ICURO: Active Deep Learning for AI Visual Measurement on Intel® Arc™ A770

A collaboration between ICURO and Intel to enable training and inference for object detection on Intel GPU

Story at a Glance

ICURO is an AI vision company that helps companies get the benefit of deep learning vision solutions. We are using the Intel® Arc™ A770 discrete GPU with the Intel® Extension for PyTorch* for training a no-code AI visual measurement solution based on object detection. The application works by supporting annotation, training, and inference all-in-one in an incremental, interleaved fashion so that annotation effort is minimized.

Since 2017, ICURO has been helping companies get the benefits of deep learning vision solutions. The company works in the domains of manufacturing, robotics, and field inspection. Its efforts assist manufacturing companies in achieving Industry 4.0 by applying AI/machine learning to their processes.

An client company asked ICURO to find a way to avoid manual effort to make visual tolerance measurements on thousands of semiconductor microscopy images. This process is more complex than defect detection because the measurements are subject to tolerance ranges and measurement points must be identified with accurate positioning. To solve that problem, ICURO created a no-code AI solution for them that leverages AI to reduce the effort, increase accuracy, and speed up turnaround time.

We call this visual measurement and it can be applied to more than just semiconductor images.

Challenge

The usual process of visual measurement is as follows:

1. Collect images, usually +400 per work-product
2. Identify the measurement units per pixel
3. Farm out to workers who annotate the images
4. Combine the results into a report

This manual process usually takes a few days. ICURO developed a vision solution that uses deep learning to measure size and distance between features in images. The user's goal is to work through a batch of images and make a report on the measurements so that tolerances can be verified. ICURO's strategy is to make a solution that applies to a wide variety of visual measurement applications.
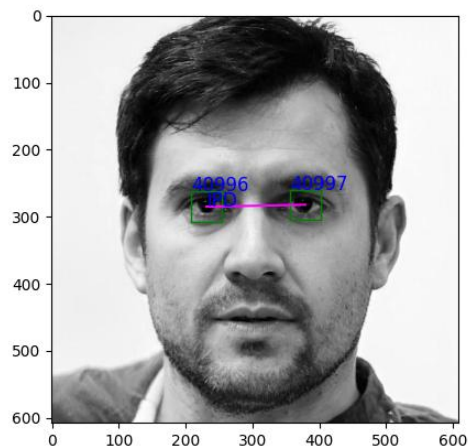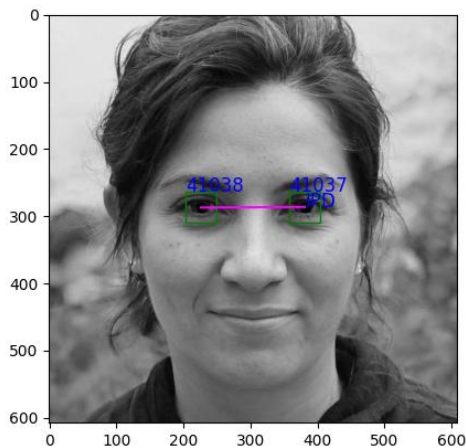
Solution

ICURO developed an interactive no-code AI application that combined annotation of images with training. The user does some annotations, triggers training, and looks at the results. This process is known as active learning and can be described with the following steps:

1. Upload a set of images
2. Annotate some of these images
3. Press Run button to trigger training and inference
4. If there are images with missing or incorrect measurements, go to step 2
5. Request report

Compared to usual deep learning projects, this is much simpler since it does not need code, management of training and test sets is fully automated, annotation and inference are done in one application instead of two, and the report is also built-in.

As an illustration of this application, we used it to train for the measurement of inter-pupil distance (IPD) of people's faces. This is needed to prepare eyeglasses so the center of focus of the lenses is centered on each eye. Here are two example images that were inferred:

(These faces were generated, not real people. For actual deployment to measure IPD, there needs to be a benchmark included in the image to gauge the number of pixels per millimeter.)

In order to deploy this solution, we needed a GPU for training and inference and the proper software to maximize performance. We turned to the Intel® Arc™ A770 discrete GPU and Intel® Extension for PyTorch*. In addition to this, we also used the Computer Vision Annotation Tool (CVAT) to create data for system testing.

Optimizing with Intel Hardware and Software Solutions

Intel® Arc™ Graphics is a family of discrete GPUs designed to address the high-performance needs of the developer market. Based on the X^e HPG microarchitecture, the Intel® Arc™ A770 features X^e cores with Intel® Xe Matrix Extensions (Intel® XMX) and dedicate matrix engines. To make optimal use of this hardware, Intel also provides the Intel® Extension for PyTorch* which extends the base PyTorch framework to enable performant training and inference on Intel GPU with minimal code changes. With IPEX, users need simply to replace previous `cuda` calls with `xpu` in order to send the model and corresponding data to the Intel GPU.

