



# Shared Certificates for Neural Network Verification



Marc Fischer<sup>1</sup>(✉) , Christian Sprecher<sup>2</sup>,  
Dimitar Iliev Dimitrov<sup>1</sup> , Gagandeep Singh<sup>3</sup> , and Martin Vechev<sup>1</sup>

<sup>1</sup> ETH Zurich, Zürich, Switzerland

{[marc.fischer](mailto:marc.fischer), [dimitar.iliev.dimitrov](mailto:dimitar.iliev.dimitrov), [martin.vechev](mailto:martin.vechev)}@inf.ethz.ch

<sup>2</sup> Nostic Solutions AG, Freienbach, Switzerland

[christian.sprecher@nostic.ch](mailto:christian.sprecher@nostic.ch)

<sup>3</sup> University of Illinois at Urbana-Champaign & VMware Research, Champaign, USA  
[ggnds@illinois.edu](mailto:ggnds@illinois.edu)

**Abstract.** Existing neural network verifiers compute a proof that each input is handled correctly under a given perturbation by propagating a symbolic abstraction of reachable values at each layer. This process is repeated from scratch independently for each input (e.g., image) and perturbation (e.g., rotation), leading to an expensive overall proof effort when handling an entire dataset. In this work, we introduce a new method for reducing this verification cost without losing precision based on a key insight that abstractions obtained at intermediate layers for different inputs and perturbations can overlap or contain each other. Leveraging our insight, we introduce the general concept of shared certificates, enabling proof effort reuse across multiple inputs to reduce overall verification costs. We perform an extensive experimental evaluation to demonstrate the effectiveness of shared certificates in reducing the verification cost on a range of datasets and attack specifications on image classifiers including the popular patch and geometric perturbations. We release our implementation at <https://github.com/eth-sri/proof-sharing>.

**Keywords:** Neural Network Verification · Local Verification · Adversarial Robustness

## 1 Introduction

The success of neural networks across a wide range of application domains [21, 30] has led to their widespread application and study. Despite this success, neural networks remain vulnerable to adversarial attacks [8, 23] which raises concerns over their trustworthiness in safety-critical settings such as autonomous driving and medical devices. To overcome this barrier, formal verification of neural networks has been proposed as a key technology in the literature [39]. As a result,

---

M. Fischer and C. Sprecher—Equal contribution.

C. Sprecher—Work performed while at ETH Zurich.

© The Author(s) 2022

S. Shoham and Y. Vizel (Eds.): CAV 2022, LNCS 13371, pp. 127–148, 2022.

[https://doi.org/10.1007/978-3-031-13185-1\\_7](https://doi.org/10.1007/978-3-031-13185-1_7)

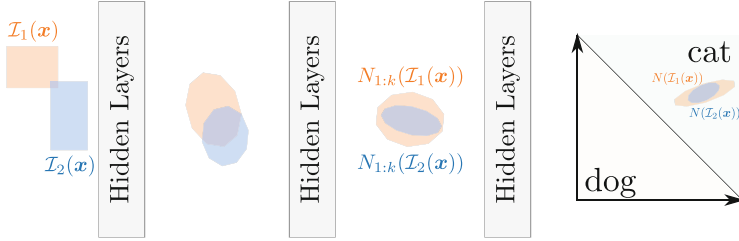
recent years have witnessed a growing interest in verifying critical safety properties of neural networks (e.g., fairness, robustness) [14, 17, 18, 31, 32, 40, 42] specified using pre and post conditions over network inputs and outputs respectively. Conceptually, existing verifiers propagate sets of inputs in the precondition captured in symbolic form (e.g., convex sets) through the network, an expensive process that produces over-approximations of all possible values at intermediate layers. The final abstraction of the output can then be used to check postconditions. The key technical challenge all existing verifiers aim to address is speeding up and scaling the certification process, i.e., faster and more efficient propagation of symbolic shapes while reducing the overapproximation error.

*This Work: Accelerating Certification via Proof Sharing.* In this work, we propose a new, complementary method for accelerating neural network verification based on the key observation that instead of treating each certification attempt in isolation as existing verifiers do, we can reuse proof effort among multiple such attempts, thus obtaining significant overall speed-ups without losing precision. Figure 1 illustrates both, standard verification and the concept of proof sharing.

In standard verification an input region  $\mathcal{I}_1(\mathbf{x})$  (orange square) is propagated from left to right, obtaining intermediate shapes at each intermediate layer (here the goal is to verify all points in the input region are classified as “cat” by the neural network  $N$ ). We observe that the abstraction obtained for a new region  $\mathcal{I}_2(\mathbf{x})$  (e.g., blue shapes) can be contained inside existing abstractions from  $\mathcal{I}_1(\mathbf{x})$ , an effect we term *proof subsumption*. This effect can be observed both between abstractions obtained from different specifications (e.g.,  $\ell_\infty$  and adversarial patches) for the same data point and between proofs for the same property but different, yet semantically similar inputs. Building on this observation, we introduce the notion of proof sharing via templates. Proof sharing works in two steps: first, we leverage abstractions from existing proofs in order to create templates, and second, we augment the verifier with these templates, stopping the expensive propagation at an intermediate layer as soon as the newly generated abstraction is included inside an existing template. Key technical ingredients to the effectiveness of our approach are fast template generation and inclusion checking techniques. We experimentally demonstrate that proof sharing can achieve significant speed-ups in challenging scenarios including proving robustness to adversarial patches [10] and geometric perturbations [3] across different neural network architectures.

*Main Contributions.* Our key contributions are:

- An introduction and formalization of the concept of proof sharing in neural network verification: the idea that some proofs capture others (Sect. 3).
- A general framework leveraging the above concept, enabling proof effort reuse via proof templates (Sect. 4).
- A thorough experimental evaluation involving verification of neural network robustness against challenging adversarial patch and geometric perturbations, demonstrating that our methods can achieve proof match rates of up 95% as well as provide non-trivial end-to-end certification speed-ups (Sect. 5).



**Fig. 1.** Visualization of neural network verification. The input regions  $\mathcal{I}_1(\mathbf{x}), \mathcal{I}_2(\mathbf{x})$  are propagated layer by layer through a neural network  $N$ . The high-dimensional convex shapes are visualized in 2d. While initially  $\mathcal{I}_1(\mathbf{x})$  and  $\mathcal{I}_2(\mathbf{x})$  only slightly overlap, at layer  $k$ ,  $N_{1:k}(\mathcal{I}_2(\mathbf{x}))$  is fully contained in  $N_{1:k}(\mathcal{I}_1(\mathbf{x}))$ . (Color figure online)

## 2 Background

Here we formally introduce the necessary background for proof sharing.

*Neural Network.* A neural network  $N$  is a function  $N : \mathbb{R}^{d_{\text{in}}} \rightarrow \mathbb{R}^{d_{\text{out}}}$ , commonly built from individual layers  $N = N_L \circ N_{L-1} \circ \dots \circ N_1$ . Throughout this text, we consider feed-forward neural networks, where each layer  $N_i(\mathbf{x}) = \max(\mathbf{A}\mathbf{x} + \mathbf{b}, 0)$  consists of an affine transformation ( $\mathbf{A}\mathbf{x} + \mathbf{b}$ ) as well as a rectified linear unit (ReLU), that applies the max with 0 elementwise. A neural network, classifying inputs into  $c$  classes, outputs  $d_{\text{out}} := c$  scores, one for each class, and assigns the class with the highest score as the predicted one. While, as is common in the neural network verification literature, we use image classification as a proxy task, many other applications work analogously. Our approach also naturally extends to other types of neural networks, if verifiers exist for these architectures. We discuss the challenges and limitations of such generalizations in Sect. 4.5. In the following, for  $k < L$ , we let  $N_{1:k}$  denote the application of the first  $k$  layers and  $N_{k+1:L}$  denote the last  $L - k$  layers respectively.

*(Local) Neural Network Verification.* Given a set of inputs and a postcondition  $\psi$ , the goal of neural network verification is to prove that  $\psi$  holds over the output of the neural network corresponding to the given set of inputs. In this work, we focus on local verification, proving that  $\psi$  holds for the network output for a given region  $\mathcal{I}(\mathbf{x}) \subseteq \mathbb{R}^{d_{\text{in}}}$  formed around the input  $\mathbf{x}$ . Formally, we state this as:

*Problem 1 (Local neural network verification).* For a region  $\mathcal{I}(\mathbf{x}) \subseteq \mathbb{R}^{d_{\text{in}}}$ , neural network  $N$ , and postcondition  $\psi$ , verify that  $\forall \mathbf{z} \in \mathcal{I}(\mathbf{x}). N(\mathbf{z}) \models \psi$ . We write  $\mathcal{I}(\mathbf{x}) \models \psi$  if  $\forall \mathbf{z} \in \mathcal{I}(\mathbf{x}). N(\mathbf{z}) \models \psi$ .

Here, we restrict ourselves to verifiers based on abstract interpretation [11, 14] as they achieve state-of-the-art precision and scalability [31, 32]. Further, many other popular verifiers [38, 42] can be formulated using abstract interpretation. These verifiers propagate  $\mathcal{I}(\mathbf{x})$  symbolically through the network  $N$  layer-by-layer using abstract transformers, which overapproximate the effect of applying the

transformations defined in the different layers on symbolic shapes. The propagation yields an abstraction of the exact shape at each layer. The verifiers finally check if the abstracted output implies  $\psi$ . This is showcased in Fig. 1, where the input regions  $\mathcal{I}_1(\mathbf{x})$  and  $\mathcal{I}_2(\mathbf{x})$  are propagated layer-by-layer through  $N$ .

