Engineering A Verifier for Deep Neural Networks

ThanhVu (Vu) Nguyen

October 15, 2024 (latest version available on Github)

Preface

Having been involved in PhD admission committees for many years, I've realized that many **international** students, especially those in smaller countries or less well-known universities, lack a clear understanding of the Computer Science PhD admission process at US universities. This confusion not only discourages students from applying but also creates the perception that getting admitted to a CS PhD program in the US is difficult compared to other countries.

So I want to share some details about the admission process and advice for those who are interested in applying for a **PhD in Computer Science in the US**. Originally, this document was intended for international students, but I have expanded it to include information that might also be useful for *US domestic students*. Moreover, while this is primarily intended for students interested in CS, it might be relevant to students from various STEM (Science, Technologies, Engineering, and Mathematics) disciplines. Furthermore, although many examples are specifics for schools that I and other contributors of this document know about, the information should be generalizable to other R1[1] institutions in the US.

This information can also help **US faculty and admission committee** gain a better understanding of international students and their cultural differences. By recognizing and leveraging these differences, CS programs in the US can attract larger and more competitive application pools from international students.

I wish you the best of luck. Happy school hunting!

> This document will be updated regularly to reflect the latest information and updates in the admission process. Its latest version is available at
>
> nguyenthanhvuh.github.io/phd-cs-us/demystify.pdf,
>
> and its LaTeX source is also on GitHub. If you have questions or comments, feel free to create new GitHub issues or discussions.

[1] An R1 institution in the US is a research-intensive university with a high level of research activity across various disciplines. Currently, 146 (out of 4000) US universities are classified as R1.

Contents

Chapter 1

Basic of Neural Network

A *neural network* (**NN**) [**?**] consists of an input layer, multiple hidden layers, and an output layer. Each layer has a number of neurons, each connected to neurons from previous layers through a predefined set of weights (derived by training the network with data). A **DNN** is an NN with at least two hidden layers.

The output of a DNN is obtained by iteratively computing the values of neurons in each layer. The value of a neuron in the input layer is the input data. The value of a neuron in the hidden layers is computed by applying an *affine transformation* to values of neurons in the previous layers, then followed by an *activation function* such as the popular Rectified Linear Unit (ReLU) activation.

For this activation, the value of a hidden neuron y is $ReLU(w_1v_1+\ldots+w_nv_n+b)$, where b is the bias parameter of y, $w_i, \ldots, w_n$ are the weights of y, $v_1, \ldots, v_n$ are the neuron values of preceding layer, $w_1v_1 + \cdots + w_nv_n + b$ is the affine transformation, and $ReLU(x) = \max(x, 0)$ is the ReLU activation. The values of a neuron in the output layer is evaluated similarly but it may skip the activation function. A ReLU activated neuron is said to be *active* if its input value is greater than zero and *inactive* otherwise.

DNN Verification Given a DNN N and a property ϕ, the *DNN verification problem* asks if ϕ is a valid property of N. Typically, ϕ is a formula of the form $\phi_{in} \Rightarrow \phi_{out}$, where ϕ_{in} is a property over the inputs of N and ϕ_{out} is a property over the outputs of N. A DNN verifier attempts to find a *counterexample* input to N that satisfies ϕ_{in} but violates ϕ_{out}. If no such counterexample exists, ϕ is a valid property of N. Otherwise, ϕ is not valid and the counterexample can be used to retrain or debug the DNN [**?**].

Example Fig. **??** shows a simple DNN with two inputs x_1, x_2, two hidden neurons x_3, x_4, and one output x_5. The weights of a neuron are shown on its incoming edges , and the bias is shown above or below each neuron. The outputs of the hidden neurons are computed the affine transformation and ReLU, e.g., $x_3 = ReLU(-0.5x_1+0.5x_2+$

1.0). The output neuron is computed with just the affine transformation, i.e., $x_5 = -x_3 + x_4 - 1$.

A valid property for this DNN is that the output is $x_5 \leq 0$ for any inputs $x_1 \in [-1, 1], x_2 \in [-2, 2]$. An invalid property for this network is that $x_5 > 0$ for those similar inputs. A counterexample showing this property violation is $\{x_1 = -1, x_2 = 2\}$, from which the network evaluates to $x_5 = -3.5$. Such properties can capture *safety requirements* (e.g., a rule in an collision avoidance system in [**?**, **?**] is "if the intruder is distant and significantly slower than us, then we stay below a certain threshold") or *local robustness* [**?**] conditions (a form of adversarial robustness stating that small perturbations of a given input all yield the same output).

1.1 Affine Transformation

1.2 Activation Function

1.3 Properties of Neural Networks

1.3.1 Robustness

1.3.2 Safety

Chapter 2

Verification of Neural Networks

2.1 Complexity

Chapter 3

Search Algorithms

Chapter 4

Constraint Solving

4.1 SMT

4.2 MILP

Chapter 5

Abstraction

5.1 Interval

5.2 Zotope

5.3 Polytope

Chapter 6

Popular Techniques and Tools

Chapter 7

Verifying the Verifiers

Chapter 8

Conclusion

Bibliography