Exercise 8

Abstract Interpretation of Neural Networks

Program Analysis for System Security and Reliability 2018
ETH Zürich

May 5, 2018

Recall that the activation function $\mathrm{ReLU}(x) = \left\{ \begin{array}{ll} x & x \geq 0 \\ 0 & \text{otherwise.} \end{array} \right\}$

Let N_θ be a two input and one hidden layer neural network defined as

$$N(x_1, x_2) = \mathrm{ReLU}(\mathrm{ReLU}(-x_1 + x_2 + 2) + \mathrm{ReLU}(x_1 - 2x_2))$$

Problem 1. Evaluate the neural network on the input $(x_1, x_2) = (-1, 1)$

Solution. 4

Problem 2. Write a short program to compute the interval analysis of this network.

Solution.

```
def relu(x):
    return [max(0,x[0]), max(0,x[1])]
def mult(s,x):
    return [min(s * x[0], s * x[1]), max(s * x[0], s * x[1])]
def add(x,y):
    return [ x[0] + y[0], x[1] + y[1]]
def addS(x,s):
    return [ x[0] + s, x[1] + s ]
def net(x,y):
    n1 = relu(add(mult(-1,x), addS(y,2)))
    n2 = relu(add(x, mult(-2, y)))
    return relu(add(n1, n2))
```

Problem 3. Apply standard interval analysis to find $N^{\#}([0, 2] \times [0, 1])$.

Solution. $[0, 5]$

Problem 4. Use interval analysis creatively to demonstrate that $5 \notin N([0,2] \times [0,1])$.

Solution. Break the intervals down into two domains, $d_1 = [0,1] \times [0,1]$ and $d_2 = [1,2] \times [0,1]$. We can use our program to find that $N^{\#}(d_1) = [1,4]$ and $N^{\#}(d_2) = [0,4]$. Thus

$$\begin{aligned} N([0,2] \times [0,1]) &= N(d_1 \cup d_2) \\ &= N(d_1) \cup N(d_2) \subseteq N^{\#}(d_1) \cup N^{\#}(d_2) \\ &\subseteq [0,4] \end{aligned}$$

Problem 5. Represent every Interval domain perfectly as a Zonotope.

Solution. Suppose $B \in$ Intervals. Define the Zonotope $a \in$ Zonotopes as follows:

$$\begin{aligned} a_0^x &= \frac{B_1^x + B_0^x}{2} \\ a_x^x &= \frac{B_1^x - B_0^x}{2} \\ a_y^x &= 0 \; \forall y \neq x \end{aligned}$$

Problem 6. Suppose we were to define the point-wise ReLU transformer for Zonotope as follows: $\text{ReLU}_n^{\#}(a)_i^n = \text{ReLU}(a_i^n)$ for error terms i and variables n. Prove that this is unsound.

Solution. All we need is the one dimensional zonotope, a where $a_0^x = 0$ and $a_1^x = -1$. $\hat{x} = -\xi_1$ Define b as $\text{ReLU}_x^{\#}(< \hat{x} >)$. Then $b_0^x = b_1^x = 0$, which implies that only 0 is a member of this zonotope, which is strict subset of $\text{ReLU}(\gamma(a))$ since $1 \in \gamma(a)$ and $\text{ReLU}(1) = 1$.

Problem 7. Design a very simple and sound point-wise transformer for ReLU for Zonotopes, and show that it is at least as precise as Interval.

Solution. The simplest is to find the interval concretization's lower and upper bound for the dimension to which ReLU is being applied, and then convert the ReLU of this interval back into a Zonotope. The proof of precision follows from the derivation of the interval concretization for Zonotope.

Specifically, let $l = a_0^x - \sum_{i=0}^{n} |a_i^x|$ and $u = a_0^x + \sum_{i=0}^{n} |a_i^x|$.

$$
\begin{aligned}
b_0^x &= \frac{u+l}{2} \\
b_{n+1}^x &= \frac{u-l}{2} \\
b_i^x &= 0 && \forall i \leq n \\
b_i^y &= a_i^y && \forall i \leq n \wedge y \neq x \\
b_{n+1}^y &= 0 && \forall y \neq x
\end{aligned}
$$