





SyReNN: A Tool for Analyzing Deep Neural Networks *

Matthew Sotoudeh  (✉) and Aditya V. Thakur  (✉)

University of California, Davis CA 95616, USA
{masotoudeh, avthakur}@ucdavis.edu



Abstract. Deep Neural Networks (DNNs) are rapidly gaining popularity in a variety of important domains. Formally, DNNs are complicated vector-valued functions which come in a variety of sizes and applications. Unfortunately, modern DNNs have been shown to be vulnerable to a variety of attacks and buggy behavior. This has motivated recent work in formally analyzing the properties of such DNNs. This paper introduces SyReNN, a tool for understanding and analyzing a DNN by computing its *symbolic representation*. The key insight is to decompose the DNN into linear functions. Our tool is designed for analyses using *low-dimensional subsets* of the input space, a unique design point in the space of DNN analysis tools. We describe the tool and the underlying theory, then evaluate its use and performance on three case studies: computing Integrated Gradients, visualizing a DNN’s decision boundaries, and patching a DNN.

Keywords: Deep Neural Networks · Symbolic representation · Integrated Gradients

1 Introduction

Deep Neural Networks (DNNs) [18] have become the state-of-the-art in a variety of applications including image recognition [53,33] and natural language processing [12]. Moreover, they are increasingly used in safety- and security-critical applications such as autonomous vehicles [31] and medical diagnosis [10,38,28,37]. These advances have been accelerated by improved hardware and algorithms.

DNNs (Section 2) are programs that compute a vector-valued function, i.e., from \mathbb{R}^n to \mathbb{R}^m . They are straight-line programs written as a concatenation of alternating linear and non-linear *layers*. The coefficients of the linear layers are learned from data via *gradient descent* during a training process. A number of different non-linear layers (called *activation functions*) are commonly used, including the *rectified linear* and *maximum pooling* functions.

Owing to the variety of application domains and deployment constraints, DNNs come in many different sizes. For instance, large image-recognition and

* Artifact available at <https://zenodo.org/record/4124489>. Extended paper available at <https://arxiv.org/abs/2101.03263>.

natural-language processing models are trained and deployed using cloud resources [33,12], medium-size models could be trained in the cloud but deployed on hardware with limited resources [31], and finally small models could be trained and deployed directly on edge devices [47,9,22,34,35]. There has also been a recent push to compress trained models to reduce their size [24]. Such smaller models play an especially important role in privacy-critical applications, such as wake word detection for voice assistants, because they allow sensitive user data to stay on the user’s own device instead of needing to be sent to a remote computer for processing.

Although DNNs are very popular, they are not perfect. One particularly concerning development is that modern DNNs have been shown to be extremely vulnerable to *adversarial examples*, inputs which are intentionally manipulated to appear unmodified to humans but become misclassified by the DNN [54,19,40,8]. Similarly, *fooling examples* are inputs that look like random noise to humans, but are classified with high confidence by DNNs [41]. Mistakes made by DNNs have led to loss of life [36,17] and wrongful arrests [26,27]. For this reason, it is important to develop techniques for analyzing, understanding, and repairing DNNs.

This paper introduces SyReNN, a tool for understanding and analyzing DNNs. SyReNN implements state-of-the-art algorithms for computing precise symbolic representations of piecewise-linear DNNs (Section 3). Given an input subspace of a DNN, SyReNN computes a symbolic representation that decomposes the behavior of the DNN into finitely-many linear functions. SyReNN implements the one-dimensional analysis algorithm of Sotoudeh and Thakur [50] and extends it to the two-dimensional setting as described in Section 4.

Key insights. There are two key insights enabling this approach, first identified in Sotoudeh and Thakur [50]. First, most popular DNN architectures today are *piecewise-linear*, meaning they can be precisely decomposed into finitely-many linear functions. This allows us to reduce their analysis to equivalent questions in linear algebra, one of the most well-understood fields of modern mathematics. Second, many applications only require analyzing the behavior of the DNN on a *low-dimensional subset* of the input space. Hence, whereas prior work has attempted to give up precision for efficiency in analyzing high-dimensional input regions [48,49,16], our work has focused on algorithms that are *both efficient and precise* in analyzing lower-dimensional regions (Section 4).

Tool design. The SyReNN tool is designed to be easy to use and extend, as well as efficient (Section 5). The core of SyReNN is written as a highly-optimized, parallel C++ server using Intel TBB for parallelization [45] and Eigen for matrix operations [23]. A user-friendly Python front-end interfaces with the PyTorch deep learning framework [44].

Use cases. We demonstrate the utility of SyReNN using three applications. The first computes *Integrated Gradients* (IG), a state-of-the-art measure used to determine which input dimensions (e.g., pixels for an image-recognition network) were most important in the final classification produced by the network (Section 6.1). The second precisely visualizes the decision boundaries of a DNN (Section 6.2). The last *patches* (repairs) a DNN to satisfy some desired specification

involving infinitely-many points (Section 6.3). Thus, SyReNN is an interesting and useful tool in the toolbox for understanding and analyzing DNNs.

Contributions. The contributions of this paper are:

- A definition of symbolic representation of DNNs (Section 3).
- An efficient algorithm for computing symbolic representations for DNNs over low-dimensional input subspaces (Section 4).
- A design of a usable and well-engineered tool implementing these ideas called SyReNN (Section 5).
- Three applications of SyReNN (Section 6).

Section 2 presents preliminaries about DNNs; Section 7 presents related work; Section 8 concludes. SyReNN is available on GitHub at <https://github.com/95616ARG/SyReNN>.

2 Preliminaries

We now formally define the notion of *DNN* we will use in this paper.

Definition 1. A Deep Neural Network (*DNN*) is a function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ which can be written $f = f_1 \circ f_2 \cdots \circ f_n$ for a sequence of layer functions f_1, f_2, \dots, f_n .

Our work is primarily concerned with the popular class of *piecewise-linear* DNNs, defined below. In this definition and the rest of this paper, we will use the term “polytope” to mean a *convex and bounded* polytope except where specified.

Definition 2. A function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is *piecewise-linear* (PWL) if its input domain \mathbb{R}^n can be partitioned into finitely-many possibly-unbounded polytopes X_1, X_2, \dots, X_k such that $f|_{X_i}$ is linear for every X_i .

The most common activation function used today is the ReLU function, a PWL activation function which is defined below.

Definition 3. The rectified linear function (*ReLU*) is a function $\text{ReLU} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ defined component-wise by

$$\text{ReLU}(\vec{v})_i := \begin{cases} 0 & \text{if } v_i < 0 \\ v_i & \text{otherwise,} \end{cases}$$

where $\text{ReLU}(\vec{v})_i$ is the i th component of the vector $\text{ReLU}(\vec{v})$ and v_i is the i th component of the vector \vec{v} .

In order to see that ReLU is PWL, we must show that its input domain \mathbb{R}^n can be partitioned such that, in each partition, ReLU is linear. In this case, we can use the orthants of \mathbb{R}^n as our partitioning: within each orthant, the signs of the components do not change hence ReLU is the linear function that just zeros out the negative components.

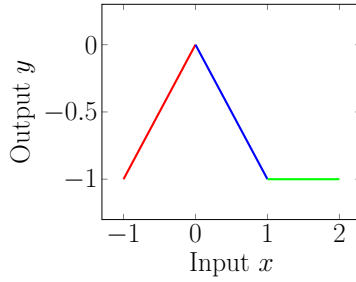


Fig. 1: Example function for which $\widehat{f}_{\uparrow[-1,2]} = \{[-1, 0], [0, 1], [1, 2]\}$.

