Caterina Urban
INRIA — Team ANTIQUE
École Normale Supérieure
45 rue d'Ulm, 75005 Paris
caterina.urban@inria.fr

Algorithmic Fairness Static Analysis for Neural Networks

M2 Research Internship Proposal, 2019-2020

Context. Nowadays, we are witnessing widespread adoption of software with far-reaching societal impact, i.e., software that assists or automates decision-making in fields such as social welfare, criminal justice, and even health care. It is not difficult to envision that in the future most of the decisions in society will be delegated to software.

However, a number of recent cases have evidenced the importance of ensuring software *fairness*[1] as well as data privacy[2]. Going forward, data science software will be subject to more and more legal regulations (e.g., the European General Data Protection Regulation adopted in 2016) as well as administrative audits.

Goals and Objectives. To meet these needs, the host team is developing a static analysis for proving *causal fairness* [3] of feed-forward multi-layer neural networks. The goal of this internship is to extend this static analysis. We envision multiple directions that can be explored:

1. The analysis is currently tailored to neural networks using ReLU activation functions. It would be interesting to generalize it to support other commonly used activation functions (e.g., *sigmoid* or *tanh*).

2. The analysis currently employs existing numerical abstract domains [1, 2, 4] which however have precision or scalability limitations. To solve this problem we aim to investigate the design of *new abstract domains* specifically adapted to the analysis of neural networks.

3. A further and more challenging direction would be to re-design the analysis to work on a *per-layer* basis in order to further improve its scalability.

In addition to addressing (some or all of) the theoretical questions mentioned above, the internship will also include the implementation of the newly designed static analyses and the experimental evaluation of their practical usefulness on neural network trained from real datasets. The implementation will be integrated into the Libra static analyzer, developed by the host team in Python. If time permits and the intern is interested, the internship could also be extended to investigate approaches for *repairing bias*.

[1] https://www.nytimes.com/2017/10/26/opinion/algorithm-compas-sentencing-bias.html

[2] https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

Prerequisites. The internship requires a background in program analysis and abstract interpretation. Knowledge of the PYTHON programming language is also required. Familiarity with neural networks is a plus, but can also be acquired during the internship.

Practical Information. The internship will take place in the INRIA research team ANTIQUE, hosted at École Normale Supérieure, Paris. A successful internship may provide opportunities for a funded PhD on a follow-up subject.

References

[1] Patrick Cousot and Radhia Cousot. Static Determination of Dynamic Properties of Programs. In *Second International Symposium on Programming*, pages 106–130, 1976.

[2] Patrick Cousot and Nicolas Halbwachs. Automatic Discovery of Linear Restraints Among Variables of a Program. In *POPL*, pages 84–96, 1978.

[3] Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual Fairness. In *NIPS*, pages 4069–4079, 2017.

[4] Antoine Miné. Symbolic Methods to Enhance the Precision of Numerical Abstract Domains. In *VMCAI*, pages 348–363, 2006.