deci.

Scaling Up AI-Enabled Video Analytics Applications on NVIDIA® Jetson™ Devices

Executive Summary

AI-enabled video analytics applications are being deployed today across many different industries such as smart city, security cameras, healthcare, smart retail, and sports tech, among others. While these applications are designed to support a wide variety of use cases, there are many shared challenges faced by AI developers across industries. Developers need to ensure that the computer vision models that power video analytics applications are accurate, can deliver real-time insights and run in a cost-efficient manner.

Deci's customers have scaled up AI-enabled video analytics solutions across various verticals using our NAS-based Deep Learning Development platform.

By Using Deci's Development Platform You Can:

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.

Deploy Your Models on Any Edge Device

Eliminate inference cloud compute cost and avoid data privacy issues by running your models directly on edge devices.

Use Case 1
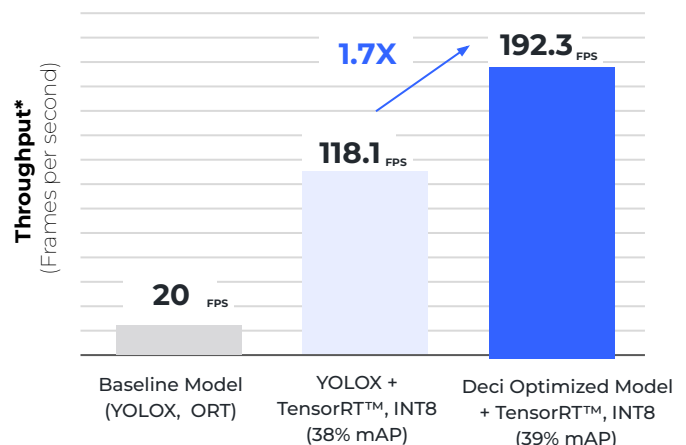### Scaling up AI-based Security Camera Solution

A security company's goal was to maximize the efficiency of their existing infrastructure by increasing the number of live video streams that can be processed in real-time on their NVIDIA Jetson Xavier NX hardware.

However, after increasing the number of video streams to be processed, their object detection model (YOLOX) did not achieve the required level of throughput (192 frames per seconds, inference batch of 8) for the solution to be viable.

By using Deci's AutoNAC engine, the customer built an architecture that delivered 1.7X acceleration reaching a throughput of 192 frames per seconds,   while also improving the accuracy by 1% mAP. This allowed the customer to double the number of video streams from 4 to 8, increasing the scalability and profitability of their video analytics solution.

1.7X Higher Throughput (Improved Accuracy)

Measured on NVIDIA Jetson Xavier NX

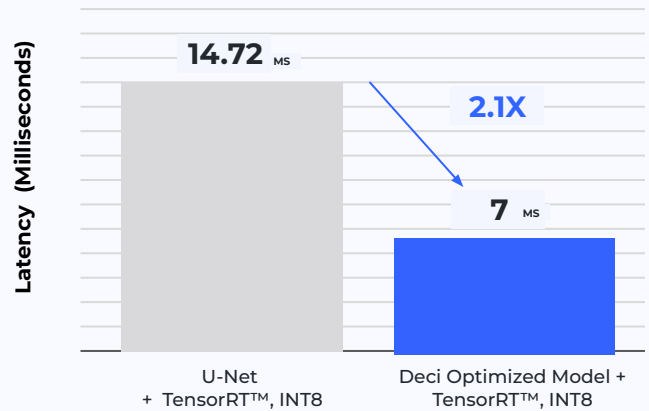*Total Model Throughput, Batch Size 8

Use Case 2
Enabling Real-Time Semantic Segmentation for an Automotive Application

An automotive company running a U-Net based segmentation model on a NVIDIA Jetson Xavier NX struggled to achieve the target throughput in production.

Using Deci's AutoNAC engine, a faster and smaller model was generated. Latency was reduced by 2.1X, model size was reduced by 3X, and memory footprint was reduced by 67% – all while maintaining the original accuracy.

2.1X Lower Latency (Same Accuracy)

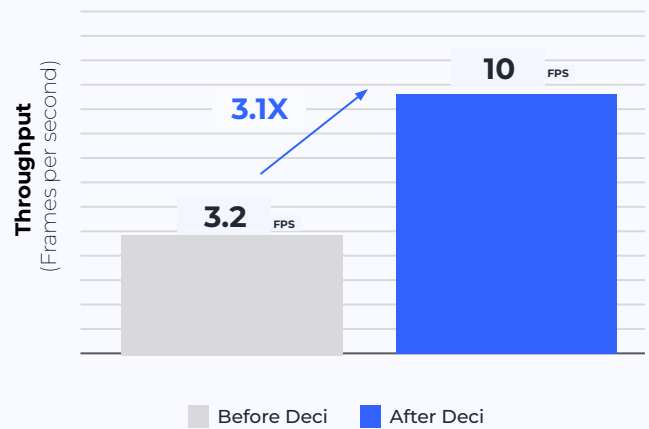Measured on NVIDIA Jetson Xavier NX

Use Case 3
Enabling a New Security Application

A defense company needed to process high resolution images for an object detection and tracking task on an NVIDIA Jetson Xavier NX device. In order for the system to become operational, the customer they needed to run in a 10 watt mode and achieve a throughput of 10 frames per second.

Using Deci's AutoNAC engine, the customer was able to increase throughput by 3.1X, and run smooth object tracking, and unlock a new security application.

3.1X Higher Throughput (Same Accuracy)

Measured on NVIDIA Jetson Xavier NX

Build Better Models Faster with Deci's Deep Learning Development Platform

The Deci platform is used by data scientists and machine learning engineers to build, optimize, and deploy highly accurate and efficient models to production. Teams can easily develop production grade models and gain unparalleled accuracy and speed tailored for any performance targets and hardware environment.

Deci is powered by AutoNAC (Automated Neural Architecture Construction), the most advanced and commercially scalable Neural Architecture Search engine in the market. AutoNAC performs a multi-constraints search to find the architecture that delivers the highest accuracy for any performance targets and hardware environment.

Use Cases

Enable Inference on Edge Devices

Simplify Development and Shorten Time to Market

Improve UX with Better Inference Performance

Improve Hardware Utilization & Reduce Inference Costs

TRY NOW

For more information, visit **deci.ai**

deci.