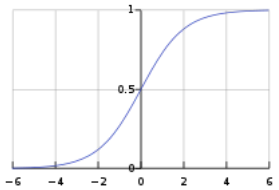
Activation function



Logistic activation function

In artificial neural networks, the **activation function** of a node defines the output of that node given an input or set of inputs. A standard integrated circuit can be seen as a digital network of activation functions that can be "ON" (1) or "OFF" (0), depending on input. This is similar to the behavior of the linear perceptron in neural networks. However, only *nonlinear* activation functions allow such networks to compute nontrivial problems using only a small number of nodes.[1]

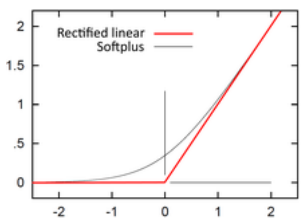
Contents

Functions

In biologically inspired neural networks, the activation function is usually an abstraction representing the rate of action potential firing in the cell.[2] In its simplest form, this function is binary—that is, either the neuron is firing or not. The function looks like $\phi(v_i)=U(v_i)$, where U is the Heaviside step function.

A line of positive slope may be used to reflect the increase in firing rate that occurs as input current increases. Such a function would be of the form $\phi(v_i)=\mu v_i$, where μ is the slope. This activation function is linear and therefore cannot make decisions.

As biological neurons cannot lower their firing rate below zero, rectified linear activation functions are used: $\phi(v_i)=\max(0,v_i)$. They introduce a non-linearity at zero that can be used for decision making.[3]



Rectified linear and Softplus activation functions

Neurons also cannot fire faster than a certain rate. Sigmoid activation functions use a second non-linearity for large inputs: $\phi(v_i)=(1+\exp(-v_i))^{-1}$. Because they are in the range between zero and one, sigmoid activations can be interpreted as probabilities. If a range from -1 to 1 is desired, the sigmoid can be scaled and shifted to yield the hyperbolic tangent activation function: $\phi(v_i)=U(v_i)\tanh(v_i)$.

Alternative structures

A special class of activation functions known as radial basis functions (RBFs) are used in RBF networks, which are extremely efficient as universal function approximators. These activation functions can take many forms, but they are usually found as one of three functions:

- Gaussian: $\phi(v_i)=\exp\left(-\frac{\|v_i-c_i\|^2}{2\sigma^2}\right)$
- Multiquadratics: $\phi(v_i)=\sqrt{\|v_i-c_i\|^2+a^2}$
- Inverse multiquadratics: $\phi(v_i)=(\|v_i-c_i\|^2+a^2)^{-1/2}$

where c_i is the vector representing the function *center* and a and σ are parameters affecting the spread of the radius.

A computationally efficient Radial Basis Function has been proposed,[4] called Square-law based RBF kernel (SQ-RBF (https://www.wolframalpha.com/input/?i=Plot%5B%7BPiecewise%5B%7B%7B1-x%5E2%2F2%2CAbs%5Bx%5D%3C%3D1%7D%2C%7B%282-Abs%5Bx%5D%29%5E2%2F2%2C1%3CAbs%5Bx%5D%3C%3D2%7D%2C%7B0%2CAbs%5Bx%5D%3E2%7D%7D%5D%2CExp%5B-x%5E2%5D%7D%2C%7Bx%2C-2.5%2C2.5%7D%5D)) which eliminates the exponential term as found in Gaussian RBF.

- SQ-RBF: $f(x)=\begin{cases}1-\frac{x^2}{2} & :|x|\le 1\\ \frac{(2-|x|)^2}{2} & :1\le|x|\le 2\\ 0 & :|x|\ge 2\end{cases}$

