

Neural Network Verification is a Programming Language Challenge

Lucas C. Cordeiro¹, Matthew L. Daggitt², Julien Girard-Satabin³, Omri Isac⁴,
Taylor T. Johnson⁵, Guy Katz⁴, Ekaterina Komendantskaya^{6,7}, Augustin
Lemesle³, Edoardo Manino¹, Artjoms Šinkarovs⁶, and Haoze Wu⁸

¹ University of Manchester, UK

² University of Western Australia, Australia

³ Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France

⁴ Hebrew University of Jerusalem, Israel

⁵ Vanderbilt University, USA

⁶ Southampton University, UK

⁷ Heriot-Watt University, UK

⁸ Amherst College, USA

Abstract. Neural network verification is a new and rapidly developing field of research. So far, the main priority has been establishing efficient verification algorithms and tools, while proper support from the programming language perspective has been considered secondary or unimportant. Yet, there is mounting evidence that insights from the programming language community may make a difference in the future development of this domain. In this paper, we formulate neural network verification challenges as programming language challenges and suggest possible future solutions.

Keywords: Neural Networks · Verification · Domain Specific Languages.

1 Introduction

Traditionally, statistical machine learning has distinguished its methods from “algorithm-driven” programming: the consensus has been that machine learning is deployed when there is example input-output data but no general algorithm for computing outputs from inputs. Thus, neural networks are commonly seen as programs that emerge from data via training, without direct human guidance on how to perform the computation. This unfortunate dichotomy has led to a divide between programming language and machine learning research that is still awaiting resolution.

The first hint that this dichotomy is not as fundamental as was thought came from the machine learning community itself. The famous paper by Szegedy et al. [113] pointed out the “intriguing” problem that even the most accurate neural networks fail to satisfy the property of *robustness*, i.e. small perturbations of their inputs should result in small changes to their output. Szegedy’s key example concerned imperceptible perturbations of pixels in an image that can

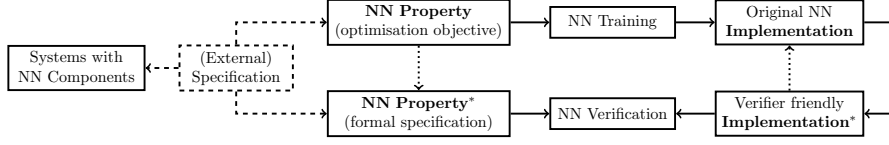


Fig. 1: Schematic representation of the state of the art in training and verifying neural networks for properties. Solid lines denote methods widely accepted by the research communities, dashed lines mean “some experimental prototypes exist”, dotted arrows mean the connection is desired but not established.

sway the neural network’s classification decisions. This lack of robustness can have safety and security implications: for example, an autonomous car’s vision unit may fail to recognise pedestrians on the road. For that reason, the problem attracted significant attention [25] but remains unresolved to this day. Partial solutions often deploy methods of *adversarial training* — i.e., training based on computing *adversarial attacks* — which augment the training set with the worst-case perturbations of the input data points with respect to the output loss of the neural network [78].

The robustness of neural networks actually yields a formal specification [26]. Given a neural network $f : \mathbb{R}^m \rightarrow \mathbb{R}^n$, f is *robust around* $\hat{x} \in \mathbb{R}^m$, if

$$\forall x, \|\hat{x} - x\| \leq \epsilon \implies \|f(\hat{x}) - f(x)\| \leq \delta, \quad (1)$$

where $\epsilon, \delta \in \mathbb{R}$ are small constants and $\|\cdot\|$ computes a vector distance. From the programming language perspective, robustness can be seen as a refinement type that refines input and output types of f , cf. [77]. At the same time, robustness is an example of a *desirable property* that neural networks cannot learn from data alone: note the quantification over vectors x that do not belong to the data set. This challenges the classical dichotomy between algorithm-driven and data-driven programming, demonstrating the inevitability of property specification in both cases.

Against this background, both the machine learning and verification communities proposed several useful methods of training for, or respectively verifying, *certain properties*⁹. Fig. 1 depicts these two groups of methods as two parallel pipelines. At the top, we include all adversarial training methods [78] that were generalised to account for arbitrary optimisation objectives, given a property informally expressed in (a fragment of) first-order logic [46, 52]. At the bottom, we include the verification pipeline which is supported by more than a dozen neural network verifiers, such as Marabou [75, 121], $\alpha\beta$ -CROWN [120], PyRAT [81], to name but a few. Unlike the machine learning approaches, it features a formal language for property specification, VNN-LIB. Furthermore, an annual competition VNN-COMP develops common standards for this domain [20, 19].

⁹ We deliberately use the term “properties” rather than “specifications” here, as the latter means the presence of a sufficiently general specification language.

However, there are several fundamental problems that prevent these emerging ideas from developing to full fruition. Firstly, both the machine learning and verification communities assume that *in theory* a neural network can be optimised for the desirable verification property. However, without any programming language support to ensure this formally, discrepancies between machine learning objectives and verification objectives have been found in the literature, even for simple robustness properties [26]. In Fig. 1, this problem is depicted by distinguishing the two versions of **NN Property** and **NN Property*** and a dotted line between them. The desirable solution is to have a single language with the relevant specification, which is then compiled down to either verification or machine learning backends.

Similarly, discrepancies have been reported between different representations of neural networks [71], e.g., using real numbers in verification and floating point numbers in training. In Fig. 1, this problem is depicted by showing two potentially disagreeing implementations, **Implementation** and **Implementation***. Ideally, we should be able to verify the actual programs, and not their idealised descriptions. Or, as an equally acceptable alternative, the solid arrow between two implementations in Fig. 1 should be reversed in the other direction – ensure that the guarantees concerning the verified neural networks extend to their actual implementations, thus establishing the connection along the bottom dotted arrow in Fig. 1.

Finally, neural networks are rarely implemented as stand-alone programs. More often, they are embedded into larger system development that, in turn, may have its own specification and verification regimes. Although the idea of a verified neural network controller is not itself new to the cyber-physical system research (cf. § 3.7), the programming language support for verification of such systems is a nascent field [116, 90].

In this light, we believe it is time to discuss how the verification and synthesis of safe neural networks fit together with general programming practices. In this “Fresh Perspectives” paper, we give an overview of the current state of the art in implementing neural network verification and explain the challenges the neural network verification community currently faces (Sec. 3). We do so by tracing different parts of the diagram in Fig. 1, and explaining the nature of the discrepancies in its different parts, from the programming language point of view. We wrap up this paper by suggesting possible ways the programming language community can help improve the state of the art (Sec. 4).

2 Neural Network Verification Properties

The problem of defining verification properties for neural networks has received substantial attention. Verification approaches started with neural networks deployed as controllers in autonomous systems [103, 74]. With time, they were generalised to cover data-dependent verification properties such as robustness [64, 43, 42, 123]. A set of standard benchmarks is revised and updated annually at the VNN-COMP; the competition reports [21, 18] provide a thorough overview of them.

Neural network verification properties can be divided into three categories.

1. **Geometric properties.** These properties are based on the geometry of the data manifold without any appeal to its possible semantic meaning. One such property is (local) *robustness*, whose definition is given in Equation 1 (see also the additional examples in [27]). Another related property is (local) *equivalence* [115], which constrains the output of two different networks to be similar under the same input, either in absolute value (ϵ -equivalence) or class prediction (top- k equivalence).
2. **Hyper-properties.** These properties require guarantees for any input, rather than just those close to the data manifold. Classic examples are global robustness [102] and global equivalence [83]. A more recent example of such properties is *confidence-based robustness* [8], which allows for some non-robust behaviour, but only for inputs close to the decision boundary. The latter complicates the specification and verification process in interesting ways (see Sec. 4).
3. **Domain-specific properties.** These properties are based on the presumed semantics of the data on which the neural network is trained. Usually, they take the form of admissible intervals on the input and output vector values.

The ACAS Xu challenge (the oldest neural network verification benchmark) best illustrates this third class of properties. It takes a neural network that models an aircraft controller: based on five input measurements between the own ship and an intruder (distance, angles, relative speeds), the neural network outputs one of five advisory actions (strong/weak left or right, clear of conflict).

When the benchmark was introduced in [74], nine properties were formulated by the engineers who designed the collision avoidance software. For instance, Property 3 states that *if the intruder is directly ahead and is moving towards the own ship, the network will not advise clear of conflict*. When written in the VNN-LIB query language [35] (see Sec. 3.1), the property is translated to real-valued intervals on the five input measurements and a constraint on the output prediction.

3 Neural Network Verification: State of the Art

In this section, we describe the state of the art in neural network verification, from the perspective of the existing programming language support, rather than the existing verification algorithms. For the latter, the tutorial [4] is available. We will proceed by tracing different arrows of Figure 1 and explaining the existing discrepancies and solutions.

3.1 Verification pipeline

Neural Network Verification Problem. Let us start with describing the common verification pipeline illustrated in Fig. 2. Given a trained neural network $f : \mathbb{R}^m \rightarrow \mathbb{R}^n$ and some network property Ξ , the *Neural Network Verification Problem* is the problem of deciding whether $\Xi(f)$ holds. Current verifiers

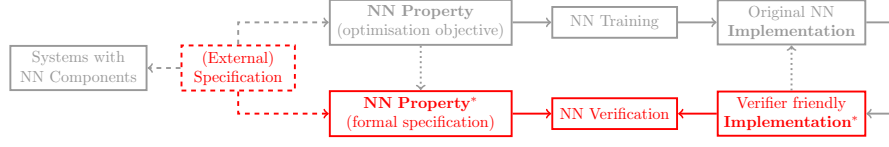


Fig. 2: Schematic representation of the neural network verification pipeline.

assume using a special format — ONNX (standing for *Open Neural Network Exchange*) [1] — to represent the neural networks. Thus, in reality, we verify $\Xi(f^*)$, where f^* is obtained from f by ONNX translation.

The verifiers typically consider properties defining a precondition on the network inputs and a postcondition on its outputs. Both conditions are most commonly linear (e.g., defined using linear bounds) and represent safe regions. Formally, let $\Xi := \langle P, Q \rangle$ where $P : \mathbb{R}^m \rightarrow \{\top, \perp\}$ and $Q : \mathbb{R}^n \rightarrow \{\top, \perp\}$. The neural network verification problem is then deciding whether $\forall x \in \mathbb{R}^m : P(x) \Rightarrow Q(f^*(x))$. Neural network verification algorithms then attempt to find a counterexample (i.e., $x \in \mathbb{R}^m$ such that $P(x) \wedge \neg Q(f^*(x))$) or conclude there is none. Several *neural network verifiers* are currently available to solve such verification problems: e.g. Marabou [75, 121], $\alpha\beta$ -CROWN [120], PyRAT [81], NNV [118, 86] and ERAN [111]. Since 2020, an annual International Verification of Neural Networks Competition (VNN-COMP) has been held, and has played an important role in consolidating the new research community and developing standards for this domain [20, 19].

