-500.

[6] Yang, Y., Zhou, J., Wong, N., & Zhang, Z. (2024, June). Loretta: Low-rank economic tensor-train adaptation for ultra-low-parameter fine-tuning of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 3161-3176).

[7] Gu, Y., Zhou, W., Iacovides, G., & Mandic, D. (2025, June). TensorLLM: Tensorising Multi-Head Attention for Enhanced Reasoning and Compression in LLMs. In *2025 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

[8] Li, Y., Guo, Z., Yin, M., & Li, B. LeSTD: LLM Compression via Learning-based Sparse Tensor Decomposition. In *The Fourteenth International Conference on Learning Representations*.

[9] Kalle, A., Rudkiewicz, T., Ouerfelli, M. O., & Tamaazousti, M. (2025). Distribution-Aware Tensor Decomposition for Compression of Convolutional Neural Networks. *arXiv preprint arXiv:2511.04494*.

Profil demandé

Le ou la candidat(e) devra posséder une solide formation en informatique, mathématiques appliquées. Une bonne maîtrise de l'algorithmique, des langages de programmation, de l'algèbre linéaire, du machine learning et deep learning est attendue.

Des compétences en programmation scientifique, notamment en C et Python, seront attendues, ainsi qu'une capacité à conduire un travail mêlant théorie, développement algorithmique et validation expérimentale.

Financement : Projet Camelia

Dossier à envoyer pour le 25 mai 2026
Début du contrat : 1er Octobre 2026
Salaire mensuel brut : 2300€

Direction de la thèse

LECLERCQ Eric (Eric.Leclercq@ube.fr) 40%

Encadrement de la thèse : co-directeur(s) et co-encadrant(s)

GINHAC Dominique (dominique.ginhac@ube.fr**) – Co-directeur 20 %**

GILLET Annabelle (annabelle.gillet@ube.fr**) – Co-encadrante 40 %**

Les candidates et candidats souhaitant postuler à cette offre de thèse sont invités à transmettre un dossier complet comprenant les documents suivants :

- un **curriculum vitae** détaillé, mettant en évidence le parcours académique, les enseignements suivis, les éventuelles expériences de recherche, les compétences techniques, ainsi que les publications, le cas échéant ;
- une **copie des relevés de notes et diplômes**, incluant notamment les résultats obtenus en Master (ou formation équivalente) ;
- une **lettre de motivation** de 1 à 2 pages, exposant l'intérêt du candidat ou de la candidate pour le

sujet de thèse, les compétences ou expériences en lien avec le projet, ainsi que ses objectifs scientifiques et professionnels ;

- une ou plusieurs **lettres de recommandation** rédigées par des référents académiques ou professionnels connaissant le travail du candidat ou de la candidate.

Les dossiers de candidature devront être envoyés par courrier électronique à E. Leclercq à l'adresse suivante : eric.leclercq@ube.fr avec copies à D. Ginhac (dominique.ginhac@ube.fr) et A. Gillet (annabelle.gillet@ube.fr) avant le 25 Mai 2026. Les dossiers incomplets ou transmis après la date limite ne seront pas examinés.

Après réception des candidatures, une phase de **pré-sélection** sera menée afin d'identifier les profils les plus prometteurs au regard de l'excellence académique, de la motivation et de l'adéquation avec les objectifs de la thèse. Les candidates et candidats retenus à l'issue de cette première étape seront contactés pour une **présentation orale suivie d'un entretien avec l'équipe encadrante.** Lors de cette audition, les candidates et candidats seront invités à présenter leur parcours, à expliciter leur motivation pour ce projet de recherche, ainsi qu'à discuter de leur compréhension du sujet et des contributions qu'ils pourraient y apporter.

Thesis title: **Tensor decompositions for optimizing neural networks**

Host Laboratory: **ICB UMR 6303 CNRS UBE**

Doctoral Specialty: **Computer Science**

Keywords: **Neural networks, Tensor modeling, Tensor decompositions**

Detailed Thesis Description

1. Summary

Neural networks optimization can rely on different methods, including compression techniques. Weight matrices compression is mainly used in this context, and few works focus on tensor decompositions.

Using tensor decomposition to compress neural networks require to model a network as one or several tensors, to choose an adapted tensor decomposition to minimize efficiency loss of the network, and to formulate original network's operations in the compressed space produced by the decomposition.

This thesis aims at investigating the potential of tensors and tensor decompositions to compress neural networks in order to optimize inference.

The objectives are to model neural networks as tensors, to select an adapted tensor decomposition depending on the network characteristics, and to formulate network operations in the compressed space in order to reduce execution time and memory consumption during inference.

2. Scientific context and state of the art

Neural network have proven their efficiency in several fields. The various model architectures allow to specialized these networks on specific tasks. However, the computational cost of training and inference of these models is high, especially with the trend consisting in increasing the number of parameters to improve model's results.

Different optimization techniques are often used in this context, that can impact the quality of the result, such as quantization [1] or pruning [2]. Weight matrices compression [3] is another optimization technique, however the benefit remains limited compared to the benefit that can be obtained with tensor compression [4].

Tensors are multi-dimensional mathematical objects allowing to generalize the notion of vectors and matrices. Tensor decomposition operators, such as CANDECOMP/PARAFAC (CP) and Tucker [5], approximate the input tensor as a set of smaller structures (for example core tensors and factor matrices), that can be used to extract latent structures, reduce dimension or reduce noise.

The use of tensor decompositions to compress neural networks remains limited [6,7,8]. Convolutional networks [9] have 3-order tensors or more, but most of neural network architectures do not directly manipulate high-order tensors.

