Towards Fast and Understandable Computations: Which "And"- and "Or"-Operations Can Be Represented by the Fastest (i.e., 1-Layer) Neural Networks? Which Activations Functions Allow Such Representations?

Kevin Alvarez[1], Julio C. Urenda[1,2], Orsolya Csiszár[3,4], Gábor Csiszár[6], József Dombi[7], György Eigner[5], and Vladik Kreinovich[1]

Departments of [1]Computer Science and [2]Mathematical Sciences
University of Texas at El Paso, El Paso, TX 79968, USA
kalvarez9@miners.utep.edu, jcurenda@utep.edu, vladik@utep.edu
[3]Faculty of Basic Sciences, University of Applied Sciences Esslingen
Esslingen, Germany
[4]Institute of Applied Mathematics, Óbuda University
[5]Institute of Biomatics and Applied Artificial Intelligence, Óbuda University
Budapest, Hungary,
orsolya.csiszar@nik.uni-obuda.hu
eigner.gyorgy@nik.uni-obuda.hu
[6]Institute of Materials Physics, University of Stuttgart
Stuttgart, Germany
gabor.csiszar@mp.imw.uni-stuttgart.de
[7]Institute of Informatics, University of Szeged
Szeged, Hungary, dombi@inf.u-szeged.hu

Abstract: We want computations to be fast, and we want them to be understandable. As we show, the need for computations to be fast naturally leads to neural networks, with 1-layer networks being the fastest, and the need to be understandable naturally leads to fuzzy logic and to the corresponding "and"- and "or"-operations. Since we want our computations to be both fast and understandable, a natural question is: which "and"- and "or"-operations of fuzzy logic can be represented by the fastest (i.e., 1-layer) neural network? And a related question is: which activation functions allow such a representation? In this paper, we provide an answer to both questions: the only "and"- and "or"-operations that can be thus represented are $\max(0, a+b-1)$ *and* $\min(a+b, 1)$*, and the only activations functions allowing such a representation are equivalent to the rectified linear function – the one used in deep learning. This result provides an additional explanation of why rectified linear neurons are so successful. With also show that with full 2-layer networks, we can compute practically any*

"and"- and "or"-operation.

Keywords: neural networks, fuzzy logic, "and"- and "or"-operations, rectified linear neurons, explainable AI

1 Formulation of the Problem

1.1 What we plan to do in this section

In this section, we not only explain our problem – we also explain *why* this problem is, in our opinion, very important.

We do not just want to formulate a technical problem listed in the title of this paper – we want to explain, from scratch, why we use neural networks and fuzzy techniques, and why it is important to relate these techniques.

We hope that these explanations will motivate the readers to continue research in this direction – in particular, to solve open problems that we listed at the end of this paper.

1.2 Computations are needed

In many application areas, we need to process data. Because of this need, computers are ubiquitous. What do we want from the computation results? First of all, we want them to be correct:

- if we are predicting weather, we want these predictions to be mostly successful,
- if we are deciding whether to give a loan to a bank's customer, we want to be sure that customers who get the loans have a high chance of repaying them, and that most customers to whom the program decided not to give the loan will not become very successful – and thus will not present our missed opportunities.

Coming up with such an algorithm is not easy, this is the main challenge. But once we have this algorithm, there are two other important challenges.

1.3 Two important challenges: computation speed and understandability

First, in most practical problems, we need to process a large amount of data – and we need to make a decision reasonably fast:

- if we predict weather, we need to take into account all the results of today's measurements of temperature, wind speed and direction, etc., in a given geographic areas, satellite images, historical data – and get the prediction of tomorrow's weather the same day: otherwise, our prediction will be useless;

- if we decide whether to give a person a loan, we need to take into account this person's financial history, financial history of similar customers, general economic situation in the region, etc. – and get the result fast, otherwise the customer may lose the business opportunity for which he/she is seeking this loan.

So, we need all the computations to be as fast as possible.

We also ideally want the computations to be understandable.

- When a weatherperson on the TV predict's tomorrow's weather, it is much more convincing if this person explains why we should expect strong winds, or, vice versa, perfect weather. These explanations may not be quantitative, usually, qualitative explanations are good enough.
- When we explain, to the person, why he/she is not getting a loan while his/her friends are, we need to have some reasonable explanations – at least to avoid lawsuits claiming gender-based, age-based, or race-based bias.

How can we achieve these two goals?

1.4 Need for fast computations leads to neural networks

A natural way to speed up computations is to perform them in parallel. In the past, only high-performance super-computers had several processors working in parallel, but nowadays, parallelism is ubiquitous: even the cheapest computers have up to four processors working in parallel. In parallel computations, all that matters is how fast computations can be performed on one of the processors – since computations on other processors are performed at the same time.

Which computations are fast? In general, computers process numbers, so, in general, any computation takes numerical inputs $x_1, \ldots, x_n$ – e.g., measurement results – and converts them into one or more numerical values y. In mathematics, a situation when to each input $x = (x_1, \ldots, x_n)$ there corresponds the result is known as a *function*, so we can say that each processor computes some function $y = f(x_1, \ldots, x_n)$.

Which functions are the easier to compute? Functions can be linear or non-linear. In general, linear functions, i.e., functions of the type

$$f(x_1, \ldots, x_n) = w_0 + w_1 \cdot x_1 + \ldots + w_n \cdot x_n \tag{1}$$

are the easiest to compute, so let us keep them in our list of easiest-to-compute functions. However, we cannot just limit ourselves to linear functions, because otherwise, if we only apply linear transformations, you will only get linear functions, but in real life, many dependencies are nonlinear. So, we need some nonlinear functions as well.

Which nonlinear functions are the easiest to compute? In general, the more inputs the function has, the longer it takes to process all these inputs. Thus, the easiest to compute are functions of one variable $y = s(z)$.