Mainstream specification languages. Most neural network verifiers have a basic query language for representing individual queries. These formats are invariably simple enough so that the type-system is implicit rather than explicit and they possess no capability to abstract over definitions. The *de-facto* standard is the VNN-LIB query language [35] which is used in VNN-COMP [11]. The language is a subset of the QFLRA fragment of the SMT-LIB language, an S-expression based language widely used in the SMT verification community as a standard input for SMT provers [13]. The goal of VNN-LIB is to model first-order logic properties on the inputs and outputs of neural networks. Fig. 3 illustrates a snippet of robustness specification written in VNN-LIB. As can be seen, VNN-LIB specification itself does not explicitly talk about the functions f or f^* , rather it is assuming that the property will be used to verify the function f^* provided in a separate ONNX file. Thus, VNN-LIB and ONNX together serve as a specification for $\Xi(f^*)$.

From a programming language perspective, there are several issues with the VNN-LIB format as a language for expressing specifications.

1. **Lack of expressivity.** VNN-LIB and ONNX are simply not expressive enough to represent all the specifications users want to write. For example, the VNN-LIB and ONNX formats can only refer to a single neural network

```

(declare-const X_0 Real)      (assert (<= X_0 0.0))
(declare-const X_1 Real)      (assert (>= X_0 0.0))
...
(declare-const X_791 Real)    (assert (<= X_791 -58.231295852661134))
(declare-const Y_0 Real)      (assert (>= X_791 -75.58388969421387))
...
(declare-const Y_8 Real)      (assert (>= Y_5 Y_1))
(declare-const Y_9 Real)      (assert (>= Y_5 Y_3))

```

Fig. 3: Snippet of robustness specification in VNN-Lib for an image data set that has input of dimension 792 and 10 classes. The specification assumes an external definition of $f^* : \mathbb{R}^{792} \rightarrow \mathbb{R}^{10}$.

at a time, which makes encoding specifications where one needs to express properties on several neural networks at once impossible. Similarly, hyper-properties [8, 28] cannot be specified in VNN-LIB without special tooling, and neither can properties involving hidden neurons. Finally, VNN-LIB only supports satisfaction queries, meaning the specification writer must manually negate universal queries before being encoded.

2. **Lack of conciseness.** The lack of abstraction and the limitation that variables cannot represent multi-dimensional tensors means that more complex properties cannot be represented concisely. Consequently, the length of the queries tend to scale with the dimensions of inputs and outputs of the network, even when the property can be expressed concisely in mathematics in constant space. For example, the full specification in Fig. 3 that encodes the single line of Eq. 1 is a couple of thousand lines long.
3. **Lack of rigour.** VNN-LIB does not have a formally defined semantics, nor does it even formally define its own syntax. Consequently, it is difficult for users to check whether their specification in VNN-LIB is correct or compliant, and impossible to prove the soundness of tools that either consume or generate VNN-LIB. Furthermore, the ONNX format that VNN-LIB relies on, also lacks a formal semantics. For example, the ONNX documentation for the convolution operator¹⁰ has no proper mathematical specification for the semantics of the operator, describing it only with the single sentence “The convolution operator consumes an input tensor and a filter, and computes the output”. Other ONNX operator descriptions like those of Convolution, Maxpool, or Add (for broadcasting) refer to external sources like Numpy, PyTorch or Tensorflow for more implementation details.
4. **Lack of dynamic bindings to datasets.** Crucial to most attempts to specify “correctness” of a neural network is the notion of the *data manifold*, i.e., the distribution of inputs that the neural network will actually encounter during operation. Usually, the data manifold is only a small subset of the actual input space. By definition, the network should never encounter inputs that lie off the data-manifold during normal operation. If it does, there is no reason to require any particular behaviour from the network, and consequently, specifications should only quantify over inputs that lie on the

¹⁰ https://onnx.ai/onnx/operators/onnx__Conv.html, accessed 21-09-2024

manifold. The problem is that, in most cases, there is no precise mathematical definition of the data manifold. Therefore, the most common approach is for the specification to approximate the manifold as the union of “small” regions around each input in the training dataset. Unfortunately, the training datasets themselves are frequently huge, anywhere from thousands to hundreds of millions of items. Therefore, it is infeasible to directly express the dataset in the specification.

This lack of rigour of the underlying specification format has been recognised as a major problem. A recent effort in the ONNX community has led to the creation of a ONNX Safety-Related Profile working group¹¹ which aims to elaborate a dedicated ONNX profile for safety-related systems. While still embryonic, this working group might answer some of the issues highlighted above.

To work around the remaining problems, the natural solution is to allow users to represent their specifications in a higher-level specification language, connecting the neural network specification to the language of the larger system in which it is embedded. Moreover, the specification language must provide some mechanism for dynamically binding variables to existing datasets in standard formats used by machine learning practitioners.

3.2 Prototypes of New Specification Languages

In response to the outlined problems, two major attempts have been made to design more principled specification languages for neural network verification. We outline the essence of both, in turn. Fig. 4 provides code snippets for illustration.

1. **CAISAR.** The CAISAR platform [51] incorporates a higher-level specification language deriving from WhyML [45]. WhyML is a typed first-order language with pattern-matching, polymorphism, and a module system. On top of that, CAISAR provides additional types of linear algebra structures common in machine learning and compiles the specification back to plain WhyML. Writing a compiler from WhyML to VNN-LIB is straightforward, allowing CAISAR to target all state-of-the-art solvers from one single specification. It can also deal with specifications involving multiple neural networks and dynamically bind variables to concrete datasets. However, it can be argued that the composability of WhyML is limited, and the lack of dependent types prevents the modelling of important properties (for instance, encoding the dimension of inputs directly in their types could prevent common runtime errors).
2. **Vehicle.** The Vehicle specification language [33, 32, 31] is a higher-order and dependently-typed functional language. The language aims to be able to express a full range of specifications and to that end it contains quantifiers as first-class language constructs, conditionals and higher-order functions over

¹¹ <https://github.com/ericjenn/working-groups/blob/ericjenn-srpgw-wg1/safety-related-profile/README.md>

```

theory MNIST

  use ieee_float.Float64
  use caisar.types.Float64WithBounds as Feature
  use caisar.types.IntWithBounds as Label
  use caisar.model.Model

  use caisar.dataset.CSV
  use caisar.robust.ClassRobustCSV

  constant model_filename: string
  constant dataset_filename: string

  constant label_bounds: Label.bounds =
    Label.{ lower = 0; upper = 9 }

  constant feature_bounds: Feature.bounds =
    Feature.{ lower = (0.0:t); upper = (1.0:t) }

[... ]
  predicate robust (f_bounds: Feature.bounds)
    (l_bounds: Label.bounds)
    (m: model) (eps: t)
    (l: Label.t)
    (e: FeatureVector.t) =
    forall perturbed_e: FeatureVector.t.
      has_length perturbed_e (length e) ->
      FeatureVector.valid f_bounds perturbed_e ->
      let perturbation = perturbed_e - e in
      bounded_by_epsilon perturbation eps ->
      advises l_bounds m perturbed_e l

[... ]
  goal robustness:
    let nn = read_model model_filename in
    let dataset = read_dataset dataset_filename in
    let eps = (0.010000000...:t) in
    robust feature_bounds label_bounds nn dataset eps
end

```

(a) CAISAR

```

type Label = Index 10
type Image = Tensor Rat [28, 28]

@network
mnist : Image -> Tensor Rat [10]

validImage : Image -> Bool
validImage x = forall i j .
  0 <= x ! i ! j <= 1

advises : Image -> Label -> Bool
advises x i = forall j .
  j != i => mnist x ! i > mnist x ! j

@parameter
epsilon : Rat

boundedByEpsilon : Image -> Bool
boundedByEpsilon x = forall i j .
  -epsilon <= x ! i ! j <= epsilon

robust : Label -> Image -> Bool
robust label image = forall perturbation .
  boundedByEpsilon perturbation and
  validImage (perturbation + image) =>
  advises label (perturbation + image)

@parameter(infer=True)
n : Nat

@dataset
images : Tensor Image [n]

@dataset
labels : Tensor Label [n]

@property
robustness : Tensor Bool [n]
robustness = foreach i .
  robustAround (images ! i) (labels ! i)

```

(b) Vehicle

Fig. 4: An extract from a local robustness specification in CAISAR and Vehicle’s input languages for the same image dataset described in Fig. 3. Note the ability to reuse predicates and definitions, the conciseness of vector-based operations, and the explicit data set bindings.

tensors such as maps and folds. The language’s dependently typed nature allows the user to encode richer properties and includes tensor size constraints that can be checked before verification by the type-checker. Vehicle also has a backend that allows connecting proofs of neural network properties to larger system specifications in Agda [32]. However, unlike CAISAR, it connects to far fewer tools and cannot allow multiple solvers to work together.

These two languages solve the problems outlined in Sec. 3.1 and provide a concrete implementation. Note, in particular, that both manage data set bindings, neural network bindings, and data validity checks in clear, explicit ways. By doing this, they are essentially building the specification languages on top of the



Fig. 5: Schematic representation of the embedding gap.

existing pipelines: in Figs. 1 and 2, this is depicted by a dashed “Specification” box towards the left side. Other specification languages exist, like NeSAL [124] (which has no implementation) or DNNP [108] (lacking quantifiers and strong typing).

3.3 The Embedding Gap

We now consider the influence of larger system verification on the neural network verification pipeline (see Fig. 5). Consider a purely symbolic program $s(\cdot)$, whose completion requires computing a complex, unknown function $\mathcal{H} : \mathcal{P} \rightarrow \mathcal{R}$ that maps objects in the *problem input space* $\mathcal{P}$ to those in the *problem output space* $\mathcal{R}$. Given an embedding function $e : \mathcal{P} \rightarrow \mathbb{R}^m$ and an unembedding function $u : \mathbb{R}^n \rightarrow \mathcal{R}$, we can approximate $\mathcal{H}$ by training a neural network $f : \mathbb{R}^m \rightarrow \mathbb{R}^n$ such that $u \circ f \circ e \approx \mathcal{H}$. We refer to $u \circ f \circ e$ as the *solution*, and refer to $\mathbb{R}^m$ and $\mathbb{R}^n$ as the *embedding input space* and *embedding output space* respectively. Unlike objects in the problem space, the vectors in the embedding space are often not directly interpretable. The complete program is then modelled as $s(u \circ f \circ e)$. Examples of u and e would be the normalization of inputs, resizing operations for images, or data augmentation operations that are commonplace in machine learning pipelines.

