OUTLIER DETECTION FROM IMAGE DATA REPRODUCIBILITY CHALLENGE

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper is mainly about a novel image outlier detection (IOD) method that based on Deep Neural Decision Forest (DNDF). While there are several exsting method for IOD, most of them are generic work such GAN, and plug them into the classical IOD method. Although these method may find some common feature of normal data, they perform poorly for the detect ”outliers” from ”inliers”. By using the DNDF method, we further modify the maximum probability provided by CNN into confidence of image. Then, setting a number K that denote how much outliers image are there in the dataset. Finally, we can get a confidence threshold (ct), by using the ct we can detect the rest outlier from the testing set.

1 INTRODUCTION

Our project based the conference paper Outlier Detection From Image Data. In this paper, the authors mentioned about while convolutional neural network (CNN) have very high accuracy in image classification, it not work well when there are outliers existing in the image dataset. Due to the generalization performance jeopardized the image outlier detection (IOD) ability of CNN, in this paper they mentioned about a method called Deep Neural Decision Forest (DNDF), which is a method combined the CNN and decision tree.

1.1 STATE-OF-THE-ART

Nowadays the outlier detection method with deep learning are usually generic network (such as a deep autoencoder or GAN) plugged in the classical outlier detection methods in stead of a deep learning method that specifically for the purpose. Cao et al. (2019)

1.2 TARGET

Our target is try to re-produce the IOD method, which is using DNDF, a method combined CNN and decision tree, to classify image and detect the outliers accurately.

2 METHODOLOGY

2.1 CONFIDENCE-BASED OUTLIER DETECTION

The confidence-based outlier detection is basically based on CNN. In CNN, the final Fully Connected (FC) layer computes a weighted sum score s_i, then the weighted sum score will be the input for the softmax activation function and generate a class probability p_i. The weighted sum score and the generated probability is vital due to it will determine whether which class the image will belong. Every images will get a probability that is the highest probability among all probability. Therefore, the confidence measure in of the confidence-based outlier detection would be the maximum weighted sum score or he maximum probability. Once we get the confidence, we sort them from small to large and select the k^{th} smallest confidence as the cutoff threshold (ct). k is intuitive parameter, it depends on how much outliers we think are in the dataset. In our project, we try to use the MNIST and CIFAR-10 dataset and will mix some image into another, therefore, if we insert 5000 CIFAR-10 image into the MNIST dataset, then the k would be 5000.

2.2 Deep Neural Decision Forest-Based Approach

While the confidence-based outlier detection feels make sense, however, due to the generalization performance of CNN, the result usually won't be good. Hence, the DNDF method based on CNN, but the final FC layer connect to the decision nodes of the decision tree.

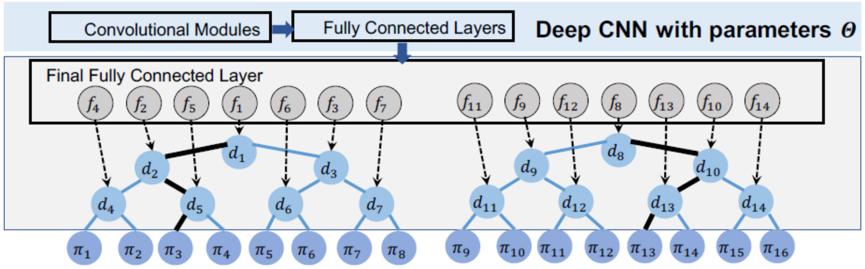

Convolutional Modules
Fully Connected Layers
Deep CNN with parameters θ
Final Fully Connected Layer
f_4 f_2 f_5 f_1 f_6 f_3 f_7 f_{11} f_9 f_{12} f_8 f_{13} f_{10} f_{14}
d_1 d_2 d_3 d_4 d_5 d_6 d_7 d_8 d_9 d_{10} d_{11} d_{12} d_{13} d_{14}
π_1 π_2 π_3 π_4 π_5 π_6 π_7 π_8 π_9 π_{10} π_{11} π_{12} π_{13} π_{14} π_{15} π_{16}

Figure 1: Deep Neural Decision Forest

$$d_n(x;\Theta) = \sigma(f_n(x;\Theta)) \tag{1}$$

$$\mu_l(x|\Theta) = \prod_{n\in N} d_n(x;\Theta)^{\mathbf{1}_{l\leftarrow n}} \bar{d}_n(x;\Theta)^{\mathbf{1}_{l\rightarrow n}} \tag{2}$$

Formular 1 is the routing decision probability. Formular 2 is how we calculate the max route based confidence.

