Title: Accelerated Linear Algebra

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

PageID: 75530149

Categories: Category:Compilers, Category:Computer library stubs, Category:Free software programmed in C++, Category:Machine learning, Category:Software using the Apache license

-----

XLA (Accelerated Linear Algebra) is an open-source compiler for machine learning developed by the OpenXLA project. XLA is designed to improve the performance of machine learning models by optimizing the computation graphs at a lower level, making it particularly useful for large-scale computations and high-performance machine learning models. Key features of XLA include:

Compilation of Computation Graphs: Compiles computation graphs into efficient machine code.

Optimization Techniques: Applies operation fusion, memory optimization, and other techniques.

Hardware Support: Optimizes models for various hardware, including CPUs, GPUs, and NPUs.

Improved Model Execution Time: Aims to reduce machine learning models' execution time for both training and inference.

Seamless Integration: Can be used with existing machine learning code with minimal changes.

XLA represents a significant step in optimizing machine learning models, providing developers with tools to enhance computational efficiency and performance.

Supported target devices

x86-64

ARM64

NVIDIA GPU

AMD GPU

Intel GPU

Apple GPU

Google TPU

AWS Trainium, Inferentia

Cerebras

Graphcore IPU

See also

TensorFlow

PyTorch

JAX

References

This computer-library -related article is a stub . You can help Wikipedia by expanding it .

v

t

e