Our end goal is to prove that $s(u \circ f \circ e)$ satisfies a property Ψ , which we will refer to as the *program property*. The natural way to proceed is to establish a *solution property* Φ and a *network property* Ξ such that the proof of Ψ is decomposable into the following three lemmas:

$$\Xi(f) \tag{2}$$

$$\forall g : \Xi(g) \Rightarrow \Phi(u \circ g \circ e) \tag{3}$$

$$\forall h : \Phi(h) \Rightarrow \Psi(s(h)) \tag{4}$$

i.e. Lemma 2 proves that the network f obeys the network property Ξ , then Lemma 3 proves that this implies $u \circ f \circ e$, the neural network lifted to the problem space, obeys the solution property Φ , and finally Lemma 4 proves that this implies $s(u \circ f \circ e)$, the neuro-symbolic program, obeys the program property Ψ .

The first issue that we run into is what we call the *embedding gap*. In Ψ , users would like to be able to model data that potentially has non-trivial semantics (for example, featuring both continuous and discrete parameters of a cyber-physical system such as velocity, stopping distance, switches etc.). However, in

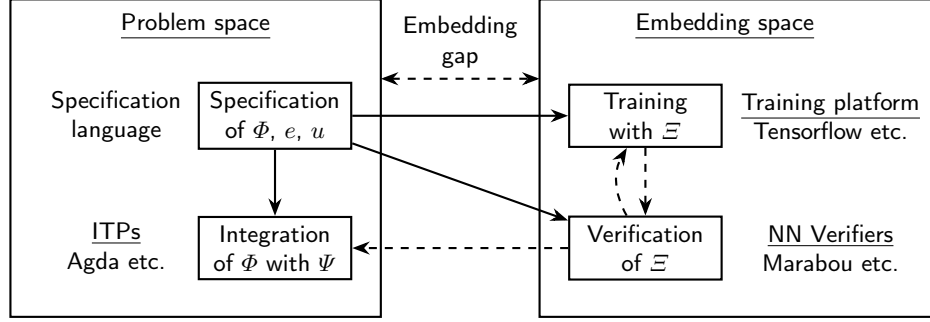


Fig. 6: Outline of Vehicle compiler backends, bridging the Embedding Gap [33, 32]. Dashed lines indicate information flow and solid lines automatic compilation.

Ξ , all values must be represented as continuous real vectors (in actuality, at the training phase, floating-point vectors, cf. Sec. 3.4). A function from the latter to the former must be highly non-surjective.

For example, consider an input type with two values, ‘Yes’ and ‘No’, encoded as real values ‘0.0’ and ‘1.0’ correspondingly. In the low-level query, one can encode that this input variable can only take two possible values using a disjunctive constraint ($x = 0.0 \vee x = 1.0$), but this does not scale well as the number of constructors in the data type grows, as each disjunction drastically increases the cost of verification. Instead, the most common current solution is to encode this as a single non-disjunctive constraint, $0.0 \leq x \leq 1$. In this case, the problem is that floating-point numbers may contain other values (e.g., ‘0.005’, ‘0.97’), which are meaningless in the chosen domain.

More generally, if users are to express specifications in Ψ , the high-level specification language must also allow users to specify the embedding and unembedding functions, e and u , as part of the specification. It should then be the responsibility of the compiler to generate suitable low-level queries representing Ξ . However, allowing the user to encode their specifications at the high-level Φ requires that the specification language compiler must be able to automatically translate from the former to the latter. The only existing attempt to provide programming language support for this was made by Vehicle [33, 32, 31] as shown in Fig. 6. In particular, Vehicle proposes a specification language to express Φ , e , u , and can compile the specification to Agda, in which more general properties of $s(\cdot)$ can be defined.

3.4 The Implementation Gap

In Sec. 3.1 we considered $\Xi(f^*)$, where f^* was an ONNX object, possibly obtained by conversion from the original implementation of f . The ONNX format has no backward translation from f^* to f , as the diagram in Fig. 7 shows. However, in the majority of neural network verification publications, authors implicitly assume that obtained verification guarantees about f^* extend to f . In

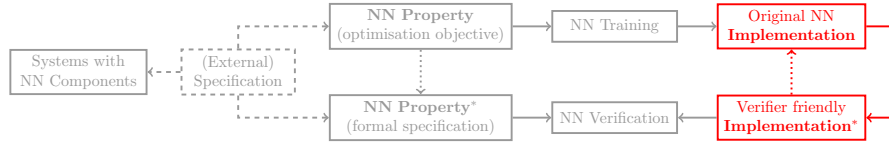


Fig. 7: Schematic representation of the implementation gap.

this section, we outline a range of problems caused by this and thus trace the right-most section of the diagram illustrated in Figs. 1 and 7.

Poor support for neural architecture conversion to ONNX. ONNX re-implementation of original neural networks remains a largely manual and un-verified procedure, which may be a source of errors. For example, neural networks contain different types of linear (e.g., fully connected, convolutional) and non-linear (e.g., ReLU, sigmoid, MaxPool) connections. Supporting the formal analysis of a new type of connection typically requires tool developers to add a new dedicated module to the codebase. For example, in verifiers based on abstract interpretation [111], this process would involve implementing the abstract transformer for the new type of connection. In SMT-based verification procedures [74, 122], the developer would need to implement the encoding, simplification, and satisfiability checking of constraints corresponding to the new connection. This process is tedious, repetitive, and error-prone. For example, the verification code for two-phase activation functions such as ReLU, Leaky ReLU, and absolute values is very similar, yet developers typically need to hard-code separate verification modules for each of these connections. Ideally, there should be automated conversion procedures with correctness guarantees.

Mismatch in numerical types. Barring experimental architectures that rely on analog computing [125], most implementations of neural networks are based on digital platforms that operate with finite-precision types such as integer and floating-point numbers. Effective conversion between real-valued types and finite-precision ones is an active research direction in machine learning [49].

The most ubiquitous numerical type in machine learning is the floating-point number [82, 14]. Indeed, the IEEE 754 single precision (32-bit) floating point type [65] is the de facto standard of libraries such as Tensorflow¹² and Pytorch¹³. Efforts to improve over the IEEE 754 standard exist, but they are often relegated to the context of hardware accelerators, where reducing the bit-size of numerical types may yield significant gains in terms of speed, memory and power consumption [119, 23].

From the verification perspective, it is well known that the safety certificates produced by real-valued neural network verifiers do not hold for floating-point

¹² <https://www.tensorflow.org>

¹³ <https://pytorch.org/>

implementations [71, 130]. Indeed, Jia and Rinard [71] propose an algorithm to search for floating-point counterexamples to real-valued safety certificates, thus invalidating them. Similarly, Zombori et al. [130] construct neural networks that contain undetectable backdoors, as long as the effects of numerical precision are neglected. Furthermore, the counterexamples produced by real-valued verifiers may not exist on a floating-point implementation of the same neural network, a phenomenon that has been reported on some VNN-COMP benchmarks [92].

Other sources of non-determinism. The current machine learning workflow, from training to inference, is not reproducible across different hardware and software platforms [101, 29, 129, 106]. This is due to a variety of reasons:

1. **Non-associativity of floating-point.** It is well-known that floating-point operations are not associative, i.e., $a + (b + c) \neq (a + b) + c$. As such, we can only verify the behaviour of a floating-point neural network if we know the *order* of all its operations. The existing *de-facto* standard ONNX does not include such a level of detail.
2. **Parallel execution.** Inference and training of neural networks are often sped up via parallel execution. Whether this is done via SIMD operations, multi-core CPUs, or GPU parallelism, it always introduces non-determinism in the results [101, 106].
3. **Auto-selection of primitives.** Modern machine learning compilers like XLA¹⁴ automatically select the most efficient algorithms depending on the computational load [101]. While PyTorch or Keras present ways to fix the behaviour of the algorithm, the ONNX runtime does not. For instance, Schögl et al. [106] report non-deterministic behaviour in the selection of convolutional algorithms on GPUs, which may alternate between explicit loop, GEMM-based, Winograd and FFT implementations.
4. **Runtime optimisations.** Machine learning frameworks may also implement runtime optimisation modifying the structure of the model itself to speed up inference or reduce memory usage, for example by fusing layers together (e.g. convolution and batch normalisation).
5. **Non-deterministic training.** The learning process itself is highly non-deterministic. Common sources include: parameter initialisation, data augmentation strategy, batch ordering, and dropout layers [101].
6. **Mathematical library rounding.** A long-standing issue in floating-point computation is incorrect roundings in the standard mathematical library `math.h`. Technically, the IEEE 754 standard recommends correct roundings [65], and there are efforts to create open-source implementations of `math.h` that abide by it [109]. However, mainstream compilers instead implement a variety of approximately-rounded algorithms [17].
7. **Low-level implementation details.** Furthermore, derived operators such as Softmax may leverage the fact that $\text{softmax}(x + c) = \text{softmax}(x)$ with constant c to increase the precision and avoid overflows. Such details can only be found in the low-level source code, even though they severely affect the precision of the computation.

¹⁴ <https://openxla.org/xla>

Overall, the end-to-end effects of the above causes of non-determinism cannot be neglected. Indeed, Pham et al. [101] reports a 2.9% difference in accuracy while reproducing the same training run on different platforms. Similarly, Cidon et al. [29] reports a 6% difference in accuracy when considering the whole image recognition pipeline, including camera noise and image processing algorithms.

From the verification perspective, certifying the safety of neural network implementations requires a different approach than high-level neural network verifiers like Marabou [121] or $\alpha\beta$ -CROWN [120]. Indeed, if we had access to the low-level implementation of every library in the machine learning pipeline, we could employ software verifiers [15] for this purpose. Unfortunately, existing software verifiers struggle to cope with the scale and complexity of neural network code [92, 89, 93]. In contrast, automated testing approaches are currently more effective [99, 56, 36], but cannot prove correctness.