Although we focus on ReLU due to its popularity and expository power, SyReNN works with a number of other popular PWL layers include MaxPool, Leaky ReLU, Hard Tanh, Fully-Connected, and Convolutional layers, as defined in [18]. PWL layers have become exceedingly common. In fact, nearly all of the state-of-the-art image recognition models bundled with Pytorch [43] are PWL.

Example 1. The DNN $f : \mathbb{R}^1 \rightarrow \mathbb{R}^1$ defined by

$$f(x) := [1 \ -1 \ -1] \text{ReLU} \left(\begin{bmatrix} 1 & -1 \\ 1 & 0 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix} \right)$$

can be broken into layers $f = f_1 \circ f_2 \circ f_3$ where

$$f_1(x) := \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix}, \quad f_2 = \text{ReLU}, \quad \text{and} \quad f_3(\vec{v}) = [1 \ -1 \ -1] \vec{v}.$$

The DNN’s input-output behavior on the domain $[-1, 2]$ is shown in Figure 1.

3 A Symbolic Representation of DNNs

We formalize the symbolic representation according to the following definition:

Definition 4. *Given a PWL function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and a bounded convex polytope $X \subseteq \mathbb{R}^n$, we define the symbolic representation of f on X , written $\widehat{f}_{\uparrow X}$, to be a finite set of polytopes $\widehat{f}_{\uparrow X} = \{P_1, \dots, P_n\}$, such that:*

1. *The set $\{P_1, P_2, \dots, P_n\}$ partitions X , except possibly for overlapping boundaries.*
2. *Each P_i is a bounded convex polytope.*
3. *Within each P_i , the function $f_{\uparrow P_i}$ is linear.*

Notably, if f is a DNN using only PWL layers, then f is PWL and so we can define $\widehat{f}_{\uparrow X}$. This symbolic representation allows one to reduce questions about

the DNN f to questions about finitely-many linear functions F_i . For example, because linear functions are convex, to verify that $\forall x \in X. f(x) \in Y$ for some polytope Y , it suffices to verify $\forall P_i \in \widehat{f_{\upharpoonright X}}. \forall \vec{v} \in \mathbf{Vert}(P_i). f(\vec{v}) \in Y$, where $\mathbf{Vert}(P_i)$ is the (finite) set of vertices for the bounded convex polytope P_i ; thus, here both of the quantifiers are over finite sets. The symbolic representation described above can be seen as a generalization of the EXACTLINE representation [50], which considered only *one-dimensional* restriction domains of interest.

Example 2. Consider again the DNN $f : \mathbb{R}^1 \rightarrow \mathbb{R}^1$ given by

$$f(x) := [1 \ -1 \ -1] \text{ReLU} \left(\begin{bmatrix} 1 & -1 \\ 1 & 0 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix} \right)$$

and the region of interest $X = [-1, 2]$. The input-output behavior of f on X is shown in Figure 1. From this, we can see that

$$\widehat{f_{\upharpoonright X}} = \{-1, 0\}, [0, 1], [1, 2]\}.$$

Within each of these partitions, the input-output behavior is *linear*, which for $\mathbb{R}^1 \rightarrow \mathbb{R}^1$ we can see visually as just a line segment. As this set fully partitions X , then, this is a valid $\widehat{f_{\upharpoonright X}}$.

4 Computing the Symbolic Representation

This section presents an efficient algorithm for computing $\widehat{f_{\upharpoonright X}}$ for a DNN f composed of PWL layers. To retain both scalability and precision, we will *require the input region X be two-dimensional*. This design choice is relatively unexplored in the neural-network analysis literature (most analyses strike a balance between precision and scalability, ignoring dimensionality). We show that, for two-dimensional X , we can use an efficient polytope representation to produce an algorithm that demonstrates good best-case and in-practice efficiency while retaining full precision. This algorithm represents a direct generalization of the approach of [50].

The difficulties our algorithm addresses arise from three areas. First, when computing $\widehat{f_{\upharpoonright X}}$ there may be exponentially many such partitions on all of \mathbb{R}^n but only a small number of them may intersect with X . Consequently, the algorithm needs to be able to find those partitions that intersect with X efficiently without explicitly listing all of the partitions on \mathbb{R}^n . Second, it is often more convenient to specify the partitioning via *hyperplanes separating the partitions* than explicit polytopes. For example, for the one-dimensional RELU function we may simply state that the line $x = 0$ separates the two partitions, because RELU is linear both in the region $x \leq 0$ and $x \geq 0$. Finally, neural networks are typically composed of sequences of linear and piecewise-linear layers, where the partitioning imposed by each layer individually may be well-understood but their composition is more complex. For example, identifying the linear partitions

of $y = \text{RELU}(4 \cdot \text{RELU}(-3x - 1) + 2)$ is non-trivial, even though we know the linear partitions of each composed function individually.

Our algorithm only requires the user to specify the hyperplanes defining the partitioning for the activation function used in each layer; our current implementation comes with support for common PWL activation functions. For example, if a RELU layer is used for an n -dimensional input vector, then the hyperplanes would be defined by the equations $x_1 = 0, x_2 = 0, \dots, x_n = 0$. It then computes the symbolic representation for a *single layer at a time*, composing them sequentially to compute the symbolic representation across the entire network.

To allow such compositions of layers, instead of directly computing $\widehat{f|_X}$, we will define another primitive, denoted by the operator \otimes and sometimes referred to as EXTEND, such that

$$\text{EXTEND}(h, \widehat{g}) = h \otimes \widehat{g} = \widehat{h \circ g}. \tag{1}$$

Consider $f = f_n \circ f_{n-1} \circ \dots \circ f_1$, and let $I : x \mapsto x$ be the identity map. I is linear across its entire input space, and, thus, $\widehat{I|_X} = \{X\}$. By the definition of $\text{EXTEND}(f_1, \cdot)$, we have $f_1 \otimes \widehat{I|_X} = (f_1 \circ I)|_X = \widehat{f_1|_X}$, where the final equality holds by the definition of the identity map I . We can then iteratively apply this procedure to inductively compute $(f_i \circ \dots \circ f_1)|_X$ from $(f_{i-1} \circ \dots \circ f_1)|_X$ like so:

$$f_i \otimes (f_{i-1} \circ \dots \circ f_1)|_X = (f_i \circ f_{i-1} \circ \dots \circ f_1)|_X$$

until we have computed $(f_n \circ f_{n-1} \circ \dots \circ f_1)|_X = \widehat{f|_X}$, which is the required symbolic representation.

4.1 Algorithm for Extend

Algorithm 1 present an algorithm for computing EXTEND for arbitrary PWL functions, where $\text{EXTEND}(h, \widehat{g}) = h \otimes \widehat{g} = \widehat{h \circ g}$.

Geometric intuition for the algorithm. Consider the RELU function (Definition 3). It can be shown that, within any orthant (i.e., when the signs of all coefficients are held constant), $\text{RELU}(\vec{x})$ is equivalent to some linear function, in particular the element-wise product of \vec{x} with a vector that zeroes out the negative-signed components. However, for our algorithm, all we need to know is that the linear partitions of RELU (in this case the orthants) are separated by hyperplanes $x_1 = 0, x_2 = 0, \dots, x_n = 0$.

