論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Improving Deep Learning Efficiency Using Reservoir Computing Inspiration
著者(和文)	LOPEZ GARCIA-ARIAS ANGEL
Author(English)	Ángel López García-Arias
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第12710号, 授与年月日:2024年3月26日, 学位の種別:課程博士, 審査員:本村 真人,一色 剛,髙橋 篤司,佐々木 広,原 祐子,藤木 大地
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第12710号, Conferred date:2024/3/26, Degree Type:Course doctor, Examiner:,,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	情報通信 系 情報通信 コース	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of （ 工学 ）
学生氏名： Student's Name	LOPEZ GARCIA-ARIAS ANGEL	審査員主査： Chief Examiner	本村 真人

要旨（英文 800 語程度）
Thesis Summary (approx.800 English Words)

Forefront deep learning (DL) models involve massive computation and energy costs due to their vast number of parameters and intensive use of power-hungry arithmetic operations. Recently, AI commercial applications have experienced rapid growth in adoption, elevating this problem to a source of environmental concern. Although there have been multiple efforts to optimize DL at both the algorithmic and hardware fronts, there is an ongoing trend for exponentially larger models.

Alternatively, research in various emerging technologies has explored solutions in different machine learning schemes and novel computation substrates. Among them, reservoir computing (RC) has emerged as a promising alternative by proving its high efficiency in early experiments.

This dissertation demonstrates RC's potential by proposing a low-cost image classifier based on cellular automata. However, it also analyzes the severe challenges RC faces to be upscaled to practical applications. Here, this dissertation proposes an approach to efficient DL based on borrowing the key efficiency elements of RC and applying them to current computer vision backbones for an immediate solution to the problem.

The Strong Lottery Ticket Hypothesis (SLTH) recently showed that training a deep neural network (DNN) without learning weights is possible by optimizing a sparse binary pruning mask—a supermask. Folding the DNN architecture beforehand into a recurrent structure results in much smaller models with higher or similar accuracy. Applying these two techniques to a residual neural network (ResNet) results in a DNN with strong similarities to a reservoir computer—the first RC-like DL model, the Hidden-Fold Network (HFN). On top of being RC-like, HFN is an accurate and tiny model that achieves competitive accuracies up to 71.92% on full-scale image classification and model size compression down to 2 MB, small enough for on-chip SRAM.

However, inefficient training makes SLTH models suffer in accuracy on large-scale datasets. This dissertation tackles this problem with Multicoated Supermasks (MSUP), a scalar supermask that raises accuracy to match counterparts with trained weights. An SLTH model that initially only reached 68.16% accuracy on ImageNet is boosted to 74.3% by enhancing the supermask. Furthermore, MSUP combines training, pruning, and quantization into a single concurrent process.

The models that result from combining HFN, MSUP, and learned signs achieve 75.28% accuracy on ImageNet with a model size of only 4 MB, setting the SOTA for SLTH models. Moreover, after analyzing the SLTH and revealing that some of its models are not RC-like, but purely quantized, this dissertation proposes the Ternary Strong Lottery Ticket Hypothesis (T-SLTH). The T-SLTH extends the SLTH to operate with three types of randomness and arbitrary connectivity, opening the door to a wide variety of RC-like DL models and offering a new framework for generic quantization.

Processing the proposed RC-like DL models in a specialized processor is necessary to take advantage of their efficiency benefits. This dissertation presents WhiteDwarf, a holistic software-hardware co-design approach that first applies a triple model compression algorithm and then uses a novel neural inference acceleration architecture for triple unstructured sparsity exploitation. A carefully designed training schedule adapts the proposed T-SLTH models to a mixed precision format of FP8 activations, INT4 weights, and FP16 normalization with virtually no loss in accuracy. Then, the resulting models are compressed using Huffman coding, achieving off-chip compression ratios as high as 290.5x. This compression stack is also extended to a modern deep-MLP architecture. The WhiteDwarf architecture then exploits sparsity at the value and bit-levels of the activations and weights to reach up to 2000x on-chip model compression.

Moreover, a multiplier-less arithmetic unit and a fine-grained clock-gating network harvest further power savings. The fabricated 40-nm CMOS WhiteDwarf chip achieves 12.24 TFLOPS/W. Additionally, it features a rich versatility in configuration aimed at research, which allows a series of ablation studies that demonstrate and quantify the efficacy of the introduced RC-like elements.

This dissertation proposes a new approach to efficient DL that can be immediately applied to real-world computer vision applications. RC-like DL models achieve competitive accuracy on full-scale image classification tasks while vastly reducing model memory size, a primary bottleneck in digital processors driving energy consumption. The results demonstrate that most of the DL computation is superfluous in multiple ways:

1) It is only necessary to learn part of the parameters, leaving the rest random. Architectures need not be deep if they are recurrent;
2) Power-hungry multiplication arithmetic operations can be almost entirely avoided;
3) The proposed approach is tested on ImageNet using ResNet and a deep-MLP architecture both algorithmically and on a fabricated ASIC, thus providing strong evidence of its viability and efficacy.