2-2022

Low-Power Computer Vision: Improve the Efficiency of Artificial Intelligence

George K. Thiruvathukal
Loyola University Chicago, gthiruvathukal@luc.edu

Yung-Hisang Lu
Purdue University

Jaeyoun Kim
Google

Yiran Chen
Duke University

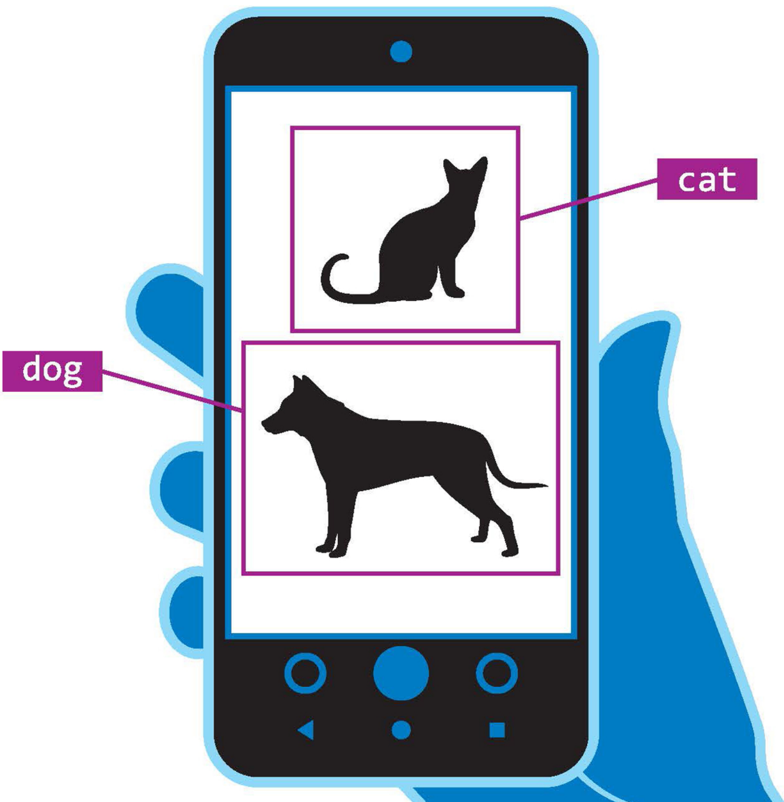
Bo Chen

Recommended Citation

Thiruvathukal, G.K., Lu, Y.-H., Kim, J., Chen, Y., & Chen, B. (Eds.). (2022). Low-Power Computer Vision: Improve the Efficiency of Artificial Intelligence (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781003162810

LOW-POWER COMPUTER VISION

IMPROVING THE EFFICIENCY OF ARTIFICIAL INTELLIGENCE



EDITED BY
GEORGE K. **THIRUVATHUKAL**
YUNG-HSIANG **LU** JAEYOUN **KIM**
YIRAN **CHEN** BO **CHEN**

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Low-Power Computer Vision

Low-Power Computer Vision

Improve the Efficiency of Artificial Intelligence

Edited by

George K. Thiruvathukal

Yung-Hsiang Lu

Jaeyoun Kim

Yiran Chen

Bo Chen

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

First edition published 2022
by CRC Press
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

and by CRC Press
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

CRC Press is an imprint of Taylor & Francis Group, LLC

Library of Congress Cataloging-in-Publication Data

Names: Thiruvathukal, George K. (George Kuriakose), editor.
Title: Low-power computer vision : improve the efficiency of artificial intelligence / edited by George K. Thiruvathukal, Yung-Hsiang Lu, Jaeyoun Kim, Yiran Chen, Bo Chen.
Description: First edition. | Boca Raton : CRC Press, [2022] | Includes bibliographical references and index.
Identifiers: LCCN 2021042753 | ISBN 9780367744700 (hbk) | ISBN 9780367755287 (pbk) | ISBN 9781003162810 (ebk)
Subjects: LCSH: Computer vision. | Low voltage systems.
Classification: LCC TA1634 .L69 2022 | DDC 006.3/7--dc23/eng/20211028
LC record available at https://lccn.loc.gov/2021042753

ISBN: 978-0-367-74470-0 (hbk)
ISBN: 978-0-367-75528-7 (pbk)
ISBN: 978-1-003-16281-0 (ebk)

DOI: 10.1201/9781003162810

Publisher's note: This book has been prepared from camera-ready copy provided by the authors.

Typeset in LM Roman
by KnowledgeWorks Global Ltd.

Contents

Foreword

Whereas electronics and computing have provided our society with unprecedented means of advancing services in this millennium, the environmental cost of using electronic technology is becoming significant. For this reason, low-energy and low-power computing has become an important area of research and development. Moreover, the miniaturization of devices, for example phones and drones, requires small energy reservoirs (i.e., low-volume, low-weight batteries). The pioneering work on digital watches of the eighties has grown up by now to a full array of hardware and software design technologies to mitigate the energy consumption of processing and storage elements in many areas.

From an application perspective, the ability of recognizing situations and actors, possibly within a complex environment, has become the key element in creating advanced systems in many domains, such as security, automated driving, and surveying. There has been a tremendous growth in the capabilities of image recognition systems in both hardware and software, and the presence of such systems is now almost ubiquitous. Nevertheless, the complexity of recognition requires a corresponding energy cost. As in the case of other electronic systems, the energy consumption may be significantly high and be an impediment to a wide use of image recognition in some domains.

As a result of the aforementioned considerations, the search for low-power computer vision systems is a key problem in both the research and development fields. There is a wide gap between the ideal minimum energy cost solutions and the current realizations. This gap is hard to quantify, as many factors come into play, ranging from the non-ideality of electronic devices (e.g., leakage current) to the choice of heuristic algorithms that approximate solutions because of the inherent computational complexity. On the bright side, this wide gap enables a continuous search for improvements within the entire design space spectrum, from circuits to algorithms, from hardware architectures to software programs.

The search for bettering energy efficiency would not be possible without realistic drivers and a world-wide participation of researchers. This is why the low-power computer vision challenge has been, and currently is, an important instrument for advancing the state of the art. The challenge was taken by some of the best groups in the world, and their effort has tackled the problem with different means and perspectives. Overall, this challenge has brought us very important results, that are fully documented in this book, and that will provide a strong impact on industry and academia.

Lausanne, March 2021

Giovanni De Micheli

Rebooting Computing and Low-Power Computer Vision

Since its start in 2013 as an initiative of IEEE Future Directions Committee, "Rebooting Computing" has provided an international, interdisciplinary environment where experts from a wide variety of computer-related fields can come together to explore novel approaches to future computing. The need for Rebooting Computing follows from the recognition that the exponential improvement in computing performance in previous decades was due primarily to transistor scaling in Moore's Law, but this is coming to an end. Radical alternative approaches are needed over the entire technology landscape, from basic devices and circuits to architectures to software, with applications from supercomputers to smartphones. Some possible newer approaches that are being explored include neuromorphic computing, approximate and stochastic computing, quantum and cryogenic computing, low-power reversible and adiabatic computing, and computing based on non-volatile memories, analog and optical systems. The initiative has now evolved to become a Task Force within the Computer Society of IEEE and continues its mission unabated.

"Rebooting Computing" spawned many innovations, including the Low-Power Image Recognition Challenge (LPIRC) in 2015, the brainchild of Prof. Yung-Hsiang Lu. LPIRC ran for several years with ever-improving performance by the teams demonstrating subsystems for image recognition at the lowest possible power. Importantly, the competition involved a multitude of students, providing inspiration and motivation to students worldwide. LPIRC was renamed as the Low-Power Computer Vision Challenge (LPCVC) in 2020 when video was also included. These challenges evaluate both accuracy and energy consumption of systems that can recognize and understand images or videos. Over the six years since the inception of the Challenge, more than 100 teams have participated. The teams have sponsorship and participation from industry, including Facebook, Google, Xilinx, ELAN Microelectronics, Amazon, Qualcomm, and Bytedance.

This book contains the collection of the solutions of the winners of the Challenge. The authors compare different options, making computer vision more efficient and explaining important design decisions. The information provides deep insight for researchers and practitioners.

Elie K. Track, CEO of nVizix LLC,
Founding Co-Chair of the IEEE Rebooting Initiative

Editors

George K. Thiruvathukal is a professor of Computer Science at Loyola University Chicago, Illinois, USA. He is also a visiting faculty at Argonne National Laboratory. His research areas include high performance and distributed computing, software engineering, and programming languages.

Yung-Hsiang Lu is a professor of Electrical and Computer Engineering at Purdue University, Indiana, USA. He is the first director of Purdue's John Martinson Engineering Entrepreneurial Center. He is a fellow of the IEEE and distinguished scientist of the ACM. His research interests include computer vision, mobile systems, and cloud computing.

Jaeyoun Kim is a technical program manager at Google, California, USA. He leads AI research projects, including MobileNets and TensorFlow Model Garden, to build state-of-the-art machine learning models and modeling libraries for computer vision and natural language processing.

Yiran Chen is a professor of Electrical and Computer Engineering at Duke University, North Carolina, USA. He is a fellow of the ACM and the IEEE. His research areas include new memory and storage systems, machine learning and neuromorphic computing, and mobile computing systems.

Bo Chen is the Director of AutoML at DJI, Guangdong, China. Before joining DJI, he was a researcher at Google, California, USA. His research interests are the co-optimization of neural network software and hardware as well as landing AI technology in products with stringent resource constraints.

I

Introduction

CHAPTER 1

Book Introduction

Yung-Hsiang Lu

Purdue University

George K. Thiruvathukal

Loyola University Chicago

Jaeyoun Kim

Google California

Yiran Chen

Duke University

Bo Chen

Da-Jiang Innovations China

CONTENTS

DOI: 10.1201/9781003162810-1

1.1 ABOUT THE BOOK

The first IEEE Low-Power Image Recognition Challenge was held in 2015. Since then, winners have presented their solutions in conferences and published detailed studies in journals. After six years of competitions, there is a rich set of knowledge about how to make computer vision efficient running on embedded computers. The organizers decided to put together this book so that researchers, engineers, and practitioners can understand what methods worked well for winning the competitions.

The book is composed of three parts: Introduction, Winners' Solutions, and Invited Articles. The first part provides a brief history of the competitions and a survey of literature. The second part includes the articles from the winners. All winners were invited to contribute to this book; this part of the book includes the articles from the winners that accepted the invitations. The third part contains articles from leaders in low-power computer vision, including authors from industry and academia.

1.2 CHAPTER SUMMARIES

1.2.1 History of Low-Power Computer Vision Challenge

Yung-Hsiang Lu (Purdue University); Xiao Hu (Purdue University); Yiran Chen (Duke University); Joe Spisak (Facebook); Gaurav Aggarwal (Facebook); Mike Zheng Shou (Facebook Research), and George K. Thiruvathukal (Loyola University Chicago)

Abstract

This chapter describes the history of IEEE History of Low-Power Computer Vision Challenge 2015–2020.

Take-aways

- Describes the history of the IEEE Low-Power Computer Vision Challenge between 2015 and 2020.
- Explains the methods to select winners and lists the winners over these years.

1.2.2 Survey on Energy-Efficient Deep Neural Networks for Computer Vision

Abhinav Goel (Purdue University); Caleb Tung (Purdue University); Xiao Hu (Purdue University); Haobo Wang (Purdue University); George Thiruvathukal (Loyola University Chicago); Yung-Hsiang Lu (Purdue University)

Abstract

Deep Neural Networks (DNNs) are greatly successful in performing many different computer vision tasks. However, the state-of-the-art DNNs are too energy, computation, and memory-intensive to be deployed on most computing devices and embedded systems. DNNs usually require server-grade CPUs and GPUs. To make computer vision more ubiquitous, recent research has focused on making DNNs more efficient. These techniques make DNNs smaller and faster through various refinements and thus are enabling computer vision on battery-powered mobile devices. Through this article, we survey the recent progress in low-power deep learning to discuss and analyze the advantages, limitations, and potential improvements to the different techniques. We particularly focus on the software-based techniques for low-power DNN inference. This survey classifies the energy-efficient DNN techniques into six broad categories: (1)Quantization, (2)Pruning, (3)Layer and Filter Compression, (4)Matrix Decomposition, (5)Neural Architecture Search, and (6)Knowledge Distillation. The techniques in each category are discussed in greater detail in this chapter.

Take-aways

- Surveys the recent progress in low-power deep learning to analyze the advantages, limitations, and potential improvements to the different techniques.
- Focus on the software-based techniques for low-power DNN inference

1.2.3 Hardware Design and Software Practices for Efficient Neural Network Inference

Yu Wang (Tsinghua University); Xuefei Ning (Tsinghua University); Shulin Zeng (Tsinghua University); Changcheng Tang (Novauto); Yi Cai (Tsinghua University); Kaiyuan Guo (Tsinghua University); Shuang Liang (Novauto); Tianyi Lu (Novauto); Hanbo Sun (Tsinghua University); Tianchen Zhao (Beihang University)

Abstract

In this chapter, we introduce our efforts in accelerating neural network inference. From the hardware design aspect, we introduce the instructions-set-architecture deep learning accelerator to support all kinds of DNN models with customized ISA and optimized software compiler. And from the algorithm aspect, we introduce several practices we have used: sensitivity-based pruning without hardware model, quantization, iterative pruning with hardware model, and neural architecture search.

Take-aways

- Discusses hardware design: An instructions-set-architecture deep learning accelerator to support all kinds of DNN models with customized ISA and optimized software compile
- Discusses software practices: Sensitivity-based pruning without hardware model, quantization, iterative pruning with hardware model, neural architecture search.

1.2.4 Progressive Automatic Design of Search Space for One-Shot Neural Architecture

Xin Xia (Bytedance Inc); Xuefeng Xiao (ByteDance Inc); XING WANG (Bytedance AI Lab)

Abstract

Neural Architecture Search (NAS) has attracted growing interest. To reduce the search cost, recent work has explored weight sharing across models and made major progress in One-Shot NAS. However, it has been observed that a model with higher one-shot model accuracy does not necessarily perform better when stand-alone trained. To address this issue, in this paper, we propose Progressive Automatic Design of search space, named PAD-NAS. Unlike previous approaches where the same

operation search space is shared by all the layers in the supernet, we formulate a progressive search strategy based on operation pruning and build a layer-wise operation search space. In this way, PAD-NAS can automatically design the operations for each layer. During the search, we also take the hardware platform constraints into consideration for efficient neural network model deployment. Extensive experiments on ImageNet show that our method can achieve state-of-the-art performance.

Take-aways

- Uses network architecture search methods to find better architectures for lower latencies and higher accuracy
- Formulates a search strategy to build a layer-wise operation search space through hierarchical operation pruning and mitigates weight coupling issue in One-Shot NAS.
- Compares the effects of different parameters on memory sizes, latency, and accuracy

1.2.5 Fast Adjustable Threshold for Uniform Neural Network Quantization

Alexander Goncharenko (Novosibirsk State University); Andrey Denisov (Expasoft); Sergey Alyamkin (Expasoft)

Abstract

The neural network quantization is highly desired procedure to perform before running neural networks on mobile devices. Quantization without fine-tuning leads to accuracy drop of the model, whereas commonly used training with quantization is done on the full set of the labeled data and therefore is both time- and resource-consuming. Real-life applications require simplification and acceleration of quantization procedure that will maintain the accuracy of full-precision neural network, especially for modern mobile neural network architectures like Mobilenet-v1, MobileNet-v2, and MNAS. Here we present two methods to significantly optimize the training with quantization procedure. The first one is introducing the trained scale factors for discretization thresholds that are separate for each filter. The second one is based on mutual rescaling of consequent depth-wise separable convolution and convolution layers. Using the proposed techniques, we quantize the modern mobile architectures of neural

networks with the set of train data of only 10% of the total ImageNet 2012 sample. Such reduction of train dataset size and small number of trainable parameters allow to fine-tune the network for several hours while maintaining the high accuracy of quantized model (accuracy drop was less than 0.5%). Ready-for-use models and code are available at: https://github.com/agoncharenko1992/FAT-fast-adjustable-threshold.

Take-aways

- Describes ways how to get an 8-bit quantized network.
- The main idea is that simple min/max quantization with calibration works poor because of outliers which spoils thresholds of quantization.
- We can adjust this thresholds by using Straight-Through Estimators. Using some tips such as Batch Normalization folding and, channel equalization (more details you can found in the paper) we can get solution as good as training with quantization from scratch but with less data and way faster.

1.2.6 Power-efficient Neural Network Scheduling on Heterogeneous system on chips (SoCs)

Ying Wang (Institute of Computing Technology, Chinese Academy of Sciences); Xuyi Cai (Institute of Computing Technology, Chinese Academy of Sciences); Xiandong Zhao (Institute of Computing Technology, Chinese Academy of Sciences)

Abstract

The powerful deep neural networks (DNNs) have been propelling the development of efficient computer vision technologies for mobile systems such as phones and drones. To enable power-efficient image processing on resource-constrained devices, many studies have been dedicated to the field of low-power DNNs from different layers of the systems. Amongst the deep stack of low-power DNN systems, task scheduling also plays an essential role as the interfacing middleware between the algorithms and the underlying hardware. Especially when heterogeneous SoCs have been widely adopted in edge and mobile scenarios as the hardware solution, an efficient DNN task scheduler is needed to reduce the implementation

overhead of DNN-based task and extract the most power from the SoC platform. This chapter will firstly exemplify DNN scheduling with the image recognition solution of LPIRC-2016 and introduce how to efficiently schedule a DNN-based visual processing task onto a typical heterogeneous SoC composed of general-purpose and specialized cores. After the elaborate task-level scheduling strategy, we will discuss the fine-grained DNN-wise scheduling policy on specialized DNN cores and show the effectiveness of memory-oriented DNN-layer scheduling. Last, since model quantization is an indispensable step to map a large-size neural network model onto the resource-thrifty mobile SoCs, we will discuss the implication of DNN quantization on the heterogeneous SoCs integrated with both integer and float-point cores, and then introduce the scheduler-friendly DNN quantizer for pure-integer hardware. Although most prior works on low-power DNNs focused their attention on efficient network and hardware architectures, it is shown that the scheduler-level optimization technology will also be critical to the energy-efficiency of the system, particularly when the algorithmic implementation is fixed and off-the-shelf hardware devices are adopted.

Take-aways

- Demonstrates the rank-1 solution of LPIRC2016 as a case study to introduce the basic coarse-grained scheduling techniques for DNN-based applications.
- Presents the memory-efficient fine-grained neural network scheduler on DNN processors.
- Introduces the scheduler-friendly quantization technique to reduce the overhead of neural network implementation on embedded SoCs.

1.2.7 Efficient Neural Architecture Search

Han Cai and Song Han (MIT)

Abstract

Designing efficient neural network architectures is a widely adopted approach to improve efficiency, besides compressing an existing deep neural network. A CNN (Convolutional Neural Network) model typically consists of convolution layers, pooling layers, and fully-connected layers,

where most of the computation comes from convolution layers. For example in ResNet-50, more than 99% multiply-accumulate operations (MACs) are from convolution layers. Therefore, designing efficient convolution layers is the core of building efficient CNN architectures. This chapter first describes the standard convolution layer and then describes three efficient variants of the standard convolution layer. Next, we present three representative manually design efficient CNN architectures, including SqueezeNet, MobileNets, and ShuffleNets. Finally, we describe automated methods for designing efficient CNN architectures.

Take-aways

- Describes the standard convolution layer and then describes three efficient variants of the standard convolution layer.
- Presents three representative manually designed efficient CNN architectures, including SqueezeNet, MobileNets, and ShuffleNets.
- Describes automated methods for designing efficient CNN architectures.

1.2.8 Design Methodology for Low-Power Image Recognition Systems Design

Soonhoi Ha (Seoul National University); EunJin Jeong (Seoul National University); Duseok Kang (Seoul National University); Jangryul Kim (Seoul National University); Donghyun Kang (Seoul National University)

Abstract

In the development of an embedded image recognition system, there are many issues to consider, such as which hardware platform and algorithm to use, how to optimize the software with resource constraints and how to optimize multiple design objectives, and so on. This chapter presents a systematic design methodology that could be applied to the design of embedded systems with a concrete example of image recognition systems. Based on the proposed methodology, we could win the first prize in LPIRC (Low-Power Image Recognition Challenge) 2017. After selecting NVIDIA Jetson TX2 as the hardware platform and Tiny YOLO as the detection algorithm, we applied the well-known software optimization techniques in a systematic way, aiming to jointly optimize speed, accuracy, and

energy. We have refined the methodology to choose a different algorithm on the same hardware platform and could build another winning solution in track 2 of LPIRC 2018. Recently new hardware platforms have been developed that contain CNN hardware accelerators as well as GPU (Graphics Processing Units), among which NVIDIA Jetson AGX Xavier is a representative example. Since it is a heterogeneous system that contains multiple hardware accelerators, how to exploit the computing power of those accelerators maximally becomes an important issue to consider in the proposed design methodology. We have developed a novel technique to maximally utilize multiple accelerators to achieve 21.7 times better score than our previous solution in LPIRC 2018, which is also presented in this chapter.

Take-aways

- First prize winning solution in LPIRC 2017 and in track2 of LPIRC 2018.
- Presents a systematic design methodology for the design of low-power image recognition systems.
- Demonstrates how to select the hardware platform and a neural network by considering the estimated performance.
- Demonstrates how to map the network onto the hardware platform aiming to maximize the throughput by pipelining.
- Shows how various software optimization techniques are then applied to each processing element.

1.2.9 Guided Design for Efficient On-device Object Detection Model

Tao Sheng and Yang Liu (Amazon)

The low-power computer vision (LPCV) challenge is an annual competition for the best technologies in image classification and object detection measured by both efficiency (execution time and energy consumption) and accuracy (precision/recall). Our Amazon team has won three awards from LPCV challenges: 1st prize for interactive object detection challenge in 2018 and 2019 and 2nd prize for interactive image classification challenge in 2018. This paper is to share our award-winning methods, which can be summarized as four major steps. First, 8-bit quantization friendly model is one of the key winning points to achieve the short execution time while maintaining the high accuracy on edge devices. Second, network architecture optimization is another winning keypoint. We optimized the network architecture to meet the 100ms latency requirement on Pixel2 phone. The third one is dataset filtering. We removed the images with small objects from the training dataset after deeply analyzing the training curves, which significantly improved the overall accuracy. And the forth one is non-maximum suppression optimization. By combining all the above steps together with the other training techniques, for example, cosine learning function and transfer learning, our final solutions were able to win the top prizes out of large number of submitted solutions across worldwide.