Quantised neural networks. Switching to integer types (uniform quantisation) [49] can help alleviate some of the above problems (e.g. non-associativity of floating point, incorrect rounding) and improve reproducibility. From the machine learning perspective, a good quantisation scheme maintains the accuracy of the original floating-point neural network. Usually, 8-bit integers are used, but more aggressive quantisation schemes exist, down to ternary [58] and binary representations [104].

From the verification perspective, integer and binary data types require fundamentally different representations than the real-valued types used by mainstream verifiers such as Marabou [121] and $\alpha\beta$ -CROWN [120]. Existing work on verifying quantised neural networks relies on either the bit-vector SMT theory [50, 12, 59] or (mixed) integer linear programming (ILP, MILP) [95, 83, 127, 63]. In contrast, verifying the robustness of some binarised neural network architectures can be encoded as a satisfiability (SAT) problem [98, 70]. Other binarised architectures can still be encoded as a real-valued verification instance [6].

3.5 Reliable Proof Production and Checking

To overcome some of the challenges raised above, neural network verifiers may accompany their results with proof certificates, attesting their soundness using an external and relatively simple checking program (the *proof checker*) [66]. Since neural network verifiers are complex software, optimised for performance and speed, their verification is commonly intractable. Thus, proof production replaces the need to directly verify the neural network verifiers, with the need to verify only the proof checker. When a safety property is violated, neural network verifiers often accompany their results with a counterexample, which can then be checked by its evaluation in the network. However, proving safety (i.e., absence of violation of the property) is not straightforward, as the DNN verification problem is NP-complete even for simple networks [74, 105]. Therefore, proving safety is a greater challenge than proving a violation, and thus requires a more complicated proof and, consequently, a more complicated proof checker.

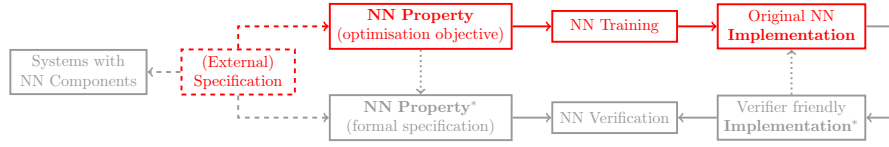


Fig. 8: Schematic representation of the neural network training pipeline.

Proof production mechanism, supporting several piecewise-linear activation functions, was implemented on top of Marabou [66, 121]. The proofs produced by Marabou are checked by a proof checker implemented within Marabou. The Marabou proof checker is implemented in C++ and uses floating points arithmetic for its computations.

When using an external proof checker, the reliability of the neural network verifier is dependent on the reliability of the proof checker. Therefore, the proof checker is expected to meet higher standards of reliability, ideally provable soundness. Functional programming languages allow the implementation of a precise checker and formal verification of its soundness. For example, a simply-typed language Imandra was deployed to check proofs produced by Marabou [37, 38]. This work also shows that computations with precise real arithmetic come at a price of limited performance. This opens up the possibility for a variety of implementations of the same checking algorithm in different programming languages, exploring the trade-off between precision and performance speed.

3.6 Property-Guided Training

Finally, we give a brief outline of the state-of-the-art in the property-guided training, which occupies the upper section of Fig. 8. This is a booming area in its own right, also known under the umbrella term of *neuro-symbolic AI*. By pointing out existing programming language discrepancies and solutions, we do not attempt to give a full survey of neuro-symbolic AI, but refer the reader to more comprehensive surveys [53, 60].

In the introduction, we have already outlined the evolution from adversarial training (seen as training for the robustness property specifically) into a more general property-driven training (for any property of choice) [46, 112, 47]. It is noteworthy that, although robust training by *projected gradient descent* [54, 88, 79] predates verification, contemporary approaches are often related to, or derived from, the corresponding verification methods by optimizing verification-inspired regularization terms.

The weakest form of property-based training boils down to translating a specification written in a subset of first-order logic into a *loss function*, that serves directly as an optimisation objective within the implementation of a training algorithm. Thus, the training algorithm optimises the neural network to satisfy the desired property. This translation method is known under the name of *differentiable logic* (or DL) [46, 112, 47]. Vehicle implements DL as one of

its backends [32] (cf. Fig. 6) and serves as a prototype of a compiler for neural network property specification languages (cf. Fig. 1). Recently, this inspired attempts at formalising different DLs in Coq [2].

There are other forms of training for robustness that come with stronger guarantees than DLs, e.g. IBP training [55, 126] and certified training [97, 128]. However, these usually have limited capacity for property specification; investigation of how these methods may fit into larger verification pipelines is warranted.

3.7 Other Directions

Verification of Cyber-Physical Systems. When following the diagram of Fig. 1, we did not impose any assumptions on the nature of properties we wish to ensure. In particular, we did not specify whether the “System” needs to be a cyber-physical system (CPS). However, CPS with machine learning components is an important safety-critical use case for neural network verification.

For example, a neural network may be utilized as a feedback controller for some plant model, typically represented as ordinary differential equations (ODEs) or generalizations thereof like hybrid automata. These are known as *neural networks control systems (NNCS)*. The introduction of constraints to describe the dynamics of a CPS requires revisiting several blocks of Fig. 1. Specifically, we need to replace the purely symbolic specifications and algorithms with those allowing for continuous variables and differential equations.

The annual International Competition on Verifying Continuous and Hybrid Systems (ARCH-COMP) has a category for this problem class, known as the AI and NNCS (AINNCS) category [87, 73, 72, 84, 85]. Several approaches for addressing the NNCS verification problem have been developed, such as implemented within software tools like CORA [76], JuliaReach [16], NNV [118, 86], OVERT [110], POLAR [61], Sherlock [41, 40], ReachNN* [62, 44], VenMAS [3], and Verisig [69, 68, 67].

More broadly, researchers have considered several strategies for the specification of properties of CPS with neural network components [48, 7, 24]. These cover significant challenges in the CPS domain, ranging from classical software verification problems to real-time systems concerns, scalability, as well as finding suitable specifications [107, 117, 116, 90]. Similarly to the standard neural network verification pipeline of Fig. 1, this area would benefit from a more principled programming language support.

Formal Specification of Probabilistic Properties. Program synthesis techniques can be valuable allies in producing correct-by-construction software and systems. In particular, the synthesis of logical formulas from a neural network and a dataset (e.g., via Inductive Logical Programming) received long-timed interest [100]. Also orthogonal to our work is Probabilistic Programming (as seen in [91]), which aims to provide a language and toolchain to express probabilistic properties of programs. It is clear that neural networks – seen as programs – would benefit from a probabilistic specification language. An early example in this direction is the ProbCompCert [114] project.

Existing Solutions	High-level				Low-level	Quantised	Software	Future	
	Vehicle [33]	CAISAR [51]	$\alpha\beta$ -Crown [120]	Marabou [75]	QEBVerif [127]	Aster [83]	CBMC [80]	ESBMC [94]	Formal Interfaces Unified Language
PL Challenges Addressed									
§3.1-3.2. Rigorous Semantics	✓	✓					✓	✓	✓
§3.3. Embedding Gap	✓								✓
§3.4. Implementation Gap		✓		✓	✓	✓	✓	✓	✓
§3.5. Proof Certificates				✓			✓*	✓*	✓
§3.6. Supports Training	✓								✓

Table 1: *Examples of existing solutions and the PL challenges they (partially) address. For the sake of variety, we include the existing solutions in four distinct categories: **high-level neural network verification DSLs** (Vehicle, CAISAR); best-performing (according to VNNCOMP) **low-level neural network verifiers** ($\alpha\beta$ -Crown, Marabou); **formalisation and synthesis of quantised neural networks** with mainstream MILP solvers (QEBVerif, Aster); and the use of **general-purpose software verifiers** (CBMC, ESBMC) in neural-network verification. For the latter, when we mark proof certificate production as ✓*, we refer to the generic proof production available for those verifiers, as opposed to the production of the Farkas witness for neural network UNSAT problems that is available in Marabou.*

Formalisation of Machine Learning. So far, research on formalisation of neural networks or optimisation algorithms has developed in isolation from the mainstream neural network verification pipelines summarised in Fig. 1. However, these two lines of research are bound to meet one day. Relevant work in formalisation of machine learning includes: verification of neural networks in Isabelle/HOL [22] and Imandra [39]; formalisation of piecewise affine activation functions in Coq [5]; providing formal guarantees of the degree to which the trained neural network will generalise to new data in Coq [10]; convergence, in this case of a single-layered perceptron in Coq [96]; verification of neural archetypes in Coq [34]. The two approaches that came the closest to unifying formalisation and verification in neural network domain are the Vehicle formalisation in Agda [9] and the formalisation of differentiable logics in Coq [2], relation of the latter to the verification pipeline of Fig. 1 is discussed in Sec. 3.6.

4 The Future Roadmap

The previous section identified five desirable programming language features that neural network verification could benefit from: rigorous semantics, support for handling the embedding and implementation gaps, generation of proof certificates, and rigorous integration of property-driven training into verification

pipelines. Currently, no single neural network verification tool or framework possesses all five features (see Table 1). Moreover, some tools considered leaders in the neural network verification market do not satisfy any. The title of this paper reflects our belief that the desirable solution – global specifications that *formally* explain the expected properties and yield a *formal* proof that the given implementation respects the specification – is a challenge in programming language design. In this section, we overview a couple of possible directions that may play a role in future solutions.

4.1 A Unified Dependently Typed Language

We believe the idealised solution to be a single language that is expressive enough to implement the machine learning pipeline and, at the same time, encode the desired properties of both the neural networks created by the pipeline and the pipeline itself. The following are a non-exhaustive list of the types of properties that should be representable:

1. theoretical results about the convergence of the training process;
2. correctness of tensor operations that underlie the training;
3. rich properties of the input data, e.g., constraining inputs to a certain range;
4. relation between input data and the weights in the network produced, e.g., robustness;
5. properties of the floating point numbers being used.

Given the complexity of encoding some of these properties (e.g. numerical stability, robustness), we believe that the expressive power of dependent types is a natural fit. We now briefly argue how such a unified dependent-typed language would allow us to make progress towards the challenges outlined in Table 1.

