Published in Secure and Private AI Math Blogging Competition

Fady Morris Milad Ebeid Follow

Aug 24, 2019 · 2 min read · Listen

Understanding Vectorized Implementation of Neural Networks

Studying the inner workings of neural networks can be difficult for beginners without clear visualization of data movement inside the network.

I've created a simple neural network example architecture and detailed matrix operations to aid in understanding the inputs, outputs and inner details of neural networks. The diagrams below and the matrix operations help to study the vectorized implementation of feed-forward propagation algorithm(used for predictions) and backpropagation algorithm(used to train the neural network).

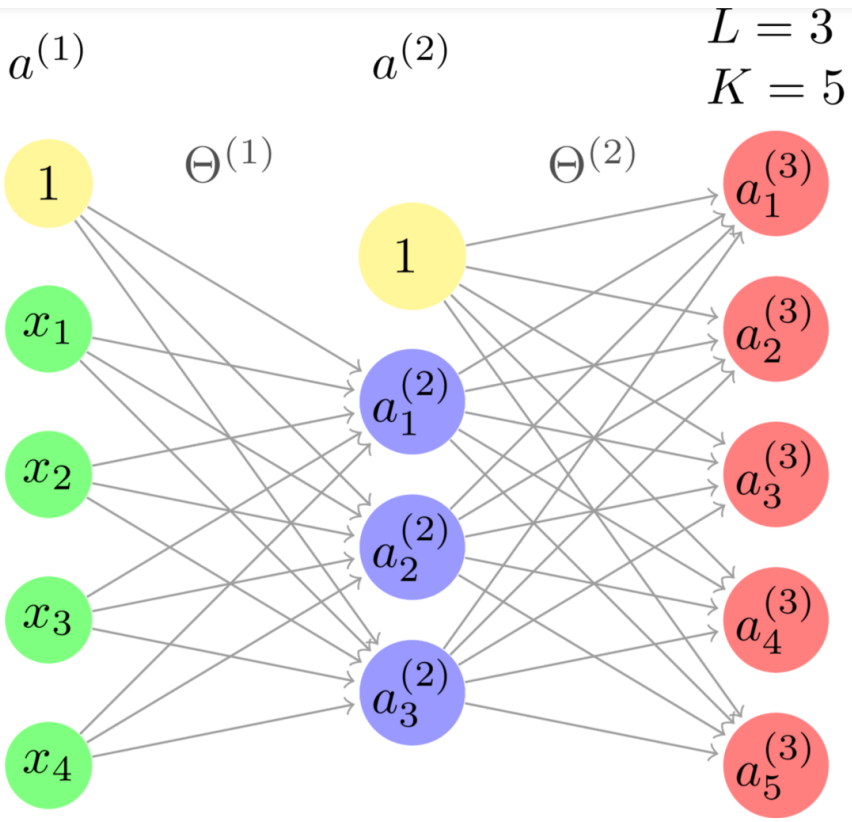
Notation :

x	:	Features vector (input vector).
y	:	Training labels.
$a^{(l)}$	:	Activation units of layer l.
$\Theta^{(l)}$	:	Weight matrix that maps layer l to layer $l+1$
$h_\Theta(x)$	:	Predicted class labels of input vector x
$D^{(l)}$	:	Gradient matrix corresponding to weight matrix $\Theta^{(l)}$
$g(z)$	:	Sigmoid function. $g(z) = \frac{1}{1+e^{-z}}$
$g'(z)$	:	The derivative of sigmoid function. $g'(z) = g(z)[1-g(z)]$

A Simple Network Architecture :

The network consists of 3 layers: an input layer, a hidden layer, and an output layer.

It has 5 output classes and accepts an input vector (x) (feature vector) with 4 features.

