

Contents

1	Introduction	1
1.1	Thesis Overview	3
1.1.1	Hardware Acceleration	4
1.1.2	Image Segmentation with Multispectral Data	6
1.1.3	Exploiting Spatio-Temporal Properties for Fast Inference on Video Data	6
1.2	Contributions and Publications	7
1.3	Outline	11
2	Implementation on GPU-Accelerated Off-The-Shelf Plat- forms	13
2.1	Introduction	14
2.2	Related Work	15
2.3	Convolutional Network Classification	17
2.3.1	Scene Labeling / Semantic Segmentation	18
2.3.2	Throughput Measurement	19
2.4	Implementation	21
2.4.1	Proposed Optimizations	22
2.5	Results & Discussion	24
2.6	Summary	26

3	Origami: Hardware Acceleration of Convolutional Networks	31
3.1	Introduction	32
3.2	Convolutional Networks	34
3.2.1	Measuring Computational Complexity	35
3.2.2	Computational Effort	36
3.3	Previous Work	37
3.3.1	Software Implementations (CPU, GPU, DSP)	39
3.3.2	FPGA Implementations	40
3.3.3	ASIC Implementations	41
3.3.4	General Neural Network Accelerators	42
3.3.5	Discussion	43
3.4	Architecture	43
3.4.1	Concept of Operation	45
3.4.2	Optimizing for Area Efficiency	48
3.4.3	Throughput	49
3.4.4	System Architecture	51
3.5	Implementation	54
3.5.1	General Implementation	54
3.5.2	Fixed-Point Analysis	55
3.5.3	System Implementation	56
3.6	Results	57
3.7	Discussion	62
3.7.1	Area Efficiency	62
3.7.2	Bandwidth Efficiency	64
3.7.3	Power Efficiency	64
3.7.4	Effects of Architecture Parameters	66
3.8	Summary	67
4	Hyperdrive: A Systolically Scalable Inference Engine for Binary-Weight CNNs	71
4.1	Introduction	72
4.2	Related Work	74
4.2.1	Software-Programmable Platforms	74
4.2.2	Co-Design of DNN Models and Hardware	75
4.2.3	FPGA and ASIC Accelerators	76
4.3	Hyperdrive Architecture	77
4.4	Computational Model	80

4.4.1	Principles of Operation	81
4.4.2	CNN Mapping	83
4.4.3	Supported Neural Network Topologies	87
4.5	Scalability to Multiple Chips	88
4.5.1	Access Pattern and Storing Scheme of the Border Memories	89
4.5.2	Border and Corner Exchange	90
4.5.3	Border and Corner Memory	91
4.5.4	Interface Implementation	92
4.6	Experimental Results	92
4.6.1	Implementation Results	94
4.6.2	Benchmarking	95
4.6.3	I/O in Multi-Chip Setup	98
4.6.4	Comparison with State-of-the-Art	100
4.7	Summary	102
5	Combinational Implementation of Binarized Neural Net- works	109
5.1	Introduction	110
5.2	Related Work	111
5.3	Combinational Hardware BNN Design	113
5.3.1	Estimating Area	115
5.4	Experimental Results	116
5.4.1	BNN Training	116
5.4.2	Synthesis Results	117
5.4.3	Energy Efficiency Evaluations	118
5.4.4	Scaling to Larger Networks	121
5.5	Summary	121
6	Hardware-Friendly Compression of the Feature Maps	123
6.1	Introduction	124
6.2	Related Work	125
6.3	Compression Algorithm	127
6.3.1	Zero/Non-Zero Encoding with RLE	127
6.3.2	Bit-Plane Compression	129
6.3.3	Hardware Suitability	129
6.4	Results	130
6.4.1	Experimental Setup	130

6.4.2	Sparsity, Activation Histogram & Data Layout	131
6.4.3	Selecting Parameters	133
6.4.4	Total Compression Factor	134
6.5	Summary	135
7	Accuracy-Compute Trade-Off by Fusing Multispectral Imaging Data	137
7.1	Introduction	138
7.2	Related Work	140
7.3	Dataset Collection	141
7.3.1	Image Sensors	141
7.3.2	Image Alignment	142
7.3.3	Data Labeling	143
7.4	Neural Network Architectures	148
7.4.1	Per-pixel Classification	148
7.4.2	Convolutional Network Targeting the Stanford Backgrounds Dataset	149
7.4.3	ResNet-inspired Convolutional Networks	151
7.4.4	Training the Convolutional Networks	151
7.5	Results and Discussion	151
7.6	Summary	156
8	CBinfer: Exploiting Frame-to-Frame Locality for Inference on Static Camera Video Streams	157
8.1	Introduction	158
8.2	Related Work	161
8.2.1	Optimized Embedded System Implementations	161
8.2.2	Approximations Trading Accuracy for Throughput	162
8.2.3	Video-based Computation Reduction	163
8.2.4	Suitable Datasets and Neural Networks	165
8.3	Methodology	166
8.3.1	Processing Steps	169
8.3.2	Memory Requirements	172
8.3.3	Closed-Loop Formulation	173
8.3.4	Fine-Grained Change-based Inference	174
8.3.5	Propagating Changes & Pooling	175
8.3.6	Threshold Selection	176
8.4	Results & Discussion	178

8.4.1	Evaluation Environment	178
8.4.2	Baseline Throughput and Computation Break- down	180
8.4.3	Threshold Selection	180
8.4.4	Throughput Evaluations	182
8.4.5	Accuracy-Throughput Trade-Off	187
8.4.6	Compute Time Breakdown	187
8.4.7	Change Propagation	188
8.4.8	Fine-grained CBinfer	193
8.4.9	Energy Efficiency	194
8.4.10	Comparison to Related Work	196
8.5	Summary	196
9	Summary and Conclusion	199
9.1	Overview of the Main Results	200
9.2	Outlook	202
	Bibliography	205
	Curriculum Vitae	233