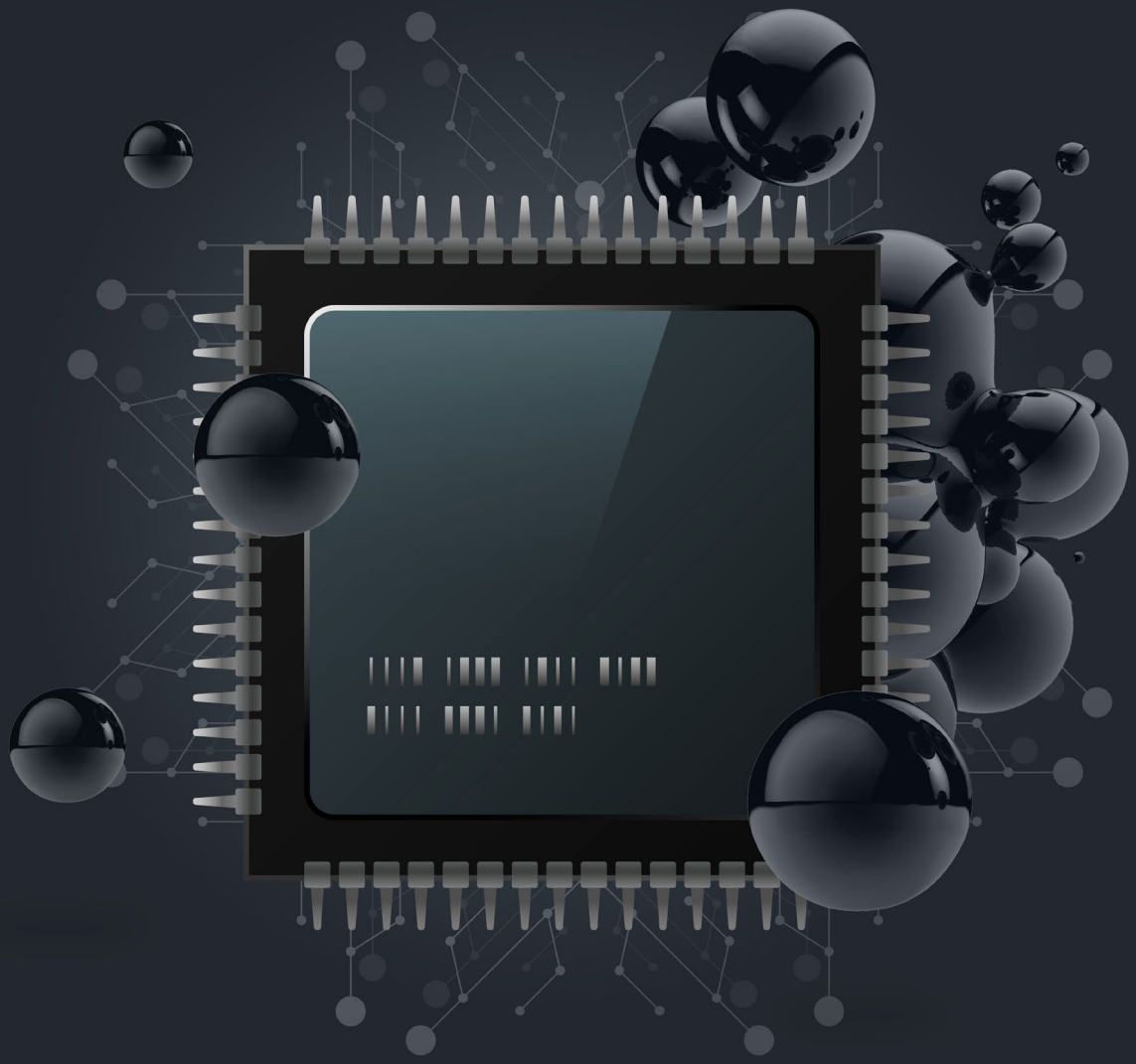
deci

The Future of CPU in Deep Learning Inference

Achieve outstanding performance without compromising on quality



KEY TAKEAWAYS:

- Deci's AutoNAC™ together with Intel's OpenVINO™ toolkit successfully reduced ResNet-50's latency by a factor of up to 11.8x and increased throughput by up to 11x–all while preserving accuracy (%1 ±).
- The Deci-Intel collaboration takes a significant step towards enabling deep learning inference at scale on CPUs.
- Companies that use large scale inference scenarios can now dramatically cut cloud costs by changing the inference hardware from GPU to CPU, and enable real-time performance on edge devices with CPUs.