For a verifier  $V$ , we let  $V(\mathcal{I}(\mathbf{x}), N)$  denote the abstraction obtained after the propagation of  $\mathcal{I}(\mathbf{x})$  through the network  $N$ . We declutter notation by overloading  $N$  and writing  $N(\mathcal{I}(\mathbf{x}))$  for the same if  $V$  is clear from context, i.e.,  $V(\mathcal{I}(\mathbf{x}), N) = N(\mathcal{I}(\mathbf{x}))$ .

We consider robustness verification, where the goal is to prove that the network classification does not change within an input region. A common input region is the  $\ell_\infty$ -bounded additive noise, defined as  $\mathcal{I}_\epsilon(\mathbf{x}) := \{\mathbf{z} \mid \|\mathbf{x} - \mathbf{z}\|_\infty \leq \epsilon\}$ . Here,  $\epsilon$  defines the size of the maximal perturbation to  $\mathbf{x}$ . The postcondition  $\psi$  denotes classification to the same class as  $\mathbf{x}$ . Throughout this paper, we consider different instantiations for  $\mathcal{I}(\mathbf{x})$  but assume that  $\psi$  denotes classification invariance (although other choices would work analogously). Due to this, we refer to  $\mathcal{I}(\mathbf{x})$  as input region and specification interchangeably. For example, in Fig. 1, the goal is to verify that all points contained in  $N(\mathcal{I}_1(\mathbf{x}))$  are classified as ‘‘cat’’.

### 3 Proof Sharing with Templates

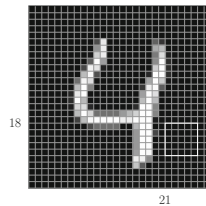
Before introducing our framework for proof sharing, we further expand the motivation example discussed in Fig. 1.

#### 3.1 Motivation: Proof Subsumption

As stated earlier, we empirically observed that for many input regions  $\mathcal{I}_i(\mathbf{x})$  and  $\mathcal{I}_j(\mathbf{x})$ , the abstraction corresponding to one region at some intermediate layer  $k$  contains that of another. Formally:

**Definition 1 (Proof Subsumption).** *For specifications  $\mathcal{I}_i(\mathbf{x}), \mathcal{I}_j(\mathbf{x})$ , we say that the proof of  $\mathcal{I}_i(\mathbf{x})$  subsumes that of  $\mathcal{I}_j(\mathbf{x})$  if at some layer  $k$ ,  $N_{1:k}(\mathcal{I}_j(\mathbf{x})) \subseteq N_{1:k}(\mathcal{I}_i(\mathbf{x}))$ , which we denote as  $\mathcal{I}_j(\mathbf{x}) \subseteq_{N,k} \mathcal{I}_i(\mathbf{x})$ .*

While not formally required, particularly interesting are cases where proof subsumption occurs despite  $\mathcal{I}_i(\mathbf{x}) \not\subseteq \mathcal{I}_j(\mathbf{x})$ . This form of proof subsumption is showcased in Fig. 1, where  $\mathcal{I}_1(\mathbf{x})$  and  $\mathcal{I}_2(\mathbf{x})$  have only a small overlap, yet  $\mathcal{I}_2(\mathbf{x}) \subseteq_{N,k} \mathcal{I}_1(\mathbf{x})$ . For another example, consider a neural network  $N$  trained as a hand-written digit classifier for the MNIST dataset [22] (example shown in Fig. 2) and the following two specifications:

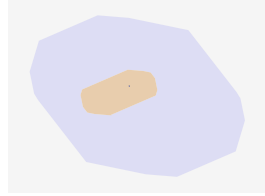


**Fig. 2.** Example of an MNIST image.  $\mathcal{I}_{5 \times 5}^{18,21}(\mathbf{x})$  signifies arbitrary change in the outlined area.

- $\ell_\infty$ -bounded perturbations: all the pixels in an input image can arbitrarily be changed independently by a small amount  $\mathcal{I}_\epsilon(\mathbf{x}) := \{\mathbf{z} \mid \|\mathbf{x} - \mathbf{z}\|_\infty \leq \epsilon\}$ ,

**Table 1.** Proof subsumption on a robust MNIST classifier with 94 % accuracy. Verif. acc. denotes the percentage of verifiable inputs from the test set for  $\ell_\infty$ -perturbations ( $\mathcal{I}_\epsilon$ ).

| $\epsilon$ | verif. acc. for $\mathcal{I}_\epsilon$ [%] | $\mathcal{I}_{2 \times 2}^{i,j}(\mathbf{x}) \subseteq_{N,k} \mathcal{I}_\epsilon(\mathbf{x})$ at layer k [%] |       |       |       |       |
|------------|--|--|-------|-------|-------|-------|
|            |  | 1  | 2     | 3     | 4     | 5     |
| 0.1        | 89.74                                      | 61.40  | 72.85 | 77.65 | 81.75 | 82.70 |
| 0.2        | 81.40                                      | 62.85  | 77.05 | 82.40 | 86.05 | 86.60 |



**Fig. 3.** The abstraction obtained for  $\mathcal{I}_\epsilon(\mathbf{x})$  (blue) contains that for  $\mathcal{I}_{2 \times 2}^{i,j}(\mathbf{x})$  (orange) (projected to  $d = 2$ ). (Color figure online)

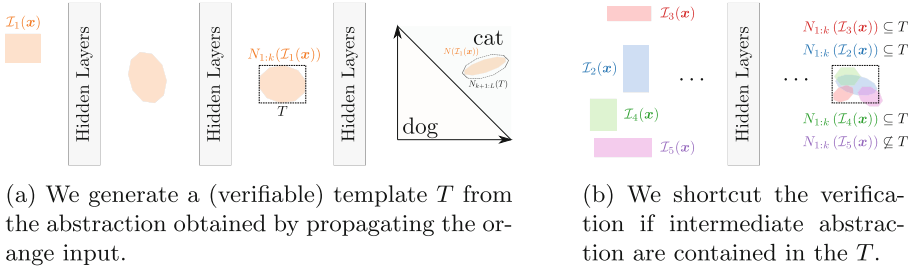
- adversarial patches [10]. A  $p \times p$  patch inside which the pixel intensity can vary arbitrarily is placed on an image at coordinates  $(i, j)$ , for which we write  $\mathcal{I}_{p \times p}^{i,j}$ . We showcase a patch in Fig. 2 and formally define them in Sect. 4.3.

Clearly  $\mathcal{I}_{p \times p}^{i,j}(\mathbf{x}) \not\subseteq \mathcal{I}_\epsilon(\mathbf{x})$  (unless  $\epsilon = 1$ ). In Table 1, we show that for a classifier (5 layers with 100 neurons each) we indeed observe proof subsumption. We report the accuracy, i.e., the rate of correct predictions on the unperturbed test data, as well as the certified accuracy, i.e., the rate of samples  $\mathbf{x}$  for which the prediction is correct and  $\mathcal{I}(\mathbf{x}) \models \psi$  is verified, for  $\mathcal{I}_\epsilon$  with  $\epsilon = 0.1$  and 0.2 over the whole test set. We also show the percentage of  $\mathcal{I}_{2 \times 2}^{i,j}(\mathbf{x})$  contained in  $\mathcal{I}_\epsilon(\mathbf{x})$  at layer  $k$ . To this end, we pick 1000 random  $\mathbf{x}$  for which  $\mathcal{I}_\epsilon(\mathbf{x})$  is verifiable and sample 2  $(i, j)$  pairs each. We utilize a Box domain verifier and a robustly trained network [24]. Figure 3 shows a patch specification  $\mathcal{I}_{2 \times 2}^{i,j}(\mathbf{x})$  (in orange) contained in the  $\ell_\infty$  specification  $\mathcal{I}_\epsilon$  (in blue) projected to 2 dimensions via PCA.

*Reasons for Proof Subsumption.* In Table 1, we observe that the rate of proof subsumption increases with larger  $\epsilon$  and  $k$ . These observations give an intuition as to why we observe proof subsumption. First, as input regions pass through the neural network, in each layer the abstractions become more imprecise. While this fundamentally limits verification, it makes the subsumption of abstractions more probable. This effect increases, when increasing  $\epsilon$  for  $\mathcal{I}_\epsilon$ . Second, and more fundamentally, while passing through the layers of a neural network, we observed that semantically similar yet distinct image inputs, e.g., two similar-looking handwritten digits, have activation vectors that grow closer in  $\ell_2$  norm as they pass through the layers of the neural network [21, 34]. This effect is a consequence of the neural network distilling low-level information (e.g., individual pixel values) into high-level concepts (e.g., the classes of digits). As specifications (and their proofs) correspond to sets of concrete inputs, a similar effect may apply. We conjecture that these two effects drive the observed proof subsumption.

### 3.2 Proof Sharing with Templates

Leveraging this insight, we introduce the idea of proof sharing via templates, showcased in Fig. 4. We use an abstraction obtained from a robustness proof



**Fig. 4.** Conceptualization of proof sharing with templates. In (a) we create a verifiable template  $T$  (black-dashed border) from specification  $N_{1:k}(\mathcal{I}_1(\mathbf{x}))$ . When verifying new specifications  $\mathcal{I}_2, \dots, \mathcal{I}_5$ , shown in (b), we can shortcut the verification of all but  $\mathcal{I}_5$  by subsuming them in  $T$ .