So, we arrive at the following computation scheme:

- first, each processor applies the fastest – linear – transformation to the data, i.e., computes the value $z = w_0 + w_1 \cdot x_1 + \ldots + w_n \cdot x_n$;
- if this is not enough, we apply the fastest non-linear transformation and compute $y = s(z)$; as a result, we get the value

$$y = s(w_0 + w_1 \cdot x_1 + \ldots + w_n \cdot x_n); \tag{2}$$

- then, if needed, we apply another linear transformation, then another nonlinear one, etc.

As a result, we get a layered computation scheme in which on each layer, each pair of processors computes the values (2), and then the results from these pairs become inputs to another layer, etc.

This scheme is what is usually known as a *neural network*; readers interested in more details can see, e.g., [6, 11, 23]. A two-part component computing the expression (2) is known as a *neuron*, and the non-linear function $s(z)$ is known as the *activation function*. So, the need for fast computations has indeed led us to neural networks. The fewer layers, the faster computations: 1-layer networks are the fastest, 2-layer networks are second fastest. This is especially important if we implement neural networks in hardware; see, e.g., [1].

Of course, to make sure that neural networks are useful, we need to check that neural networks can indeed describe any possible continuous dependence with any desired accuracy, i.e., in precise terms, that for every continuous function $y = f(x_1, \ldots, x_n)$ on a bounded domain and for every desired accuracy $\varepsilon > 0$, there exists a function which is ε-close to $f(x_1, \ldots, x_n)$ and which can be represented by a neural network. Such *universal approximation* results are indeed known for many different activation functions; see, e.g., [6, 7, 19, 26].

Neural networks have been very successful in practical applications. Which activation function should we use? Traditionally, the most widely used neural networks used *sigmoid* activation functions $s(z) = 1/(1 + \exp(-z))$. Lately, it turned out that even more successful are *deep* neural networks [11] that use *rectified linear* functions $s(z) = \max(0, z)$.

Comments.

- Deep neural networks not only use a different activation function, they also use a large number of layers. This makes the computations somewhat slower that for traditional "shallow" (few-layers) neural networks, but this slowing down is needed to provide a better approximation accuracy; see, e.g., [4, 21, 23, 22] for a detailed explanation of this need.
- Another case when sacrificing speed can improve accuracy is recurrent neural networks that work iteratively: Hopfield networks [13], Elman networks [10], Kohonen's self organizing maps [16], fuzzy cognitive maps (see, e.g., [27, 49]), and other similar schemes (see, e.g., [9]).

- How many neurons do we need to get a good approximation? The space of all continuous function is infinite-dimensional. This means, crudely speaking, that to precisely describe a generic function, we need to use infinitely many parameters. The more parameters we use, the more accurately we can approximate each function. For neural networks, this means that the more neurons we allow, the more accurate is the resulting approximation.
- How can we prove universal approximation results? Many of these proofs use Stone's generalization [46] of the classical Weierstrass's Theorem [52] according to which each continuous function can be approximated, with any given accuracy, by a polynomial.
- Interestingly, by using appropriate activation functions, we can get not only an ε-*approximation* to the desired function $f(x_1,\ldots,x_n)$, but also the *exact* representation of this function. This possibility follows from the unexpected Kolmogorov's solution [17] to the 13th Hilbert problem [12], one of the 23 problems that 19 century mathematicians left to the 20th century to solve. According to Kolmogorov's theorem, every continuous function on a bounded domain can be represented as a composition of addition and functions of one variable; see, e.g., [30, 31]. This result – as well as its improvements and generalizations such as [36, 45] – underlies the theorems about exact representation of functions by neural networks; see, e.g., [28].

  It is worth mentioning that the corresponding activation functions cannot be smooth. This fact relates these functions to another Weiestrass's result – that there exist continuous functions which are nowhere differentiable [51]. Weierstrass's functions are actually historically the first examples of what is now called a *fractal*; see, e.g., [33].
- It is also worth mentioning that the universal approximation result for neural networks has applications beyond neural networks themselves: e.g., it can explain complexity of collective decision making [48] and – on the qualitative level – the existence and properties of quarks [20].

1.5 Need for understandability leads to fuzzy techniques

Neural networks can compute any dependence – and we can train them to fit any given data, but the problem is that the resulting recommendations come with no justification. As we have mentioned, it is desirable to make our recommendations understandable – i.e., justified, explainable by words from natural language.

Understandability means that we should be able to describe the computations by using words from natural language. One of the main challenges in coming up with such a description is that natural language is imprecise (fuzzy), so it is difficult to find the relation between imprecise words from natural language and precise algorithms. In solving this challenge, it is natural to use the experience of researchers who came up with such a relationship from the other side of it: by trying to translate natural-language knowledge into precise terms.

This experience led to the design on fuzzy logic by Lotfi Zadeh; readers interested in

more details can see, e.g., [5, 15, 34, 42, 43, 55]. Lotfi Zadeh, a specialist in control and an author of a successful textbook on control, noticed, in the early 1960s, a puzzling phenomenon: that human-led control often leads to much better results than even the optimal automatic control. The answer to this puzzle was clear: humans use additional knowledge which was not taken into account when the automatic controllers were designed. The reason why this additional knowledge was not taken into account is that this knowledge is not described in precise terms, it is described by using imprecise words from natural language. For example, an operator may say: if the pressure drops a little bit, increase a little bit the flow of the chemical into the chamber; here, "a little bit" does not have a precise meaning. Zadeh invented a methodology for translating this "fuzzy" knowledge into precise terms, a methodology that he called *fuzzy logic*, or, more generally, *fuzzy techniques*.