Take-aways:

- Discusses the methods involved in the winning solutions over the years.
- Explains the impacts of each method (quantization, architecture search, hyperparameter tuning)
- Reduces the resolutions to improve performance

1.2.10 Quantizing Neural Networks for Low-Power Computer Vision

Markus Nagel (Qualcomm); Marios Fournarakis (Qualcomm); Rana Ali Amjad (Qualcomm); Yelysei Bondarenko (Qualcomm); Mart van Baalen (Qualcomm); Tijmen Blankevoort (Qualcomm)

Abstract

Over the last years, Neural Networks (NNs) have been widely adapted in Computer Vision (CV) applications. While for many tasks they outperform traditional CV algorithms they often come at a high compute cost. Even mobile friendly architectures such as MobileNet still require hundreds of million floating point operations. To further reduce the energy efficiency and latency of NNs, quantization can be used to replace the original floating-point operations with low bit fixed-point operations. In this chapter we introduce NN quantization for low-power computer vision. Afterward we highlight recent advances in post-training quantization, a class of algorithms that can be applied to pretrained NNs and do not require any expert knowledge. In the last part we will focus on

quantization-aware training, a technique that trains NNs with simulated quantization operations.

Take-aways

- Introduces neural network quantization
- Serves as a practical guide to quantization simulation with HW considerations
- Introduces state-of-the-art post-training quantization (PTQ) techniques that are easy to use.
- Introduces state-of-the-art quantization-aware training (QAT) approaches that result in best performance.
- Defines standard PTQ and QAT pipeline and evaluates them on several computer vision models and tasks.

1.2.11 A Practical Guide to Designing Efficient Mobile Architectures

Mark Sandler and Andrew Howard (Google)

Abstract

In this chapter we overview a set of basic techniques that can be applied when designing and fine-tuning efficient architectures. We establish basic principles that practitioners can use when adapting existing architectures to particular applications. While a lot of modern research has been dedicated to network architecture search, the basic design principles are often poorly understood. Our goal here is to build a solid foundation and demystify the reasoning about image neural networks from a practical perspective. From our experience, such a foundation is indispensable for both designing new architecture search spaces, as well as for practical tuning of existing architectures to new hardware and/or problems, without relying on opaque Network Architecture Search (NAS) techniques.

Take-aways

- Introduces a set of basic techniques for adapting and fine-tuning existing model architectures to different hardware and problems.

- Provides an in-depth overview of several types of multipliers that enable a user to independently adjust resource consumption such as model size, memory requirements, and energy consumption.
- Demonstrates more specialized ways to fine-tune individual layers.
- Demonstrates ways to phase in custom nonlinearities that have limited support on existing hardware.

1.2.12 A Survey of Quantization Methods for Efficient Neural Network Inference

Amir Gholami (UC Berkeley); Sehoon Kim (University of California, Berkeley); Zhen Dong (UC Berkeley); Zhewei Yao (University of California, Berkeley); Michael Mahoney (University of California, Berkeley); Kurt Keutzer (EECS, UC Berkeley)

Abstract

As soon as abstract mathematical computations were adapted to computation on digital computers, the problem of efficient representation, manipulation, and communication of the numerical values in those computations arose. Strongly related to the problem of numerical representation is the problem of quantization: in what manner should a set of continuous real-valued numbers be distributed over a fixed discrete set of numbers to minimize the number of bits required and also to maximize the accuracy of the attendant computations? This perennial problem of quantization is particularly relevant whenever memory and/or computational resources are severely restricted, and it has come to the forefront in recent years due to the remarkable performance of Neural Network models in computer vision, natural language processing, and related areas. Moving from floating-point representations to low-precision fixed integer values represented in four bits or less holds the potential to reduce the memory footprint and latency by a factor of 16x; and, in fact, reductions of 4x to 8x are often realized in practice in these applications. Thus, it is not surprising that quantization has emerged recently as an important and very active sub-area of research in the efficient implementation of computations associated with Neural Networks. In this article, we survey approaches to the problem of quantizing the numerical values in deep Neural Network computations, covering the advantages/disadvantages of current methods. With this survey and its organization, we hope to have

presented a useful snapshot of the current research in quantization for Neural Networks and to have given an intelligent organization to ease the evaluation of future research in this area.

Take-aways

- As soon as abstract mathematical computations were adapted to computation on digital computers, the problem of efficient representation, manipulation, and communication of the numerical values in those computations arose.
- Strongly related to the problem of numerical representation is the problem of quantization, which is the main focus of this chapter.
- We will first introduce the basic concepts of quantization, and then discuss the advanced methods, as well as open problems in this area.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Power-efficient Neural Network Scheduling

Zhao, X. , Wang, Y. , Cai, X. , Liu, C. , and Zhang, L. (2019, September). Linear symmetric quantization of neural networks for low-precision integer hardware. In International Conference on Learning Representations.

Wang, Y. , Quan, Z. , Li, J. , Han, Y. , Li, H. , and Li, X. (2018, March). A retrospective evaluation of energy-efficient object detection solutions on embedded devices. In 2018 Design, Automation and Test in Europe Conference and Exhibition (DATE) (pp. 709–714). IEEE.

Zhao, X. , Wang, Y. , Liu, C. , Shi, C. , Tu, K. , and Zhang, L. (2020, July). BitPruner: network pruning for bit-serial accelerators. In 2020 57th ACM/IEEE Design Automation Conference (DAC) (pp. 1–6). IEEE.

Wang, C. , Wang, Y. , Han, Y. , Song, L. , Quan, Z. , Li, J. , and Li, X. (2017, January). CNN-based object detection solutions for embedded heterogeneous multicore SoCs. In 2017 22nd Asia and South Pacific Design Automation Conference (ASP-DAC) (pp. 105–110). IEEE.

Xu, H. , Wang, Y. , Wang, Y. , Li, J. , Liu, B. , and Han, Y. (2019, November). ACG-Engine: An Inference Accelerator for Content Generative Neural Networks. In 2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD) (pp. 1–7). IEEE.

Xu, D. , Liu, C. , Wang, Y. , Tu, K. , He, B. , and Zhang, L. (2020). Accelerating generative neural networks on unmodified deep learning processors—A software approach. IEEE Transactions on Computers, 69(8), 1172–1184.

Cai, X. , Wang, Y. , Zhang, L. (2021) Optimus: Towards Optimal Layer-Fusion on Deep Learning Processors In The 22st ACM SIGPLAN/SIGBED Conference on Languages, Compilers, and Tools for Embedded Systems.

Chen, W. , Wang, Y. , Lin, G. , Gao, C. , Liu, C. , and Zhang, L. . CHaNAS: Coordinated Search for Network Architecture and Scheduling Policy. In The 22st ACM SIGPLAN/SIGBED Conference on Languages, Compilers, and Tools for Embedded Systems.

References

Yung-Hsiang Lu , Alexander C. Berg , and Yiran Chen . Low-power image recognition challenge. AI Magazine, 39(2):87–88, Jul. 2018.

Yung-Hsiang Lu . Low-power image recognition. Nature Machine Intelligence, 1(4):199–199, Apr 2019.

K. Gauen , R. Rangan , A. Mohan , Y. Lu , W. Liu , and A. C. Berg . Low-power image recognition challenge. In 22nd Asia and South Pacific Design Automation Conference (ASP-DAC), pages 99–104, 2017.

Y. Lu , A. M. Kadin , A. C. Berg , T. M. Conte , E. P. DeBenedictis , R. Garg , G. Gingade , B. Hoang , Y. Huang , B. Li , J. Liu , W. Liu , H. Mao , J. Peng , T. Tang , E. K. Track , J. Wang , T. Wang , Y. Wang , and J. Yao . Rebooting computing and low-power image recognition challenge. In *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)* , pages 927–932, 2015.

S. Alyamkin , M. Ardi , A. C. Berg , A. Brighton , B. Chen , Y. Chen , H. Cheng , Z. Fan , C. Feng , B. Fu , K. Gauen , A. Goel , A. Goncharenko , X. Guo , S. Ha , A. Howard , X. Hu , Y. Huang , D. Kang , J. Kim , J. G. Ko , A. Kondratyev , J. Lee , S. Lee , S. Lee , Z. Li , Z. Liang , J. Liu , X. Liu , Y. Lu , Y. Lu , D. Malik , H. H. Nguyen , E. Park , D. Repin , L. Shen , T. Sheng , F. Sun , D. Svitov , G. K. Thiruvathukal , B. Zhang , J. Zhang , X. Zhang , and S. Zhuo . Low-power computer vision: Status, challenges, and opportunities. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 9(2):411–421, 2019.

M. Ardi , A. C. Berg , B. Chen , Y. Chen , Y. Chen , D. Kang , J. Lee , S. Lee , Y. Lu , Y. Lu , and F. Sun . Special session: 2018 low-power image recognition challenge and beyond. In *IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)* , pages 154–157, 2019.

K. Gauen , R. Dailey , Y. Lu , E. Park , W. Liu , A. C. Berg , and Y. Chen . Three years of low-power image recognition challenge: Introduction to special session. In Design, Automation Test in Europe Conference Exhibition (DATE *)*, pages 700–703, 2018.

Karen Simonyan and Andrew Zisserman . Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs], April 2015. arXiv: 1409.1556.

Andrew G. Howard , Menglong Zhu , Bo Chen , Dmitry Kalenichenko , Weijun Wang , Tobias Weyand , Marco Andreetto , and Hartwig Adam . MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv:1704.04861 [cs], April 2017. arXiv: 1704.04861.

Song Han , Huizi Mao , and William J. Dally . Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding *arXiv:1510.00149 [cs]*, October 2015. arXiv: 1510.00149.

Joseph Redmon and Ali Farhadi . YOLOv3: An Incremental Improvement. arXiv:1804.02767 [cs], April 2018. arXiv: 1804.02767.

Yann LeCun , Yoshua Bengio , and Geoffrey Hinton . Deep learning. Nature, 521(7553):436–444, May 2015. Number: 7553 Publisher: Nature Publishing Group.

Forrest N. Iandola , Song Han , Matthew W. Moskewicz , Khalid Ashraf , William J. Dally , and Kurt Keutzer . SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. arXiv:1602.07360 [cs], November 2016. arXiv: 1602.07360.

Anup Mohan , Kent Gauen , Yung-Hsiang Lu , Wei Wayne Li , and Xuemin Chen . Internet of video things in 2030: A world with many cameras. In IEEE International Symposium on Circuits and Systems (ISCAS), pages 1–4, Baltimore, MD, USA, May. IEEE.

K. Kumar and Y. Lu . Cloud Computing for Mobile Users: Can Offloading Computation Save Energy? Computer, 43(4):51–56, April 2010. Conference Name: Computer.

S. Alyamkin , M. Ardi , A. C. Berg , A. Brighton , B. Chen , Y. Chen , H. Cheng , Z. Fan , C. Feng , B. Fu , K. Gauen , A. Goel , A. Goncharenko , X. Guo , S. Ha , A. Howard , X. Hu , Y. Huang , D. Kang , J. Kim , J. G. Ko , A. Kondratyev , J. Lee , S. Lee , S. Lee , Z. Li , Z. Liang , J. Liu , X. Liu , Y. Lu , Y. Lu , D. Malik , H. H. Nguyen , E. Park , D. Repin , L. Shen , T. Sheng , F. Sun , D. Svitov , G. K. Thiruvathukal , B. Zhang , J. Zhang , X. Zhang , and S. Zhuo . Low-Power Computer Vision: Status, Challenges, and Opportunities. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 9(2):411–421, June 2019.

Forrest Iandola and Kurt Keutzer . Small neural nets are beautiful: enabling embedded systems with small deep-neural-network architectures. In *Proceedings of the Twelfth IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis Companion* , CODES, pages 1–10, New York, NY, USA, October 2017. Association for Computing Machinery.

Hao Li , Asim Kadav , Igor Durdanovic , Hanan Samet , and Hans Peter Graf . Pruning Filters for Efficient ConvNets. *arXiv:1608.08710 [cs]*, August 2016. arXiv: 1608.08710.

A. Goel , C. Tung , Y.-H. Lu , and G. K. Thiruvathukal . A Survey of Methods for Low-Power Deep Learning and Computer Vision. In IEEE 6th World Forum on Internet of Things (WF-IoT), pages 1–6, June 2020.

Isha Ghodgaonkar , Subhankar Chakraborty , Vishnu Banna , Shane Allcroft , Mohammed Metwaly , Fischer Bordwell , Kohsuke Kimura , Xinxin Zhao , Abhinav Goel , Caleb Tung , Akhil Chinnakotla , Minghao Xue , Yung-Hsiang Lu , Mark Daniel Ward , Wei Zakharov , David S. Ebert , David M. Barbarash , and George K. Thiruvathukal . Analyzing Worldwide Social Distancing through Large-Scale Computer Vision. arXiv:2008.12363 [cs], August 2020. arXiv: 2008.12363.

Y. LeCun , B. Boser , J. S. Denker , D. Henderson , R. E. Howard , W. Hubbard , and L. D. Jackel . Backpropagation applied to handwritten zip code recognition. Neural Computation, 1(4):541–551, 1989.

Raghuraman Krishnamoorthi . Quantizing deep convolutional networks for efficient inference: A whitepaper. arXiv:1806.08342 [cs, stat], June 2018. arXiv: 1806.08342.

Abhinav Goel , Sara Aghajanzadeh , Caleb Tung , Shuo-Han Chen , George K. Thiruvathukal , and Yung-Hsiang Lu . Modular Neural Networks for Low-Power Image Classification on Embedded Devices. ACM Transactions on Design Automation of Electronic Systems, 26(1):1:1–1:35, October 2020.

Mahdi Nazemi , Ghasem Pasandi , and Massoud Pedram . NullaNet: Training Deep Neural Networks for Reduced-Memory-Access Inference. arXiv:1807.08716 [cs, stat], August 2018. arXiv: 1807.08716.

Jerry Bowles . How Athena Security found a brand new market in the midst of the COVID-19 pandemic, August 2020. Section: Governing identity privacy and security.

S. Aghajanzadeh , R. Naidu , S.-H. Chen , C. Tung , A. Goel , Y.-H.. Lu , and G. K. Thiruvathukal . Camera Placement Meeting Restrictions of Computer Vision. In *IEEE International Conference on Image Processing (ICIP)* , pages 3254–3258, October 2020. ISSN: 2381-8549.

M. Ardi , A. C. Berg , B. Chen , Y. Chen , Y. Chen , D. Kang , J. Lee , S. Lee , Y. Lu , Y. Lu , and F. Sun . Special Session: 2018 Low-Power Image Recognition Challenge and Beyond. In *IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)* , pages 154–157, March 2019.

Kent Gauen , Ryan Dailey , Yung-Hsiang Lu , Eunbyung Park, Wei Liu, Alexander C. Berg, and Yiran Chen. Three years of low-power image recognition challenge: Introduction to special session. In Design, Automation & Test in Europe Conference & Exhibition (DATE), pages 700–703, Dresden, Germany, March 2018. IEEE.

K. Gauen , R. Rangan , A. Mohan , Y. Lu , W. Liu , and A. C. Berg . Low-power image recognition challenge. In *22nd Asia and South Pacific Design Automation Conference (ASP-DAC)* , pages 99–104, January 2017. ISSN: 2153-697X.

Y. Lu , A. M. Kadin , A. C. Berg , T. M. Conte , E. P. DeBenedictis , R. Garg , G. Gingade , B. Hoang , Y. Huang , B. Li , J. Liu , W. Liu , H. Mao , J. Peng , T. Tang , E. K. Track , J. Wang , T. Wang , Y. Wang , and J. Yao . Rebooting computing and low-power image recognition challenge. In *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)* , pages 927–932, 2015.

Yung-Hsiang Lu . Low-power image recognition. Nature Machine Intelligence, 1(4):199–199, April 2019. Number: 4 Publisher: Nature Publishing Group.

Yung-Hsiang Lu , Alexander C. Berg , and Yiran Chen . Low-Power Image Recognition Challenge. AI Magazine, 39(2):87–88, July 2018. Number: 2.

J. Wu , C. Leng , Y. Wang , Q. Hu , and J. Cheng . Quantized Convolutional Neural Networks for Mobile Devices. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4820–4828, June 2016. ISSN: 1063-6919.

Markus Nagel , Rana Ali Amjad , Mart van Baalen , Christos Louizos , and Tijmen Blankevoort . Up or Down? Adaptive Rounding for Post-Training Quantization. *arXiv:2004.10568 [cs, stat]*, June 2020. arXiv: 2004.10568.

M. Nagel , M. V. Baalen , T. Blankevoort , and M. Welling . Data-Free Quantization Through Weight Equalization and Bias Correction. In *IEEE/CVF International Conference on Computer Vision (ICCV)* , pages 1325–1334, October 2019. ISSN: 2380-7504.

M. Sandler , A. Howard , M. Zhu , A. Zhmoginov , and L. Chen . MobileNetV2: Inverted Residuals and Linear Bottlenecks. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4510–4520, June 2018. ISSN: 2575-7075.

K. He , X. Zhang , S. Ren , and J. Sun . Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* , pages 770–778, June 2016. ISSN: 1063-6919.

Barret Zoph , Vijay Vasudevan , Jonathon Shlens , and Quoc V. Le . Learning Transferable Architectures for Scalable Image Recognition. *arXiv:1707.07012 [cs, stat]*, April 2018. arXiv: 1707.07012.

Min Lin , Qiang Chen , and Shuicheng Yan . Network In Network. *arXiv:1312.4400 [cs]*, March 2014. arXiv: 1312.4400.

Matthieu Courbariaux , Yoshua Bengio , and Jean-Pierre David . Training deep neural networks with low precision multiplications. *arXiv:1412.7024 [cs]*, September 2015. arXiv: 1412.7024.

Naigang Wang , Jungwook Choi , Daniel Brand , Chia-Yu Chen , and Kailash Gopalakrishnan . Training Deep Neural Networks with 8-bit Floating Point Numbers. arXiv:1812.08011 [cs, stat], December 2018. arXiv: 1812.08011.

A. Goel , Z. Liu , and R. D. Blanton . CompactNet: High Accuracy Deep Neural Network Optimized for On-Chip Implementation. In *2018 IEEE International Conference on Big Data (Big Data)* , pages 4723–4729, December 2018.

Ruizhou Ding , Zeye Liu , Ting-Wu Chin , Diana Marculescu , and R. D. (Shawn) Blanton . FLightNNs: Lightweight Quantized Deep Neural Networks for Fast and Accurate Inference. In Proceedings of the 56th Annual Design Automation Conference, DAC, pages 1–6, New York, NY, USA, June. Association for Computing Machinery.

Matthieu Courbariaux , Itay Hubara , Daniel Soudry , Ran El-Yaniv , and Yoshua Bengio . Binarized Neural Networks: Training Deep Neural Networks with Weights and Activations Constrained to +1 or −1. *arXiv:1602.02830 [cs]*, February 2016. arXiv: 1602.02830.

Mohammad Rastegari , Vicente Ordonez , Joseph Redmon , and Ali Farhadi . XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks. *arXiv:1603.05279 [cs]*, March 2016. arXiv: 1603.05279.

Christos Louizos , Matthias Reisser , Tijmen Blankevoort , Efstratios Gavves , and Max Welling . Relaxed Quantization for Discretized Neural Networks. *arXiv:1810.01875 [cs, stat]*, October 2018. arXiv: 1810.01875.

Yaohui Cai , Zhewei Yao , Zhen Dong , Amir Gholami , Michael W. Mahoney , and Kurt Keutzer . ZeroQ: A Novel Zero Shot Quantization Framework. pages 13169–13178, 2020.

Shuchang Zhou , Yuxin Wu , Zekun Ni , Xinyu Zhou , He Wen , and Yuheng Zou . DoReFa-Net: Training Low Bitwidth Convolutional Neural Networks with Low Bitwidth Gradients. *arXiv:1606.06160 [cs]*, February 2018. arXiv: 1606.06160.

Yann LeCun , John S. Denker , and Sara A. Solla . Optimal Brain Damage. In D. S. Touretzky , editor, Advances in Neural Information Processing Systems 2, pages 598–605. Morgan-Kaufmann, 1990.

B. Hassibi , D.G. Stork , and G.J. Wolff . Optimal Brain Surgeon and general network pruning. In IEEE International Conference on Neural Networks, pages 293–299, San Francisco, CA, USA, 1993. IEEE.

Xiaohan Ding , Guiguang Ding , Yuchen Guo , Jungong Han , and Chenggang Yan . Approximated Oracle Filter Pruning for Destructive CNN Width Optimization. January 2019.

Song Han , Jeff Pool , John Tran , and William J. Dally . Learning both weights and connections for efficient neural networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, NIPS, page 1135â€“1143, Cambridge, MA, USA, 2015. MIT Press.

Alex Krizhevsky , Ilya Sutskever , and Geoffrey E. Hinton . ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems, 25:1097–1105, 2012.

Ruichi Yu , Ang Li , Chun-Fu Chen , Jui-Hsin Lai , Vlad I. Morariu , Xintong Han , Mingfei Gao , Ching-Yung Lin , and Larry S. Davis . NISP: Pruning Networks Using Neuron Importance Score Propagation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9194–9203, Salt Lake City, UT, June 2018. IEEE.

Suraj Srinivas and R. Venkatesh Babu . Data-free parameter pruning for Deep Neural Networks. *arXiv:1507.06149 [cs]*, July 2015. arXiv: 1507.06149.

P. Panda , A. Ankit , P. Wijesinghe , and K. Roy . FALCON: Feature Driven Selective Classification for Energy-Efficient Image Recognition. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 36(12):2017–2029, December 2017. Conference Name: IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems.

P. Panda and K. Roy . Semantic driven hierarchical learning for energy-efficient image classification. In Design, Automation Test in Europe Conference Exhibition (DATE), pages 1582–1587, March 2017. ISSN: 1558-1101.

Deboleena Roy , Priyadarshini Panda , and Kaushik Roy . Tree-CNN: A hierarchical Deep Convolutional Neural Network for incremental learning. Neural Networks, 121:148–160, January 2020.