Support-vector machines (SVMs) can effectively utilize a class of activation functions that includes both sigmoids and RBFs. In this case, the input is transformed to reflect a decision boundary hyperplane based on a few training inputs called *support-vectors* x. The activation function for the hidden layer of these machines is referred to as the *inner product kernel*, $K(v_i,x)=\phi(v_i)$. The support-vectors are represented as the centers in RBFs with the kernel equal to the activation function, but they take a unique form in the perceptron as

$$\phi(v_i)=\tanh\left(\beta_1+\beta_0\sum_j v_{i,j}x_j\right),$$

where β_0 and β_1 must satisfy certain conditions for convergence. These machines can also accept arbitrary-order polynomial activation functions where

$$\phi(v_i)=\left(1+\sum_j v_{i,j}x_j\right)^p.$$ [5]

Activation function having types:

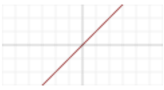
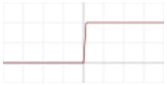
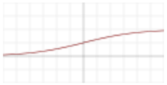
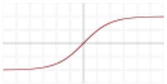
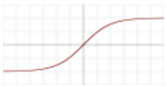
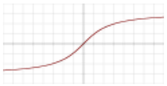
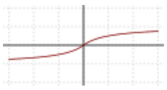
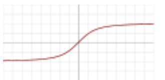
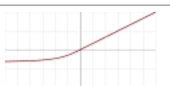
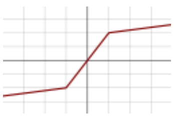
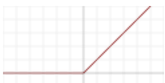
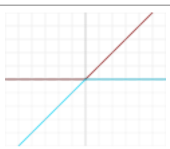
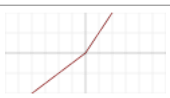
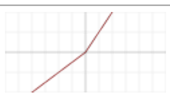
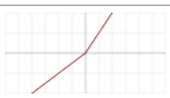
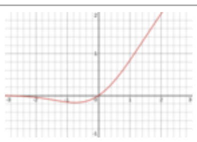
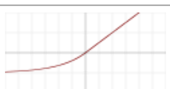
- Identity function
- Binary step function
- Bipolar step function
- Sigmoidal function
    - Binary sigmoidal function
    - Bipolar sigmoidal function
- Ramp function

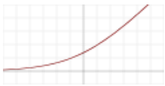
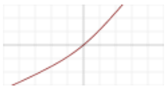
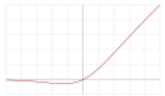
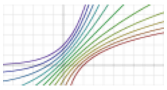
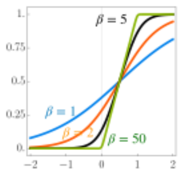
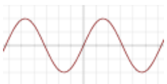
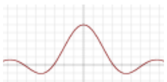
Comparison of activation functions

Some desirable properties in an activation function include:

- Nonlinear – When the activation function is non-linear, then a two-layer neural network can be proven to be a universal function approximator.[6] This is known as the Universal Approximation Theorem. The identity activation function does not satisfy this property. When multiple layers use the identity activation function, the entire network is equivalent to a single-layer model.
- Range – When the range of the activation function is finite, gradient-based training methods tend to be more stable, because pattern presentations significantly affect only limited weights. When the range is infinite, training is generally more efficient because pattern presentations significantly affect most of the weights. In the latter case, smaller learning rates are typically necessary.
- Continuously differentiable – This property is desirable (RELU is not continuously differentiable and has some issues with gradient-based optimization, but it is still possible) for enabling gradient-based optimization methods. The binary step activation function is not differentiable at 0, and it differentiates to 0 for all other values, so gradient-based methods can make no progress with it.[7]
- Monotonic – When the activation function is monotonic, the error surface associated with a single-layer model is guaranteed to be convex.[8]
- Smooth functions with a monotonic derivative – These have been shown to generalize better in some cases.
- Approximates identity near the origin – When activation functions have this property, the neural network will learn efficiently when its weights are initialized with small random values. When the activation function does not approximate identity near the origin, special care must be used when initializing the weights.[9] In the table below, activation functions where $f(0) = 0$ and $f'(0) = 1$ and f' is continuous at 0 are indicated as having this property.

