Exercise 7

Robustness of Neural Networks

Program Analysis for System Security and Reliability 2018
ETH Zurich

April 27, 2018

In this exercise, we will consider the verification of *robustness* property for neural networks. We first define robustness below:

Robustness A robustness property for a neural network is a pair (X, C) consisting of a robustness region X and robustness condition C. A network is robust if all samples from X satisfy the condition C.

Consider the neural network in Figure 1 which takes input x at neuron v_{11} and produces output y at neuron v_{31} and has one hidden layer with two neurons v_{21} and v_{22}. The labels on the edges show the weights of the connections and the biases are assumed to be zero.

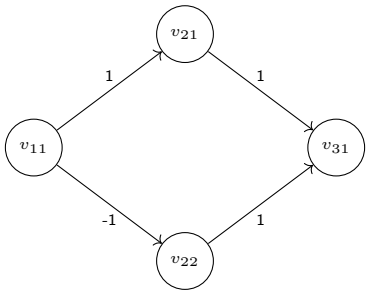

Figure 1: Original neural network.

For convenience, we split both neurons in the hidden layer into two to represent affine transformation and RELU separately. The modified network is shown in Figure 2.

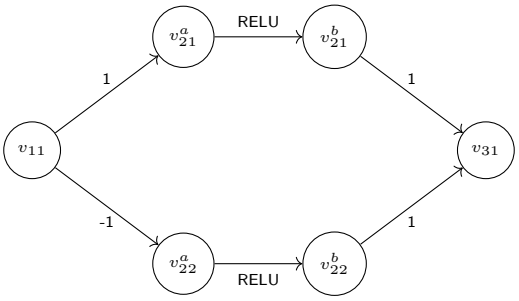

Figure 2: Neural network after splitting the neurons in the hidden layer.

Consider,

$$X = \{x \in \mathbb{R} \mid x \in [0.5, 1]\} \text{ and,}$$
$$C = \{y \in \mathbb{R} \mid y \in (0.4, 1]\}.$$

Show that the given neural network is robust for (X, C) using Reluplex.