Reliable and Interpretable Artificial Intelligence project report

Simone Barbaro, Guillaume Wang

1 Verification and lambdas optimization

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 relaxation of the neural network applied to the input image.

Then the network is robust to the input image if $\forall (o_0, ..., o_9) \in Z, o_i < o_t \forall i$.

Let $Z'_i = Z_i - Z_t$ for $i = 0...9$. Then the condition specified above is equivalent to $\forall i \forall x_i \in Z'_i \; x_i < 0$. Which is true $\iff$ $\max_i \max_{x_i \in Z'_i} x_i < 0$.

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

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

Then we can compute $L(Z)$ after propagating the input zonotope through the network and check whether or not $L(Z) < 0$.

If it is, we can return that the input is verified, otherwise we can use a gradient based algorithm to optimize the lambdas of the ReLU transformations with respect to $L(Z)$ since this loss function is differentiable. Our search for the optimal lambdas consists on repeating this step until $L(Z) < 0$.

2 Zonotope transformations for neural network layers

As we have already seen in class, linear transformations can be computed exactly for zonotopes. Linear and normalization layers compute linear transformations of their input, so they are both exact. Flatten layers simply reshape the input and for zonotopes it's enough to reshape the error terms in the same way to obtain an exact transformation.

ReLU transformations are computed as indicated in the project description.

Finally, convolution layers can be compute exactly as well by showing that a convolution on a zonotope is equivalent to applying the same convolution to the center and to each error tensor separately and then applying the eventual bias of the convolutional layer to the center only.