Performance Results & Benefits

In close collaboration with Intel, the ICURO team was able to optimize the object detection backbone of their AI visual measurement solution using the Intel® Extension for PyTorch* to accelerate training and inference on the Intel® Arc™ A770 discrete GPU. This ended up being a more cost-efficient solution while providing comparable performance to similar, alternative GPUs on the market.

We conducted a test on the object detection neural network using 40 annotated images all 608 high and 608 wide, with 100 epochs of training. The first epoch consumes a significant amount of time as it includes model loading. We focus on the time required to process each image in every epoch as our key performance indicator. This KPI possesses the ability to scale for a variable number of epochs, and our empirical data provides estimates for the time needed to process different image sizes. Owing to these factors, we disregard the time of the first epoch, concentrating instead on the time required to process each image for one epoch. This resulted in an average time per epoch that amounted to 4.878 sec.
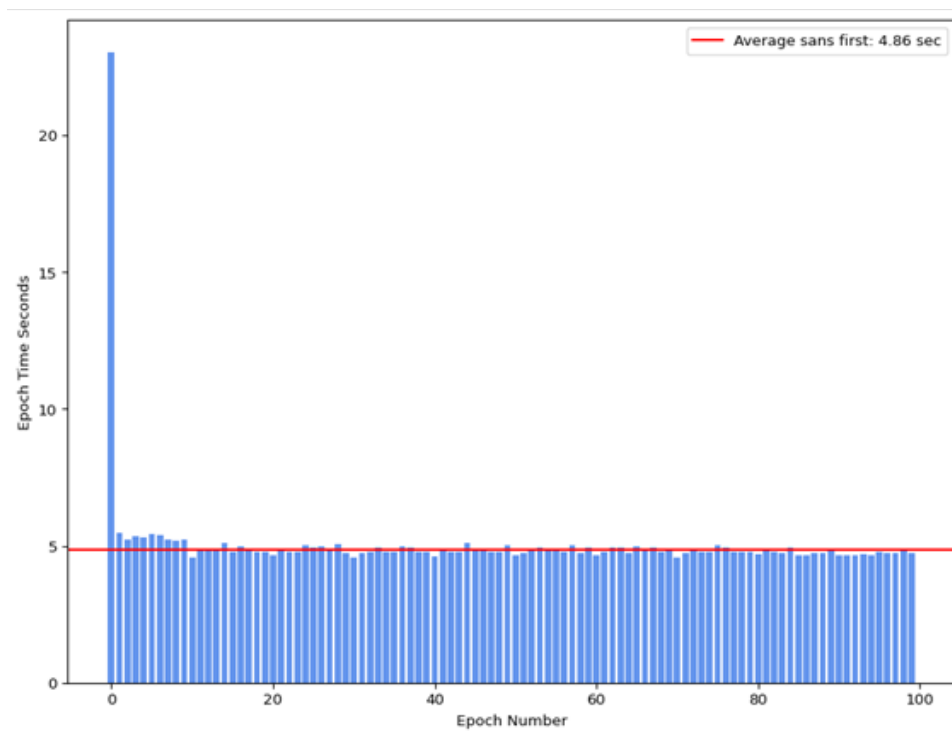

Figure 1: Time per epoch (s) during training

Now that we know the per epoch time on the A770, we can reliably estimate training times that figure into the active learning approach. This gives us assurance the A770 will be able to support the end-user experience we want in our product.

"ICURO and Intel have collaborated in the realm of robotics and visual analytical tools. The collaboration has reached new heights with the aid of Intel® Extension for PyTorch and AI Tools powered by oneAPI. The integration of Intel® Arc™ A770 graphics card has unleashed the power of active deep learning. Intel® Developer Cloud has played an invaluable role in facilitating testing and benchmarking. The collaboration between ICURO and Intel is transformative and offers extraordinary possibilities."

— Paul Baclace, chief AI Architect, ICURO

Summary & Conclusion

ICURO developed a no-code AI visual measurement solution based on an object detection model. This solution was trained in an active learning fashion where an initial sparse set of annotations is used for training and inference and the user is iteratively queried to analyze for missing or incorrect measurements. Using Intel® Extension for PyTorch* made it easy to enable and deploy this solution on the Intel® Arc™ A770 discrete GPU for performant training and inference. Through this, we were able to achieve an average time of less than 5 seconds per epoch, which is comparable to alternative GPUs on the market. This performance ensures that our solution ultimately meets our high standards of efficiency and effectiveness to deliver a superior end-user experience.

About ICURO

ICURO is a renowned AI company that specializes in providing comprehensive AI solutions, covering everything from strategic business models to complete system deployment. Their proven AI platforms unify the power of distributed computing, sensing, and actuation to deliver enterprise growth outcomes across multiple industries and markets

Solution Ingredients

- Intel® Arc™ A770 graphics card
- Intel® Xe Matrix Extensions (Intel® XMX)
- Intel® Extension for PyTorch* (IPEX)

Learn More

Intel AIML Ecosystem

Intel® Extension for PyTorch*

Intel® Arc™ Graphics