Enabling Real Time Semantic Segmentation for a Video Conferencing Application

Executive Summary

The world of video conferencing is constantly evolving, with new innovations emerging every day. There is a concerted effort among companies to gain a competitive edge by enhancing their applications with AI capabilities. Developers need to be able to enhance their applications while still maintaining the user experience. A large enterprise customer was struggling to achieve real time people segmentation to power its video conferencing application. By working with the Deci platform the customer was able to quickly achieve the desired performance goals and deliver a superior user experience.

3x
Latency Acceleration

4.47x
Reduction in Model File Size

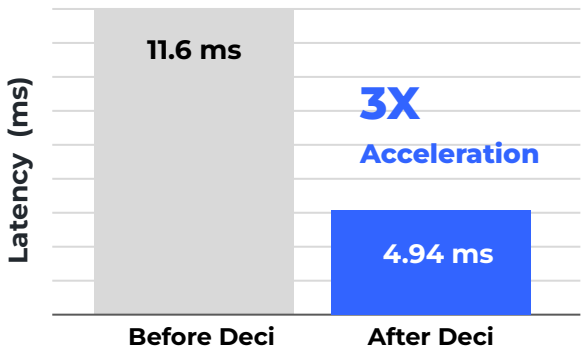
22%
Reduction in Memory Footprint

The Challenge

The customer sought to improve the latency of a person segmentation model (“Stacked-Hourglass”) which was trained on Face Synthetics, an animated dataset of facial images. The model powered a conference room application that was not achieving the targeted real time latency on their desired hardware, a Qualcomm® Snapdragon™ 888 board. The customer wanted to reduce the model’s latency to achieve real time performance but also wanted to preserve the model’s accuracy level.

The Solution

By leveraging the Deci platform the customer generated a custom model architecture that was tailored for the use case and the Qualcomm board. The new segmentation model reduced the latency by 3x from 11.6ms to 4.94 ms. In addition, the model file size was reduced by 4.47x and the memory footprint was reduced by 22%, all while preserving the original model accuracy.



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

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**