Given a two-dimensional convex bounded polytope X , the execution of the algorithm for $f = \text{RELU}$ can be visualized as follows. We pick some vertex v of X , and begin traversing the boundary of the polytope in counter-clockwise order. If we hit an orthant boundary (corresponding to some hyperplane $x_i = 0$), it implies that the behavior of the function behaves differently at the points of the polytope to one side of the boundary from those at the other side of the boundary. Thus, we *partition X into X_1 and X_2* , where X_1 lies to one side of the hyperplane and X_2 lies to the other side. We recursively apply this

procedure to X_1 and X_2 until the resulting polytopes all lie on exactly one side of every hyperplane (orthant boundary). But lying on exactly one side of every hyperplane (orthant boundary) implies each polytope lies entirely within a linear partition of the function (a single orthant), hence the application of the function on that polytope is linear, and hence we have our partitioning.

Functions used in algorithm. Given a two-dimensional bounded convex polytope X , $\text{Vert}(X)$ returns a list of its vertices in counter-clockwise order, repeating the initial vertex at the end. Given a set of points X , $\text{ConvexHull}(X)$ represents their convex hull (the smallest bounded polytope containing every point in X). Given a scalar value x , $\text{Sign}(x)$ computes the sign of that value (i.e., -1 if $x < 0$, $+1$ if $x > 0$, and 0 if $x = 0$).

Algorithm description. The key insight of the algorithm is to recursively partition the polytopes until such a partition lies entirely within a linear region of the function f . Algorithm 1 begins by constructing a queue containing the polytopes of $\widehat{g}_{\perp X}$. Each iteration either removes a polytope from the queue that lies entirely in one linear region (placing it in Y), or splits (partitions) some polytope into two smaller polytopes that get put back into the queue. When we pop a polytope P from the queue, Line 6 iterates over all hyperplanes $N_k \cdot x = b_k$ defining the piecewise-linear partitioning of f , looking for any for which some vertex V_i lies on the positive side of the hyperplane and another vertex V_j lies on the negative side of the hyperplane. If none exist (Line 7), by convexity we are guaranteed that the entire polytope lies entirely on one side with respect to every hyperplane, meaning it lies entirely within a linear partition of f . Thus, we can add it to Y and continue. If two such vertices are found (starting Line 10), then we can find “extreme” i and j indices such that V_i is the last vertex in a counter-clockwise traversal to lie on the same side of the hyperplane as V_1 and V_j is the last vertex lying on the opposite side of the hyperplane. We then call $\text{SplitPlane}()$ (Algorithm 2) to actually partition the polytope on opposite sides of the hyperplane, adding both to our worklist.

In the best case, each partition is in a single orthant: the algorithm never calls $\text{SplitPlane}()$ at all — it merely iterates over all of the n input partitions, checks their v vertices, and appends to the resulting set (for a best-case complexity of $O(nv)$). In the worst case, it splits each polytope in the queue on each face, resulting in exponential time complexity. As we will show in Section 6, this exponential worst-case behavior is not encountered in practice, thus making SyReNN a practical tool for DNN analysis.

Please see the extended version of this paper for a worked example of the algorithm’s execution.

4.2 Representing Polytopes

We close this section with a discussion of implementation concerns when representing the convex polytopes that make up the partitioning of $\widehat{f}_{\perp X}$. In standard computational geometry, bounded polytopes can be represented in two equivalent forms:

Algorithm 1: $f \otimes \widehat{g}_{\uparrow X}$ for two-dimensional X . f is defined by hyperplanes $N_1 \cdot x = b_1$ through $N_m \cdot x = b_m$ such that, within any partition imposed by the hyperplanes f is equivalent to some affine function.

Input: $\widehat{g}_{\uparrow X} = \{P_1, \dots, P_n\}$.

Output: $f \circ \widehat{g}_{\uparrow X}$

```

1  $W \leftarrow \text{ConstructQueue}(\widehat{g}_{\uparrow X})$ 
2  $Y \leftarrow \emptyset$ 
3 while  $W$  not empty do
4    $P \leftarrow \text{Pop}(W)$ 
5    $V \leftarrow \text{Vert}(P)$ 
6    $K \leftarrow \{N_k \mid \exists i, j : \text{Sign}(N_k \cdot g(V_i) - b_k) > 0 \wedge \text{Sign}(N_k \cdot g(V_j) - b_k) < 0\}$ 
7   if  $K = \emptyset$  then
8      $Y \leftarrow Y \cup \{P\}$ 
9     continue
10   $N, b \leftarrow$  any element from  $K$ 
11   $i \leftarrow \arg \max_i \{\text{Sign}(N \cdot g(V_i) - b) = \text{Sign}(N \cdot g(V_1) - b)\}$ 
12   $j \leftarrow \arg \max_j \{\text{Sign}(N \cdot g(V_j) - b) \neq \text{Sign}(N \cdot g(V_i) - b)\}$ 
13  for  $V' \in \text{SplitPlane}(V, g, i, j, N, b)$  do
14     $W \leftarrow \text{Push}(W, \text{ConvexHull}(V'))$ 
15 return  $Y$ 

```

1. The *half-space* or *H-representation*, which encodes the polytope as an intersection of finitely-many half-spaces. (Each half-space being defined as a halfspace defined by an affine inequality $Ax \leq b$.)
2. The *vertex* or *V-representation*, which encodes the polytope as a set of finitely many points; the polytope is then taken to be the convex hull of the points (i.e., smallest convex shape containing all of the points).

Certain operations are more efficient when using one representation compared to the other. For example, finding the intersection of two polytopes in an H-representation can be done in linear time by concatenating their representative half-spaces, but the same is not possible in V-representation.

There are two main operations on polytopes we need perform in our algorithms: (i) splitting a polytope with a hyperplane, and (ii) applying an affine map to all points in the polytope. In general, the first is more efficient in an H-representation, while the latter is more efficient in a V-representation. However, when restricted to two-dimensional polygons, the former is also efficient in a V-representation, as demonstrated by Algorithm 2, helping to motivate our use of the V-representation in our algorithm.

Furthermore, the two polytope representations have different resiliency to floating-point operations. In particular, H-representations for polytopes in \mathbb{R}^n are notoriously difficult to achieve high-precision with, because the error introduced from using floating point numbers gets arbitrarily large as one goes in a particular direction along any hyperplane face. Ideally, we would like the

Algorithm 2: SplitPlane(V, g, i, j, N, b)

Input: V , the vertices of the polytope in the input space of g . A function g . i is the index of the last vertex lying on the same side of the orthant face as V_1 . j is the index of the last vertex lying on the opposite side of the orthant face as V_1 . N and b define the hyperplane $N \cdot x = b$ to split on.

Output: $\{P_1, P_2\}$, two sets of vertices whose convex hulls form a partitioning of V such that each lies on only one side of the $N \cdot x = b$ hyperplane.

- 1 $p_i \leftarrow V_i + \frac{b - N \cdot g(V_i)}{N \cdot (g(V_{i+1}) - g(V_i))} (V_{i+1} - V_i)$
 - 2 $p_j \leftarrow V_j + \frac{b - N \cdot g(V_j)}{N \cdot (g(V_{j+1}) - g(V_j))} (V_{j+1} - V_j)$
 - 3 $A \leftarrow \{p_i, p_j\} \cup \{v \in V \mid \text{Sign}(N \cdot v - b) = \text{Sign}(N \cdot V_i - b)\}$
 - 4 $B \leftarrow \{p_i, p_j\} \cup \{v \in V \mid \text{Sign}(N \cdot v - b) = \text{Sign}(N \cdot V_j - b)\}$
 - 5 **return** $\{A, B\}$
-

hyperplane to be most accurate in the region of the polytope itself, which corresponds to choosing the magnitude of the norm vector correctly. Unfortunately, to our knowledge, there is no efficient algorithm for computing the ideal floating point H-representation of a polytope, although libraries such as APRON [30] are able to provide reasonable results for low-dimensional spaces. However, because neural networks utilize extremely high-dimensional spaces (often hundreds or thousands of dimensions) and we wish to iteratively apply our analysis, we find that errors from using floating-point H-representations can quickly multiply and compound to become infeasible. By contrast, floating-point inaccuracies in a V-representation are directly interpretable as slightly misplacing the vertices of the polytope; no “localization” process is necessary to penalize inaccuracies close to the polytope more than those far away from it.

