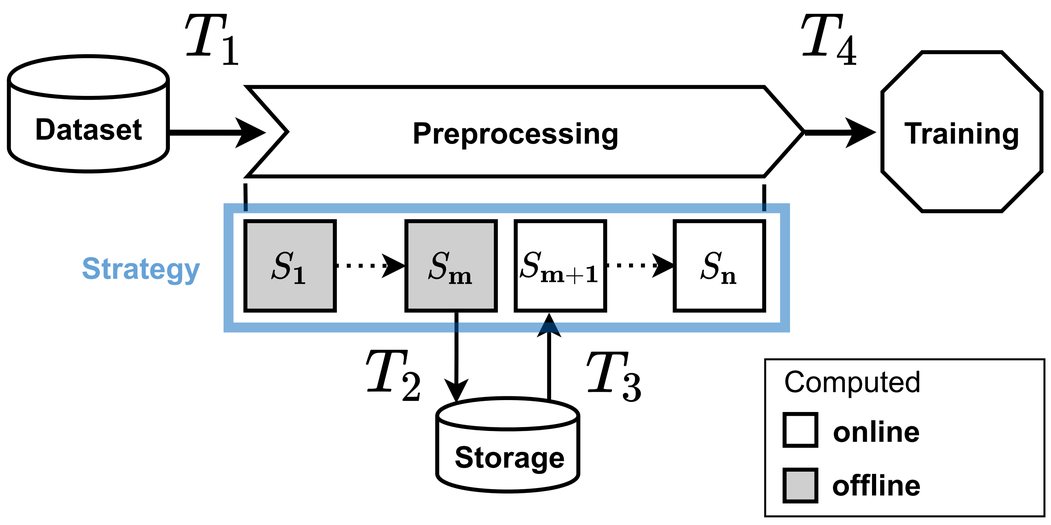
Where Is My Training Bottleneck? Hidden Trade-Offs in Deep Learning Preprocessing Pipelines

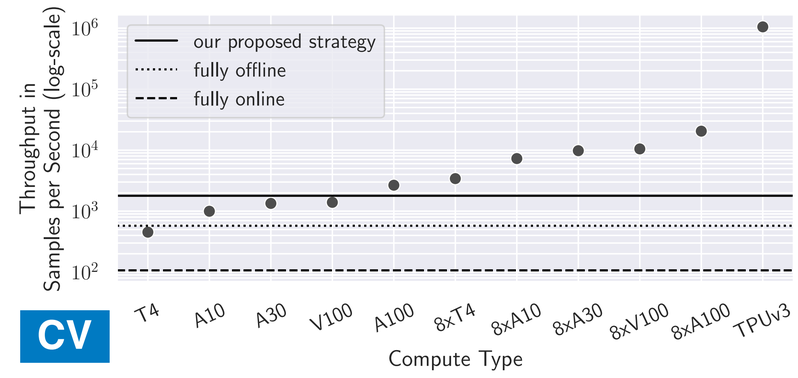
Alexander Isenko, Ruben Mayer, Jeffrey Jedele, Hans-Arno Jacobsen (University of Toronto)

Introduction



General Deep Learning preprocessing setup with different theoretical throughputs T_1-T_4