$N_{1:k}(\mathcal{I}_1(\mathbf{x}))$  at layer  $k$  to create a template  $T$ . After ensuring that  $T$  is verifiable, it can be used to shortcut the verification of other regions, e.g., of  $\mathcal{I}_2(\mathbf{x}), \dots, \mathcal{I}_5(\mathbf{x})$ . Formally we decompose proof sharing into two sub-problems: (i) the generation of proof templates and (ii) the matching of abstractions corresponding to other properties to these templates. For simplicity, here we only consider templates at a single layer  $k$  of the neural network and we show an extension to multiple layers in Sect. 4.3.

Our goal is to construct a template  $T$  at layer  $k$  that implies the postcondition and captures abstractions at layer  $k$  obtained from propagating several  $\mathcal{I}_i(\mathbf{x})$ . As it is challenging to find a single  $T$  that captures abstractions corresponding to many input regions, yet remains verifiable, we allow a set of templates  $\mathcal{T}$ . We state this formally as:

*Problem 2 (Template Generation).* For a given neural network  $N$ , input  $\mathbf{x}$  and set of specifications  $\mathcal{I}_1, \dots, \mathcal{I}_r$ , layer  $k$  and a postcondition  $\psi$ , find a set of templates  $\mathcal{T}$  with  $|\mathcal{T}| \leq m$  such that:

$$\begin{aligned} \arg \max_{\mathcal{T}} \sum_{i=1}^r \left[ \bigvee_{T \in \mathcal{T}} N_{1:k}(\mathcal{I}_i(\mathbf{x})) \subseteq T \right] & \quad (1) \\ \text{s.t. } \forall T \in \mathcal{T}. N_{k+1:L}(T) \models \psi. & \end{aligned}$$

Intuitively, Eq. (1) aims to find a set  $\mathcal{T}$  of templates  $T$  at layer  $k$ , such that the maximal amount (via the sum) of specifications  $\mathcal{I}_1, \dots, \mathcal{I}_r$  is contained in at least one template  $T$  (via the disjunction) while ensuring that the individual  $T$  are still verifiable (via the constraint on the second line). As neural network verification required by the constraints of Eq. (1), is NP-complete [17], computing an exact solution to Problem 2 is computationally infeasible. Therefore, we compute an approximate solution to Eq. (1). In general, Problem 2 does not necessarily require that the templates  $T$  are created from previous proofs. However, building on proof subsumption, as discussed in Sect. 3.1, in Sect. 4 we will infer the templates from previously obtained abstractions.

To leverage proof sharing once the templates  $\mathcal{T}$  are obtained, we need to be able to match an abstraction  $S = N_{1:k}(\mathcal{I}(\mathbf{x}))$  verified using proof transfer to a template in  $\mathcal{T}$ :

*Problem 3 (Template Matching).* Given a set of templates  $\mathcal{T}$  at layer  $k$  of a neural network  $N$ , and a new input region  $\mathcal{I}(\mathbf{x})$ , determine whether there exists a  $T \in \mathcal{T}$  such that  $S \subseteq T$ , where  $S = N_{1:k}(\mathcal{I}(\mathbf{x}))$ .

Together, Problems 2 and 3 outline a general framework for proof sharing, permitting many instantiations. We note that Problems 2 and 3 present an inherent precision vs. speed trade-off: Problem 3 can be solved most efficiently for small values of  $m = |\mathcal{T}|$  and simpler representations of  $T$  (allowing faster checking of  $S \subseteq T$ ) at the cost of lower proof matching rates. Alternatively, Eq. (1) can be maximized by large  $m$  and  $T$  represented by complex abstractions, thus attaining high precision but expensive template generation and matching.

*Beyond Proof Sharing on the Same Input.* In this section, we focused on proof sharing for different specifications of the same input  $\mathbf{x}$ . However, we observed that proof sharing is even possible between specifications defined on different inputs  $\mathbf{x}$  and  $\mathbf{x}'$ . To facilitate the use of templates in this setting, Eq. (1) in Problem 2 can be adapted to consider an input distribution.

## 4 Efficient Verification via Proof Sharing

We now consider an instantiation of proof sharing where we are given an input  $\mathbf{x}$  and properties  $\mathcal{I}_1, \dots, \mathcal{I}_r$  to verify. Our general approach, based on Problems 2 and 3, is shown in Algorithm 1. In this section, we first discuss Algorithm 1 in general. We then describe the possible choices of abstract domains and their implications on the algorithm, followed by a discussion on template generation for two different specific problems. Finally, we conclude the section with a discussion on the conditions for effective proof sharing verification.

In Algorithm 1, we first create the set of templates  $\mathcal{T}$  (Line 1, discussed shortly) and subsequently verify  $\mathcal{I}_1, \dots, \mathcal{I}_r$  using  $\mathcal{T}$ . Here, we consider two, potentially identical, verifiers  $V_T$  and  $V_S$ , where  $V_T$  is used to create the templates  $\mathcal{T}$  and  $V_S$  is used to propagate input regions up to the template layer  $k$ . For each  $\mathcal{I}_i$  we propagate it up to layer  $k$  (Line 4) to obtain  $S = N_{1:k}(\mathcal{I}_i(\mathbf{x}))$  and check if we can match it to a template  $T_j \in \mathcal{T}$  (Line 6) using an inclusion check. If a match is found, then we conclude that  $N(\mathcal{I}_i(\mathbf{x})) \models \psi$  and set the verification output  $v_i$  to True. If this is not the case (Line 11) we verify  $N(\mathcal{I}_i(\mathbf{x})) \models \psi$  directly by checking  $V_S(S, N_{k+1:L}) \models \psi$ . If the template generation fails, we revert to verifying  $\mathcal{I}_i$  by applying  $V_S$  in the usual way (omitted in Algorithm 1).

*Soundness.* As long as the templates  $T$  are sound, this procedure is sound, i.e. Algorithm 1 only returns  $v_i = \text{True}$  if  $\forall z \in \mathcal{I}_i(\mathbf{x}). N(z) \models \psi$  holds. Formally:

**Theorem 1.** *Algorithm 1 is sound if  $\forall T \in \mathcal{T}, z \in T. N_{k+1:L}(z) \models \psi$  and  $V_S$  is sound.*

This holds by the construction of the algorithm:

*Proof.* For a given  $\mathbf{x}$  and  $\mathcal{I}_i$ , Algorithm 1 only claims  $v_i = \text{True}$  if either the check in (i) Line 6 or (ii) Line 11 succeeds. Since  $V_S$  is sound, we know that  $\forall \mathbf{z} \in \mathcal{I}_i(\mathbf{x}). N_{1:k}(\mathbf{z}) \in S$ . Therefore in case (i) by our requirement on  $T$  as well as  $S \subseteq T$  it follows that  $\forall \mathbf{z} \in \mathcal{I}_i(\mathbf{x}). N(\mathbf{z}) \models \psi$ . In case (ii) we execute Line 12 and the same property holds due to the soundness of  $V_S$ .

Importantly, Theorem 1 shows that the generation process of  $\mathcal{T}$  does not affect the overall soundness as long as the set of templates  $\mathcal{T}$  fulfills the condition in Theorem 1. In particular, that means that when solving Problem 2, it suffices to show the side condition  $(\forall T \in \mathcal{T}. N_{k+1:L}(T) \models \psi)$  holds, while heuristically approximating the actual optimization criteria. We let  $V_T$  denote the verifier used to ensure this property in GEN\_TEMPLATES.

*Precision.* We say a verifier  $V_1$  is more precise than another verifier  $V_2$  on  $N$  if out of a set of specifications it can verify some that  $V_2$  can not.

**Theorem 2.** *If  $V_S(V_S(\mathcal{I}_i(\mathbf{x}), N_{1:k}), N_{k+1:L}) = V_S(\mathcal{I}_i(\mathbf{x}), N)$ , then Algorithm 1 is at least as precise as  $V_S$ .*

*Proof.* Since, even if the inclusion check in Line 6 fails, due to Line 12 we output  $v_i = V_S(V_S(\mathcal{I}_i(\mathbf{x}), N_{1:k}), N_{k+1:L}) \models \psi$  (Line 12), which by our requirement equals  $v_i = V_S(\mathcal{I}_i(\mathbf{x}), N) \models \psi$ . Therefore we have at least the precision of  $V_S$ .

The required property holds for any verifier  $V_S$  for which the abstractions of all network layers depends only on the abstractions from previous layers and is fulfilled for all verifiers considered in this paper. For verifiers  $V_S$  that do not fulfill the required property, potential losses in precision can be remedied (at the cost of runtime) by using  $V_S(\mathcal{I}_i(\mathbf{x}), N_{1:L})$  in Line 12. Interestingly, it is even possible to increase the precision of Algorithm 1 over  $V_S$  by creating templates  $T$  that are verified with a more precise verifier  $V_T$ . However, in this discussion, we restrict ourselves to speed gains. We believe that obtaining precision gains requires instantiating our framework with a significantly different approach than that taken for improving speed which is the main focus of our work. We leave this as an interesting item for future work.

