deci

Irisity’s 6.5x performance Boost Leads to Cost-Efficient, Real-Time Video Analytics on Edge CPUs

Results:

6.5x
Throughput Increase

Run and efficiently scale on a wide range of Intel CPUs

Significantly reduce development time

IRISITY
powered by Agent Vi

> **Using Deci, we swiftly developed a model that enabled us to expand our offering and further scale our solution on existing CPU infrastructure with significant cost-efficiency.**
>
> *Zvika Ashani, CTO at Irisiry*

Introduction

Video analytics powered by AI is revolutionizing industries by extracting insights from visual data streams. Using deep learning algorithms, video content is automatically analyzed and interpreted to detect objects, people, and events.

AI video analytics applications are rapidly expanding across sectors, offering benefits such as enhanced security capabilities, improved operational efficiency, and data-driven decision-making.

Irisity is a leading provider of AI-powered video analytics software serving a wide range of use cases including critical infrastructure security monitoring, fire and smoke detection, slip and fall identification, and traffic monitoring. Its solutions are powered by advanced algorithms that recognize activities and incidents, collect operational intelligence, and generate business insights for better decision-making and optimal operation.

Overview

- Video Analytics Platform
- Object Detection
- Intel CPUs N2-standard-4, GCP

Irisity was looking to improve their offering by further increasing the throughput of their object detection model while maintaining its high accuracy. Doing so would enable Irisity to scale its solutions even more cost-effectively on existing CPU infrastructure, thus reducing operational costs for their customers.

In addition, faster models also deliver better user experience with real time insights and alerts. The ability to generate insights within seconds is key, especially for video analytics applications that address use cases such as security of critical infrastructure personal, health-care and other time sensitive use cases.

Solution Overview

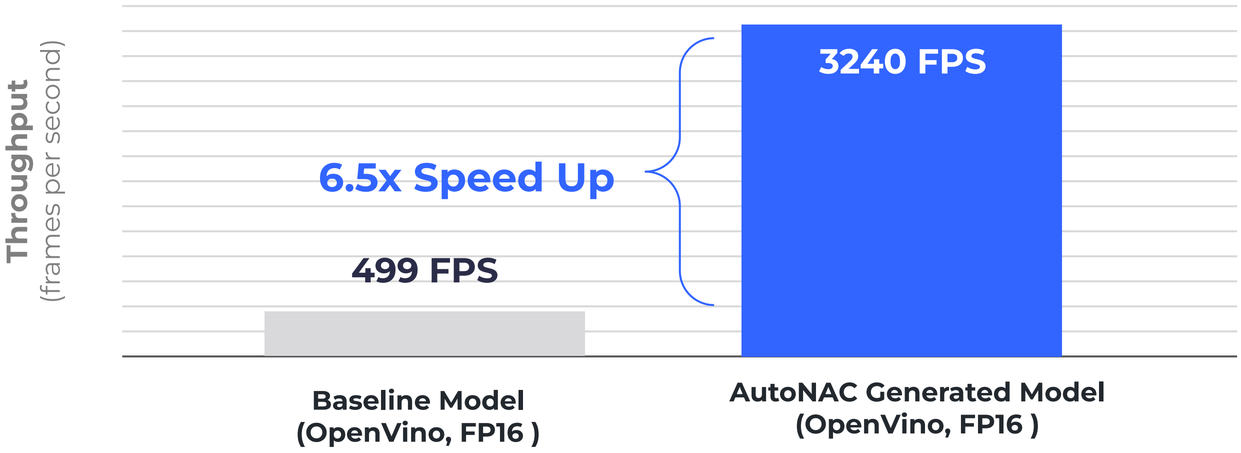
Leveraging Deci's platform, Irisity's team developed an enhanced object detection model that outperformed its predecessor in terms of inference performance, all while retaining the original model's accuracy. The new model, generated with Deci's Automated Neural Architecture Search (AutoNAC) engine, showcased an impressive throughput increase, moving from 499 frames per second to 3,240 – a 6.5x improvement. The team utilized Deci's SuperGradients training library to train this new model and further optimized it using Infery, Deci's SDK for inference optimization and deployment.