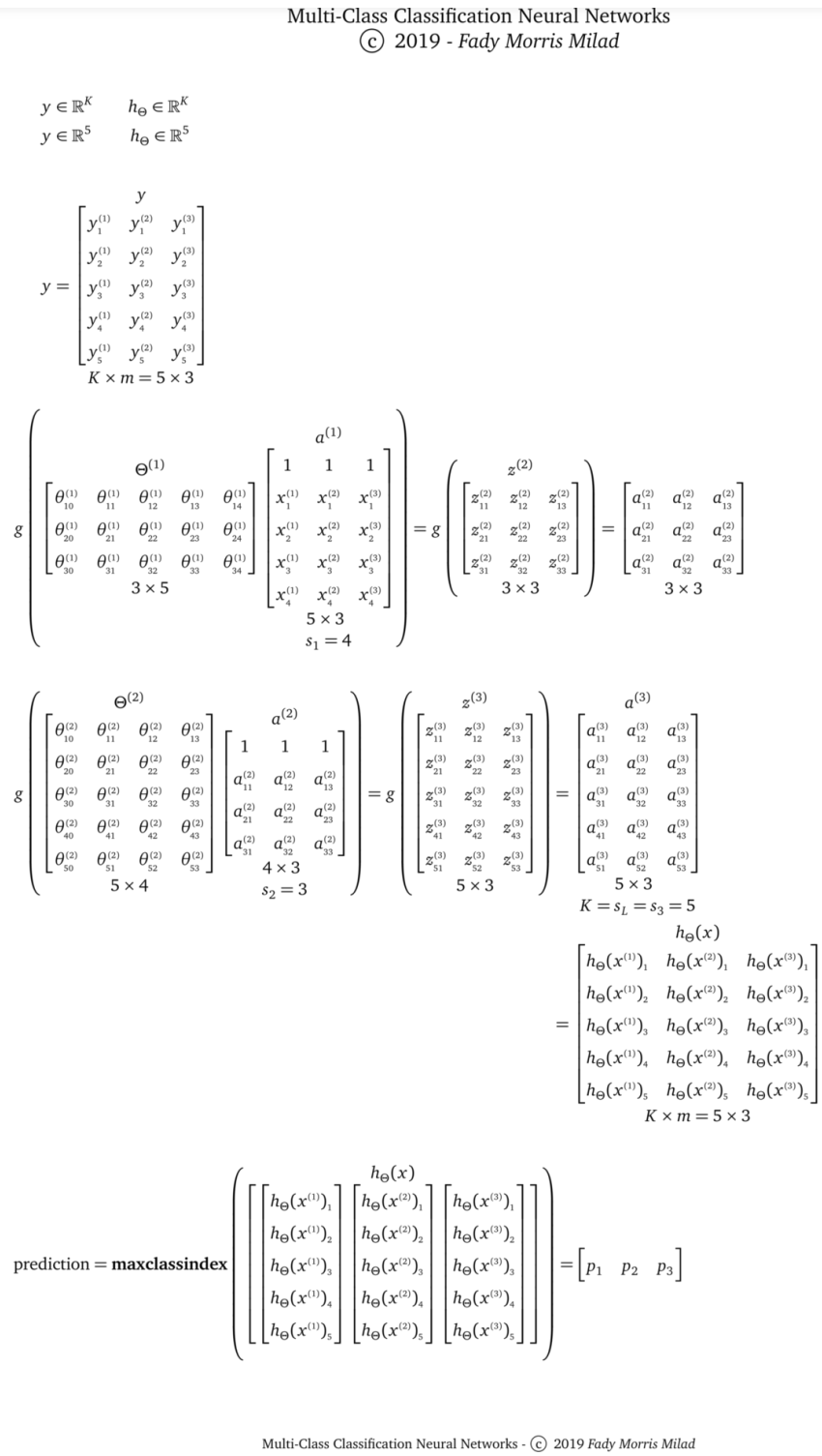


A Simple Neural Network Architecture, The Network consists of three layers : (1) An input layer. (2) A hidden layer. (3) An output layer.

Feed Forward Propagation (Using the network for predictions) :

The feed-forward propagation takes an input feature vector (x) and outputs predicted classes (hΘ)corresponding to each training example, here we have 3 training examples.

> *Note: In our example, the bias unit (b), which is stated in other literature, is substituted by appending 1 as the first element in the feature vector and every activation layer (a) , and adding a corresponding θo in the weight matrix.*



Feed-forward Algorithm

The backpropagation (training the neural network) :

The backpropagation algorithm calculates the error matrix (Δ) and the gradient matrix (D) corresponding to every weight matrix (Θ). The gradient matrix can be used by optimization algorithms to optimize the weights of the network.

