WHITE PAPER

Artificial Intelligence
Deep Learning
Intel® AI Builders

Manufacturing Package Fault Detection Using Deep Learning

Executive Summary

Intel's Software and Services Group engineers recently worked with assembly and test factory engineers on a proof of concept focused on adopting deep-learning technology based on Caffe* for manufacturing package fault detection. The results proved that neural network technology can be applied to silicon manufacturing. They also showed that the Intel® architecture platform has competitive performance and can easily be used to provide both neural network training and inference support.

Background

Silicon packaging, one aspect of semiconductor manufacturing, is a complex and expensive process that requires high quality. During the packaging process, various factors, such as fingerprints, scratches, and stains, can cause cosmetic damage. These damages need to be manually inspected to determine whether they exceed the threshold of allowable damages. As the final checkpoint on the product line with more than 11 criteria rules, the inspection process is subjective due to the complexity of the damage scenarios and human error and inconsistency.

A deep neural network has been proven to outperform traditional methods in terms of image processing. Although topologies such as GoogLeNet have shown good accuracy on general ImageNet tests, questions remain as to whether these topologies can be used for high-quality manufacturing tests. Typically neural network training is done on a GPU, and whether the Intel® architecture platform can provide similar capability is yet another question. This proof of concept (PoC) provided positive answers to these questions.

Problem Statement

This PoC aimed to reduce the human review rate for package cosmetic damage at the final inspection point, while keeping the false negative ratio at the same level as the human rate. The input was package photos, and the goal was to perform binary classification on each of them, indicating whether the package was rejected or passed. Manual inspection followed a set of rejection criteria on damages of particular shapes and location and sizes exceeding a particular threshold. The manual inspection required a low false negative, and the majority of the input photos that were inspected passed. The unbalanced pass/reject input photo number ratio, coupled with complicated judgment criteria rules made the manual inspection work tedious and prone to errors.

Table of Contents

Solutions

SSG first proposed GoogLeNet V1 topology based on the convolutional neural network (CNN). This topology balances the training/inference time and testing accuracy, making it well suited for use as image classification. To tailor the topology, we took only the green channel input and reduced the Full Connection layer class from 1,000 to 2 for binary classification. The required top-1 accuracy is much stricter than that for the standard GoogLeNet V1 on ImageNet-1k (approximately 68.7 percent).

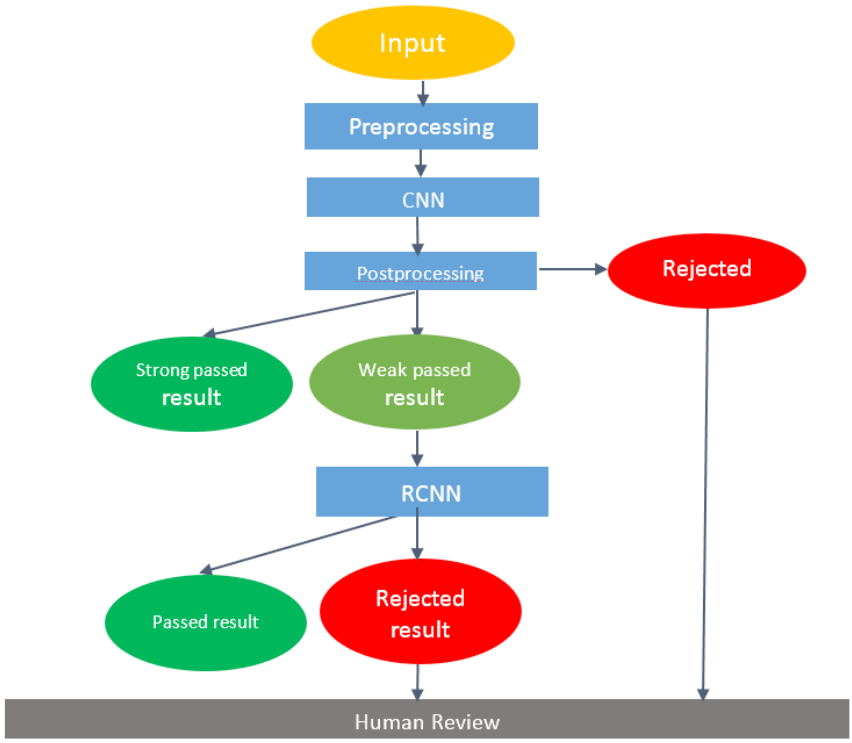
The training was supervised learning. We took about 4,000 images as input, labeled them as either passed or rejected, and then rotated each image 36 times, 10 degrees each time, for data augmentation purposes. We did not augment the images with different scales, since the damage criteria were size sensitive. We fed the input into GoogLeNet V1 for training. For the output result, we classified each original image by 36 images result ensemble.