Introduction

Deep learning models are growing in complexity, and driving increased compute demand for AI on the cloud, in data centers, and at the edge.

Being able to run more efficient AI models on CPUs rather than GPUs has the potential to dramatically open up the market for the application of AI—delivering new practical solutions for cameras and drones, autonomous cars, computer-aided diagnosis for medical imaging, and more.

This requires deep learning inference on CPUs to have low latency for real-time applications and high throughput for cost-effective inference at scale when it comes to datacenters or cloud.

But can CPUs really become the new AI dynamo for deep learning computing?

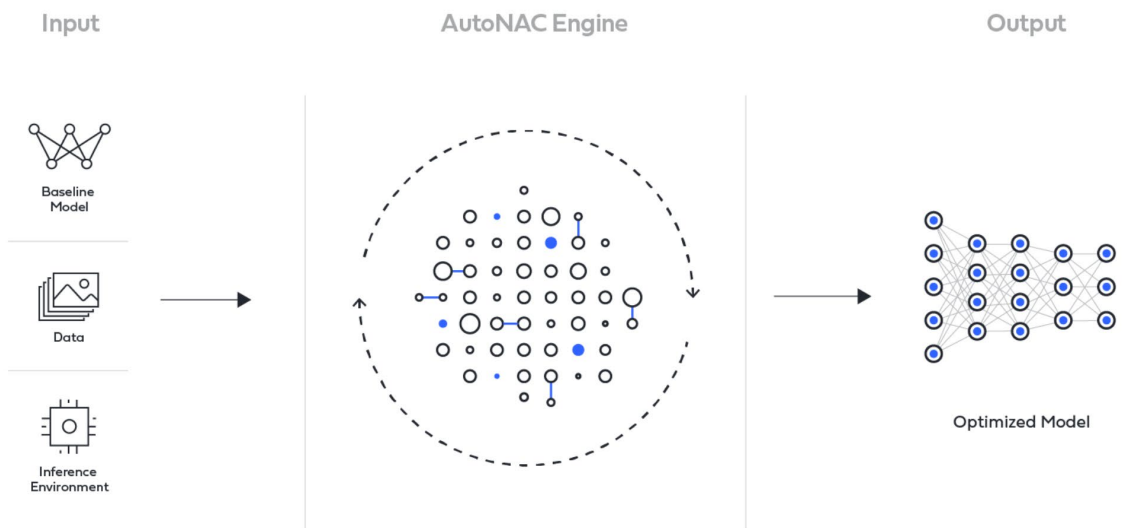
The Problem

CPUs are inherent inside everything from large data centers to smartphones. They are also cheaper than GPUs and generally consume less power. However, in many real-world applications of neural networks, the inference on CPU devices is difficult due to its high latency or low throughput. The ResNet-50 deep neural network is one example of this challenge, and has the potential to open a new world of opportunities if we could only speed up its inference without losing out on accuracy.

Accelerating Inference with AutoNAC™

Deci's platform, featuring its AutoNAC™ technology, performs a smart high-speed search across a huge set of neural network architectures to aggressively speed up runtime, while preserving accuracy—and can serve as an ideal solution for production-grade inference on CPU devices. We accomplish this by optimizing the fit between the neural network structure, the user's dataset, and the target computing hardware.