---

**Algorithm 1:** Neural Network Verification Utilizing Proof Templates

---

**Input:**  $\mathbf{x}, \mathcal{I}_1, \dots, \mathcal{I}_r, k, \psi$ , verifiers  $V_S, V_T$

**Result:**  $v_1, \dots, v_r$  indicating

$v_i := (N(\mathcal{I}_i(\mathbf{x})) \models \psi)$

```

1  $\mathcal{T} \leftarrow \text{GEN\_TEMPLATES}(\mathbf{x}, N, k, \psi, V_S, V_T)$ 
2  $v_1, \dots, v_r \leftarrow \text{False}$ 
3 for  $i \leftarrow 1$  to  $r$  do
4    $S \leftarrow V_S(\mathcal{I}_i(\mathbf{x}), N_{1:k})$ 
5   for  $T_j \in \mathcal{T}$  do
6     if  $S \subseteq T_j$  then
7        $v_i \leftarrow \text{True}$ 
8       break
9     end
10  end
11  if  $\neg v_i$  then
12     $v_i \leftarrow (V_S(S, N_{k+1:L}) \models \psi)$ 
13  end
14 end
15 return  $v_1, \dots, v_r$ 

```

---



*Run-Time.* Here, we aim to characterize the run-time of Algorithm 1 as well as its speed-up over conventional verification. For an input  $\mathbf{x}$ , (keeping the other parameters fixed), the expected run time is

$$t_{PS} = t_{\mathcal{T}} + r(t_S + t_{\subseteq} + (1 - \rho)t_{\psi}) \quad (2)$$

where  $t_{\mathcal{T}}$  is the expected time required to generate the templates at Line 1,  $r$  is the number of specifications to be verified,  $t_S$  is the expected time to compute  $S$  (Line 4),  $t_{\subseteq}$  is the time to check  $S \subseteq T$  for  $T \in \mathcal{T}$  until a match is found (Line 5 to Line 10),  $\rho \in [0, 1]$  is the rate of specifications where a template is found and  $t_{\psi}$  is the time required to check  $\psi$  on the network output corresponding to  $S$  (Line 12). This time is minimized if the individual expected run times  $t_{\mathcal{T}}, t_S, t_{\psi}$  are minimal and  $\rho$  is large (i.e., close to 1). Unfortunately, computing the template match rate  $\rho$  analytically is challenging and requires global reasoning over the neural network for all valid inputs, which are not clearly defined. However, our empirical analysis (in Sect. 5) shows that  $\rho$  is higher when templates are created at later layers (as in Sect. 3.1).

To determine the speed-up compared to a baseline standard verifier, we make the simplifying assumption that there is a single verifier  $V = V_S = V_T$  that has expected run-time  $\nu$  for each layer. Thus, the expected run-time for the conventional verifier is  $t_{BL} = rL\nu$ . We have  $t_{\mathcal{T}} = \lambda m L \nu$ ,  $t_S = k\nu$ ,  $t_{\psi} = (L - k)\nu$ ,  $t_{\subseteq} = \eta m$  and ultimately  $t_{PS} = (m + r(1 - \rho))L\nu + r\rho k\nu + r\eta m$  for constants  $\lambda \in \mathbb{R}_{>0}$ , which indicates the overhead in generating one template over just verifying it, and  $\eta \in \mathbb{R}_{>0}$  which denotes the time required to perform an inclusion check for one template. As this phrasing shows, Algorithm 1 has the same asymptotic runtime as the base verifier  $V$ . Further, this formulation allows us to write our expected speed-up as  $\frac{t_{BL}}{t_{PS}} = \frac{r}{\lambda m + \eta r m / L \mu + r \rho k / L + r(1 - \rho)}$ . This speed-up is maximized when  $k$  is small compared to  $L$ , i.e., templates are placed early in the neural network, the matching rate  $\rho$  is close to 1, and  $m, \lambda, \eta$  are small, i.e., generation and matching are fast. Unfortunately, these requirements are at odds with each other: as we show in Sect. 5, higher  $m$  leads to higher matching rate  $\rho$  and  $\rho$  is naturally higher for templates later in the neural network (higher  $k$ ). Thus high speed-ups require careful hyper-parameter choices.

To showcase how we can achieve good templates as well as fast matching, we next discuss the choice of the abstract domain to be used in the propagation and the representation of the templates. Then we discuss the template generation procedure and instantiate it for the verification of robustness to adversarial patches and geometric perturbations.

## 4.1 Choice of Abstract Domain

To solve Problems 2 and 3 in a way that minimizes the expected runtime and maximizes the overall precision, the choice of abstract domain is crucial. Here we briefly review common choices of abstract domains for neural network verification and how they are suited to our problem. Geometrically these domains can be thought of as a convex abstraction of the set of vectors representing reachable

values at each layer of the neural network. We say that an abstraction  $a_1$  is more precise than another abstraction  $a_2$ , if and only if  $a_1 \subseteq a_2$ , i.e., all points in  $a_1$  occur in  $a_2$ . Similarly, we say that a domain is more precise than another if it can express all abstractions in the other domain.

The Box (or Interval) domain [14, 16, 24] abstracts sets in  $d$  dimensions as  $B = \{\mathbf{a} + \text{diag}(\mathbf{d})\mathbf{e} \mid \mathbf{e} \in [-1, 1]^d\}$  with center  $\mathbf{a} \in \mathbb{R}^d$  and width  $\mathbf{d} \in \mathbb{R}_{\geq 0}^d$ . The Zonotope domain [14, 15, 24, 31, 40] uses relaxations  $Z$  of the form

$$Z = \{\mathbf{a} + \mathbf{A}\mathbf{e} \mid \mathbf{e} \in [-1, 1]^q\}, \quad (3)$$

parametrized with  $\mathbf{a} \in \mathbb{R}^d$  and  $\mathbf{A} \in \mathbb{R}^{d \times q}$ .

A third common choice are (restricted) convex Polyhedra  $P$  [12, 32, 42]. Here, we consider  $P$  to be in the DeepPoly (DP) domain [32, 42]. Generally, Boxes are less precise, i.e. certify fewer properties, than Zonotopes or Polyhedra.

For efficient proof sharing, we require a fast inclusion check  $S \subseteq T$ , which is challenging in our context due to the high dimensionality  $d$  of the intermediate neural network layers. While we point the interested reader to [29] for a detailed discussion, we summarize the key results in Table 2. There,  $\checkmark$  denotes feasibility, i.e. low polynomial runtime (usually  $2d$  compar-

**Table 2.** Feasibility of  $S \subseteq T$  for Box  $B$ , Zonotope  $Z$  (with order reduction) and DP Polyhedra  $P$ .

|     |     | $T$          |          |              |                |
|-----|-----|--------------|----------|--------------|----------------|
|     |     | $B$          | $Z$      | $\alpha(Z)$  | $P$            |
| $S$ | $B$ | $\checkmark$ | $\times$ | $\checkmark$ | $(\checkmark)$ |
|     | $Z$ | $\checkmark$ | $\times$ | $\checkmark$ | $\times$       |
|     | $P$ | $\checkmark$ | $\times$ | $\checkmark$ | $(\checkmark)$ |

isons, sometimes with an additional matrix multiplication),  $\times$  denotes infeasibility, e.g. exponential run time. If  $T$  is a Box all checks are simple as it suffices to compute the outer bounding box of  $S$  and compare the  $2d$  constraints. If  $T$  is a DP Polyhedra these checks require a linear program (LP) to be solved. While the size of this LP permits a low theoretical time complexity, in case  $S$  is a Box or DP Polyhedra, in practice, we consider calling an LP solver too expensive (denoted as  $(\checkmark)$ ). For Zonotopes these checks are generally infeasible, as they require enumeration of the faces or corners, which is computationally expensive for large  $d$  and  $P$ . While Zonotopes can be encoded as Polyhedra (but not necessarily DP Polyhedra) and the same LP inclusion check as for  $P$  could be used, the resulting LP would require exponentially many variables due to the previously mentioned enumeration. However, by placing constraints on the matrix  $\mathbf{A}$  in Eq. (3) these inclusion checks can be performed efficiently. The mapping of a Zonotope to such a restricted Zonotope is called order reduction via outer-approximation [19, 29].

In particular, for a Zonotope  $Z$  we consider the order reduction  $\alpha_{\text{Box}}$  to its outer bounding box (where  $\mathbf{A}$  is diagonal) and note that other choices of  $\alpha$  are possible (e.g. the reduction to affine transformations of a hyperbox).

For a general Zonotope  $Z$  its outer bounding box  $Z' = \alpha_{\text{Box}}(Z)$  can be easily obtained. The center of  $Z'$  is  $\mathbf{a}$ , the center of  $Z$ . The width  $\mathbf{d} \in \mathbb{R}_{\geq 0}^d$  is given as  $d_i = \sum_{j=1}^q |A_{i,j}|$ .  $Z'$  is represented as either a Box or a Zonotope (with

$\mathbf{A} = \text{diag}(\mathbf{d})$ . To check  $S \subseteq Z'$  for a general Zonotope  $S$  it suffices to check  $\alpha_{\text{Box}}(S) \subseteq Z'$  which reduces to the simple inclusion check for boxes.

Based on the above discussion we will use the Zonotope domain to represent all abstractions, and use verifiers  $V_S = V_T$  that propagate these zonotopes using the state-of-the-art DeepZ transformers [31]. To permit efficient inclusion checks we apply  $\alpha_{\text{Box}}$  on the resulting zonotopes to obtain the Box templates  $T$ , which we treat as a special case of Zonotopes.