3. Scientific Problems

Several scientific problems structure this subject.

The first problem consists in modeling neural network architectures with tensors. Indeed, most of neural networks are composed of matrices, but do not present higher order structures. Neural networks' structures can be rearranged (for example by stacking weight matrices of different layers in a 3-order tensor) in order to model the network with tensors.

The second problem consists in selecting a tensor decomposition that will not induce a loss of performance of the neural network. As the different tensor decompositions produce results having different characteristics and format, their impact on the network can vary. It is thus essential to identify adapted

decompositions depending on the network characteristics.

The third problem consists in formulating the network operations in the compressed space produced by the decomposition. The compressed space being greatly reduced compared to the original tensor, the profit on execution time and memory consumption will be greater if most of the operations are performed in the compressed space. However, as tensor decompositions produce a set of structures, operations that were executed on a single tensor in the original network must be formulated on this set.

4. Thesis Objectives

The thesis aims at exploring the use of tensor decomposition to reduce execution time and memory consumption of neural networks, in order to reduce energetic needs of such AI techniques.

A first step consists in experimenting the modeling of neural networks as tensors and the application of different tensor decompositions on this modeling, as well as studying expressivity capabilities of network's operations in the compressed space. This study will be carried on predefined neural network architectures (such as GPT). These works can lead to the automatization of the application of the best modeling and tensor decomposition on any neural network architecture depending on the network characteristics.

5. Thesis Context

This thesis is part of the CAMELIA project, that aims at conceiving conjointly hardware and software required for the training of large AI models and for the inference. To do so, optimizations are defined on hardware, software and algorithms, in order to offer new execution opportunities while having performances as main preoccupation, regarding energetic consumption and execution time.

One of the targeted project of CAMELIA focuses on developing the software environment necessary to train and use large AI models on hardware accelerators conceived in other targeted projects of CAMELIA. It aims at developing optimized compilation solutions for new computing paradigms studied in the project (precision, sparsity, etc.), to interface these solutions with existing development environment (PyTorch, TensorFlow, etc.) and to facilitate the adoption of these tools by AI engineers and researchers to ease the prototyping of new ideas.

Bibliographic References

[1] Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M. W., & Keutzer, K. (2022). A survey of quantization methods for efficient neural network inference. In *Low-power computer vision* (pp. 291-326). Chapman and Hall/CRC.

[2] Cheng, H., Zhang, M., & Shi, J. Q. (2024). A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(12), 10558-10578.

[3] Wang, X., Zheng, Y., Wan, Z., & Zhang, M. SVD-LLM: Truncation-aware Singular Value Decomposition for Large Language Model Compression. In *The Thirteenth International Conference on Learning Representations*.

[4] Novikov, A., Podoprikhin, D., Osokin, A., & Vetrov, D. P. (2015). Tensorizing neural networks. *Advances in neural information processing systems*, *28*.

[5] Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review*, *51*(3), 455-500.

[6] Yang, Y., Zhou, J., Wong, N., & Zhang, Z. (2024, June). Loretta: Low-rank economic tensor-train adaptation for ultra-low-parameter fine-tuning of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 3161-3176).

[7] Gu, Y., Zhou, W., Iacovides, G., & Mandic, D. (2025, June). TensorLLM: Tensorising Multi-Head Attention for Enhanced Reasoning and Compression in LLMs. In *2025 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

[8] Li, Y., Guo, Z., Yin, M., & Li, B. LeSTD: LLM Compression via Learning-based Sparse Tensor Decomposition. In *The Fourteenth International Conference on Learning Representations*.

[9] Kalle, A., Rudkiewicz, T., Ouerfelli, M. O., & Tamaazousti, M. (2025). Distribution-Aware Tensor Decomposition for Compression of Convolutional Neural Networks. *arXiv preprint arXiv:2511.04494*.

Required Profile

The candidate must have a strong background in computer science or applied mathematics. Proficiency in algorithmic, programming languages, linear algebra, machine and deep learning is expected.

Skills in scientific programming, especially with C and Python, are expected, along with the ability to carry out work combining theory, algorithmic development and experimental validation.

Funding: Camelia project

Application deadline: 25 May 2026
Contract start date: 1st October 2026
Monthly gross salary: 2300€

Thesis Supervisor

LECLERCQ Eric (Eric.Leclercq@ube.fr) 40%

Thesis Co-supervisors

GINHAC Dominique (dominique.ginhac@ube.fr**) 20 %**

GILLET Annabelle (annabelle.gillet@ube.fr**) 40 %**

Candidates wishing to apply for this PhD position are invited to submit a complete application including the following documents:

- a **curriculum vitae** detailing academic background, courses taken, any research experience, technical skills, and publications if applicable;
- a **copy of transcripts and diplomas**, including Master degree results (or equivalent);
- a **cover letter** of 1 to 2 pages outlining the candidate's interest in the thesis topic, relevant skills or experience, as well as scientific and professional objectives;
- **recommendation letters** written by academic or professional references familiar with the candidate's work.

Applications must be sent by email to E. Leclercq at: eric.leclercq@ube.fr with copies to D. Ginhac (dominique.ginhac@ube.fr) and A. Gillet (annabelle.gillet@ube.fr) before 25 May 2026. Incomplete applications or those received after the deadline will not be considered.

After receipt of applications, a phase of **pre-selection** will be conducted based on academic excellence, motivation, and alignment with the thesis objectives. Candidates selected at this stage will be contacted for an **oral presentation followed by an interview with the supervision team**. During this interview, candidates will be invited to present their background, clarify their motivation, discuss their understanding of the topic, and outline their potential contributions.