His main point is that in contrast to exact statements like "pressure is below 1.2 atmospheres" – which is always either true or false – about the statements that include natural-language words – like "the drop from 1.3 to 1.2 means that the pressure dropped a little bit" – experts are not sure. The smaller the drop, the larger the expert's degree of confidence that this statement is true. For each value of the corresponding quantity (e.g., pressure), we can gauge the expert's degree of confidence in the corresponding statement by asking the expert to mark it on a scale, e.g., from 0 to 10. The resulting mark depends on what scale we use: from 0 to 5 or from 0 to 10 or form 0 to any other number. To make these estimates uniform, a reasonable idea is to divide the mark by the largest number on the scale, so that, e.g., 7 on a scale from 0 to 10 becomes $7/10 = 0.7$. In this new scale, 1 means that the expert is absolutely confident that this statement is true, 0 means that the expert is absolutely confident that the statement is false, and values between 0 and 1 correspond to intermediate degrees of confidence.

The reason why this methodology is called fuzzy *logic* is that in addition to simple statements – like the ones above – expert knowledge often contains statements that include *logical connectives* like "and" and "or". For example, an expert can recommend a certain action if the pressure dropped a little bit *and* the temperature increased somewhat. How can we gauge our degree of certainty in such composite statements? It would be great if we could similarly ask the expert to estimate his/her degree of confidence for all possible pairs of values (pressure, temperature). If we have a composite statement combining three or four different statements, we would need to consider all possible triples or quadruples. Even if we consider a reasonable number 20-30 of possible values of each quantity, it makes sense to ask the expert about all 30 values, but asking about all $30^4 = 810000$ possible quadruples is not realistic. Since we cannot directly elicit the degree of confidence in all such composite statements directly from the expert, we need to be able to estimate this degree based on whatever information we can elicit – i.e., based on the expert's degrees of confidence in the component statements.

In precise terms, we need a procedure that would take, as input, the degrees of confidence a and b in two statements A and B and return an estimate for the expert's degree of confidence in a composite statement $A\,\&\,B$. We will denote this estimate by $f_{\&}(a,b)$. The corresponding function $f_{\&}$ is known as an *"and"-operation*, or,

for historical reason, a *t-norm*.

Since the statements "A and B" and "B and A" mean the same thing, it is reasonable to require that for these two statements, we have the same degree of confidence, i.e., that $f_{\&}(a,b) = f_{\&}(b,a)$. In other words, an "and"-operation must be commutative.

When A is false, clearly $A\,\&\,B$ is false too, so we must have $f_{\&}(0,b) = 0$ for all b. When A is true, our degree of confidence in $A\,\&\,B$ is the same as our degree of confidence in B, i.e., we must have $f_{\&}(1,b) = b$.

Similarly, we need a procedure that would take, as input, the degrees of confidence a and b in two statements A and B and return an estimate for the expert's degree of confidence in a composite statement $A \vee B$. We will denote this estimate by $f_{\vee}(a,b)$. The corresponding function $f_{\vee}$ is known as an *"or"-operation*, or, for historical reason, a *t-conorm*.

Since the statements "A or B" and "B or A" mean the same thing, it is reasonable to require that for these two statements, we have the same degree of confidence, i.e., that $f_{\vee}(a,b) = f_{\vee}(b,a)$. In other words, an "or"-operation must be commutative.

When A is true, clearly $A \vee B$ is true too, so we must have $f_{\vee}(1,b) = 1$ for all b. When A is false, our degree of confidence in $A \vee B$ is the same as our degree of confidence in B, i.e., we must have $f_{\vee}(0,b) = b$.

Fuzzy logic can help translate expert rules of the type "if $A_i(x)$ then $B_i(u)$" related the input x with the control value u – rules that are formulated by using natural-language terms $A_i(x)$ and $B_i(u)$ (such as "x is small") – into precise recommendations. Indeed, for any given input x, the value u is a reasonable control if one of the rules is applicable, i.e., if either $A_1(x)$ is true and $B_1(u)$ holds, or $A_2(x)$ is true and $B_2(u)$ holds, etc.:

$$(A_1(x)\,\&\,B_1(u)) \vee (A_2(x)\,\&\,B_2(u)) \vee \ldots$$

We can elicit, from the expert, degrees to which the statements $A_i(x)$ and $B_i(u)$ hold for different values x and u – the resulting functions are known as *membership functions.* After that, we can use appropriate "and"- and "or"-operations to come up with a degree to which, for given input x, the control u is reasonable. Then, if needed, we can combine these degrees into a single recommendation $\bar{u}(x_1,\ldots,x_n)$ corresponding to the given input $(x_1,\ldots,x_n)$.

It is know that functions $\bar{u}(x_1,\ldots,x_n)$ corresponding to different rules and different membership functions are also universal approximators; see, e.g., [2, 3, 8, 18, 24, 25, 37, 40, 44, 50, 53, 54].

Comments.

- Similarly to the case of neural networks, the more rules we allow, the more accurate is the approximation: if we fix the number of rules, we can only achieve a limited approximation accuracy; see, e.g., [14, 35, 47].
- Similar universal approximation results are known for fuzzy neural networks that combine fuzzy and neural techniques; see, e.g., [29, 32].

- Also similarly to the neural network case, it is possible not only to *approximate* any continuous function by an appropriate system, but also to represent any function *exactly* – by using non-smooth ("fractal") membership functions motivated by the above-mentioned Kolmogorov's theorem; see, e.g., [38, 39].

1.6 Natural questions

As we have mentioned earlier, we want our computations to be both fast and understandable. Understandable means that we have to use some "and"- and "or"-operations. We thus want these operations to be fast. The fastest possible computations are computations on a 1-layer neural network, in which thus "and"-operation is computed by a single neuron, and in which the "or"-operation can also be computed by a single neuron. So, natural questions are:

- which "and"- and "or"-operations can be computed by a 1-layer neural network, and
- what activation functions allow computing "and"- and "or"-operations by such neural networks.

1.7 What we do in this paper

In this paper, we provide answers to both questions, namely:

- we show that the only "and"- and "or"-operations which can be computed by a 1-layer neural network are $\max(0, a+b-1)$ and $\min(a+b, 1)$, and
- we show that the only activation function allowing such fast computations are equivalent to *rectified linear neurons* – which probably provides some explanations for the current success of such activation functions.

