

Contents

List of Figures	v
List of Tables	vii
List of Algorithms	ix
List of Abbreviations	xi
List of Symbols	xvii
1 Introduction	1
1.1 Motivation	1
1.2 Objectives	3
1.3 Contribution	4
1.4 Thesis Outline	5
2 Background	9
2.1 Neural Networks for Computer Vision	9
2.1.1 Convolutional Neural Network	9
2.1.2 Training of Neural Networks	14
2.2 Datasets for Image Classification and Semantic Segmentation	15
2.2.1 Image Classification	15
2.2.2 Semantic Segmentation	17
2.3 Optimization Methods for Convolutional Neural Networks	18
2.3.1 Pruning	19
2.3.2 Quantization	21
2.3.3 Binarization	22
2.3.4 Efficient Convolution Algorithms	23
3 Hardware Aware Performant Perception of Neural Networks	25
3.1 Tripartite Representation of HW-CNN Co-Design	26
3.2 Related Work	30
3.2.1 DarwinAI's Structural Level Deployment	30
3.2.2 The Hardware Vendors Design Process	31
3.2.3 Full-Custom Optimization and Hardware	34
3.3 Design Methodologies for Lightweight Neural Networks	35
3.3.1 HW-Independent Dense Compression of Neural Networks	36
3.3.2 SIMD-based Application of Irregular Pruned and Quantized CNNs	38

3.3.3	OpenCL-based High Throughput CNN Implementation	40
3.3.4	Meet-in-the-Middle Custom Accelerator Design for BNNs	41
4	Resource-aware Optimization	43
4.1	Related Work	43
4.2	Resource-aware Element-wise Pruning and Quantization	46
4.2.1	Binary Search Accelerated Weight Pruning	47
4.2.2	Pruning Order	47
4.2.3	Magnitude-based Pruning and Fine-tuning	49
4.2.4	Clustering-based Weight Sharing	51
4.2.5	Fixed-Point Quantization of Cluster Centroids and Activations	52
4.2.6	Compressed Sparse Weights	53
4.3	Experimental Results	53
4.3.1	Memory-based and Performance-based Compression	53
4.3.2	Low Latency Applications	55
4.3.3	Comparison with State-of-the-Art	56
4.4	Discussion	57
5	Parallelized SIMD-based Algorithm for Pruned Convolutional Layers	59
5.1	Related Work	59
5.2	Dense-Sparse Convolution	61
5.2.1	Dense Convolution	62
5.2.2	Sparse Convolution	65
5.2.3	Kernel-wise Joint Dense-Sparse Convolution	66
5.3	Experimental Results	67
5.3.1	Ablation Study	68
5.3.2	Exploration of Thread-Level Parallelism	70
5.3.3	Comparison with State-of-the-Art	72
5.4	Discussion	73
6	Winograd Convolution and Quantization	75
6.1	Related Work	75
6.2	Winograd Convolution using OpenCL	77
6.2.1	High Level Synthesis Architecture Design	77
6.2.2	WinoCNN Building Blocks and Units	78
6.3	Experimental Results	81
6.3.1	Ablation Study	81
6.3.2	Comparison with State-of-the-Art	83
6.4	Discussion	85
7	ALF: Autoencoder-based Low-Rank Filter-Sharing	87
7.1	Related Work	87
7.2	Sparse Autoencoder-based Filter-wise Pruning	89
7.2.1	Structural Elements of ALF-Block	90

7.2.2	Training Flow of an ALF-Block based CNN	91
7.2.3	Algorithmic Implementation	94
7.2.4	Deployment of an ALF-Block	95
7.3	Experimental Results	95
7.3.1	Design Space Exploration	96
7.3.2	Layer-wise Analysis for an Embedded Application	99
7.3.3	Comparison with State-of-the-Art	100
7.4	Discussion	102
8	Binarized Drivable Area Detection Neural Network	105
8.1	Related Work	105
8.2	Binary Drivable Area Detection	108
8.2.1	Binary Encoder for Feature Extraction	108
8.2.2	Binary Bottleneck with Enlarged Receptive Field	109
8.2.3	Binary Decoder for Semantic Predictions	111
8.2.4	Training and Algorithmic Implementation	111
8.3	Experimental Results	113
8.3.1	Design Space Exploration	113
8.3.2	Training Binary Drivable Area Detection with Automatic Annotations	116
8.3.3	Comparison with the State-of-the-Art	117
8.4	Discussion	119
9	Runtime Configurable Processing Element for BNNs	121
9.1	Related Work	121
9.2	Processing Element Providing Two Modes of Operation	122
9.2.1	Binary Operating Mode	124
9.2.2	Fixed-Point Operating Mode	126
9.2.3	Mode Switching and Partial Sum Accumulation	126
9.3	Experimental Results	128
9.3.1	Resource Utilization Analysis	128
9.3.2	Dynamic Power Analysis	130
9.4	Discussion	131
10	Conclusion	133
	Bibliography	135