Quantified DNN learning algorithms with limited hardware overhead for Edge implementation

Period

Oct 2022 – Oct 2025

Director

(Denis Trystram : Denis.Trystram@imag.fr)

Sylvain Bouveret : sylvain.bouveret@imag.fr

CEA supervisor

Thomas Mesquida (LSTA) : thomas.mesquida@cea.fr

Host laboratory in CEA

DRT/LIST/DSCIN/LSTA – Lab for Systems and Advanced Technologies

The CEA Tech is a world-class research institute. The teams of engineers and researchers are motivated to develop a broad portfolio of advanced technologies that can be transferred to industrial partners. The research activities span communication technologies, energy and health.

At the center of the CEA Tech, the center for “technology research”, is the LIST institute which leads the activities in the realm of digital and intelligent systems including R&D programs for advanced manufacturing, embedded systems and artificial intelligence. We provide technology solutions for our partners in the transportation, energy, health, security and defense sectors, allowing them to improve the overall competiveness.

Within the LIST institute, is the DSCIN (Digital Systems and Circuits for Circuits) department whose mission is to create technologies for:

- Embedded digital compute systems
- Integrated components and systems for intelligent objects and wireless communications

and to offer a link between technology and applications based on a design platform for embedded systems, especially for IoT applications, autonomous vehicles, artificial intelligence (AI) and high performance computing (HPC).

Within the DSCIN department, the LSTA (Lab for Systems and Advanced Technologies) has the mission is to study, design and implement architectures for multi-core systems and accelerators The team works with advanced CMOS processes (such as 7nm), advanced 2.5D/3D integration technologies and non-volatile memories. Applications span from HPC, AI to control circuits for quantum computers.

Context

Intelligence at the Edge aims to push the computation performed on the data to the edge for energy and security reasons. This results in the implementation of co-optimized hardware for the inference of artificial neural networks (ANN) of varying depths (DNN) and aims at a calculation as close as possible to the creation of useful information. The supported networks are trained offline and their parameters exported to the support.

Two main aspects are necessary for this hardware to adapt to a particular/peculiar environment or to refine its knowledge: the direction of learning, or how to define the target to reach, and the optimization of the network parameters, or how to minimize the error with respect to the defined target. The most commonly used learning algorithms have the drawback of requiring a much larger amount of memory than during the inference phases. Indeed, all the intermediate results of the network must be stored so that the gradient back-propagation can be performed. The extra cost of learning compared to pure inference is consequent and the goal of this PhD is to minimize it in the framework of quantized artificial neural networks.

Some methods have progressively proposed to move away from these algorithms in order to minimize the overhead. [1] proposes to replace the transpositions of the weight matrices in the gradient back-propagation phase by random matrices generated at the beginning of the training. [2] pushes the idea further by proposing a direct backpropagation of the error to each parameter of the network, eliminating the sequential aspect of this phase. [3] uses the same principle starting from the target and not from the error. The update can then be done during the propagation phase and not afterwards, thus considerably reducing the memory requirements. [4-7] also propose methods that allow updating during the propagation phase, based on similarity and/or prediction criteria at different granularities in the network.

Work description

The goal of this PhD is to propose, implement and validate DNN learning algorithms by optimizing the associated memory and energy requirements. In addition to this, there is the constraint of strong quantification of DNN parameters for frugal inference, which is not yet taken into account in the literature in this context. These algorithmic studies can be integrated into the laboratory's hardware simulation platforms. An ENSIMAG profile with a strong AI background is preferred.

Expected contributions are:

- Critical analysis of existing algorithms (3 months)
- Implementation of tools for material impact estimation (3 months)
- Proposal and implementation of alternative algorithms (18 months)
    - Empirical validation of the algorithms and proof of convergence
    - Evaluation in the framework of strong parameter quantification
- Validation of the proposed system
    - (Optional) FPGA proof of concept (3 months)
    - (Optional) Integration to the SystemC flexible DNN inference model (3 months)
- Valorization of the work (scientific publications in conferences and journals, patents)
- Writing and defense of the thesis (6 months)

References

[1] T. P. Lillicrap et al., “Random synaptic feedback weights support error backpropagation for deep learning,” 2016.

[2] A. Nøkland, “Direct feedback alignment provides learning in deep neural networks,” 2016.

[3] C. Frenkel et al., “Learning without feedback: direct random target projection as a feedback-alignment algorithm with layerwise feedforward training” 2019.

[4] W. D. Kurt Ma et al., “The HSIC Bottleneck: Deep Learning without Back-Propagation” 2019.

[5]A. Nøkland et al., “Training Neural Networks with Local Error Signals” 2019.

[6] Belilovsky et al., “Greedy Layerwise Learning Can Scale to ImageNet” 2019.

[7] Lee et al., “Dynamic Block-Wise Local Learning Algorithm for Efficient Neural Network Training” 2021.

Algorithmes d'apprentissage DNN quantifiés à surcoût matériel limité en vue d'une implémentation Edge

Période

Oct 2022 – Oct 2025

Directeur

(Denis Trystram : Denis.Trystram@imag.fr)

Sylvain Bouveret : sylvain.bouveret@imag.fr

Encadrant CEA

Thomas Mesquida (LSTA) : thomas.mesquida@cea.fr

Laboratoire d'accueil au CEA

DRT/LIST/DSCIN/LSTA – Laboratoire Systèmes-sur-puce et Technologies Avancées

CEA Tech est un leader mondial de la recherche technologique. Les équipes d'ingénieurs chercheurs sont mobilisées pour bâtir et transférer à des partenaires industriels des portefeuilles de technologies répondant aux besoins des filières technologiques dans les domaines de l'information, de la communication, de l'énergie et de la santé.