Topology	Number of image classes	Number of training input images	Top-1 accuracy
Customized GoogLeNet V1	2 (passed or rejected)	14,400 (4000*36)	Human-level false negative (far greater than 68.7 percent)
Standard GoogLeNet V1	1,000 (ImageNet-1k)	1.2 million (ImageNet-1k)	68.7 percent

In addition to CNN, we added region-based CNN (RCNN) to meet the strict false negative rate goal, which is typical for manufacturing. This goal is hard to achieve with a single CNN classification model. A major problem is that the human decision of "passed" or "rejected" is likely inconsistent with the class boundaries (for example, damages with a size around the threshold). These ambiguous labels confuse CNN. We first tackled the problem by sacrificing the false positive rate to reduce the false negative rate. To do this, we relabeled the input images to put images around the boundary into the rejected class. However, the results were still not satisfying. Even though the false negatives went down, the false positives went up rapidly. Eventually we decided to add RCNN to enhance the detection accuracy.

RCNN can detect an object's location, size, and type. We used the ZFNet-based Faster RCNN model. We leveraged the class probability output of CNN and further categorized the image samples into three classes: strong passed, rejected, and weak passed. The weak passed category, which has low confidence per CNN probability, is likely a result of ambiguous training, so we took the weak passed class output and fed it into the RCNN network. Any input with a detectable defect was put in the rejected class; otherwise the input was put in the strong passed class.

All rejected images were inspected again by humans, who served as the final gatekeepers.



Results

The final output from the two concatenated networks was revealing. The false negative rate consistently met the expected human-level accuracy with the months' live manufacturing data testing. The false positive rate was approximately 30 percent, which means that 70 percent of the manual inspection effort was saved. The human inspectors used to identify one rejected image out of 10 input images. Now that rate is one out of three, which inspires them to work with lower workload.

We compared the prediction result from pure CNN to the CNN and RCNN approach. RCNN helped to reduce the false negative rate to the strict target, while still keeping the false positive within an acceptable range. Since human inspectors may also misclassify the image samples around the criteria boundary, the model also served as a cross-check to educate human inspectors and improve their inspection quality.

The deep neural network solution based on Caffe optimized for Intel architecture was developed by SSG and handed over to TMG engineers, who played the role of domain experts. Even without artificial intelligence knowledge, TMG engineers are able to easily fine-tune the network with new samples. Our practice showed that by gradually adding new unseen samples to the training set and retraining the models, this deep learning-based solution can show consistent performance over time.

Intel® Architecture for Training

The solution was initially designed to run on a GPU, but later a decision was made to migrate it to the Intel architecture platform. The requirement is to finish the GoogLeNet V1 training within 15 hours. With a little migration effort, we can run GoogLeNet V1 based on Caffe optimized for Intel architecture on the Intel® Xeon® processor E5-2699 and Intel® Xeon® Platinum 8180 processor platforms. Both these platforms can provide time to train within the requirement range.

Hardware	TFlops	Software	Batch Size	Images per Second	Time to Train
Intel® Xeon® processor E5-2699	3.1	Berkeley Vision and Learning Center Caffe with CPU mode	36	1.5	55 days (estimated)
Intel Xeon processor E5-2699	3.1	Caffe optimized for Intel® architecture with the Intel® Math Kernel Library (Intel® MKL)	36	173	11.5 hours
Intel® Xeon® Platinum 8180 processor (2.5 GHz 8180)	8.2	Caffe optimized for Intel architecture with the Intel MKL	36	266	7.5 hours

We later decided to migrate the training to multi-node IA environment, leveraging the multi-node training feature provided by Intel Optimized Caffe. We run the training on 4 node and 8 nodes with OPA and Ethernet connection respectively. The IA multi-node training shows great performance with very high scaling efficiency, thus resulting in great TTT (Time To Train). The training can complete within 1 hour on 8-node Intel® Xeon® Platinum 8180 processor platforms

Hardware	# of Nodes	Connection	Software	Batch Size	Scaling Efficiency
Intel® Xeon® Platinum 8180 processor (2.5 GHz 8180)	4	OPA	Berkeley Vision and Learning Center Caffe with CPU mode	36	96.92%
Intel® Xeon® Platinum 8180 processor (2.5 GHz 8180)	8	OPA	Caffe optimized for Intel® architecture with the Intel® Math Kernel Library (Intel® MKL)	36	96.61%
Intel® Xeon® Platinum 8180 processor (2.5 GHz 8180)	4	10Gb Ethernet	Caffe optimized for Intel architecture with the Intel MKL	36	93.69%
Intel® Xeon® Platinum 8180 processor (2.5 GHz 8180)	8	10Gb Ethernet	Caffe optimized for Intel architecture with the Intel MKL	36	91.84%

Conclusion

This PoC demonstrated that deep-learning technology can be applied to the manufacturing field with high-quality requirements. The architecture of CNN can learn the sophisticated features from the input images for classification. The combination of CNN and RCNN can provide low false negative rates with a reasonable false positive rate.

The PoC also proved that the Intel architecture platform can be used for real deep-learning application training. Both the Intel Xeon processor E5-2699 and Intel Xeon Platinum 8180 processor platforms can meet user requirements. Caffe optimized for Intel architecture improved the training performance by 100 times on Intel Xeon Platinum 8180 over Berkeley Vision and Learning Center Caffe on Intel Xeon processor E5-2699. This significant performance improvement, plus the great multi-node performance has made training on Intel architecture a reality.