

Natural Language Processing Neural Networks

Felipe Bravo-Marquez

August 21, 2019

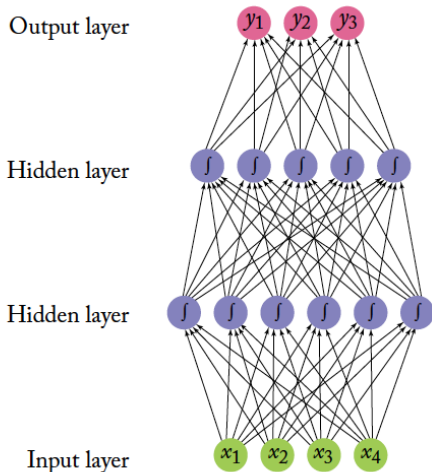
Introduction to Neural Networks

- Very popular machine learning models formed by units called **neurons**.
- A neuron is a computational unit that has scalar inputs and outputs.
- Each input has an associated weight w .
- The neuron multiplies each input by its weight, and then sums them (other functions such as **max** are also possible).
- It applies an activation function g (usually non-linear) to the result, and passes it to its output.
- Multiple layers can be stacked.

Activation Functions

- The nonlinear activation function g has a crucial role in the network's ability to represent complex functions.
- Without the nonlinearity in g , the neural network can only represent linear transformations of the input.

Feedforward Network with two Layers



Feedforward Network Neural Networks

- The feedforward network from the picture is a stack of linear models separated by nonlinear functions.
- The values of each row of neurons in the network can be thought of as a vector.
- The input layer is a 4-dimensional vector (\vec{x}), and the layer above it is a 6-dimensional vector (\vec{h}^1).
- The fully connected layer can be thought of as a linear transformation from 4 dimensions to 6 dimensions.
- A fully connected layer implements a vector-matrix multiplication, $\vec{h} = \vec{x}W$.
- The weight of the connection from the i -th neuron in the input row to the j -th neuron in the output row is $W_{[i,j]}$.
- The values of \vec{h} are transformed by a nonlinear function g that is applied to each value before being passed on as input to the next layer.

⁰Vectors are assumed to be row vectors and superscript indices correspond to network layers.

Neural Networks as Mathematical Functions

- The Multilayer Perceptron (MLP) from the figure is called MLP2 because it has two hidden layers.
- A simpler model would be MLP1, a multilayer perceptron of one hidden layer:

$$NN_{MLP1}(\vec{x}) = g(\vec{x}W^1 + \vec{b}^1)W^2 + \vec{b}^2 \quad (1)$$
$$\vec{x} \in \mathcal{R}^{in}, W^1 \in \mathcal{R}^{d_{in} \times d_1}, \vec{b}^1 \in \mathcal{R}^1, W^2 \in \mathcal{R}^{d_1 \times d_{out}}, \vec{b}^2 \in \mathcal{R}^{d_{out}}$$

- Here W^1 and \vec{b}^1 are a matrix and a bias term for the first linear transformation of the input.
- The function g is a nonlinear function that is applied element-wise (also called a nonlinearity or an activation function).
- W^2 and \vec{b}^2 are the matrix and bias term for a second linear transform.
- When describing a neural network, one should specify the dimensions of the layers (d_1) and the input (d_{in}).

Neural Networks as Mathematical Functions

- MLP2 can be written as the following mathematical function:

$$\begin{aligned} NN_{MLP2}(\vec{x}) &= \vec{y} \\ \vec{h}^1 &= g^1(\vec{x}W^1 + \vec{b}^1) \\ \vec{h}^2 &= g^2(\vec{h}^1W^2 + \vec{b}^2) \\ \vec{y} &= \vec{h}^2W^3 \\ \vec{y} &= (g^2(g^1(\vec{x}W^1 + \vec{b}^1)W^2 + \vec{b}^2))W^3. \end{aligned} \tag{2}$$

- The matrices and the bias terms that define the linear transformations are the parameters of the network.
- Like in linear models, it is common to refer to the collection of all parameters as Θ .