Au sein de CEA Tech, le pôle « recherche technologique » du CEA, l'Institut LIST dédie ses activités aux systèmes numériques intelligents avec des programmes de R&D dans le manufacturing avancé, les systèmes embarqués, et l'intelligence ambiante. Nous accompagnons nos partenaires dans les domaines des transports, de l'industrie, de l'énergie, de la santé, de la sécurité et de la défense, pour transférer les technologies issues de l'innovation et améliorer leur compétitivité.

Intégré au LIST, le Département des Systèmes et Circuits Intégrés Numériques (DSCIN), a pour mission de créer des technologies :

- De systèmes numériques de calcul intégrés ou embarqués
- De composants intégrés et systèmes d'objets intelligents et communicants sans fil

et de proposer une offre assurant le lien entre technologie et applications, basée sur les plateformes de Conception et Systèmes embarqués, en particulier sur les domaines de l'Internet des Objets, des véhicules autonomes, de l'intelligence artificielle et du calcul à haute performance (HPC).

Au sein de ce département, le Laboratoire Systèmes-sur-puce et Technologies Avancées (LSTA) a pour mission d'étudier, concevoir et implémenter des architectures multi-cœurs et des accélérateurs haute performance. Il exploite pour cela les dernières technologies avancées disponibles : CMOS jusqu'au nœud 7nm, intégration 2.5D/3D, mémoires non-volatiles. Les domaines applicatifs sont ceux du calcul haute performance (HPC – High Performance Computing), de l'intelligence artificielle (IA) et du quantique (contrôle numérique de circuits quantiques CMOS).

Contexte

L'intelligence à l'Edge vise à pousser le calcul effectué sur les données en périphérie pour des raisons énergétiques et de sécurité. En découle l'implémentation de matériel co-optimisé pour l'inférence de réseaux de neurones artificiels (ANN) plus ou moins profonds (DNN) et qui vise un calcul au plus près de la création de l'information utile. Les réseaux supportés sont entrainés offline et leurs paramètres exportés vers le support.

Deux aspects principaux sont nécessaires pour que ce matériel s'adapte à un environnement particulier ou affine ses connaissances : la direction de l'apprentissage, ou comment définir la cible à atteindre, et l'optimisation des paramètres du réseau, ou comment minimiser l'erreur vis-à-vis de la cible définie. Les algorithmes d'apprentissage les plus utilisés ont pour défaut de nécessiter une quantité de mémoire bien plus importante que lors des phases d'inférence. En effet, l'ensemble des résultats intermédiaires du réseau doivent être stockées pour que la rétro-propagation du gradient puisse être effectuée. Le surcout lié à l'apprentissage vis-à-vis de l'inférence pure est conséquent et le but de ce PhD est de le minimiser dans le cadre de réseaux de neurones artificiels quantifiés.

Certaines méthodes ont progressivement proposé de s'éloigner de ces algorithmes pour en minimiser les surcouts. [1] propose de remplacer les transposées de matrices de poids dans la phase de rétropropagation du gradient par des matrices aléatoires générées en début d'apprentissage. [2] pousse l'idée plus loin en proposant une rétropropagation directe de l'erreur vers chaque paramètre du réseau, éliminant l'aspect séquentiel de cette phase. [3] reprends le même principe à partir de la cible et non de l'erreur. La mise à jour peut alors se faire durant la phase de propagation et non après, allégeant considérablement les besoins mémoire qui lui sont liés. [4-7] proposent eux aussi des méthodes permettant la mise à jour en phase de propagation, se basant sur des critères de similarité et/ou de prédiction à différentes granularités dans le réseau.

Description du sujet

Le but de ce PhD est de proposer, implémenter et valider des algorithmes d'apprentissage DNN en optimisant les besoins mémoires et énergétiques associés. A cela s'ajoute la contrainte de quantification forte des paramètres du DNN pour une inférence frugale qui n'est pas encore prise en compte dans la littérature dans ce contexte. Ces études algorithmiques pourront être intégrées aux plateformes de simulation matérielles du laboratoire. Un profil type ENSIMAG avec une bonne base IA est recherché.

Les contributions attendues sont :

- Analyse critique des algorithmes existants (3 mois)
- Implémentation d'outils pour estimation des impacts matériels (3 mois)
- Proposition et implémentation d'algorithmes alternatifs (18 mois)
    - Validation empirique des algorithmes et preuve de convergence
    - Evaluation dans le cadre de quantification forte des paramètres
- Validation du système proposé
    - (Optionnel) Preuve de concept FPGA (3 mois)
    - (Optionnel) Intégration au modèle SystemC d'inférence DNN flexible (3 mois)
- Valorisation des travaux (publications scientifiques dans des conférences et journaux, brevets)
- Rédaction et soutenance de la thèse (6 mois)

Références

[1] T. P. Lillicrap et al., “Random synaptic feedback weights support error backpropagation for deep learning,” 2016.

[2] A. Nøkland, “Direct feedback alignment provides learning in deep neural networks,” 2016.

[3] C. Frenkel et al., “Learning without feedback: direct random target projection as a feedback-alignment algorithm with layerwise feedforward training” 2019.

[4] W. D. Kurt Ma et al., “The HSIC Bottleneck: Deep Learning without Back-Propagation” 2019.

[5]A. Nøkland et al., “Training Neural Networks with Local Error Signals” 2019.

[6] Belilovsky et al., “Greedy Layerwise Learning Can Scale to ImageNet” 2019.

[7] Lee et al., “Dynamic Block-Wise Local Learning Algorithm for Efficient Neural Network Training” 2021.