

Contents

1	Introduction	1
1.1	Trends in deep learning	5
1.2	Research directions in deep learning	8
1.3	Contributions	14
1.4	Thesis structure	16
1.5	List of publications	18
2	Benchmarking today's systems	21
2.1	Benchmarks for transprecision computing	22
2.1.1	PageRank	24
2.1.2	BLSTM	25
2.1.3	GLQ	28
2.2	Summary and conclusion	29
3	Approximate computing	31
3.1	Approximate computing techniques	33
3.1.1	The use of datatypes	33
3.1.2	Loop perforation	34
3.1.3	Task skipping and memoization	36
3.1.4	Using multiple inexact program versions	37
3.1.5	Stochastic computing	38
3.2	Applying approximate computing	40
3.3	Selected results on benchmarks	43
3.3.1	PageRank	43
3.3.2	BLSTM	48
3.3.3	GLQ	51

3.4	Summary and conclusion	51
4	Core transprecision concepts	55
4.1	Number formats	55
4.1.1	Fixed-point	56
4.1.2	IEEE 754 floating-point	57
4.1.3	Logarithmic number system (LNS)	59
4.2	The transprecision system view	60
4.2.1	Transprecision concepts	63
4.2.2	Reduced precision as root-cause	66
4.2.3	Transprecision computing in current solutions	68
4.3	Summary and conclusion	70
5	Emulating numerical behavior of applications	73
5.1	The floatx library	74
5.1.1	Related work	74
5.1.2	Interface and design goals	75
5.1.3	The choice of C++	77
5.1.4	The floatx class template	78
5.1.5	Operations on floatx objects	81
5.1.6	The floatxr class template	85
5.1.7	Notes on concurrency	86
5.1.8	Advanced properties and performance of floatx	87
5.2	Numerical analysis of applications	89
5.2.1	PageRank	89
5.2.2	BLSTM	93
5.2.3	GLQ	96
5.3	Summary and conclusion	100
6	Floatx for deep learning	103
6.1	Integrating TP into deep learning	104
6.1.1	Arithmetic free and elementary kernels	107
6.1.2	High precision accumulator assumption	107
6.1.3	The intrinsic versus extrinsic approach	109
6.2	Numerical analysis of deep learning models	114
6.2.1	Reference models	115
6.2.2	Numerical analysis	117
6.3	Summary and conclusion	122

7 Optimization for transprecision configurations	125
7.1 Searching transprecision configurations	126
7.2 Search heuristics	128
7.3 Results on reference problem instance	134
7.4 Heuristic search performance	141
7.5 Summary and conclusion	147
8 Optimization for IoT devices with given constraints	149
8.1 Related work for network architecture search	151
8.2 Narrow-space architecture search	152
8.2.1 Narrow-space and sampling law definition	153
8.2.2 Precision analysis	158
8.2.3 Performance characterization on hardware	159
8.2.4 Fast cognitive design algorithms	163
8.2.5 Statistical properties of generated networks	165
8.2.6 Training setup	167
8.3 Results	168
8.4 Summary and conclusion	173
9 Conclusions	175
9.1 Summary of main results	176
9.2 Outlook and future work	181
A Use case: Efficient video classification	185
A.1 Related work	186
A.2 Practical video classification systems	187
A.2.1 Global video descriptor (GVD)	189
A.2.2 Frame based classification (FBC)	189
A.2.3 Long short-term memory (LSTM)	189
A.3 Computational complexity	191
A.4 Temporal subsampling	191
A.5 Evaluation and results	193
A.6 Results and discussion	195
A.7 Summary and conclusion	196

B Notation and acronyms	199
Symbols	199
Operators	200
Acronyms	201
Bibliography	205
Curriculum Vitae	231