Representation Power

- [Hornik et al., 1989] and [Cybenko, 1989] showed that a multilayer perceptron of one hidden layer (MLP1) is a universal approximator.
- MLP1 can approximate all continuous functions on a closed and bounded subset of \mathcal{R}^n .
- This may suggest there is no reason to go beyond MLP1 to more complex architectures.
- The result does not say how easy or hard it is to set the parameters based on training data and a specific learning algorithm.
- It also does not guarantee that a training algorithm will find the correct function generating our training data.
- Finally, it does not state how large the hidden layer should be.

Representation Power

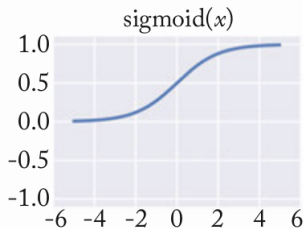
- In practice, we train neural networks on relatively small amounts of data using local search methods.
- We also use hidden layers of relatively modest sizes (up to several thousands).
- The universal approximation theorem does not give any guarantees under these conditions.
- However, there is definitely benefit in trying out more complex architectures than MLP1.
- In many cases, however, MLP1 does indeed provide strong results.

Activation Functions

- The nonlinearity g can take many forms.
- There is currently no good theory as to which nonlinearity to apply in which conditions.
- Choosing the correct nonlinearity for a given task is for the most part an empirical question.

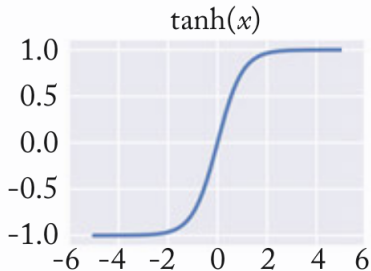
Sigmoid

- The sigmoid activation function $\sigma(x) = \frac{1}{1+e^{-x}}$ is an S-shaped function, transforming each value x into the range $[0, 1]$.
- The sigmoid was the canonical nonlinearity for neural networks since their inception.
- Is currently considered to be deprecated for use in internal layers of neural networks, as the choices listed next prove to work much better empirically.



Hyperbolic tangent (tanh)

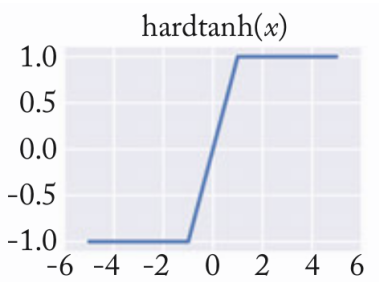
- The hyperbolic tangent $\tanh(x) = \frac{e^{2x}-1}{e^{2x}+1}$ activation function is an S-shaped function, transforming the values x into the range $[-1, 1]$.



Hard tanh

- The hard-tanh activation function is an approximation of the tanh function which is faster to compute and to find derivatives thereof:

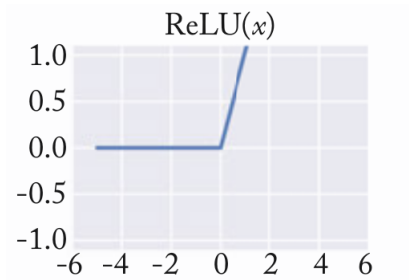
$$\text{hardtanh}(x) = \begin{cases} -1 & x < -1 \\ 1 & x > 1 \\ x & \text{otherwise.} \end{cases}$$



ReLU

- The rectifier activation function [Glorot et al., 2011], also known as the rectified linear unit is a very simple activation function.
- It is easy to work with and was shown many times to produce excellent results.
- The ReLU unit clips each value $x < 0$ at 0.

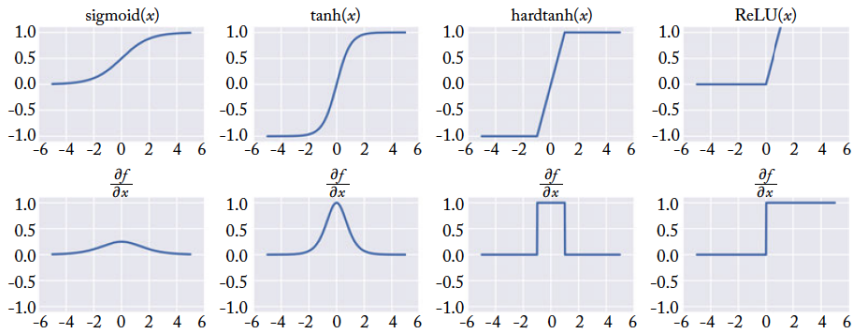
$$\text{ReLU}(x) = \max(0, x)$$



- It performs well for many tasks, especially when combined with the dropout regularization technique (to be explained later).

