Why Rectified Linear Neurons: Two Convexity-Related Explanations

Jonatan Contreras[1], Martine Ceberio[1], Olga Kosheleva[1], Vladik Kreinovich[1], and Nguyen Hoang Phuong[2]

[1]University of Texas at El Paso, 500 W. University
El Paso, Texas 79968, USA
jmcontreras2@utep.edu, mceberio@utep.edu, olgak@utep.edu
vladik@utep.edu
[2]Thang Long University, Hanoi, Vietnam
nhphuong2008@gmail.com

1. Rectified linear neurons are very successful

- At present, the most successful machine learning technique is deep neural networks.
- In general, in neural networks, signals go through two types of transformations:
  - linear transformations and
  - non-linear transformation described by the so-called *activation function* $x \mapsto s(x)$.
- Deep neural networks mostly used *rectified linear* (ReLU) activation functions

$$s_0(x) = \max(0, x).$$

- The main reason for this choice is that empirically, these activation function have been most successful.

2. But why are they successful?

- From the theoretical viewpoint, this empirical success is a challenge.
- Why are these activations functions more successful than others?
- Are there activation functions that we have not tried yet – which will be even more successful?

3. Important comment

- We have linear transformations before and after each application of an activation function; so:
  - the same results that we obtain by using rectified linear activation function $s_0(x)$
  - can also obtained by neurons that use shifted and scale versions of this function:

$$s_1(x) = b_0 + b_1 \cdot x + b_2 \cdot s_0(a_0 + a_1 \cdot x).$$

- Here, $s_1(x) = s_0 + a_- \cdot (x - x_0)$ for $x \le x_0$, and

$$s_1(x) = s_0 + a_+ \cdot (x - x_0) \text{ for } x \ge x_0.$$

4. What is known and what we do in this paper

- There are some theoretical explanations of why rectified linear neurons are so successful.
- It was proven that the rectified linear activation functions are, in some reasonable sense, optimal.
- This explanation is based on the idea that:
  - the relative quality of different data processing techniques,
  - in particular, the relative quality of neural networks using different activation functions,
  - should not change if we change all the numerical values by changing the measuring units
  - and/or by changing the starting points for measuring the corresponding quantities.
- In this talk, we provide yet another theoretical explanation for this empirical success.

5. Need for optimization

- In practice, we always want to find the best possible solution.
- In precise terms, which solution is better and which is worse is usually described in numerical terms:
  - by assigning a number to each possible solution,
  - so that a solution with the largest (or smallest) value of this numerical characteristic is the best.
- The mapping that assigns such a number to each alternative x is known as the *objective function* $f(x)$.
- A company tries to maximize its profit.
- An environmental agency tries to minimize the overall pollution, etc.

6. Need for optimization (cont-d)

- In general, as the above examples show, we can have both maximization and minimization problems.
- However, the problem of maximizing an objective function $f(x)$ is equivalent to minimizing the function

$$g(x) \stackrel{\text{def}}{=} -f(x).$$

- Thus, all optimization problems can be easily reduced to minimizations.
- So, without losing generality, mathematicians usually only talk about minimization problems.

7. Need for convex optimization

- In general, optimization is NP-hard; this means that:
  - unless P=NP (which most computer scientists believe to be impossible),
  - no feasible algorithm can solve all optimization problems.
- There is an important class of optimization problems:
  - for which optimization is feasible:
  - the class of all *convex* optimization problems.
- There, the minimized functions $f(x)$ is convex, i.e., for all x, x', and $\alpha \in [0,1]$:

$$f(\alpha \cdot x + (1-\alpha) \cdot x') \le \alpha \cdot f(x) + (1-\alpha) \cdot f(x').$$

8. Need for convex optimization (cont-d)

- Moreover, it has been proven that:
  - convex functions are, in some reasonable sense,
  - the largest class of functions for which optimization is feasible.
- Once we add some non-convex functions to this problem, the optimization problem becomes NP-hard.
- This result will underlie our two explanations.