Another difference is in the space complexity of the representation. In general, H-representations can be more space-efficient for common shapes than V-representations. However, when the polytope lies in a low-dimensional subspace of a larger space, the V-representation is usually significantly more efficient.

Thus, V-representations are a good choice for low-dimensionality polytopes embedded in high-dimensional space, which is exactly what we need for analyzing neural networks with two-dimensional restriction domains of interest. This is why we designed our algorithms to rely on Vert(X), so that they could be directly computed on a V-representation.

The 2D algorithm described above can be seen as implementing the recursive case of a more general, n -dimensional version of the algorithm that recurses on each of the $(n - 1)$ -dimensional facets. Please see the extended version of this paper for more details.

5 SyReNN tool

This section provides more details about the design and implementation of our tool, SyReNN (Symbolic Representations of Neural Networks), which computes

$\widehat{f}_{\uparrow X}$, where f is a DNN using only piecewise-linear layers and X is a union of one- or two-dimensional polytopes. The tool is available under the MIT license at <https://github.com/95616ARG/SyReNN> and in the PyPI package `pysyrenn`.

Input and output format. SyReNN supports reading DNNs from two standard formats: ERAN (a textual format used by the ERAN project [1]) as well as ONNX (an industry-standard format supporting a wide variety of different models) [42]. Internally, the input DNN is described as an instance of the `Network` class, which is itself a list of sequential `Layers`. A number of layer types are provided by SyReNN, including `FullyConnectedLayer`, `ConvolutionalLayer`, and `ReLULayer`. To support more complicated DNN architectures, we have implemented a `ConcatLayer`, which represents a concatenation of the output of two different layers. The input region of interest, X , is defined as a polytope described by a list of its vertices in counter-clockwise order. The output of the tool is the symbolic representation $\widehat{f}_{\uparrow X}$.

Overall Architecture. We designed SyReNN in a client-server architecture using gRPC [20] and protocol buffers [21] as a standard method of communication between the two. This architecture allows the bulk of the heavy computation to be done in efficient C++ code, while allowing user-friendly interfaces in a variety of languages. It also allows practitioners to run the server remotely on a more powerful machine if necessary. The C++ server implementation uses the Intel TBB library for parallelization. Our official front-end library is written in Python, and available as a package on PyPI so installation is as simple as `pip install pysyrenn`. The entire project can be built using the Bazel build system, which manages dependencies using checksums.

Server Architecture. The major algorithms are implemented as a gRPC server written in C++. When a connection is first made, the server initializes the state with an empty DNN $f(x) = x$. During the session, three operations are permitted: (i) append a layer g so that the current session’s DNN is updated from f_0 to $f_1(x) := g(f_0(x))$, (ii) compute $\widehat{f}_{\uparrow X}$ for a one-dimensional X , or (iii) compute $\widehat{f}_{\uparrow X}$ for a two-dimensional X . We have separate methods for one- and two-dimensional X , because the one-dimensional case has specific optimizations for controlling memory usage. The `SegmentedLine` and `UPolytope` types are used to represent one- and two-dimensional partitions of X , respectively. When operation (1) is performed, a new instance of the `LayerTransformer` class is initialized with the relevant parameters and added to a running `vector` of the current layers. When operation (2) is performed, a new queue of `SegmentedLines` is constructed, corresponding to X , and the before-allocated `LayerTransformers` are applied sequentially to compute $\widehat{f}_{\uparrow X}$. In this case, extra control is provided to automatically gauge memory usage and pause computation for portions of X until more memory is made available. Finally, when operation (3) is performed, a new instance of `UPolytope` is initialized with the vertices of X and the `LayerTransformers` are again applied sequentially to compute $\widehat{f}_{\uparrow X}$.

Client Architecture. Our Python client exposes an interface for defining DNNs similar to the popular Sequential-Network Keras API [11]. Objects repre-

sent individual layers in the network, and they can be combined sequentially into a `Network` instance. The key addition of our library is that this `Network` exposes methods for computing $\widehat{f}_{\uparrow X}$ given a V-representation description of X . To do this, it invokes the server and passes a layer-by-layer description of f followed by the polytope X , then parses the response $\widehat{f}_{\uparrow X}$.

Extending to support different layer types. Different layer types are supported by sub-classing the `LayerTransformer` class. Instances of this class expose a method for computing `EXTEND(h, \cdot)` for the corresponding layer h . To simplify implementation, two sub-classes of `LayerTransformer` are provided: one for entirely-linear layers (such as fully-connected and convolutional layers), and one for piecewise-linear layers. For fully-linear layers, all that needs to be provided is a method computing the layer function itself. For piecewise-linear layers, two methods need to be provided: (1) computing the layer function itself, and (2) one describing the hyperplanes which separate the linear regions. The base class then directly implements Algorithm 1 for that layer. This architecture makes supporting new layers a straight-forward process.

Float Safety. Like Reluplex [32], SyReNN uses floating-point arithmetic to compute $\widehat{f}_{\uparrow X}$ efficiently. Unfortunately, this means that in some cases its results will not be entirely precise when compared to a real-valued or multiple-precision version of the algorithm. Approaches for addressing this are discussed in the extended version of this paper.

6 Applications of SyReNN

This section presents the use of SyReNN in three example case studies.

6.1 Integrated Gradients

A common problem in the field of *explainable machine learning* is understanding *why* a DNN made the prediction it did. For example, given an image classified by a DNN as a ‘cat,’ why did the DNN decide it was a cat instead of, say, a dog? Were there particular pixels which were particularly important in deciding this? Integrated Gradients (IG) [52] is the state-of-the-art method for computing such *model attributions*.

Definition 5. Given a DNN f , the integrated gradients along dimension i for input x and baseline x' is defined to be:

$$IG_i(x) \stackrel{\text{def}}{=} (x_i - x'_i) \times \int_{\alpha=0}^1 \frac{\partial f(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha. \quad (2)$$

The computed value $IG_i(x)$ determines relatively how important the i th input (e.g., pixel) was to the classification.

However, exactly computing this integral requires a symbolic, closed form for the gradient of the network. Until [50], it was not known how to compute

such a closed-form and so IGs were always only *approximated* using a sampling-based approach. Unfortunately, because it was unknown how to compute the true value, there was no way for practitioners to determine how accurate their approximations were. This is particularly concerning in fairness applications where an accurate attribution is exceedingly important.

In [50], it was recognized that, when $X = \text{ConvexHull}(\{x, x'\})$, $\widehat{f}_{|X}$ can be used to *exactly* compute $IG_i(x)$. This is because within each partition of $\widehat{f}_{|X}$ the gradient of the network is *constant* because it behaves as a linear function, and hence the integral can be written as the weighted sum of such finitely-many gradients.¹ Using our symbolic representation, the exact IG can thus be computed as follows:

$$\sum_{\text{ConvexHull}(\{y_i, y'_i\}) \in \widehat{f}_{| \text{ConvexHull}(\{x, x'\})}} (y'_i - y_i) \times \frac{\partial f(0.5 \times (y_i + y'_i))}{\partial x_i} \quad (3)$$

Where here y_i, y'_i are the endpoints of the segment with y_i closer to x and y'_i closest to x' .

