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 robustness analysis during forward mode where certifying output stability is given 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. 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`[3] provides us with a clear method to certify output bounds for a given neural network. Let us consider the situation where we are given a trained neural network $f(\mathbf{I})$ on some set $\mathbf{I}$. Then, under some input $x_0 \in \mathbf{I}$ and perturbation ϵ, we need to verify that the output of f still satisfies the same result i.e. we want to verify that the neural network is invariant to a ϵ perturbation of the input. This can formally be represented as:

$$\forall x \in \mathbb{B}_\infty(x_0, \epsilon): \quad \text{argmax}(f(x)) = \text{argmax}(f(x_0))$$

where, $\mathbb{B}_\infty$ is a ball of ϵ radius around the input provided by its $\mathbf{L}_\infty$ norm. This is critical in a lot of scenarios as with many supervised learning problems, it is infeasible to verify the output of a neural network against all possible test sets. We want to formally prove that the model is invariant to these input changes. [1, 2, 4] were some of the earliest works in this domain.

1.1 Definition of the Abstract Dual Domain

We define the Abstract Dual domain $\hat{\mathcal{D}}$ as a product space of two affine forms representing the range of values and the range of gradients across a set.

Definition 1.1. An **Abstract Dual Number** $\mathcal{X} \in \hat{\mathcal{D}}$ is a pair:

$$\mathcal{X} = \langle \hat{x}_{val}, \hat{x}_{grad} \rangle$$

where $\hat{x}_{val}$ is the affine form of the neuron values and $\hat{x}_{grad}$ is the affine form of the partial derivatives with respect to the input.

1.2 Propagation Rules

To propagate $\mathcal{X}$ through a neural network, we define abstract transformers for each layer type. 3 **Linear Layers:** For a weight matrix W and bias b, the transformation is exactly determined by the linearity of the dual algebra:

$$\mathcal{Y} = \langle W\hat{x}_{val} + b, W\hat{x}_{grad} \rangle$$

Non-linear Activations: For a smooth activation σ, we apply a linear relaxation. Let $[l, u]$ be the interval range of $\hat{x}_{val}$. We bound the derivative σ' over this interval:

$$\hat{y}_{grad} = [\inf_{z \in [l,u]} \sigma'(z), \sup_{z \in [l,u]} \sigma'(z)] \cdot \hat{x}_{grad}$$

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

2 Overview

3 Contributions

3.1 Theoretical Results: Soundness and Instability

4 Formal Proof of Soundness

In this section, we establish the mathematical validity of the Abstract Dual domain. We demonstrate that the interval of gradients produced by our forward-mode propagation soundly over-approximates the true range of partial derivatives of the neural network f over the input region X_0.

4.1 Foundational Definitions

Let $f : \mathcal{R}^n \to \mathcal{R}^k$ be a neural network. Let $X_0 \subseteq \mathcal{R}^n$ be a centrally symmetric input region represented by the affine form $\hat{x}_{val}$. We define the **Concrete Derivative Set** as follows:

$$\mathcal{D}(f, X_0) = \{\nabla f(x) \in \mathcal{R}^{k \times n} \mid x \in X_0\} \tag{1}$$

Definition 4.1 (Soundness of Abstract Duals). An Abstract Dual Number $\mathcal{X} = \langle \hat{x}_{val}, \hat{x}_{grad} \rangle$ is considered **sound** with respect to a function f and input region X_0 if and only if:

(1) $\forall x \in X_0 : f(x) \in \gamma(\hat{x}_{val})$
(2) $\forall x \in X_0 : \nabla f(x) \in \gamma(\hat{x}_{grad})$

where $\gamma(\hat{x})$ denotes the concretization function mapping an affine form to its corresponding subset of $\mathcal{R}$.

4.2 Linear Layer Soundness

Lemma 4.2 (Soundness of Linear Transformers). *Given a sound abstract dual $\mathcal{X}$ and a linear transformation $y = Wx + b$, the abstract transformer $\mathcal{Y} = W\mathcal{X} + b$ is sound.*

Proof. By the fundamental rules of multi-variable calculus, if y is a linear function of x defined by $y = Wx + b$, the Jacobian matrix is constant: $\nabla_x y = W$. By the chain rule, for a composite function $y(f(x))$, we have:

$$\nabla y = W \cdot \nabla f(x) \tag{2}$$

In our abstract domain, we define:

$$\hat{y}_{val} = W\hat{x}_{val} + b, \quad \hat{y}_{grad} = W\hat{x}_{grad} \tag{3}$$

