Title: DL Boost

URL: https://en.wikipedia.org/wiki/DL_Boost

PageID: 67200581

Categories: Category:Deep learning, Category:SIMD computing, Category:X86 architecture

-----

Intel 's Deep Learning Boost (DL Boost) is a marketing name for instruction set architecture (ISA) features on the x86-64 designed to improve performance on deep learning tasks such as training and inference. [1]

Features

DL Boost consists of two sets of features:

AVX-512 VNNI , 4VNNIW, or AVX-VNNI : fast multiply-accumulation mainly for convolutional neural networks .

AVX-512 BF16: lower-precision bfloat16 floating-point numbers for generally faster computation. Operations provided include conversion to/from float32 and dot product .

DL Boost features were introduced in the Cascade Lake architecture.

A TensorFlow -based benchmark run on the Google Cloud Platform Compute Engine shows improved performance and reduced cost compared to previous CPUs and to GPUs, especially for small batch sizes. [2]

Notes

External links

Deep Learning Boost at Intel

Andres Rodrigues et al., "Lower Numerical Precision Deep Learning Inference and Training", Intel White paper [3]

Intel and ML (2017), from Intel's Developer Relations Division