Fuzzy Techniques Explain the Effectiveness of ReLU Activation Function in Deep Learning

Julio Urenda[1], Vladik Kreinovich[2], and Olga Kosheleva[3]

[1,2]Departments of [1]Mathematics, [2]Computer Science and [3]Teacher Education
University of Texas at El Paso, El Paso, Texas 79968, USA
jcurenda@utep.edu, vladik@utep.edu, olgak@utep.edu

1. Problem

- In an artificial neural network, each neuron transforms the inputs $x_1, \ldots, x_n$ into a value $y = s(w_1 \cdot x_1 + \ldots + w_n \cdot x_n + w_0..$
- Here w_i are real numbers and $s(x)$ is a function – usually strictly increasing – that is called *activation function.*
- Traditionally, neural networks used a function $s(x) = 1/(1+\exp(-x))$ coming from biological neurons.
- Lately, it turns out that the use of a Rectified Linear Unit (ReLU) $s(x) = \max(0, x)$ leads to a much more effective learning.
- The use of ReLU was one of the factors contributing to the spectacular successes of deep learning.
- Why ReLU is so effective is, however, to a large extent a mystery.
- This is the problem that we deal with in this talk.

2. Towards an explanation: first step

- It is natural to require that if the values x and x' are close, then the values $s(x)$ and $s(x')$ should be close.
- In fuzzy terms, this means that:
  - the degree of closeness between $s(x)$ and $s(x')$ should be at least as large as
  - the degree of closeness between x and x':

$$\mu(|s(x) - s(x')|) \geq \mu(|x - x'|).$$

- Here $\mu(|x - x'|)$ is the membership function corresponding to "close".
- Since the function μ is clearly decreasing, this implies that

$$|s(x) - s(x')| \leq |x - x'|.$$

- In terms of the derivative $s'(x)$, this means that the absolute value of this derivative should not exceed 1.
- Since $s(x)$ is monotonic, we should have $0 \leq s'(x) \leq 1$.

3. Towards an explanation: final step

- In general, according to calculus, the maximum of a function $F(X)$ in a given region is attained:
  - either at the local maximum, where all partial derivatives of this function are 0s,
  - or at the border of this region.
- In situations where we have many constraints, each of which decreases the region size, the resulting region is very small.
- So, the probability that it contains a local maximum is also very small.
- Thus, in most cases, the maximum is attained at the border, i.e., where (at least) one the inequality constraints turns into an equality.
- Thus, to find the maximum of a function in a region, it is, in most cases, sufficient to find its maximum on the border of this region.

4. Towards an explanation: final step (cont-d)

- We can apply the same argument to the function $F(X)$ restricted to the border.
- We then conclude that most probably, the maximum is attained when another inequality constraint becomes an equality, etc.
- In general, the maximum is most probably attained at the point where most – if not all – inequality constraints turn into equalities.
- Let us apply this conclusion to our case.
- The set of all activation functions is determined by inequality constraints $0 \le s'(x) \le 1$ corresponding to different x; thus:
  - whatever optimality criterion we use,
  - the optimal activation function most probably corresponds to the situation when each of these inequalities turns into an equality,
  - i.e., when for each x, we either have $s'(x) = 0$ or $s'(x) = 1$.

5. Towards an explanation: final step (cont-d)

- In regions where $s'(x) = 0$, the function $s(x)$ is constant.
- In regions where $s'(x) = 1$, we have $s(x) = x + c$ for some constant c.
- Thus, the optimal activation function must consists of regions in which it is either constant or have the form $s(x) = x + c$.
- We cannot have only one such region: then $s(x)$ would be linear, and we will not be able to represent non-linear functions.
- The simplest case when we have non-linear functions is when we have two regions:
  - on one of them $s(x)$ is constant,
  - on another one $s(x) = x + c$.
- Each such two-region function is linearly equivalent to ReLU.
- Thus, we indeed have a fuzzy-based explanation for the success of ReLU.

6. Acknowledgments

This work was supported in part by:

- National Science Foundation grants 1623190, HRD-1834620, HRD-2034030, and EAR-2225395;
- AT&T Fellowship in Information Technology;
- program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478, and
- a grant from the Hungarian National Research, Development and Innovation Office (NRDI).