deci

2.1x Acceleration Leads To Cloud Costs Savings and Improved User Experience For BRIA's Gen AI Platform

Results:

50%
Lower Cost of Serving

2.1X
Lower Latency

Improved User Experience

BRIA

> **"Controlling our inference cloud spend without compromising on performance is key for our business success. Deci enabled us to scale our workloads while reducing costs and improving our users' experience."**
>
> *Dr. Yair Adato, Founder & CEO at BRIA*

Introduction

The demand for high-quality visual content continues to grow across industries. Visual generative AI foundation models such as latent or cascaded diffusion models, SAM and others, provide the ability to generate or customize realistic images. This is opening up new possibilities for many new applications. These foundation models have the potential to streamline and automate tasks that would otherwise require significant time and resources to complete manually.

BRIA provides engineers, AI teams, and researchers with safe and legal visual generative AI capabilities. With an access to trained models, source code, and comprehensive API suites, companies can enhance their products and services using BRIA's Visual Generative AI Platform. BRIA's solutions, having been trained on the world's largest, fully licensed, high-quality training set, eliminate all legal risks for commercial enterprise use.

Overview

- Visual Generative AI Platform
- Empower developers with pre-trained foundation models, source-code and API suites for commercial use
- NVIDIA A10G GPU (g5.xlarge instance, AWS)

Cloud cost reduction is also a top priority for BRIA. Foundation models are larger and their inference process is more complex compared to classical AI models. To generate a new sample, a model performs several inference iterations. The combination of extremely large models and the iterative nature of the inference results in a significantly higher demand for compute power and overall inference costs. In order to run inference at a high scale, BRIA aims to reduce its cloud expenditure by optimizing inference time and increasing GPU utilization. Moreover, faster models are not only more cost-effective, but they also increase client satisfaction by reducing the latency of inference.

Solution Overview

Using the Deci's platform, BRIA's team was able to optimize the inference performance of their BRIA 1.4 and Segment Anything models and reduce their inference cloud cost by 50%.

BRIA's team used Deci's Infery library to easily perform a hybrid compilation and selective quantization in their complex diffusers and transformers based architectures.

Infery automatically profiles the architecture's sub-components and layers and then leverages the optimal production orientated framework and quantization level for each one, all while taking into account the inference hardware characteristics. Using Infery, BRIA's team was able to maximize the acceleration potential of their complex models while saving valuable time and effort.

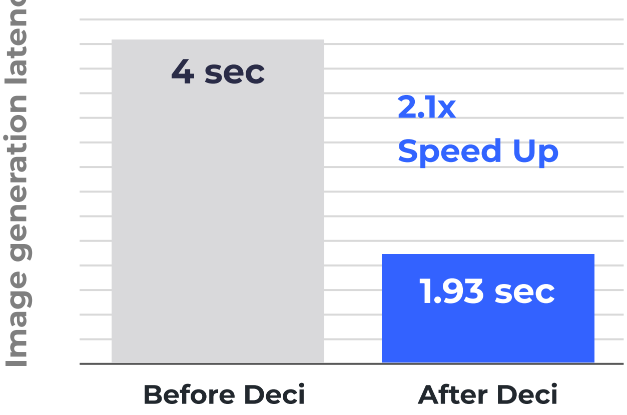
Infery's optimization module was easily integrated into BRIA's CI/CD pipeline. In its production environment, BRIA uses Infery's deployment module, which includes advanced inference capabilities, integrated as a backend inference engine with NVIDIA Triton server.

BRIA's Results Overview

The optimized model lead to an improved user experience and a significant reduction in cloud cost.

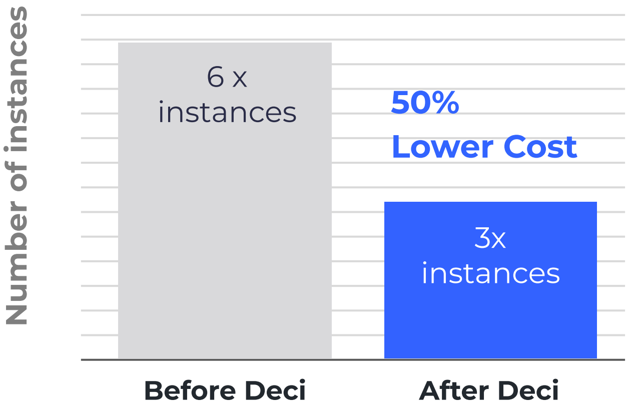
2.1x Faster Inference of Stable Diffusion