Deci's deep learning platform was up to the task of boosting the ResNet-50 model's performance and we were keen to collaborate with Intel to fuse Deci's optimization with their OpenVINO toolkit. But we needed an objective measurement to determine just how well the AutoNAC solution performs against other possibilities to confirm that this was the right choice for Intel and other companies seeking similar solutions.



An Objective Benchmark - MLPerf

MLPerf is a non-profit organization that provides fair and standardized benchmarks for measuring training and inference performance of machine learning hardware, software, and services. This venue enabled us to test our core technology on a fair benchmark.

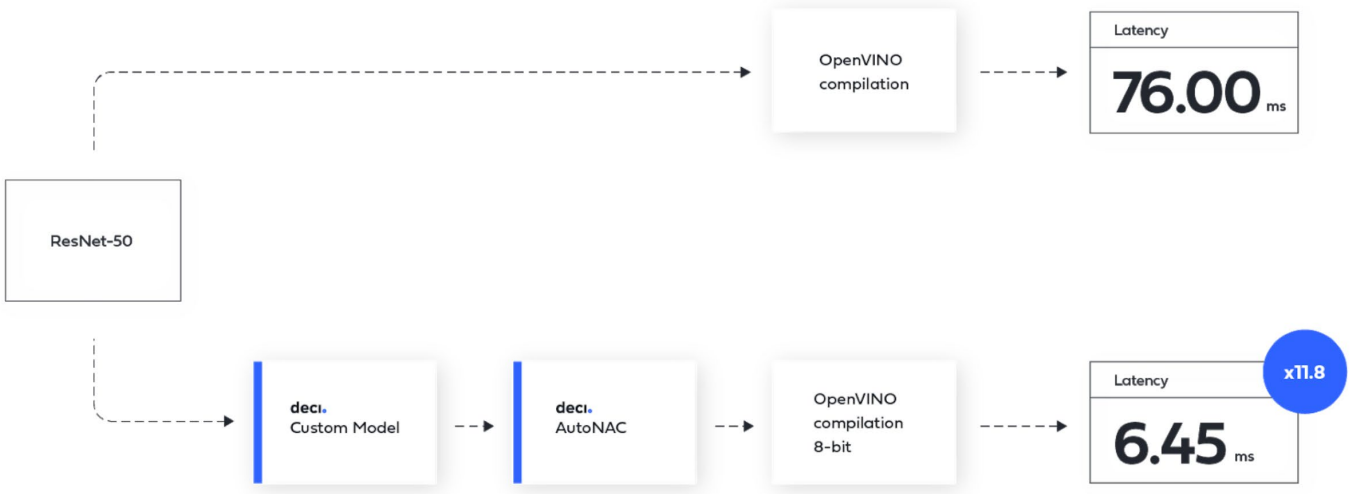
With its broad approach to machine learning supported by both industry and research academia, MLPerf is primarily used for assessing workloads, with over 40 organizations coming together to decide on a consistent set of benchmarks for ML workflows.

There are two independent benchmark suites in MLPerf: training and inference. Each inference benchmark is defined by a model, dataset, quality target, and latency constraint. Every benchmark also contains two divisions: Open and Closed. While the Closed division is restricted to benchmarking a given neural architecture on specific hardware platforms and optimization techniques, the Open division is intended to foster innovation and therefore allows modification of the neural architecture. Deci collaborated with Intel in a submission to the Open, ImageNet, ResNet-50 track including results for the Single Stream and Offline scenarios, which demonstrate latency and throughput, respectively. We ran our models on three different types of hardware: single and 8 core 2.8GHz Intel Cascade Lake CPU (n2-highcpu-2, n2-highcpu-16 GCP instances, respectively), and 1.4GHz Intel quad core i5-8257U Macbook Pro 2019.

The Results

According to the MLPerf rules, our goal was to maximally reduce the latency, or increase the throughput, while staying within 1% of ResNet-50 original accuracy.

AutoNAC together with Intel's OpenVINO toolkit successfully reduced ResNet-50's latency by a factor of up to 11.8x and increased throughput by up to 11x–all while preserving accuracy (%1).



