deci

Optimize Deep Learning Models for Fast & Cost Efficient Cloud Inference at Scale

Introduction

Running your deep learning models in cloud environments is a costly matter. High inference costs can dramatically cut down your product's profitability. To win, you need to ensure that your AI infused applications can be highly accurate, fast, and cost efficient. Achieving a sweet spot among these three parameters is often a struggle for AI teams who end up relying on more powerful and expensive cloud instances to support their applications' inference.

At the end of the day, companies face the dilemma of executing model inference at a lower operating margin but with high performance, or sacrificing the user experience with poorly-performing deep learning models.

With Deci, you can boost your models' performance and maximize hardware utilization to cut down inference time and cost on existing hardware as well as migrate workloads to more affordable cloud instances. Below are three case studies that demonstrate how companies were able to improve performance and cut their cloud cost by optimizing their models with the Deci platform.

Use Case 1

Enabling Cost Efficient Inference of Facial Liveness Classification at Scale

A customer developing a security application powered by a facial liveness classification model was looking to gain real time performance as well as cost efficiently scale their business. The customer's original classification model was deployed on an NVIDIA V100 GPU and suffered from high latency.

By using the Deci platform, the customer built a new model architecture which delivered a significant acceleration of latency while maintaining the original model's accuracy.

Results

67.7% Cloud Cost Reduction

3.9x Latency Acceleration

3.9X Lower Latency

Measured on NVIDIA V100 GPU

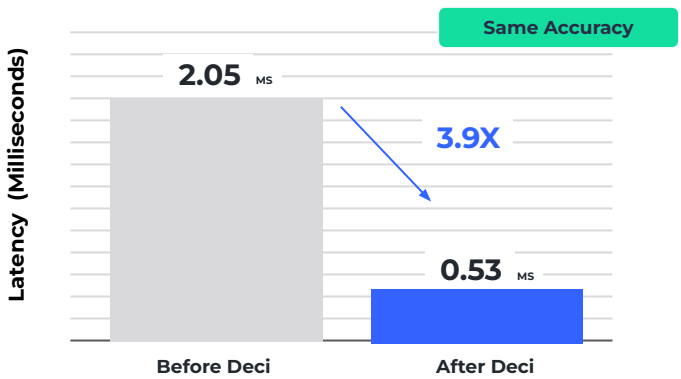


Figure 1: Inference performance measured end to end on an NVIDIA V100 GPU.

67.7% Lower Inference Cost

Measured on NVIDIA V100 GPU

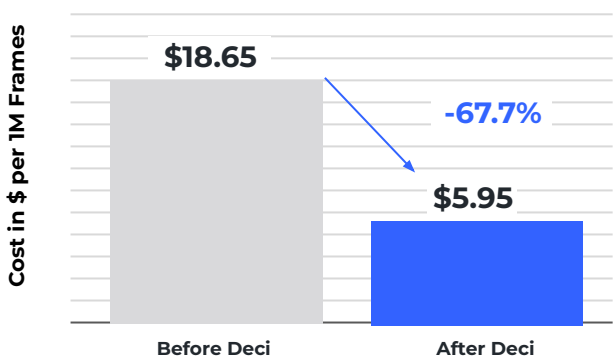


Figure 2: Inference cost is calculated per 1 million frames on NVIDIA V100 GPU based on a $0.752 hourly on demand rate on an AWS machine (g4dn.2xlarge).

Use Case 2

Reducing Cloud Costs and Improving UX for a Text Summarization Application

A customer developing an AI platform for text summarization was struggling to achieve satisfactory latency performance on a model powering their application. This led to a poor user experience as well as high cloud costs. The model was deployed on NVIDIA T4 GPU.

The customer used Deci's compilation and quantization tools to easily optimize the model performance and significantly reduce cloud cost as well as improve the user experience.

Results

- **68%** Cloud Cost Reduction
- **3.92x** Latency Acceleration
- **50%** Model Size Reduction

3.92X Lower Latency

Measured on NVIDIA T4 GPU

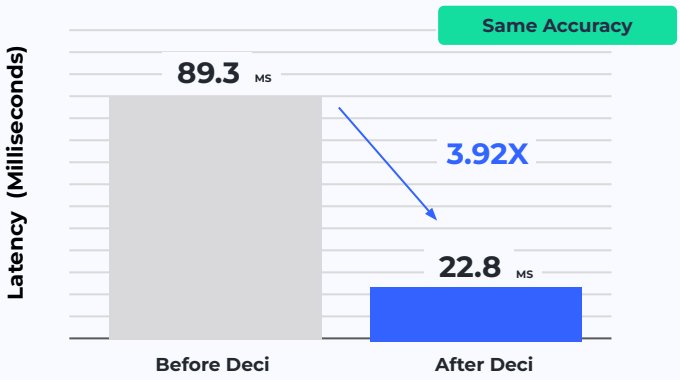


Figure 1: Inference performance measured end to end on an NVIDIA T4 GPU and calculated for a batch size of 1.

68% Lower Inference Cost

Measured on NVIDIA T4 GPU

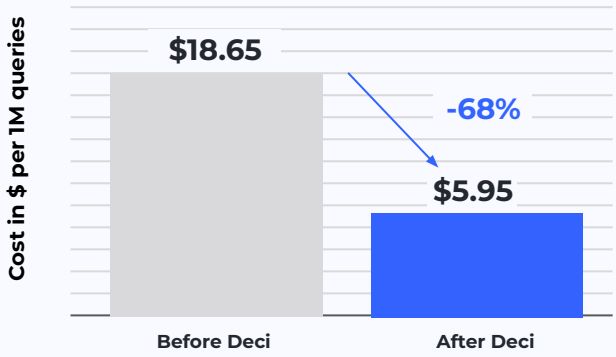


Figure 2: Inference cost is calculated per 1 million queries on NVIDIA T4 GPU based on a $0.752 hourly on demand rate on an AWS machine (g4dn.2xlarge).