1. **Rigorous semantics.** The meta-theory of dependent types is well studied [30] so defining rigorous semantics for the language should not be a significant challenge. Furthermore, by implementing all the components in a single language, the friction of aligning the semantics of the different components in the system is significantly reduced.
2. **Embedding gap.** In a dependently typed language, from one perspective there would be no embedding gap, as any representation changes must be stated explicitly as type conversions. However, from another perspective, working in such a language does not address the fundamental problem of translating the proofs from the problem space to the embedding space. It merely moves the work from the external proofs into the type-conversion functions. Nevertheless, the expressive power of dependent types is more than sufficient to implement the partial solutions proposed by Vehicle (Sec. 3.3).
3. **Implementation gap.** The implementation gap (Sec. 3.4) will be resolved, as implementations will respect their types. For example, consider numerical types. If our specification assumes infinite reals, there is no way to instantiate an implementation that uses machine floats, as we will not be able to prove that machine floats is a valid representation of reals. If our specification is

weak and we do not require properties of the operations or other equalities, then machine floats may be a valid data type for the chosen constraints. If we envision implementation to operate on machine floats, we can describe all the properties of interest (e.g., lack of associativity) in the data type. We must understand that we cannot use external libraries such as XLA or OpenBLAS without verifying them, as this will break all the formal guarantees in our specification. Either these libraries have to be verified formally or we can synthesise the code with similar runtime properties directly from our specification in a type-preserving way.

4. **Proof certificates.** In a dependently-typed language, where the specifications are encoded as types, there is no need for separate proof certificates or proof checkers. In particular, the terms themselves act as the proof certificates and the type-checker acts as the proof checker.
5. **Supports training.** Although at the moment training is carried out in non-dependently typed languages, there is nothing stopping training (e.g. automatic differentiation or similar algorithms) from being implemented in dependently-typed language. Not only would such an implementation significantly reduce the friction between training and verification, it would also facilitate the integrating property-driven training and verification by viewing it as code synthesis problem. The key idea here is that generating code from a formal specification is much easier than checking whether the given code respects the specification.

4.2 Formal Interfaces

However, we are not naive as to the difficulty of implementing such a unified framework. Firstly, it will be an uphill battle to overcome the significant first-mover advantage of existing tools in their respective domains, e.g. training frameworks in Python, C and others and neural network verifiers. Even leaving that aside, work has shown that checking such complex type-based specifications in an efficient manner is still a challenging problem (e.g. Kokke et al. [77]).

Therefore, we believe a more realistic short-term goal is to keep the overarching maximally expressive specification language, but design a compiler that can utilise existing tools to achieve certain subgoals. In particular, we should follow the approach of industry, where the use of many disparate systems is common. In such an environment, the designers of these individual components should not only rigorously pre-define their interfaces, but also provide full formal specifications about their behaviour and, ideally, provide proof certificates that the output satisfies the specification as part of the interface. This would allow the compiler to specify the expected behaviour of a given module and let the programmer choose the best implementation (provided it respects the specification) at that abstraction level.

One possible inspiration for the design of such interfaces could come from the rich literature of behavioural interface specification language (BISL) [57]. A BISL is a family of languages used to specify the expected behaviour of a program at the *function* level, providing a fine-grained level of control on how

to precisely describe the function. BISLs usually follow a Hoare triplet-inspired formalism: the programmer should specify the precondition and the control flow; automated provers using weakest-precondition calculus or SAT can then automatically derive preconditions. Drawing from well-known languages like SPARK and Eiffel, it would then become easy to specify invariants on *several functions* at once. Such properties could then be translated back to the original program language (and then checked with the type system) or - if it results in a program that is impossible to represent - checked using external provers. This approach would be representative of what is being done in the industry for critical systems for decades, with JML, Why3 [45] or SPARK.

5 Conclusion

We have given some support for our main thesis – that *neural network verification is increasingly becoming a programming language challenge*. We hope this paper will provoke a stimulating discussion of this topic, helping the programming language community explore the opportunities presented by this new domain. Although we have supported our arguments with references to existing approaches, this is not a survey paper, and we make no claims of bibliographic completeness.

6 Acknowledgements

M. Daggitt and E. Komendantskaya acknowledge the partial support of the EPSRC grant AISEC: AI Secure and Explainable by Construction (EP/T026960/1). E. Komendantskaya was supported by ARIA: Mathematics for Safe AI grant. L. Cordeiro and E. Manino acknowledge the support of the EPSRC grant EnnCore: End-to-End Conceptual Guarding of Neural Architectures (EP/T026995/1). J. Girard-Satabin and Augustin Lemesle were supported by the French Agence Nationale de la Recherche (ANR) grant ANR-23-DEGR-0001 as part of the France 2030 programme. The work of Isac and Katz was partially funded by the European Union (ERC, VeriDeL, 101112713). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

1. Open Neural Network Exchange format, <https://onnx.ai/>, accessed on 30.01.2022
2. Affeldt, R., Bruni, A., Komendantskaya, E., Slusarz, N., Stark, K.: Taming Differentiable Logics with Coq Formalisation. In: Interactive Theorem Provers (ITP) 2024 (2024)

3. Akintunde, M.E., Botoeva, E., Kouvaros, P., Lomuscio, A.: Formal verification of neural agents in non-deterministic environments. In: International Conference on Autonomous Agents and Multiagent Systems, AAMAS. pp. 25–33 (2020)
4. Albarghouthi, A.: Introduction to neural network verification (2021), <https://arxiv.org/abs/2109.10317>
5. Aleksandrov, A., Völlinger, K.: Formalizing piecewise affine activation functions of neural networks in Coq. In: 15th International NASA Symposium on Formal Methods (NFM 2023), Houston, TX, USA, May 16–18, 2023. Lecture Notes in Computer Science, vol. 13903, pp. 62–78. Springer (2023). https://doi.org/10.1007/978-3-031-33170-1_4, https://doi.org/10.1007/978-3-031-33170-1_4
6. Amir, G., Wu, H., Barrett, C., Katz, G.: An smt-based approach for verifying binarized neural networks. In: Groote, J.F., Larsen, K.G. (eds.) Tools and Algorithms for the Construction and Analysis of Systems. pp. 203–222. Springer International Publishing, Cham (2021)
7. Astorga, A., Hsieh, C., Madhusudan, P., Mitra, S.: Perception contracts for safety of ml-enabled systems. *Proc. ACM Program. Lang.* **7**(OOPSLA2) (Oct 2023). <https://doi.org/10.1145/3622875>
8. Athavale, A., Bartocci, E., Christakis, M., Maffei, M., Nickovic, D., Weissenbacher, G.: Verifying global two-safety properties in neural networks with confidence (2024), <https://arxiv.org/abs/2405.14400>
9. Atkey, R., Daggett, M.L., Kokke, W.: Vehicle formalisation (2024), <https://github.com/vehicle-lang/vehicle-formalisation>
10. Bagnall, A., Stewart, G.: Certifying the true error: Machine learning in Coq with verified generalization guarantees. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 2662–2669 (2019)
11. Bak, S., Liu, C., Johnson, T.: The Second International Verification of Neural Networks Competition (VNN-COMP 2021): Summary and Results (2021), technical Report. <http://arxiv.org/abs/2109.00498>
12. Baranowski, M., He, S., Lechner, M., Nguyen, T.S., Rakamarić, Z.: An smt theory of fixed-point arithmetic. In: Peltier, N., Sofronie-Stokkermans, V. (eds.) Automated Reasoning. pp. 13–31. Springer International Publishing, Cham (2020)
13. Barrett, C., Fontaine, P., Tinelli, C.: The Satisfiability Modulo Theories Library (SMT-LIB). www.SMT-LIB.org (2016)
14. Baskin, C., Liss, N., Schwartz, E., Zheltonozhskii, E., Giryes, R., Bronstein, A.M., Mendelson, A.: Uniq: Uniform noise injection for non-uniform quantization of neural networks. *ACM Trans. Comput. Syst.* **37**(1–4) (mar 2021). <https://doi.org/10.1145/3444943>, <https://doi.org/10.1145/3444943>
15. Beyer, D.: Competition on software verification and witness validation: Sv-comp 2023. In: Sankaranarayanan, S., Sharygina, N. (eds.) Tools and Algorithms for the Construction and Analysis of Systems. pp. 495–522. Springer Nature Switzerland, Cham (2023)
16. Bogomolov, S., Forets, M., Frehse, G., Potomkin, K., Schilling, C.: JuliaReach: A toolbox for set-based reachability. In: Proc. of the 22nd ACM International Conference on Hybrid Systems: Computation and Control. p. 39–44 (2019). <https://doi.org/10.1145/3302504.3311804>
17. Brisebarre, N., Hanrot, G., Muller, J.M., Zimmermann, P.: Correctly-rounded evaluation of a function: why, how, and at what cost? (2024), <https://hal.science/hal-04474530/document>

18. Brix, C., Bak, S., Johnson, T.T., Wu, H.: The fifth international verification of neural networks competition (vnn-comp 2024): Summary and results (2024), <https://arxiv.org/abs/2412.19985>
19. Brix, C., Bak, S., Liu, C., Johnson, T.T.: The Fourth International Verification of Neural Networks Competition (VNN-COMP 2023): Summary and Results. CoRR **abs/2312.16760** (2023). <https://doi.org/10.48550/ARXIV.2312.16760>, <https://doi.org/10.48550/arXiv.2312.16760>
20. Brix, C., Müller, M.N., Bak, S., Johnson, T.T., Liu, C.: First three years of the international verification of neural networks competition (VNN-COMP). Int. J. Softw. Tools Technol. Transf. **25**(3), 329–339 (2023). <https://doi.org/10.1007/S10009-023-00703-4>, <https://doi.org/10.1007/s10009-023-00703-4>
21. Brix, C., Müller, M.N., Bak, S., Johnson, T.T., Liu, C.: First three years of the international verification of neural networks competition (vnn-comp). International Journal on Software Tools for Technology Transfer **25**(3), 329–339 (2023)
22. Brucker, A.D., Stell, A.: Verifying feedforward neural networks for classification in Isabelle/HOL. In: 25th International Symposium on Formal Methods (FM 2023), Lübeck, Germany, March 6–10, 2023. Lecture Notes in Computer Science, vol. 14000, pp. 427–444. Springer (2023). https://doi.org/10.1007/978-3-031-27481-7_24, https://doi.org/10.1007/978-3-031-27481-7_24
23. Burgess, N., Milanovic, J., Stephens, N., Monachopoulos, K., Mansell, D.: Bfloat16 processing for neural networks. In: 2019 IEEE 26th Symposium on Computer Arithmetic (ARITH). pp. 88–91 (2019). <https://doi.org/10.1109/ARITH.2019.00022>
24. Calinescu, R., Imrie, C., Mangal, R., Rodrigues, G.N., Păsăreanu, C., Santana, M.A., Vázquez, G.: Controller synthesis for autonomous systems with deep-learning perception components. IEEE Transactions on Software Engineering **50**(6), 1374–1395 (2024). <https://doi.org/10.1109/TSE.2024.3385378>
25. Carlini, N.: A complete list of all (arxiv) adversarial example papers (2019)
26. Casadio, M., Komendantskaya, E., Daggett, M.L., Kokke, W., Katz, G., Amir, G., Refaeli, I.: Neural network robustness as a verification property: A principled case study. In: Computer Aided Verification (CAV 2022). Lecture Notes in Computer Science, Springer (2022)
27. Casadio, M., Komendantskaya, E., Daggett, M.L., Kokke, W., Katz, G., Amir, G., Refaeli, I.: Neural network robustness as a verification property: A principled case study. In: Computer Aided Verification (CAV 2022). Lecture Notes in Computer Science, Springer (2022)
28. Christakis, M., Eniser, H.F., Hoffmann, J., Singla, A., Wüstholtz, V.: Specifying and testing k -safety properties for machine-learning models (2022), <https://arxiv.org/abs/2206.06054>
29. Cidon, E., Pergament, E., Asgar, Z., Cidon, A., Katti, S.: Characterizing and taming model instability across edge devices. In: Smola, A., Dimakis, A., Stolica, I. (eds.) Proceedings of Machine Learning and Systems. vol. 3, pp. 624–636 (2021), https://proceedings.mlsys.org/paper_files/paper/2021/file/5190e987c46a346974e351f96997d640-Paper.pdf
30. Coquand, T., Huet, G.: The calculus of constructions. Ph.D. thesis, Inria (1986)
31. Daggett, M.L., Kokke, W., Atkey, R., Arnaboldi, L., Komendantskaya, E.: Vehicle: Interfacing neural network verifiers with interactive theorem provers (2022). <https://doi.org/10.48550/ARXIV.2202.05207>, <https://arxiv.org/abs/2202.05207>