The following table compares the properties of several activation functions that are functions of one fold x from the previous layer or layers:

Name	Plot	Equation	Derivative (with respect to *x*)	Rang
Identity		$f(x)=x$	$f'(x)=1$	$(-\infty,\infty)$
Binary step		$f(x)=\begin{cases}0 & \text{for } x<0\\ 1 & \text{for } x\geq 0\end{cases}$	$f'(x)=\begin{cases}0 & \text{for } x\neq 0\\ ? & \text{for } x=0\end{cases}$	$\{0,1\}$
Logistic (a.k.a. Sigmoid or Soft step)		$f(x)=\sigma(x)=\frac{1}{1+e^{-x}}$ [1]	$f'(x)=f(x)(1-f(x))$	$(0,1)$
TanH		$f(x)=\tanh(x)=\frac{(e^{x}-e^{-x})}{(e^{x}+e^{-x})}$	$f'(x)=1-f(x)^2$	$(-1,1)$
SQNL[10]		$f(x)=\begin{cases}1 & : x>2.0\\ x-\frac{x^2}{4} & : 0\leq x\leq 2.0\\ x+\frac{x^2}{4} & : -2.0\leq x<0\\ -1 & : x<-2.0\end{cases}$	$f'(x)=1\mp\frac{x}{2}$	$(-1,1)$
ArcTan		$f(x)=\tan^{-1}(x)$	$f'(x)=\frac{1}{x^2+1}$	$\left(-\frac{\pi}{2},\frac{\pi}{2}\right)$
ArSinH		$f(x)=\sinh^{-1}(x)=\ln\left(x+\sqrt{x^2+1}\right)$	$f'(x)=\frac{1}{\sqrt{x^2+1}}$	$(-\infty,\infty)$
ElliotSig[11][12] Softsign[13][14]		$f(x)=\frac{x}{1+\lvert x\rvert}$	$f'(x)=\frac{1}{(1+\lvert x\rvert)^2}$	$(-1,1)$
Inverse square root unit (ISRU)[15]		$f(x)=\frac{x}{\sqrt{1+\alpha x^2}}$	$f'(x)=\left(\frac{1}{\sqrt{1+\alpha x^2}}\right)^3$	$\left(-\frac{1}{\sqrt{\alpha}},\right.$
Inverse square root linear unit (ISRLU)[15]		$f(x)=\begin{cases}\frac{x}{\sqrt{1+\alpha x^2}} & \text{for } x<0\\ x & \text{for } x\geq 0\end{cases}$	$f'(x)=\begin{cases}\left(\frac{1}{\sqrt{1+\alpha x^2}}\right)^3 & \text{for } x<0\\ 1 & \text{for } x\geq 0\end{cases}$	$\left(-\frac{1}{\sqrt{\alpha}},\right.$
Piecewise Linear Unit (PLU)[16]		$f(x)=\max(\alpha(x+c)-c,\min(\alpha(x-c)+c,x))$	$f'(x)=\begin{cases}\alpha & \text{for } \lvert x\rvert>c\\ 1 & \text{for } \lvert x\rvert<c\end{cases}$	$(-\infty,\infty)$
Rectified linear unit (ReLU)[17]		$f(x)=\begin{cases}0 & \text{for } x\leq 0\\ x & \text{for } x>0\end{cases}$	$f'(x)=\begin{cases}0 & \text{for } x\leq 0\\ 1 & \text{for } x>0\end{cases}$	$[0,\infty)$
Bipolar rectified linear unit (BReLU)[18]		$f(x_i)=\begin{cases}ReLU(x_i) & \text{if } i \bmod 2=0\\ -ReLU(-x_i) & \text{if } i \bmod 2\neq 0\end{cases}$	$f'(x_i)=\begin{cases}ReLU'(x_i) & \text{if } i \bmod 2=0\\ ReLU'(-x_i) & \text{if } i \bmod 2\neq 0\end{cases}$	$(-\infty,\infty)$
Leaky rectified linear unit (Leaky ReLU)[19]		$f(x)=\begin{cases}0.01x & \text{for } x<0\\ x & \text{for } x\geq 0\end{cases}$	$f'(x)=\begin{cases}0.01 & \text{for } x<0\\ 1 & \text{for } x\geq 0\end{cases}$	$(-\infty,\infty)$
Parameteric rectified linear unit (PReLU)[20]		$f(\alpha,x)=\begin{cases}\alpha x & \text{for } x<0\\ x & \text{for } x\geq 0\end{cases}$	$f'(\alpha,x)=\begin{cases}\alpha & \text{for } x<0\\ 1 & \text{for } x\geq 0\end{cases}$	$(-\infty,\infty)$
Randomized leaky rectified linear unit (RReLU)[21]		$f(\alpha,x)=\begin{cases}\alpha x & \text{for } x<0\\ x & \text{for } x\geq 0\end{cases}$ [3]	$f'(\alpha,x)=\begin{cases}\alpha & \text{for } x<0\\ 1 & \text{for } x\geq 0\end{cases}$	$(-\infty,\infty)$
Gaussian Error Linear Unit (GELU)[22]		$f(x)=x\Phi(x)=x(1+\text{erf}(x/\sqrt{2}))/2$	$f'(x)=\Phi(x)+x\phi(x)$	$(-0.17,\infty$
Exponential linear unit (ELU)[23]		$f(\alpha,x)=\begin{cases}\alpha(e^{x}-1) & \text{for } x\leq 0\\ x & \text{for } x>0\end{cases}$	$f'(\alpha,x)=\begin{cases}f(\alpha,x)+\alpha & \text{for } x\leq 0\\ 1 & \text{for } x>0\end{cases}$	$(-\alpha,\infty)$
Scaled exponential linear unit (SELU)[24]		$f(\alpha,x)=\lambda\begin{cases}\alpha(e^{x}-1) & \text{for } x<0\\ x & \text{for } x\geq 0\end{cases}$ with $\lambda=1.0507$ and $\alpha=1.67326$	$f'(\alpha,x)=\lambda\begin{cases}\alpha(e^{x}) & \text{for } x<0\\ 1 & \text{for } x\geq 0\end{cases}$	$(-\lambda\alpha,\infty)$