Abhinav Goel , Caleb Tung , Sara Aghajanzadeh , Isha Ghodgaonkar , Shreya Ghosh , George K. Thiruvathukal , and Yung-Hsiang Lu . Low-power object counting with hierarchical neural networks. In *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design* , ISLPED, page 163â€“168, New York, NY, USA, 2020. Association for Computing Machinery.

Yu Cheng , Duo Wang , Pan Zhou , and Tao Zhang . A Survey of Model Compression and Acceleration for Deep Neural Networks. *arXiv:1710.09282 [cs]*, June 2020. arXiv: 1710.09282.

Wei Wen , Chunpeng Wu , Yandan Wang , Yiran Chen , and Hai Li . Learning Structured Sparsity in Deep Neural Networks. In *Advances in Neural Information Processing Systems 29* , pages 2074–2082. 2016.

Pruning Tutorial â€” PyTorch Tutorials 1.7.1 documentation.

Tejalal Choudhary , Vipul Mishra , Anurag Goswami , and Jagannathan Sarangapani . A comprehensive survey on model compression and acceleration. Artificial Intelligence Review, 53(7):5113–5155, October 2020.

Hongyang Li , Wanli Ouyang , and Xiaogang Wang . Multi-Bias Non-linear Activation in Deep Neural Networks. arXiv:1604.00676 [cs], April 2016. arXiv: 1604.00676.

Andrew Howard , Mark Sandler , Grace Chu , Liang-Chieh Chen , Bo Chen , Mingxing Tan , Weijun Wang , Yukun Zhu , Ruoming Pang , Vijay Vasudevan , Quoc V. Le , and Hartwig Adam . Searching for MobileNetV3. *arXiv:1905.02244 [cs]*, November 2019. arXiv: 1905.02244.

G. Huang , S. Liu , L. v d Maaten , and K. Q. Weinberger . CondenseNet: An Efficient DenseNet Using Learned Group Convolutions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* , pages 2752–2761, June 2018. ISSN: 2575-7075.

Xiangyu Zhang , Xinyu Zhou , Mengxiao Lin , and Jian Sun . ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. *arXiv:1707.01083 [cs]*, December 2017. arXiv: 1707.01083.

C. Szegedy , Wei Liu , Yangqing Jia , P. Sermanet , S. Reed , D. Anguelov , D. Erhan , V. Vanhoucke , and A. Rabinovich . Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* , pages 1–9, June 2015. ISSN: 1063-6919.

B. Wu , A. Wan , X. Yue , P. Jin , S. Zhao , N. Golmant , A. Gholaminejad , J. Gonzalez , and K. Keutzer . Shift: A Zero FLOP, Zero Parameter Alternative to Spatial Convolutions. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9127–9135, June 2018. ISSN: 2575-7075.

Han Cai , Chuang Gan , Ligeng Zhu , and Song Han . TinyTL: Reduce Memory, Not Parameters for Efficient On-Device Learning. Advances in Neural Information Processing Systems, 33, 2020.

Max Jaderberg , Andrea Vedaldi , and Andrew Zisserman . Speeding up Convolutional Neural Networks with Low Rank Expansions. In *Proceedings of the British Machine Vision Conference* , pages 88.1–88.13, Nottingham, 2014. British Machine Vision Association.

Emily Denton , Wojciech Zaremba , Joan Bruna , Yann LeCun , and Rob Fergus . Exploiting linear structure within convolutional networks for efficient evaluation. In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1, NIPS, pages 1269–1277, Cambridge, MA, USA, December 2014. MIT Press.

Tamara G. Kolda and Brett W. Bader . Tensor Decompositions and Applications. SIAM Review, 51(3):455–500, August 2009. Publisher: Society for Industrial and Applied Mathematics.

Cheng Tai , Tong Xiao , Yi Zhang , Xiaogang Wang , and Weinan E. Convolutional neural networks with low-rank regularization. *arXiv:1511.06067 [cs, stat]*, February 2016. arXiv: 1511.06067.

Vadim Lebedev , Yaroslav Ganin , Maksim Rakhuba , Ivan Oseledets , and Victor Lempitsky . Speeding-up Convolutional Neural Networks Using Fine-tuned CP-Decomposition. *arXiv:1412.6553 [cs]*, April 2015. arXiv: 1412.6553.

Huanrui Yang , Minxue Tang , Wei Wen , Feng Yan , Daniel Hu , Ang Li , Hai Li , and Yiran Chen . Learning Low-Rank Deep Neural Networks via Singular Vector Orthogonality Regularization and Singular Value Sparsification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* , pages 678–679, 2020.

Hongkai Xiong , Yuhui Xu , Yuxi Li , Shuai Zhang , Yiran Chen , Yingyong Qi , Botao Wang , Wei Wen , and Weiyao Lin . TRP: Trained Rank Pruning for Efficient Deep Neural Networks. In *International Joint Conferences on Artificial Intelligence* , volume 1, pages 977–983, July 2020. ISSN: 1045-0823.

Jose M. Alvarez and Mathieu Salzmann . Compression-aware Training of Deep Networks. Advances in Neural Information Processing Systems, 30:856–867, 2017.

Wei Wen , Cong Xu , Chunpeng Wu , Yandan Wang , Yiran Chen , and Hai Li . Coordinating Filters for Faster Deep Neural Networks. In *Proceedings of the IEEE International Conference on Computer Vision* , pages 658–666, 2017.

Xiaohan Ding , Guiguang Ding , Yuchen Guo , and Jungong Han . Centripetal SGD for Pruning Very Deep Convolutional Networks With Complicated Structure. pages 4943–4953, 2019.

Pengzhen Ren , Yun Xiao , Xiaojun Chang , Po-Yao Huang , Zhihui Li , Xiaojiang Chen , and Xin Wang . A Comprehensive Survey of Neural Architecture Search: Challenges and Solutions. *arXiv:2006.02903 [cs, stat]*, January 2021. arXiv: 2006.02903.

Barret Zoph and Quoc V. Le. Neural Architecture Search with Reinforcement Learning. arXiv:1611.01578 [cs], February 2017. arXiv: 1611.01578.

Bowen Baker , Otkrist Gupta , Nikhil Naik , and Ramesh Raskar. Designing Neural Network Architectures using Reinforcement Learning. *arXiv:1611.02167 [cs]*, March 2017. arXiv: 1611.02167.

Esteban Real , Alok Aggarwal , Yanping Huang , and Quoc V. Le. Regularized Evolution for Image Classifier Architecture Search. *arXiv:1802.01548 [cs]*, February 2019. arXiv: 1802.01548.

Mingxing Tan , Bo Chen , Ruoming Pang , Vijay Vasudevan , Mark Sandler , Andrew Howard , and Quoc V. Le. MnasNet: Platform-Aware Neural Architecture Search for Mobile. *arXiv:1807.11626 [cs]*, May 2019. arXiv: 1807.11626.

Bichen Wu , Xiaoliang Dai , Peizhao Zhang , Yanghan Wang , Fei Sun , Yiming Wu , Yuandong Tian , Peter Vajda , Yangqing Jia , and Kurt Keutzer. FBNet: Hardware-Aware Efficient ConvNet Design via Differentiable Neural Architecture Search. *arXiv:1812.03443 [cs]*, May 2019. arXiv: 1812.03443.

Han Cai , Ligeng Zhu , and Song Han. ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware. *arXiv:1812.00332 [cs, stat]*, February 2019. arXiv: 1812.00332.

Y. Chen , G. Meng , Q. Zhang , S. Xiang , C. Huang , L. Mu , and X. Wang . RENAS: Reinforced Evolutionary Neural Architecture Search. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* , pages 4782–4791, June 2019. ISSN: 2575-7075.

Hanxiao Liu , Karen Simonyan , and Yiming Yang . DARTS: Differentiable Architecture Search. *arXiv:1806.09055 [cs, stat]*, April 2019. arXiv: 1806.09055.

Yuhui Xu , Lingxi Xie , Xiaopeng Zhang , Xin Chen , Guo-Jun Qi , Qi Tian , and Hongkai Xiong . PC-DARTS: Partial Channel Connections for Memory-Efficient Architecture Search. *arXiv:1907.05737 [cs]*, April 2020. arXiv: 1907.05737.

Hanlin Chen , Li'an Zhuo , Baochang Zhang , Xiawu Zheng , Jianzhuang Liu , Rongrong Ji , David Doermann , and Guodong Guo. Binarized Neural Architecture Search for Efficient Object Recognition. *arXiv:2009.04247 [cs]*, September 2020. arXiv: 2009.04247.

Y. Li , X. Jin , J. Mei , X. Lian , L. Yang , C. Xie , Q. Yu , Y. Zhou , S. Bai , and A. L. Yuille . Neural Architecture Search for Lightweight Non-Local Networks. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 10294–10303, 2020.

Mohammad Loni , Ali Zoljodi , Sima Sinaei , Masoud Daneshtalab , and Mikael SjÃ¶din . NeuroPower: Designing Energy Efficient Convolutional Neural Network Architecture for Embedded Systems. In Artificial Neural Networks and Machine Learning â€“ICANN: Theoretical Neural Computation, Lecture Notes in Computer Science, pages 208–222, Cham, 2019. Springer International Publishing.

Hieu Pham , Melody Guan , Barret Zoph , Quoc Le , and Jeff Dean . Efficient neural architecture search via parameters sharing. In Jennifer Dy and Andreas Krause , editors, *International Conference on Machine Learning* , volume 80 of *Proceedings of Machine Learning Research*, pages 4095–4104. PMLR, 10–15 Jul 2018.

Dimitrios Stamoulis , Ruizhou Ding , Di Wang , Dimitrios Lymberopoulos , Bodhi Priyantha , Jie Liu , and Diana Marculescu . Single-Path NAS: Designing Hardware-Efficient ConvNets in Less Than 4 Hours. In Ulf Brefeld , Elisa Fromont , Andreas Hotho , Arno Knobbe , Marloes Maathuis , and CÃ©line Robardet , editors, *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pages 481–497, Cham, 2020. Springer International Publishing.

Zeyuan Allen-Zhu , Yuanzhi Li , and Yingyu Liang . Learning and Generalization in Overparameterized Neural Networks, Going Beyond Two Layers. September 2019.

Alon Brutzkus and Amir Globerson . Why do Larger Models Generalize Better? A Theoretical Perspective via the XOR Problem. In *International Conference on Machine Learning* , pages 822–830. PMLR, May 2019. ISSN: 2640-3498.

Zhuozhuo Tu , Fengxiang He , and Dacheng Tao . Understanding Generalization in Recurrent Neural Networks. In *International Conference on Learning Representations* , April 2020.

Jianping Gou , Baosheng Yu , Stephen John Maybank , and Dacheng Tao . Knowledge Distillation: A Survey. *arXiv:2006.05525 [cs, stat]*, October 2020. arXiv: 2006.05525.

Jimmy Ba and Rich Caruana . Do Deep Nets Really Need to be Deep? Advances in Neural Information Processing Systems, 27:2654–2662, 2014.

Cristian Bucilua , Rich Caruana , and Alexandru Niculescu-Mizil . Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* , KDD ‘06, pages 535–541, New York, NY, USA, August 2006. Association for Computing Machinery.

Geoffrey Hinton , Oriol Vinyals , and Jeff Dean . Distilling the Knowledge in a Neural Network. *arXiv:1503.02531 [cs, stat]*, March 2015. arXiv: 1503.02531.

X. Liu , X. Wang , and S. Matwin . Improving the Interpretability of Deep Neural Networks with Knowledge Distillation. In *IEEE International Conference on Data Mining Workshops (ICDMW)* , pages 905–912, November 2018. ISSN: 2375-9259.

Jang Hyun Cho and Bharath Hariharan . On the Efficacy of Knowledge Distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* , pages 4794–4802, October 2019.

Byeongho Heo , Minsik Lee , Sangdoo Yun , and Jin Young Choi . Knowledge Transfer via Distillation of Activation Boundaries Formed by Hidden Neurons. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01):3779–3787, July 2019. Number: 01.

S. Ahn , S. X. Hu , A. Damianou , N. D. Lawrence , and Z. Dai . Variational Information Distillation for Knowledge Transfer. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 9155–9163, June 2019. ISSN: 2575-7075.

J. Yim , D. Joo , J. Bae , and J. Kim . A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* , pages 7130–7138, July 2017. ISSN: 1063-6919.

Seyed Iman Mirzadeh , Mehrdad Farajtabar , Ang Li , Nir Levine , Akihiro Matsukawa , and Hassan Ghasemzadeh . Improved Knowledge Distillation via Teacher Assistant. Proceedings of the AAAI Conference on Artificial Intelligence, 34(04):5191–5198, April 2020. Number: 04.

Theodore S. Nowak and Jason J. Corso . Deep Net Triage: Analyzing the Importance of Network Layers via Structural Compression. *arXiv:1801.04651 [cs]*, March 2018. arXiv: 1801.04651.

Mingxing Tan and Quoc V. Le . EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv:1905.11946 [cs, stat]*, September 2020. arXiv: 1905.11946.

Han Cai , Chuang Gan , Tianzhe Wang , Zhekai Zhang , and Song Han . Once-for-all: Train one network and specialize it for efficient deployment. *arXiv preprint arXiv:1908.09791*, 2019.

Jian-Hao Luo , Hao Zhang , Hong-Yu Zhou , Chen-Wei Xie , Jianxin Wu , and Weiyao Lin . Thinet: Pruning cnn filters for a thinner net. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(10):2525–2538, 2019.

Yuhui Xu , Yuxi Li , Shuai Zhang , Wei Wen , Botao Wang , Yingyong Qi , Yiran Chen , Weiyao Lin , and Hongkai Xiong . Trained Rank Pruning for Efficient Deep Neural Networks. *arXiv preprint arXiv:1812.02402*, 2018.

Kuan Wang , Zhijian Liu , Yujun Lin , Ji Lin , and Song Han . HAQ: Hardware-Aware Automated Quantization With Mixed Precision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8604–8612, Long Beach, CA, USA, June 2019. IEEE.

Tien-Ju Yang , Yu-Hsin Chen , Joel Emer , and Vivienne Sze . A method to estimate the energy consumption of deep neural networks. In Asilomar Conference on Signals, Systems, and Computers, pages 1916–1920, 2017.

Chenzhuo Zhu , Song Han , Huizi Mao , and William J. Dally . Trained Ternary Quantization. *arXiv:1612.01064 [cs]*, February 2017. arXiv: 1612.01064.

Matthieu Courbariaux , Yoshua Bengio , and Jean-Pierre David . BinaryConnect: Training Deep Neural Networks with binary weights during propagations. *arXiv:1511.00363 [cs]*, April 2016. arXiv: 1511.00363.

Kuan Wang , Zhijian Liu , Yujun Lin , Ji Lin , and Song Han . Haq: Hardware-aware automated quantization with mixed precision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , pages 8612–8620, 2019.

Luis Guerra , Bohan Zhuang , Ian Reid , and Tom Drummond . Automatic Pruning for Quantized Neural Networks. arXiv:2002.00523 [cs], February 2020. arXiv: 2002.00523.

F. Tung and G. Mori . CLIP-Q: Deep Network Compression Learning by In-parallel Pruning-Quantization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* , pages 7873–7882, June 2018. ISSN: 2575-7075.

Xiaofan Lin , Cong Zhao , and Wei Pan . Towards accurate binary convolutional neural network. In International Conference on Neural Information Processing Systems, NIPS, pages 344–352, Red Hook, NY, USA, December 2017. Curran Associates Inc.

Asit Mishra and Debbie Marr . Apprentice: Using Knowledge Distillation Techniques To Improve Low-Precision Network Accuracy. *arXiv:1711.05852 [cs]*, November 2017. arXiv: 1711.05852.

Jangho Kim , Yash Bhalgat , Jinwon Lee , Chirag Patel , and Nojun Kwak . QKD: Quantization-aware Knowledge Distillation. *arXiv:1911.12491 [cs]*, November 2019. arXiv: 1911.12491.

Sridhar Swaminathan , Deepak Garg , Rajkumar Kannan , and Frederic Andres . Sparse low rank factorization for deep neural network compression. Neurocomputing, 398:185–196, July 2020.

Michael Zhu and Suyog Gupta . To prune, or not to prune: exploring the efficacy of pruning for model compression. *arXiv:1710.01878 [cs, stat]*, November 2017. arXiv: 1710.01878.

Mingxing Tan , Ruoming Pang , and Quoc V. Le . EfficientDet: Scalable and Efficient Object Detection. pages 10781–10790, 2020.

Alex Krizhevsky, Geoffrey . Learning multiple layers of features from tiny images. Technical Report TR-2009, University of Toronto, Toronto, 2009.

Yuval Netzer , Tao Wang , Adam Coates , Alessandro Bissacco , Bo Wu , and Andrew Y Ng . Reading digits in natural images with unsupervised feature learning. 2011.

Gregory Cohen , Saeed Afshar , Jonathan Tapson , and Andre van Schaik . EMNIST: an extension of MNIST to handwritten letters. February 2017.

Gregory Griffin , Alex Holub , and Pietro Perona . Caltech-256 object category dataset. 2007.

J. Deng , W. Dong , R. Socher , L. Li , Kai Li , and Li Fei-Fei . ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* , pages 248–255, June 2009. ISSN: 1063-6919.

Tsung-Yi Lin , Michael Maire , Serge Belongie , James Hays , Pietro Perona , Deva Ramanan , Piotr Dollar , and C. Lawrence Zitnick . Microsoft COCO: Common Objects in Context. In Computer Vision â€“ECCV, Lecture Notes in Computer Science, pages 740–755, Cham, 2014. Springer International Publishing.

M. Everingham , L. Van Gool , C. K. I. Williams , J. Winn , and A. Zisserman . The PASCAL Visual Object Classes Challenge 2012 (VOC) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

Gabriel J. Brostow , Jamie Shotton , Julien Fauqueur , and Roberto Cipolla . Segmentation and Recognition Using Structure from Motion Point Clouds. In David Forsyth , Philip Torr , and Andrew Zisserman , editors, Computer Vision â€“ECCV 2008, Lecture Notes in Computer Science, pages 44–57, Berlin, Heidelberg, 2008. Springer.

Liang Zheng , Liyue Shen , Lu Tian , Shengjin Wang , Jingdong Wang , and Qi Tian . Scalable Person Re-Identification: A Benchmark. pages 1116–1124, 2015.

Liang Zheng , Zhi Bie , Yifan Sun , Jingdong Wang , Chi Su , Shengjin Wang , and Qi Tian . MARS: A Video Benchmark for Large-Scale Person Re-Identification. In Computer Vision â€“ECCV 2016, pages 868–884, 2016.

Peng Wang , Bingliang Jiao , Lu Yang , Yifei Yang , Shizhou Zhang , Wei Wei , and Yanning Zhang . Vehicle Re-Identification in Aerial Imagery: Dataset and Approach. pages 460–469, 2019.

A Geiger , P Lenz , C Stiller , and R Urtasun . Vision meets robotics: The KITTI dataset. International Journal of Robotics Research, 32(11):1231–1237, September 2013.

Marius Cordts , Mohamed Omran , Sebastian Ramos , Timo Rehfeld , Markus Enzweiler , Rodrigo Benenson , Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. pages 3213–3223, 2016.

Kaiming He , Xiangyu Zhang , Shaoqing Ren , and Jian Sun . Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* , pages 770–778, 2016.

Dario Amodei , Sundaram Ananthanarayanan , Rishita Anubhai , Jingliang Bai , Eric Battenberg , Carl Case , Jared Casper , Bryan Catanzaro , Qiang Cheng , Guoliang Chen , Jie Chen , Jingdong Chen , Zhijie Chen , Mike Chrzanowski , Adam Coates , Greg Diamos, Ke Ding , Niandong Du , Erich Elsen , Jesse Engel , Weiwei Fang , Linxi Fan , Christopher Fougner , Liang Gao, Caixia Gong, Awni Hannun , Tony Han , Lappi Vaino Johannes , Bing Jiang , Cai Ju , Billy Jun , Patrick LeGresley , Libby Lin , Junjie Liu , Yang Liu , Weigao Li , Xiangang Li , Dongpeng Ma , Sharan Narang , Andrew Ng , Sherjil Ozair , Yiping Peng , Ryan Prenger , Sheng Qian , Zongfeng Quan , Jonathan Raiman , Vinay Rao , Sanjeev Satheesh , David Seetapun , Shubho Sengupta , Kavya Srinet , Anuroop Sriram , Haiyuan Tang , Liliang Tang , Chong Wang , Jidong Wang , Kaifu Wang , Yi Wang , Zhijian Wang , Zhiqian Wang , Shuang Wu , Likai Wei , Bo Xiao , Wen Xie , Yan Xie , Dani Yogatama , Bin Yuan , Jun Zhan , and Zhenyao Zhu . Deep speech 2: end-to-end speech recognition in English and mandarin. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48* , ICML'16, pages 173–182, New York, NY, USA, June 2016. JMLR.org .

K. Guo , Shulin Zeng , J. Yu , Y. Wang , and H. Yang . [dl] a survey of fpga-based neural network inference accelerators. ACM Trans. Reconfigurable Technol. Syst., 12:2:1–2:26, 2019.

Christian Szegedy , Wei Liu , Yangqing Jia , Pierre Sermanet , Scott Reed , Dragomir Anguelov , Dumitru Erhan , Vincent Vanhoucke , and Andrew Rabinovich . Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* , pages 1–9, 2015.