## 4.2 Template Generation

We now discuss instantiations for GEN\_TEMPLATES in Algorithm 1. Recall from Sect. 3.1 the idea of proof subsumption, i.e. that abstractions for some specification contain abstractions for other specifications. Building on this, we relax the Problem 2 in order to create  $m$  templates  $T_j$  from intermediate abstractions  $N_{1:k}(\hat{\mathcal{I}}_i(\mathbf{x}))$  for some  $\hat{\mathcal{I}}_1, \dots, \hat{\mathcal{I}}_m$ . Note that  $\hat{\mathcal{I}}_j$  are not necessarily directly related to the specifications  $\mathcal{I}_1, \dots, \mathcal{I}_r$  that we want to verify. For a chosen layer  $k$ , input  $\mathbf{x}$ , number of templates  $m$  and verifiers  $V_S$  and  $V_T$  we optimize

$$\arg \max_{\hat{\mathcal{I}}_1, \dots, \hat{\mathcal{I}}_m} \sum_{i=1}^r \left[ \bigvee_{j=1}^m V_S(\mathcal{I}_i(\mathbf{x}), N_{1:k}) \subseteq T_j \right] \quad (4)$$

where  $T_j = \alpha_{\text{Box}}(V_T(\hat{\mathcal{I}}_j(\mathbf{x}), N_{1:k}))$

s.t.  $V_T(T_j, N_{k+1:L}) \models \psi$  for  $j \in 1, \dots, m$ .

As originally in Problem 2 (Eq. (1)) we aim to find a set of templates such that the intermediate shapes at layer  $k$  for most of the  $r$  specifications are covered by at least one template  $T$ . In contrast to Eq. (1), we tie  $T_j$  to the specifications  $\hat{\mathcal{I}}_j$ . This alone does not make the problem easier to tackle. However, next, we will discuss how to generate application-specific parametric  $\hat{\mathcal{I}}_j$  and solve Eq. (4) by optimizing over their parameters, allowing us to solve template generation much more efficiently than in Eq. (1).

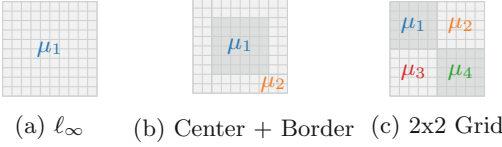
## 4.3 Robustness to Adversarial Patches

We now instantiate the above scheme in order to verify the robustness of image classifiers against adversarial patches [10]. Consider an attacker that is allowed to arbitrarily change any  $p \times p$  patch of the image, as showcased earlier in Fig. 2. For such a patch over pixel positions  $([i, i+p-1] \times [j, j+p-1])$ , the corresponding perturbation is

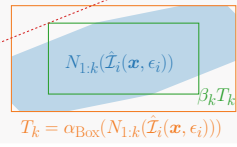
$$\mathcal{I}_{p \times p}^{i,j}(\mathbf{x}) := \{ \mathbf{z} \in [0, 1]^{h \times w} \mid \mathbf{z}_{\pi_{i,j}^C} = \mathbf{x}_{\pi_{i,j}^C} \}$$

$$\text{with } \pi_{i,j} = \left\{ (k, l) \mid \begin{array}{l} k \in i, \dots, i+p-1 \\ l \in j, \dots, j+p-1 \end{array} \right\}$$

where  $h$  and  $w$  denote the height and width of the input  $\mathbf{x}$ . Here  $\pi_{i,j}$  denotes the parts of the image affected by the patch, and  $\pi_{i,j}^C$  its complement, i.e., the



**Fig. 5.** Example splits  $\mu$  for  $10 \times 10$  pixels.



**Fig. 6.** Example Template. (Color figure online)

unaffected part of the image. To prove robustness for an arbitrarily placed  $p \times p$  patch, however, one must consider the perturbation set  $\mathcal{I}_{p \times p}(\mathbf{x}) := \cup_{i,j} \mathcal{I}_{p \times p}^{i,j}(\mathbf{x})$ .

To prove robustness for  $\mathcal{I}_{p \times p}$ , existing approaches [10] separately verify  $\mathcal{I}_{p \times p}^{i,j}(\mathbf{x})$  for all  $i \in \{1, \dots, h - p + 1\}, j \in \{1, \dots, w - p + 1\}$ . For example, with  $p = 2$  and a  $28 \times 28$  MNIST image, this approach requires 729 individual proofs. Because the different proofs for  $\mathcal{I}_{p \times p}$  share similarities, this is an ideal candidate for proof sharing. We utilize Algorithm 1 and check  $\wedge_i v_i$  at the end to speed up this process. For template generation, we solve Eq. (4) for  $m$  templates with an input perturbation  $\hat{\mathcal{I}}_i$  per template.

We empirically found that (recall Table 1) setting  $\hat{\mathcal{I}}_i$  to an  $\ell_\infty$  region  $\mathcal{I}_{\epsilon_i}$  to work particularly well to capture a majority of patch perturbations  $\mathcal{I}_{p \times p}^{i,j}$  at intermediate layers. Specifically, we found that setting  $\epsilon_i$  to the maximally verifiable value for this input to work particularly well.

To further increase the number of specifications contained in a set of templates  $\mathcal{T}$ , we use  $m$  template perturbations of the form

$$\hat{\mathcal{I}}_i(\mathbf{x}) := \{\mathbf{z} \mid \|\mathbf{x}_{\mu_i} - \mathbf{z}_{\mu_i}\|_\infty \leq \epsilon_i \wedge \mathbf{x}_{\mu_i^C} = \mathbf{z}_{\mu_i^C}\},$$

where  $\mu_i$  denotes a subset of pixels of the input image and  $\mu_i^C$  its complement and we maximize  $\epsilon_i$  in a best-effort manner. In particular, we consider  $\mu_1, \dots, \mu_m$ , such that they partition the set of pixels in the image (e.g., in Fig. 5).

As noted earlier, this generation procedure needs to be fast, yet obtain  $\mathcal{T}$  to which many abstractions match in order to obtain speed-ups. Thus, we consider small  $m$ , and fixed patterns  $\mu_1, \dots, \mu_m$ . For each  $\hat{\mathcal{I}}_i$ , we aim to find the largest  $\epsilon_i$  which can still be verified in order to maximize the number of matches. Note that for  $m = 1$ , this is equivalent to the  $\ell_\infty$  input perturbation  $\mathcal{I}_\epsilon$  with the maximally verifiable  $\epsilon$  for the given image.

Concretely, we can perform binary search over  $\epsilon_i$  in order find a large  $\epsilon_i$ , still satisfying  $N_{k+1:L}(\alpha_{\text{Box}}(N_{1:k}(\hat{\mathcal{I}}_i))) \models \psi$ . Verification with our chosen DeepZ Zonotopes is not monotonous in  $\epsilon_i$  due to the non-monotonic transformers used for non-linearities (e.g., ReLU). This renders the application of binary search a best-effort approximation. As we don't require a formal maximum but rather aim to solve a surrogate for Problem 2, this still works well in practice. Further note that, applying  $\alpha_{\text{Box}}$  to templates introduces imprecision, i.e.  $V_{\mathcal{T}}$  might not be able to prove properties over templates that it could without the application of  $\alpha_{\text{Box}}$ . However, Theorem 2 (which only requires properties of  $V_S$ ) still applies.

*Templates at Multiple Layers.* We can extend this approach to obtain templates at multiple layers without a large increase in computational cost. With templates at multiple layers, we first try to match the propagated shape against the earliest template layer and upon failure propagate it further to the next, where we again attempt to match the template. In Algorithm 1, this means repeating the block from Line 4 to Line 10 for each template layer before going on to the check on Line 11.

The full template generation procedure is given in Algorithm 2. First, we perform a binary search over  $\epsilon_i$  (Line 6) to find the largest  $\epsilon_i$ , for which the specification is verifiable. Then for each layer  $k$  in the set of layers  $K$  at which we are creating templates we create a box  $T_k$  from the Zonotope. As this  $T_k$  may not be verifiable, due to the imprecision added in  $\alpha_{\text{Box}}$ , we then perform another binary search for the largest scaling factor  $\beta_k$  (Line 10), which is applied to the matrix  $\mathbf{A}$  in Eq. (3). We denote this operation as  $\beta_k T_k$ . We show an example for a single layer  $k$  in Fig. 6. The blue area outlines the Zonotope found via Line 6, which is verifiable as it is fully on one side of the decision boundary (red, dashed). After applying  $\alpha_{\text{Box}}$  (orange), however, is not (crosses the decision boundary). By scaling it with  $\beta_k$  the shape is verifiable again (green) and used as a template.