S-shaped rectified linear activation unit (SReLU)[25]		$f_{t_l,a_l,t_r,a_r}(x) = \begin{cases} t_l + a_l(x - t_l) & \text{for } x \le t_l \\ x & \text{for } t_l < x < t_r \\ t_r + a_r(x - t_r) & \text{for } x \ge t_r \end{cases}$ t_l, a_l, t_r, a_r are parameters.	$f'_{t_l,a_l,t_r,a_r}(x) = \begin{cases} a_l & \text{for } x \le t_l \\ 1 & \text{for } t_l < x < t_r \\ a_r & \text{for } x \ge t_r \end{cases}$	$(-\infty, \infty)$
Adaptive piecewise linear (APL)[26]		$f(x) = \max(0, x) + \sum_{s=1}^{S} a_i^s \max(0, -x + b_i^s)$	$f'(x) = H(x) - \sum_{s=1}^{S} a_i^s H(-x + b_i^s)$[4]	$(-\infty, \infty)$
SoftPlus[27]		$f(x) = \ln(1 + e^x)$	$f'(x) = \frac{1}{1 + e^{-x}}$	$(0, \infty)$
Bent identity		$f(x) = \frac{\sqrt{x^2 + 1} - 1}{2} + x$	$f'(x) = \frac{x}{2\sqrt{x^2 + 1}} + 1$	$(-\infty, \infty)$
Sigmoid Linear Unit (SiLU)[22] (AKA SiL[28] and Swish-1[29])		$f(x) = \frac{x}{1 + e^{-x}}$	$f'(x) = \frac{1 + e^{-x} + xe^{-x}}{(1 + e^{-x})^2}$	$[\approx -0.278$
SoftExponential[30]		$f(\alpha, x) = \begin{cases} -\frac{\ln(1 - \alpha(x + \alpha))}{\alpha} & \text{for } \alpha < 0 \\ x & \text{for } \alpha = 0 \\ \frac{e^{\alpha x} - 1}{\alpha} + \alpha & \text{for } \alpha > 0 \end{cases}$	$f'(\alpha, x) = \begin{cases} \frac{1}{1 - \alpha(\alpha + x)} & \text{for } \alpha < 0 \\ e^{\alpha x} & \text{for } \alpha \ge 0 \end{cases}$	$(-\infty, \infty)$
Soft Clipping [31][32]		$f(\alpha, x) = \frac{1}{\alpha} \log \frac{1 + e^{\alpha x}}{1 + e^{\alpha(x-1)}}$	$f'(\alpha, x) = \frac{1}{2} \sinh\left(\frac{p}{2}\right) \operatorname{sech}\left(\frac{px}{2}\right) \operatorname{sech}\left(\frac{p}{2}(1 - x)\right)$	$(0, 1)$
Squashing[32][33][34]		$S_{a,\lambda}^{(\beta)}(x) = \frac{1}{\lambda\beta} \ln \frac{1 + e^{\beta(x - (a - \lambda/2))}}{1 + e^{\beta(x - (a + \lambda/2))}} = \frac{1}{\lambda\beta} \ln \frac{\sigma_{a+\lambda/2}^{(-\beta)}(x)}{\sigma_{a-\lambda/2}^{(-\beta)}(x)}.$	$\frac{\partial S_{a,\lambda}^{(\beta)}(x)}{\partial x} = \frac{1}{\lambda} \left(\sigma_{a-\lambda/2}^{(\beta)}(x) - \sigma_{a+\lambda/2}^{(\beta)}(x)\right)$	$(0, 1)$
Sinusoid[35]		$f(x) = \sin(x)$	$f'(x) = \cos(x)$	$[-1, 1]$
Sinc		$f(x) = \begin{cases} 1 & \text{for } x = 0 \\ \frac{\sin(x)}{x} & \text{for } x \ne 0 \end{cases}$	$f'(x) = \begin{cases} 0 & \text{for } x = 0 \\ \frac{\cos(x)}{x} - \frac{\sin(x)}{x^2} & \text{for } x \ne 0 \end{cases}$	$[\approx -.217$
Gaussian		$f(x) = e^{-x^2}$	$f'(x) = -2xe^{-x^2}$	$(0, 1]$
SQ-RBF		$f(x) = \begin{cases} 1 - \frac{x^2}{2} & : \lvert x \rvert \le 1 \\ \frac{(2 - \lvert x \rvert)^2}{2} & : 1 \le \lvert x \rvert \le 2 \\ 0 & : \lvert x \rvert \ge 2 \end{cases}$	$f'(x) = \begin{cases} -x & : \lvert x \rvert \le 1 \\ x - 2\operatorname{sgn}(x) & : 1 \le \lvert x \rvert \le 2 \\ 0 & : \lvert x \rvert \ge 2 \end{cases}$	$[0, 1]$