Jifeng Dai , Yi Li , Kaiming He , and Jian Sun . R-fcn: Object detection via region-based fully convolutional networks. In Advances in neural information processing systems, pages 379–387,

2016.
Haoxiang Li , Zhe Lin , Xiaohui Shen , Jonathan Brandt , and Gang Hua . A convolutional neural network cascade for face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* , pages 5325–5334, 2015.
Jure Zbontar and Yann LeCun . Stereo matching by training a convolutional neural network to compare image patches. Journal of Machine Learning Research, 17:1–32, 2016.
Jiantao Qiu , Jie Wang , Song Yao , Kaiyuan Guo , Boxun Li , Erjin Zhou , Jincheng Yu , Tianqi Tang , Ningyi Xu , Sen Song , et al. Going deeper with embedded fpga platform for convolutional neural network. In *Proceedings of the 2016 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays* , pages 26–35, 2016.
Bernard Bosi , Guy Bois , and Yvon Savaria . Reconfigurable pipelined 2-d convolvers for fast digital signal processing. VLSI, 7(3):299–308, 1999.
Srimat Chakradhar , Murugan Sankaradas , Venkata Jakkula , et al. A dynamically configurable coprocessor for conv olutional neural networks. In ACM SIGARCH Computer Architecture News, volume 38, pages 247–257. ACM, 2010.
Vinayak Gokhale , Jonghoon Jin , Aysegul Dundar , et al. A 240 g-ops/s mobile coprocessor for deep neural networks. In CVPRW, pages 682–687, 2014.
Chen Zhang , Peng Li , Guangyu Sun , Yijin Guan , Bingjun Xiao , and Jason Cong . Optimizing FPGA-based Accelerator Design for Deep Convolutional Neural Networks. In *Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays* , FPGA '15, pages 161–170, New York, NY, USA, February 2015. Association for Computing Machinery.
Huimin Li , Xitian Fan , Li Jiao , Wei Cao , Xuegong Zhou , and Lingli Wang . A high performance fpga-based accelerator for large-scale convolutional neural networks. International Conference on Field Programmable Logic and Applications (FPL), 2016, pp. 1–9, doi: 10.1109/FPL.2016.7577308.
Xiaowei Xu , Xinyi Zhang , Bei Yu , X Sharon Hu , Christopher Rowen , Jingtong Hu , and Yiyu Shi . Dac-sdc low power object detection challenge for uav applications. IEEE transactions on pattern analysis and machine intelligence, 2019.
Song Han , Huizi Mao , and William J Dally . Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *International Conference on Learning Representations (ICLR)* , 2016.
Wei Wen , Chunpeng Wu , Yandan Wang , Yiran Chen , and Hai Li . Learning structured sparsity in deep neural networks. *neural information processing systems* , pages 2074–2082, 2016.
Hao Li , Asim Kadav , Igor Durdanovic , Hanan Samet , and Hans Peter Graf . Pruning filters for efficient convnets. arXiv preprint arXiv:1608.08710, 2016.
Hengyuan Hu , Rui Peng , Yu-Wing Tai , and Chi-Keung Tang . Network trimming: A data-driven neuron pruning approach towards efficient deep architectures. *arXiv preprint arXiv:1607.03250*, 2016.
Jian-Hao Luo , Jianxin Wu , and Weiyao Lin . Thinet: A filter level pruning method for deep neural network compression. In *Proceedings of the IEEE international conference on computer vision* , pages 5058–5066, 2017.
Geoffrey Hinton , Oriol Vinyals , and Jeff Dean . Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
Shitao Tang , Litong Feng , Wenqi Shao , Zhanghui Kuang , Wei Zhang , and Yimin Chen . Learning efficient detector with semi-supervised adaptive distillation. *arXiv preprint arXiv:1901.00366*, 2019.
Yihui He , Ji Lin , Zhijian Liu , Hanrui Wang , Li-Jia Li , and Song Han . Amc: Automl for model compression and acceleration on mobile devices. In *Proceedings of the European Conference on Computer Vision (ECCV)* , pages 784–800, 2018.
Tien-Ju Yang , Andrew Howard , Bo Chen , Xiao Zhang , Alec Go , Mark Sandler , Vivienne Sze , and Hartwig Adam . Netadapt: Platform-aware neural network adaptation for mobile applications. In *Proceedings of the European Conference on Computer Vision (ECCV)* , pages 285–300, 2018.
Zechun Liu , Haoyuan Mu , Xiangyu Zhang , Zichao Guo , Xin Yang , Kwang-Ting Cheng , and Jian Sun . Metapruning: Meta learning for automatic neural network channel pruning. In *Proceedings of the IEEE International Conference on Computer Vision* , pages 3296–3305, 2019.

Ning Liu, Xiaolong Ma, Zhiyuan Xu, Yanzhi Wang, Jian Tang , and Jieping Ye . Autocompress: An automatic dnn structured pruning framework for ultra-high compression rates. In *Proceedings of the AAAI Conference on Artificial Intelligence* , 2020.

Hieu Pham , Melody Y Guan , Barret Zoph , Quoc V Le , and Jeff Dean . Efficient neural architecture search via parameter sharing. In *International Conference on Machine Learning (ICML)* , 2018.

Shulin Zeng , Hanbo Sun , Y. Xing , Xuefei Ning , Y. Shan , X. Chen , Yu Wang , and Hua zhong Yang . Black box search space profiling for accelerator-aware neural architecture search. Asia and South Pacific Design Automation Conference (ASP-DAC), pages 518–523, 2020.

Zichao Guo , Xiangyu Zhang , Haoyuan Mu , Wen Heng , Zechun Liu , Yichen Wei , and Jian Sun . Single path one-shot neural architecture search with uniform sampling. *arXiv preprint arXiv:1904.00420*, 2019.

Chenxi Liu , Barret Zoph , Maxim Neumann , Jonathon Shlens , Wei Hua , Li-Jia Li , Li Fei-Fei , Alan Yuille , Jonathan Huang , and Kevin Murphy . Progressive neural architecture search. In *Proceedings of the European Conference on Computer Vision (ECCV)* , pages 19–34, 2018.

Hanxiao Liu , Karen Simonyan , and Yiming Yang . Darts: Differentiable architecture search. In ICLR, 2019.

Mingxing Tan and Quoc V Le . Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv: 1905.11946, 2019.

Barret Zoph and Quoc V Le . Neural architecture search with reinforcement learning. arXiv preprint arXiv:1611.01578, 2016.

Barret Zoph , Vijay Vasudevan , Jonathon Shlens , and Quoc V Le . Learning transferable architectures for scalable image recognition. In CVPR, 2018.

Yukang Chen , Tong Yang , Xiangyu Zhang , Gaofeng Meng , Xinyu Xiao , and Jian Sun . Detnas: Backbone search for object detection. In NeurIPS, pages 6638–6648, 2019.

Mingxing Tan , Ruoming Pang , and Quoc V Le . Efficientdet: Scalable and efficient object detection. arXiv preprint arXiv:1911.09070, 2019.

Wuyang Chen , Xinyu Gong , Xianming Liu , Qian Zhang , Yuan Li , and Zhangyang Wang . Fasterseg: Searching for faster real-time semantic segmentation. In ICLR, 2020.

Chenxi Liu , Liang-Chieh Chen , Florian Schroff , Hartwig Adam , Wei Hua , Alan L Yuille , and Li Fei-Fei . Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In CVPR, 2019.

Vladimir Nekrasov , Hao Chen , Chunhua Shen , and Ian Reid . Fast neural architecture search of compact semantic segmentation models via auxiliary cells. In CVPR, 2019.

Ronald J Williams . Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning, 8(3-4):229–256, 1992.

Esteban Real , Alok Aggarwal , Yanping Huang , and Quoc V Le . Regularized evolution for image classifier architecture search. In *Proceedings of the aaai conference on artificial intelligence* , volume 33, pages 4780–4789, 2019.

Andrew Brock , Theodore Lim , James Millar Ritchie , and Nicholas J Weston . Smash: One-shot model architecture search through hypernetworks. In ICLR, 2018.

Sirui Xie , Hehui Zheng , Chunxiao Liu , and Liang Lin . Snas: stochastic neural architecture search. *arXiv preprint arXiv:1812.09926*, 2018.

Xinbang Zhang , Zehao Huang , and Naiyan Wang . You only search once: Single shot neural architecture search via direct sparse optimization. *arXiv preprint arXiv:1811.01567*, 2018.

Han Cai , Ligeng Zhu , and Song Han . ProxylessNAS: Direct neural architecture search on target task and hardware. In ICLR, 2019.

Bichen Wu , Xiaoliang Dai , Peizhao Zhang , Yanghan Wang , Fei Sun , Yiming Wu , Yuandong Tian , Peter Vajda , Yangqing Jia , and Kurt Keutzer . Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In CVPR, 2019.

Gabriel Bender , Pieter-Jan Kindermans , Barret Zoph , Vijay Vasudevan , and Quoc V. Le . Understanding and simplifying one-shot architecture search. In ICML, 2018.

Xiangxiang Chu , Bo Zhang , Ruijun Xu , and Jixiang Li . Fairnas: Rethinking evaluation fairness of weight sharing neural architecture search. arXiv preprint arXiv:1907.01845, 2019.

George Adam and Jonathan Lorraine . Understanding neural architecture search techniques. arXiv preprint arXiv:1904. *00438* , 2019.

Mark Fleischer . The measure of pareto optima applications to multi-objective metaheuristics. In *International Conference on Evolutionary Multi-Criterion Optimization* , pages 519–533.

Springer, 2003.
Martn Abadi , Ashish Agarwal , Paul Barham , Eugene Brevdo , Zhifeng Chen , Craig Citro , Greg S Corrado , Andy Davis , Jeffrey Dean , Matthieu Devin , et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603 . 04467* , 2016.
Mark Sandler , Andrew Howard , Menglong Zhu , Andrey Zhmoginov , and Liang-Chieh Chen . Mobilenetv2: Inverted residuals and linear bottlenecks. In CVPR, 2018.
Shan You , Tao Huang , Mingmin Yang , Fei Wang , Chen Qian , and Changshui Zhang . Greedynas: Towards fast one-shot nas with greedy supernet. In CVPR, 2020.
Yiming Hu , Yuding Liang , Zichao Guo , Ruosi Wan , Xiangyu Zhang , Yichen Wei , Qingyi Gu , and Jian Sun . Angle-based search space shrinking for neural architecture search. In ECCV, 2020.
Ilija Radosavovic , Raj Prateek Kosaraju , Ross Girshick , Kaiming He , and Piotr Dollár . Designing network design spaces. In CVPR, 2020.
Xin Chen , Lingxi Xie , Jun Wu , and Qi Tian . Progressive differentiable architecture search: Bridging the depth gap between search and evaluation. In ICCV, 2019.
Shen Yan , Biyi Fang , Faen Zhang , Yu Zheng , Xiao Zeng , Mi Zhang , and Hui Xu . Hm-nas: Efficient neural architecture search via hierarchical masking. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* , 2019.
Song Han , Huizi Mao , and William J Dally . Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In ICLR, 2016.
Yihui He , Xiangyu Zhang , and Jian Sun . Channel pruning for accelerating very deep neural networks. In ICCV, 2017.
Zhuang Liu , Jianguo Li , Zhiqiang Shen , Gao Huang , Shoumeng Yan , and Changshui Zhang . Learning efficient convolutional networks through network slimming. In ICCV, 2017.
Kalyanmoy Deb . A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: Nsga-2. IEEE Trans. Evol. Comput., 6(2):182–197, 2002.
Zhichao Lu , Ian Whalen , Vishnu Boddeti , Yashesh Dhebar , Kalyanmoy Deb , Erik Goodman , and Wolfgang Banzhaf . Nsga-net: neural architecture search using multi-objective genetic algorithm. In *Proceedings of the Genetic and Evolutionary Computation Conference* , pages 419–427. ACM, 2019.
Jia Deng , Wei Dong , Richard Socher , Li Jia Li , and Fei Fei Li . Imagenet: a large-scale hierarchical image database. In CVPR, 2009.
Jie Hu , Li Shen , and Gang Sun . Squeeze-and-excitation networks. In CVPR, 2018.
Prajit Ramachandran , Barret Zoph , and Quoc V Le . Searching for activation functions. arXiv preprint arXiv:1710.05941, 2017.
Hongyi Zhang , Moustapha Cisse , Yann N Dauphin , and David Lopez-Paz . mixup: Beyond empirical risk minimization. In ICLR, 2018.
Ekin D. Cubuk , Barret Zoph , Dandelion Mane , Vijay Vasudevan , and Quoc V. Le . Autoaugment: Learning augmentation strategies from data. In CVPR, 2019.
Andrew G Howard , Menglong Zhu , Bo Chen , Dmitry Kalenichenko , Weijun Wang , Tobias Weyand , Marco Andreetto , and Hartwig Adam . Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.
Ningning Ma , Xiangyu Zhang , Hai-Tao Zheng , and Jian Sun . Shufflenet v2: Practical guidelines for efficient cnn architecture design. In ECCV, pages 116–131, 2018.
Xin Chen , Lingxi Xie , Jun Wu , and Qi Tian . Progressive darts: Bridging the optimization gap for nas in the wild. arXiv preprint arXiv:1912.10952, 2019.
Mingxing Tan , Bo Chen , Ruoming Pang , Vijay Vasudevan , Mark Sandler , Andrew Howard , and Quoc V Le . Mnasnet: Platform-aware neural architecture search for mobile. In CVPR, 2019.
Jiemin Fang , Yuzhu Sun , Qian Zhang , Yuan Li , Wenyu Liu , and Xinggang Wang . Densely connected search space for more flexible neural architecture search. In CVPR, 2020.
Dimitrios Stamoulis , Ruizhou Ding , Di Wang , Dimitrios Lymberopoulos , Bodhi Priyantha , Jie Liu , and Diana Marculescu . Single-path nas: Designing hardware-efficient convnets in less than 4 hours. arXiv preprint arXiv:1904.02877, 2019.
Andrew Howard , Mark Sandler , Grace Chu , Liang-Chieh Chen , Bo Chen , Mingxing Tan , Weijun Wang , Yukun Zhu , Ruoming Pang , Vijay Vasudevan , et al. Searching for mobilenetv3. In ICCV, 2019.

Maurice G Kendall . A new measure of rank correlation. Biometrika, 30(1/2):81–93, 1938.
A. G. Howard , M. Zhu , B. Chen , D. Kalenichenko , W. Wang , T. Weyand , M. Andreetto , and H. Adam . Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.
M. Sandler , A.G. Howard , M. Zhu , A. Zhmoginov , and L.C. Chen . Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR *)*, 2018.
M. Tan , B. Chen , R. Pang , V. Vasudevan , and Q. V. Le . Mnasnet: Platform-aware neural architecture search for mobile. arXiv preprint arXiv:1807.11626, 2018.
J. H. Lee , S. Ha , S. Choi , W. Lee , and S. Lee . Quantization for rapid deployment of deep neural networks. arXiv preprint arXiv:1810.05488, 2018.
B. Jacob , S. Kligys , B. Chen , M. Zhu , M. Tang , A. Howard , H. Adam , and D. Kalenichenko . Quantization and training of neural networks for efficient integer-arithmetic only inference. In Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
G. Hinton , O. Vinyals , and J. Dean . Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.
A. Mishra and D. Marr . Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy. arXiv preprint arXiv:1711.05852, 2017.
A. Mishra , E. Nurvitadhi , J. J. Cook , and D. Marr . Wrpn: Wide reduced-precision networks. arXiv preprint arXiv:1709.01134, 2017.
NVIDIA TensorRT SDK. https://developer.nvidia.com/tensorrt, 2018.
M. Abadi , A. Agarwal , P. Barham , E. Brevdo , Z. Chen , C. Citro , G. S. Corrado , A. Davis , J. Dean , M. Devin , S. Ghemawat , I. Goodfellow , A. Harp , G. Irving , M. Isard , Y. Jia , R. Jozefowicz , L. Kaiser , M. Kudlur , J. Levenberg , D. Mane , R. Monga , S. Moore , D. Murray , C. Olah , M. Schuster , J. Shlens , B. Steiner , I. Sutskever , K. Talwar , P. Tucker , V. Vanhoucke , V. Vasudevan , F. Viegas , O. Vinyals , P. Warden , M. Wattenberg , M. Wicke , Y. Yu , and X. Zheng . Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467, 2016.
Neta Zmora , Guy Jacob , Lev Zlotnik , Bar Elharar , and Gal Novik . Neural network distiller: A python package for dnn compression research. arXiv preprint arXiv:1910.12232, 2019.
M. Courbariaux , Y. Bengio , and J. David . Training deep neural networks with low precision multiplications. In *International Conference on Learning Representations (ICLR 2015)* , 2015.
I. Hubara , M. Courbariaux , D. Soudry , R. El-Yaniv , and Y. Bengio . Binarized neural networks. In Advances in Neural Information Processing Systems (NIPS *)*, pages 4107–4115, 2016.
M. Rastegari , V. Ordonez , J. Redmon , and A. Farhadi . Xnor-net: Imagenet classification using binary convolutional neural networks. In European Conference on Computer Vision (ECCV), pages 525–542. Springer, 2016.
Sh. Zhou , Y. Wu , Z. Ni , X. Zhou , H. Wen , and Y. Zou . Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. arXiv preprint arXiv:1606. *06160* , 2016.
Y. Bengio , N. Leonard , and A. C. Courville . Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308. *3432* , 2013.
M.D. McDonnell . Training wide residual networks for deployment using a single bit for each weight. In *International Conference on Learning Representations (ICLR)* , 2018.
Sh. Zhu , X. Dong , and H. Su . Binary ensemble neural network: more bits per network or more networks per bit? *arXiv preprint arXiv:1806* . *07550* , 2018.
Ch. Baskin , N. Liss , Y. Chai , E. Zheltonozhskii , E. Schwartz , R. Giryes , A. Mendelson , and A. M. Bronstein . Nice: Noise injection and clamping estimation for neural network quantization. arXiv preprint arXiv:1810. *00162* , 2018.
O. Russakovsky , J. Deng , H. Su , J. Krause , S. Satheesh , S. Ma , Z. Huang , A. Karpathy , A. Khosla , M. Bernstein , A. C. Berg , and L. Fei-Fei . Imagenet large scale visual recognition challenge. arXiv preprint arXiv:1409. *0575* , 2014.
S. Ioffe and C. Szegedy . Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML 2015)* , 2015.
D. P. Kingma and J. L. Ba . Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)* , 2015.

T. Sheng , C. Feng , S. Zhuo , X. Zhang , L. Shen , and M. Aleksic . A quantization-friendly separable convolution for mobilenets. arXiv preprint arXiv:1803. *08607* , 2018.
https://github.com/tensorflow/tensorflow/blob/61c6c84964b4aec-80aeace187aab8cb2c3e55a72/tensorflow/lite/g3doc/models.md.
https://git.io/jo4gj.
Suyog Gupta and Berkin Akin . Accelerator-aware neural network design using automl. 2020.
Weiwen Jiang , Xinyi Zhang , Edwin Hsing-Mean Sha , Lei Yang , Qingfeng Zhuge , Yiyu Shi , and Jingtong Hu . Accuracy vs. efficiency: Achieving both through fpga-implementation aware neural architecture search. In *Proceedings of the 56th Annual Design Automation Conference, DAC, Las Vegas, NV, USA* , page 5. ACM, 2019.
Weiwei Chen , Ying Wang , Shuang Yang , Chen Liu , and Lei Zhang . Towards best-effort approximation: applying nas to general-purpose approximate computing. In Design, Automation & Test in Europe Conference & Exhibition (DATE), pages 1315–1318. IEEE, 2020.
Xiandong Zhao , Ying Wang , Xuyi Cai , Cheng Liu , and Lei Zhang . Linear symmetric quantization of neural networks for low-precision integer hardware. In *International Conference on Learning Representations* , 2019.
Ying Wang , Huawei Li , Long Cheng , and Xiaowei Li . A qos-qor aware cnn accelerator design approach. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 38(11):1995–2007, 2018.
Tianshi Chen , Zidong Du , Ninghui Sun , Jia Wang , Chengyong Wu , Yunji Chen , and Olivier Temam . Diannao: a small-footprint high-throughput accelerator for ubiquitous machine-learning. Architectural support for programming languages and operating systems, 49(4):269–284, 2014.
Y. Chen , J. Emer , and V. Sze . Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks. IEEE Micro, pages 1–1, 2018.
Ying Wang , Jie Xu , Yinhe Han , Huawei Li , and Xiaowei Li . Deepburning: automatic generation of fpga-based learning accelerators for the neural network family. In *Proceedings of the 53rd Annual Design Automation Conference* , page 110. ACM, 2016.
Lili Song , Ying Wang , Yinhe Han , Xin Zhao , Bosheng Liu , and Xiaowei Li . C-brain: A deep learning accelerator that tames the diversity of cnns through adaptive data-level parallelization. In *Proceedings of the 53rd Annual Design Automation Conference* , pages 1–6, 2016.
Weiwei Chen , Ying Wang , Shuang Yang , Chen Liu , and Lei Zhang. You only search once: a fast automation framework for single-stage dnn/accelerator co-design. In Design, Automation & Test in Europe Conference & Exhibition (DATE), pages 1283–1286. IEEE, 2020.
Ting Hu , Ying Wang , Lei Zhang , and Tingting He . A tightly-coupled light-weight neural network processing units with risc-v core. 2017.
Erik Debenedictis , Yung-Hsiang Lu , Alan Kadin , Alexander Berg , Thomas Conte , Rachit Garg , Ganesh Gingade , Bichlien Hoang , Yongzhen Huang , Boxun Li , et al. Rebooting computing and low-power image recognition challenge. Technical report, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2016.
Srimat T Chakradhar and Anand Raghunathan . Best-effort computing: Re-thinking parallel software and hardware. In *Design Automation Conference* , pages 865–870. IEEE, 2010.
Cheng Wang , Ying Wang , Yinhe Han , Lili Song , Zhenyu Quan , Jiajun Li , and Xiaowei Li . Cnn-based object detection solutions for embedded heterogeneous multicore socs. In *Asia and South Pacific Design Automation Conference (ASP-DAC)* , pages 105–110. IEEE, 2017.
Jasper RR Uijlings , Koen EA Van De Sande , Theo Gevers , and Arnold WM Smeulders . Selective search for object recognition. International journal of computer vision, 104(2):154–171, 2013.
Ming-Ming Cheng , Ziming Zhang , Wen-Yan Lin , and Philip Torr . Bing: Binarized normed gradients for objectness estimation at 300fps. In *Proceedings of the IEEE conference on computer vision and pattern recognition* , pages 3286–3293, 2014.
Ross Girshick , Jeff Donahue , Trevor Darrell , and Jitendra Malik . Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* , pages 580–587, 2014.
Pierre Sermanet , David Eigen , Xiang Zhang , Michaël Mathieu , Rob Fergus , and Yann LeCun . Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312. *6229* , 2013.

Kaiming He , Xiangyu Zhang , Shaoqing Ren , and Jian Sun . Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 37(9):1904–1916, 2015.

Ross Girshick . Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision* , pages 1440–1448, 2015.

