Reliable and Interpretable Artificial Intelligence
project report

Simone Barbaro, Guillaume Wang

December 2019 (HS2019)

1 Zonotope representation and transformation

We represent a zonotope Z by its center $a_0 \in R^d$ and a non-negative tensor $A \in R^{k \times d}$ representing the coefficient of the k error terms.

Zonotope propagation through the neural network is straightforward using the transformations presented during the course. For convolutional layers, it suffices to apply the convolution to A itself (excluding bias). The proof is not reproduced here due to space restrictions.

2 Loss function

Let $[o_0, o_2, ...o_9]$ be the output layer of the neural network (the logits for the MNIST digit classification). Let Z be the zonotope region at the output layer, for a given input region, and for given ReLU-transformation parameters λ.

Then the network is verifiably robust on the input region if:

$$\forall (o_0, ..., o_9) \in Z, \forall i \in \{0, ..., 9\}, o_i \leq o_t$$
$$\forall (x_0, ..., x_9) \in Z', \forall i \in \{0, ..., 9\}, x_i \leq 0$$
$$\max_{x \in Z'} \max_i x_i \leq 0$$
$$\max_i \max_{x \in Z'_i} x_i \leq 0$$

where Z' is the zonotope of the "violations" $[o_0 - o_t, ..., o_0 - o_t]$, and Z'_i is the zonotope of $o_i - o_t$.

Since Z'_i are one-dimensional zonotopes, the innermost max can be computed in $O(1)$ by assigning all the error terms to the sign of the corresponding coefficients.

Recall that Z'_i depends on the ReLU-transformation parameters λ. The loss function:

$$L(\lambda) = \max_i \max_{x_i \in Z'_i} x_i$$

can be computed by propagating the input zonotope through the network. The network is verifiably robust if there exists λ such that $L(\lambda) \leq 0$.

3 Learning lambda

$L(\lambda)$ is differentiable, so we use gradient-based methods to minimize it.