^ Here, H is the Heaviside step function.
^ α is a stochastic variable sampled from a uniform distribution at training time and fixed to the expectation value of the distribution at test time.
^ ^ ^ Here, σ is the logistic function.
^ $\alpha > 0$ for the range to hold true.

The following table lists activation functions that are not functions of a single fold x from the previous layer or layers:

Name	Equation	Derivatives	Range	Order of continuity
Softmax	$f_i(\vec{x}) = \frac{e^{x_i}}{\sum_{j=1}^{J} e^{x_j}}$ for $i = 1, \ldots, J$	$\frac{\partial f_i(\vec{x})}{\partial x_j} = f_i(\vec{x})(\delta_{ij} - f_j(\vec{x}))$[5]	$(0, 1)$	C^∞
Maxout[36]	$f(\vec{x}) = \max_i x_i$	$\frac{\partial f}{\partial x_j} = \begin{cases} 1 & \text{for } j = \underset{i}{\operatorname{argmax}}\, x_i \\ 0 & \text{for } j \ne \underset{i}{\operatorname{argmax}}\, x_i \end{cases}$	$(-\infty, \infty)$	C^0

^ Here, δ_{ij} is the Kronecker delta.

See also

- Logistic function
- Rectifier (neural networks)
- Stability (learning theory)