Measured on NVIDIA A10G GPU



50% Lower Compute Cost

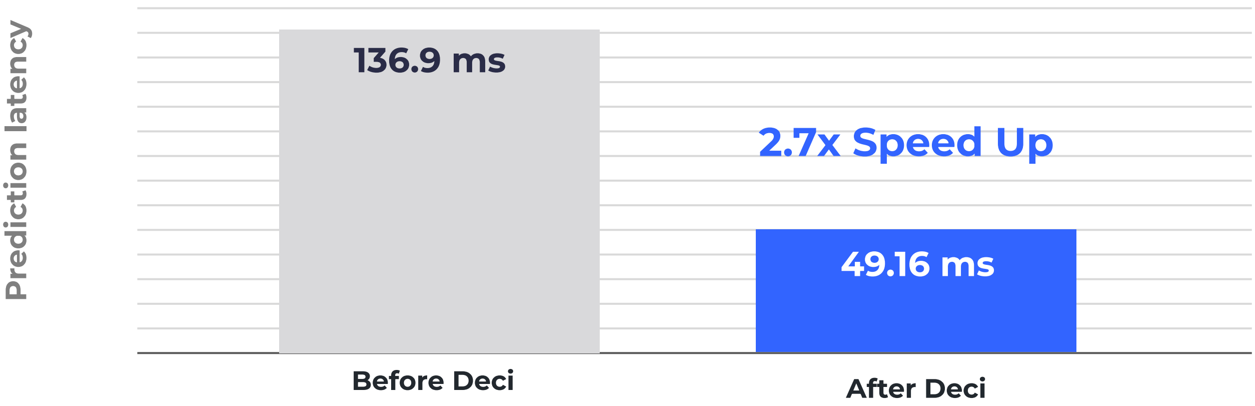
On demand g5.xlarge AWS instances



*50 steps of BRIA 1.4 on A10G instance. 512x512 image resolution.

2.7x Faster Inference of Segment Anything Model

Measured on NVIDIA A10G GPU



Elevate Your AI Capabilities with Deci's Powerful Foundation Models and Developer Tools

Power your AI solutions with the world's most efficient foundation models. Customize Deci’s models for your task with a comprehensive suite of fine-tuning, optimization, and runtime tools for seamless deployment.

Launch your Gen AI Apps Faster

Use enterprise-grade models. Lower risk, shorten dev time from months to days.

Cut your Training and Fine-Tuning Cost

Deci’s NAS generated foundation models are trained and fine-tuned 2x faster compared to other Gen AI models.

Scale Your Gen AI Inference Cost-Effectively

Save up to 80% on your inference cost. Migrate workloads to affordable & widely available HW.

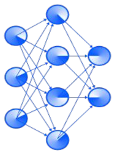
Improve User Experience with Better Inference Speed

Ship better products and delight users with up to 5x lower latency performance compared to other SOTA models.

Deci Platform

Foundation or Custom Models

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



SaaS

AutoNAC

Neural Architecture Search Engine

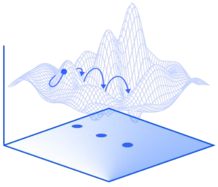
On Prem

DataGradients™

Dataset analyzer

Train or Fine-tune

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



On Prem

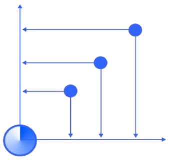
SuperGradients™

PyTorch Training library

Training Frameworks

Optimize & Run

Apply acceleration techniques. Run self-hosted inference anywhere.



On Prem

Infery

Optimization & Inference Engine SDK

Runtime Frameworks

Why Deci?

Unparalleled Performance

The world's most efficient and cost effective models, generated with Deci's AutoNAC engine. Get unrestricted access, host anywhere you’d like.

Control Over Model Quality & Customization

Fine-tune, and adjust content filters. Gain competitive edge through additional advanced model customizations.

Enterprise Ready Secure & Compliant

Self-hosted inference. No vendor lock-in. Ideal for enterprises and for handling sensitive data.

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.