Certifying Differential Invariants of Neural Networks using Abstract Duals

CHANDRA KANTH NAGESH, University of Colorado Boulder, USA

Neural networks are increasingly used in safety-critical domains, necessitating formal guarantees about their properties under perturbations to inputs. Existing robust verification techniques, typified by `DeepPoly`, primarily focus on the robustness analysis during forward mode learning where certifying output stability is typified by $f(\mathbf{I}) \subseteq [\mathbf{y}_L, \mathbf{y}_R]$ for an $\mathbf{L}_\infty$-constrained input set $\mathbf{I} = \{\mathbf{x} \| \mathbf{x} - \mathbf{x}_0\|_\infty \leq \epsilon\}$. While such methods employ a sophisticated polyhedral abstract domain (combining intervals and affine forms) to generate sound, tight bounds, this entire class of analysis still fails to provide sound guarantees over the behaviour of the network. This oversight creates a critical verification gap for gradient-dependent systems where the Backward Pass Verification Problem which can be described formally as bounding the Jacobian $\mathbf{J}(\mathbf{x}) = \nabla_\mathbf{x} f(\mathbf{x})$ such that $\mathbf{J}(\mathbf{x}) \subseteq [\mathbf{J}_L, \mathbf{J}_R]$ for all $\mathbf{x} \in \mathbf{I}$ is essential.

1 Introduction

`DeepPoly`[1]

2 Overview

3 [Contribution 1]

4 [Contribution 2]

5 Evaluation

6 Related Work

7 Conclusion

Acknowledgments

TBD

References

[1] Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. 2019. An abstract domain for certifying neural networks, Vol. 3. 1–30. doi:10.1145/3290354

Author's Contact Information: Chandra Kanth Nagesh, ckn@colorado.edu, University of Colorado Boulder, Boulder, Colorado, USA.