## ETIENNE Dupuis

# Titres

* Weight-Sharing Methods for Retraining-Free CNN Compression
* Méthode de partage de poids pour compresser des réseaux de neurones profonds sans réentrainement

# Mots Clés

* Convolutional Neural Network, Deep Learning, Computer vision, Design Space Exploration, Approximate Computing, Weight-Sharing
* Réseaux de neuronnes convolutionels, Apprentissage profond, Vision par ordinateur, exploration d’espace de conception, calcul approché, partage de poids

# English Abstract

The outstanding performance achieved by Convolutional Neural Network (CNN) comes at the cost of extremely high computational requirements, making them out of reach for most low-power embedded devices [1]. Most of the energy cost during CNN inference comes from memory access as analyzed in [1]. This work focuses on reducing the memory footprint of CNNs to improve energy efficiency. The Approximate Computing (AxC) paradigm leverages the inherent error-resilience of CNNs to improve energy efficiency by relaxing the need for fully accurate operations. CNNs have a high degree of redundancy in terms of their structure and parameters [2], and this redundancy is not always necessary for an accurate prediction. This observation has paved the way for several highly recognized approximation techniques [1] such as pruning, quantization, low-rank factorization, and Weight Sharing (WS) [2]. WS aims to group weights into buckets or clusters sharing the same value. It allows a significant reduction of the CNN memory footprint by storing shared values in a dedicated data structure, where original weight values in the weight matrix are replaced by their corresponding indexes, represented with fewer bits. To make an example, by reducing the number of different shared values (k) to 256, the indexes can be encoded with only log2(k) = 8 bits, allowing a 4 Compression Rate (CR) compared to the original 32-bit counterpart. Shared weight values can be determined by using a clustering algorithm like the K-means.

WS, as well as the others approximation techniques, came at the cost of a certain accuracy loss. Although the standard application of WS [2]–[5], requires the retraining of the network to recover accuracy loss, it has been proven [6] that it is also possible to optimize the number of shared values to each layer’s resilience while avoiding the costly retraining step.

Exhaustively exploring every combination results in O(|krange|N ) complexity, with krange being the set of the possible number of shared values and N the number of layers. As an example, a toy CNN with N = 5 layers and a krange = [1, 256], results in a number of scoring steps equal to 256⁵ = 1.1x10¹². Thus, the exploration of the complete solution space would require more than three decades when considering an optimistic 1 ms evaluation step. To make things worse, the complexity rises exponentially with N , which is more than a factor of ten higher in recent CNNs.

In this thesis, we propose a heuristic approach to achieve a scalable retrainingfree WS compression. The main contributions can be summarized as (1) The extensive study of the complexity of the weight-sharing optimization; (2) A novel automatic two-step heuristic optimization to retraining-free weight-sharing.

The proposed method for CNN compression can efficiently explore the large design space to produce a set of solutions that offer very interesting tradeoffs between AL and CR. These solutions achieve more than a 5 compression over the baseline memory footprint in multiple state-of-the-art computer vision CNNs on the challenging Imagenet dataset under MLPerf [7] quality target

constraints. Importantly, these compression results are achieved while avoiding the prohibitively costly retraining step, which is commonly used in prior works. The thesis dissemination consists of 4 publications in international conferences with proceedings, 1 journal Elsevier, 1 chapter of a book edited by Springer, and 4 poster dissemination in summits or workshops. The early works using greedy exploration on LeNet/MNIST were published on [8], a generalization to the Imagenet class CNNs with the use of a proxy metrics to accelerate exploration were published in [9], an extension adding a multi-objective exploration was published in [6], the new divide & conquer approach results on MNIST were published in [10] and the generalization to Imagenet is published

in [11]. Finally, a survey on AxC for deep learning was published in [12].