Using the trained model from the AutoNAC, we compiled the model with Intel's OpenVINO. Specifically, we used the OpenVINO post-training optimization toolkit to perform 8-bit quantization on the model and convert the network's trained weights to 8-bit precision. The selected optimal model was robust to 8-bit quantization and the accuracy degradation was negligible (around 0.3%). Moreover, the resulting speedup was more than 2.5x in terms of run-time, relative to vanilla compilation.

To further distinguish the improvements due to the AutoNAC, we compared a compiled 8-bit ResNet-50 precision to our submitted model. This comparison showed an improvement in latency of between 2.6x and 3.3x. What's more, our throughput scenario submissions improved ResNet-50 vanilla's throughput by a factor of 6.9x to 11x. Isolating the net effect of AutoNAC, when compared to a compiled 8-bit ResNet-50 precision, we can see an improvement factor in the range of 2.7x to 3.1x.

Our submission to MLPerf proved that our AutoNAC technology reduces runtime while preserving the accuracy of the base model. When combined with off-the-shelf compilers such as Intel's OpenVINO, Deci's AutoNAC can achieve more than 11x improvement over the base models.

Latency (ms)

Hardware	ResNet-50 OpenVINO 32-bit	ResNet-50 OpenVINO 8-bit	Deci	Deci's Boost
1.4GHz 8th-generation Intel quad core i5 Macbook Pro 2019	83	83	7	**11.8x**
1 Intel Cascade Lake Core	76	21	6.45	**3. 3 - 11.8x**
8 Intel Cascade Lake Cores	11	5.5	2.13	**2.6 - 5.16x**

Single stream scenario - Latency (ms). Three hardware types were tested. ResNet-50 Vanilla is ResNet-50 compiled with OpenVINO to 32-bit. The third column shows the same hardware with compilation to 8-bit, and the column labelled Deci shows the custom model after AutoNAC with 8-bit compilation.

Throughput

Hardware	ResNet-50 OpenVINO 32-bit	ResNet-50 OpenVINO 8-bit	Deci	Deci's Boost
1.4GHz 8th-generation Intel quad core i5 Macbook Pro 2019	30	30	207	**6.9x**
1 Intel Cascade Lake Core	14	50	154	**3.1 - 11x**
8 Intel Cascade Lake Cores	110	410	1092	**2.7 - 9.9x**

Offline scenario - Throughput (images per second). Three hardware types were tested. ResNet-50 Vanilla is ResNet-50 compiled with OpenVINO to 32-bit. The third column is the same with compilation to 8-bit, and the Deci column shows the custom model after AutoNAC with 8-bit compilation.

The Future of CPU Inference

The Deci-Intel collaboration takes a significant step towards enabling deep learning inference at scale on CPUs. In this collaboration with Intel, we used our technology, the Automated Neural Architecture Constructor (AutoNAC), and integrated it with Intel's OpenVINO toolkit to cut the inference latency and boost throughput of their ResNet-50 model.

This submission shows that the future is bright. Going beyond the remarkable results achieved at MLPerf, this kind of performance improvement enables extensive use of deep learning inference on CPU devices. Companies that use large scale inference scenarios can now dramatically cut cloud costs by changing the inference hardware from GPU to CPU, and enable real-time performance on edge devices with CPUs.

> *Intel and Deci partnered to break a new record at the MLPerf benchmark, accelerating deep learning by 11x on Intel's Cascade Lake CPU. That's amazing! Deci's platform and technology have what it takes to unleash a whole new world of opportunities for deep learning inference on CPUs*
>
> **Guy Boudoukh, Deep Learning Research, Intel AI Research**

About the Deci Platform

Deci's deep learning platform automatically gears up neural networks for production and top performance—on any hardware. Powered by AutoNAC™ technology, the platform enables data scientists to:

- Accelerate model inference (throughput/latency) by up to 10x, on any hardware, while maintaining accuracy
- Cut compute costs, including cloud and edge, by up to 80%
- Seamlessly deploy and serve models to any production environment