We also show that if we allow linear pre-processing after a single neuron, then we also represent $\min(a,b)$ and $\max(a,b)$. If we allow several neurons in a 2-layer network, then, in effect, we can compute any "and"- and "or"-operations.

2 Definitions and the Main Results

Definition 1. *By an* "and"-operation*, we mean a function*

$$f_{\&} : [0,1] \times [0,1] \to [0,1] \tag{3}$$

for which the following properties are satisfied:

- $f_{\&}(a,b) = f_{\&}(b,a)$ *for all a and b,*
- $f_{\&}(0,b) = 0$ *and* $f_{\&}(1,b) = b$ *for all b.*

Definition 2. *By an* "or"-operation*, we mean a function*

$$f_{\vee} : [0,1] \times [0,1] \to [0,1] \tag{4}$$

for which the following properties are satisfied:

- $f_\vee(a,b) = f_\vee(b,a)$ *for all* a *and* b,
- $f_\vee(0,b) = b$ *and* $f_\vee(1,b) = 1$ *for all* b.

Comment. Usually, for both "and"- and "or"-operations, other properties are required as well – namely, continuity, monotonicity, and associativity – but for our main results, we do not need these additional properties.

Definition 3. *We say that a function* $f(x_1,\ldots,x_n)$ *can be represented by a 1-layer neural network if this function can be represented in the form*

$$f(x_1,\ldots,x_n) = s(w_0 + w_1 \cdot x_1 + \ldots + w_n \cdot x_n) \tag{5}$$

for some function $s(z)$ *and for some values* w_i. *The corresponding function* $s(z)$ *is called an* activation function.

Definition 4. *By a* rectified linear *function, we mean a function*

$$s_0(z) = \max(0,z). \tag{6}$$

Definition 5. *We say that two activation functions* $s_1(z)$ *and* $s_2(z)$ *are* equivalent *if for some constants* a_{ij} *and* b_{ij}, *we have*

$$s_1(z) = a_{10} + a_{12} \cdot s_2(b_{10} + b_{11} \cdot z) + a_{1z} \cdot z \tag{7}$$

and

$$s_2(z) = a_{20} + a_{21} \cdot s_1(b_{20} + b_{21} \cdot z) + a_{2z} \cdot z \tag{8}$$

for all z.

Comment. This way, the corresponding multi-layer neural networks represent, in effect, the same class of functions, since each non-linear layer is equivalent to adding extra linear transformations before and after the non-linear layer representing another activation function.

Theorem 1. *The only "and"-operation that can be represented by a 1-layer neural network is* $\max(0,a+b-1)$, *and all activation functions allowing such a representation are equivalent to the rectified linear function.*

Theorem 2. *The only "or"-operation that can be represented by a 1-layer neural network is* $\min(a+b,1)$, *and all activation functions allowing such a representation are equivalent to the rectified linear function.*

Comment. These results provide another explanation for why rectified linear activation functions are so successful in deep neural networks.

2.1 Proof of Theorem 1

Let us consider an "and"-operation $f_{\&}(a,b)$ which can be represented by a 1-layer neural network. By definition of such a representation, this means that $f_{\&}(a,b) = s(w_0 + w_a \cdot a + w_b \cdot b)$ for some function $s(z)$ and for some coefficients w_i.

By definition of an "and"-operation, we have $f_{\&}(a,b) = f_{\&}(b,a)$ for all a and b. Thus, the expression $s(w_0 + w_a \cdot a + w_b \cdot b)$ should not change if we swap a and b: $s(w_0 + w_a \cdot a + w_b \cdot b) = s(w_0 + w_a \cdot b + w_b \cdot a)$. Therefore, we must have $w_a = w_b$, i.e., $f_{\&}(a,b) = s(w_0 + w_a \cdot a + w_a \cdot b)$, and thus,

$$f_{\&}(a,b) = s(w_0 + w_a \cdot (a+b)). \tag{9}$$

Let us introduce an auxiliary function $t(z) \stackrel{\text{def}}{=} s(w_0 + w_a \cdot z)$. This function is, by the definition of equivalence, equivalent to $s(z)$. In terms of this auxiliary function, the formula (9) takes the following simplified form:

$$f_{\&}(a,b) = t(a+b). \tag{10}$$

For $a = 0$, by definition of an "and"-operation, we have $f_{\&}(0,b) = 0$ for all $b \in [0,1]$, thus $t(z) = 0$ for all $z \in [0,1]$.

For $a = 1$, by definition of an "and"-operation, we have $f_{\&}(1,b) = b$ for all $b \in [0,1]$, thus $t(1+b) = b$ for all $b \in [0,1]$. For $z = 1+b$, we have $z \in [1,2]$ and $b = z-1$, thus $t(z) = z-1$ for all $z \in [1,2]$. So, we have:

- $t(z) = 0$ for $z \in [0,1]$, and
- $t(z) = z-1$ for $z \in [1,2]$.

These two cases can be combined into a single formula

$$t(z) = \max(0, z-1). \tag{11}$$

Substituting this expression for $t(z)$ into the formula (10), we conclude that $f_{\&}(a,b) = \max(0, a+b-1)$. So, this "and"-operation is indeed the only one that can be represented by a 1-layer neural network.

Which activation functions can be used for this representation? From the formula (11), we can see that $t(z)$ is indeed equivalent to the rectified linear activation function. Since the original function $s(z)$ is equivalent to $t(z)$, we can conclude that $s(z)$ is also equivalent to the rectified linear activation function. Thus, the 1-layer representation of an "and"-operation is only possible if we use rectified linear neurons.

The theorem is proven.

2.2 Proof of Theorem 2

Let us now consider an "or"-operation $f_{\vee}(a,b)$ which can be represented by a 1-layer neural network. By definition of such a representation, this means that $f_{\vee}(a,b) = s(w_0 + w_a \cdot a + w_b \cdot b)$ for some function $s(z)$ and for some coefficients w_i.