32. Daggitt, M.L., Kokke, W., Atkey, R., Slusarz, N., Arnaboldi, L., Komendantskaya, E.: Vehicle: Bridging the embedding gap in the verification of neuro-symbolic programs. CoRR **abs/2401.06379** (2024). <https://doi.org/10.48550/ARXIV.2401.06379>, <https://doi.org/10.48550/arXiv.2401.06379>
33. Daggitt, M.L., Kokke, W., Komendantskaya, E., Atkey, R., Arnaboldi, L., Slusarz, N., Casadio, M., Coke, B., Lee, J.: The vehicle tutorial: Neural network verification with vehicle. In: Narodytska, N., Amir, G., Katz, G., Isac, O. (eds.) Proceedings of the 6th Workshop on Formal Methods for ML-Enabled Autonomous Systems, FoMLAS@CAV 2023, Paris, France, July 17-18, 2023. Kalpa Publications in Computing, vol. 16, pp. 1–5. EasyChair (2023). <https://doi.org/10.29007/5S2X>, <https://doi.org/10.29007/5s2x>
34. De Maria, E., Bahrami, A., l’Yvonnet, T., Felty, A., Gaffé, D., Ressouche, A., Grammont, F.: On the use of formal methods to model and verify neuronal archetypes. *Frontiers of Computer Science* **16**(3), 1–22 (2022)
35. Demarchi, S., Guidotti, D., Pulina, L., Tacchella, A.: Supporting standardization of neural networks verification with vnn-lib and coconet. In: 6th Workshop on Formal Methods for ML-Enabled Autonomous Systems (Jul 2023)
36. Deng, Z., Meng, G., Chen, K., Liu, T., Xiang, L., Chen, C.: Differential testing of cross deep learning framework APIs: Revealing inconsistencies and vulnerabilities. In: 32nd USENIX Security Symposium (USENIX Security 23). pp. 7393–7410. USENIX Association, Anaheim, CA (Aug 2023), <https://www.usenix.org/conference/usenixsecurity23/presentation/deng-zizhuang>
37. Desmartin, R., Isac, O., Komendantskaya, E., Stark, K., Passmore, G., Katz, G.: A Certified Proof Checker for Deep Neural Network Verification. In: <https://arxiv.org/abs/2405.10611> (2024)
38. Desmartin, R., Isac, O., Passmore, G.O., Stark, K., Komendantskaya, E., Katz, G.: Towards a Certified Proof Checker for Deep Neural Network Verification. In: Glück, R., Kafle, B. (eds.) Logic-Based Program Synthesis and Transformation - 33rd International Symposium, LOPSTR 2023, Cascais, Portugal, October 23–24, 2023, Proceedings. Lecture Notes in Computer Science, vol. 14330, pp. 198–209. Springer (2023). https://doi.org/10.1007/978-3-031-45784-5_13, https://doi.org/10.1007/978-3-031-45784-5_13
39. Desmartin, R., Passmore, G.O., Komendantskaya, E., Daggit, M.: Checkinn: Wide range neural network verification in imandra. In: PPDP 2022: 24th International Symposium on Principles and Practice of Declarative Programming, Tbilisi, Georgia, September 20 - 22, 2022. pp. 3:1–3:14. ACM (2022). <https://doi.org/10.1145/3551357.3551372>, <https://doi.org/10.1145/3551357.3551372>
40. Dutta, S., Chen, X., Sankaranarayanan, S.: Reachability analysis for neural feedback systems using regressive polynomial rule inference. In: ACM International Conference on Hybrid Systems: Computation and Control, HSCC. pp. 157–168 (2019). <https://doi.org/10.1145/3302504.3311807>
41. Dutta, S., Jha, S., Sankaranarayanan, S., Tiwari, A.: Learning and verification of feedback control systems using feedforward neural networks. *IFAC-PapersOnLine* **51**(16), 151 – 156 (2018). <https://doi.org/10.1016/j.ifacol.2018.08.026>, iFAC Conference on Analysis and Design of Hybrid Systems ADHS 2018
42. Dutta, S., Jha, S., Sankaranarayanan, S., Tiwari, A.: Output range analysis for deep feedforward neural networks. In: Dutle, A., Muñoz, C., Narkawicz, A. (eds.)

- NASA Formal Methods. pp. 121–138. Springer International Publishing, Cham (2018)
43. Ehlers, R.: Formal verification of piece-wise linear feed-forward neural networks. In: D’Souza, D., Narayan Kumar, K. (eds.) *Automated Technology for Verification and Analysis*. pp. 269–286. Springer International Publishing, Cham (2017)
 44. Fan, J., Huang, C., Li, W., Chen, X., Zhu, Q.: Reachnn*: A tool for reachability analysis of neural-network controlled systems. In: *International Symposium on Automated Technology for Verification and Analysis (ATVA)* (2020)
 45. Filliâtre, J.C., Paskevich, A.: Why3 - Where Programs Meet Provers. In: Felleisen, M., Gardner, P. (eds.) *Programming Languages and Systems*. pp. 125–128. *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37036-6_8
 46. Fischer, M., Balunovic, M., Drachsler-Cohen, D., Gehr, T., Zhang, C., Vechev, M.T.: DL2: training and querying neural networks with logic. In: Chaudhuri, K., Salakhutdinov, R. (eds.) *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA. Proceedings of Machine Learning Research*, vol. 97, pp. 1931–1941. PMLR (2019), <http://proceedings.mlr.press/v97/fischer19a.html>
 47. Flinkow, T., Pearlmutter, B.A., Monahan, R.: Comparing differentiable logics for learning with logical constraints (2024), <https://arxiv.org/abs/2407.03847>
 48. Fremont, D.J., Dreossi, T., Ghosh, S., Yue, X., Sangiovanni-Vincentelli, A.L., Seshia, S.A.: Scenic: a language for scenario specification and scene generation. In: *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*. p. 63–78. *PLDI 2019*, Association for Computing Machinery, New York, NY, USA (2019). <https://doi.org/10.1145/3314221.3314633>
 49. Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M.W., Keutzer, K.: A survey of quantization methods for efficient neural network inference. In: *Low-Power Computer Vision*, pp. 291–326. Chapman and Hall/CRC (2022)
 50. Giacobbe, M., Henzinger, T.A., Lechner, M.: How many bits does it take to quantize your neural network? In: Biere, A., Parker, D. (eds.) *Tools and Algorithms for the Construction and Analysis of Systems*. pp. 79–97. Springer International Publishing, Cham (2020)
 51. Girard-Satabin, J., Alberti, M., Bobot, F., Chihani, Z., Lemesle, A.: Caesar: A platform for characterizing artificial intelligence safety and robustness. In: *AI Safety. CEUR-Workshop Proceedings*, Vienne, Austria (Jul 2022), <https://hal.archives-ouvertes.fr/hal-03687211>
 52. Giunchiglia, E., Stoian, M.C., Lukasiewicz, T.: Deep learning with logical constraints. In: *Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*. pp. 5478–5485. *International Joint Conferences on Artificial Intelligence Organization* (7 2022). <https://doi.org/10.24963/ijcai.2022/767>, <https://doi.org/10.24963/ijcai.2022/767>, survey Track
 53. Giunchiglia, E., Stoian, M.C., Lukasiewicz, T.: Deep learning with logical constraints. In: Raedt, L.D. (ed.) *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*. pp. 5478–5485. *ijcai.org* (2022). <https://doi.org/10.24963/ijcai.2022/767>, <https://doi.org/10.24963/ijcai.2022/767>
 54. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples (2015)
 55. Goyal, S., Dvijotham, K.D., Stanforth, R., Bunel, R., Qin, C., Uesato, J., Arandjelovic, R., Mann, T., Kohli, P.: Scalable verified training for provably robust