- Softmax function
- Swish function

References

1. Hinkelmann, Knut. "Neural Networks, p. 7" (http://didattica.cs.unicam.it/lib/exe/fetch.php?media=didattica:magistrale:kebi:ay_1718:ke-11_neural_networks.pdf) (PDF). *University of Applied Sciences Northwestern Switzerland*.
2. Hodgkin, A. L.; Huxley, A. F. (1952-08-28). "A quantitative description of membrane current and its application to conduction and excitation in nerve" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1392413). *The Journal of Physiology*. **117** (4): 500–544. doi:10.1113/jphysiol.1952.sp004764 (https://doi.org/10.1113%2Fjphysiol.1952.sp004764). PMC 1392413 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1392413). PMID 12991237 (https://pubmed.ncbi.nlm.nih.gov/12991237).
3. Behnke, Sven (2003). *Hierarchical Neural Networks for Image Interpretation* (https://www.researchgate.net/publication/220688219). Lecture Notes in Computer Science. **2766**. Springer. doi:10.1007/b11963 (https://doi.org/10.1007%2Fb11963). ISBN 978-3-540-40722-5.
4. Wuraola, Adedamola; Patel, Nitish (2018), "Computationally Efficient Radial Basis Function", *2018 International Conference on Neural Information Processing (ICONIP)*, Siem reap Cambodia: Springer, pp. 103–112, doi:10.1007/978-3-030-04179-3_9 (https://doi.org/10.1007%2F978-3-030-04179-3_9)
5. Haykin, Simon S. (1999). *Neural Networks: A Comprehensive Foundation* (https://books.google.com/books?id=bX4pAQAAMAAJ). Prentice Hall. ISBN 978-0-13-273350-2.
6. Cybenko, G. (December 1989). "Approximation by superpositions of a sigmoidal function". *Mathematics of Control, Signals, and Systems*. **2** (4): 303–314. doi:10.1007/BF02551274 (https://doi.org/10.1007%2FBF02551274). ISSN 0932-4194 (https://www.worldcat.org/issn/0932-4194).
7. Snyman, Jan (3 March 2005). *Practical Mathematical Optimization: An Introduction to Basic Optimization Theory and Classical and New Gradient-Based Algorithms* (https://books.google.com/books?id=0tFmf_UKl7oC). Springer Science & Business Media. ISBN 978-0-387-24348-1.
8. Wu, Huaiqin (2009). "Global stability analysis of a general class of discontinuous neural networks with linear growth activation functions". *Information Sciences*. **179** (19): 3432–3441. doi:10.1016/j.ins.2009.06.006 (https://doi.org/10.1016%2Fj.ins.2009.06.006).
9. Sussillo, David; Abbott, L. F. (2014-12-19). "Random Walk Initialization for Training Very Deep Feedforward Networks". arXiv:1412.6558 (https://arxiv.org/abs/1412.6558) [cs.NE (https://arxiv.org/archive/cs.NE)].
10. Wuraola, Adedamola; Patel, Nitish (2018), "SQNL:A New Computationally Efficient Activation Function" (https://ieeexplore.ieee.org/document/8489043), *2018 International Joint Conference on Neural Networks (IJCNN)*, Rio Rio de Janeiro, Brazil: IEEE, pp. 1–7
11. Elliot, David L. (1993), "A better activation function for artificial neural networks", *ISR Technical Report TR 93-8, University of Maryland, College Park, MD 20742.*, CiteSeerX 10.1.1.46.7204 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.46.7204)
12. "elliotsig, Elliot symmetric sigmoid transfer function" (https://www.mathworks.com/help/deeplearning/ref/elliotsig.html), *command introduced in Matlab R2012b, Matlab Documentation, MathWorks.*
13. Bergstra, James; Desjardins, Guillaume; Lamblin, Pascal; Bengio, Yoshua (2009). "Quadratic polynomials learn better image features". Technical Report 1337" (https://web.archive.org/web/20180925121919/http://www.iro.umontreal.ca/~lisa/publications2/index.php/attachments/single/205). *Département d'Informatique et de Recherche Opérationnelle, Université de Montréal*. Archived from the original (http://www.iro.umontreal.ca/~lisa/publications2/index.php/attachments/single/205) on 2018-09-25.
14. Glorot, Xavier; Bengio, Yoshua (2010), "Understanding the difficulty of training deep feedforward neural networks" (http://jmlr.org/proceedings/papers/v9/glorot10a/glorot10a.pdf) (PDF), *International Conference on Artificial Intelligence and Statistics (AISTATS'10)*, Society for Artificial Intelligence and Statistics
15. Carlile, Brad; Delamarter, Guy; Kinney, Paul; Marti, Akiko; Whitney, Brian (2017-11-09). "Improving Deep Learning by Inverse Square Root Linear Units (ISRLUs)". arXiv:1710.09967 (https://arxiv.org/abs/1710.09967) [cs.LG (https://arxiv.org/archive/cs.LG)].
