# Repairing Confusion and Bias Errors for DNN-Based Image Classifiers

Yuchi Tian
Columbia University
United States
yuchi.tian@columbia.edu

## ABSTRACT

Recent works in DNN testing show that DNN based image classifiers are susceptible to confusion and bias errors. A DNN model, even robust trained model can be highly confused between certain pair of objects or highly bias towards some object than others. In this paper, we propose a differentiable distance metric, which is highly correlated with confusion errors. We propose a repairing approach by increasing the distance between two classes during retraining the model to reduce the confusion errors. We evaluate our approaches on both single-label and multi-label classification models and datasets. Our results show that our approach effectively reduce confusion errors with very slight accuracy reduce.

## CCS CONCEPTS

• **Software and its engineering** → **Software testing and debugging**; • **Computing methodologies** → *Neural networks.*

## KEYWORDS

DNNs, image classifiers, confusion, bias

**ACM Reference Format:**
Yuchi Tian. 2020. Repairing Confusion and Bias Errors for DNN-Based Image Classifiers. In *Proceedings of the 28th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '20), November 8–13, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3368089.3418776

## 1 PROBLEM AND MOTIVATION

Recent works in DNN testing show that DNN based image classifiers are susceptible to confusion and bias errors[1, 5, 6, 8]. For example, a state-of-the-art CIFAR-10 model[11] is most confused between *dog* and *cat*; models[12] trained in COCO Gender dataset are more likely to predict woman for an indoor activity and man for outdoor sports. Confusions result from two objects' frequent appearing together in the training data while bias result from that an object appears more frequently with one object in the training data than the other object. NAPVD (Neuron Activation Probability Vector Distance) is proposed to identify confusions and bias errors for DNN based image classifiers[8]. As a logical next step, we work on repairing confusion errors and bias errors. We first propose a differentiable distance metric correlating with confusion errors and a distance difference metric correlating with bias errors based on NAPVD and then propose including these metrics into loss functions to repair confusion and bias errors when re-training. A related research problem is adversarial training for increasing general adversarially robust accuracy[2, 7, 9, 10]. Our work focuses on reducing confusions and bias for target pairs and triples.


*ESEC/FSE '20, November 8–13, 2020, Virtual Event, USA*

ACM ISBN 978-1-4503-7043-1/20/11…$15.00
https://doi.org/10.1145/3368089.3418776

## 2 BACKGROUND

### 2.1 Confusion/Bias Errors

Confusion and bias errors are proposed in DeepInspect[8].

#### 2.1.1 *Confusion Errors.*

$$\text{type1conf}(x, y) = \text{mean}(\text{P}(x|y), \text{P}(y|x))$$

The confusion error definition describes DNN's probability to misclassfy class y as x and vice-versa.

#### 2.1.2 *Bias Errors.*

**Confusion disparity** (cd) is used to indicate bias errors among three classes.

$$\text{cd}(x, y, z) = |confusion(x, z) - confusion(y, z)|,$$

This definition measure differences in confusion error rate between $x$ and $z$ and between $y$ and $z$. With presence of z, if the model more likely mis-classifies it to be x rather than y, then the model is bias toward x than y given z.

### 2.2 Non-differentiable NAPVD

Tian et. al[8] proposed NAPVD (Neuron Activation Probability Vector Distance) to measure the class level relation perceived by a pre-trained DNN model. Intuitively, if images in two different classes always activate same set of neurons for a model under test, then this model cannot distinguish between these two classes. Based on this intuition, they defined the perception of each class from a given model under test as follows,

$$P(C) = [A(n_1)/N, A(n_2)/N, ..., A(n_t)/N],$$

where N is number of images in class C, $A(n_i)$ means how many times the neuron $n_i$ is activated when feeding these N images to the model under test.

The distance between two classes is the L2 norm between these two classes' perception vectors.

$$NAPVD(x, y) = L2\_norm(P(x), P(y))$$

They showed that the smaller $NAPVD(x, y)$, the model will be more confused between x and y. The larger $NAPVD(x, z) - NAPVD(y, z)$, the model will be bias towards y than x in presence of z. They leveraged this correlation to detect confusion and bias errors.

## 3 APPROACH AND CONTRIBUTION

NAPVD distance metric cannot be used in loss function because $A(n_i)$ is not differentiable. Instead we proposed a differentiable distance metric based on NAPVD.

### 3.1 Differentiable Distance Metric

We proposed the following definition of perception of class C from a model under test.

$$P_{new}(C) = [S(n_1)/N, S(n_2)/N, ..., S(n_t)/N],$$

where $S(n_i)$ is the sum of each output of neuron $n_i$, given N input images. Then, the distance metric we proposed is as follows,

$$D_{new}(x, y) = L2\_norm(P_{new}(x), P_{new}(y))$$

In this definition, $P_{new}(C)$ and $D_{new}$ are differentiable.

