ROBUSTNESS, BIAS ASSESSMENT, AND BIAS REMOVAL IN NEURAL NETWORKS PREDICTIONS

AUTHORS

Ettore Merlo - ettore.merlo@polymtl.ca *Computer and Software Engineering Dept., Polytechnique Montreal*
Benjamin Djian - benjamin.djian@polymtl.ca *Computer and Software Engineering Dept., Polytechnique Montreal*
Sébastien Gambs - gambs.sebastien@uqam.ca *Computer Science Dept., UQAM*

INTRODUCTION

Deep Neural Network models infer characteristics from input data to predict output classes.

Despite machine learning impressive performance in various domains, the incorporation of **personal data** within training sets, in some contexts, may reflect historically biased human decisions or social values. It represents a risk of **unintended biases** and **discrimination against demographic groups** characterized by **ethically sensitive attributes**, such as gender, race, age, and so on.

Such biases must be assessed and removed for fair decision-making.

OBJECTIVES

We present an original approach based on **Computational Profile Likelihood (CPL)** [1], [2] to assess potential bias in neural network decisions according to some ethically sensitive attributes and to remove such a bias.

Bias removal can be performed by **post-processing neural network decisions** toward the desired ethical outcome profiles of sensitive attributes.

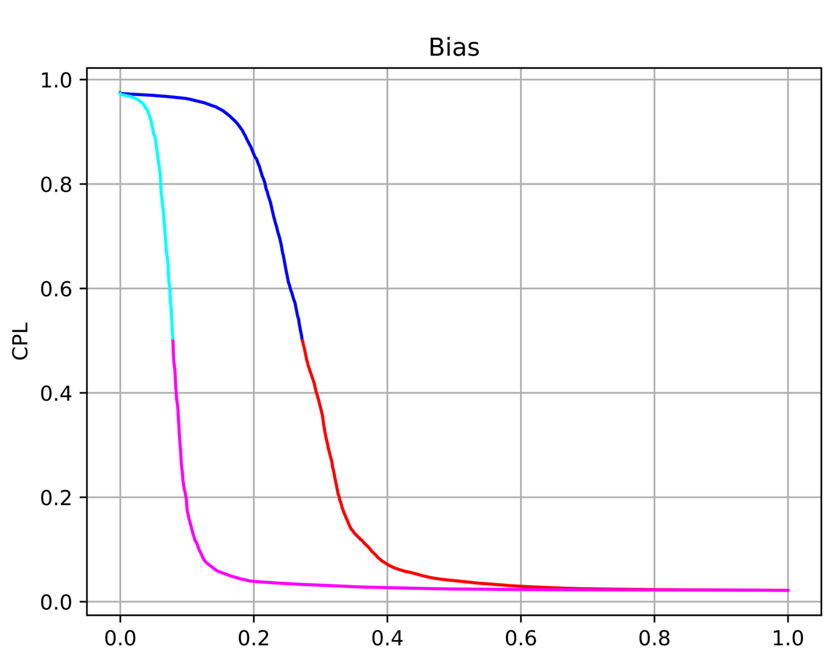

Figure 1
Initial predictions

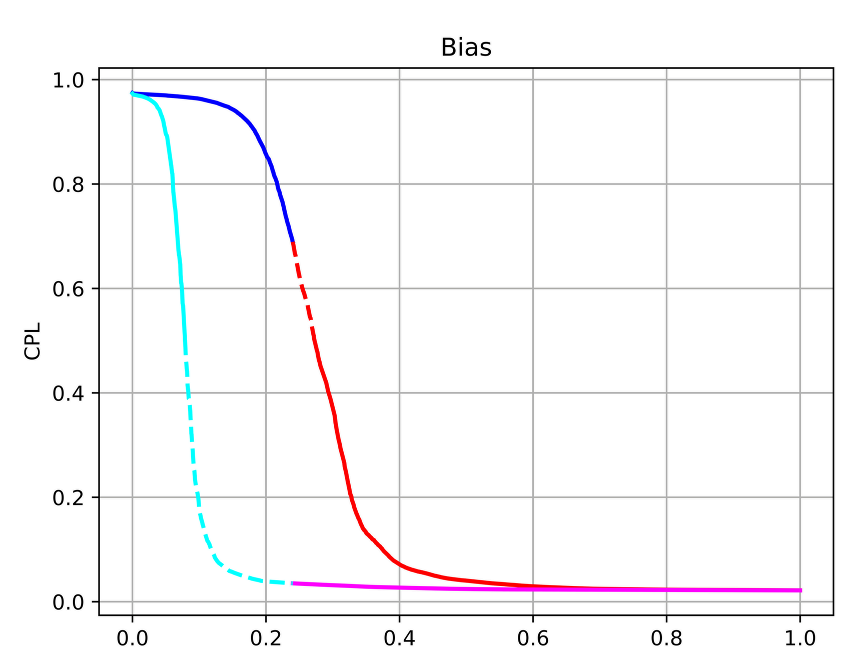

Figure 2
Bias removal

METHOD

Initially designed for Out-Of-Distribution (OOD) detection, **Computational Profile Likelihood (CPL)** estimates the conditional probability of a network internal neuron excitation levels during predictions.
CPL distributions observed during training are compared in contrast with those observed when processing new inputs.