By definition of an "or"-operation, we have $f_\vee(a,b) = f_\vee(b,a)$ for all a and b. Thus, the expression $s(w_0 + w_a \cdot a + w_b \cdot b)$ should not change if we swap a and b: $s(w_0 + w_a \cdot a + w_b \cdot b) = s(w_0 + w_a \cdot b + w_b \cdot a)$. Therefore, we must have $w_a = w_b$, i.e., $f_\vee(a,b) = s(w_0 + w_a \cdot a + w_a \cdot b)$, and thus,

$$f_\vee(a,b) = s(w_0 + w_a \cdot (a+b)). \tag{12}$$

Similar to the proof of Theorem 1, let us introduce an auxiliary function $t(z) \stackrel{\text{def}}{=} s(w_0 + w_a \cdot z)$. This function is, by the definition of equivalence, equivalent to $s(z)$. In terms of this auxiliary function, the formula (12) takes the following simplified form:

$$f_\vee(a,b) = t(a+b). \tag{13}$$

For $a = 0$, by definition of an "or"-operation, we have $f_\vee(0,b) = b$ for all $b \in [0,1]$, thus $t(z) = z$ for all $z \in [0,1]$.

For $a = 1$, by definition of an "or"-operation, we have $f_\&(1,b) = 1$ for all $b \in [0,1]$, thus $t(1+b) = 1$ for all $b \in [0,1]$. For $z = 1+b$, we have $z \in [1,2]$ and $b = z-1$, thus $t(z) = 1$ for all $z \in [1,2]$. So, we have:

- $t(z) = z$ for $z \in [0,1]$, and
- $t(z) = 1$ for $z \in [1,2]$.

These two cases can be combined into a single formula

$$t(z) = \min(z,1). \tag{14}$$

Substituting this expression for $t(z)$ into the formula (13), we conclude that $f_\vee(a,b) = \min(1, a+b)$. So, this "or"-operation is indeed the only one that can be represented by a 1-layer neural network.

Which activation functions can be used for this representation? One can easily see that the expression (14) can be represented in an equivalent form

$$t(z) = 1 - \max(1-z,0), \tag{15}$$

so $t(z)$ is indeed equivalent to the rectified linear activation function. Since the original function $s(z)$ is equivalent to $t(z)$, we can conclude that $s(z)$ is also equivalent to the rectified linear activation function. Thus, the 1-layer representation of an "or"-operation is only possible if we use rectified linear neurons.

The theorem is proven.

3 Two-Layer Networks and the Auxiliary Result

3.1 What about other "and"- and "or"-operations?

In this paper, we have shown that only the operations $f_\&(a,b) = \max(0, a+b-1)$ and $f_\vee(a,b) = \min(a+b,1)$ can be represented by 1-layer neural networks. How many layers do we need to represent general "and"- and "or"-operations?

It is known – see, e.g., [41] – that for every continuous "and"- (or "or"-) operation $f(a,b)$ and for every $\varepsilon > 0$, then exists a function $F(z)$ for which an "and"- (or, respectively, "or"-) operation

$$g(a,b) = F^{-1}(F(a) + F(b)) \tag{16}$$

satisfies the property $|f(a,b) - g(a,b)| \le \varepsilon$ for all a and b. (Of course, for this result to be true, it is not sufficient to have the above simplified definitions of "and"- and "or"-operations: we also need to assume associativity and monotonicity.)

For very small ε, the operations $f(a,b)$ and $g(a,b)$ are practically indistinguishable. So, from practical viewpoint, every "and"-operation and every "or"-operation can be represented in the form (16). Every function of this form can be computed by a 2-layer neural network:

- in the first layer, we use the inputs a and b to compute the values $a' = F(a)$ and $b' = F(b)$;
- then, in the second layer, we compute the value $F^{-1}(a' + b')$, which is exactly the desired value $F^{-1}(F(a) + F(b))$.

So, from the practical viewpoint, every "and"-operation and every "or"-operation can be computed by a 2-layer neural network.

For example, a widely used "and"-operation $f_{\&}(a,b) = a \cdot b$ can be computed as $\exp(\ln(a) + \ln(b))$, with $F(z) = \ln(z)$ and the inverse function $F^{-1}(z) = \exp(z)$. Similarly, a widely used "or"-operation $f_{\vee}(a,b) = a + b - a \cdot b$ can be computed in the form (16) with $F(z) = \ln(1-z)$ and $F^{-1}(z) = 1 - \exp(z)$.

3.2 When is it sufficient to have a single neuron with linear post-processing?

We have shown that, from the practical viewpoint, all "and"- and "or"-operations can be represented by a 2-layer neural network. Interestingly, some "and"- and "or"-operations $f(a,b)$ can be represented by a single neuron if we allow an additional linear post-processing. For example, one can easily see that $\min(a,b) = b - \max(0, b-a)$ and $\max(a,b) = a + \max(0, b-a)$.

It turns out that these are the only "and"- and "or"-operations which can be thus represented.

Definition 6. *We say that a continuous monotonic associative "and"-operation* $f_{\&}(a,b)$ can be computed by a single neuron with linear post-processing *if we have*

$$f_{\&}(a,b) = c_0 + c_a \cdot a + c_b \cdot b + s(w_0 + w_a \cdot a + w_b \cdot b). \tag{17}$$

Definition 7. *We say that a continuous monotonic associative "or"-operation* $f_{\vee}(a,b)$ can be computed by a single neuron with linear post-processing *if we have*

$$f_{\vee}(a,b) = c_0 + c_a \cdot a + c_b \cdot b + s(w_0 + w_a \cdot a + w_b \cdot b). \tag{18}$$

Theorem 3. *The only "and"-operations that can be computed by a single neuron with linear post-processing are* $\max(0, a+b-1)$ *and* $\min(a,b)$. *All activation functions allowing such a computation are equivalent to the rectified linear function.*

Theorem 4. *The only "or"-operations that can be computed by a single neuron with linear post-processing are* $\min(a+b,1)$ *and* $\max(a,b)$. *All activation functions allowing such a computation are equivalent to the rectified linear function.*

3.3 Proof of Theorems 3 and 4

First of all, let us somewhat simplify the expressions (17) and (18) for the corresponding operation $f(a,b)$.