Irisity's Results Overview

The optimized model enabled Irisity to scale their solution by extending its support for a wide range of CPUs. With the help of the Deci platform, the Irisity team not only shorten their development process and minimized associated risks, but also effortlessly achieved their desired performance benchmarks while keeping stringent data privacy in check.

6.5x Faster Inference

Models measured on GCP Intel CPU instance N2-standard-4 (batch size 4)

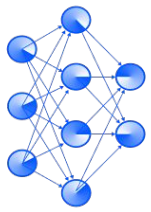


* Maintaining accuracy F1 ± <1%

Deci Platform

Foundation or Custom Models

Choose an ultra-performant pre-trained model or generate a custom one.



SaaS

AutoNAC

Neural Architecture Search Engine

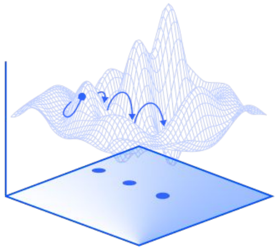
On Prem

DataGradients™

Dataset analyzer

Train or Fine-tune

Use Deci's library & custom recipe to train on-prem.



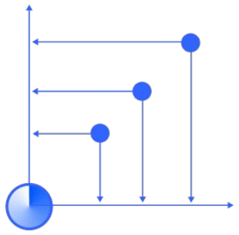
On Prem

SuperGradients™

PyTorch Training library

Optimize & Run

Apply acceleration techniques. Run self-hosted inference anywhere.



On Prem

Infery

Optimization & Inference Engine SDK

Training Frameworks

Runtime Frameworks

AutoNAC™ - The Most Advanced Neural Architecture Search Technology

Deci is powered by the groundbreaking Automated Neural Architecture Construction (AutoNAC™) technology. Deci's AutoNAC engine performs a multi-constraints search to find the optional model architecture that delivers the highest accuracy for pre-defined speed, model size, inference hardware and use case requirements. This cutting-edge technology underlies all our foundational models, expertly designed to deliver cost-effective inference performance. YOLO-NAS, globally recognized as the superior foundation model for object detection, DeciCoder, DeciLM 6B, DeciDiffusion and our groundbreaking achievements at MLPerf, are just a few illustrative examples of AutoNAC's extraordinary performance and capabilities.

Main Capabilities Overview

Gain Superior Performance with Custom Architectures

Build accurate and efficient architectures tailored to your hardware and application's performance targets with Deci's Neural Architecture Search technology.

Maximize Accuracy with Advanced Training Techniques

Train models with SuperGradients. Leverage custom recipes and advanced training techniques with one line of code.

Simplify Runtime Optimization

Easily compile and quantize your models (FP16/INT8) and evaluate different production settings with a click of a button.

Streamline Deployment with 3 Lines of Code

Deploy your models with Infery, Deci's simple-to-use, unified, model inference API. Streamline deployment and boost serving performance. Compatible with multiple frameworks and hardware types.

Why Deci?

Achieve Real-Time Inference on Edge Devices

Improve latency and throughput, and reduce model size by up to 5X while maintaining the model's accuracy.

Process More Video Streams on Less Devices

Maximize hardware utilization and cost-efficiently scale your solution at the Edge.

Reduce Development Effort & Risks

Simplify development with automated tools that guarantee success.

About Deci

Deci enables deep learning to live up to its true potential by using AI to build better AI. With Deci's platform, AI developers can easily build, optimize, and deploy highly accurate and efficient models to any environment including cloud, edge, and mobile. Leading enterprises are using Deci to boost their deep learning models' performance, shorten development cycles, enable new use cases on edge devices, and reduce computing costs.

BOOK A DEMO

deci