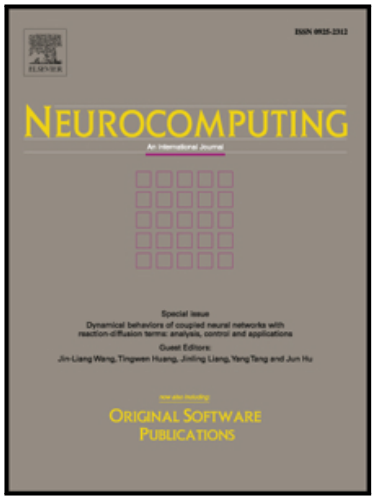
Communicated by Bo Shen

Accepted Manuscript



Training Deep Neural Networks with Discrete State Transition

Guoqi Li, Lei Deng, Lei Tian, Haotian Cui, Wentao Han, Jing Pei, Luping Shi

PII: S0925-2312(17)31186-4
DOI: 10.1016/j.neucom.2017.06.058
Reference: NEUCOM 18657

To appear in: *Neurocomputing*

Received date: 9 August 2016
Revised date: 26 January 2017
Accepted date: 26 June 2017

Please cite this article as: Guoqi Li, Lei Deng, Lei Tian, Haotian Cui, Wentao Han, Jing Pei, Luping Shi, Training Deep Neural Networks with Discrete State Transition, *Neurocomputing* (2017), doi: 10.1016/j.neucom.2017.06.058

Training Deep Neural Networks with Discrete State Transition

Guoqi Li*, Lei Deng*, Lei Tian, Haotian Cui, Wentao Han, Jing Pei and Luping Shi

Abstract

Deep neural networks have been achieving booming breakthroughs in various artificial intelligence tasks, however they are notorious for consuming unbearable hardware resources, training time and power. The emerging pruning/binarization methods, which aim at both decreasing overheads and retaining high performance, seem to promise applications on portable devices. However, even with these most advanced algorithms, we have to save the full-precision weights during the gradient descent process which remains size and power bottlenecks of memory access and the resulting computation. To address this challenge, we propose a unified ***discrete state transition (DST)*** framework by introducing a probabilistic projection operator that constrains the weight matrices in a discrete weight space (DWS) with configurable number of states, throughout the whole training process. The experimental results over various data sets including MNIST, CIFAR10 and SVHN show the effectiveness of this framework. The direct transition between discrete states significantly saves memory for storing weights in full precision, as well as simplifies the computation of weight updating. The proposed DST framework is hardware friendly as it can be easily implemented by a wide range of emerging portable devices, including binary, ternary and multiple-level memory devices. This work paves the way for on-chip learning on various portable devices in the near future.

Keywords: Deep Learning, Neural Network Applications, Discrete State Transition, Discrete Weight Space

1. INTRODUCTION

In the past few years, big datasets, various learning models and GPUs have driven deep neural networks (DNNs) to achieve best results one after another in a wide range of artificial intelligence (AI) tasks, including computer vision [1]-[4], speech recognition [5]-[7], natural language processing [8]-[10], and the incredible success of the AlphaGo [11]. In addition, the applications of neural networks or deep neural networks have also been extended to other disciplines such as physical and chemical problems [12]-[18]. However, behind the appealing demonstrations, the overheads of hardware resources, training time and power consumption are terribly expensive. As a result,

* Equal contributions. G. Li, L. Deng, L.Tian, J. Pei and L. Shi are with the Department of Precision Instrument, Center for Brain Inspired Computing Research, Tsinghua University, Beijing, China, 100084 (email: liguoqi@mail.tsinghua.edu.cn, peij@mail.tsinghua.edu.cn and lpshi@mail.tsinghua.edu.cn). H. Cui is Department of Biomedical Engineering, Tsinghua University, Beijing, China,100084. W. Han is with the Department of Computer Science, Tsinghua University, Beijing, China,100084.