We cannot have $w_a = w_b = 0$ because then, the function $f(a,b)$ would be linear, and it is easy to show that no linear function can satisfy all the requirements of an "and"-operation or of an "or"-operation. Thus, either $w_a \neq 0$ or $w_b \neq 0$ (or both).

If $w_a = 0$, then, due to commutativity of $f(a,b)$, we can swap a and b and get an expression with $w_a \neq 0$. Thus, without losing generality, we can assume that $w_a \neq 0$.

We can thus introduce an auxiliary function $t(z) = c_0 + s(w_0 + w_a \cdot z)$. In terms of this auxiliary function, formulas (17) and (18) take the form

$$f(a,b) = c_a \cdot a + c_b \cdot b + t(a + k \cdot b), \tag{19}$$

where $k \stackrel{\text{def}}{=} w_b / w_a$.

If $k = 1$, then the expression $t(a + k \cdot b)$ is symmetric with respect to a and b. Since for both types of operations, the function $f(a,b)$ is commutative, we thus conclude that the difference

$$c_a \cdot a + c_b \cdot b = f(a,b) - t(a+b) \tag{20}$$

is also commutative. Therefore, $c_a = c_b$, hence the whole expression (19) depends only on the sum $a+b$, i.e., has the form $F(a+b)$ for some function $F(z)$. This means that each such function is computable by a 1-layer neural network, and all "and"- and "or"-operations which can be thus represented have been described in Theorems 1 and 2.

To complete the proof, it is therefore necessary to consider the case when $k \neq 1$, i.e., when the lines $a + k \cdot b = \text{const}$ are not parallel to the diagonal $a = b$ of the square $[0,1] \times [0,1]$. Each line $a + k \cdot b = \text{const}$ intersects the borderline of the square at two points. On the borderline – i.e., when one of the values a and b is equal to 0 or to 1 – the value of an "and"- or "or"-operation is uniquely determined by the corresponding Definition (Definition 1 or Definition 2). Since the function $f(a,b)$ is linear on this line, its values for all the points from this line are uniquely determined by the values at these two borderline points. Thus, for each k, we uniquely determine all the values $f(a,b)$ for all the pairs (a,b).

One can check that the only case when the resulting function is commutative and associative is the case $k = -1$, in which case we indeed get $\min(a,b)$ and $\max(a,b)$. We can also easily check that in both case, the activation function $t(z)$ is indeed equivalent to the rectified linear function. The theorems are proven.

3.4 Remaining open problems

It is known (see, e.g., [6]) that functions represented as linear combinations of the results of 1-neuron layer are universal approximators – i.e., for each continuous function on a bounded domain and for each accuracy $\varepsilon > 0$, we can find a neural network which computes the given function with the desired accuracy. In general, the more accuracy we require, the more neurons we need. So, to achieve perfect accuracy – i.e., exact computations – we will need *potentially infinite* number of neurons.

Interestingly, for some "and"- and "or"-operations, we can have perfect accuracy with a *limited number* of neurons: e.g., the operation $a \cdot b$ can be computed by a 2-neuron network, as

$$a \cdot b = \frac{1}{4} \cdot (a+b)^2 - \frac{1}{4} \cdot (a-b)^2. \tag{21}$$

The operation $a + b - a \cdot b$ can be computed by a 3-neuron network:

$$a + b - a \cdot b = (a+b) - \frac{1}{4} \cdot (a+b)^2 - \frac{1}{4} \cdot (a-b)^2. \tag{22}$$

It would be interesting to describe all such "and"- and "or"-operations. Maybe $a \cdot b$ and $a + b - a \cdot b$ are the only such operations?

4 Conclusions

We would like our computations to be fast and understandable. As we show in this paper, the need for the computations to be fast naturally leads to neural networks, and the need for the computations to be understandable – i.e., describable by words from natural language – naturally invokes techniques relating imprecise natural-language words with numerical recommendations – techniques of fuzzy logic. The need to use both neural and fuzzy techniques necessitates analyzing when fuzzy "and"- and "or"-operations – the main building blocks of fuzzy techniques – can be implemented by the fastest possible (1-layer) neural network, and which activation functions can be used for such an implementation.

Interestingly, the answer is that we need to use min, max, and related fuzzy operations $\min(a+b,1)$ and $\max(a+b-1,0)$ – which are indeed among the most successfully used fuzzy techniques, and the corresponding activation function is the rectified linear function – the activation function which is successfully used in deep learning. These result provide a possible explanation of why neural networks that use rectified linear activation function are so successful.

Acknowledgements

This work was supported in part by the grant TUDFO/47138-1/2019-ITM from the Ministry of Technology and Innovation, Hungary, and by the US National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science) and HRD-1242122 (Cyber-ShARE Center of Excellence).

The authors are greatly thankful to the anonymous referees for their thorough reading and valuable suggestions.

References

[1] N. Ádám, A. Baláž, E. Pietriková, E. Chovancová, and P. Fecil′ak: The impact of data representations on hardware based MLP network implementation, Acta Polytechnica Hungarica, 2018, Vol. 15, No. 2, pp. 69–88.

[2] M. M. Afravi and V. Kreinovich: From fuzzy universal approximation to fuzzy universal representation: it all depends on the continuum hypothesis, Proceedings of the Joint 17th Congress of International Fuzzy Systems Association and 9th International Conference on Soft Computing and Intelligent Systems IFSA-SCIS'2017, Otsu, Japan, June 27–30, 2017.

