deci.

Boost Manufacturing Efficiency With Powerful Deep Learning Models

Introduction

Computer vision is already transforming manufacturing processes, with numerous applications such as quality inspection, safety monitoring, defect detection, supply chain management, and more. Computer vision-based solutions deliver the intelligence to simplify processes, drive new efficiencies, and empower faster decision-making. However, AI developers are still facing challenges in developing and deploying such solutions. The inability to run real-time inference, high false alarms due to low model accuracy, and the failure to deploy on CPUs or edge devices are just some of the barriers to production faced by AI teams in manufacturing companies today.

With Deci, you can boost your models' performance and maximize hardware utilization to deliver accurate and cost-efficient inference on cloud or edge devices. Below are two case studies of manufacturing companies that significantly improved their models' performance, resulting in better production efficiency, scalable edge deployment, and higher profitability.

Case Study 1

Improving Inspection Quality & Production Capacity for a Defect Detection Application

1.6x Throughput Increase

62% Model Size Reduction

13% Memory Footprint Reduction

The Challenge

A manufacturing company specializing in materials engineering for the semiconductor industry sought to improve their defect detection inspection process and capacity by accelerating their model's runtime performance.

The company's semantic segmentation model was not achieving the desired throughput on the targeted inference hardware, a NVIDIA A100 GPU (40GB).

The Solution

The team used the Deci platform and its Neural Architecture Search engine to build a custom segmentation model that delivered a 1.6x increased throughput, a 62% reduction in model size, and a 13% reduction in memory footprint while maintaining the accuracy.

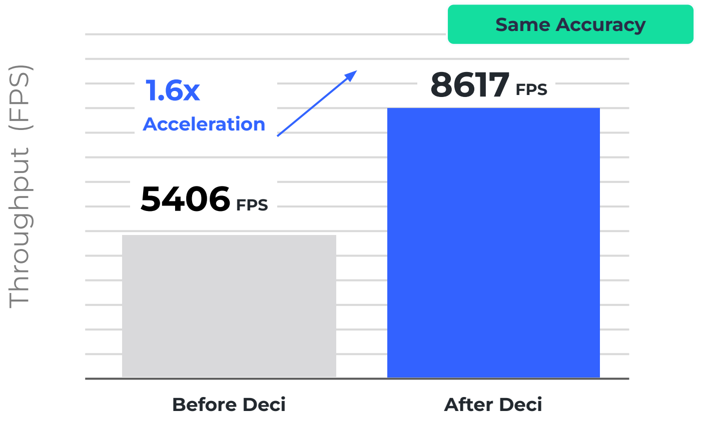
The team trained the new model on-premise using Deci's open-source training library called SuperGradients and then compiled and quantized the model to TensorRT FP16 using Deci's platform. The company was able to quickly improve its visual inspection process, resulting in increased production capacity and better profit margins.

Overview

- Defect Detection in Manufacturing
- Segmentation
- NVIDIA A100, 40 GB

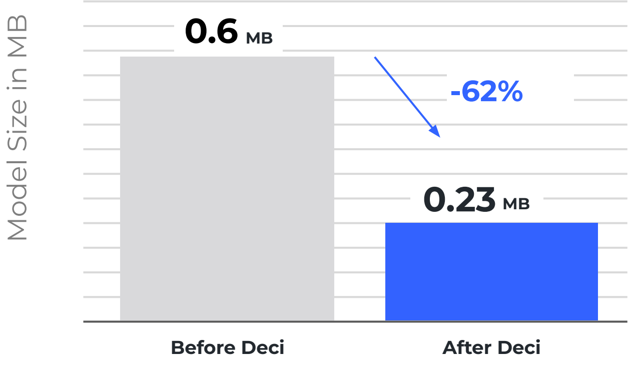
1.6x Higher Throughput

Measured on NVIDIA A100 GPU, 40GB



62% Smaller Model Size

Measured on NVIDIA A100 GPU, 40GB



Case Study 2

Enabling Real Time Inference on Edge Devices for a PPE Inspection Solution

+14.7% Accuracy Increase

4% Throughput Increase

80% Shorter Development Time

Overview

- Safety Personal Protective Equipment (PPE) Identification
- Object Detection
- All i5 Intel CPUs

The Challenge

A company enabling manufacturers to digitally transform operations with a connected, IoT-native, no-code platform sought to enhance its safety personal protective equipment (PPE) identification offering by improving the model's accuracy and latency on CPUs and other resource-constrained edge devices.