- image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4842–4851 (2019)
56. Guo, Q., Xie, X., Li, Y., Zhang, X., Liu, Y., Li, X., Shen, C.: Audee: Automated testing for deep learning frameworks. In: Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering. p. 486–498. ASE '20, Association for Computing Machinery, New York, NY, USA (2021). <https://doi.org/10.1145/3324884.3416571>, <https://doi.org/10.1145/3324884.3416571>
 57. Hatcliff, J., Leavens, G.T., Leino, K.R.M., Müller, P., Parkinson, M.: Behavioral interface specification languages. *ACM Comput. Surv.* **44**(3) (Jun 2012). <https://doi.org/10.1145/2187671.2187678>, <https://doi.org/10.1145/2187671.2187678>
 58. He, Z., Fan, D.: Simultaneously optimizing weight and quantizer of ternary neural network using truncated gaussian approximation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
 59. Henzinger, T.A., Lechner, M., Žikelić, Đ.: Scalable verification of quantized neural networks. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 3787–3795 (2021)
 60. Hitzler, P., Sarker, M.: *Neuro-symbolic Artificial Intelligence: The State of the Art*. IOS Press (2022)
 61. Huang, C., Fan, J., Chen, X., Li, W., Zhu, Q.: POLAR: A polynomial arithmetic framework for verifying neural-network controlled systems. In: International Symposium on Automated Technology for Verification and Analysis (ATVA) (2022)
 62. Huang, C., Fan, J., Li, W., Chen, X., Zhu, Q.: Reachnn: Reachability analysis of neural-network controlled systems. *ACM Transactions on Embedded Computing Systems (TECS)* **18**(5s), 1–22 (2019)
 63. Huang, P., Wu, H., Yang, Y., Daukantas, I., Wu, M., Zhang, Y., Barrett, C.: Towards efficient verification of quantized neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 21152–21160 (2024)
 64. Huang, X., Kwiatkowska, M., Wang, S., Wu, M.: Safety verification of deep neural networks (2017)
 65. IEEE: Ieee standard for floating-point arithmetic. *IEEE Std 754-2019* (Revision of IEEE 754-2008) pp. 1–84 (2019). <https://doi.org/10.1109/IEEESTD.2019.8766229>
 66. Isac, O., Barrett, C., Zhang, M., Katz, G.: Neural Network Verification with Proof Production. In: Proc. 22nd Int. Conf. on Formal Methods in Computer-Aided Design (FMCAD). pp. 38–48 (2022)
 67. Ivanov, R., Carpenter, T., Weimer, J., Alur, R., Pappas, G.J., Lee, I.: Verisig 2.0: Verification of neural network controllers using taylor model preconditioning. In: International Conference on Computer-Aided Verification (2021)
 68. Ivanov, R., Carpenter, T.J., Weimer, J., Alur, R., Pappas, G.J., Lee, I.: Verifying the safety of autonomous systems with neural network controllers. *ACM Trans. Embed. Comput. Syst.* **20**(1) (Dec 2020). <https://doi.org/10.1145/3419742>
 69. Ivanov, R., Weimer, J., Alur, R., Pappas, G.J., Lee, I.: Verisig: Verifying safety properties of hybrid systems with neural network controllers. In: International Conference on Hybrid Systems: Computation and Control. p. 169–178. HSCC, ACM (2019). <https://doi.org/10.1145/3302504.3311806>
 70. Jia, K., Rinard, M.: Efficient exact verification of binarized neural networks. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) *Advances in Neural Information Processing Systems*. vol. 33, pp. 1782–1795. Curran

- Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/1385974ed5904a438616ff7bdb3f7439-Paper.pdf
71. Jia, K., Rinard, M.: Exploiting verified neural networks via floating point numerical error. In: Drăgoi, C., Mukherjee, S., Namjoshi, K. (eds.) *Static Analysis*. pp. 191–205. Springer International Publishing, Cham (2021)
72. Johnson, T.T., Lopez, D.M., Benet, L., Forets, M., Guadalupe, S., Schilling, C., Ivanov, R., Carpenter, T.J., Weimer, J., Lee, I.: Arch-comp21 category report: Artificial intelligence and neural network control systems (ainncs) for continuous and hybrid systems plants. In: Frehse, G., Althoff, M. (eds.) *8th International Workshop on Applied Verification of Continuous and Hybrid Systems (ARCH21)*. EPiC Series in Computing, vol. 80, pp. 90–119. EasyChair (2021). <https://doi.org/10.29007/kfk9>, <https://easychair.org/publications/paper/Jq4h>
73. Johnson, T.T., Lopez, D.M., Musau, P., Tran, H.D., Botoeva, E., Leofante, F., Maleki, A., Sidrane, C., Fan, J., Huang, C.: Arch-comp20 category report: Artificial intelligence and neural network control systems (ainncs) for continuous and hybrid systems plants. In: Frehse, G., Althoff, M. (eds.) *ARCH20. 7th International Workshop on Applied Verification of Continuous and Hybrid Systems (ARCH20)*. EPiC Series in Computing, vol. 74, pp. 107–139. EasyChair (2020). <https://doi.org/10.29007/9xgv>, <https://easychair.org/publications/paper/Jvwg>
74. Katz, G., Barrett, C., Dill, D.L., Julian, K., Kochenderfer, M.J.: Reluplex: An efficient smt solver for verifying deep neural networks. In: *International conference on computer aided verification*. pp. 97–117. Springer (2017)
75. Katz, G., Huang, D., Ibeling, D., Julian, K., Lazarus, C., Lim, R., Shah, P., Thakoor, S., Wu, H., Zeljić, A., Dill, D., Kochenderfer, M., Barrett, C.: The Marabou Framework for Verification and Analysis of Deep Neural Networks, pp. 443–452 (07 2019)
76. Kochdumper, N., Schilling, C., Althoff, M., Bak, S.: Open- and closed-loop neural network verification using polynomial zonotopes. In: *NASA Formal Methods*. pp. 16–36. Springer (2023)
77. Kokke, W., Komendantskaya, E., Kienitz, D., Atkey, R., Aspinall, D.: Neural networks, secure by construction - an exploration of refinement types. In: d. S. Oliveira, B.C. (ed.) *Programming Languages and Systems - 18th Asian Symposium, APLAS 2020, Fukuoka, Japan, November 30 - December 2, 2020, Proceedings*. Lecture Notes in Computer Science, vol. 12470, pp. 67–85. Springer (2020). https://doi.org/10.1007/978-3-030-64437-6_4, https://doi.org/10.1007/978-3-030-64437-6_4
78. Kolter, Z., Madry, A.: Adversarial robustness—theory and practice. *NeurIPS 2018 tutorial* (2018), available at <https://adversarial-ml-tutorial.org/>
79. Kolter, Z., Madry, A.: Adversarial robustness: Theory and practice. *Tutorial at NeurIPS* p. 3 (2018)
80. Kroening, D., Tautschnig, M.: CBMC-C bounded model checker. In: *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*. pp. 389–391. Springer (2014)
81. Lemesle, A., Lehmann, J., Gall, T.L.: Neural network verification with pyrat (2024), <https://arxiv.org/abs/2410.23903>
82. Li, Y., Dong, X., Wang, W.: Additive powers-of-two quantization: An efficient non-uniform discretization for neural networks. In: *International Conference on Learning Representations* (2020), <https://openreview.net/forum?id=BkgXT24tDS>

83. Lohar, D., Jeangoudoux, C., Volkova, A., Darulova, E.: Sound mixed fixed-point quantization of neural networks. *ACM Trans. Embed. Comput. Syst.* **22**(5s) (sep 2023). <https://doi.org/10.1145/3609118>, <https://doi.org/10.1145/3609118>
84. Lopez, D.M., Althoff, M., Benet, L., Chen, X., Fan, J., Forets, M., Huang, C., Johnson, T.T., Ladner, T., Li, W., Schilling, C., Zhu, Q.: Arch-comp22 category report: Artificial intelligence and neural network control systems (ainncs) for continuous and hybrid systems plants. In: Frehse, G., Althoff, M., Schoitsch, E., Guiochet, J. (eds.) *Proceedings of 9th International Workshop on Applied Verification of Continuous and Hybrid Systems (ARCH22)*. EPiC Series in Computing, vol. 90, pp. 142–184. EasyChair (2022). <https://doi.org/10.29007/wfgr>, <https://easychair.org/publications/paper/C1J8>
85. Lopez, D.M., Althoff, M., Forets, M., Johnson, T.T., Ladner, T., Schilling, C.: Arch-comp23 category report: Artificial intelligence and neural network control systems (ainncs) for continuous and hybrid systems plants. In: Frehse, G., Althoff, M. (eds.) *Proceedings of 10th International Workshop on Applied Verification of Continuous and Hybrid Systems (ARCH23)*. EPiC Series in Computing, vol. 96, pp. 89–125. EasyChair (2023). <https://doi.org/10.29007/x38n>
86. Lopez, D.M., Choi, S.W., Tran, H.D., Johnson, T.T.: NNV 2.0: The neural network verification tool. In: Enea, C., Lal, A. (eds.) *Computer Aided Verification*. pp. 397–412. Springer Nature Switzerland, Cham (2023)
87. Lopez, D.M., Musau, P., Tran, H.D., Dutta, S., Carpenter, T.J., Ivanov, R., Johnson, T.T.: Arch-comp19 category report: Artificial intelligence and neural network control systems (ainncs) for continuous and hybrid systems plants. In: Frehse, G., Althoff, M. (eds.) *ARCH19. 6th International Workshop on Applied Verification of Continuous and Hybrid Systems*. EPiC Series in Computing, vol. 61, pp. 103–119. EasyChair (2019). <https://doi.org/10.29007/rgv8>, <https://easychair.org/publications/paper/BFKs>
88. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: *International Conference on Learning Representations* (2018)
89. Magalhães, J.W.d.S., Woodruff, J., Polgreen, E., O’Boyle, M.F.P.: C2taco: Lifting tensor code to taco. In: *Proceedings of the 22nd ACM SIGPLAN International Conference on Generative Programming: Concepts and Experiences*. p. 42–56. GPCE 2023, Association for Computing Machinery, New York, NY, USA (2023). <https://doi.org/10.1145/3624007.3624053>, <https://doi.org/10.1145/3624007.3624053>
90. Mandal, U., Amir, G., Wu, H., Daukantas, I., Newell, F.L., Ravaioli, U.J., Meng, B., Durling, M., Ganai, M., Shim, T., Katz, G., Barrett, C.W.: Formally verifying deep reinforcement learning controllers with lyapunov barrier certificates. *CoRR abs/2405.14058* (2024). <https://doi.org/10.48550/ARXIV.2405.14058>, <https://doi.org/10.48550/arXiv.2405.14058>
91. Manhaeve, R., Dumančić, S., Kimmig, A., Demeester, T., De Raedt, L.: Neural probabilistic logic programming in deepproblog. *Artificial Intelligence* **298**, 103504 (2021). <https://doi.org/https://doi.org/10.1016/j.artint.2021.103504>, <https://www.sciencedirect.com/science/article/pii/S0004370221000552>
92. Manino, E., Menezes, R.S., Shmarov, F., Cordeiro, L.C.: NeuroCodeBench: a Plain C Neural Network Benchmark for Software Verification. In: *Workshop on Automated Formal Reasoning for Trustworthy AI Systems* (2023)
93. Matos, J.B.P., de Lima Filho, E.B., Bessa, I., Manino, E., Song, X., Cordeiro, L.C.: Counterexample guided neural network quantization refinement. *IEEE Transac-*