9. How is all this related to neural networks

- One of the main applications of neural networks is to make decisions.
- For this purpose, we need to train the neural network to predict:
  - for each possible action,
  - the consequences of this action.
- In other words, we want:
  - given the parameters x that characterize the possible decision,
  - to compute the value $f(x)$ of the objective function that characterizes this decision.

10. Relation to neural networks (cont-d)

- For the simplest neural networks, this means that:
  - we approximate the original function $f(x_1, \ldots, x_n)$
  - by a linear combination of the output of non-liner neurons:

$$f(x_1, \ldots, x_n) = \sum_{k=1}^{K} W_k \cdot s\left(\sum_{i=1}^{n} w_{ki} \cdot x_i - w_{k0}\right) - W_0.$$

- For multi-layer neural networks, the corresponding expression is more complicated.

11. Towards resulting natural requirements on the activation function

- First, we train the neural network to compute the value of the objective function.
- A natural next step is to find the alternative x that minimizes this objective function.
- As we have mentioned, optimization is only feasible for convex objective functions.
- So, it makes sense to make sure that the neural approximation preserve convexity as much as possible.
- In other words:
  - if the actual activation function is convex,
  - we want this convexity to be, in some reasonable sense, preserved in an approximating expressions.

12. First requirement

- The above idea means, in particular, that:
  - for the simplest case when one neuron is sufficient,
  - the activation function $s(x)$ itself must be convex.
- The rectified linear activation function itself is convex, so it satisfies this requirement.
- On the other hand, there are many other convex functions.
- So this requirement does not uniquely determine the rectified linear function.
- For this unique determination, we need to come up with additional requirement(s).

13. Second requirement

- It is known that if functions $f_1(x), \ldots, f_n(x)$ are convex, then their convex combination is also convex:

$$f(x) = w_1 \cdot f_1(x) + \ldots + w_K \cdot f_K(x), \quad w_k \geq 0, \quad \sum_{k=1}^{K} w_k = 1.$$

- Moreover, any linear combination with non-negative coefficients is convex.
- On the other hand:
  - if we allow even one of the coefficients to be negative,
  - then we already get non-convex functions.

14. Second requirement (cont-d)

- So:
  - the only way to make sure that a linear combination of convex functions is convex
  - is to make sure that all the coefficients w_k are non-negative.
- It is therefore reasonable to require that:
  - every convex function $f(x)$ – at least every convex function of one variable,
  - be representable as a linear combination of activation functions with non-negative coefficients.
- This is our second requirement.
- Let us analyze what are the activation functions that satisfy this requirement.

15. Let us recall the usual calculus-based characteristics of convexity

- It is known that:
  - a differentiable function $f(x)$ is convex
  - if and only if its second derivative $f''(x)$ is everywhere non-negative $f''(x) \geq 0$.
- Not all convex functions are everywhere differentiable.
- E.g., the rectified linear activation function $s_0(x)$ is not differentiable at the point $x = 0$.
- However, for such function, we can consider, as derivatives, *generalized functions*.
- They are also known as *Schwartz distributions*.
- They are limits of usual functions.

16. Convexity (cont-d)

- The most well-known generalized function is a *delta-function* $\delta(x)$:
  - which is equal to 0 for all $x \neq 0$ and
  - which tends to ∞ at $x = 0$.
- Such functions are used in physics to describe, e.g., point-wise particles and objects.
- In particular:
  - the derivative $s_0'(x)$ of the rectified linear function is equal to 0 for $x \leq 0$ an to 1 for $x > 0$, and
  - the second derivative is exactly the delta-function.
- For a linear combination of functions,
  - its second derivative is equal to
  - the linear combination of its second derivatives:

$$f''(x) = w_1 \cdot f_1''(x) + \ldots + w_K \cdot f_K''(x).$$

17. Convexity (cont-d)

- So, in terms of the second derivatives, the above second requirement means that:
  - every non-negative (generalized) function can be represented as a linear combination of t
  - the functions corresponding to second derivative of the activation function $s(x)$
  - and of its shifted and scaled versions $s(a_0 + a_1 \cdot x)$.

