

2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing – NeurIPS Edition (EMC2-NIPS 2019)

**Vancouver, British Columbia, Canada
13 December 2019**



**IEEE Catalog Number: CFP19AT6-POD
ISBN: 978-1-6654-2419-6**

**Copyright © 2019 by the Institute of Electrical and Electronics Engineers, Inc.
All Rights Reserved**

Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

For other copying, reprint or republication permission, write to IEEE Copyrights Manager, IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854. All rights reserved.

****** This is a print representation of what appears in the IEEE Digital Library. Some format issues inherent in the e-media version may also appear in this print version.***

IEEE Catalog Number:	CFP19AT6-POD
ISBN (Print-On-Demand):	978-1-6654-2419-6
ISBN (Online):	978-1-6654-2418-9

Additional Copies of This Publication Are Available From:

Curran Associates, Inc
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: (845) 758-0400
Fax: (845) 758-2633
E-mail: curran@proceedings.com
Web: www.proceedings.com

CURRAN ASSOCIATES INC.
proceedings
.com

2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS Edition (EMC2-NIPS) EMC2-NIPS 2019

Table of Contents

EMC2-NIPS 2019 Preface .viii.....
EMC2-NIPS 2019 Reviewers .ix.....
EMC2-NIPS 2019 Abstracts of Invited Talks .x.....

Technical Papers

Exploring Bit-Slice Sparsity in Deep Neural Networks for Efficient ReRAM-Based Deployment .1...
Jingyang Zhang (Duke University), Huanrui Yang (Duke University), Fan Chen (Duke University), Yitu Wang (Fudan University), and Hai Li (Duke University)

Discovering Low-Precision Networks Close to Full-Precision Networks for Efficient Inference .6.....
Jeffrey L. McKinstry (IBM Research), Steven K. Esser (IBM Research), Rathinakumar Appuswamy (IBM Research), Deepika Bablani (IBM Research), John V. Arthur (IBM Research), Izzet B. Yildiz (IBM Research), and Dharmendra S. Modha (IBM Research)

QPyTorch: A Low-Precision Arithmetic Simulation Framework .10.....
Tianyi Zhang (Cornell University), Zhiqiu Lin (Cornell University), Guandao Yang (Cornell University), and Christopher De Sa (Cornell University)

Trained Rank Pruning for Efficient Deep Neural Networks .14.....
Yuhui Xu (Shanghai Jiao Tong University), Yuxi Li (Shanghai Jiao Tong University), Shuai Zhang (Qualcomm AI Research), Wei Wen (Duke University), Botao Wang (Qualcomm AI Research), Wenrui Dai (Shanghai Jiao Tong University), Yingyong Qi (Qualcomm AI Research), Yiran Chen (Duke University), Weiyao Lin (Shanghai Jiao Tong University), and Hongkai Xiong (Shanghai Jiao Tong University)

Pushing the Limits of RNN Compression .18.....
Urmish Thakker (Army ML Research Lab), Igor Fedorov (Army ML Research Lab), Jesse Beu (Army ML Research Lab), Dibakar Gope (Army ML Research Lab), Chu Zhou (Army ML Research Lab), Ganesh Dasika (Army ML Research Lab), and Matthew Mattina (Army ML Research Lab)

YOLO Nano: A Highly Compact You Only Look Once Convolutional Neural Network for Object Detection .22.....	
	<i>Alexander Wong (University of Waterloo, Canada; DarwinAI Corp., Canada), Mahmoud Famuori (University of Waterloo, Canada; DarwinAI Corp., Canada), Mohammad Javad Shafiee (University of Waterloo, Canada; DarwinAI Corp., Canada), Francis Li (DarwinAI Corp., Canada), Brendan Chwyl (DarwinAI Corp., Canada), and Jonathan Chung (DarwinAI Corp., Canada)</i>
Instant Quantization of Neural Networks Using Monte Carlo Methods .26.....	
	<i>Gonçalo Mordido (Hasso Plattner Institute, Germany), Matthijs Van Keirsbilck (NVIDIA, Germany), and Alexander Keller (NVIDIA, Germany)</i>
Progressive Stochastic Binarization of Deep Networks .31.....	
	<i>David Hartmann (Johannes Gutenberg-University of Mainz, Germany) and Michael Wand (Johannes Gutenberg-University of Mainz, Germany)</i>
Q8BERT: Quantized 8Bit BERT .36.....	
	<i>Ofir Zafrir (Intel Labs, Israel), Guy Boudoukh (Intel Labs, Israel), Peter Izsak (Intel Labs, Israel), and Moshe Wasserblat (Intel Labs, Israel)</i>
Towards Co-Designing Neural Network Function Approximators with In-SRAM Computing .40...	
	<i>Shamma Nasrin (University of Illinois at Chicago), Dīaa Badawi (University of Illinois at Chicago), Ahmet Enis Cetin (University of Illinois at Chicago), Wilfred Gomes (Intel Corp.), and Amit Ranjan Trivedi (University of Illinois at Chicago)</i>
Training Compact Models for Low Resource Entity Tagging Using Pre-Trained Language Models ... 44	
	<i>Peter Izsak (Intel Labs, Israel), Shira Guskin (Intel Labs, Israel), and Moshe Wasserblat (Intel Labs, Israel)</i>
Algorithm-Hardware Co-Design for Deformable Convolution .48.....	
	<i>Qijing Huang (University of California, Berkeley), Dequan Wang (University of California, Berkeley), Yizhao Gao (University of Chinese Academy of Science), Yaohui Cai (Peking University), Zhen Dong (University of California, Berkeley), Bichen Wu (University of California, Berkeley), Kurt Keutzer (University of California, Berkeley), and John Wawrzyniek (University of California, Berkeley)</i>
Bit Efficient Quantization for Deep Neural Networks .52.....	
	<i>Prateeth Nayak (Latent AI), David Zhang (SRI International), and Sek Chai (Latent AI)</i>
Spoken Language Understanding on the Edge .57.....	
	<i>Alaa Saade (Snips, France), Alice Coucke (Snips, France), Alexandre Caulier (Snips, France), Joseph Dureau (Snips, France), Adrien Ball (Snips, France), Théodore Bluche (Snips, France), David Leroy (Snips, France), Clément Doumouro (Snips, France), Thibault Gisselbrecht (Snips, France), Francesco Caltagirone (Snips, France), Thibaut Lavril (Snips, France), and Mael Primet (Snips, France)</i>
Neural Networks Weights Quantization: Target None-Retraining Ternary (TNT) .62.....	
	<i>Tianyu Zhang (Webank, China), Lei Zhu (Harbin Engineering University, China), Qian Zhao (University of Hyogo, Japan), and Kilho Shin (Gakushuin University, Japan)</i>

On Hardware-Aware Probabilistic Frameworks for Resource Constrained Embedded Applications
66

Laura I. Galindez Olascoaga (KU Leuven & UC Berkeley), Wannes Meert (KU Leuven), Nimish Shah (KU Leuven), Guy Van den Broeck (UCLA), and Marian Verhelst (KU Leuven)

Author Index 71