Use Case 3

Enabling Real Time NLP Inference on CPUs

A multinational manufacturer and distributor of electricity and gas was looking to improve the latency of an automatic extraction of text from documents and images using OCR and NLP.

Using Deci's AutoNAC engine the customer was able to gain 8.3x faster latency compared to its original model while also improving the accuracy from 77.17% to 80.21% (word-level).

Results

- **85%** Cloud Cost Reduction
- **8.3x** Latency Acceleration
- **+3%** Accuracy Increase

8.3X Lower Latency

Intel(R) Xeon(R) Platinum 8175M CPU

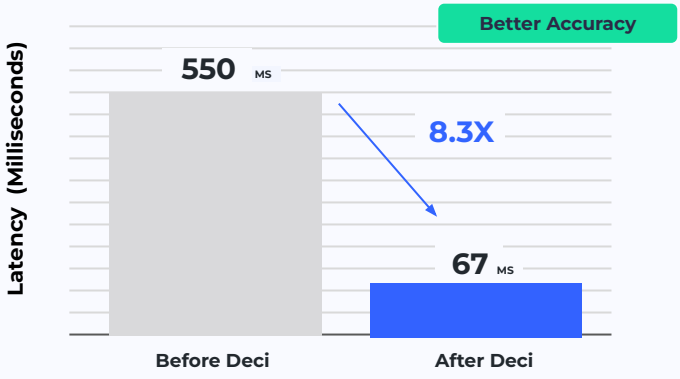


Figure 1: Inference performance measured end to end on Intel(R) Xeon(R) Platinum 8175M CPU based on and calculated for a batch size of 32.

85% Lower Inference Cost

Intel(R) Xeon(R) Platinum 8175M CPU

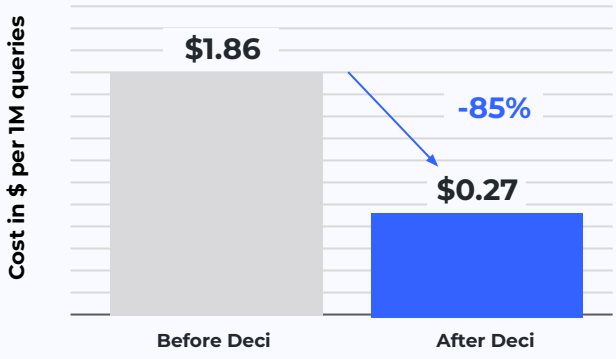


Figure 2: Inference cost is calculated per 1 million queries on Intel(R) Xeon(R) Platinum 8175M CPU based on a $0.38 hourly on demand rate on an AWS machine (m5.2xlarge).

By Using Deci's Platform You Can:

Improve Hardware Utilization and Reduce Inference Cost

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

Improve UX with Better Inference Performance

Maximize hardware utilization and cost-efficiently scale your solution.

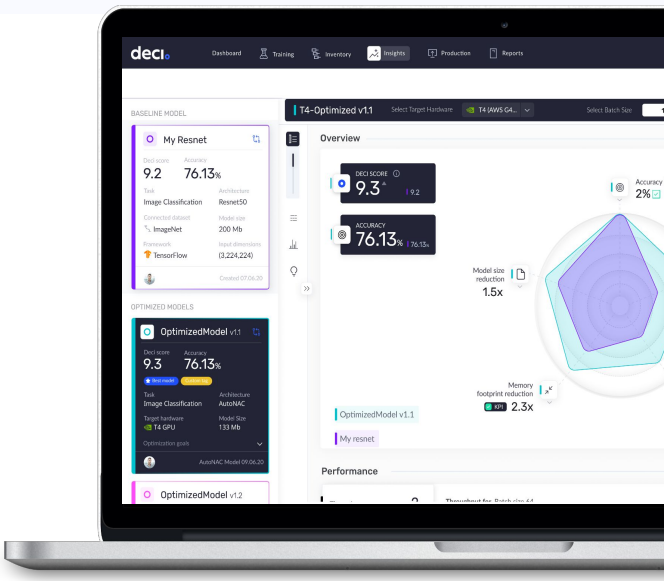
Simplify Development and Shorten Time to Market

Automate model selection and optimization. Eliminate uncertainty, guarantee success in production, and reach production faster.

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.



Feature Highlight

Easily Find the Best Hardware for the Job

Benchmark your models' expected inference performance across multiple hardware types on Deci's online hardware fleet. Get actionable insights for the ideal hardware and production settings.

Gain Superior Performance with Custom Architectures

Build accurate & efficient architectures tailored for the application, hardware, and performance targets with Deci's AutoNAC engine.

Simplify Inference Runtime Optimization

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

Maximize Accuracy with Advanced Training Techniques

Train models with SuperGradients and leverage custom training recipes with one line of code.

Automate Model Benchmarking

Easily benchmark and compare the performance of different models on your target inference hardware against other hardware including GPUs, CPUs, and commercial edge devices.

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 model serving performance with parallelism and concurrent execution. Compatible with multiple frameworks and hardware types.

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

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

deci