Reliable and Interpretable Artificial Intelligence

Course Project

Rohit Kaushik, Pratyush Singh

Our Approach:

We have defined ϵ's for each of the 784 pixels and choose the slope (λ) for the minimum zonotope area. However, we have not trained λ to find its optimal value. But our approach to train lambda was to choose the loss function using the difference between the lower bound of the target minus the sum of the upper bound of the other outputs. And then apply the backpropagation algorithm to find the optimal λ value. The approach seems viable to us but took a time greater than 2 minutes because of which we had to exclude the λ-training process.

We calculate bound by calculation sum of absolute values of ϵ's and then for lower bound we subtract it from the actual value and for the upper bound we add it to the actual value.

We have defined three abstract layers for Linear, Convolutional and Relu Layer. The Relu Layer makes use of DeepZ approximation. We calculate lower bound and upper bound for each of the neuron in Relu Layer and use the appropriate approximation. We had to define Linear and Convolutional abstract layer since the forward function will be slightly different. The difference is because of the fact that bias would only be added for forwarding of image batch and not for error terms. We have used pytorch vectorized operations so that forwards are fast.