#### 4.4 Geometric Robustness

Geometric robustness verification [3, 13, 28, 32] aims to verify the robustness of neural networks against geometric transformations such as image rotations or translations. These transformations typically include an interpolation operation. For example consider rotation  $R_\gamma$  of an image by  $\gamma \in \Gamma$  degrees for an interval  $\Gamma$  (e.g.,  $\gamma \in [-5, 5]$ ), for which we consider the specification  $\mathcal{I}_\Gamma(\mathbf{x}) := \{R_\gamma(\mathbf{x}) \mid \gamma \in \Gamma\}$ . We note that, unlike  $\ell_\infty$  and patch verification, the input regions for geometric transformations are non-linear and have no closed-form solutions. Thus, an overapproximation of the input region must be obtained [3]. For large  $\Gamma$ , the approximate input region  $\mathcal{I}_\Gamma(\mathbf{x})$ , can be too coarse resulting in imprecise verification. Hence, in order to assert  $\psi$  on  $\mathcal{I}_\Gamma$ , existing state-of-the-art approaches [3], split  $\Gamma$  into  $r$  smaller ranges  $\Gamma_1, \dots, \Gamma_r$  and then verify the resulting  $r$  specifications  $(\mathcal{I}_{\Gamma_i}, \psi)$  for  $i \in 1, \dots, r$ . These smaller perturbations share similarities facilitating proof sharing. We instantiate our approach similar to Sect. 4.3. A key difference to Sect. 4.3 is that while  $\mathbf{x} \in \mathcal{I}_{p \times p}^{i,j}(\mathbf{x})$  for all  $i, j$  in patches, here

---

#### Algorithm 2: Online Template Generation for Patches

---

**Input:**  $\mathbf{x}, N, \mu_1, \dots, \mu_m, K, \psi, V_T$   
**Result:**  $\mathcal{T}^k$  for  $k \in K$

- 1  $\mathcal{T}^k \leftarrow \{\}$  for  $k \in K$
- 2 **for**  $i \leftarrow 1$  **to**  $m$  **do**
- 3      $\hat{\mathcal{I}}_i(\mathbf{x}, \epsilon) := \{\mathbf{z} \mid \|\mathbf{x}_{\mu_i} - \mathbf{z}_{\mu_i}\| \leq \epsilon$
- 4          $\wedge \mathbf{x}_{\mu_i^c} = \mathbf{z}_{\mu_i^c}\}$
- 5      $f(\epsilon) := V_T(\hat{\mathcal{I}}_i(\mathbf{x}, \epsilon), N) \models \psi$
- 6      $\epsilon_i \leftarrow \text{bin\_search}(\epsilon, f(\epsilon))$
- 7     **for**  $k \in K$  **do**
- 8          $T_k \leftarrow \alpha_{\text{Box}}(V_T(\hat{\mathcal{I}}_i(\mathbf{x}, \epsilon_i), N_{1:k}))$
- 9          $g(\beta_k) := V_T(\beta T_k, N_{k+1:L}) \models \psi$
- 10         $\beta_k \leftarrow \text{bin\_search}(\beta, g(\beta))$
- 11         $\mathcal{T}^k \leftarrow \mathcal{T}^k \cup \{\beta_k T_k\}$
- 12     **end**
- 13 **end**
- 14 **return**  $\mathcal{T}^k$  for  $k \in K$

---

in general  $\mathbf{x} \notin \mathcal{I}_{\Gamma_i}(\mathbf{x})$  for most  $i$ . Therefore, the individual perturbations  $\mathcal{I}_i(\mathbf{x})$  do not overlap. To account for this, we consider  $m$  templates and split  $\Gamma$  into  $m$  equally sized chunks (unrelated to the  $r$  splits) obtaining the angles  $\gamma_1, \dots, \gamma_m$  at the center of each chunk. For  $m$  templates we then consider the perturbations  $\hat{\mathcal{I}}_i := \mathcal{I}_{\epsilon_i}(R_{\gamma_i}(\mathbf{x}))$ , denoting the  $\ell_\infty$  perturbation of size  $\epsilon_i$  around the  $\gamma_i$  degree rotated  $\mathbf{x}$ . To find the template we employ a procedure analogous to Algorithm 2.

#### 4.5 Requirements for Proof Sharing

Now, we discuss the requirements on the neural network  $N$  such that proof sharing via templates works well. For simplicity, we discuss simple per-dimension box bounds propagation for  $V_S$  and  $V_T$ . However, similar arguments can be made for more complex relational abstractions such as Zonotopes or Polyhedra.

In order for an abstraction  $S$  to match to a template  $T$ , we need to show interval inclusion for each dimension. For a particular dimension  $i$  this can occur in two ways: (i) when both  $S$  and  $T$  are just a point in that dimension and these points coincide, e.g.,  $a_i^S = a_i^T$ , or (ii) when  $a_i^S \pm d_i^S \subseteq a_i^T \pm d_i^T$ . While particularly in ReLU networks, the first case can occur after a ReLU layer sets values to zero, we focus our analysis here on the second case as it is more common. In this case, the width of  $T$  in that dimension  $d_i^T$  must be sufficient to cover  $S$ . Ignoring case (i) and letting  $\text{supp}(T)$  denote the dimensions in which  $d_i^T > 0$ , we can pose that  $\text{supp}(S) \subseteq \text{supp}(T)$  as a necessary condition for inclusion. While it is in general hard to argue about the magnitudes of these values, this approach still provides an intuition. When starting from input specifications  $\text{supp}(\mathcal{I}) \not\subseteq \text{supp}(\hat{\mathcal{I}})$ ,  $\text{supp}(S) \subseteq \text{supp}(T)$  can only occur if during propagation through the neural network  $N_{1:k}$  the mass in  $\text{supp}(\hat{\mathcal{I}})$  can “spread out” sufficiently to cover  $\text{supp}(S)$ . In the fully connected neural networks that we discuss here, the matrices of linear layers provide this possibility. However, in networks that only read part of the input at a time such as recurrent neural networks, or convolutional neural networks in which only locally neighboring inputs feed into the respective output in the next layer, these connections do not necessarily exist. This makes proof sharing hard until layers later in the neural network, that regionally or globally pool information. As this increases the depth of the layer  $k$  at which proof transfer can be applied, this also decreases the potential speed-up of proof transfer. This could be alleviated by different ways of creating templates, which we plan to investigate in the future.

## 5 Experimental Evaluation

We now experimentally evaluate the effectiveness of our algorithms from Sect. 4.

### 5.1 Experimental Setup

We consider the verification of robustness to adversarial patch attacks and geometric transformations in Sect. 5.2 and Sect. 5.3, respectively. We define specifications on the first 100 test set images each from the MNIST [22] and the

**Table 3.** Rate of  $\mathcal{I}_{2 \times 2}^{i,j}$  matched to templates  $\mathcal{T}$  for  $\mathcal{I}_{2 \times 2}$  patch verification for different combinations of template layers  $k$ ,  $7 \times 200$  networks, using  $m = 1$  template.

| template at layer $k$ | 1    | 2    | 3    | 4    | 5    | 6    | 7    | patch verif. [%] |
|-----------------------|------|------|------|------|------|------|------|------------------|
| MNIST                 | 18.6 | 85.6 | 94.1 | 95.2 | 95.5 | 95.7 | 95.7 | 97.0             |
| CIFAR                 | 0.1  | 27.1 | 33.7 | 34.4 | 34.2 | 34.2 | 34.3 | 42.2             |

**Table 4.** Average verification time in seconds per image for  $\mathcal{I}_{2 \times 2}$  patches for different combinations of template layers  $k$ ,  $7 \times 200$  networks, using  $m = 1$  template.

|       | Baseline | Proof Sharing, template layer $k$ |      |      |      |      |             |             |       |
|-------|----------|-----------------------------------|------|------|------|------|-------------|-------------|-------|
|       |          | 1                                 | 2    | 3    | 4    | 1+3  | 2+3         | 2+4         | 2+3+4 |
| MNIST | 2.10     | 1.94                              | 1.15 | 1.22 | 1.41 | 1.27 | <b>1.09</b> | 1.10        | 1.14  |
| CIFAR | 3.27     | 2.98                              | 2.53 | 2.32 | 2.47 | 2.35 | 2.49        | <b>2.42</b> | 2.55  |

CIFAR-10 dataset [20] (“CIFAR”) as with repetitions and parameter variations the overall runtime becomes high. We use DeepZ [31] as the baseline verifier as well as for  $V_S$  and  $V_T$  [31]. Throughout this section, we evaluate proof sharing for two networks on two common datasets: We use a seven layer neural network with 200 neurons per layer (“ $7 \times 200$ ”) and a nine layer network with 500 neurons per layer (“ $9 \times 500$ ”) for both the MNIST [22] and CIFAR datasets [20], both utilizing ReLU activations. These architectures are similar to the fully-connected ones used in the ERAN and Mnistfc VNN-Comp categories [2].

For MNIST, we train 100 epochs, enumerating all patch locations for each sample, and for CIFAR we train for 600 with 10 random patch locations, as outlined in [10] with interval training [16, 24]. On MNIST the  $7 \times 200$  and the  $9 \times 500$  achieve a natural accuracy of 98.3% and 95.3% respectively. For CIFAR, these values are 48.8% and 48.1% respectively. Our implementation utilizes PyTorch [25] and is evaluated on Ubuntu 18.04 with an Intel Core i9-9900K CPU and 64 GB RAM. For all timing results, we provide the mean over three runs.

## 5.2 Robustness Against Adversarial Patches

For MNIST, containing  $28 \times 28$  images, as outlined in Sect. 4.3, in order to verify inputs to be robust against  $2 \times 2$  patch perturbations, 729 individual perturbations must be verified. Only if all are verified, the overall property can be verified for a given image. Similarly, for CIFAR, containing  $32 \times 32$  color images, there are 961 individual perturbations (the patch is applied over all color channels).