Motivation

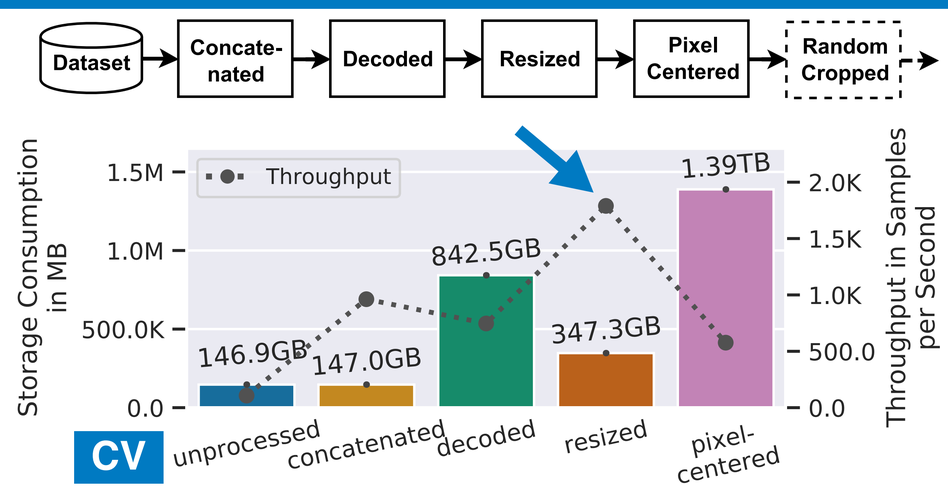


Training throughput of ResNet-50 for different hardware configurations

Strategy	Throughput	Storage Consumption
fully online	107 samples/s	146 GB
fully offline	576 samples/s	**1535 GB**
our proposed strategy	**1789 samples/s** (3x ↓)	494 GB (0.3x ↑)

Insight #1

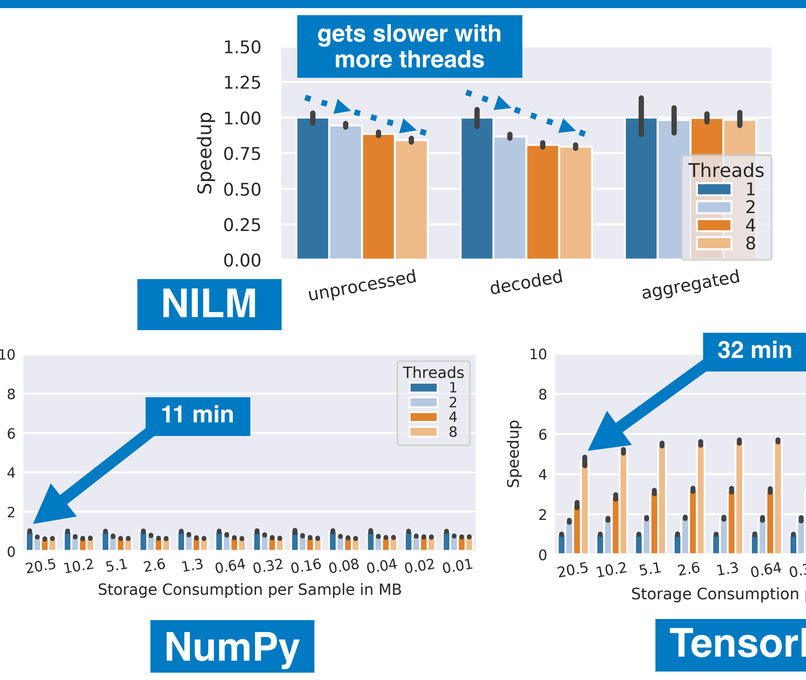
Fully Offline Preprocessing Is Not Necessarily The Best Performing Strategy



Storage consumption vs. throughput comparison for the CV pipeline

Insight #2

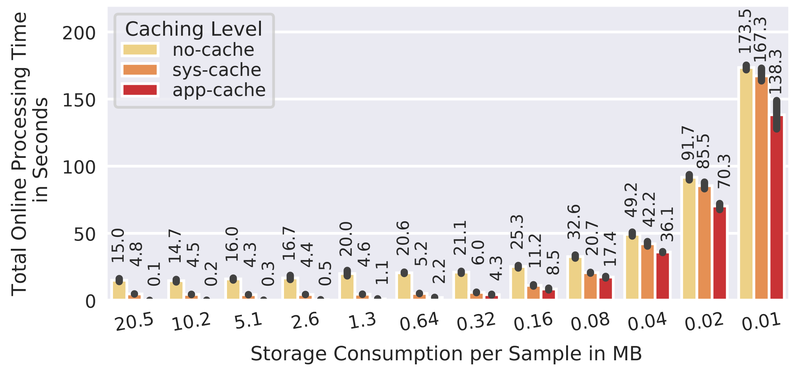
Native Implementations Are Not Necessarily The Best



Parallelization analysis of the NILM pipeline with 1 to 8 threads

Insight #3

Small Sample Sizes Can Negatively Affect Performance



Online processing time for different caching levels and sample sizes of a synthetic 15 GB dataset

Lower is better!