16. Nicolae, Andrei (2018-09-03). "PLU: The Piecewise Linear Unit Activation Function". arXiv:1809.09534 (https://arxiv.org/abs/1809.09534) [cs.NE (https://arxiv.org/archive/cs.NE)].
17. Nair, Vinod; Hinton, Geoffrey E. (2010), "Rectified Linear Units Improve Restricted Boltzmann Machines" (http://dl.acm.org/citation.cfm?id=3104322.3104425), *27th International Conference on International Conference on Machine Learning*, ICML'10, USA: Omnipress, pp. 807–814, ISBN 9781605589077
18. Eidnes, Lars; Nøkland, Arild (2018). "Shifting Mean Activation Towards Zero with Bipolar Activation Functions". *International Conference on Learning Representations (ICLR) Workshop*. arXiv:1709.04054 (https://arxiv.org/abs/1709.04054). Bibcode:2017arXiv170904054E (https://ui.adsabs.harvard.edu/abs/2017arXiv170904054E).
19. Maas, Andrew L.; Hannun, Awni Y.; Ng, Andrew Y. (June 2013). "Rectifier nonlinearities improve neural network acoustic models". *Proc. ICML*. **30** (1). S2CID 16489696 (https://api.semanticscholar.org/CorpusID:16489696).
20. He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian (2015-02-06). "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". arXiv:1502.01852 (https://arxiv.org/abs/1502.01852) [cs.CV (https://arxiv.org/archive/cs.CV)].
21. Xu, Bing; Wang, Naiyan; Chen, Tianqi; Li, Mu (2015-05-04). "Empirical Evaluation of Rectified Activations in Convolutional Network". arXiv:1505.00853 (https://arxiv.org/abs/1505.00853) [cs.LG (https://arxiv.org/archive/cs.LG)].
22. Hendrycks, Dan; Gimpel, Kevin (2016). "Gaussian Error Linear Units (GELUs)". arXiv:1606.08415 (https://arxiv.org/abs/1606.08415) [cs.LG (https://arxiv.org/archive/cs.LG)].
23. Clevert, Djork-Arné; Unterthiner, Thomas; Hochreiter, Sepp (2015-11-23). "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)". arXiv:1511.07289 (https://arxiv.org/abs/1511.07289) [cs.LG (https://arxiv.org/archive/cs.LG)].
24. Klambauer, Günter; Unterthiner, Thomas; Mayr, Andreas; Hochreiter, Sepp (2017-06-08). "Self-Normalizing Neural Networks". *Advances in Neural Information Processing Systems*. **30** (2017). arXiv:1706.02515 (https://arxiv.org/abs/1706.02515). Bibcode:2017arXiv170602515K (https://ui.adsabs.harvard.edu/abs/2017arXiv170602515K).
25. Jin, Xiaojie; Xu, Chunyan; Feng, Jiashi; Wei, Yunchao; Xiong, Junjun; Yan, Shuicheng (2015-12-22). "Deep Learning with S-shaped Rectified Linear Activation Units". arXiv:1512.07030 (https://arxiv.org/abs/1512.07030) [cs.CV (https://arxiv.org/archive/cs.CV)].
26. Forest Agostinelli; Matthew Hoffman; Peter Sadowski; Pierre Baldi (21 Dec 2014). "Learning Activation Functions to Improve Deep Neural Networks". arXiv:1412.6830 (https://arxiv.org/abs/1412.6830) [cs.NE (https://arxiv.org/archive/cs.NE)].
27. Glorot, Xavier; Bordes, Antoine; Bengio, Yoshua (2011). "Deep sparse rectifier neural networks" (http://proceedings.mlr.press/v15/glorot11a/glorot11a.pdf) (PDF). *International Conference on Artificial Intelligence and Statistics*.
28. Elfwing, Stefan; Uchibe, Eiji; Doya, Kenji (2018). "Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning". *Neural Networks*. **107**: 3–11. arXiv:1702.03118 (https://arxiv.org/abs/1702.03118). doi:10.1016/j.neunet.2017.12.012 (https://doi.org/10.1016%2Fj.neunet.2017.12.012). PMID 29395652 (https://pubmed.ncbi.nlm.nih.gov/29395652).
29. Ramachandran, Prajit; Zoph, Barret; Le, Quoc V (2017). "Searching for Activation Functions". arXiv:1710.05941 (https://arxiv.org/abs/1710.05941) [cs.NE (https://arxiv.org/archive/cs.NE)].
30. Godfrey, Luke B.; Gashler, Michael S. (2016-02-03). "A continuum among logarithmic, linear, and exponential functions, and its potential to improve generalization in neural networks". *7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management: KDIR*. **1602**: 481–486. arXiv:1602.01321 (https://arxiv.org/abs/1602.01321). Bibcode:2016arXiv160201321G (https://ui.adsabs.harvard.edu/abs/2016arXiv160201321G).
31. Klimek, Matthew D.; Perelstein, Maxim (2018-10-26). "Neural Network-Based Approach to Phase Space Integration". arXiv:1810.11509 (https://arxiv.org/abs/1810.11509) [hep-ph (https://arxiv.org/archive/hep-ph)].