We now investigate the two main parameters of Algorithm 2: the masks  $\mu_1, \dots, \mu_m$  and the layers  $k \in K$ . We first study the impact of the layer  $k$  used for creating the template. To this end, we consider the  $7 \times 200$  networks,

**Table 5.**  $\mathcal{I}_{2 \times 2}$  patch verification with templates at the 2nd & 3rd layer of the  $7 \times 200$  networks for different masks.

| Method/Mask       | m | patch matched [%] | t [s]       |
|-------------------|---|-------------------|-------------|
| Baseline          | - | -                 | 2.14        |
| L-infinity        | 1 | 94.1              | <b>1.11</b> |
| Center + Border   | 2 | 94.6              | 1.41        |
| $2 \times 2$ Grid | 4 | <b>95.0</b>       | 3.49        |

**Table 6.**  $\mathcal{I}_{2 \times 2}$  patch verification with templates generated on the second and third layer using the  $\ell_\infty$ -mask. Verification times are given for the baseline  $t^{BL}$  and for applying proof sharing  $t^{PS}$  in seconds per image.

| Dataset | Net            | verif. acc. [%] | $t^{BL}$ | $t^{PS}$    | patch mat. [%] | patch verif. [%] |
|---------|----------------|-----------------|----------|-------------|----------------|------------------|
| MNIST   | $7 \times 200$ | 81.0            | 2.10     | <b>1.10</b> | 94.1           | 97.0             |
|         | $9 \times 500$ | 66.0            | 2.70     | <b>1.32</b> | 93.0           | 95.3             |
| CIFAR   | $7 \times 200$ | 29.0            | 3.28     | <b>2.45</b> | 33.7           | 42.2             |
|         | $9 \times 500$ | 28.0            | 5.48     | <b>4.48</b> | 34.2           | 46.2             |

use  $m = 1$  (covering the whole image; equivalent to  $\hat{\mathcal{I}}_\epsilon$ ). Table 3 shows the corresponding template matching rates, and the overall percentage of individual patches that can be verified “patches verif.” (The overall percentage of images for which  $\mathcal{I}_{2 \times 2}$  is true is reported as “verif.” in Table 6.) Table 4 shows the corresponding verification times (including the template generation). We observe that many template matches can already be made at the second or third layer. As creating templates simultaneously at the second and third layer works well for both datasets, we utilize templates at these layers in further experiments.

Next, we investigate the impact of the pixel masks  $\mu_1, \dots, \mu_m$ . To this end, we consider three different settings, as showcased in Fig. 5 earlier: (i) the full image ( $\ell_\infty$ -mask as before;  $m = 1$ ), (ii) “center + border” ( $m = 2$ ), where we consider the  $6 \times 6$  center pixel as one group and all others as another, and (iii) the  $2 \times 2$  grid ( $m = 4$ ) where we split the image into equally sized quarters.

As we can see in Table 5, for higher  $m$  more patches can be matched to the templates, indicating that our optimization procedure is a good approximation to Problem 2, which only considers the number of templates matched. Yet, for  $m > 1$  the increase in matching rate  $p$  does not offset the additional time in template generation and matching. Thus,  $m = 1$  results in a better trade-off. This result highlights the trade-offs discussed throughout Sect. 3 and Sect. 4. Based on this investigation we now, in Table 6, evaluate all networks and datasets using  $m = 1$  and template generation at layers 2 and 3. In all cases, we obtain a speed up between 1.2 to  $2 \times$  over the baseline verifier. Going from  $2 \times 2$  to  $3 \times 3$  patches speed ups remain around 1.6 and 1.3 for the two datasets respectively.



**Table 7.** Speed-ups achievable in the setting of Table 3.  $t^{BL}$  the baseline.

| Layer $k$                                       |  | speedup at layer $k$ |      |      |      |
|---|--|----------------------|------|------|------|
|   |  | 1                    | 2    | 3    | 4    |
| realized  | $t^{BL}/t^{PS}$                        | 1.08                 | 1.83 | 1.72 | 1.49 |
| optimal   | $t^{BL}/(t_T + rt_S + rt_{\subseteq})$ | 3.75                 | 2.51 | 1.92 | 1.56 |
| optimal, no $\subseteq$                         | $t^{BL}/(t_T + rt_S)$                  | 4.02                 | 2.68 | 2.01 | 1.62 |
| optimal, no gen $\mathcal{T}$ ., no $\subseteq$ | $t^{BL}/rt_S$                          | 4.57                 | 2.90 | 2.13 | 1.69 |

*Comparison with Theoretically Achievable Speed-Up.* Finally, we want to determine the maximal possible speed-up with proof sharing and see how much of this potential is realized by our method. To this end we investigate the same setting and network as in Table 3. We let  $t^{BL}$  and  $t^{PS}$  denote the runtime of the base verifier without and with proof sharing respectively. Similar to the discussion in Sect. 4 we can break down  $t^{PS}$  into  $t_T$  (template generation time),  $t_S$  (time to propagate one input to layer  $k$ ),  $t_{\subseteq}$  (time to perform template matching) and  $t_{\psi}$  (time to verify  $S$  if no match). Table 7 shows different ratios of these quantities. For all, we assume a perfect matching rate at layer  $k$  and calculate the achievable speed-up for patch verification on MNIST. Comparing the optimal and realized results, we see that at layers 3 and 4 our template generation algorithm, despite only approximately solving Problem 2 achieves near-optimal speed-up. By removing the time for template matching and template generation we can see that, at deeper layers, speeding up  $t_{\subseteq}$  and  $t_T$  only yield diminishing returns.

### 5.3 Robustness Against Geometric Perturbations

For the verification of geometric perturbations, we take 100 images from the MNIST dataset and the  $7 \times 200$  neural network from Sect. 5.2. In Table 8, we consider an input region with  $\pm 2^\circ$  rotation,  $\pm 10\%$  contrast and  $\pm 1\%$  brightness change, inspired by [3]. To verify this region, similar to existing approaches [3], we choose to split the rotation into  $r$  regions, each yielding a Box specification over the input. Here we use  $m = 1$ , a single template, with the largest verifiable  $\epsilon$  found via binary search. We observe that as we increase  $r$ , the verification rate increases, but also the speed ups. Proof sharing enables significant speed-up between 1.6 to  $2.9\times$ .

Finally, we investigate the impact of the number of templates  $m$ . To this end, we consider a setting with a large parameter space:  $\pm 40^\circ$  rotation generated input region with  $r = 200$ . In Table 9, we evaluate this for  $m$  templates obtained from the  $\ell_\infty$  input perturbation around  $m$  equally spaced rotations, where we apply binary search to find  $\epsilon_i$  tailored for each template. Again we observe that  $m > 1$  allows more templates matches. However, in this setting the relative increase is much larger than for patches, thus making  $m = 3$  faster than  $m = 1$ .

**Table 8.**  $\pm 2^\circ$  rotation,  $\pm 10\%$  contrast and  $\pm 1\%$  brightness change split into  $r$  perturbations on 100 MNIST images. Verification rate, rate of splits matched and verified along with the run time of Zonotope  $t^{BL}$  and proof sharing  $t^{PS}$ .

| $r$ | verif. [%] | splits verif. [%] | splits matched [%] | $t^{BL}$ | $t^{PT}$     |
|-----|------------|-------------------|--------------------|----------|--------------|
| 4   | 73.0       | 87.3              | 73.1               | 3.06     | <b>1.87</b>  |
| 6   | 91.0       | 94.8              | 91.0               | 9.29     | <b>3.81</b>  |
| 8   | 93.0       | 95.9              | 94.2               | 20.64    | <b>7.48</b>  |
| 10  | 95.0       | 96.5              | 94.9               | 38.50    | <b>13.38</b> |

**Table 9.**  $\pm 40^\circ$  rotation split into 200 perturbations evaluated on MNIST. The verification rate is just 15 %, but 82.1 % of individual splits can be verified.

| Method        | $m$ | splits matched [%] | $t$ [s]     |
|---------------|-----|--------------------|-------------|
| Baseline      | -   | -                  | 11.79       |
| Proof Sharing | 1   | 38.0               | 9.15        |
|               | 2   | 41.1               | 9.21        |
|               | 3   | 58.5               | <b>8.34</b> |

## 5.4 Discussion

We have shown that proof sharing can achieve speed-ups over conventional execution. While the speed-up analysis (see Sect. 4 and Table 7) put a ceiling on what is achievable in particular settings, we are optimistic that proof sharing can be an important tool for neural network robustness analysis. In particular, as the size of certifiable neural networks continues to grow, the potential for gains via proof sharing is equally growing. Further, the idea of proof effort reuse can enable efficient verification of larger disjunctive specifications such as the patch or geometric examples considered here. Besides the immediately useful speed-ups, the concept of proof sharing is interesting in its own right and can provide insights into the learning mechanisms of neural networks.

## 6 Related Work

Here, we briefly discuss conceptually related work:

*Incremental Model Checking* The field of model checking aims to show whether a formalized model, e.g. of software or hardware, adheres to a specification. As neural network verification can also be cast as model checking, we review incremental model checking techniques which utilize a similar idea to proof sharing: reuse partial previous computations when checking new models or specifications. Proof sharing has been applied for discovering and reusing lemmas when proving theorems for satisfiability [6], Linear Temporal Logic [7], and modal  $\mu$ -calculus [33].

Similarly, caching solvers [35] for Satisfiability Modulo Theories cache obtained results or even the full models used to obtain the solution, with assignments for all variables, allowing for faster verification of subsequent queries. For program analysis tasks that deal with repeated similar inputs (e.g. individual commits in a software project) can leverage partial results [41], constraints [36] precision information [4,5] from previous runs.

