EMERGING TECH CONFERENCE – Edge Intelligence

Volume 01, 2022, Page 118 – 120

Proceedings of Emerging Tech Conference:
Edge Intelligence 2022

Bio-inspired, resource-efficient design approach for designing Deep Learning models

Nikos Fragoulis*, Ilias Theodorakopoulos, Christos Theocharatos

Irida Labs, Patras InnoHub – Kastritsiou 4, Magoula Patras 26504, Greece

nfrag@iridalabs.com

Abstract

In this paper we present a new building block for designing efficient Convolutional Neural Networks. The goal is to offer a new path in the design of deep Convolutional Neural Networks (CNNs), by formulating a differentiable extension of standard convolution featuring additional mechanisms for competitive inhibition/excitation and stochastic activation with tunable probability. The method pursues the hypothesis that in CNN models featuring such blocks, the optimization of the learning tasks will drive the creation of modes that are information-rich, in a process like the one observed in biological neural networks. Furthermore, the stochastic nature of neuronal activity if appropriately modelled and augmented with sparsity-inducing mechanisms, has the potential to enable the training of models with parametrized levels of sparsity, offering the capacity to control inference/complexity tradeoff on-the-fly, without any need for additional finetuning. Such capabilities for inference with adjustable level of thoroughness/speed can facilitate new applications such as rapid content search in large databases, energy-aware decision-making etc.

1. Introduction

This work aims to explore new paths in the design of deep Artificial Neural Networks (ANNs) and especially Convolutional Neural Networks, which could enable the creation of a new family of highly efficient and adaptive models, easily integrated in any application that relies on the predictive capabilities of deep learning. Inspired by recent advances in the field of biological Interneurons that highlight the importance of inhibition and random connectivity to the encoding efficiency of neuronal circuits, the project aims to create a new paradigm for bio-inspired convolutional layers with enhanced functional diversity, yet compatible with standard training algorithms and high-throughput inference H/W.

Multiple evidence [1,2] suggest that diversity of inhibitory neural mechanisms is likely to contribute to an optimized network encoding capacity, while mediating lateral competition which can be useful in generating sparse codes and solve constraint-satisfaction problems. Additionally, randomness in connectivity can further enhance the encoding reliability of the networks. Such behavior could be of great significance in practical applications utilizing ANNs,

since encoding efficiency has a significant impact in both the predictive accuracy of an ANN, as also to its ability to adapt to different domains and problems.

Inhibition in biological networks can have various forms, two of the most prominent being feedback and feed-forward inhibition. Although the actual dynamics of biological inhibition can be rather complex, these two mechanisms have already some simplified analogues in CNNs architectures such as channel attention [3] and kernel pruning [4]. In this work we pursue the hypothesis that by combining such mechanisms with the conventional convolutional operator, we can significantly enhance the computational capabilities of the fundamental building block of CNNs, enabling the design of models that deliver richer representations in tight budgets. This hypothesis states that by creating competition between the inhibition and excitation mechanisms, the optimization of the learning task will drive the model to create traffic-controlling modes that are information-rich, similar to what is observed in the biological neural networks. Tools that enable modeling of stochastic neuronal activity and formulation of sparsity-inducing constraints [4] in a differentiable manner, could enable training the above model with standard and well-established Deep Learning tools and techniques.

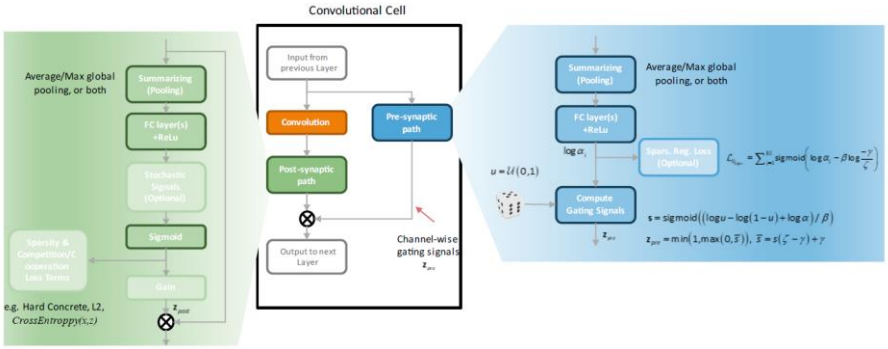

Figure 1: A possible configuration with independent pre- and post-synaptic paths and gating mechanisms