Experiments have been performed using the dataset **Adult Census Income** [3], extensively studied in the context of fairness [4]. This dataset has 45, 222 instances of census information of people. The task of prediction concerns the "income" feature and the two output classes are "**High Revenues (HR)**" and "**Low Revenues (LR)**".
We consider the "sex" feature as the sensitive attribute, with two possible values: "**Women**" and "**Men**".

We used CPL to **assess and remove gender bias** in neural network predictions trained on the dataset Adult Census Income. CPL results for bias removal have been compared to those obtained from the "Reject Option Classification" (ROC) [5].

ACKNOWLEDGEMENTS

The authors wish to thank the industrial and academic partners of the "DEpendable & Explainable Learning" (DEEL) project, CRIAQ, and the National Science and Engineering Research Council of Canada (NSERC) for funding the project CRDPJ 537462-18.

RESULTS AND DISCUSSION

Figure 1: The CPL analysis revealed a **negative bias against high-income (HR) women**, with approximately 8% of women initially predicted as HR (cyan line) with respect to LR women (magenta line). In contrast, there is a **favorable bias towards high-income (HR) men**, with around 30% of men predicted as HR (blue line) in contrast to LR men (red line).

Figure 2: The CPL approach can **completely remove gender bias**. This is achieved by **increasing the number of HR women** (dashed cyan line) and decreasing that of LR women (magenta line), while at the same time **decreasing the number of HR men** (blue line) and increasing that of LR men (dashed red line). An **equal target ratio of 25%** HR women (cyan line) and HR men (blue line) is obtained after bias removal.

Both CPL and ROC methods perfectly reach decision correction up to an equal and sought target ratio of 25%. CPL increases network classification precision by filtering out less reliable predictions [1, 2]. Therefore, **CPL-based bias assessment and removal are more robust against noise or input anomalies** than ROC corrections. **CPL corrections also preserve the relative ranking** of sorted CPL.

Common corrections between CPL and ROC can be considered as **highly reliable** and represent about **55% of corrections**.

CONCLUSIONS

CPL has been proven effective in the identification of OOD computations. This method is robust against unusual input cases and natural anomalies by filtering out 70% to 90% of misclassifications of adversarial and unusual inputs.

We have used CPL to evaluate and assess gender bias between Male and Female individuals in the "Adult Census Income" database. Results show a negative bias against high-income women - about 8% only of women were originally predicted at high income - and a corresponding positive bias towards high-income men - about 30% of men were originally predicted at high income.

Bias can be totally removed from model predictions using CPL. After bias removal, an equal target ratio of 25% high-income Males and Females is reached.
We also compared CPL and ROC approaches and found that these methods share about 55% of common highly reliable bias corrections.
removed from model predictions using CPL. We also compared CPL and ROC approaches and found that these methods share about 55% of common highly reliable bias corrections.

Although preliminary results are promising, additional studies are necessary to further investigate and generalize these findings.

BIBLIOGRAPHY

[1] M. Marhaba, E. Merlo, F. Khomh, and G. Antoniol, "Identification of out-of-distribution cases of cnn using class-based surprise adequacy," in 1st International Conference on AI Engineering - Software Engineering for AI (CAIN), Pittsburgh, PA, USA. IEEE/ACM, 2022.
[2] E. Merlo, M. Marhaba, F. Khomh, H. B. Braiek, and G. Antoniol, "Models of computational profiles to study the likelihood of DNN metamorphic test cases," CoRR, vol. abs/2107.13491, 2021. [Online]. Available: https://arxiv.org/abs/2107.13491
[3] Becker,Barry and Kohavi,Ronny. (1996). Adult. UCI Machine Learning Repository. https://doi.org/10.24432/C5XW20.
[4] S. Gambs and R. C. Ngueveu, "Fair mapping," CoRR, 2023. [Online]. Available: https://arxiv.org/abs/2209.00617
[5] F. Kamiran, A. Karim and X. Zhang, "Decision Theory for Discrimination-Aware Classification," 2012 IEEE 12th International Conference on Data Mining, Brussels, Belgium, 2012, pp. 924-929, doi: 10.1109/ICDM.2012.45.