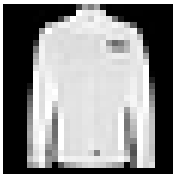
universität
innsbruck

Causes and Effects of Unanticipated Numerical Deviations in Neural Network Inference Frameworks

Alexander Schlögl Nora Hofer Rainer Böhme

37th Conference on Neural Information Processing Systems · New Orleans, LA · Dec 2023

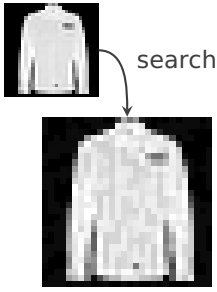
Numerical Deviations



Platform	Label	Confidence
Intel i7-9700	Shirt	0.98742348
Intel e3-1250	Shirt	0.98742348
AMD TR 2950x	Shirt	0.987423466

Schlögl, A., Kupek, T., and Böhme, R. Forensicability of Deep Neural Network Inference Pipelines.
IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 2515–2519, 2021.

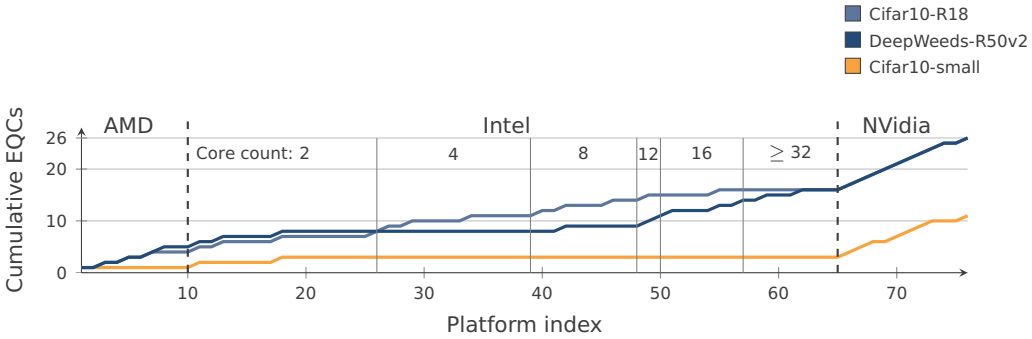
Numerical Deviations



Platform	Label	Confidence
Intel i7-9700	Shirt	0.50000011
Intel e3-1250	Shirt	0.50000011
AMD TR 2950x	Coat	0.49999996

Schlögl, A., Kupek, T., and Böhme, R. iNNformant: Boundary Samples as Telltale Watermarks.
ACM Workshop on Information Hiding and Multimedia Security, pp. 81—86, 2021.

Results



See Tables SUP-1 and SUP-2 in the paper for all hardware, model, and input details.

Causes of Deviations on CPUs

Floating-point addition is not associative

$(a + b) + c$

a	1.414213562373095145475
b	2.718281828459045090796 $+$
	4.132495390832140458315
c	3.141592653589793115998 $+$
	7.274088044421933574313

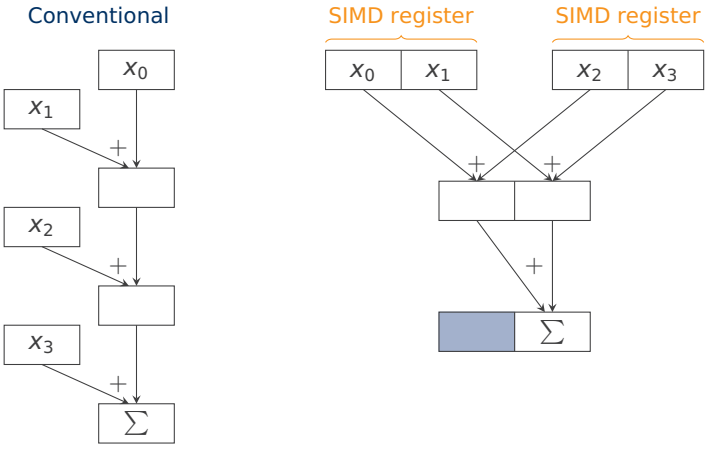
$a + (b + c)$

b	2.718281828459045090796
c	3.141592653589793115998 $+$
	5.859874482048837762704
a	1.414213562373095145475 $+$
	7.274088044421932686134

Example values: $a = \texttt{float32}(\sqrt{2})$, $b = \texttt{float32}(e)$, $c = \texttt{float32}(\pi)$.

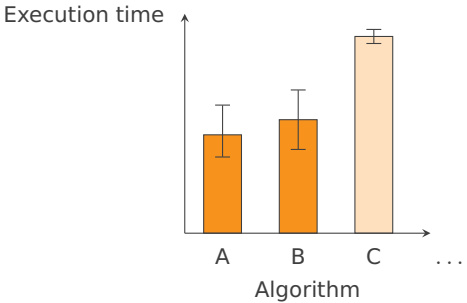
Causes of Deviations on CPUs

Parallelization changes aggregation order



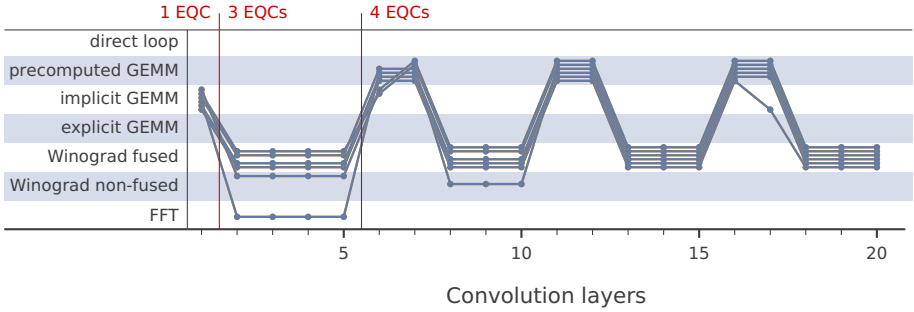
Causes of Deviations on GPUs

Microbenchmarks determine algorithm selection



Causes of Deviations on GPUs

Algorithm selection by layer per session



Source: Figure 3 (b) in our paper.

Implications

Broader effects of platform-specific deviations

System correctness

Replicability

Security

Forensics

Thank You

Authors

Alexander Schlögl, Nora Hofer, Rainer Böhme

Research code

https://github.com/uibk-iNNference

Funding notice

UNCOVER

Parts of this work were funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101021687, and a PhD Fellowship from NEC Laboratories Europe (2021–2022).