- tions on Computer-Aided Design of Integrated Circuits and Systems **43**(4), 1121–1134 (2024). <https://doi.org/10.1109/TCAD.2023.3335313>
94. Menezes, R.S., Aldughaim, M., Farias, B., Li, X., Manino, E., Shmarov, F., Song, K., Brauße, F., Gadelha, M.R., Tihanyi, N., Korovin, K., Cordeiro, L.C.: *Esbmc v7.4: Harnessing the power of intervals*. In: Finkbeiner, B., Kovács, L. (eds.) *Tools and Algorithms for the Construction and Analysis of Systems*. pp. 376–380. Springer Nature Switzerland, Cham (2024)
95. Mistry, S., Saha, I., Biswas, S.: An milp encoding for efficient verification of quantized deep neural networks. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **41**(11), 4445–4456 (2022). <https://doi.org/10.1109/TCAD.2022.3197697>
96. Murphy, C., Gray, P., Stewart, G.: Verified perceptron convergence theorem. In: *Proceedings of the 1st ACM SIGPLAN International Workshop on Machine Learning and Programming Languages*. pp. 43–50 (2017)
97. Müller, M.N., Eckert, F., Fischer, M., Vechev, M.: Certified training: Small boxes are all you need (2023)
98. Narodytska, N., Kasiviswanathan, S., Ryzhyk, L., Sagiv, M., Walsh, T.: Verifying properties of binarized deep neural networks. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 32 (2018)
99. Odena, A., Olsson, C., Andersen, D., Goodfellow, I.: TensorFuzz: Debugging neural networks with coverage-guided fuzzing. In: Chaudhuri, K., Salakhutdinov, R. (eds.) *Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research*, vol. 97, pp. 4901–4911. PMLR (09–15 Jun 2019), <https://proceedings.mlr.press/v97/odena19a.html>
100. Payani, A., Fekri, F.: Inductive Logic Programming via Differentiable Deep Neural Logic Networks. Tech. rep. (Jun 2019). <https://doi.org/10.48550/arXiv.1906.03523>, <http://arxiv.org/abs/1906.03523>, zSCC: 0000039 arXiv:1906.03523 [cs] type: article
101. Pham, H.V., Qian, S., Wang, J., Lutellier, T., Rosenthal, J., Tan, L., Yu, Y., Nagappan, N.: Problems and opportunities in training deep learning software systems: an analysis of variance. In: *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering*. p. 771–783. ASE ’20, Association for Computing Machinery, New York, NY, USA (2021). <https://doi.org/10.1145/3324884.3416545>, <https://doi.org/10.1145/3324884.3416545>
102. Prach, B., Brau, F., Buttazzo, G., Lampert, C.H.: 1-lipschitz layers compared: Memory speed and certifiable robustness. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 24574–24583 (June 2024)
103. Pulina, L., Tacchella, A.: An abstraction-refinement approach to verification of artificial neural networks. In: Touili, T., Cook, B., Jackson, P. (eds.) *Computer Aided Verification*. pp. 243–257. Springer Berlin Heidelberg, Berlin, Heidelberg (2010)
104. Qin, H., Gong, R., Liu, X., Bai, X., Song, J., Sebe, N.: Binary neural networks: A survey. *Pattern Recognition* **105**, 107281 (2020). <https://doi.org/https://doi.org/10.1016/j.patcog.2020.107281>, <https://www.sciencedirect.com/science/article/pii/S0031320320300856>
105. Sälzer, M., Lange, M.: Reachability Is NP-Complete Even for the Simplest Neural Networks. In: *Proc. 15th Int. Conf. on Reachability Problems (RP)*. pp. 149–164 (2021)

106. Schlögl, A., Hofer, N., Böhme, R.: Causes and effects of unanticipated numerical deviations in neural network inference frameworks. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) *Advances in Neural Information Processing Systems*. vol. 36, pp. 56095–56107. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/af076c3bdbf935b81d808e37c5ede463-Paper-Conference.pdf
107. Seshia, S.A., Sadigh, D., Sastry, S.S.: Toward verified artificial intelligence. *Commun. ACM* **65**(7), 46–55 (Jun 2022). <https://doi.org/10.1145/3503914>
108. Shriver, D., Elbaum, S., Dwyer, M.B.: DNNV: A framework for deep neural network verification. In: Silva, A., Leino, K.R.M. (eds.) *Computer Aided Verification*. pp. 137–150. Springer International Publishing, Cham (2021)
109. Sibidanov, A., Zimmermann, P., Glondou, S.: The core-math project. In: 2022 IEEE 29th Symposium on Computer Arithmetic (ARITH). pp. 26–34 (2022). <https://doi.org/10.1109/ARITH54963.2022.00014>
110. Sidrane, C., Kochenderfer, M.J.: OVERT: Verification of nonlinear dynamical systems with neural network controllers via overapproximation. *Safe Machine Learning workshop at ICLR* (2019)
111. Singh, G., Gehr, T., Püschel, M., Vechev, M.: An abstract domain for certifying neural networks. *Proceedings of the ACM on Programming Languages* **3**(POPL), 1–30 (2019)
112. Slusarz, N., Komendantskaya, E., Daggett, M.L., Stewart, R.J., Stark, K.: Logic of differentiable logics: Towards a uniform semantics of DL. In: Piskac, R., Voronkov, A. (eds.) *LPAR 2023: Proceedings of 24th International Conference on Logic for Programming, Artificial Intelligence and Reasoning*, Manizales, Colombia, 4–9th June 2023. *EPiC Series in Computing*, vol. 94, pp. 473–493. EasyChair (2023). <https://doi.org/10.29007/C1NT>, <https://doi.org/10.29007/c1nt>
113. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks (2014)
114. Tassarotti, J., Tristan, J.B.: Verified density compilation for a probabilistic programming language. *Proc. ACM Program. Lang.* **7**(PLDI) (Jun 2023). <https://doi.org/10.1145/3591245>, <https://doi.org/10.1145/3591245>
115. Teuber, S., Büning, M.K., Kern, P., Sinz, C.: Geometric path enumeration for equivalence verification of neural networks. In: 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI). pp. 200–208 (2021). <https://doi.org/10.1109/ICTAI52525.2021.00035>
116. Teuber, S., Mitsch, S., Platzer, A.: Provably safe neural network controllers via differential dynamic logic. *CoRR* **abs/2402.10998** (2024). <https://doi.org/10.48550/ARXIV.2402.10998>, <https://doi.org/10.48550/arXiv.2402.10998>
117. Tran, H.D., Xiang, W., Johnson, T.T.: Verification approaches for learning-enabled autonomous cyber-physical systems. *IEEE Design & Test* **39**(1), 24–34 (2022). <https://doi.org/10.1109/MDAT.2020.3015712>
118. Tran, H.D., Yang, X., Lopez, D.M., Musau, P., Nguyen, L.V., Xiang, W., Bak, S., Johnson, T.T.: NNV: The neural network verification tool for deep neural networks and learning-enabled cyber-physical systems. In: 32nd International Conference on Computer-Aided Verification (CAV’20) (7 2020)
119. Wang, N., Choi, J., Brand, D., Chen, C.Y., Gopalakrishnan, K.: Training deep neural networks with 8-bit floating point numbers. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.)

- Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018), https://proceedings.neurips.cc/paper_files/paper/2018/file/335d3d1cd7ef05ec77714a215134914c-Paper.pdf
120. Wang, S., Zhang, H., Xu, K., Lin, X., Jana, S., Hsieh, C.J., Kolter, J.Z.: Beta-crown: Efficient bound propagation with per-neuron split constraints for neural network robustness verification. *Advances in Neural Information Processing Systems* **34**, 29909–29921 (2021)
121. Wu, H., Isac, O., Zeljic, A., Tagomori, T., Daggitt, M.L., Kokke, W., Refaeli, I., Amir, G., Julian, K., Bassan, S., Huang, P., Lahav, O., Wu, M., Zhang, M., Komendantskaya, E., Katz, G., Barrett, C.W.: Marabou 2.0: A Versatile Formal Analyzer of Neural Networks. In: *Computer Aided Verification (CAV)* (2024)
122. Wu, H., Zeljić, A., Katz, G., Barrett, C.: Efficient neural network analysis with sum-of-infeasibilities. In: *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*. pp. 143–163. Springer (2022)
123. Xiang, W., Tran, H.D., Johnson, T.T.: Output reachable set estimation and verification for multilayer neural networks. *IEEE transactions on neural networks and learning systems* **29**(11), 5777–5783 (2018)
124. Xie, X., Kersting, K., Neider, D.: Neuro-symbolic verification of deep neural networks. In: Raedt, L.D. (ed.) *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*. pp. 3622–3628. International Joint Conferences on Artificial Intelligence Organization (7 2022). <https://doi.org/10.24963/ijcai.2022/503>, <https://doi.org/10.24963/ijcai.2022/503>, main Track
125. Yao, P., Wu, H., Gao, B., Tang, J., Zhang, Q., Zhang, W., Yang, J.J., Qian, H.: Fully hardware-implemented memristor convolutional neural network. *Nature* **577**(7792), 641–646 (2020)
126. Zhang, H., Chen, H., Xiao, C., Goyal, S., Stanforth, R., Li, B., Boning, D., Hsieh, C.J.: Towards stable and efficient training of verifiably robust neural networks. In: *8th International Conference on Learning Representations, ICLR 2020* (2020)
127. Zhang, Y., Song, F., Sun, J.: Qebverif: Quantization error bound verification of neural networks. In: Enea, C., Lal, A. (eds.) *Computer Aided Verification*. pp. 413–437. Springer Nature Switzerland, Cham (2023)
128. Zhang, Y., Albarghouthi, A., D’Antoni, L.: Robustness to programmable string transformations via augmented abstract training. In: *Proceedings of the 37th International Conference on Machine Learning*. pp. 11023–11032 (2020)
129. Zhuang, D., Zhang, X., Song, S., Hooker, S.: Randomness in neural network training: Characterizing the impact of tooling. In: Marculescu, D., Chi, Y., Wu, C. (eds.) *Proceedings of Machine Learning and Systems*. vol. 4, pp. 316–336 (2022), https://proceedings.mlsys.org/paper_files/paper/2022/file/427e0e886ebf87538afdf0badb805b7f-Paper.pdf
130. Zombori, D., Bánhelyi, B., Csendes, T., Megyeri, I., Jelasity, M.: Fooling a complete neural network verifier. In: *International Conference on Learning Representations* (2021), <https://openreview.net/forum?id=4IwieFS441>