2. Basic Architecture

The basic scheme is based on a generalization of the standard convolutional layer that includes mechanisms of stochastic and/or deterministic group inhibition of the convolution's output in channel level, based on both the input and the output of the convolutional neurons, in a form of a new convolutional cell (Figure 1). This approach is followed such that the different data paths inside the cell to resemble the pre- and post-synaptic inhibition behaviors observed in biological neurons. The adopted formulation for this mechanism can facilitate a variety of loss terms, designed to induce different behaviors on the inhibition signals, such as sparsity and competition/cooperation dynamics. Also, it can facilitate different variations of the dependency and effect of pre and post-synaptic signals that inherit different qualities to the learned representations.

3. Preliminary Results

3.1. Accuracy in Image Classification Tasks

Basic variations of the proposed convolutional cell were implemented in multiple established Deep Learning Frameworks, and preliminary experiments were conducted with various CNN architectures (LeNet5, AlexCifar, wide-ResNet) on small image classification tasks (MNIST Digits,

CIFAR10, CIFAR100). Results indicate that networks featuring the proposed convolutional operator can achieve convergence under comparable conditions (learning hyper-parameters, weights' initialization, etc.) to those utilizing conventional convolutional layers. Furthermore, the extended convolutional operator can deliver moderate accuracy gains for the investigated tasks easily, without extensive hyper-parameter tuning.

3.2. Models with Adjustable Inference Effort

In addition, preliminary experiments were also conducted to test the suitability of the proposed scheme for training CNN models that have characteristics of dynamic (data-dependent) kernel pruning and dynamic adjustment of their accuracy/complexity balance. To test this setting, we used an architecture of a small CNN with 3 convolutional layers with and one FC layer, trained on the CIFAR10 task. by substituting the second and third conv layers with the proposed conv cells equipped only with pre-synaptic signal path and L0 regularization in the pre-synaptic stochastic gates, the new model was trained using the same learning hyperparameters. The resulting model was evaluated using a threshold on the predicted parameter $\log\alpha_j$ of each kernel, indicating whether the respective kernel can be omitted from computation, thus reducing the required inference operations. Results indicate that via the proposed formulation the trained model can indeed exhibit the desired characteristics, achieving the best accuracy of 73.13% for a threshold of -7 (~4% higher than the reference model) in 67% of the reference Multiply-Accumulation (MAC) operations. By just moving the threshold, the same model can achieve faster inference with less accuracy, achieving the reference accuracy at ~50 % of the reference MAC count.

Acknowledgments

This work in the context of project "NeurAdapt", has indirectly received funding from the European Union's Horizon 2020 research and innovation action programme, via the AI4Media Open Call #1 issued and executed under the AI4Media project (Grant Agreement no. 951911).

References

[1] Tsai K-T, Hu C-K, Li K-W, Hwang W-L, Chou Y-H. Circuit variability interacts with excitatory-inhibitory diversity of interneurons to regulatenetwork encoding capacity. Sci Rep 2018; 8:8027.

[2] Jonke Z, Habenschuss S, Maass W. Solving Constraint SatisfactionProblems with Networks of Spiking Neurons. Front Neurosci 2016; 10:118.

[3] Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks. In 2018IEEE/CVF Conf. on Computer Vision and Pattern Recognition, Jun. 2018, pp. 7132–7141.

[4] Louizos C, Welling M, Kingma D P. Learning Sparse Neural Networksthrough L_0 Regularization. Int. Conf. on Learning Representations, Feb.2018. [5] Jacobs, G., & Maun, J. C. (2019, June). Identifying washing machineconsumption in supervised global electric consumption. In 2019 IEEEMilan PowerTech (pp. 1-6). IEEE.