### 3.2 Repairing Confusion and Bias Errors

We proposed the following two loss functions to reduce confusion errors and bias errors respectively,

$$Loss_{confusion} = Loss_{orig} - \lambda D_{new}(x, y)$$

$$Loss_{bias} = Loss_{orig} + \lambda\ abs(D_{new}(x, z) - D_{new}(y, z))$$

To optimize on $Loss_{confusion}$ is to reduce the original loss and at the same time increase the distance between class x and class y. Similarly, to optimize on $Loss_{bias}$ means to reduce the original loss and at the same time reducing the distance difference between (x, z) and (y, z).

## 4 PRELIMINARY RESULTS

We evaluate our approaches on a state-of-the-art ResNet-50 model trained using CutMix[11] in CIFAR-10[3] dataset and a pre-trained ResNet-50 model[12] in COCO[4] dataset.

### 4.1 CIFAR-10

We compute pairwise confusion errors on the model under test and find that *dog* and *cat* is the most confusing pair. After applying our repair, the *dog* and *cat* confusion is reduced by 14% (from 0.064 to 0.055) while the accuracy is only reduced from 0.9484 to 0.9417.

### 4.2 COCO

We find that the model under test is most confused between *bus* and *person*. Figure 1 shows that after applying our proposed loss function at epoch 30, the confusion between *bus* and *person* is reduced from 0.5146 to 0.1557 while the mean precision is increased from 0.6532 to 0.6560.

## 5 CONCLUSIONS AND FUTURE WORK

The preliminary results show that our work can effectively reduce confusion errors. The confusion reduction for CIFAR-10 model is not as significant as COCO model because CIFAR-10 is a easy task and its model under test is very accurate. In future work, we will include the evaluation of the performance in reducing bias errors.

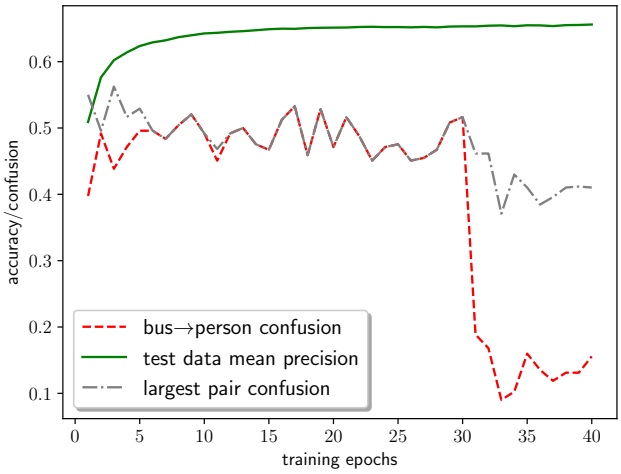


**Figure 1: Confusion error before and after repairing**

## ACKNOWLEDGMENTS

Special thanks to Baishakhi Ray, Vicente Ordonez and Ziyuan Zhong for useful discussions. This work is supported in part by NSF CCF-1845893, CNS-1842456, and CCF-1822965.

## REFERENCES

[1] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *FAT*.
[2] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. 2019. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning*. 1802–1811.
[3] Alex Krizhevsky. 2012. Learning Multiple Layers of Features from Tiny Images. *University of Toronto* (05 2012).
[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.
[5] Shiqing Ma, Yingqi Liu, Wen-Chuan Lee, Xiangyu Zhang, and Ananth Grama. 2018. MODE: Automated Neural Network Model Debugging via State Differential Analysis and Input Selection. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (Lake Buena Vista, FL, USA) *(ESEC/FSE 2018)*. ACM, New York, NY, USA, 175–186. https://doi.org/10.1145/3236024.3236082
[6] Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana. 2017. Deepxplore: Automated whitebox testing of deep learning systems. In *proceedings of the 26th Symposium on Operating Systems Principles*. 1–18.
[7] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. 2019. Adversarial training for free!. In *Advances in Neural Information Processing Systems*. 3358–3369.
[8] Yuchi Tian, Ziyuan Zhong, Vicente Ordonez, Gail Kaiser, and Baishakhi Ray. 2020. Testing DNN Image Classifier for Confusion & Bias Errors. In *International Conference of Software Engineering (ICSE)*.
[9] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2019. Robustness may be at odds with accuracy. In *International Conference on Learning Representations (ICLR)*.
[10] Shiqi Wang, Yizheng Chen, Ahmed Abdou, and Suman Jana. 2018. MixTrain: Scalable Training of Formally Robust Neural Networks. *CoRR* abs/1811.02625 (2018). arXiv:1811.02625 http://arxiv.org/abs/1811.02625
[11] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. In *International Conference on Computer Vision (ICCV)*.
[12] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2941–2951.