[3] M. Afravi and V. Kreinovich: Fuzzy systems are universal approximators for random dependencies: a simplified proof, In: M. Ceberio and V. Kreinovich (eds.): Decision Making under Constraints, Springer Verlag, Cham, Switzerland, 2020, pp. 1–5.

[4] C. Baral, O. Fuentes, and V. Kreinovich: Why deep neural networks: a possible theoretical explanation, In: M. Ceberio and V. Kreinovich (eds.): Constraint Programming and Decision Making: Theory and Applications, Springer Verlag, Berlin, Heidelberg, 2018, pp. 1–6.

[5] R. Belohlavek, J. W. Dauben, and G. J. Klir: Fuzzy Logic and Mathematics: A Historical Perspective, Oxford University Press, New York, 2017.

[6] C. M. Bishop: Pattern Recognition and Machine Learning, Springer, New York, 2006.

[7] E. K. Blum and L. K. Li: Approximation theory and feedforward networks, Neural Networks, 1991, Vol. 4, No. 4, pp. 511–515.

[8] J. L. Castro: Fuzzy logic controllers are universal approximators, IEEE Transactions on Systems, Man, and Cybernetics, 1995, Vol. 25, pp. 629–635.

[9] Chiu-Hsiung Chen, Chang-Chih Chung, Fei Chao, Chih-Min Lin, and I. J. Rudas: Intelligent robust control for uncertain nonlinear multivariable systems using recurrent cerebellar model neural networks, Acta Polytechnica Hungarica, 2015, Vol. 12, No. 5, pp. 7–33.

[10] J. L. Elman: Finding structure in time, Cognitive Science, 1990, Vol. 14, pp. 179-211.

[11] I. Goodfellow, Y. Bengio, and A. Courville: Deep Leaning, MIT Press, Cambridge, Massachusetts, 2016.

[12] D. Hilbert: Mathematical Problems, Bulletin of the American Mathematical Society, 1902, Vol. 8, No. 10, pp. 437–479.

[13] J. J. Hopfield: Neural networks and physical systems with emergent collective computational abilities, Proceedings of the National Academy of Sciences of the USA, 1992, Vol. 79, No. 8, pp. 2554–2558.

[14] E. P. Klement, L.T. Kóczy, and B. Moser: Are fuzzy systems universal approximators?, International Journal of General Systems, 1999, Vol. 28, No. 2–3, pp. 259–282.

[15] G. Klir and B. Yuan: Fuzzy Sets and Fuzzy Logic, Prentice Hall, Upper Saddle River, New Jersey, 1995.

[16] T. Kohonen: Self-organized formation of topologically correct feature maps, Biological Cybernetics, 1982, Vol. 43, No. 1, pp. 59–69.

[17] A. N. Kolmogorov: On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition, Dokl. Akad. Nauk. SSSR, 1957, Vol. 114, pp. 953–956 (in Russian).

[18] B. Kosko: Fuzzy systems as universal approximators, In: Proceedings of the IEEE International Conference on Fuzzy Systems FUZZ-IEEE'92, San Diego, California, 1992, pp. 1153–1162.

[19] V. Kreinovich: Arbitrary nonlinearity is sufficient to represent all functions by neural networks: a theorem, Neural Networks, 1991, Vol. 4, 381–383.

[20] V. Kreinovich, Fundamental properties of pair-wise interactions naturally lead to quarks and quark confinement: a theorem motivated by neural universal approximation results, In: M. Ceberio and V. Kreinovich (eds.): How Uncertainty-Related Ideas Can Provide Theoretical Explanation for Empirical Dependencies, Springer, Cham, Switzerland, to appear.

[21] V. Kreinovich: From traditional neural networks to deep learning: towards mathematical foundations of empirical successes, In: S. N. Shahbazova, J. Kacprzyk, V. E. Balas, and V. Kreinovich (eds.): Proceedings of the World Conference on Soft Computing, Baku, Azerbaijan, May 29–31, 2018.

[22] V. Kreinovich and O. Kosheleva: Optimization under uncertainty explains empirical success of deep learning heuristics", In: P. Pardalos, V. Rasskazova, and M. N. Vrahatis (eds.): Black Box Optimization, Machine Learning and No-Free Lunch Theorems, Springer, Cham, Switzerland, to appear.

[23] V. Kreinovich and O. Kosheleva: Deep learning (partly) demystified, Proceedings of the 4th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence ISMSI'2020, Thimpu, Bhutan, April 18–19, 2020.

[24] V. Kreinovich, G. C. Mouzouris, and H. T. Nguyen: Fuzzy rule based modeling as a universal approximation tool, In: H. T. Nguyen and M. Sugeno (eds.): Fuzzy Systems: Modeling and Control, Kluwer, Boston, MA, 1998, pp. 135–195.

[25] V. Kreinovich, H. T. Nguyen, and Y. Yam: Fuzzy systems are universal approximators for a smooth function and its derivatives, International Journal of Intelligent Systems, 2000, Vol. 15, No. 6, pp. 565–574.

[26] V. Kreinovich and O. Sirisaengtaksin: 3-layer neural networks are universal approximators for functionals and for control strategies, Neural, Parallel, and Scientific Computations, 1993, Vol. 1, pp. 325–346.

[27] V. Kreinovich and C. Stylios: Why Fuzzy Cognitive Maps are efficient, International Journal of Computers, Communications, & Control, 2015, Vol. 10, No. 6, pp. 825–833.

[28] V. Kůrková, Kolmogorov's theorem and multilayer neural networks, Neural Networks, 1992, Vol. 5, pp. 501-506.

[29] A. Lemos, V. Kreinovich, W. Caminhas, and F. Gomide: Universal approximation with uninorm-based fuzzy neural networks, Proceedings of the 30th Annual Conference of the North American Fuzzy Information Processing Society NAFIPS'2011, El Paso, Texas, March 18–20, 2011.