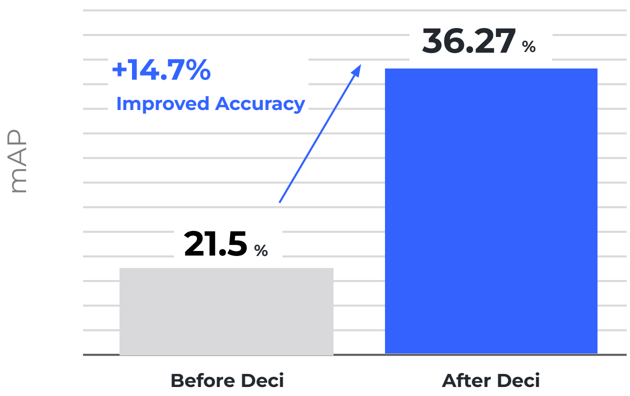
The company needed to deploy its models on a wide range of old-generation CPUs, such as Intel Cascade Lake CPU and Intel Tiger Lake CPU, among others, to utilize existing server infrastructure. However, their object detection model was not achieving the desired accuracy and latency on such CPUs.

The Solution

The team used the Deci platform and its Neural Architecture Search engine to develop a custom object detection model that delivered a 14.7% increase in accuracy and a 4% boost in throughput. The company deployed the model on a wide range of hardware environments and enhanced its safety PPE identification solution. By using Deci early in the development process, the team succeeded in reaching production within a few weeks instead of months while minimizing their development risk and effort.

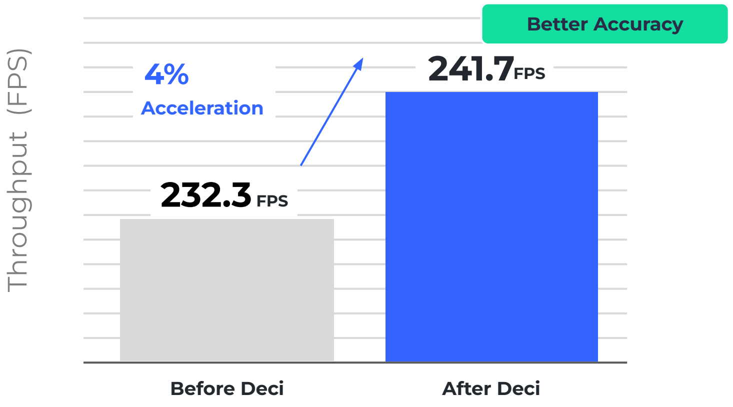
+14.7% Accuracy Increase

Measured on Intel CPU Cascade Lake (c5.4xlarge)



4% Higher Throughput

Measured on Intel CPU Cascade Lake (c5.4xlarge)



By Using Deci's Platform You Can:

Achieve Real Time Inference to Detect Defects as They Happen

Improve latency and throughput, reduce memory footprint while maintaining the model's accuracy.

Reduce False Alarms with Better Model Accuracy

Increase inspection and production capacity with better accuracy.

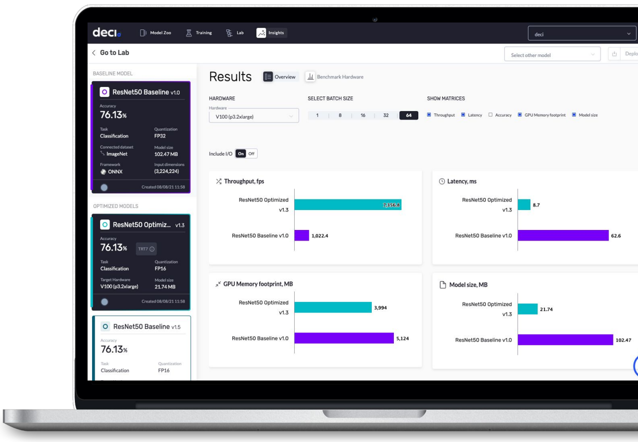
Eliminate Development Risk and Reach Deployment Faster

Eliminate endless trial and error iterations. Go from data to a production ready model in days.

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.

Deci is powered by AutoNAC (Automated Neural Architecture Construction), the most advanced and commercially scalable Neural Architecture Search technology in the market. Teams use AutoNAC to develop custom, production-grade models that deliver unparalleled accuracy and speed for any performance targets and hardware environments.



BASELINE MODEL

ResNet50 Baseline v1.0
Accuracy 76.13%
Task: Classification
Quantization: FP32
Connected dataset: ImageNet
Model size: 102.47 MB
Framework: ONNX
Input dimensions: (3,224,224)
Created 08/08/21 11:58

OPTIMIZED MODELS

ResNet50 Optimiz... v1.3
Accuracy 76.13% TRT7
Task: Classification
Quantization: FP16
Target Hardware: V100 (p3.2xlarge)
Model size: 21.74 MB
Created 08/08/21 11:58

ResNet50 Baseline v1.5
Accuracy 76.13%
Task: Classification
Quantization: FP16

HARDWARE
Hardware: V100 (p3.2xlarge)