*Proof Sharing Between Networks.* In neural network verification, some approaches abstract the network to achieve speed-ups in verification. These simplifications are constructed in a way that the proof can be adapted for the original neural network [1,43]. Similarly, another family of approaches analyzes the difference between two closely related neural networks by utilizing their structural similarity [26,27]. Such approaches can be used to reuse analysis results between neural network modifications, e.g. fine-tuning [9,37].

In contrast to these works, we do not modify the neural network, but achieve speed-ups rather by only considering the relaxations obtained in the proofs. [37] additionally consider small changes to the input, however, these are much smaller than the difference in specification we consider here.

## 7 Conclusion

We introduced the novel concept of proof sharing in the context of neural network verification. We showed how to instantiate this concept, achieving speed-ups of up to 2 to 3 x for patch verification and geometric verification. We believe that the ideas introduced in this work can serve as a solid foundation for exploring methods that effectively share proofs in neural network verification.

## References

1. Ashok, P., Hashemi, V., Křetínský, J., Mohr, S.: DeepAbstract: neural network abstraction for accelerating verification. In: Hung, D.V., Sokolsky, O. (eds.) ATVA 2020. LNCS, vol. 12302, pp. 92–107. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-59152-6\\_5](https://doi.org/10.1007/978-3-030-59152-6_5)
2. Bak, S., Liu, C., Johnson, T.T.: The second international verification of neural networks competition. arXiv preprint abs/2109.00498 (2021)
3. Balunovic, M., Baader, M., Singh, G., Gehr, T., Vechev, M.T.: Certifying geometric robustness of neural networks. In: Neural Information Processing Systems (NIPS) (2019)
4. Beyer, D., Löwe, S., Novikov, E., Stahlbauer, A., Wendler, P.: Precision reuse for efficient regression verification. In: Symposium on the Foundations of Software Engineering (SIGSOFT) (2013)
5. Beyer, D., Wendler, P.: Reuse of verification results - conditional model checking, precision reuse, and verification witnesses. In: Bartocci, E., Ramakrishnan, C.R. (eds.) SPIN 2013. LNCS, vol. 7976, pp. 1–17. Springer, Heidelberg (2013). [https://doi.org/10.1007/978-3-642-39176-7\\_1](https://doi.org/10.1007/978-3-642-39176-7_1)

6. Bradley, A.R.: SAT-based model checking without unrolling. In: Jhala, R., Schmidt, D. (eds.) VMCAI 2011. LNCS, vol. 6538, pp. 70–87. Springer, Heidelberg (2011). [https://doi.org/10.1007/978-3-642-18275-4\\_7](https://doi.org/10.1007/978-3-642-18275-4_7)
7. Bradley, A.R., Somenzi, F., Hassan, Z., Zhang, Y.: An incremental approach to model checking progress properties. In: International Conference on Formal Methods in Computer-Aided Design (FMCAD) (2011)
8. Brown, T.B., Mané, D., Roy, A., Abadi, M., Gilmer, J.: Adversarial patch. arXiv preprint [abs/1712.09665](https://arxiv.org/abs/1712.09665) (2017)
9. Cheng, C., Yan, R.: Continuous safety verification of neural networks. In: Design, Automation and Test in Europe Conference and Exhibition (2021)
10. Chiang, P., Ni, R., Abdelkader, A., Zhu, C., Studer, C., Goldstein, T.: Certified defenses for adversarial patches. In: Proceedings of International Conference on Learning Representations (ICLR) (2020)
11. Cousot, P., Cousot, R.: Abstract interpretation: A unified lattice model for static analysis of programs by construction or approximation of fixpoints. In: Proceedings of Principles of Programming Languages (POPL) (1977)
12. Cousot, P., Halbwachs, N.: Automatic discovery of linear restraints among variables of a program. In: Proceedings of Principles of Programming Languages (POPL) (1978)
13. Fischer, M., Baader, M., Vechev, M.T.: Certified defense to image transformations via randomized smoothing. In: Neural Information Processing Systems (NIPS) (2020)
14. 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: Symposium on Security and Privacy (S&P) (2018)
15. Goubault, E., Putot, S.: A zonotopic framework for functional abstractions. *Formal Methods Syst. Des.* **47**(3), 302–360 (2016). <https://doi.org/10.1007/s10703-015-0238-z>
16. Goyal, S., et al.: On the effectiveness of interval bound propagation for training verifiably robust models. arXiv preprint [abs/1810.12715](https://arxiv.org/abs/1810.12715) (2018)
17. Katz, G., Barrett, C., Dill, D.L., Julian, K., Kochenderfer, M.J.: Reluplex: an efficient SMT solver for verifying deep neural networks. In: Majumdar, R., Kunčák, V. (eds.) CAV 2017. LNCS, vol. 10426, pp. 97–117. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-63387-9\\_5](https://doi.org/10.1007/978-3-319-63387-9_5)
18. Katz, G., et al.: The Marabou framework for verification and analysis of deep neural networks. In: Dillig, I., Tasiran, S. (eds.) CAV 2019. LNCS, vol. 11561, pp. 443–452. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-25540-4\\_26](https://doi.org/10.1007/978-3-030-25540-4_26)
19. Kopetzki, A., Schürmann, B., Althoff, M.: Methods for order reduction of zonotopes. In: Conference on Decision and Control (CDC) (2017)
20. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
21. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Neural Information Processing Systems (NIPS) (2012)
22. LeCun, Y., et al.: Handwritten digit recognition with a back-propagation network. In: Neural Information Processing Systems (NIPS) (1989)
23. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: Proceedings of International Conference on Learning Representations (ICLR) (2018)

24. Mirman, M., Gehr, T., Vechev, M.T.: Differentiable abstract interpretation for provably robust neural networks. In: Proceedings of International Conference on Machine Learning (ICML), vol. 80 (2018)
25. Paszke, A., et al.: Pytorch: an imperative style, high-performance deep learning library. In: Neural Information Processing Systems (NIPS) (2019)
26. Paulsen, B., Wang, J., Wang, C.: RELUDIFF: differential verification of deep neural networks. In: International Conference on Software Engineering (ICSE) (2020)
27. Paulsen, B., Wang, J., Wang, J., Wang, C.: NEURODIFF: scalable differential verification of neural networks using fine-grained approximation. In: Conference on Automated Software Engineering (ASE) (2020)
28. Pei, K., Cao, Y., Yang, J., Jana, S.: Towards practical verification of machine learning: the case of computer vision systems. arXiv preprint abs/1712.01785 (2017)
29. Sadraddini, S., Tedrake, R.: Linear encodings for polytope containment problems. In: Conference on Decision and Control (CDC) (2019)
30. Silver, D., et al.: Mastering the game of go without human knowledge. *Nature* **550**(7676) (2017)
31. Singh, G., Gehr, T., Mirman, M., Püschel, M., Vechev, M.T.: Fast and effective robustness certification. In: Neural Information Processing Systems (NIPS) (2018)
32. Singh, G., Gehr, T., Püschel, M., Vechev, M.T.: An abstract domain for certifying neural networks. *PACMPL* **3**(POPL) (2019)
33. Sokolsky, O.V., Smolka, S.A.: Incremental model checking in the modal mu-calculus. In: Dill, D.L. (ed.) CAV 1994. LNCS, vol. 818, pp. 351–363. Springer, Heidelberg (1994). <https://doi.org/10.1007/3-540-58179-0-67>
34. Szegedy, C., et al.: Intriguing properties of neural networks. In: Proceedings of International Conference on Learning Representations (ICLR) (2014)
35. Taljaard, J., Geldenhuys, J., Visser, W.: Constraint caching revisited. In: Lee, R., Jha, S., Mavridou, A., Giannakopoulou, D. (eds.) NFM 2020. LNCS, vol. 12229, pp. 251–266. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-55754-6\\_15](https://doi.org/10.1007/978-3-030-55754-6_15)
36. Visser, W., Geldenhuys, J., Dwyer, M.B.: Green: reducing, reusing and recycling constraints in program analysis. In: Symposium on the Foundations of Software Engineering (SIGSOFT) (2012)
37. Wei, T., Liu, C.: Online verification of deep neural networks under domain or weight shift. arXiv preprint abs/2106.12732 (2021)
38. Weng, T., et al.: Towards fast computation of certified robustness for ReLu networks. In: Proceedings of International Conference on Machine Learning (ICML), vol. 80 (2018)
39. Wing, J.M.: Trustworthy AI. *Commun. ACM* **64**(10) (2021)
40. Wong, E., Kolter, J.Z.: Provable defenses against adversarial examples via the convex outer adversarial polytope. In: Proceedings of International Conference on Machine Learning (ICML), vol. 80 (2018)
41. Yang, G., Dwyer, M.B., Rothermel, G.: Regression model checking. In: International Conference on Software Maintenance (ICSM) (2009)
42. Zhang, H., Weng, T., Chen, P., Hsieh, C., Daniel, L.: Efficient neural network robustness certification with general activation functions. In: Neural Information Processing Systems (NIPS) (2018)
43. Zhong, Y., Ta, Q.-T., Luo, T., Zhang, F., Khoo, S.-C.: Scalable and modular robustness analysis of deep neural networks. In: Oh, H. (ed.) APLAS 2021. LNCS, vol. 13008, pp. 3–22. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-89051-3\\_1](https://doi.org/10.1007/978-3-030-89051-3_1)

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://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.