Shaoqing Ren , Kaiming He , Ross Girshick , and Jian Sun . Faster r-cnn: Towards real-time object detection with region proposal networks. arXiv preprint arXiv:1506.01497, 2015.

Ken Chatfield , Karen Simonyan , Andrea Vedaldi , and Andrew Zisserman . Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531* , 2014.

Song Han , Huizi Mao , and William J Dally . Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *international conference on learning representations* , 2016.

Hongpeng Zhou , Minghao Yang , Jun Wang , and Wei Pan . Bayesnas: A bayesian approach for neural architecture search. In *International Conference on Machine Learning* , pages 7603–7613. PMLR, 2019.

Stephen Boyd , Neal Parikh , and Eric Chu . Distributed optimization and statistical learning via the alternating direction method of multipliers. Now Publishers Inc, 2011.

Ning Liu , Xiaolong Ma , Zhiyuan Xu , Yanzhi Wang , Jian Tang , and Jieping Ye . Autocompress: An automatic dnn structured pruning framework for ultra-high compression rates. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 4876–4883, 2020.

Yuqiao Liu , Yanan Sun , Bing Xue , Mengjie Zhang , and Gary Yen . A survey on evolutionary neural architecture search. arXiv preprint arXiv:2008. *10937* , 2020.

Songyun Qu , Bing Li , Ying Wang , Dawen Xu , Xiandong Zhao , and Lei Zhang . Raqu: An automatic high-utilization cnn quantization and mapping framework for general-purpose rram accelerator. In *ACM/IEEE Design Automation Conference (DAC)* , pages 1–6. IEEE, 2020.

Ying Wang , Huawei Li , and Xiaowei Li . Real-time meets approximate computing: An elastic cnn inference accelerator with adaptive trade-off between qos and qor. In *ACM/EDAC/IEEE Design Automation Conference (DAC)* , pages 1–6. IEEE, 2017.

Martn Abadi , Ashish Agarwal , Paul Barham , Eugene Brevdo , Zhifeng Chen , Craig Citro , Greg S. Corrado , Andy Davis , Jeffrey Dean , Matthieu Devin , Sanjay Ghemawat , Ian Goodfellow , Andrew Harp , Geoffrey Irving , Michael Isard , Yangqing Jia , Rafal Jozefowicz , Lukasz Kaiser , Manjunath Kudlur , Josh Levenberg , Dan Mané , Rajat Monga , Sherry Moore , Derek Murray , Chris Olah , Mike Schuster , Jonathon Shlens , Benoit Steiner , Ilya Sutskever , Kunal Talwar , Paul Tucker , Vincent Vanhoucke , Vijay Vasudevan , Fernanda Viégas , Oriol Vinyals , Pete Warden , Martin Wattenberg , Martin Wicke , Yuan Yu , and Xiaoqiang Zheng . TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

Adam Paszke , Sam Gross , Francisco Massa , Adam Lerer , James Bradbury , Gregory Chanan , Trevor Killeen , Zeming Lin , Natalia Gimelshein , Luca Antiga , et al. Pytorch: An imperative style, high-performance deep learning library. arXiv preprint arXiv:1912.01703, 2019.

Tianqi Chen , Mu Li , Yutian Li , Min Lin , Naiyan Wang , Minjie Wang , Tianjun Xiao , Bing Xu , Chiyuan Zhang , and Zheng Zhang . Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. arXiv preprint arXiv:1512.01274, 2015.

Frank Seide and Amit Agarwal . Cntk: Microsoft's open-source deep-learning toolkit. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* , pages 2135–2135, 2016.

Tianqi Chen , Thierry Moreau , Ziheng Jiang , Lianmin Zheng , Eddie Q. Yan , Haichen Shen , Meghan Cowan , Leyuan Wang , Yuwei Hu , Luis Ceze , Carlos Guestrin , and Arvind Krishnamurthy . TVM: an automated end-to-end optimizing compiler for deep learning. In 13th USENIX Symposium on Operating Systems Design and Implementation, *OSDI, Carlsbad, CA, USA* , pages 578–594. USENIX Association, 2018.

Nicolas Vasilache , Oleksandr Zinenko , Theodoros Theodoridis , Priya Goyal , Zachary DeVito , William S. Moses , Sven Verdoolaege , Andrew Adams , and Albert Cohen . Tensor comprehensions: Framework-agnostic high-performance machine learning abstractions. In arXiv:1802.04730, 2018.

Nadav Rotem , Jordan Fix , Saleem Abdulrasool , Garret Catron , Summer Deng , Roman Dzhabarov , Nick Gibson , James Hegeman , Meghan Lele , Roman Levenstein , et al. Glow: Graph lowering compiler techniques for neural networks. arXiv preprint arXiv:1805.00907, 2018.

Scott Cyphers , Arjun K. Bansal , Anahita Bhiwandiwalla , Jayaram Bobba , Matthew Brookhart , Avijit Chakraborty , Will Constable , Christian Convey , Leona Cook , Omar Kanawi , Robert Kimball , Jason Knight , Nikolay Korovaiko , Varun Kumar , Yixing Lao , Christopher R. Lishka , Jaikrishnan Menon , Jennifer Myers , Sandeep Aswath Narayana , Adam Procter and Tristan J. Webb . Intel ngraph: An intermediate representation, compiler, and executor for deep learning. arXiv preprint arXiv:1801.08058, 2018.

Chris Leary and Todd Wang . Xla: Tensorflow, compiled. *TensorFlow Dev Summit* , 2017.

Yu-Hsin Chen , Tien-Ju Yang , Joel S. Emer , and Vivienne Sze . Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices. IEEE J. Emerg. Sel. Topics Circuits Syst., pages 292–308, 2019.

Manoj Alwani , Han Chen , Michael Ferdman , and Peter Milder . Fused-layer cnn accelerators. In *Microarchitecture (MICRO), Annual IEEE/ACM International Symposium on* , pages 1–12. IEEE, 2016.

Qingcheng Xiao , Yun Liang , Liqiang Lu , Shengen Yan , and Yu-Wing Tai . Exploring heterogeneous algorithms for accelerating deep convolutional neural networks on fpgas. In *Proceedings of the 54th Annual Design Automation Conference 2017* , pages 1–6, 2017.

Li Zhou , Hao Wen , Radu Teodorescu , and David HC Du . Distributing deep neural networks with containerized partitions at the edge. In 2nd {USENIX} Workshop on Hot Topics in Edge Computing (HotEdge 19), 2019.

Yu Xing , Shuang Liang , Lingzhi Sui , Zhen Zhang , Jiantao Qiu , Xijie Jia , Xin Liu , Yushun Wang , Yi Shan , and Yu Wang . Dnnvm: End-to-end compiler leveraging operation fusion on fpga-based cnn accelerators. In *Proceedings of the ACM/SIGDA International Symposium on Field-Programmable Gate Arrays* , pages 187–188, 2019.

Abhinav Jangda and Uday Bondhugula . An effective fusion and tile size model for optimizing image processing pipelines. ACM SIGPLAN Notices, 53(1):261–275, 2018.

Shixuan Zheng , Xianjue Zhang , Daoli Ou , Shibin Tang , Leibo Liu , Shaojun Wei , and Shouyi Yin . Efficient scheduling of irregular network structures on cnn accelerators. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 39(11):3408–3419, 2020.

Yu-Hsin Chen , Tushar Krishna , Joel Emer , and Vivienne Sze . Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks. In ISSCC. IEEE, 2016.

Xuan Yang , Mingyu Gao , Qiaoyi Liu , Jeff Setter , Jing Pu , Ankita Nayak , Steven Bell , Kaidi Cao , Heonjae Ha , Priyanka Raina , Christos Kozyrakis and Mark Horowitz . Interstellar: Using Halide's Scheduling Language to Analyze DNN Accelerators. In *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS ‘20, pages 369–383, New York, NY, USA, March 2020. Association for Computing Machinery.

Alex Krizhevsky , Ilya Sutskever , and Geoffrey E Hinton . Imagenet classification with deep convolutional neural networks. In NIPS, pages 1097–1105, 2012.

Karen Simonyan and Andrew Zisserman . Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409. *1556* , 2014.

Forrest N Iandola , Matthew W Moskewicz , Khalid Ashraf , et al. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and < 1mb model size. arXiv preprint arXiv:1602.07360, 2016.

Xiaoming Chen , Yinhe Han , and Yu Wang . Communication lower bound in convolution accelerators. In *IEEE International Symposium on High Performance Computer Architecture (HPCA)* , pages 529–541. IEEE, 2020.

Orest Kupyn , Volodymyr Budzan , Mykola Mykhailych , Dmytro Mishkin , and Jiri Matas . Deblurgan: Blind motion deblurring using conditional adversarial networks. ArXiv e-prints, 2017.

Casey Chu , Andrey Zhmoginov , and Mark Sandler . Cyclegan, a master of steganography. arXiv preprint arXiv:1712.02950, 2017.

Jun-Yan Zhu , Taesung Park , Phillip Isola , and Alexei A Efros . Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* , pages 2223–2232, 2017.

Haobo Xu , Ying Wang , Yujie Wang , Jiajun Li , Bosheng Liu , and Yinhe Han . Acg-engine: An inference accelerator for content generative neural networks. In *2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)* , pages 1–7. IEEE, 2019.

Norman P Jouppi , C.S. Young , Nishant Patil , David A Patterson , Gaurav Agrawal , Raminder Bajwa , Sarah Bates , Suresh K Bhatia , Nan Boden , Al Borchers , et al. In-datacenter performance analysis of a tensor processing unit. International symposium on computer

architecture, 45(2):1–12, 2017.
Jinmook Lee , Changhyeon Kim , Sang Hoon Kang , Dongjoo Shin , Sangyeob Kim , and Hoijun Yoo . Unpu: A 50.6tops/w unified deep neural network accelerator with 1b-to-16b fully-variable weight bit-precision. pages 218–220, 2018.
Patrick Judd , Jorge Albericio , Tayler H Hetherington , Tor M Aamodt , and Andreas Moshovos . Stripes: bit-serial deep neural network computing. *international symposium on microarchitecture* , pages 1–12, 2016.
Jorge Albericio , Alberto Delmás , Patrick Judd , Sayeh Sharify , Gerard O'Leary , Roman Genov , and Andreas Moshovos . Bit-pragmatic deep neural network computing. In *Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture* , pages 382–394, 2017.
Eunhyeok Park , Junwhan Ahn , and Sungjoo Yoo . Weighted-entropy-based quantization for deep neural networks. pages 7197–7205, 2017.
Dongqing Zhang , Jiaolong Yang , Dongqiangzi Ye , and Gang Hua . Lq-nets: Learned quantization for highly accurate and compact deep neural networks. *european conference on computer vision* , pages 373–390, 2018.
Shuchang Zhou , Zekun Ni , Xinyu Zhou , He Wen , Yuxin Wu , and Yuheng Zou . Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. arXiv: Neural and Evolutionary Computing, 2016.
Christos Louizos , Matthias Reisser , Tijmen Blankevoort , Efstratios Gavves , and Max Welling . Relaxed quantization for discretized neural networks. *international conference on learning representations* , 2019.
Asit K Mishra , Eriko Nurvitadhi , Jeffrey J Cook , and Debbie Marr . Wrpn: Wide reduced-precision networks. *international conference on learning representations* , 2018.
Jungwook Choi . Pact: Parameterized clipping activation for quantized neural networks. *arXiv: Computer Vision and Pattern Recognition* , 2018.
Chenzhuo Zhu , Song Han , Huizi Mao , and William J Dally . Trained ternary quantization. *international conference on learning representations* , 2017.
Raghuraman Krishnamoorthi . Quantizing deep convolutional networks for efficient inference: A whitepaper. arXiv preprint arXiv:1806. *08342* , 2018.
Joseph Redmon , Santosh Divvala , Ross Girshick , and Ali Farhadi . You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* , pages 779–788, 2016.
Zidong Du , Robert Fasthuber , Tianshi Chen , Paolo Ienne , Ling Fei Li , Tao Luo , Xiaobing Feng , Yunji Chen , and Olivier Temam . Shidiannao: shifting vision processing closer to the sensor. International symposium on computer architecture, 43(3):92–104, 2015.
Y. Wang , J. Xu , Y. Han , H. Li , and X. Li . Deepburning: Automatic generation of fpga-based learning accelerators for the neural network family. In *ACM/EDAC/IEEE Design Automation Conference (DAC)* , pages 1–6, June 2016.
S. Han , X. Liu , H. Mao , J. Pu , A. Pedram , M. A. Horowitz , and W. J. Dally . Eie: Efficient inference engine on compressed deep neural network. In *ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)* , pages 243–254, June 2016.
Aojun Zhou , Anbang Yao , Yiwen Guo , Lin Xu , and Yurong Chen . Incremental network quantization: Towards lossless cnns with low-precision weights. CoRR, abs/1702.03044, 2017.
Zhaowei Cai , Xiaodong He , Jian Sun , and Nuno Vasconcelos. Deep learning with low precision by half-wave gaussian quantization. In CVPR, 2017.
Yoshua Bengio , Nicholas Léonard , and Aaron Courville . Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv: *1308* .3432, 2013.
Adam Paszke , Sam Gross , Soumith Chintala , Gregory Chanan , Edward Yang , Zachary DeVito , Zeming Lin , Alban Desmaison , Luca Antiga , and Adam Lerer . Automatic differentiation in PyTorch. In NIPS Autodiff Workshop, 2017.
Mark Sandler , Andrew Howard , Menglong Zhu , Andrey Zhmoginov , and Liang-Chieh Chen . Mobilenetv2: Inverted residuals and linear bottlenecks, 2018.
Joseph Redmon and Ali Farhadi . Yolo9000: Better, faster, stronger. pages 6517–6525, 2017.
Xiandong Zhao , Ying Wang , Xuyi Cai , Cheng Liu , and Lei Zhang . Linear symmetric quantization of neural networks for low-precision integer hardware. 2020.
Ying Wang , Zhenyu Quan , Jiajun Li , Yinhe Han , Huawei Li , and Xiaowei Li . A retrospective evaluation of energy-efficient object detection solutions on embedded devices. In *Design,*

Automation & Test in Europe Conference & Exhibition (DATE) , pages 709–714. IEEE, 2018.
Xiandong Zhao , Ying Wang , Cheng Liu , Cong Shi , Kaijie Tu , and Lei Zhang . Bitpruner: network pruning for bit-serial accelerators. In *ACM/IEEE Design Automation Conference (DAC)* , pages 1–6. IEEE, 2020.
Dawen Xu , Cheng Liu , Ying Wang , Kaijie Tu , Bingsheng He , and Lei Zhang . Accelerating generative neural networks on unmodified deep learning processorsâ€"a software approach. IEEE Transactions on Computers, 69(8):1172–1184, 2020.
Xuyi Cai , Ying Wang , and Lei Zhang . Optimus: towards optimal layer-fusion on deep learning processors. In *Proceedings of the 22nd ACM SIGPLAN/SIGBED International Conference on Languages, Compilers, and Tools for Embedded Systems* , pages 67–79, 2021.
Weiwei Chen , Ying Wang , Gangliang Lin , Chengsi Gao , Cheng Liu , and Lei Zhang . Chanas: coordinated search for network architecture and scheduling policy. In *Proceedings of the 22nd ACM SIGPLAN/SIGBED International Conference on Languages, Compilers, and Tools for Embedded Systems* , pages 42–53, 2021.
Xiangyu Zhang , Xinyu Zhou , Mengxiao Lin , and Jian Sun . ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In CVPR, 2018.
Christian Szegedy , Vincent Vanhoucke , Sergey Ioffe , Jon Shlens , and Zbigniew Wojna . Rethinking the Inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* , pages 2818–2826, 2016.
Min Lin , Qiang Chen , and Shuicheng Yan . Network in network. arXiv preprint arXiv:1312.4400, 2013.
Dongyoon Han , Jiwhan Kim , and Junmo Kim . Deep pyramidal residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* , pages 5927–5935, 2017.
Barret Zoph , Vijay Vasudevan , Jonathon Shlens , and Quoc V Le . Learning transferable architectures for scalable image recognition. arXiv preprint arXiv:1707.07012, 2017.
Esteban Real , Alok Aggarwal , Yanping Huang , and Quoc V Le . Regularized evolution for image classifier architecture search. arXiv preprint arXiv:1802. *01548* , 2018.
Bowen Baker , Otkrist Gupta , Nikhil Naik , and Ramesh Raskar . Designing neural network architectures using reinforcement learning. arXiv preprint arXiv:1611.02167, 2016.
Zhao Zhong , Junjie Yan , and Cheng-Lin Liu . Practical network blocks design with q-learning. arXiv preprint arXiv:1708.05552, 2017.
Hanxiao Liu , Karen Simonyan , Oriol Vinyals , Chrisantha Fernando , and Koray Kavukcuoglu . Hierarchical representations for efficient architecture search. In ICLR, 2018.
Lei Deng , Guoqi Li , Song Han , Luping Shi , and Yuan Xie . Model compression and hardware acceleration for neural networks: A comprehensive survey. Proceedings of the IEEE, 108(4):485–532, 2020.
Han Cai , Tianyao Chen , Weinan Zhang , Yong Yu , and Jun Wang . Efficient Architecture Search by Network Transformation. In AAAI, 2018.
Han Cai , Jiacheng Yang , Weinan Zhang , Song Han , and Yong Yu . Path-Level Network Transformation for Efficient Architecture Search. In ICML, 2018.
Thomas Elsken , Jan Hendrik Metzen , and Frank Hutter . Efficient multi-objective neural architecture search via lamarckian evolution. In *International Conference on Learning Representations* , 2018.
Andrew Brock , Theodore Lim , James M Ritchie , and Nick Weston . Smash: one-shot model architecture search through hypernetworks. arXiv preprint arXiv:1708.05344, 2017.
Hieu Pham , Melody Guan , Barret Zoph , Quoc Le , and Jeff Dean . Efficient neural architecture search via parameters sharing. In *International Conference on Machine Learning* , pages 4095–4104. PMLR, 2018.
Golnaz Ghiasi , Tsung-Yi Lin , and Quoc V Le . Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* , pages 7036–7045, 2019.
Mingxing Tan , Ruoming Pang , and Quoc V Le . Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , pages 10781–10790, 2020.
Liang-Chieh Chen , Maxwell Collins , Yukun Zhu , George Papandreou , Barret Zoph , Florian Schroff , Hartwig Adam , and Jon Shlens . Searching for efficient multi-scale architectures for dense image prediction. In Advances in neural information processing systems, pages

8699–8710, 2018.
Han Cai , Chuang Gan , Tianzhe Wang , Zhekai Zhang , and Song Han . Once for all: Train one network and specialize it for efficient deployment. In *International Conference on Learning Representations* , 2020.
Mingxing Tan , Bo Chen , Ruoming Pang , Vijay Vasudevan , and Quoc V Le . Mnasnet: Platform-aware neural architecture search for mobile. arXiv preprint arXiv:1807.11626, 2018.
Emma Strubell , Ananya Ganesh , and Andrew McCallum . Energy and policy considerations for deep learning in nlp. In ACL, 2019.
Jiahui Yu , Pengchong Jin , Hanxiao Liu , Gabriel Bender , Pieter-Jan Kindermans , Mingxing Tan , Thomas Huang , Xiaodan Song , Ruoming Pang , and Quoc Le . Bignas: Scaling up neural architecture search with big single-stage models. arXiv preprint arXiv:2003.11142, 2020.
Jiahui Yu and Thomas Huang . Autoslim: Towards one-shot architecture search for channel numbers. arXiv preprint arXiv:1903.11728, 2019.
S. Ren , K. He , R. Girshick , and J. Sun . Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of NIPS, pages 91–99, 2015.
Joseph Redmon , Santosh Divvala , Ross Girshick , and Ali Farhadi . You only look once: Unified, real-time object detection. In Proceedings of CVPR, IEEE, pages 779–788, 2016.
Joseph Redmon and Ali Farhadi . Yolo9000: better, faster, stronger. arXiv preprint 1612.08242, 2016.
Joseph Redmon and Ali Farhadi , YOLOv3: An Incremental Improvement, https://arxiv.org/abs/1804.02767.
Wei Liu , Dragomir Anguelov , Dumitru Erhan , Christian Szegedy , Scott Reed , Cheng-Yang Fu , and Alexander C Berg . Ssd: Single shot multibox detector. In Proceedings of ECCV, pages 21–37. Springer, 2016.
Tsung-Yi Lin , Priya Goyal , Ross Girshick , Kaiming He , and Piotr Dollár . Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* , 2017.
J. Hosang , R. Benenson , and B. Schiele . Learning non-maximum suppression. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 6469–6477, 2017.
Y. D. Kim , E. Park , S. Yoo , T. Choi , L. Yang , and D. Shin . Compression of deep convolutional neural networks for fast and low power mobile applications. arXiv preprint 1511.06530, 2015.
tim. lewis. Openmp.
Eunhyeok Park , Junwhan Ahn , and Sungjoo Yoo . Weighted-entropy-based quantization for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* , pages 5456–5464, 2017.
T. Y. Lin , M. Maire , S. Belongie , J. Hays , P. Perona , D. Ramanan , P. Dollár , and C. L. Zitnick. Microsoft coco: Common objects in context. In *Proceedings of ECCV*, pages 740–755. Springer, 2014.
J. Redmon and A. Farhadi . Yolo website. https://pjreddie.com/darknet/yolo/, 2016–2018.
T. Y. Lin , P. Dollár , R. Girshick , K. He , B. Hariharan , and S. Belongie . Feature pyramid networks for object detection. In Proceedings of CVPR, IEEE, volume 1, page 4, 2017.
G. Huang , Z. Liu , K. Q. Weinberger , and L. van der Maaten . Densely connected convolutional networks. arXiv preprint arXiv:1608.06993, 2016.
Facebook. Caffe2 detectron api. https://github.com/facebook-research/Detectron, 2018.
J. Redmon . Yolo: Real-time object detection. http://pjreddie.com/darknet/yolo/, 2013–2016.
NVIDIA Jetson , October 2015.
M. Verucchi et al. A systematic assessment of embedded neural networks for object detection. ETFA, 2020.
Alexey Bochkovskiy , Chien-Yao Wang , and Hong-Yuan Mark Liao . YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv:2004.10934 [cs, eess], April 2020. arXiv: 2004.10934.
Suyog Gupta and Berkin Akin . Accelerator-aware neural network design using automl. arXiv preprint arXiv:2003.02838, 2020.
Tsung-Yi Lin , Michael Maire , Serge J. Belongie , Lubomir D. Bourdev , Ross B. Girshick , James Hays , Pietro Perona , Deva Ramanan , Piotr Dollár , and C. Lawrence Zitnick . Microsoft COCO: common objects in context. CoRR, abs/1405.0312, 2014.
Kent Gauen , Ryan Dailey , Yung-Hsiang Lu , Eunbyung Park , Wei Liu , Alexander C. Berg , and Yiran Chen . Three years of low-power image recognition challenge: Introduction to special

session. DATE,. 2018.83420998, 2018.