Since affine arithmetic is a linear abstraction, it is exact for linear transformations (i.e., $\gamma(W\hat{x} + b) = \{Wx + b \mid x \in \gamma(\hat{x})\}$). Given that $\mathcal{X}$ is sound, $\gamma(\hat{x}_{grad})$ contains all possible values of $\nabla f(x)$. Therefore, $\gamma(W\hat{x}_{grad})$ must contain $W \cdot \nabla f(x)$. This concludes that $\mathcal{Y}$ is sound. □

4.3 Activation Function Soundness

Lemma 4.3 (Soundness of Non-linear Activation Transformers). *Let σ be a differentiable activation function (e.g., Sigmoid, Tanh). The abstract dual transformer $\sigma^{\#}(\mathcal{X})$ defined by interval derivative bounding is sound.*

Proof. Consider the composite function $h(x) = \sigma(f(x))$. By the univariate chain rule applied element-wise:

$$\nabla h(x) = \sigma'(f(x)) \cdot \nabla f(x) \tag{4}$$

Let $[l, u]$ be the interval range of the values $\hat{x}_{val}$. We compute the derivative range Σ' as:

$$\Sigma' = \left[\inf_{z \in [l,u]} \sigma'(z), \sup_{z \in [l,u]} \sigma'(z) \right] \tag{5}$$

The abstract gradient is then computed as the product $\hat{h}_{grad} = \Sigma' \cdot \hat{x}_{grad}$. By the Mean Value Theorem, for any $x \in X_0$, the concrete value $f(x)$ must lie within $[l, u]$. Consequently, the concrete derivative $\sigma'(f(x))$ must be contained within the interval Σ'. Using the soundness property of interval-affine multiplication, the resulting affine form $\hat{h}_{grad}$ is guaranteed to enclose the product of any value in Σ' and any vector in $\gamma(\hat{x}_{grad})$. Thus, $\forall x \in X_0 : \nabla h(x) \in \gamma(\hat{h}_{grad})$. □

4.4 Main Theorem: Lipschitz Certification

Theorem 4.4 (Global Lipschitz Soundness). *The Lipschitz constant K_{comp} derived from the output Abstract Dual $\mathcal{Y}_{out}$ is a sound upper bound for the function f over the region X_0.*

Proof. By induction over the network depth L, using Lemmas 1 and 2, the final gradient affine form $\hat{y}_{grad}$ at the output layer is a sound over-approximation of the set $\mathcal{D}(f, X_0)$. The global Lipschitz constant for f with respect to the L_∞ norm is defined as:

$$K = \sup_{x \in X_0} \|\nabla f(x)\|_1 \tag{6}$$

For any affine form $\hat{g} = \alpha_0 + \sum_{i=1}^{n} \alpha_i \epsilon_i$, the supremum of its absolute value is soundly bounded by:

$$\sup |\gamma(\hat{g})| \leq |\alpha_0| + \sum_{i=1}^{n} |\alpha_i| \tag{7}$$

We define K_{comp} as the sum of these absolute coefficients for the output gradient affine form. It follows that $K_{comp} \geq K$. By the Mean Value Theorem:

$$\forall x \in X_0 : \|f(x) - f(x_0)\|_\infty \leq K_{comp} \cdot \|x - x_0\|_\infty \tag{8}$$

Thus, if the safety condition $f(x_0)_{target} - f(x_0)_{other} > K_{comp} \cdot \epsilon$ is satisfied, the classification is guaranteed to be invariant within the ϵ-ball. □

Acknowledgments

TBD

References

[1] Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin T. Vechev. 2018. AI2: Safety and Robustness Certification of Neural Networks with Abstract Interpretation. *2018 IEEE Symposium on Security and Privacy (SP)* (2018), 3–18. https://api.semanticscholar.org/CorpusID:206579396

[2] Gagandeep Singh, Timon Gehr, Matthew Mirman, Markus Püschel, and Martin Vechev. 2018. Fast and Effective Robustness Certification. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/f2f446980d8e971ef3da97af089481c3-Paper.pdf

[3] 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

[4] Tsui-Wei Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Duane S. Boning, Inderjit S. Dhillon, and Luca Daniel. 2018. Towards Fast Computation of Certified Robustness for ReLU Networks. *ArXiv* abs/1804.09699 (2018). https://api.semanticscholar.org/CorpusID:13750928