Implementation. The helper class `IntegratedGradientsHelper` is provided by our Python client library. It takes as input a DNN f and a set of (x, x') input-baseline pairs and then computes IG for each pair.

Empirical Results. In [50] SyReNN was used to show conclusively that existing sampling-based methods were insufficient to adequately approximate the true IG. This realization led to changes in the official IG implementation to use the more-precise trapezoidal sampling method we argued for.

Timing Numbers. In those experiments, we used SyReNN to compute $\widehat{f}_{|X}$ for three different DNNs f , namely the small, medium, and large convolutional models from [1]. For each DNN, we ran SyReNN on 100 one-dimensional lines. The 100 calls to SyReNN completed in 20.8 seconds for the small model, 183.3 for the medium model, and 615.5 for the big model. Tests were performed on an Intel Core i7-7820X CPU at 3.60GHz with 32GB of memory.

6.2 Visualization of DNN Decision Boundaries

Whereas IG helps understand why a DNN made a particular prediction about a single input point, another major task is *visualizing* the decision boundaries of a DNN on *infinitely-many* input points. Figure 2 shows a visualization of an ACAS Xu DNN [31] which takes as input the position of an airplane and an approaching attacker, then produces as output one of five advisories instructing the plane, such as “clear of conflict” or to move “weak left.” Every point in the diagram represents the relative position of the approaching plane, while the color indicates the advisory.

¹ As noted in [50], this technically requires a slight strengthening of the definition of $\widehat{f}_{|X}$ which is satisfied by our algorithms as defined above.

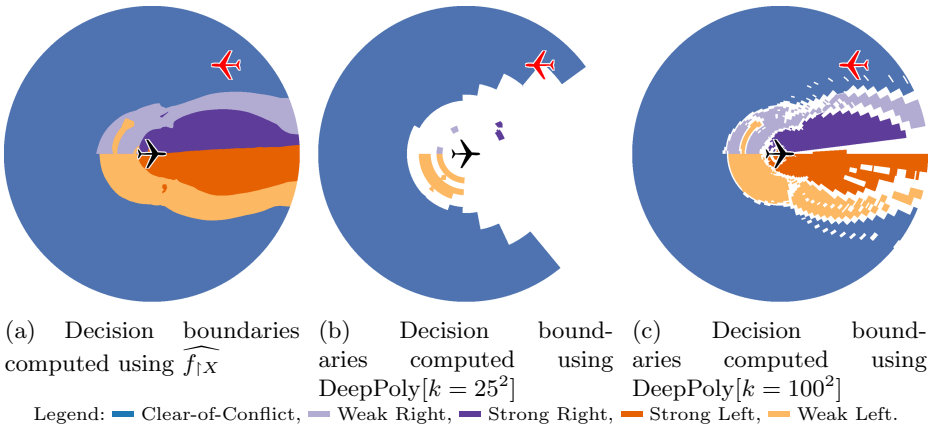


Fig. 2: Visualization of decision boundaries for the ACAS Xu network. Using SyReNN (left) quickly produces the exact decision boundaries. Using abstract interpretation-based tools like DeepPoly (middle and right) are slower and produce only imprecise approximations of the decision boundaries.

One approach to such visualizations is to simply sample finitely-many points and extrapolate the behavior on the entire domain from those finitely-many points. However, this approach is imprecise and risks missing vital information because there is no way to know the correct sampling density to use to identify all important features.

Another approach is to use a tool such as DeepPoly [49] to over-approximate the output range of the DNN. However, because DeepPoly is a relatively coarse over-approximation, there may be regions of the input space for which it cannot state with confidence the decision made by the network. In fact, the approximations used by DeepPoly are extremely coarse. A naïve application of DeepPoly to this problem results in it being unable to make claims about *any* of the input space of interest. In order to utilize it, we must *partition* the space and run DeepPoly within each partition, which significantly slows down the analysis. Even when using 25^2 partitions, Figure 2b shows that most of the interesting region is still unclassifiable with DeepPoly (shown in white). Only with 100^2 partitions can DeepPoly effectively approximate the decision boundaries, although it is still quite imprecise.

By contrast, \widehat{f}_{1X} can be used to *exactly* determine the decision boundaries on any 2D polytope subset of the input space, which can then be plotted. This is shown in Figure 2a. Furthermore, as shown in Table 1, the approach using \widehat{f}_{1X} is *significantly* faster than that using ERAN, even as we get the *precise* answer instead of an approximation. Such visualizations can be particularly helpful in identifying issues to be fixed using techniques such as those in Section 6.3.

Table 1: Comparing the performance of DNN visualization using SyReNN versus DeepPoly for the ACAS Xu network [31]. $\widehat{f}_{\uparrow X}$ size is the number of partitions in the symbolic representation. SyReNN time is the time taken to compute $\widehat{f}_{\uparrow X}$ using SyReNN. DeepPoly[k] time is the time taken to compute DeepPoly for approximating decision boundaries with k partitions. Each scenario represents a different two-dimensional slice of the input space; within each slice, the heading of the intruder relative to the ownship along with the speed of each involved plane is fixed.

Scenario	$\widehat{f}_{\uparrow X}$ size	SyReNN time (secs)	DeepPoly time (secs)		
			$k = 25^2$	$k = 55^2$	$k = 100^2$
Head-On, Slow	33200	10.9	9.1	43.2	141.3
Head-On, Fast	30769	10.2	8.2	39.0	128.0
Perpendicular, Slow	37251	12.5	9.2	42.9	141.7
Perpendicular, Fast	33931	11.4	8.2	39.2	127.5
Opposite, Slow	36743	12.1	9.8	46.7	152.5
Opposite, Fast	38965	13.0	9.5	45.2	147.3
-Perpendicular, Slow	36037	11.9	9.5	45.0	146.4
-Perpendicular, Fast	33208	10.9	8.3	39.5	130.2

Implementation. The helper class `PlanesClassifier` is provided by our Python client library. It takes as input a DNN f and an input region X , then computes the decision boundaries of f on X .

Timing Numbers. Timing comparisons are given in Table 1. We see that SyReNN is quite performant, and the exact SyReNN can be computed more quickly than even a mediocre approximation from DeepPoly using 55^2 partitions. Tests were performed on a dedicated Amazon EC2 c5.metal instance, using BenchExec [5] to limit the number of CPU cores to 16 and RAM to 16GB.

6.3 Patching of DNNs

We have now seen how SyReNN can be used to visualize the behavior of a DNN. This can be particularly useful for identifying buggy behavior. For example, in Figure 2a we can see that the decision boundary between “strong right” and “strong left” is not symmetrical.

The final application we consider for SyReNN is *patching DNNs* to correct undesired behavior. Patching is described formally in [51]. Given an initial network N and a *specification* ϕ describing desired constraints on the input/output, the goal of patching is to find a small modification to the parameters of N producing a new DNN N' that satisfies the constraints in ϕ .