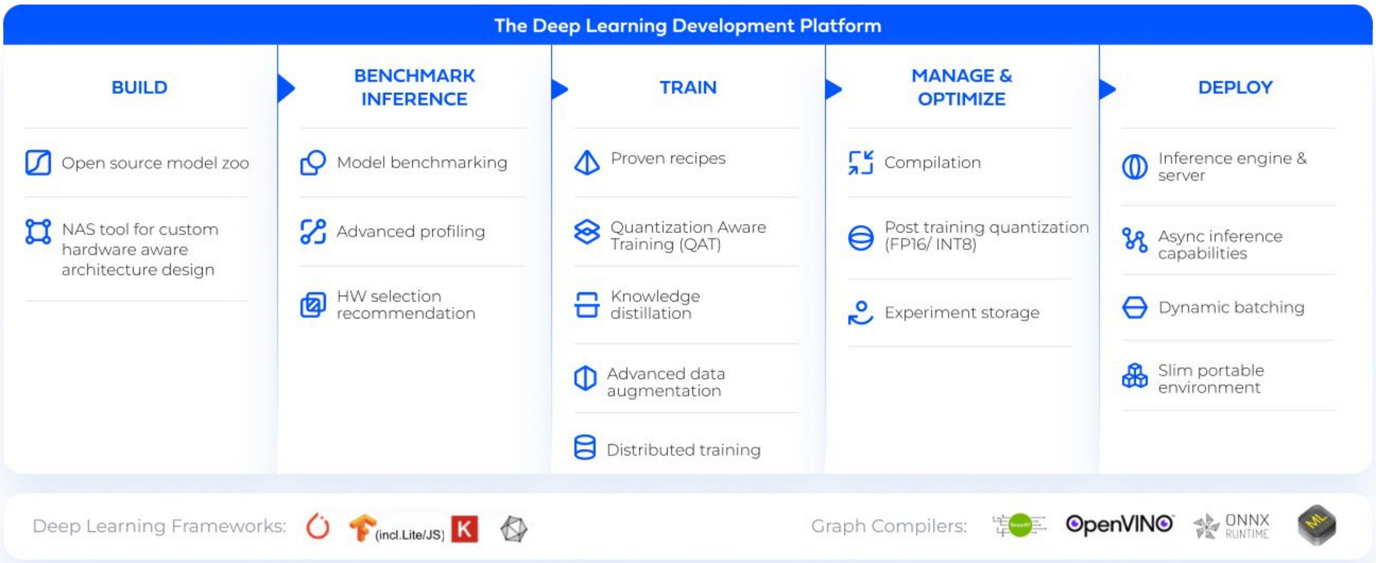
SELECT BATCH SIZE: 1 | 8 | 16 | 32 | **64**

SHOW MATRICES: ☒ Throughput ☒ Latency □ Accuracy ☒ GPU Memory footprint ☒ Model size

Include I/O: **On** Off

Deci's Deep Learning Platform

Powered by Neural Architecture Search



The Deep Learning Development Platform

BUILD	BENCHMARK INFERENCE	TRAIN	MANAGE & OPTIMIZE	DEPLOY
Open source model zoo	Model benchmarking	Proven recipes	Compilation	Inference engine & server
NAS tool for custom hardware aware architecture design	Advanced profiling	Quantization Aware Training (QAT)	Post training quantization (FP16/ INT8)	Async inference capabilities
	HW selection recommendation	Knowledge distillation	Experiment storage	Dynamic batching
		Advanced data augmentation		Slim portable environment
		Distributed training		

Deep Learning Frameworks: (incl.Lite/JS)

Graph Compilers: OpenVINO ONNX RUNTIME

Main Capabilities Overview

Gain Superior Performance with Custom Architectures

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

Simplify Runtime Optimization

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

Easily Find the Best Hardware for the Job

Benchmark your models' inference performance across multiple hardware types with Deci's online hardware fleet. Get actionable insights and select the optimal hardware.

Maximize Accuracy with Advanced Training Techniques

Train models with SuperGradients. Leverage custom recipes and advanced training techniques (e.g., knowledge distillation, quantization-aware training) with one line of code.

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 with parallelism and concurrent execution. Compatible with multiple frameworks and hardware types.

Automate Model Benchmarking

Easily measure and compare the performance of various models on your inference hardware.

For Everyone in the AI-Driven Organization

Tech Executives

Reduce time to market by 80% and lower development cost by 30%. Supercharge your AI teams with advanced tools.

Data Scientists

Deliver state-of-the-art models, faster than ever, without worrying about performance or model size. Focus on your core competency: solving business problems with AI.

ML Engineers

Easily optimize models for various hardware types with a few clicks. Seamlessly deploy and maximize application performance with advanced serving capabilities.

Product Leaders

Unlock new use cases and release new features to production faster without compromising on quality.

Amir Bar
Head of Software & Algorithms
APPLIED MATERIALS

> *"We have been working with Deci on optimizing the performance of our AI model, and managed to reduce its GPU inference time by 33%. This was done on an architecture that was already optimized. We will continue using the Deci platform to build more powerful AI models to increase our inspection and production capacity with better accuracy and higher throughput."*

Book a Demo

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

deci.