Multi-Class Classification Neural Networks

$$\underset{5\times 3}{\begin{bmatrix} \delta_{11}^{(3)} & \delta_{12}^{(3)} & \delta_{13}^{(3)} \\ \delta_{21}^{(3)} & \delta_{22}^{(3)} & \delta_{23}^{(3)} \\ \delta_{31}^{(3)} & \delta_{32}^{(3)} & \delta_{33}^{(3)} \\ \delta_{41}^{(3)} & \delta_{42}^{(3)} & \delta_{43}^{(3)} \\ \delta_{51}^{(3)} & \delta_{52}^{(3)} & \delta_{53}^{(3)} \end{bmatrix}}^{\delta^{(3)}} = \underset{5\times 3}{\begin{bmatrix} h_\Theta(x^{(1)})_1 & h_\Theta(x^{(2)})_1 & h_\Theta(x^{(3)})_1 \\ h_\Theta(x^{(1)})_2 & h_\Theta(x^{(2)})_2 & h_\Theta(x^{(3)})_2 \\ h_\Theta(x^{(1)})_3 & h_\Theta(x^{(2)})_3 & h_\Theta(x^{(3)})_3 \\ h_\Theta(x^{(1)})_4 & h_\Theta(x^{(2)})_4 & h_\Theta(x^{(3)})_4 \\ h_\Theta(x^{(1)})_5 & h_\Theta(x^{(2)})_5 & h_\Theta(x^{(3)})_5 \end{bmatrix}}^{h_\Theta(x)} - \underset{5\times 3}{\begin{bmatrix} y_1^{(1)} & y_1^{(2)} & y_1^{(3)} \\ y_2^{(1)} & y_2^{(2)} & y_2^{(3)} \\ y_3^{(1)} & y_3^{(2)} & y_3^{(3)} \\ y_4^{(1)} & y_4^{(2)} & y_4^{(3)} \\ y_5^{(1)} & y_5^{(2)} & y_5^{(3)} \end{bmatrix}}^{y}$$

$$\underset{\substack{4\times 3 \\ = size\ of\ [a^{(2)}]}}{\begin{bmatrix} \delta_{11}^{(2)} & \delta_{12}^{(2)} & \delta_{13}^{(2)} \\ \delta_{21}^{(2)} & \delta_{22}^{(2)} & \delta_{23}^{(2)} \\ \delta_{31}^{(2)} & \delta_{32}^{(2)} & \delta_{33}^{(2)} \\ \delta_{41}^{(2)} & \delta_{42}^{(2)} & \delta_{43}^{(2)} \end{bmatrix}}^{\delta^{(2)}} = \underset{4\times 5}{\begin{bmatrix} \theta_{10}^{(2)} & \theta_{20}^{(2)} & \theta_{30}^{(2)} & \theta_{40}^{(2)} & \theta_{50}^{(2)} \\ \theta_{11}^{(2)} & \theta_{21}^{(2)} & \theta_{31}^{(2)} & \theta_{41}^{(2)} & \theta_{51}^{(2)} \\ \theta_{12}^{(2)} & \theta_{22}^{(2)} & \theta_{32}^{(2)} & \theta_{42}^{(2)} & \theta_{52}^{(2)} \\ \theta_{13}^{(2)} & \theta_{23}^{(2)} & \theta_{33}^{(2)} & \theta_{43}^{(2)} & \theta_{53}^{(2)} \end{bmatrix}}^{(\Theta^{(2)})^T} \underset{5\times 3}{\begin{bmatrix} \delta_{11}^{(3)} & \delta_{12}^{(3)} & \delta_{13}^{(3)} \\ \delta_{21}^{(3)} & \delta_{22}^{(3)} & \delta_{23}^{(3)} \\ \delta_{31}^{(3)} & \delta_{32}^{(3)} & \delta_{33}^{(3)} \\ \delta_{41}^{(3)} & \delta_{42}^{(3)} & \delta_{43}^{(3)} \\ \delta_{51}^{(3)} & \delta_{52}^{(3)} & \delta_{53}^{(3)} \end{bmatrix}}^{\delta^{(3)}} \odot \overbrace{\underset{4\times 3}{\begin{bmatrix} 1 & 1 & 1 \\ a_{11}^{(2)} & a_{12}^{(2)} & a_{13}^{(2)} \\ a_{21}^{(2)} & a_{22}^{(2)} & a_{23}^{(2)} \\ a_{31}^{(2)} & a_{32}^{(2)} & a_{33}^{(2)} \end{bmatrix}}^{a^{(2)}} \odot \left(1 - \underset{4\times 3}{\begin{bmatrix} 1 & 1 & 1 \\ a_{11}^{(2)} & a_{12}^{(2)} & a_{13}^{(2)} \\ a_{21}^{(2)} & a_{22}^{(2)} & a_{23}^{(2)} \\ a_{31}^{(2)} & a_{32}^{(2)} & a_{33}^{(2)} \end{bmatrix}}^{a^{(2)}}\right)}^{g'(z^{(2)}) = g(z^{(2)})\left[1 - g(z^{(2)})\right]}$$