Activation Functions

- As a rule of thumb, both ReLU and tanh units work well, and significantly outperform the sigmoid.
- You may want to experiment with both tanh and ReLU activations, as each one may perform better in different settings.
- The figure from below shows the shapes of the different activations functions, together with the shapes of their derivatives.



Backpropagation

- Backpropagation is an efficient technique for evaluating the gradient of a loss function L for a feed-forward neural network with respect to all its parameters [Bishop, 2006].¹
- Those parameters are: $W^1, \vec{b}^1, \dots, W^m, \vec{b}^m$, for a network of m layers.
- Recall that superscripts are used to denote layer indexes (not exponentiations).
- For simplicity, we will assume that L is calculated over a single example.
- In a general feed-forward network, each unit computes a weighted sum of its inputs in the form:

$$\vec{h}_{[j]}^l = \left(\sum_i w_{[i,j]}^l \times \vec{z}_{[i]}^{(l-1)} \right) + \vec{b}_{[j]}^l \quad (3)$$

- The variable $\vec{z}_{[i]}^{(l-1)}$ is an input that sends a connection to unit $\vec{h}_{[j]}^l$, $w_{[i,j]}^l$ is the weight associated with that connection, and l is the layer index.
- The biases vectors $\vec{b}_{[j]}^l$ can be excluded from (eq.3) and included to the weight matrix $w_{[i,j]}^l$ by introducing an extra unit, or input, with activation fixed at +1.

¹The following slides on backpropagation are based on [Bishop, 2006], we adapted the notation to be consistent with [Goldberg, 2017].

Backpropagation

- The inputs at layer l , $\vec{z}_{[l]}^{(l-1)}$ are the result of applying the activation function g to units from the previous layer:

$$\vec{z}_{[l]}^l = g(\vec{h}_{[l]}^l) \quad (4)$$

- For the input layer ($l = 0$), \vec{z} corresponds to the input vector $\vec{z} = \vec{x}$

$$\vec{z}_{[l]}^0 = \vec{x}_{[l]} \quad (5)$$

- For each instance in the training set, we supply the corresponding input vector \vec{x} to the network.
- Next we calculate the activations of all of the hidden and output units in the network by successive application of (eq.3) and (eq.4).
- This process is often called forward propagation because it can be regarded as a forward flow of information through the network.

Backpropagation

- Now consider the evaluation of the derivative of L with respect to a weight $W_{[i,j]}^l$.
- Assuming that the loss L is calculated over a single example, we can note that L depends on the weight $W_{[i,j]}^l$ only via the summed input $\vec{h}_{[j]}^l$.
- We can therefore apply the chain rule for partial derivatives to give

$$\frac{\partial L}{\partial W_{[i,j]}^l} = \frac{\partial L}{\partial \vec{h}_{[j]}^l} \times \frac{\partial \vec{h}_{[j]}^l}{\partial W_{[i,j]}^l} \quad (6)$$

Backpropagation

- We now introduce a useful notation:

$$\vec{\delta}_{[l]}^l \equiv \frac{\partial L}{\partial \vec{h}_{[l]}^l} \quad (7)$$

- Using (3), we can write

$$\frac{\partial \vec{h}_{[l]}^l}{\partial \mathbf{W}_{[i,j]}^l} = \vec{z}_{[i]}^{(l-1)} \quad (8)$$

- Substituting (7) and (8) into (6), we then obtain

$$\frac{\partial L}{\partial \mathbf{W}_{[i,j]}^l} = \vec{\delta}_{[l]}^l \times \vec{z}_{[i]}^{(l-1)} \quad (9)$$

Backpropagation

- Equation (9) tells us that the required derivative is obtained simply by multiplying the value of $\vec{\delta}_{[j]}^l$ by the value of $\vec{z}_{[j]}^{(l-1)}$.
- Thus, in order to evaluate the derivatives, we need only to calculate the value of $\vec{\delta}_{[j]}^l$ for each hidden and output unit in the network, and then apply (9).
- Calculating $\vec{\delta}_{[j]}^m$ for output units ($l = m$), is usually straightforward, since activation units $\vec{h}_{[j]}^m$ are directly observed in the loss expression.
- The same applies for shallow linear models.