Yung-Hsiang Lu , Alexander C. Berg , and Yiran Chen . Low-power image recognition challenge. AI Magazine, Vol 39 No 2: Summer 2018, 2018.

K. Simonyan et al. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556.

C. Szegedy , W. Liu , and Y. Jia . Going deeper with convolutions. CVPR, arXiv:1409.4842, 2015.

Kaiming He , Xiangyu Zhang , Shaoqing Ren , and Jian Sun . Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* , June 2016.

Andrew G. Howard , Menglong Zhu , Bo Chen , Dmitry Kalenichenko , Weijun Wang , Tobias Weyand , Marco Andreetto , and Hartwig Adam . Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv, abs/1704.04861, 2017.

Andrew G. Howard , Menglong Zhu , Bo Chen , Dmitry Kalenichenko , Weijun Wang , Tobias Weyand , Marco Andreetto , and Hartwig Adam . Mobilenetv2: Inverted residuals and linear bottlenecks. arXiv, abs/1801.04381, 2018.

Benoit Jacob , Skirmantas Kligys , Bo Chen , Menglong Zhu , Matthew Tang , Andrew G. Howard , Hartwig Adam , and Dmitry Kalenichenko . Quantization and training of neural networks for efficient integer-arithmetic-only inference. CoRR, abs/1712.05877, 2017.

Tao Sheng , Chen Feng , Shaojie Zhuo , Xiaopeng Zhang , Liang Shen , and Mickey Aleksic . A quantization-friendly separable convolution for mobilenets. *arXiv*, arXiv:1803.08607, 2018.

Ilya Loshchilov and Frank Hutter . SGDR: stochastic gradient descent with restarts. CoRR, abs/1608.03983, 2016.

M. Horowitz . 1.1 computing's energy problem (and what we can do about it). In *IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)* , pages 10–14, 2014.

Benoit Jacob , Skirmantas Kligys , Bo Chen , Menglong Zhu , Matthew Tang , Andrew Howard , Hartwig Adam , and Dmitry Kalenichenko . Quantization and training of neural networks for efficient integer-arithmetic-only inference. *Conference on Computer Vision and Pattern Recognition (CVPR)* , 2018.

Bita Rouhani , Daniel Lo , Ritchie Zhao , Ming Liu , Jeremy Fowers , Kalin Ovtcharov , Anna Vinogradsky , Sarah Massengill , Lita Yang , Ray Bittner , Alessandro Forin , Haishan Zhu , Taesik Na , Prerak Patel , Shuai Che , Lok Chand Koppaka , Xia Song , Subhojit Som , Kaustav Das , Saurabh Tiwary , Steve Reinhardt , Sitaram Lanka , Eric Chung , and Doug Burger . Pushing the limits of narrow precision inferencing at cloud scale with microsoft floating point. In Neural Information Processing Systems (NeurIPS). ACM, November 2020.

Pierre Stock , Armand Joulin , Rémi Gribonval , Benjamin Graham , and Hervé Jégou . And the bit goes down: Revisiting the quantization of neural networks. CoRR, abs/1907.05686, 2019.

Marcelo Gennari do Nascimento , Roger Fawcett , and Victor Adrian Prisacariu . Dsconv: Efficient convolution operator. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* , October 2019.

Sergey Ioffe and Christian Szegedy . Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei , editors, Proceedings of the 32nd International Conference on Machine Learning, volume 37 of *Proceedings of Machine Learning Research* , pages 448–456, Lille, France, 07–09 Jul 2015. PMLR.

Raghuraman Krishnamoorthi . Quantizing deep convolutional networks for efficient inference: A whitepaper. arXiv preprint arXiv:1806. *08342* , 2018.

Prajit Ramachandran , Barret Zoph , and Quoc V. Le . Searching for activation functions. CoRR, abs/1710.05941, 2017.

Mart van Baalen , Christos Louizos , Markus Nagel , Rana Ali Amjad , Ying Wang , Tijmen Blankevoort , and Max Welling . Bayesian bits: Unifying quantization and pruning. In H. Larochelle , M. Ranzato , R. Hadsell , M. F. Balcan , and H. Lin , editors, Advances in Neural Information Processing Systems, volume 33, pages 5741–5752. Curran Associates, Inc., 2020.

Zhen Dong , Zhewei Yao , Amir Gholami , Michael W. Mahoney , and Kurt Keutzer . HAWQ: hessian aware quantization of neural networks with mixed-precision. *International Conference on Computer Vision (ICCV)* , 2019.

Stefan Uhlich , Lukas Mauch , Fabien Cardinaux , Kazuki Yoshiyama , Javier Alonso Garcia , Stephen Tiedemann , Thomas Kemp , and Akira Nakamura . Mixed precision dnns: All you need is a good parametrization. In *International Conference on Learning Representations* , 2020.

Ron Banner , Yury Nahshan , and Daniel Soudry . Post training 4-bit quantization of convolutional networks for rapid-deployment. *Neural Information Processing Systems (NeuRIPS)* , 2019.

Markus Nagel , Mart van Baalen , Tijmen Blankevoort , and Max Welling . Data-free quantization through weight equalization and bias correction. *International Conference on Computer Vision (ICCV)* , 2019.

Tao Sheng , Chen Feng , Shaojie Zhuo , Xiaopeng Zhang , Liang Shen , and Mickey Aleksic . A quantization-friendly separable convolution for mobilenets. In 1st Workshop on Energy Efficient Machine Learning and Cognitive Computing for Embedded Applications (EMC2 *)*, 2018.

Mark Sandler , Andrew Howard , Menglong Zhu , Andrey Zhmoginov , and Liang-Chieh Chen . Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* , pages 4510–4520, 2018.

Eldad Meller , Alexander Finkelstein , Uri Almog , and Mark Grobman . Same, same but different: Recovering neural network quantization error through weight factorization. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA* , pages 4486–4495, 2019.

Alexander Finkelstein , Uri Almog , and Mark Grobman . Fighting quantization bias with bias. arXiv preprint arxiv:1906. *03193* , 2019.

Markus Nagel , Rana Ali Amjad , Mart Van Baalen , Christos Louizos , and Tijmen Blankevoort . Up or down? Adaptive rounding for post-training quantization. In Hal DaumÃ© III and Aarti Singh , editors, Proceedings of the 37th International Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, pages 7197–7206. PMLR, 13–18 Jul 2020.

Suyog Gupta , Ankur Agrawal , Kailash Gopalakrishnan , and Pritish Narayanan . Deep learning with limited numerical precision. *International Conference on Machine Learning, ICML* , 2015.

Bert Moons , Parham Noorzad , Andrii Skliar , Giovanni Mariani , Dushyant Mehta , Chris Lott , and Tijmen Blankevoort . Distilling optimal neural networks: Rapid search in diverse spaces. arXiv preprint arXiv:2012. *08859* , 2020.

Steven K. Esser , Jeffrey L. McKinstry , Deepika Bablani , Rathinakumar Appuswamy , and Dharmendra S. Modha . Learned step size quantization. *International Conference on Learning Representations (ICLR)* , 2020.

Sambhav R. Jain , Albert Gural , Michael Wu , and Chris Dick . Trained uniform quantization for accurate and efficient neural network inference on fixed-point hardware. CoRR, abs/1903.08066, 2019.

Yash Bhalgat , Jinwon Lee , Markus Nagel , Tijmen Blankevoort , and Nojun Kwak . Lsq+: Improving low-bit quantization through learnable offsets and better initialization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* , June 2020.

Jiwei Yang , Xu Shen , Jun Xing , Xinmei Tian , Houqiang Li , Bing Deng , Jianqiang Huang , and Xian-sheng Hua . Quantization networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* , June 2019.

Mingxing Tan and Quoc Le . EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov , editors, Proceedings of the 36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019.

Gao Huang , Shichen Liu , Laurens van der Maaten , and Kilian Q. Weinberger . Condensenet: An efficient densenet using learned group convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* , June 2018.

M. Horowitz . 1.1 computing's energy problem (and what we can do about it). In IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), pages 10–14, 2014.

Andrew G. Howard , Menglong Zhu , Bo Chen , Dmitry Kalenichenko , Weijun Wang , Tobias Weyand , Marco Andreetto , and Hartwig Adam . Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR, abs/1704.04861, 2017.

Karen Simonyan and Andrew Zisserman . Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations* , 2015.

Xiangyu Zhang , Xinyu Zhou , Mengxiao Lin , and Jian Sun . Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on*

Computer Vision and Pattern Recognition (CVPR) , June 2018.
Mingxing Tan and Quoc V. Le . Efficientnet: Rethinking model scaling for convolutional neural networks. CoRR, abs/1905.11946, 2019.
Andrew Howard , Mark Sandler , Grace Chu , Liang-Chieh Chen , Bo Chen , Mingxing Tan , Weijun Wang , Yukun Zhu , Ruoming Pang , Vijay Vasudevan , et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* , pages 1314–1324, 2019.
Francois Chollet . Xception: Deep learning with depthwise separable convolutions. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.
Mark Sandler , Jonathan Baccash , Andrey Zhmoginov , and Andrew Howard . Non-discriminative data or weak model? on the relative importance of data and model resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops* , Oct 2019.
Saining Xie , Ross Girshick , Piotr Dollar , Zhuowen Tu , and Kaiming He . Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* , July 2017.
Sergey Zagoruyko and Nikos Komodakis . Wide Residual Networks. In *British Machine Vision Conference 2016* , York, France, January 2016. British Machine Vision Association.
Imagenet.
Wenzhe Shi , Jose Caballero , Ferenc Huszar , Johannes Totz , Andrew P. Aitken , Rob Bishop , Daniel Rueckert , and Zehan Wang . Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR *)*, June 2016.
Mehdi SM Sajjadi , Raviteja Vemulapalli , and Matthew Brown . Frame-recurrent video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* , pages 6626–6634, 2018.
Tien-Ju Yang , Maxwell D Collins , Yukun Zhu , Jyh-Jing Hwang , Ting Liu , Xiao Zhang , Vivienne Sze , George Papandreou , and Liang-Chieh Chen . Deeperlab: Single-shot image parser. arXiv preprint arXiv:1902.05093, 2019.
Adam Paszke , Sam Gross , Francisco Massa , Adam Lerer , James Bradbury , Gregory Chanan , Trevor Killeen , Zeming Lin , Natalia Gimelshein , Luca Antiga , Alban Desmaison , Andreas Kopf , Edward Yang , Zachary DeVito , Martin Raison , Alykhan Tejani , Sasank Chilamkurthy , Benoit Steiner , Lu Fang , Junjie Bai , and Soumith Chintala . Pytorch: An imperative style, high-performance deep learning library. In H. Wallach , H. Larochelle , A. Beygelzimer , F. Alché-Buc , E. Fox , and R. Garnett , editors, Advances in Neural Information Processing Systems *32* , pages 8024–8035. Curran Associates, Inc., 2019.
Christian Szegedy , Sergey Ioffe , Vincent Vanhoucke , and Alex A. Alemi . Inception-v4, inception-resnet and the impact of residual connections on learning. In ICLR 2016 Workshop, 2016.
Radford M. Neal . Connectionist learning of belief networks. Artif. Intell., 56(1):71â€“113, July 1992.
Yann LeCun , Léon Bottou , Genevieve B. Orr , and Klaus-Robert Müller . Efficient backprop. In Neural Networks: Tricks of the Trade, This Book is an Outgrowth of a 1996 NIPS Workshop, page 9â€“50, Berlin, Heidelberg, 1998. Springer-Verlag.
Richard Hahnloser , Rahul Sarpeshkar , Misha Mahowald , Rodney Douglas , and H. Seung . Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. Nature, 405:947–951, 07 2000.
Richard Hahnloser and H. Sebastian Seung . Permitted and forbidden sets in symmetric threshold-linear networks. In T. Leen , T. Dietterich , and V. Tresp , editors, Advances in Neural Information Processing Systems, volume 13. MIT Press, 2001.
K. Fukushima . Visual feature extraction by a multilayered network of analog threshold elements. IEEE Transactions on Systems Science and Cybernetics, 5(4):322–333, 1969.
Xavier Glorot , Antoine Bordes , and Yoshua Bengio . Deep sparse rectifier neural networks. In Geoffrey Gordon , David Dunson , and Miroslav Dudík , editors, Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, Fort Lauderdale, FL, USA, 11–13 Apr 2011. JMLR Workshop and Conference Proceedings.
Chigozie Nwankpa , Winifred Ijomah , Anthony Gachagan , and Stephen Marshall . Activation functions: Comparison of trends in practice and research for deep learning. CoRR,

abs/1811.03378, 2018.
Dan Hendrycks and Kevin Gimpel . Bridging nonlinearities and stochastic regularizers with gaussian error linear units. CoRR, abs/1606.08415, 2016.
Elfwing Stefan , Uchibe Eiji , and Doya Kenji . Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. Neural Networks, Jan 2018.
R. Avenash and P. Vishawanth . Semantic segmentation of satellite images using a modified cnn with hard-swish activation function. In VISIGRAPP, 2019.
Amir Gholami , Michael W Mahoney , and Kurt Keutzer . An integrated approach to neural network design, training, and inference. *Univ. California, Berkeley* , *Berkeley, CA, USA* , *Tech. Rep*, 2020.
Yani Ioannou , Duncan Robertson , Roberto Cipolla , and Antonio Criminisi . Deep roots: Improving cnn efficiency with hierarchical filter groups. In *Proceedings of the IEEE conference on computer vision and pattern recognition* , pages 1231–1240, 2017.
Franck Mamalet and Christophe Garcia . Simplifying convnets for fast learning. In *International Conference on Artificial Neural Networks* , pages 58–65. Springer, 2012.
Bichen Wu , Alvin Wan , Xiangyu Yue , Peter Jin , Sicheng Zhao , Noah Golmant , Amir Gholaminejad , Joseph Gonzalez , and Kurt Keutzer . Shift: A zero flop, zero parameter alternative to spatial convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* , pages 9127–9135, 2018.
Tara N Sainath , Brian Kingsbury , Vikas Sindhwani , Ebru Arisoy , and Bhuvana Ramabhadran . Low-rank matrix factorization for deep neural network training with high-dimensional output targets. In *IEEE international conference on acoustics, speech and signal processing* , pages 6655–6659. IEEE, 2013.
PP Kanjilal , PK Dey , and DN Banerjee . Reduced-size neural networks through singular value decomposition and subset selection. Electronics Letters, 29(17):1516–1518, 1993.
Qibin Zhao , Masashi Sugiyama , Longhao Yuan , and Andrzej Cichocki . Learning efficient tensor representations with ring-structured networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* , pages 8608–8612. IEEE, 2019.
Gao Huang , Zhuang Liu , Laurens Van Der Maaten , and Kilian Q Weinberger . Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* , pages 4700–4708, 2017.
Dilin Wang , Meng Li , Chengyue Gong , and Vikas Chandra . Attentivenas: Improving neural architecture search via attentive sampling. arXiv preprint arXiv:2011. *09011* , 2020.
Thomas Elsken , Jan Hendrik Metzen , and Frank Hutter . Neural architecture search: A survey. J. Mach. Learn. Res., 20(55):1–21, 2019.
Amir Gholami , Kiseok Kwon , Bichen Wu , Zizheng Tai , Xiangyue Yue , Peter Jin , Sicheng Zhao , and Kurt Keutzer . SqueezeNext: Hardware-aware neural network design. Workshop paper in CVPR, 2018.
Yann LeCun , John S Denker , and Sara A Solla . Optimal brain damage. In Advances in neural information processing systems, pages 598–605, 1990.
Babak Hassibi and David G Stork . Second order derivatives for network pruning: Optimal brain surgeon. Morgan Kaufmann, 1993.
Xin Dong , Shangyu Chen , and Sinno Jialin Pan . Learning to prune deep neural networks via layer-wise optimal brain surgeon. arXiv preprint arXiv:1705.07565, 2017.
Namhoon Lee , Thalaiyasingam Ajanthan , and Philip HS Torr . Snip: Single-shot network pruning based on connection sensitivity. arXiv preprint arXiv:1810.02340, 2018.
Xia Xiao , Zigeng Wang , and Sanguthevar Rajasekaran . Autoprune: Automatic network pruning by regularizing auxiliary parameters. In Advances in Neural Information Processing Systems, pages 13681–13691, 2019.
Sejun Park , Jaeho Lee , Sangwoo Mo , and Jinwoo Shin . Lookahead: a far-sighted alternative of magnitude-based pruning. arXiv preprint arXiv:2002.04809, 2020.
Yihui He , Ji Lin , Zhijian Liu , Hanrui Wang , Li-Jia Li , and Song Han . Amc: Automl for model compression and acceleration on mobile devices. In *Proceedings of the European Conference on Computer Vision (ECCV)* , pages 784–800, 2018.
Ruichi Yu , Ang Li , Chun-Fu Chen , Jui-Hsin Lai , Vlad I Morariu , Xintong Han , Mingfei Gao , Ching-Yung Lin , and Larry S Davis . Nisp: Pruning networks using neuron importance score propagation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern*

Recognition , pages 9194–9203, 2018.
Shaohui Lin , Rongrong Ji , Yuchao Li , Yongjian Wu , Feiyue Huang , and Baochang Zhang . Accelerating convolutional networks via global & dynamic filter pruning. In IJCAI, pages 2425–2432, 2018.
Zehao Huang and Naiyan Wang . Data-driven sparse structure selection for deep neural networks. In *Proceedings of the European conference on computer vision (ECCV)* , pages 304–320, 2018.
Chenglong Zhao , Bingbing Ni , Jian Zhang , Qiwei Zhao , Wenjun Zhang , and Qi Tian . Variational convolutional neural network pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* , pages 2780–2789, 2019.
Shixing Yu , Zhewei Yao , Amir Gholami , Zhen Dong , Michael W Mahoney , and Kurt Keutzer . Hessian-aware pruning and optimal neural implant. arXiv preprint arXiv:2101.08940, 2021.
Aydin Buluc and John R Gilbert . Challenges and advances in parallel sparse matrix-matrix multiplication. In *2008 37th International Conference on Parallel Processing* , pages 503–510. IEEE, 2008.
Trevor Gale , Erich Elsen , and Sara Hooker . The state of sparsity in deep neural networks. arXiv preprint arXiv:1902. *09574* , 2019.
Davis Blalock , Jose Javier Gonzalez Ortiz , Jonathan Frankle , and John Guttag . What is the state of neural network pruning? arXiv preprint arXiv:2003. *03033* , 2020.
Torsten Hoefler , Dan Alistarh , Tal Ben-Nun , Nikoli Dryden , and Alexandra Peste . Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. arXiv preprint arXiv:2102. *00554* , 2021.
Andrey Kuzmin , Markus Nagel , Saurabh Pitre , Sandeep Pendyam , Tijmen Blankevoort , and Max Welling . Taxonomy and evaluation of structured compression of convolutional neural networks. arXiv preprint arXiv:1912.09802, 2019.
Adriana Romero , Nicolas Ballas , Samira Ebrahimi Kahou , Antoine Chassang , Carlo Gatta , and Yoshua Bengio . Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550, 2014.
Asit Mishra and Debbie Marr . Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy. In *International Conference on Learning Representations* , 2018.
Yuncheng Li , Jianchao Yang , Yale Song , Liangliang Cao , Jiebo Luo , and Li-Jia Li . Learning from noisy labels with distillation. In *Proceedings of the IEEE International Conference on Computer Vision* , pages 1910–1918, 2017.
Junho Yim , Donggyu Joo , Jihoon Bae , and Junmo Kim . A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* , pages 4133–4141, 2017.
Antonio Polino , Razvan Pascanu , and Dan Alistarh . Model compression via distillation and quantization. In *International Conference on Learning Representations* , 2018.
Sungsoo Ahn , Shell Xu Hu , Andreas Damianou , Neil D Lawrence , and Zhenwen Dai . Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , pages 9163–9171, 2019.
Hongxu Yin , Pavlo Molchanov , Jose M Alvarez , Zhizhong Li , Arun Mallya , Derek Hoiem , Niraj K Jha , and Jan Kautz . Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , pages 8715–8724, 2020.
Ron Banner , Itay Hubara , Elad Hoffer , and Daniel Soudry . Scalable methods for 8-bit training of neural networks. Advances in neural information processing systems, 2018.
Naigang Wang , Jungwook Choi , Daniel Brand , Chia-Yu Chen , and Kailash Gopalakrishnan . Training deep neural networks with 8-bit floating point numbers. Advances in neural information processing systems, 2018.
Jangho Kim , KiYoon Yoo , and Nojun Kwak . Position-based scaled gradient for model quantization and sparse training. Advances in neural information processing systems, 2020.
Fartash Faghri , Iman Tabrizian , Ilia Markov , Dan Alistarh , Daniel Roy , and Ali Ramezani-Kebrya . Adaptive gradient quantization for data-parallel sgd. Advances in neural information processing systems, 2020.
Brian Chmiel , Liad Ben-Uri , Moran Shkolnik , Elad Hoffer , Ron Banner , and Daniel Soudry . Neural gradients are near-lognormal: improved quantized and sparse training. In *International Conference on Learning Representations* , 2021.

Matthieu Courbariaux , Yoshua Bengio , and Jean-Pierre David . Training deep neural networks with low precision multiplications. arXiv preprint arXiv:1412.7024, 2014.

Suyog Gupta , Ankur Agrawal , Kailash Gopalakrishnan , and Pritish Narayanan . Deep learning with limited numerical precision. In *International conference on machine learning* , pages 1737–1746. PMLR, 2015.

