

Achieve Production-Grade Performance, Faster.

Simplify and accelerate the development of deep learning with advanced tools to build, optimize, and deploy highly accurate and efficient models.

The Barrier to Production

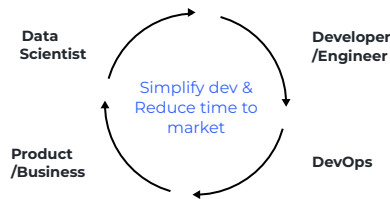
Deep learning-powered advancements in AI have led to innovations that have the potential to revolutionize companies across many industries. But prolonged development cycles, high computing costs, and impracticality of deployment are major roadblocks in the path to AI implementation.

Deci's Deep Learning Platform

Deci is a deep learning development platform that enables AI practitioners to easily build, optimize, and deploy highly accurate and efficient models to any environment including cloud, edge, and mobile. Powered by its proprietary automated Neural Architecture Construction technology (AutoNAC), Deci harnesses AI to eliminate production-related bottlenecks across the deep learning lifecycle.

Shorten Development Time

Go from data to production-ready models in days. No more manual trial & error iterations, intense engineering efforts, or time wasted on boilerplate & pipeline code. Reach production faster and empower your teams with advanced tools.

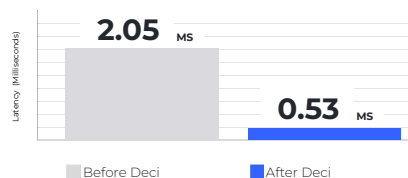


Accelerate Inference Performance

Eliminate dev risk & uncertainty. Easily accelerate inference performance on any hardware by up to 5x without compromising on accuracy. Reduce the latency of real-time applications or increase throughput for large scale inference.

3.9x Faster Latency (Same Accuracy)

Inference HW: NVIDIA V100 GPU | Task: Facial liveness classification

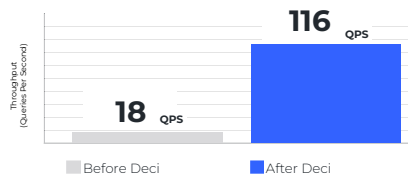


Scale Up Inference Efficiently

Dramatically increase inference hardware utilization to make the most of your current hardware or move to one that's more affordable. Enable new use cases on CPU and edge devices or cut up to 80% of your cloud compute costs.

6.44x Higher Throughput (Same Accuracy)

Inference HW: AMD EPYC 7773X | Task: NLP, Question answering



Deep Learning Across the Board

Deep Learning Tasks

- Computer Vision: Image Classification, Object Detection, Semantic Segmentation
- NLP

Deployment Options

- Public Cloud: AWS, GCP, Azure
- Datacenter
- Edge / Server
- Mobile
- Browser

Inference Hardware

- CPU
- GPU
- ARM-based
- Mobile AI Chip
- Specialized AI Chip

Sample Use Cases

- Video analytics
- Security cameras
- Autonomous cars
- Image/video editing
- Medical diagnostics
- Manufacturing QA
- Robotics
- Media & entertainment

Vadim Zhuk,
Senior VP R&D

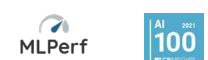
RingCentral

"At RingCentral, we strive to provide our customers with the best AI-based experiences. With Deci's platform, we were able to exceed our deep learning performance goals while shortening our development cycles. Working with Deci allows us to launch superior products faster."

Trusted by Industry Leaders

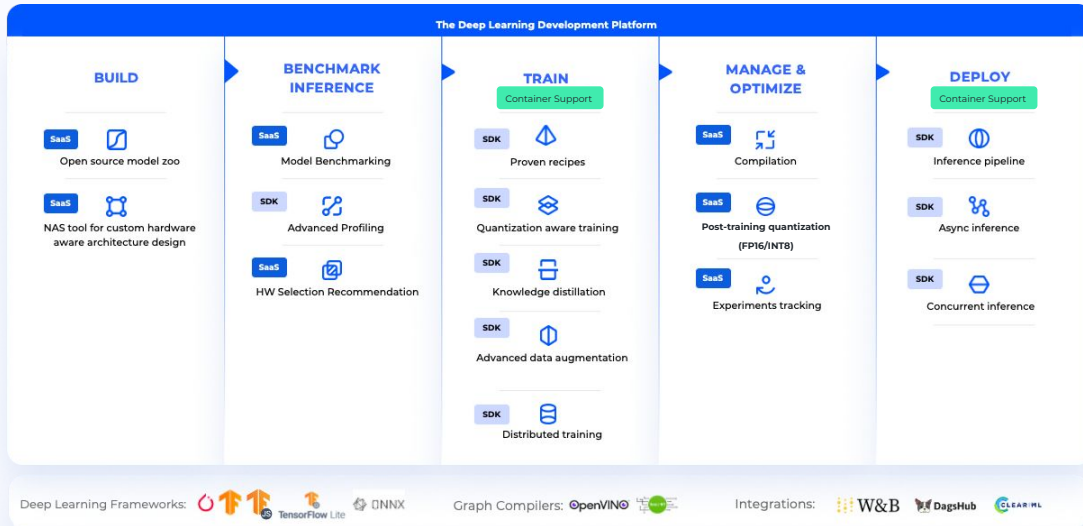


Recognition and Awards



Production-Aware Development with Deci's Deep Learning Platform

Go from data to production-ready model in 14 days



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 AutoNAC engine.



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.



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



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.



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.

80%

Shorter development process on average

30%

Lower development costs on average

5x

Inference acceleration

5x

Inference & training cost reduction

Pallav Vyas
Senior Engineering Manager, Document AI & Innovation

"At Adobe, we are committed to redefining the possibilities of digital experiences and delivering excellent AI-based solutions across a wide range of cloud and edge environments. By using Deci, we significantly shortened our time to market and transitioned inference workloads from cloud to edge devices. As a result we improved the user experience and dramatically reduced our spend on cloud inference cost."

BOOK A DEMO

For more information, please contact:

sales@deci.ai