Backpropagation

- To evaluate the $\vec{\delta}_{[l]}^l$ for hidden units, we again make use of the chain rule for partial derivatives:

$$\vec{\delta}_{[l]}^l \equiv \frac{\partial L}{\partial \vec{h}_{[l]}^l} = \sum_k \left(\frac{\partial L}{\partial \vec{h}_{[k]}^{l+1}} \times \frac{\partial \vec{h}_{[k]}^{l+1}}{\partial \vec{h}_{[l]}^l} \right) \quad (10)$$

- The sum runs over all units $\vec{h}_{[k]}^{l+1}$ to which unit $\vec{h}_{[l]}^l$ sends connections.
- We assume that connections go only to consecutive layers in the network (from layer l to layer $(l + 1)$).
- The units $\vec{h}_{[k]}^{l+1}$ could include other hidden units and/or output units.
- If we now substitute the definition of $\vec{\delta}_{[l]}^l$ given by (eq.7) into (eq.10), we get

$$\vec{\delta}_{[l]}^l \equiv \frac{\partial L}{\partial \vec{h}_{[l]}^l} = \sum_k \left(\vec{\delta}_{[k]}^{l+1} \times \frac{\partial \vec{h}_{[k]}^{l+1}}{\partial \vec{h}_{[l]}^l} \right) \quad (11)$$

Backpropagation

- Now, for expression $\vec{h}_{[k]}^{l+1}$ we can go to its definition (eq.3):

$$\vec{h}_{[k]}^{(l+1)} = \left(\sum_i w_{[i,k]}^{l+1} \times \vec{z}_{[i]}^l \right) + \vec{b}_{[k]}^{(l+1)}$$

- Now, we replace (eq.4) ($\vec{z}_{[i]}^l = g(\vec{h}_{[i]}^l)$) into previous equation and we obtain:

$$\vec{h}_{[k]}^{(l+1)} = \left(\sum_i w_{[i,k]}^{l+1} \times g(\vec{h}_{[i]}^l) \right) + \vec{b}_{[k]}^{(l+1)}$$

- Now when calculating $\frac{\partial \vec{h}_{[k]}^{l+1}}{\partial \vec{h}_{[j]}^l}$ all the terms in the summation where $i \neq j$ get canceled out.
- Hence:

$$\frac{\partial \vec{h}_{[k]}^{l+1}}{\partial \vec{h}_{[j]}^l} = w_{[j,k]}^{l+1} \times g'(\vec{h}_{[j]}^l) \quad (12)$$

Backpropagation

- Now, if we substitute (eq.12) into (eq.11)

$$\vec{\delta}_{[j]}^l \equiv \frac{\partial L}{\partial \vec{h}_{[j]}^l} = \sum_k \left(\vec{\delta}_{[k]}^{l+1} \times w_{[j,k]}^{l+1} \times g'(\vec{h}_{[j]}^l) \right) \quad (13)$$

- Since $g'(\vec{h}_{[j]}^l)$ doesn't depend on k we can obtain the following backpropagation formula:

$$\vec{\delta}_{[j]}^l = g'(\vec{h}_{[j]}^l) \times \sum_k \left(\vec{\delta}_{[k]}^{l+1} \times w_{[j,k]}^{l+1} \right) \quad (14)$$

- Which tells us that the value of δ for a particular hidden unit can be obtained by propagating the δ 's backwards from units higher up in the network. [Bishop, 2006].

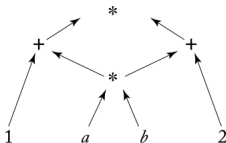
Backpropagation

The backpropagation procedure can be summarized as follows.