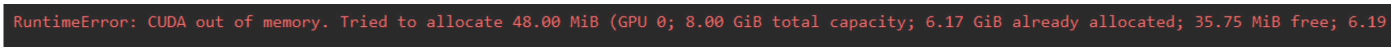
3 Experiment Setup

Our setup based a DNDF we find on GitHub. The original function of the DNDF is classification, we further modify it output become confidence level, thus we can use it to detect outliers.

For CNN method, we direct using the maximum probability provided by the model. For DNDF method, we using max route based confidence.

4 Results

We were try to merge dataset such as CIFAR-10 and CIFAR-100 so we can further detect the outliers, however our program suffer from the error below:



```
RuntimeError: CUDA out of memory. Tried to allocate 48.00 MiB (GPU 0; 8.00 GiB total capacity; 6.17 GiB already allocated; 35.75 MiB free; 6.19
```

Figure 2: Error we did not fixed

Table 1: Confidence threshold comparison between CNN and DNDF on MNIST

PART	**DESCRIPTION**
CNN	DNDF
1.4193049×10^{-4}	0.9998

While we try to fix by changing the batch size, the error is still there, so we change our plan in to run MNIST dataset on DNDF and CNN, then compare the confidence threshold and see is there are large difference between the two models.

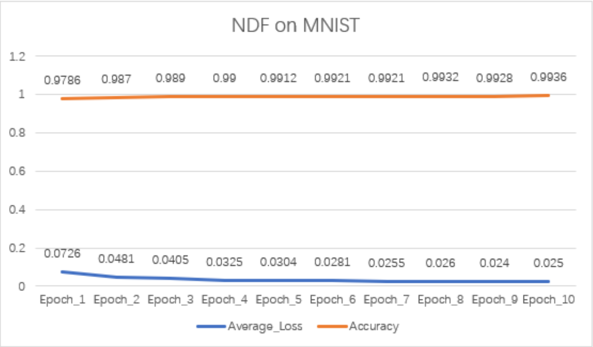


Figure 3: Result of DNDF

As shown in the above figure, we operate the DNDF model on MNIST dataset, and the model setting is that number of trees is 12, depth of tree is 8. It is known that with more trees, the better accuracy will be obtained, but GPU has reached its limitation. After running the forest, the sorted max route based confidence is shown as a list, [0.2651, 0.3131, 0.3263, ……, 1.0000, 1.0000, 1.0000] The total length of this list is 60000, which corresponds to the number of images in the MNIST training dataset. Each number represents that how confident the classifier to each image. To detect outliers, we need to set a threshold, however, we do not need to directly define a threshold but define a more intuitive parameter k. It means that training image x_k has the k^{th} smallest confidence among all objects in the training dataset. In this case, we set k = 5000, and calculated threshold value is 0.9998, which means that when using this trained model on other dataset, images in this dataset whose max route based confidence is less than 0.9998 will be considered outliers.

We also runs the MNIST dataset on CNN, and the confidence threshold at k=5000 is 1.4193049×10^{-4}.

5 Conclusion

In conclusion, in this project we try to reproduce a IOD method. However, when we training the CIFAR-10 and CIFAR-100 we met some error, and cannot be fixed. Therefore, we changed our direction to run CNN and DNDF separately and see is there any difference between the confidence threshold. And from the result we can see that the ct provided by CNN and DNDF are significant different. So we think that even we did not completely reproduce the method, but since the ct existing large difference, the method may be able to detect image outliers.

REFERENCES

Lei Cao, Yizhou Yan, Samuel Madden, and Elke Rundensteiner. Outlier detection from image data, 2019. URL https://openreview.net/forum?id=HygTE309t7.