18. Now we can prove that only rectified linear activation f-n satisfies both requirements

- Suppose that the second derivative $s''(x)$ differs from 0 for at least two different values $x \neq x'$.
- Then this property remains true for any convex combination of shifted and scaled versions of $s(x)$.
- Thus, this way, we will never get $f(x)$ for which $f''(x)$ is non-zero only for one value x.
- For example, we will never get the rectified linear function $s_0(x)$.
- On the other hand:
  - if we select the rectified linear function $s_0(x)$ as an activation function,
  - then we have $s_0''(x) = \delta(x)$.

19. Proof (cont-d)

- In this case, any $f''(x) \geq 0$ can be represented as a linear combination of shifted versions of $s_0''(x)$:

$$f''(x) = \int f''(y) \cdot \delta(x - y)\, dy = \int f''(y) \cdot s_0''(x - y)\, dy.$$

- Thus, the function $f(x)$:
  - can be represented as a similar linear combination of the shifted versions of $s_0(x)$
  - plus possibly some linear terms:

$$f(x) = b_0 + b_1 \cdot x + \int f''(y) \cdot s_0(x - y)\, dy.$$

- In general, our second requirement is satisfied:
  - by any convex function
  - for which the second derivative is different from 0 only for one value $x = x_0$.

20. Proof (cont-d)

- This second derivative can therefore be described as

$$s''(x) = c \cdot \delta(x - x_0) \text{ for some } c > 0.$$

- Integrating this equality twice, we conclude that

$$s(x) = b_0 + b_1 \cdot x + c \cdot s_0(x - x_0).$$

- One can check that this is exactly the desired expression; so indeed:
  - the above two natural convexity-related requirements
  - naturally lead to the rectified linear activation functions.

21. Let us consider a more general setting

- Out of the above two requirements:
  - the first one looks more convincing,
  - the second one is somewhat less convincing.
- Let us therefore consider a more general setting.
- We still postulate the first requirement – i.e., we still consider only convex activation functions,
- However:
  - instead of postulating the second requirement,
  - we want to find the activation function which is the best in some sense.

22. General setting (cont-d)

- This means that:
  - the corresponding objective functional $F(s)$ – describing the relative qualities of different $s(x)$,
  - attains its smallest possible value.

23. What calculus tells us

- In general, a maximum or minimum of a function on a multi-D domain is attained:
  - either inside this domain – it which case it is a stationary point of this function,
  - or on its boundary.
- When the domain is relatively small:
  - the probability that a global stationary point is inside this domain is very small,
  - so it is reasonable to assume that the minimum is attained on the boundary.
- This general conclusion can be applied to our case.
- We optimize a functional $F(s)$ on the domain of all convex functions s.
- Most functions are not convex.

24. What calculus tells us (cont-d)

- So, in the space of all possible functions, the domain of all convex functions is indeed small.
- Similarly:
  - if the domain's boundary contains a flat face-type part – as when the domain is a polytope,
  - then it is reasonable to assume that the minimum is attained not in the interior of this face,
  - it is attained on its boundary.
- In general, we can conclude that:
  - the minimum is most probably attained at one of the *extreme points* of the original domain,
  - i.e., at a point that cannot be represented as a convex combination of other points from this domain.

25. What this implies for optimal activation functions

- We want to select an activation function.
- In this case, the domain is the set of all convex functions.
- What are the extreme elements of this domain?
- We have already shown that:
  - any convex function $s(x)$ whose second derivative differs from 0 at least 2 different points can be
  - represented as a convex combination of other convex functions – shifted rectified linear ones.

26. What this implies for optimal activation functions (cont-d)

- Hence, such functions $s(x)$ are not extreme elements of our domain; thus:
  - the only extreme elements of this domain are
  - convex functions whose second derivative differs from 0 only at one point.
- These functions are, as we have shown, exactly rectified linear functions.
- With high probability, only extreme elements can be optimal.
- So, we conclude that with high probability:
  - only rectified linear functions can be optimal,
  - no matter what optimality criterion we used.

27. Acknowledgments

This work was supported in part by:

- the National Science Foundation grants:
  - 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in CS), and
  - HRD-1834620 and HRD-2034030 (CAHSI Includes),
- and by the AT&T Fellowship in Information Technology.

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.