1. Apply an input vector \vec{x} to the network and forward propagate through the network using (eq.3) and (eq.4) to find the activations of all the hidden and output units.
2. Evaluate the $\vec{\delta}_{[j]}^m$ for all the output units (recall that the derivatives involved here are easy to calculate).
3. Backpropagate the $\vec{\delta}_{[k]}^{(l+1)}$ using (eq.14) to obtain $\vec{\delta}_{[j]}^l$ for each hidden unit in the network. We go from higher to lower layers in the network.
4. Use (eq.9) ($\frac{\partial L}{\partial w_{[i,j]}^l} = \vec{\delta}_{[j]}^l \times \vec{z}_{[i]}^{(l-1)}$) to evaluate the required derivatives.

The Computation Graph Abstraction

- One can compute the gradients of the various parameters of a network by hand and implement them in code.
- This procedure is cumbersome and error prone.
- For most purposes, it is preferable to use automatic tools for gradient computation [Bengio, 2012].
- A computation graph is a representation of an arbitrary mathematical computation (e.g., a neural network) as a graph.
- Consider for example a graph for the computation of $(a * b + 1) * (a * b + 2)$:



- The computation of $a * b$ is shared.
- The graph structure defines the order of the computation in terms of the dependencies between the different components.

The Computation Graph Abstraction

- The computation graph abstraction allows us to:
 1. Easily construct arbitrary networks.
 2. Evaluate their predictions for given inputs (forward pass)

Algorithm 5.3 Computation graph forward pass.

```
1: for i = 1 to N do
2:   Let  $a_1, \dots, a_m = \pi^{-1}(i)$ 
3:    $v(i) \leftarrow f_i(v(a_1), \dots, v(a_m))$ 
```

3. Compute gradients for their parameters with respect to arbitrary scalar losses (backward pass or backpropagation).

Algorithm 5.4 Computation graph backward pass (backpropagation).

```
1:  $d(N) \leftarrow 1$   $\triangleright \frac{\partial N}{\partial N} = 1$ 
2: for i = N-1 to 1 do
3:    $d(i) \leftarrow \sum_{j \in \pi(i)} d(j) \cdot \frac{\partial f_j}{\partial i}$   $\triangleright \frac{\partial N}{\partial i} = \sum_{j \in \pi(i)} \frac{\partial N}{\partial j} \frac{\partial j}{\partial i}$ 
```

- The backpropagation algorithm (backward pass) is essentially following the chain-rule of differentiation².

²A comprehensive tutorial on the backpropagation algorithm over the computational graph abstraction:

<https://colah.github.io/posts/2015-08-Backprop/>

Deep Learning Frameworks

Several software packages implement the computation-graph model. All these packages support all the essential components (node types) for defining a wide range of neural network architectures.

- TensorFlow (<https://www.tensorflow.org/>): an open source software library for numerical computation using data-flow graphs originally developed by the Google Brain Team.
- Keras: High-level neural network API that runs on top of Tensorflow as well as other backends (<https://keras.io/>).
- PyTorch: open source machine learning library for Python, based on Torch, developed by Facebook's artificial-intelligence research group. It supports dynamic graph construction, a different computation graph is created from scratch for each training sample. (<https://pytorch.org/>)

Questions?

Thanks for your Attention!

References I



Bishop, C. M. (2006).
Pattern recognition and machine learning.
springer.



Cybenko, G. (1989).
Approximation by superpositions of a sigmoidal function.
Mathematics of control, signals and systems, 2(4):303–314.



Glorot, X., Bordes, A., and Bengio, Y. (2011).
Deep sparse rectifier neural networks.
In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323.



Goldberg, Y. (2017).
Neural network methods for natural language processing.
Synthesis Lectures on Human Language Technologies, 10(1):1–309.



Goodfellow, I., Bengio, Y., and Courville, A. (2016).
Deep learning.
MIT press.



Hornik, K., Stinchcombe, M., and White, H. (1989).
Multilayer feedforward networks are universal approximators.
Neural networks, 2(5):359–366.