The key theory behind DNN patching we will use was developed in [51]. The key realization of that work is that, for a certain DNN architecture, correcting the network behavior on an infinite, 2D region X is exactly equivalent to correcting

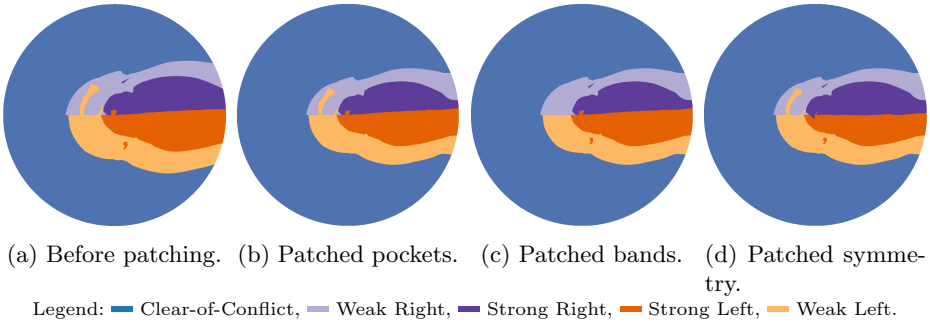


Fig. 3: Network patching.

its behavior on *the finitely-many vertices* $\text{Vert}(P_i)$ for each of the finitely-many $P_i \in \widehat{f_{\uparrow X}}$. Hence, SyReNN plays a key role in enabling efficient DNN patching.

For this case study, we patched the same aircraft collision-avoidance DNN visualized in Section 6.2. We patched the DNN three times to correct three different buggy behaviors of the network: (i) remove “Pockets” of strong left/strong right in regions that are otherwise weak left/weak right; (ii) remove the “Bands” of weak-left advisory behind and to the left of the plane; and (iii) enforce “Symmetry” across the horizontal. The DNNs before and after patching with different specifications are shown in Figure 3.

Implementation The helper class `NetPatcher` is provided by our Python client library. It takes as input a DNN f and pairs of input region, output label X_i, Y_i , then computes a new DNN f' which maps all points in each X_i into Y_i .

Timing Numbers. As in Section 6.2, computing $\widehat{f_{\uparrow X}}$ for use in patching took approximately 10 seconds.

7 Related Work

The related problem of exact reach set analysis for DNNs was investigated in [58]. However, the authors use an algorithm that relies on explicitly enumerating all exponentially-many (2^n) possible signs at each RELU layer. By contrast, our algorithm adapts to the actual input polytopes, efficiently restricting its consideration to activations that are actually possible.

Hanin and Rolnick [25] prove theoretical properties about the cardinality of $\widehat{f_{\uparrow X}}$ for RELU networks, showing that $|\widehat{f_{\uparrow X}}|$ is expected to grow polynomially with the number of nodes in the network for randomly-initialized networks.

Thrun [55] and Bastani et al.[4] extract symbolic rules meant to approximate DNNs, which can approximate the symbolic representation $\widehat{f_{\uparrow X}}$.

In particular, the ERAN [1] tool and underlying DeepPoly [49] domain were designed to verify the non-existence of adversarial examples. Breutel et al. [6] give an iterative refinement algorithm for an overapproximation of the weakest precondition as a polytope where the required output is also a polytope.

Scheibler et al. [46] verify the safety of a machine-learning controller using the SMT-solver iSAT3, but support small unrolling depths and basic safety properties. Zhu et al. [60] use a synthesis procedure to generate a safe deterministic program that can enforce safety conditions by monitoring the deployed DNN and preventing potentially unsafe actions. The presence of adversarial and fooling inputs for DNNs as well as applications of DNNs in safety-critical systems has led to efforts to verify and certify DNNs [3,32,14,29,16,7,57,49,2]. *Approximate reachability analysis* for neural networks safely overapproximates the set of possible outputs [16,58,59,57,13,56].

Prior work in the area of network patching focuses on enforcing constraints on the network during training. DiffAI [39] is an approach to train neural networks that are certifiably robust to adversarial perturbations. DL2 [15] allows for training and querying neural networks with logical constraints.

8 Conclusion and Future Work

We presented SyReNN, a tool for understanding and analyzing DNNs. Given a piecewise-linear network and a low-dimensional polytope subspace of the input subspace, SyReNN computes a symbolic representation that decomposes the behavior of the DNN into finitely-many linear functions. We showed how to efficiently compute this representation, and presented the design of the corresponding tool. We illustrated the utility of SyReNN on three applications: computing exact IG, visualizing the behavior of DNNs, and patching (repairing) DNNs.

In contrast to prior work, SyReNN explores a unique point in the design space of DNN analysis tools. Instead of trading off precision of the analysis for efficiency, SyReNN focuses on analyzing DNN behavior on *low-dimensional subspaces* of the domain, for which we can provide *both* efficiency and precision.

We plan on extending SyReNN to make use of GPUs and other massively-parallel hardware to more quickly compute $\widehat{f_{\downarrow X}}$ for large f or X . Techniques to support input polytopes that are greater than two dimensional is also a ripe area of future work. We may also be able to take advantage of the fact that non-convex polytopes can be represented efficiently in 2D. Extending algorithms for $\widehat{f_{\downarrow X}}$ to handle architectures such as Recurrent Neural Networks (RNNs) will open up new application areas for SyReNN.

Acknowledgements. We thank the anonymous reviewers for their feedback and suggestions on this work. This material is based upon work supported by a Facebook Probability and Programming award.

References

1. ETH robustness analyzer for neural networks (ERAN). <https://github.com/eth-sri/eran> (2019), accessed: 2019-05-01

2. Anderson, G., Pailoor, S., Dillig, I., Chaudhuri, S.: Optimization and abstraction: a synergistic approach for analyzing neural network robustness. In: McKinley, K.S., Fisher, K. (eds.) *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation, PLDI 2019, Phoenix, AZ, USA, June 22-26, 2019*. pp. 731–744. ACM (2019). <https://doi.org/10.1145/3314221.3314614>, <https://doi.org/10.1145/3314221.3314614>
3. Bastani, O., Ioannou, Y., Lampropoulos, L., Vytiniotis, D., Nori, A.V., Criminisi, A.: Measuring neural net robustness with constraints. In: Lee, D.D., Sugiyama, M., von Luxburg, U., Guyon, I., Garnett, R. (eds.) *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. pp. 2613–2621 (2016), <http://papers.nips.cc/paper/6339-measuring-neural-net-robustness-with-constraints>
4. Bastani, O., Pu, Y., Solar-Lezama, A.: Verifiable reinforcement learning via policy extraction. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*. pp. 2499–2509 (2018), <http://papers.nips.cc/paper/7516-verifiable-reinforcement-learning-via-policy-extraction>
5. Beyer, D.: Reliable and reproducible competition results with benchexec and witnesses (report on SV-COMP 2016). In: Chechik, M., Raskin, J. (eds.) *Tools and Algorithms for the Construction and Analysis of Systems - 22nd International Conference, TACAS 2016, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2016, Eindhoven, The Netherlands, April 2-8, 2016, Proceedings. Lecture Notes in Computer Science, vol. 9636*, pp. 887–904. Springer (2016). https://doi.org/10.1007/978-3-662-49674-9_55, https://doi.org/10.1007/978-3-662-49674-9_55
6. Breutel, S., Maire, F., Hayward, R.: Extracting interface assertions from neural networks in polyhedral format. In: *ESANN 2003, 11th European Symposium on Artificial Neural Networks, Bruges, Belgium, April 23-25, 2003, Proceedings*. pp. 463–468 (2003), <https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2003-72.pdf>
7. Bunel, R., Turkaslan, I., Torr, P.H.S., Kohli, P., Mudigonda, P.K.: A unified view of piecewise linear neural network verification. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*. pp. 4795–4804 (2018), <http://papers.nips.cc/paper/7728-a-unified-view-of-piecewise-linear-neural-network-verification>
8. Carlini, N., Wagner, D.A.: Audio adversarial examples: Targeted attacks on speech-to-text. In: *2018 IEEE Security and Privacy Workshops, SP Workshops 2018, San Francisco, CA, USA, May 24, 2018*. pp. 1–7. IEEE Computer Society (2018). <https://doi.org/10.1109/SPW.2018.00009>, <https://doi.org/10.1109/SPW.2018.00009>
9. Chen, J., Ran, X.: Deep learning with edge computing: A review. *Proc. IEEE* **107**(8), 1655–1674 (2019). <https://doi.org/10.1109/JPROC.2019.2921977>, <https://doi.org/10.1109/JPROC.2019.2921977>
10. Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., Kalinin, A.A., Do, B.T., Way, G.P., Ferrero, E., Agapow, P.M., Zietz, M., Hoffman, M.M., Xie, W., Rosen, G.L., Lengerich, B.J., Israeli, J., Lanchantin, J., Woloszynek, S., Carpenter, A.E., Shrikumar, A., Xu, J., Cofer, E.M., Lavender, C.A., Turaga, S.C., Alexandari, A.M., Lu, Z., Harris, D.J., DeCaprio, D., Qi, Y., Kundaje, A., Peng, Y., Wiley, L.K., Segler,