32. Csiszár, Orsolya; Csiszár, Gábor; Dombi, József (2020-04-28). "Interpretable neural networks based on continuous-valued logic and multicriteria decision operators" (http://www.sciencedirect.com/science/article/pii/S0950705120302896). *Knowledge-Based Systems*. **199**: 105972. arXiv:1910.02486 (https://arxiv.org/abs/1910.02486). doi:10.1016/j.knosys.2020.105972 (https://doi.org/10.1016%2Fj.knosys.2020.105972). ISSN 0950-7051 (https://www.worldcat.org/issn/0950-7051).
33. Urenda, Julio; Csiszár, Orsoly; Csiszár, Gábor; Dombi, József; Kosheleva, Olga; Kreinovich, Vladik; Eigner, György (2020-02-01). "Why Squashing Functions in Multi-Layer Neural Networks" (https://scholarworks.utep.edu/cs_techrep/1398). *Departmental Technical Reports (CS)*.
34. Dombi, József; Gera, Zsolt (September 2005). "The approximation of piecewise linear membership functions and Łukasiewicz operators". *Fuzzy Sets and Systems*. **154** (2): 275–286. doi:10.1016/j.fss.2005.02.016 (https://doi.org/10.1016%2Fj.fss.2005.02.016).
35. Gashler, Michael S.; Ashmore, Stephen C. (2014-05-09). "Training Deep Fourier Neural Networks To Fit Time-Series Data". arXiv:1405.2262 (https://arxiv.org/abs/1405.2262) [cs.NE (https://arxiv.org/archive/cs.NE)].
36. Goodfellow, Ian J.; Warde-Farley, David; Mirza, Mehdi; Courville, Aaron; Bengio, Yoshua (2013). "Maxout Networks". *JMLR Workshop and Conference Proceedings*. **28** (3): 1319–1327. arXiv:1302.4389 (https://arxiv.org/abs/1302.4389). Bibcode:2013arXiv1302.4389G (https://ui.adsabs.harvard.edu/abs/2013arXiv1302.4389G).