Boris Ginsburg , Sergei Nikolaev , Ahmad Kiswani , Hao Wu , Amir Gholaminejad , Slawomir Kierat , Michael Houston , and Alex Fit-Florea . Tensor processing using low precision format, December 28 2017. US Patent App. 15/624,577.

Paulius Micikevicius , Sharan Narang , Jonah Alben , Gregory Diamos , Erich Elsen , David Garcia , Boris Ginsburg , Michael Houston , Oleksii Kuchaiev , Ganesh Venkatesh , et al. Mixed precision training. arXiv preprint arXiv:1710.03740, 2017.

Warren S McCulloch and Walter Pitts . A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5(4):115–133, 1943.

Rufin VanRullen and Christof Koch . Is perception discrete or continuous? Trends in cognitive sciences, 7(5):207–213, 2003.

James Tee and Desmond P Taylor . Is information in the brain represented in continuous or discrete form? IEEE Transactions on Molecular, Biological and Multi-Scale Communications, 6(3):199–209, 2020.

A Aldo Faisal , Luc PJ Selen , and Daniel M Wolpert . Noise in the nervous system. Nature reviews neuroscience, 9(4):292–303, 2008.

Rishidev Chaudhuri and Ila Fiete . Computational principles of memory. Nature neuroscience, 19(3):394, 2016.

Kenneth W Latimer , Jacob L Yates , Miriam LR Meister , Alexander C Huk , and Jonathan W Pillow . Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. Science, 349(6244):184–187, 2015.

Lav R Varshney and Kush R Varshney . Decision making with quantized priors leads to discrimination. Proceedings of the IEEE, 105(2):241–255, 2016.

Mel Win Khaw , Luminita Stevens , and Michael Woodford . Discrete adjustment to a changing environment: Experimental evidence. Journal of Monetary Economics, 91:88–103, 2017.

Lav R Varshney , Per Jesper Sjöström , and Dmitri B Chklovskii . Optimal information storage in noisy synapses under resource constraints. Neuron, 52(3):409–423, 2006.

John Z Sun , Grace I Wang , Vivek K Goyal , and Lav R Varshney . A framework for bayesian optimality of psychophysical laws. Journal of Mathematical Psychology, 56(6):495–501, 2012.

Robert M. Gray and David L. Neuhoff . Quantization. IEEE transactions on information theory, 44(6):2325–2383, 1998.

S. M. Stigler . The History of Statistics: The Measurement of Uncertainty before 1900. Harvard University Press, Cambridge, 1986.

Bernhard Riemann . Ueber die Darstellbarkeit einer Function durch eine trigonometrische Reihe, volume 13. Dieterich, 1867.

William Fleetwood Sheppard . On the calculation of the most probable values of frequency-constants, for data arranged according to equidistant division of a scale. Proceedings of the London Mathematical Society, 1(1):353–380, 1897.

Claude E Shannon . A mathematical theory of communication. *The* Bell system technical journal, 27(3):379–423, 1948.

David A Huffman . A method for the construction of minimum-redundancy codes. Proceedings of the IRE, 40(9):1098–1101, 1952.

Claude E Shannon . Coding theorems for a discrete source with a fidelity criterion. IRE Nat. Conv. Rec, 4(142-163):1, 1959.

JG Dunn . The performance of a class of n dimensional quantizers for a gaussian source. In *Proc. Columbia Symp. Signal Transmission Processing* , pages 76–81, 1965.

AE Gamal , L Hemachandra , Itzhak Shperling , and V Wei . Using simulated annealing to design good codes. IEEE Transactions on Information Theory, 33(1):116–123, 1987.

William H Equitz . A new vector quantization clustering algorithm. IEEE transactions on acoustics, speech, and signal processing, 37(10):1568–1575, 1989.

Kenneth Rose , Eitan Gurewitz , and Geoffrey Fox . A deterministic annealing approach to clustering. Pattern Recognition Letters, 11(9):589–594, 1990.

BM Oliver , JR Pierce , and Claude E Shannon . The philosophy of pcm. Proceedings of the IRE, 36(11):1324–1331, 1948.

William Ralph Bennett . Spectra of quantized signals. The Bell System Technical Journal, 27(3):446–472, 1948.

L.N. Trefethen and D. Bau III . Numerical Linear Algebra. SIAM, Philadelphia, 1997.

Benoit Jacob , Skirmantas Kligys , Bo Chen , Menglong Zhu , Matthew Tang , Andrew Howard , Hartwig Adam , and Dmitry Kalenichenko . Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* , 2018.

Hao Wu , Patrick Judd , Xiaojie Zhang , Mikhail Isaev , and Paulius Micikevicius . Integer quantization for deep learning inference: Principles and empirical evaluation. arXiv preprint arXiv:2004. *09602* , 2020.

Yash Bhalgat , Jinwon Lee , Markus Nagel , Tijmen Blankevoort , and Nojun Kwak . Lsq+: Improving low-bit quantization through learnable offsets and better initialization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* , pages 696–697, 2020.

Jeffrey L McKinstry , Steven K Esser , Rathinakumar Appuswamy , Deepika Bablani , John V Arthur , Izzet B Yildiz , and Dharmendra S Modha. Discovering low-precision networks close to full-precision networks for efficient embedded inference. arXiv preprint arXiv:1809.04191, 2018.

Szymon Migacz . Nvidia 8-bit inference with tensorrt. *GPU Technology Conference* , 2017.

Zhewei Yao , Zhen Dong , Zhangcheng Zheng , Amir Gholami , Jiali Yu , Eric Tan , Leyuan Wang , Qijing Huang , Yida Wang , Michael W Mahoney , et al. Hawqv3: Dyadic neural network quantization. *International Conference on Machine Learning* , 2021.

Wonyong Sung , Sungho Shin , and Kyuyeon Hwang . Resiliency of deep neural networks under quantization. arXiv preprint arXiv:1511.06488, 2015.

Sungho Shin , Kyuyeon Hwang , and Wonyong Sung . Fixed-point performance analysis of recurrent neural networks. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* , pages 976–980. IEEE, 2016.

Yoni Choukroun , Eli Kravchik , Fan Yang , and Pavel Kisilev . Low-bit quantization of neural networks for efficient inference. In ICCV Workshops, pages 3009–3018, 2019.

Ritchie Zhao , Yuwei Hu , Jordan Dotzel , Christopher De Sa , and Zhiru Zhang . Improving neural network quantization without retraining using outlier channel splitting. *Proceedings of Machine Learning Research* , 2019.

Rundong Li , Yan Wang , Feng Liang , Hongwei Qin , Junjie Yan , and Rui Fan . Fully quantized network for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* , 2019.

Jungwook Choi , Zhuo Wang , Swagath Venkataramani , Pierce I-Jen Chuang , Vijayalakshmi Srinivasan , and Kailash Gopalakrishnan . Pact: Parameterized clipping activation for quantized neural networks. arXiv preprint arXiv:1805. *06085* , 2018.

Chenzhuo Zhu , Song Han , Huizi Mao , and William J Dally . Trained ternary quantization. arXiv preprint arXiv:1612. *01064* , 2016.

Dongqing Zhang , Jiaolong Yang , Dongqiangzi Ye , and Gang Hua . Lq-nets: Learned quantization for highly accurate and compact deep neural networks. In European conference on computer vision (ECCV *)*, 2018.

Steven K Esser , Jeffrey L McKinstry , Deepika Bablani , Rathinakumar Appuswamy , and Dharmendra S Modha . Learned step size quantization. arXiv preprint arXiv:1902. *08153* , 2019.

Sheng Shen , Zhen Dong , Jiayu Ye , Linjian Ma , Zhewei Yao , Amir Gholami , Michael W Mahoney , and Kurt Keutzer . Q-BERT: Hessian based ultra low precision quantization of bert. In AAAI, pages 8815–8821, 2020.

Ashish Vaswani , Noam Shazeer , Niki Parmar , Jakob Uszkoreit , Llion Jones , Aidan N Gomez , Łukasz Kaiser , and Illia Polosukhin . Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.

Shuchang Zhou , Zekun Ni , Xinyu Zhou , He Wen , Yuxin Wu , and Yuheng Zou . Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. arXiv preprint arXiv:1606. *06160* , 2016.

Qijing Huang , Dequan Wang , Zhen Dong , Yizhao Gao , Yaohui Cai , Tian Li , Bichen Wu , Kurt Keutzer , and John Wawrzynek . Codenet: Efficient deployment of input-adaptive object detection on embedded fpgas. In The 2021 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, pages 206–216, 2021.

Moran Shkolnik , Brian Chmiel , Ron Banner , Gil Shomron , Yuri Nahshan , Alex Bronstein , and Uri Weiser . Robust quantization: One model to rule them all. *Advances in neural information processing systems* , 2020.

Yunchao Gong , Liu Liu , Ming Yang , and Lubomir Bourdev . Compressing deep convolutional networks using vector quantization. arXiv preprint arXiv:1412. *6115* , 2014.

Jiaxiang Wu , Cong Leng , Yuhang Wang , Qinghao Hu , and Jian Cheng . Quantized convolutional neural networks for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* , pages 4820–4828, 2016.

Lu Hou , Quanming Yao , and James T Kwok . Loss-aware binarization of deep networks. arXiv preprint arXiv:1611. *01600* , 2016.

Zhouhan Lin , Matthieu Courbariaux , Roland Memisevic , and Yoshua Bengio . Neural networks with few multiplications. arXiv preprint arXiv:1510. *03009* , 2015.

Daisuke Miyashita , Edward H Lee , and Boris Murmann . Convolutional neural networks using logarithmic data representation. arXiv preprint arXiv:1603. *01025* , 2016.

Yoojin Choi , Mostafa El-Khamy , and Jungwon Lee . Towards the limit of network quantization. arXiv preprint arXiv:1612. *01543* , 2016.

Zhaowei Cai , Xiaodong He , Jian Sun , and Nuno Vasconcelos . Deep learning with low precision by half-wave gaussian quantization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* , pages 5918–5926, 2017.

Eunhyeok Park , Sungjoo Yoo , and Peter Vajda . Value-aware quantization for training and inference of neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)* , pages 580–595, 2018.

Peisong Wang , Qinghao Hu , Yifan Zhang , Chunjie Zhang , Yang Liu , and Jian Cheng . Two-step quantization for low-bit neural networks. In *Proceedings of the IEEE Conference on computer vision and pattern recognition* , pages 4376–4384, 2018.

Yongkweon Jeon , Baeseong Park , Se Jung Kwon , Byeongwook Kim , Jeongin Yun , and Dongsoo Lee . Biqgemm: matrix multiplication with lookup table for binary-coding-based quantized dnns. arXiv preprint arXiv:2005.09904, 2020.

Sangil Jung , Changyong Son , Seohyung Lee , Jinwoo Son , Jae-Joon Han , Youngjun Kwak , Sung Ju Hwang , and Changkyu Choi . Learning to quantize deep networks by optimizing quantization intervals with task loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , pages 4350–4359, 2019.

Jiwei Yang , Xu Shen , Jun Xing , Xinmei Tian , Houqiang Li , Bing Deng , Jianqiang Huang , and Xian-sheng Hua . Quantization networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , pages 7308–7316, 2019.

Julian Faraone , Nicholas Fraser , Michaela Blott , and Philip HW Leong . Syq: Learning symmetric quantization for efficient deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* , pages 4300–4309, 2018.

Frederick Tung and Greg Mori . Clip-q: Deep network compression learning by in-parallel pruning-quantization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* , pages 7873–7882, 2018.

Aojun Zhou , Anbang Yao , Kuan Wang , and Yurong Chen . Explicit loss-error-aware quantization for low-bit deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* , pages 9426–9435, 2018.

Zhaohui Yang , Yunhe Wang , Kai Han , Chunjing Xu , Chao Xu , Dacheng Tao , and Chang Xu . Searching for low-bit weights in quantized neural networks. Advances in neural information processing systems, 2020.

Zhenyu Liao , Romain Couillet , and Michael W Mahoney . Sparse quantized spectral clustering. *International Conference on Learning Representations* , 2021.

Chaim Baskin , Eli Schwartz , Evgenii Zheltonozhskii , Natan Liss , Raja Giryes , Alex M Bronstein , and Avi Mendelson . Uniq: Uniform noise injection for non-uniform quantization of neural networks. arXiv preprint arXiv:1804. *10969* , 2018.

Yuhang Li , Xin Dong , and Wei Wang . Additive powers-of-two quantization: An efficient non-uniform discretization for neural networks. arXiv preprint arXiv:1909. *13144* , 2019.

Shubham Jain , Swagath Venkataramani , Vijayalakshmi Srinivasan , Jungwook Choi , Kailash Gopalakrishnan , and Leland Chang . Biscaled-dnn: Quantizing long-tailed datastructures with two scale factors for deep neural networks. In *2019 56th ACM/IEEE Design Automation Conference (DAC)* , pages 1–6. IEEE, 2019.

Jun Fang , Ali Shafiee , Hamzah Abdel-Aziz , David Thorsley , Georgios Georgiadis , and Joseph H Hassoun . Post-training piecewise linear quantization for deep neural networks. In European Conference on Computer Vision, pages 69–86. Springer, 2020.

Aojun Zhou , Anbang Yao , Yiwen Guo , Lin Xu , and Yurong Chen . Incremental network quantization: Towards lossless cnns with low-precision weights. arXiv preprint arXiv:1702. *03044* , 2017.

Chen Xu , Jianqiang Yao , Zhouchen Lin , Wenwu Ou , Yuanbin Cao , Zhirong Wang , and Hongbin Zha . Alternating multi-bit quantization for recurrent neural networks. arXiv preprint arXiv:1802. *00150* , 2018.

Yiwen Guo , Anbang Yao , Hao Zhao , and Yurong Chen . Network sketching: Exploiting binary structure in deep cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* , pages 5955–5963, 2017.

Itay Hubara , Matthieu Courbariaux , Daniel Soudry , Ran Elyaniv , and Yoshua Bengio. Binarized neural networks. *neural information processing systems* , pages 4107–4115, 2016.

Wei Tang , Gang Hua , and Liang Wang . How to train a compact binary neural network with high accuracy? In *Proceedings of the AAAI Conference on Artificial Intelligence* , volume 31, 2017.

Xiaofan Lin , Cong Zhao , and Wei Pan . Towards accurate binary convolutional neural network. arXiv preprint arXiv:1711. *11294* , 2017.

Matthieu Courbariaux , Yoshua Bengio , and Jean-Pierre David . BinaryConnect: Training deep neural networks with binary weights during propagations. In *Advances in neural information processing systems* , pages 3123–3131, 2015.

Mohammad Rastegari , Vicente Ordonez , Joseph Redmon , and Ali Farhadi . Xnor-net: Imagenet classification using binary convolutional neural networks. 2016.

Philipp Gysel , Mohammad Motamedi , and Soheil Ghiasi . Hardware-oriented approximation of convolutional neural networks. arXiv preprint arXiv:1604.03168, 2016.

Philipp Gysel , Jon Pimentel , Mohammad Motamedi , and Soheil Ghiasi . Ristretto: A framework for empirical study of resource-efficient inference in convolutional neural networks. IEEE transactions on neural networks and learning systems, 29(11):5784–5789, 2018.

Shyam A Tailor , Javier Fernandez-Marques , and Nicholas D Lane . Degree-quant: Quantization-aware training for graph neural networks. *International Conference on Learning Representations* , 2021.

Renkun Ni , Hongmin Chu , Oscar Castañeda , Ping-yeh Chiang , Christoph Studer , and Tom Goldstein . Wrapnet: Neural net inference with ultra-low-resolution arithmetic. arXiv preprint arXiv:2007.13242, 2020.

Yu Bai , Yu-Xiang Wang , and Edo Liberty . Proxquant: Quantized neural networks via proximal operators. *International Conference on Learning Representations* , 2018.

Penghang Yin , Jiancheng Lyu , Shuai Zhang , Stanley Osher , Yingyong Qi , and Jack Xin . Understanding straight-through estimator in training activation quantized neural nets. arXiv preprint arXiv:1903. *05662* , 2019.

Frank Rosenblatt . The perceptron, a perceiving and recognizing automaton Project Para. Cornell Aeronautical Laboratory, 1957.

Frank Rosenblatt . Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, Cornell Aeronautical Lab Inc Buffalo NY, 1961.

Bohan Zhuang , Chunhua Shen , Mingkui Tan , Lingqiao Liu , and Ian D Reid . Towards effective low-bitwidth convolutional neural networks. *computer vision and pattern recognition* , pages 7920–7928, 2018.

Pierre Stock , Angela Fan , Benjamin Graham , Edouard Grave , Rémi Gribonval , Herve Jegou , and Armand Joulin . Training with quantization noise for extreme model compression. In *International Conference on Learning Representations* , 2021.

Shangyu Chen , Wenya Wang , and Sinno Jialin Pan . Metaquant: Learning to quantize by learning to penetrate non-differentiable quantization. In H. Wallach , H. Larochelle , A. Beygelzimer , F. Alché-Buc , E. Fox , and R. Garnett , editors, Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019.

Angela Fan , Pierre Stock , Benjamin Graham , Edouard Grave , Rémi Gribonval , Hervé Jégou , and Armand Joulin . Training with quantization noise for extreme model compression. arXiv e-prints, pages arXiv–2004, 2020.

Zechun Liu , Baoyuan Wu , Wenhan Luo , Xin Yang , Wei Liu , and Kwang-Ting Cheng . Bi-real net: Enhancing the performance of 1-bit cnns with improved representational capability and advanced training algorithm. In *Proceedings of the European conference on computer vision (ECCV)* , pages 722–737, 2018.

Cong Leng , Zesheng Dou , Hao Li , Shenghuo Zhu , and Rong Jin . Extremely low bit neural network: Squeeze the last bit out with admm. In *Proceedings of the AAAI Conference on Artificial Intelligence* , volume 32, 2018.

Eirikur Agustsson and Lucas Theis . Universally quantized neural compression. *Advances in neural information processing systems* , 2020.

Abram L Friesen and Pedro Domingos . Deep learning as a mixed convex-combinatorial optimization problem. arXiv preprint arXiv:1710. *11573* , 2017.

Dong-Hyun Lee , Saizheng Zhang , Asja Fischer , and Yoshua Bengio . Difference target propagation. In *Joint european conference on machine learning and knowledge discovery in databases*, pages 498–515. Springer, 2015.

Eric Jang , Shixiang Gu , and Ben Poole . Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611. *01144* , 2016.

Yoojin Choi , Mostafa El-Khamy , and Jungwon Lee . Learning low precision deep neural networks through regularization. arXiv preprint arXiv:1809.00095, 2, 2018.

Maxim Naumov , Utku Diril , Jongsoo Park , Benjamin Ray , Jedrzej Jablonski , and Andrew Tulloch . On periodic functions as regularizers for quantization of neural networks. arXiv preprint arXiv:1811. *09862* , 2018.

Milad Alizadeh , Arash Behboodi , Mart van Baalen , Christos Louizos , Tijmen Blankevoort , and Max Welling . Gradient l1 regularization for quantization robustness. arXiv preprint arXiv:2002. *07520* , 2020.

Lei Deng , Peng Jiao , Jing Pei , Zhenzhi Wu , and Guoqi Li . Gxnor-net: Training deep neural networks with ternary weights and activations without full-precision memory under a unified discretization framework. Neural Networks, 100:49–58, 2018.

Zhi-Gang Liu and Matthew Mattina . Learning low-precision neural networks without straight-through estimator (STE). arXiv preprint arXiv:1903. *01061* , 2019.

Markus Nagel , Rana Ali Amjad , Mart Van Baalen , Christos Louizos , and Tijmen Blankevoort . Up or down? adaptive rounding for post-training quantization. In *International Conference on Machine Learning* , pages 7197–7206. PMLR, 2020.

Ron Banner , Yury Nahshan , Elad Hoffer , and Daniel Soudry . Post-training 4-bit quantization of convolution networks for rapid-deployment. arXiv preprint arXiv:1810. *05723* , 2018.

Eldad Meller , Alexander Finkelstein , Uri Almog , and Mark Grobman . Same, same but different: Recovering neural network quantization error through weight factorization. In *International Conference on Machine Learning* , pages 4486–4495. PMLR, 2019.

Jun Fang , Ali Shafiee , Hamzah Abdel-Aziz , David Thorsley , Georgios Georgiadis , and Joseph Hassoun . Near-lossless post-training quantization of deep neural networks via a piecewise linear approximation. arXiv preprint arXiv:2002. *00104* , 2020.

Jun Haeng Lee , Sangwon Ha , Saerom Choi , Won-Jo Lee , and Seungwon Lee . Quantization for rapid deployment of deep neural networks. arXiv preprint arXiv:1810. *05488* , 2018.

Markus Nagel , Mart van Baalen , Tijmen Blankevoort , and Max Welling . Data-free quantization through weight equalization and bias correction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* , pages 1325–1334, 2019.

Yaohui Cai , Zhewei Yao , Zhen Dong , Amir Gholami , Michael W Mahoney , and Kurt Keutzer . Zeroq: A novel zero shot quantization framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , pages 13169–13178, 2020.

Yuhang Li , Ruihao Gong , Xu Tan , Yang Yang , Peng Hu , Qi Zhang , Fengwei Yu, Wei Wang , and Shi Gu . Brecq: Pushing the limit of post-training quantization by block reconstruction. *International Conference on Learning Representations* , 2021.

Xiangyu He and Jian Cheng . Learning compression from limited unlabeled data. In *Proceedings of the European Conference on Computer Vision (ECCV)* , pages 752–769, 2018.

Sahaj Garg , Anirudh Jain , Joe Lou , and Mitchell Nahmias . Confounding tradeoffs for neural network quantization. arXiv preprint arXiv:2102. *06366* , 2021.

Sahaj Garg , Joe Lou , Anirudh Jain , and Mitchell Nahmias . Dynamic precision analog computing for neural networks. *arXiv preprint arXiv:2102.06365* , 2021.

Itay Hubara , Yury Nahshan , Yair Hanani , Ron Banner , and Daniel Soudry . Improving post training neural quantization: Layer-wise calibration and integer programming. arXiv preprint arXiv:2006. *10518* , 2020.

Alexander Finkelstein , Uri Almog , and Mark Grobman . Fighting quantization bias with bias. arXiv preprint arXiv:1906. *03193* , 2019.

Matan Haroush , Itay Hubara , Elad Hoffer , and Daniel Soudry . The knowledge within: Methods for data-free model compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , pages 8494–8502, 2020.