# French Abstract

Les performances exceptionnelles atteintes par les Convolutional Neural Network (CNN) se font au prix d’exigences de calcul extrêmement élevées, les rendant hors de portée de la plupart des dispositifs embarqués à faible puissance [1]. La majeure partie du coût énergétique pendant l’inférence CNN provient de l’accès à la mémoire, comme analysé dans [1]. Ce travail se concentre sur la réduction de l’empreinte mémoire des CNNs pour améliorer l’efficacité

énergétique. Le paradigme Approximate Computing (AxC) tire parti de la résistance aux erreurs inhérente aux CNNs pour améliorer l’efficacité énergétique en assouplissant le besoin d’opérations totalement précises. Les CNNs ont un haut degré de redondance en termes de structure et de paramètres [2], et cette redondance n’est pas toujours nécessaire pour une prédiction précise. Cette observation a ouvert la voie à plusieurs techniques d’approximation très reconnues [1] telles que l’élagage, la quantification, la factorisation à faible rang et la Weight Sharing (WS) [2]. WS vise à regrouper les poids dans des buckets ou clusters partageant la même valeur. Il permet une réduction significative de l’empreinte mémoire de CNN en stockant les valeurs partagées dans une structure de données dédiée, ou` les valeurs de poids originales dans la matrice de poids sont remplacées par leurs index correspondants, représentés avec moins de bits. Pour prendre un exemple, en réduisant le nombre de valeurs partagées différentes (k) à 256, les index peuvent être codés avec seulement log2(k) = 8 bits, permettant un ratio de compression CR de 4 par rapport à l’originale avec les valeurs sur 32 bits. Les valeurs de poids partagées peuvent être déterminées à l’aide d’un algorithme de regroupement comme les K-means. Le WS, ainsi que les autres techniques d’approximation, se font au prix d’une certaine perte de précision. Bien que l’application standard de WS [2]– [5], nécessite le ré-entrainement du réseau pour récupérer la perte de précision, il a été prouvé [6] qu’il est également possible d’optimiser le nombre de valeurs partagées à la résilience de chaque couche tout en évitant l’étape coûteuse du ré-entrainement.

L’exploration exhaustive de chaque combinaison entraîne une complexité de O(|krange|N ), avec |krange| étant l’ensemble du nombre possible de valeurs partagées et N le nombre de couches. A` titre d’exemple, un CNN jouet avec N = 5 couches et un krange = [1, 256], résulte en un nombre d’étapes de notation égal à 256⁵ = 1.1x10¹². Ainsi, l’exploration de l’espace complet des solutions nécessiterait plus de trois décennies en considérant un pas d’évaluation optimiste de 1 ms. Pour aggraver les choses, la complexité augmente exponentiellement avec N , qui est plus d’un facteur dix dans les CNN récents. Dans cette thèse, nous proposons une approche heuristique pour réaliser une compression WS sans ré-entraînement. Les principales contributions peuvent être résumées comme suit : (1) L’étude approfondie de la complexité de l’optimisation du partage du poids ; (2) Une nouvelle optimisation heuristique automatique en deux étapes pour le partage du poids sans ré-entrainement.

La méthode proposée pour la compression CNN peut explorer efficacement le vaste espace de conception pour produire un ensemble de solutions qui offrent des compromis très intéressants entre AL et CR. Ces solutions permettent d’obtenir une compression de plus de 5 par rapport à l’empreinte mémoire de base dans plusieurs CNN de vision par ordinateur de pointe sur le jeu de données Imagenet sous des contraintes de qualité MLPerf [7]. Il est important de noter que ces résultats de compression sont obtenus tout en évitant l’étape de réapprentissage prohibitive et coûteuse, qui est couramment utilisée dans les travaux antérieurs.

La diffusion de la thèse consiste en 4 publications dans des conférences internationales avec actes, 1 journal ACM, 1 chapitre d’un livre, et 4 posters diffusés dans des sommets ou des ateliers. Les premiers travaux utilisant l’exploration sur LeNet/MNIST ont été publiés sur [8], une généralisation à la classe Imagenet CNNs avec l’utilisation d’une métrique proxy pour accélérer l’exploration ont été publiés dans [9], une extension ajoutant une exploration multi-objectif a

été publiée dans citeDUPUIS2021FastWS, les résultats de la nouvelle approche ”diviser pour régner” sur MNIST ont été publiés dans citeDupuis2021FastExplo et la généralisation à Imagenet est publiée dans citeDupuis2022HeuristicExplo. Enfin, une enquête sur AxC pour l’apprentissage profond a été publiée dans [12].