MATH 6380P Final Project: Kaggle in-class Contest: Nexperia Image Classication II

Ganghua FAN

Hong Kong University of Science and Technology

gfanab@connect.ust.hk

December 13, 2020

Introduction

Nexperia is one of the biggest Semi-conductor company in the world. They produce billions of semi-conductors every year. Meanwhile, a lot of unqualified devices are mixed with the good ones. Mass production makes it diffcult for human workers to examine all of the products. Therefore, we would like to use **modern machine learning methods**, particularly **deep learning**, to help Nexperia pick out as many defect devices as possible while preserving the good ones.

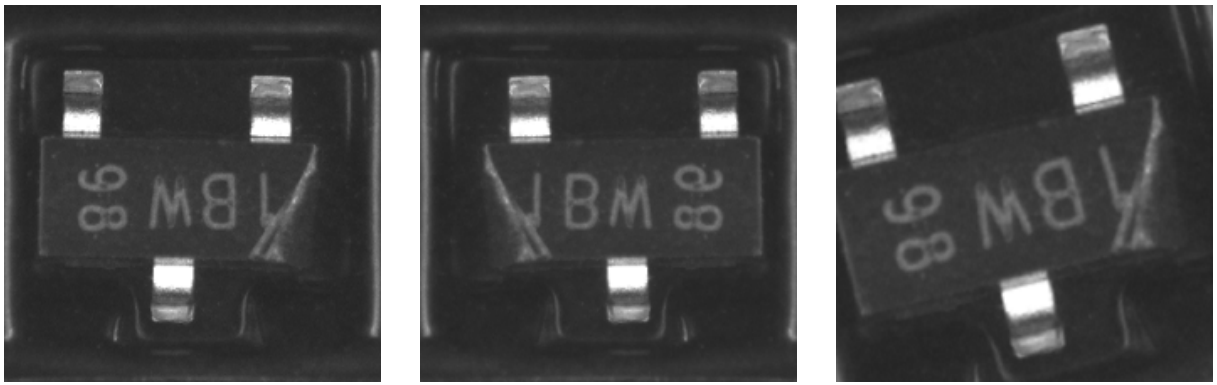
The Nexperia image dataset in the Kaggle contest contain 34457 train images (27420 good and 7039 bad) and 3830 test images with similar good-to-bad ratio.

challenges and methods

- efficient classifier: ResNet18 with fine-tune
- imbalanced data (good-to-bad ratio in training images is about 4:1): data augment methods like images transformation, Generative adversarial network(GAN)
- anomaly detection: one-class support vector machine(OCSVM), k-nearest neighbor, local outlier factor

data augment

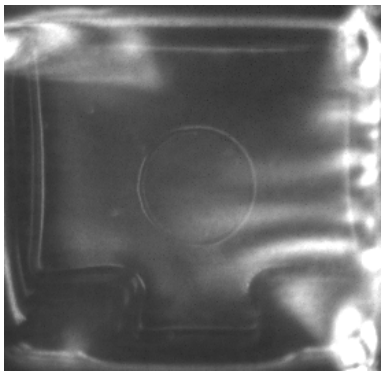
I choose image transformations like horizontal flipping, rotation a small angle to enlarge the defect images in training data. This method is simple, efficient and will not bring extra information that might pollute the original defect data.



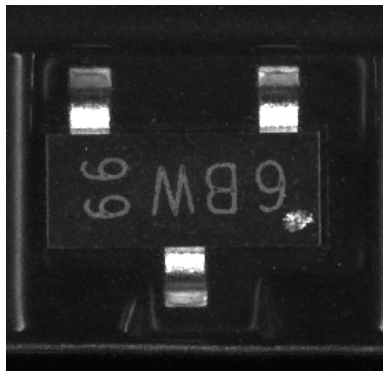
(a) origin (b) horizontal flip (c) rotation

anomaly detection

This real world dataset contains noisy labels, especially the images labeled as "good" possibly being "bad" ones in fact.
Since we do not have ground truth on which labels are wrong, unsupervised anomaly detection technique, like OCSVM, is used to detect anomalies after using Scattering Net to extract fecture. Here are some outliers OCSVM found:



(d) outlier in bad



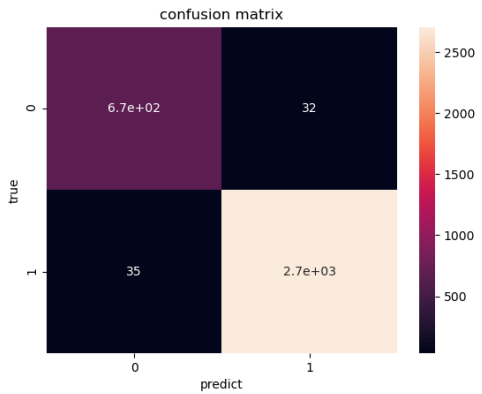
(e) outlier in good

Experiment result: Accuracy Comparison

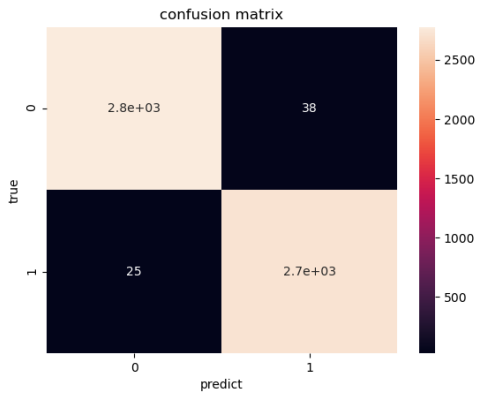
I take 10% of training data out to do the validation. The learning rate of fine-tune is 0.001. Best AUC is 97.64% obtained by ResNet18 fine-tune on augment data.

ResNet18	Without fine-tune		With fine-tune		
Training data	Original	Augment	Original	Augment	without outlier, Augment
Val accuracy	86.68%	80.98%	98.06%	**98.87%**	98.60%

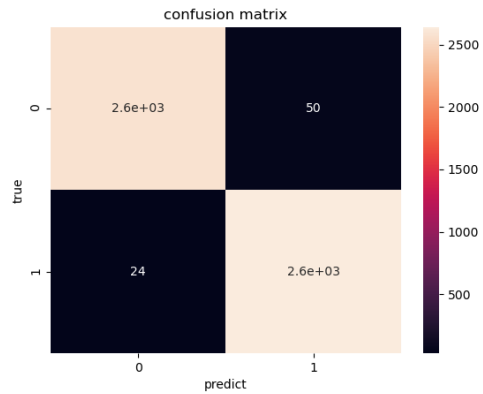
Experiment result: Confusing Matrices



(f) origin data



(g) augment data



(h) augment without outliers

Figure: Confusion Matrices. 1 stands for good semiconductor and 0 stands for defect one.

Analysis

- Fine-tune can extremely improve the accuracy of pre-trained deep neural network.
- Data augmentation can benefit the performance of ResNet18 with fine-tune, but not much. ResNet18 on original data performs better that I have thought, maybe the imbalance of this dataset(4:1) is not that horrible to be the dominant condition.
- The experiment of anomaly detection does find some outliers, but "augmentation after deleting outliers" has lower accuracy than "augmentation only" one. I think deleting outliers may lead to lose some information and the anomaly detection process can also introduce errors.

Conclusion and Future work

This work uses **pre-trained ResNet18** to classify semiconductors into good class and defect class. I perform experiments on **original data**, **augmented data** and **outlier detected data** to find the effect of imbalance and outliers on the classification. It turns out balanced data(after augment the defect images) has better performance with best AUC score of 97.64%. While removing outliers before getting balanced can also improve the accuracy, the limited improvement shows that anomaly detection can also bring errors and is still a big challenge. Future work can focus on more reliable anomaly detection techniques.

The End