Jacob Devlin , Ming-Wei Chang , Kenton Lee , and Kristina Toutanova . Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810. *04805* , 2018.

Dan Hendrycks and Kevin Gimpel . Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606* . *08415*, 2016.

Prajit Ramachandran , Barret Zoph , and Quoc V Le . Swish: a self-gated activation function. arXiv preprint arXiv:1710.05941, 7:1, 2017.

Hanting Chen , Yunhe Wang , Chang Xu , Zhaohui Yang , Chuanjian Liu , Boxin Shi , Chunjing Xu , Chao Xu , and Qi Tian . Data-free learning of student networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* , pages 3514–3522, 2019.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. arXiv preprint arXiv:1406. *2661* , 2014.

Sergey Ioffe and Christian Szegedy . Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International conference on machine learning* , pages 448–456, 2015.

Yoojin Choi , Jihwan Choi , Mostafa El-Khamy , and Jungwon Lee . Data-free network quantization with adversarial knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* , pages 710–711, 2020.

Shoukai Xu , Haokun Li , Bohan Zhuang , Jing Liu , Jiezhang Cao , Chuangrun Liang , and Mingkui Tan . Generative low-bitwidth data free quantization. In *European Conference on Computer Vision*, pages 1–17. Springer, 2020.

Xiangyu He , Qinghao Hu , Peisong Wang , and Jian Cheng . Generative zero-shot network quantization. arXiv preprint arXiv:2101. *08430* , 2021.

Jianfei Chen , Yu Gai , Zhewei Yao , Michael W Mahoney , and Joseph E Gonzalez . A statistical framework for low-bitwidth training of deep neural networks. arXiv preprint arXiv:2010. *14298* , 2020.

Sehoon Kim , Amir Gholami , Zhewei Yao , Michael W Mahoney , and Kurt Keutzer . I-bert: Integer-only bert quantization. *International conference on machine learning* , 2021.

Darryl Lin , Sachin Talathi , and Sreekanth Annapureddy . Fixed point quantization of deep convolutional networks. In *International conference on machine learning* , pages 2849–2858. PMLR, 2016.

Mark Horowitz . 1.1 computing's energy problem (and what we can do about it). In 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), pages 10–14. IEEE, 2014.

Jimmy Lei Ba , Jamie Ryan Kiros , and Geoffrey E Hinton . Layer normalization. arXiv preprint arXiv:1607. *06450* , 2016.

Maxim Naumov , Dheevatsa Mudigere , Hao-Jun Michael Shi , Jianyu Huang , Narayanan Sundaraman , Jongsoo Park , Xiaodong Wang , Udit Gupta , Carole-Jean Wu , Alisson G Azzolini , et al. Deep learning recommendation model for personalization and recommendation systems. arXiv preprint arXiv:1906. *00091* , 2019.

Yiren Zhou , Seyed-Mohsen Moosavi-Dezfooli , Ngai-Man Cheung , and Pascal Frossard . Adaptive quantization for deep neural network. arXiv preprint arXiv:1712. *01048* , 2017.

Kuan Wang , Zhijian Liu , Yujun Lin , Ji Lin , and Song Han . HAQ: Hardware-aware automated quantization. *In Proceedings of the IEEE conference on computer vision and pattern recognition* , 2019.

Zhen Dong , Zhewei Yao , Amir Gholami , Michael W Mahoney , and Kurt Keutzer . Hawq: Hessian aware quantization of neural networks with mixed-precision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* , pages 293–302, 2019.

Mart van Baalen , Christos Louizos , Markus Nagel , Rana Ali Amjad , Ying Wang , Tijmen Blankevoort , and Max Welling . Bayesian bits: Unifying quantization and pruning. *Advances in neural information processing systems* , 2020.

Huanrui Yang , Lin Duan , Yiran Chen , and Hai Li . Bsq: Exploring bit-level sparsity for mixed-precision neural network quantization. arXiv preprint arXiv:2102. *10462* , 2021.

Zhongnan Qu , Zimu Zhou , Yun Cheng , and Lothar Thiele . Adaptive loss-aware quantization for multi-bit networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* , June 2020.

Tianzhe Wang , Kuan Wang , Han Cai , Ji Lin , Zhijian Liu , Hanrui Wang , Yujun Lin , and Song Han . Apq: Joint search for network architecture, pruning and quantization policy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , pages 2078–2087, 2020.

Peng Hu , Xi Peng , Hongyuan Zhu , Mohamed M Sabry Aly , and Jie Lin . Opq: Compressing deep neural networks with one-shot pruning-quantization. 2021.

Lin Ning , Guoyang Chen , Weifeng Zhang , and Xipeng Shen . Simple augmentation goes a long way: {ADRL} for {dnn} quantization. In *International Conference on Learning Representations* , 2021.

Hai Victor Habi , Roy H Jennings , and Arnon Netzer . Hmq: Hardware friendly mixed precision quantization block for cnns. arXiv preprint arXiv:2007. *09952* , 2020.

Manuele Rusci , Marco Fariselli , Alessandro Capotondi , and Luca Benini . Leveraging automated mixed-low-precision quantization for tiny edge microcontrollers. In IoT Streams for Data-Driven Predictive Maintenance and IoT, Edge, and Mobile for Embedded Machine Learning, pages 296–308. Springer, 2020.

Bichen Wu , Yanghan Wang , Peizhao Zhang , Yuandong Tian , Peter Vajda , and Kurt Keutzer . Mixed precision quantization of convnets via differentiable neural architecture search. arXiv preprint arXiv:1812. *00090* , 2018.

Zhen Dong , Zhewei Yao , Daiyaan Arfeen , Amir Gholami , Michael W. Mahoney , and Kurt Keutzer . HAWQ-V2: Hessian aware trace-weighted quantization of neural networks. *Advances in neural information processing systems* , 2020.

Tien-Ju Yang , Andrew Howard , Bo Chen , Xiao Zhang , Alec Go , Mark Sandler , Vivienne Sze , and Hartwig Adam . Netadapt: Platform-aware neural network adaptation for mobile applications. In *Proceedings of the European Conference on Computer Vision (ECCV)* , pages 285–300, 2018.

Ying Wang , Yadong Lu , and Tijmen Blankevoort . Differentiable joint pruning and quantization for hardware efficiency. In European Conference on Computer Vision, pages 259–277. Springer, 2020.

Benjamin Hawks , Javier Duarte , Nicholas J Fraser , Alessandro Pappalardo , Nhan Tran , and Yaman Umuroglu . Ps and qs: Quantization-aware pruning for efficient low latency neural network inference. arXiv preprint arXiv:2102. *11289* , 2021.

Prad Kadambi , Karthikeyan Natesan Ramamurthy , and Visar Berisha . Comparing fisher information regularization with distillation for dnn quantization. Advances in neural information processing systems, 2020.

Jianming Ye , Shiliang Zhang , and Jingdong Wang . Distillation guided residual learning for binary convolutional neural networks. arXiv preprint arXiv:2007. *05223* , 2020.

Wonpyo Park , Dongju Kim , Yan Lu , and Minsu Cho . Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , pages 3967–3976, 2019.

Shan You , Chang Xu , Chao Xu , and Dacheng Tao . Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* , pages 1285–1294, 2017.

Antti Tarvainen and Harri Valpola . Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. arXiv preprint arXiv:1703. *01780* , 2017.

Elliot J Crowley , Gavin Gray , and Amos J Storkey . Moonshine: Distilling with cheap convolutions. In NeurIPS, pages 2893–2903, 2018.

Linfeng Zhang , Jiebo Song , Anni Gao , Jingwei Chen , Chenglong Bao , and Kaisheng Ma . Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* , pages 3713–3722, 2019.

Minje Kim and Paris Smaragdis . Bitwise neural networks. arXiv preprint arXiv:1601.06071, 2016.

Xiangguo Zhang , Haotong Qin , Yifu Ding , Ruihao Gong , Qinghua Yan , Renshuai Tao , Yuhang Li , Fengwei Yu , and Xianglong Liu . Diversifying sample generation for accurate data-free quantization. *CVPR* , 2021.

Se Jung Kwon , Dongsoo Lee , Byeongwook Kim , Parichay Kapoor , Baeseong Park , and Gu-Yeon Wei . Structured compression by weight encryption for unstructured pruning and quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , pages 1909–1918, 2020.

Haichuan Yang , Shupeng Gui , Yuhao Zhu , and Ji Liu . Automatic neural network compression by sparsity-quantization joint learning: A constrained optimization-based approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , pages 2178–2188, 2020.

Qing Jin , Linjie Yang , and Zhenyu Liao . Adabits: Neural network quantization with adaptive bit-widths. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , pages 2146–2156, 2020.

Haotong Qin , Ruihao Gong , Xianglong Liu , Mingzhu Shen , Ziran Wei , Fengwei Yu , and Jingkuan Song . Forward and backward information retention for accurate binary neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , pages 2250–2259, 2020.

Bohan Zhuang , Chunhua Shen , Mingkui Tan , Lingqiao Liu , and Ian Reid . Structured binary neural networks for accurate image classification and semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , pages 413–422, 2019.

Ziwei Wang , Jiwen Lu , Chenxin Tao , Jie Zhou , and Qi Tian . Learning channel-wise interactions for binary convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , pages 568–577, 2019.

Chunlei Liu , Wenrui Ding , Xin Xia , Baochang Zhang , Jiaxin Gu , Jianzhuang Liu , Rongrong Ji , and David Doermann . Circulant binary convolutional networks: Enhancing the performance of 1-bit dcnns with circulant back propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , pages 2691–2699, 2019.

Shilin Zhu , Xin Dong , and Hao Su . Binary ensemble neural network: More bits per network or more networks per bit? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , pages 4923–4932, 2019.

Yinghao Xu , Xin Dong , Yudian Li , and Hao Su . A main/subsidiary network framework for simplifying binary neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , pages 7154–7162, 2019.

Zhezhi He and Deliang Fan . Simultaneously optimizing weight and quantizer of ternary neural network using truncated gaussian approximation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , pages 11438–11446, 2019.

Felix Juefei-Xu , Vishnu Naresh Boddeti , and Marios Savvides . Local binary convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* , pages 19–28, 2017.

Yueqi Duan , Jiwen Lu , Ziwei Wang , Jianjiang Feng , and Jie Zhou . Learning deep binary descriptor with multi-quantization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* , pages 1183–1192, 2017.

Xiaodi Wang , Baochang Zhang , Ce Li , Rongrong Ji , Jungong Han , Xianbin Cao , and Jianzhuang Liu . Modulated convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* , pages 840–848, 2018.

Qinghao Hu , Gang Li , Peisong Wang , Yifan Zhang , and Jian Cheng . Training binary weight networks via semi-binary decomposition. In *Proceedings of the European Conference on Computer Vision (ECCV)* , pages 637–653, 2018.

Koen Helwegen , James Widdicombe , Lukas Geiger , Zechun Liu , Kwang-Ting Cheng , and Roeland Nusselder . Latent weights do not exist: Rethinking binarized neural network optimization. *Advances in neural information processing systems* , 2019.

Dongsoo Lee , Se Jung Kwon , Byeongwook Kim , Yongkweon Jeon , Baeseong Park , and Jeongin Yun . Flexor: Trainable fractional quantization *Advances in neural information processing systems* , 2020.

Alexander Shekhovtsov , Viktor Yanush , and Boris Flach . Path sample-analytic gradient estimators for stochastic binary networks. *Advances in neural information processing systems* , 2020.

Kai Jia and Martin Rinard . Efficient exact verification of binarized neural networks. Advances in neural information processing systems, 2020.

Mingbao Lin , Rongrong Ji , Zihan Xu , Baochang Zhang , Yan Wang , Yongjian Wu , Feiyue Huang , and Chia-Wen Lin . Rotated binary neural network. Advances in neural information processing systems, 2020.

Hyungjun Kim , Kyungsu Kim , Jinseok Kim , and Jae-Joon Kim . Binaryduo: Reducing gradient mismatch in binary activation network by coupling binary activations. *International Conference on Learning Representations* , 2020.

Yuhang Li , Ruihao Gong , Fengwei Yu , Xin Dong , and Xianglong Liu . Dms: Differentiable dimension search for binary neural networks. *International Conference on Learning Representations* , 2020.

Kai Han , Yunhe Wang , Yixing Xu , Chunjing Xu , Enhua Wu , and Chang Xu . Training binary neural networks through learning with noisy supervision. In *International Conference on Machine Learning* , pages 4017–4026. PMLR, 2020.

Haotong Qin , Zhongang Cai , Mingyuan Zhang , Yifu Ding , Haiyu Zhao , Shuai Yi, Xianglong Liu , and Hao Su . Bipointnet: Binary neural network for point clouds. *International Conference on Learning Representations* , 2021.

Adrian Bulat , Brais Martinez , and Georgios Tzimiropoulos . High-capacity expert binary networks. *International Conference on Learning Representations* , 2021.

James Diffenderfer and Bhavya Kailkhura . Multi-prize lottery ticket hypothesis: Finding accurate binary neural networks by pruning a randomly weighted network. In *International Conference on Learning Representations* , 2021.

Cheng Fu , Shilin Zhu , Hao Su , Ching-En Lee , and Jishen Zhao . Towards fast and energy-efficient binarized neural network inference on fpga. In FPGA, 2019.

Jiaxin Gu , Junhe Zhao , Xiaolong Jiang , Baochang Zhang , Jianzhuang Liu , Guodong Guo , and Rongrong Ji . Bayesian optimized 1-bit cnns. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* , pages 4909–4917, 2019.

Mingzhu Shen , Kai Han , Chunjing Xu , and Yunhe Wang . Searching for accurate binary neural architectures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops* , 2019.

Milad Alizadeh , Javier Fernández-Marqués , Nicholas D Lane , and Yarin Gal . An empirical study of binary neural networks' optimisation. In *International Conference on Learning Representations* , 2018.

Fengfu Li , Bo Zhang , and Bin Liu . Ternary weight networks. arXiv preprint arXiv:1605.04711, 2016.

Diwen Wan , Fumin Shen , Li Liu , Fan Zhu , Jie Qin , Ling Shao , and Heng Tao Shen . Tbn: Convolutional neural network with ternary inputs and binary weights. In *Proceedings of the European Conference on Computer Vision (ECCV)* , pages 315–332, 2018.

Haotong Qin , Ruihao Gong , Xianglong Liu , Xiao Bai , Jingkuan Song , and Nicu Sebe . Binary neural networks: A survey. Pattern Recognition, 105:107281, 2020.

Zefan Li , Bingbing Ni , Wenjun Zhang , Xiaokang Yang , and Wen Gao . Performance guaranteed network acceleration via high-order residual quantization. In *Proceedings of the IEEE international conference on computer vision* , pages 2584–2592, 2017.

Qinghao Hu , Peisong Wang , and Jian Cheng . From hashing to cnns: Training binary weight networks via hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence* , volume 32, 2018.

Asit Mishra , Eriko Nurvitadhi , Jeffrey J Cook , and Debbie Marr . WRPN: Wide reduced-precision networks. In *International Conference on Learning Representations* , 2018.

Ting-Wu Chin , Pierce I-Jen Chuang , Vikas Chandra , and Diana Marculescu . One weight bitwidth to rule them all. *Proceedings of the European Conference on Computer Vision (ECCV)* , 2020.

Mingzhu Shen , Xianglong Liu , Ruihao Gong , and Kai Han . Balanced binary neural networks with gated residual. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* , pages 4197–4201. IEEE, 2020.

Adrian Bulat and Georgios Tzimiropoulos . Xnor-net++: Improved binary neural networks. *British Machine Vision Conference* , 2019.

Brais Martinez , Jing Yang , Adrian Bulat , and Georgios Tzimiropoulos . Training binary neural networks with real-to-binary convolutions. arXiv preprint arXiv:2003. *11535* , 2020.

Lu Hou and James T. Kwok . Loss-aware weight quantization of deep networks. In *International Conference on Learning Representations* , 2018.

Ruizhou Ding , Ting-Wu Chin , Zeye Liu , and Diana Marculescu . Regularizing activation distribution for training binarized deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , pages 11408–11417, 2019.

Xiuyi Chen , Guangcan Liu , Jing Shi , Jiaming Xu , and Bo Xu . Distilled binary neural network for monaural speech separation. In *International Joint Conference on Neural Networks (IJCNN)* , pages 1–8. IEEE, 2018.

Sajad Darabi , Mouloud Belbahri , Matthieu Courbariaux , and Vahid Partovi Nia . Bnn+: Improved binary network training. 2018.

Ruihao Gong , Xianglong Liu , Shenghu Jiang , Tianxiang Li , Peng Hu , Jiazhen Lin , Fengwei Yu , and Junjie Yan . Differentiable soft quantization: Bridging full-precision and low-bit neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* , pages 4852–4861, 2019.

Adrian Bulat , Georgios Tzimiropoulos , Jean Kossaifi , and Maja Pantic . Improved training of binary networks for human pose estimation and image recognition. arXiv preprint arXiv:1904.05868, 2019.

Zhe Xu and Ray CC Cheung . Accurate and compact convolutional neural networks with trained binarization. *British Machine Vision Conference* , 2019.

Penghang Yin , Shuai Zhang , Jiancheng Lyu , Stanley Osher , Yingyong Qi , and Jack Xin . Blended coarse gradient descent for full quantization of deep neural networks. Research in the Mathematical Sciences, 6(1):14, 2019.

Wei Zhang , Lu Hou , Yichun Yin , Lifeng Shang , Xiao Chen , Xin Jiang , and Qun Liu . Ternarybert: Distillation-aware ultra-low bit bert. arXiv preprint arXiv:2009.12812, 2020.

Haoli Bai , Wei Zhang , Lu Hou , Lifeng Shang , Jing Jin , Xin Jiang , Qun Liu , Michael Lyu , and Irwin King . Binarybert: Pushing the limit of bert quantization. arXiv preprint arXiv:2012.15701, 2020.

Jing Jin , Cai Liang , Tiancheng Wu , Liqin Zou , and Zhiliang Gan . Kdlsq-bert: A quantized bert combining knowledge distillation with learned step size quantization. arXiv preprint arXiv:2101.05938, 2021.

Yinhan Liu , Myle Ott , Naman Goyal , Jingfei Du , Mandar Joshi , Danqi Chen , Omer Levy , Mike Lewis , Luke Zettlemoyer , and Veselin Stoyanov . RoBERTa: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

Alec Radford , Karthik Narasimhan , Tim Salimans , and Ilya Sutskever . Improving language understanding by generative pre-training, 2018.

Alec Radford , Jeffrey Wu , Rewon Child , David Luan , Dario Amodei , and Ilya Sutskever . Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.

Tom B Brown , Benjamin Mann , Nick Ryder , Melanie Subbiah , Jared Kaplan , Prafulla Dhariwal , Arvind Neelakantan , Pranav Shyam , Girish Sastry , Amanda Askell , et al. Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 2020.

Dana Harry Ballard . An introduction to natural computation. MIT press, 1999.

Herve Jegou , Matthijs Douze , and Cordelia Schmid . Product quantization for nearest neighbor search. IEEE transactions on pattern analysis and machine intelligence, 33(1):117–128, 2010.

Eirikur Agustsson , Fabian Mentzer , Michael Tschannen , Lukas Cavigelli , Radu Timofte , Luca Benini , and Luc Van Gool . Soft-to-hard vector quantization for end-to-end learning compressible representations. arXiv preprint arXiv:1704.00648, 2017.

Julieta Martinez , Shobhit Zakhmi , Holger H Hoos , and James J Little . Lsq++: Lower running time and higher recall in multi-codebook quantization. In *Proceedings of the European Conference on Computer Vision (ECCV)* , pages 491–506, 2018.

Lopamudra Mukherjee , Sathya N Ravi , Jiming Peng , and Vikas Singh . A biresolution spectral framework for product quantization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* , pages 3329–3338, 2018.

Kuilin Chen and Chi-Guhn Lee . Incremental few-shot learning via vector quantization in deep embedded space. In *International Conference on Learning Representations* , 2021.

Pierre Stock , Armand Joulin , Rémi Gribonval , Benjamin Graham , and Hervé Jégou . And the bit goes down: Revisiting the quantization of neural networks. arXiv preprint arXiv:1907.05686, 2019.

Liangzhen Lai , Naveen Suda , and Vikas Chandra . CMSIS-NN: Efficient neural network kernels for arm cortex-m cpus. arXiv preprint arXiv:1801.06601, 2018.

Eric Flamand , Davide Rossi , Francesco Conti , Igor Loi , Antonio Pullini , Florent Rotenberg , and Luca Benini . Gap-8: A risc-v soc for ai at the edge of the iot. In *IEEE International Conference on Application-specific Systems, Architectures and Processors (ASAP)* , pages 1–4. IEEE, 2018.

Jeff Johnson . Rethinking floating point for deep learning. arXiv preprint arXiv:1811.01721, 2018.

Naveen Mellempudi , Sudarshan Srinivasan , Dipankar Das , and Bharat Kaul . Mixed precision training with 8-bit floating point. arXiv preprint arXiv:1905.12334, 2019.

Shuang Wu . Training and inference with integers in deep neural networks. *international conference on learning representations* , 2018.

Hamed F Langroudi , Zachariah Carmichael , David Pastuch , and Dhireesha Kudithipudi . Cheetah: Mixed low-precision hardware & software co-design framework for dnns on the edge. arXiv preprint arXiv:1908. *02386* , 2019.

Léopold Cambier , Anahita Bhiwandiwalla , Ting Gong , Mehran Nekuii , Oguz H Elibol , and Hanlin Tang . Shifted and squeezed 8-bit floating point format for low-precision training of deep neural networks. arXiv preprint arXiv:2001.05674, 2020.

Wuwei Lin . Automating optimization of quantized deep learning models on cuda: https://tvm.apache.org/2019/04/29/opt-cuda-quantized, 2019.

Dave Salvator , Hao Wu , Milind Kulkarni , and Niall Emmart . Int4 precision for ai inference: https://developer.nvidia.com/blog/int4-for-ai-inference/, 2019.

Animesh Jain , Shoubhik Bhattacharya , Masahiro Masuda , Vin Sharma , and Yida Wang . Efficient execution of quantized deep learning models: A compiler approach. arXiv preprint arXiv:2006.10226, 2020.