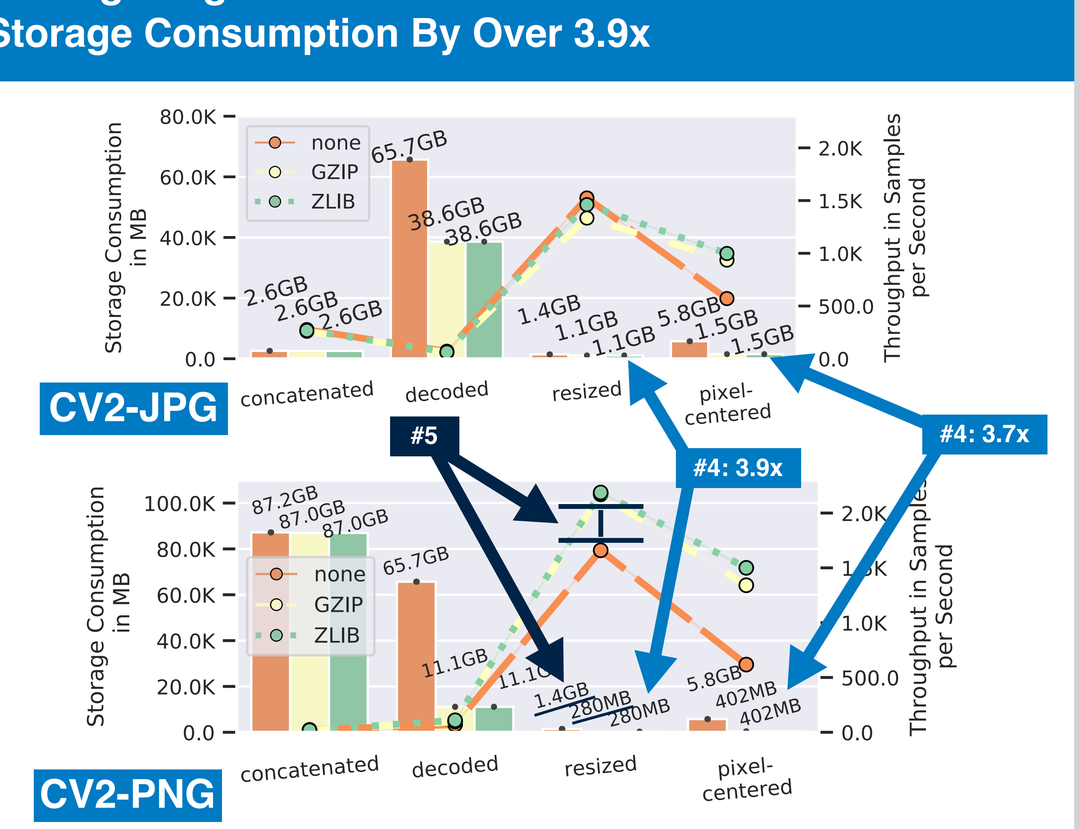
Pipeline	System-Level	Application-Level	Sample Size
CV2-JPG	3.3x	15.2x	1.18 MB
CV2-PNG	3.5x	14.5x	1.18 MB
FLAC	4.2x	8.0x	0.41 MB
MP3	1.6x	2.2x	0.08 MB
NILM	1.1x	1.4x	0.01 MB

Bigger = Faster

Throughput increase compared to no cache runs

Insight #4

Storing Images As PNG vs. JPG Can Reduce Storage Consumption By Over 3.9x



GZIP, ZLIB and no compression runs for the CV2-JPG and CV2-PNG dataset

Insight #5

Compression Can Reduce Storage Consuption AND Increase Throughput At The Same Time

github.com/cirquit/presto

alex.isenko@tum.de