- M.H.S., Boca, S.M., Swamidass, S.J., Huang, A., Gitter, A., Greene, C.S.: Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface* **15**(141), 20170387 (2018). <https://doi.org/10.1098/rsif.2017.0387>
11. Chollet, F., et al.: Keras. <https://keras.io> (2015)
 12. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. pp. 4171–4186. Association for Computational Linguistics (2019). <https://doi.org/10.18653/v1/n19-1423>, <https://doi.org/10.18653/v1/n19-1423>
 13. Dutta, S., Jha, S., Sankaranarayanan, S., Tiwari, A.: Output range analysis for deep feedforward neural networks. In: Dutle, A., Muñoz, C.A., Narkawicz, A. (eds.) *NASA Formal Methods - 10th International Symposium, NFM 2018, Newport News, VA, USA, April 17-19, 2018, Proceedings. Lecture Notes in Computer Science*, vol. 10811, pp. 121–138. Springer (2018). https://doi.org/10.1007/978-3-319-77935-5_9, https://doi.org/10.1007/978-3-319-77935-5_9
 14. Ehlers, R.: Formal verification of piece-wise linear feed-forward neural networks. In: D’Souza, D., Kumar, K.N. (eds.) *Automated Technology for Verification and Analysis - 15th International Symposium, ATVA 2017, Pune, India, October 3-6, 2017, Proceedings. Lecture Notes in Computer Science*, vol. 10482, pp. 269–286. Springer (2017). https://doi.org/10.1007/978-3-319-68167-2_19, https://doi.org/10.1007/978-3-319-68167-2_19
 15. Fischer, M., Balunovic, M., Drachler-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>
 16. Gehr, T., Mirman, M., Drachler-Cohen, D., Tsankov, P., Chaudhuri, S., Vechev, M.T.: AI2: safety and robustness certification of neural networks with abstract interpretation. In: *2018 IEEE Symposium on Security and Privacy, SP 2018, Proceedings, 21-23 May 2018, San Francisco, California, USA*. pp. 3–18. IEEE Computer Society (2018). <https://doi.org/10.1109/SP.2018.00058>, <https://doi.org/10.1109/SP.2018.00058>
 17. Gonzales, R.: Feds say self-driving uber suv did not recognize jaywalking pedestrian in fatal crash. NPR <https://www.npr.org/2019/11/07/777438412/feds-say-self-driving-uber-suv-did-not-recognize-jaywalking-pedestrian-in-fatal> (Nov 2019), accessed: 2020-06-06
 18. Goodfellow, I., Bengio, Y., Courville, A.: *Deep Learning*. MIT Press (2016), <http://www.deeplearningbook.org>
 19. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: Bengio, Y., LeCun, Y. (eds.) *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (2015), <http://arxiv.org/abs/1412.6572>
 20. Google: grpc: A high-performance, open source universal rpc framework. ”<https://grpc.io/> (2020)
 21. Google: Protocol buffers - google’s data interchange format. <https://developers.google.com/protocol-buffers/> (2020)

22. Gopinath, S., Ghanathe, N., Seshadri, V., Sharma, R.: Compiling kb-sized machine learning models to tiny iot devices. In: McKinley, K.S., Fisher, K. (eds.) Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation, PLDI 2019, Phoenix, AZ, USA, June 22-26, 2019. pp. 79–95. ACM (2019). <https://doi.org/10.1145/3314221.3314597>, <https://doi.org/10.1145/3314221.3314597>
23. Guennebaud, G., Jacob, B., et al.: Eigen v3. <http://eigen.tuxfamily.org> (2010)
24. Han, S., Mao, H., Dally, W.J.: Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding. In: Bengio, Y., LeCun, Y. (eds.) 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings (2016), <http://arxiv.org/abs/1510.00149>
25. Hanin, B., Rolnick, D.: Complexity of linear regions in deep networks. 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. 2596–2604. PMLR (2019), <http://proceedings.mlr.press/v97/hanin19a.html>
26. Hern, A.: Facebook translates 'good morning' into 'attack them', leading to arrest. <https://www.theguardian.com/technology/2017/oct/24/facebook-palestine-israel-translates-good-morning-attack-them-arrest> (Jun 2017), accessed: 2020-06-06
27. Hill, K.: Wrongfully accused by an algorithm. New York Times. <https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html> (Jun 2020), accessed: 2020-06-06
28. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L.H., Aerts, H.J.: Artificial intelligence in radiology. *Nature Reviews Cancer* p. 1 (2018)
29. Huang, X., Kwiatkowska, M., Wang, S., Wu, M.: Safety verification of deep neural networks. In: Majumdar, R., Kuncak, V. (eds.) Computer Aided Verification - 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017, Proceedings, Part I. Lecture Notes in Computer Science, vol. 10426, pp. 3–29. Springer (2017). https://doi.org/10.1007/978-3-319-63387-9_1, https://doi.org/10.1007/978-3-319-63387-9_1
30. Jeannet, B., Miné, A.: Apron: A library of numerical abstract domains for static analysis. In: Bouajjani, A., Maler, O. (eds.) Computer Aided Verification, 21st International Conference, CAV 2009, Grenoble, France, June 26 - July 2, 2009. Proceedings. Lecture Notes in Computer Science, vol. 5643, pp. 661–667. Springer (2009). https://doi.org/10.1007/978-3-642-02658-4_52, https://doi.org/10.1007/978-3-642-02658-4_52
31. Julian, K.D., Kochenderfer, M.J., Owen, M.P.: Deep neural network compression for aircraft collision avoidance systems. *CoRR* **abs/1810.04240** (2018), <http://arxiv.org/abs/1810.04240>
32. Katz, G., Barrett, C.W., Dill, D.L., Julian, K., Kochenderfer, M.J.: Reluplex: An efficient SMT solver for verifying deep neural networks. In: Majumdar, R., Kuncak, V. (eds.) Computer Aided Verification - 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017, Proceedings, Part I. Lecture Notes in Computer Science, vol. 10426, pp. 97–117. Springer (2017). https://doi.org/10.1007/978-3-319-63387-9_5, https://doi.org/10.1007/978-3-319-63387-9_5
33. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Bartlett, P.L., Pereira, F.C.N., Burges, C.J.C.,