[30] G. G. Lorentz: Approximation of Functions, Holt, Reinhard and Winston, New York, 1965.

[31] G. G. Lorentz: The 13th problem of Hilbert, In: F. Browder (ed.): Mathematical Developments Arising from Hilbert's Problems, American Mathematical Society, Providence, Rhode Island, 1976, Vol. 2, pp. 419-430.

[32] R. Lovassy, L. T. Kóczy, and L. Gál: Function approximation performance of fuzzy neural networks, Acta Polytechnica Hungarica, 2010, Vol. 7, No. 4, pp. 25–38.

[33] B. B. Mandelbrot: The Fractal Geometry of Nature, Henry Holt and Company, New York, 1983.

[34] J. M. Mendel: Uncertain Rule-Based Fuzzy Systems: Introduction and New Directions, Springer, Cham, Switzerland, 2017.

[35] B. Moser: Sugeno controllers with a bounded number of rules are nowhere dense, International Journal of General Systems, 1999, Vol. 28, No. 3, pp. 269–277.

[36] M. Nakamura, R. Mines, and V. Kreinovich: Guaranteed intervals for Kolmogorov's theorem (and their possible relation to neural networks), Interval Computations, 1993, No. 3, pp. 183–199.

[37] H. T. Nguyen and V. Kreinovich: On approximations of controls by fuzzy systems, Proceedings of the Fifth International Fuzzy Systems Association World Congress IFSA'93, Seoul, Korea, July 1993, pp. 1414–1417.

[38] H. T. Nguyen and V. Kreinovich: Kolmogorov's Theorem and its impact on soft computing, In: R. R. Yager and J. Kacprzyk (eds.), The Ordered Weighted Averaging Operators: Theory and Applications, Kluwer, Boston, MA, 1997, pp. 3–17.

[39] H. T. Nguyen, V. Kreinovich, and D. Sprecher: Normal forms for fuzzy logic – an application of Kolmogorov's theorem, International Journal on Uncertainty, Fuzziness, and Knowledge-Based Systems, 1996, Vol. 4, No. 4, pp. 331–349.

[40] H. T. Nguyen, V. Kreinovich, and O. Sirisaengtaksin: Fuzzy control as a universal control tool, Fuzzy Sets and Systems, 1996, Vol. 80, No. 1, pp. 71–86.

[41] H. T. Nguyen, V. Kreinovich, and P. Wojciechowski: Strict Archimedean t-norms and t-conorms as universal approximators, International Journal of Approximate Reasoning, 1998, Vol. 18, Nos. 3–4, pp. 239–249.

[42] H. T. Nguyen, C. L. Walker, and E. A. Walker: A First Course in Fuzzy Logic, Chapman and Hall/CRC, Boca Raton, Florida, 2019.

[43] V. Novák, I. Perfilieva, and J. Močkoř: Mathematical Principles of Fuzzy Logic, Kluwer, Boston, Dordrecht, 1999.

[44] I. Perfilieva and V. Kreinovich: A new universal approximation result for fuzzy systems, which reflects CNF–DNF duality, International Journal of Intelligent Systems, 2002, Vol. 17. No. 12, pp. 1121–1130.

[45] D. A. Sprecher: On the structure of continuous functions of several variables, Transactions of the American Mathematical Society, 1965, Vol. 115, pp. 340–355.

[46] M. H. Stone: A generalized Weierstrass approximation theorem, Mathematics Magazine, 1948, Vol. 21, pp. 167–184 and 237–254.

[47] D. Tikk: On nowhere denseness of certain fuzzy controllers containing prerestricted number of rules, Tatra Mountains Mathematical Publications, 1999, Vol. 16, pp. 369–377.

[48] R. Trejo and V. Kreinovich: Complexity of collective decision making explained by neural network universal approximation theorem, In: G. Alefeld and R. A. Trejo (eds.), Interval Computations and its Applications to Reasoning Under Uncertainty, Knowledge Representation, and Control Theory. Proceedings of MEXICON'98, Workshop on Interval Computations, 4th World Congress on Expert Systems, México City, México, 1998.

[49] J. Vaščák and L. Madarász: Function approximation performance of fuzzy neural networks, Acta Polytechnica Hungarica, 2010, Vol. 7, No. 3, pp. 109–122.

[50] L.-X. Wang and J. M. Mendel: Fuzzy basis functions, universal approximation and orthogonal least squares learning, IEEE Transactions on Neural Networks, 1992, Vol. 3, pp. 807–814.

[51] K. Weierstraß: Über continuirliche Functionen eines reellen Arguments,die für keinen Werth des letzeren einen bestimmten Differentialquotienten besitzen (On continuous functions of a real argument which possess a definite derivative for no value of the argument), Königlich Preussichen Akademie der Wissenschaften, 1872; reprinted in: Mathematische Werke von Karl Weierstrass, Mayer & Mueller, Berlin, Germany, 1895, Vol. 2, pp. 71–74.

[52] K. Weierstraß, Über die analytische Darstellbarkeit sogenannter willkürlicher Functionen einer reellen Veränderlichen, Sitzungsberichte der Akademie zu Berlin, 1885, pp. 633–639 and 789–805.

[53] R. R. Yager and V. Kreinovich: Universal approximation theorem for uninorm-based fuzzy systems modeling, Fuzzy Sets and Systems, 2003, Vol. 140, No. 2, pp. 331–339.

[54] Y. Yam, H. T. Nguyen, and V. Kreinovich, Multi-resolution techniques in the rules-based intelligent control systems: a universal approximation result, Proceedings of the 14th IEEE International Symposium on Intelligent Control/Intelligent Systems and Semiotics ISIC/ISAS'99, Cambridge, Massachusetts, September 15–17, 1999, pp. 213–218.

[55] L. A. Zadeh: Fuzzy sets, Information and Control, 1965, Vol. 8, pp. 338–353.