- Bottou, L., Weinberger, K.Q. (eds.) *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012*. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States. pp. 1106–1114 (2012), <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks>
34. Kumar, A., Seshadri, V., Sharma, R.: Shiftry: RNN inference in 2kb of RAM. *Proc. ACM Program. Lang.* **4**(OOPSLA), 182:1–182:30 (2020). <https://doi.org/10.1145/3428250>, <https://doi.org/10.1145/3428250>
 35. Kusupati, A., Singh, M., Bhatia, K., Kumar, A., Jain, P., Varma, M.: Fastgrnn: A fast, accurate, stable and tiny kilobyte sized gated recurrent neural network. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, December 3-8, 2018, Montréal, Canada. pp. 9031–9042 (2018)
 36. Lee, D.: US opens investigation into Tesla after fatal crash. BBC. <https://www.bbc.co.uk/news/technology-36680043> (Jul 2016), accessed: 2020-06-06
 37. Mendelson, E.B.: Artificial intelligence in breast imaging: potentials and limitations. *American Journal of Roentgenology* **212**(2), 293–299 (2019)
 38. Miotto, R., Wang, F., Wang, S., Jiang, X., Dudley, J.T.: Deep learning for healthcare: review, opportunities and challenges. *Briefings Bioinform.* **19**(6), 1236–1246 (2018). <https://doi.org/10.1093/bib/bbx044>, <https://doi.org/10.1093/bib/bbx044>
 39. Mirman, M., Gehr, T., Vechev, M.T.: Differentiable abstract interpretation for provably robust neural networks. In: Dy, J.G., Krause, A. (eds.) *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholm, Sweden, July 10-15, 2018*. Proceedings of Machine Learning Research, vol. 80, pp. 3575–3583. PMLR (2018), <http://proceedings.mlr.press/v80/mirman18b.html>
 40. Moosavi-Dezfooli, S., Fawzi, A., Frossard, P.: Deepfool: A simple and accurate method to fool deep neural networks. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. pp. 2574–2582. IEEE Computer Society (2016). <https://doi.org/10.1109/CVPR.2016.282>, <https://doi.org/10.1109/CVPR.2016.282>
 41. Nguyen, A.M., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. pp. 427–436. IEEE Computer Society (2015). <https://doi.org/10.1109/CVPR.2015.7298640>, <https://doi.org/10.1109/CVPR.2015.7298640>
 42. ONNX: Open neural network exchange. <https://onnx.ai/> (2020)
 43. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
 44. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d’Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, December 8-14, 2019, Vancouver, BC, Canada. pp. 8024–8035 (2019), <http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library>

45. Reinders, J.: Intel threading building blocks: outfitting C++ for multi-core processor parallelism. " O'Reilly Media, Inc." (2007)
46. Scheibler, K., Winterer, L., Wimmer, R., Becker, B.: Towards verification of artificial neural networks. In: Heinkel, U., Kriesten, D., Rößler, M. (eds.) Methoden und Beschreibungssprachen zur Modellierung und Verifikation von Schaltungen und Systemen, MBMV 2015, Chemnitz, Germany, March 3-4, 2015. pp. 30–40. Sächsische Landesbibliothek (2015)
47. Sharma, H., Park, J., Mahajan, D., Amaro, E., Kim, J.K., Shao, C., Mishra, A., Esmailzadeh, H.: From high-level deep neural models to fpgas. In: 49th Annual IEEE/ACM International Symposium on Microarchitecture, MICRO 2016, Taipei, Taiwan, October 15-19, 2016. pp. 17:1–17:12. IEEE Computer Society (2016). <https://doi.org/10.1109/MICRO.2016.7783720>, <https://doi.org/10.1109/MICRO.2016.7783720>
48. Singh, G., Gehr, T., Mirman, M., Püschel, M., Vechev, M.T.: Fast and effective robustness certification. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada. pp. 10825–10836 (2018), <http://papers.nips.cc/paper/8278-fast-and-effective-robustness-certification>
49. Singh, G., Gehr, T., Püschel, M., Vechev, M.T.: An abstract domain for certifying neural networks. Proc. ACM Program. Lang. **3**(POPL), 41:1–41:30 (2019). <https://doi.org/10.1145/3290354>, <https://doi.org/10.1145/3290354>
50. Sotoudeh, M., Thakur, A.V.: Computing linear restrictions of neural networks. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. pp. 14132–14143 (2019), <http://papers.nips.cc/paper/9562-computing-linear-restrictions-of-neural-networks>
51. Sotoudeh, M., Thakur, A.V.: Correcting deep neural networks with small, generalizing patches. In: Workshop on Safety and Robustness in Decision Making (2019)
52. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. Proceedings of Machine Learning Research, vol. 70, pp. 3319–3328. PMLR (2017), <http://proceedings.mlr.press/v70/sundararajan17a.html>
53. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 2818–2826. IEEE Computer Society (2016). <https://doi.org/10.1109/CVPR.2016.308>, <https://doi.org/10.1109/CVPR.2016.308>
54. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. In: Bengio, Y., LeCun, Y. (eds.) 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings (2014), <http://arxiv.org/abs/1312.6199>
55. Thrun, S.: Extracting rules from artificial neural networks with distributed representations. In: Tesauro, G., Touretzky, D.S., Leen, T.K. (eds.) Advances in Neural Information Processing Systems 7, [NIPS Conference, Denver, Colorado, USA, 1994]. pp. 505–512. MIT Press (1994)

56. Wang, S., Pei, K., Whitehouse, J., Yang, J., Jana, S.: Formal security analysis of neural networks using symbolic intervals. In: Enck, W., Felt, A.P. (eds.) 27th USENIX Security Symposium, USENIX Security 2018, Baltimore, MD, USA, August 15-17, 2018. pp. 1599–1614. USENIX Association (2018), <https://www.usenix.org/conference/usenixsecurity18/presentation/wang-shiqi>
57. Weng, T., Zhang, H., Chen, H., Song, Z., Hsieh, C., Daniel, L., Boning, D.S., Dhillon, I.S.: Towards fast computation of certified robustness for relu networks. In: Dy, J.G., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholm, Sweden, July 10-15, 2018. Proceedings of Machine Learning Research, vol. 80, pp. 5273–5282. PMLR (2018), <http://proceedings.mlr.press/v80/weng18a.html>
58. Xiang, W., Tran, H., Johnson, T.T.: Reachable set computation and safety verification for neural networks with relu activations. CoRR **abs/1712.08163** (2017), <http://arxiv.org/abs/1712.08163>
59. Xiang, W., Tran, H., Rosenfeld, J.A., Johnson, T.T.: Reachable set estimation and safety verification for piecewise linear systems with neural network controllers. In: 2018 Annual American Control Conference, ACC 2018, Milwaukee, WI, USA, June 27-29, 2018. pp. 1574–1579. IEEE (2018). <https://doi.org/10.23919/ACC.2018.8431048>, <https://doi.org/10.23919/ACC.2018.8431048>
60. Zhu, H., Xiong, Z., Magill, S., Jagannathan, S.: An inductive synthesis framework for verifiable reinforcement learning. In: McKinley, K.S., Fisher, K. (eds.) Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation, PLDI 2019, Phoenix, AZ, USA, June 22-26, 2019. pp. 686–701. ACM (2019). <https://doi.org/10.1145/3314221.3314638>, <https://doi.org/10.1145/3314221.3314638>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