$$\underset{5\times 4 = size\ of\ [\Theta^{(2)}]}{\begin{bmatrix} \Delta_{10}^{(2)} & \Delta_{11}^{(2)} & \Delta_{12}^{(2)} & \Delta_{13}^{(2)} \\ \Delta_{20}^{(2)} & \Delta_{21}^{(2)} & \Delta_{22}^{(2)} & \Delta_{23}^{(2)} \\ \Delta_{30}^{(2)} & \Delta_{31}^{(2)} & \Delta_{32}^{(2)} & \Delta_{33}^{(2)} \\ \Delta_{40}^{(2)} & \Delta_{41}^{(2)} & \Delta_{42}^{(2)} & \Delta_{43}^{(2)} \\ \Delta_{50}^{(2)} & \Delta_{51}^{(2)} & \Delta_{52}^{(2)} & \Delta_{53}^{(2)} \end{bmatrix}}^{\Delta^{(2)}} = \underset{5\times 3}{\begin{bmatrix} \delta_{11}^{(3)} & \delta_{12}^{(3)} & \delta_{13}^{(3)} \\ \delta_{21}^{(3)} & \delta_{22}^{(3)} & \delta_{23}^{(3)} \\ \delta_{31}^{(3)} & \delta_{32}^{(3)} & \delta_{33}^{(3)} \\ \delta_{41}^{(3)} & \delta_{42}^{(3)} & \delta_{43}^{(3)} \\ \delta_{51}^{(3)} & \delta_{52}^{(3)} & \delta_{53}^{(3)} \end{bmatrix}}^{\delta^{(3)}} \underset{3\times 4}{\begin{bmatrix} 1 & a_{11}^{(2)} & a_{21}^{(2)} & a_{31}^{(2)} \\ 1 & a_{12}^{(2)} & a_{22}^{(2)} & a_{32}^{(2)} \\ 1 & a_{13}^{(2)} & a_{23}^{(2)} & a_{33}^{(2)} \end{bmatrix}}^{(a^{(2)})^T}$$

$$\underset{3\times 5 = size\ of\ [\Theta^{(1)}]}{\begin{bmatrix} \Delta_{10}^{(1)} & \Delta_{11}^{(1)} & \Delta_{12}^{(1)} & \Delta_{13}^{(1)} & \Delta_{14}^{(1)} \\ \Delta_{20}^{(1)} & \Delta_{21}^{(1)} & \Delta_{22}^{(1)} & \Delta_{23}^{(1)} & \Delta_{24}^{(1)} \\ \Delta_{30}^{(1)} & \Delta_{31}^{(1)} & \Delta_{32}^{(1)} & \Delta_{33}^{(1)} & \Delta_{34}^{(1)} \end{bmatrix}}^{\Delta^{(1)}} = \underset{3\times 3}{\begin{bmatrix} \delta_{21}^{(2)} & \delta_{22}^{(2)} & \delta_{23}^{(2)} \\ \delta_{31}^{(2)} & \delta_{32}^{(2)} & \delta_{33}^{(2)} \\ \delta_{41}^{(2)} & \delta_{42}^{(2)} & \delta_{43}^{(2)} \end{bmatrix}}^{\delta^{(2)}[2:4;1:3]} \underset{3\times 5}{\begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & x_4^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & x_4^{(2)} \\ 1 & x_1^{(3)} & x_2^{(3)} & x_3^{(3)} & x_4^{(3)} \end{bmatrix}}^{(a^{(1)})^T}$$

$$\underset{5\times 4 = size\ of\ [\Theta^{(2)}]}{\begin{bmatrix} D_{10}^{(2)} & D_{11}^{(2)} & D_{12}^{(2)} & D_{13}^{(2)} \\ D_{20}^{(2)} & D_{21}^{(2)} & D_{22}^{(2)} & D_{23}^{(2)} \\ D_{30}^{(2)} & D_{31}^{(2)} & D_{32}^{(2)} & D_{33}^{(2)} \\ D_{40}^{(2)} & D_{41}^{(2)} & D_{42}^{(2)} & D_{43}^{(2)} \\ D_{50}^{(2)} & D_{51}^{(2)} & D_{52}^{(2)} & D_{53}^{(2)} \end{bmatrix}}^{\substack{D^{(2)} \\ (\Theta^{(2)}\ \text{gradient})}} = \frac{1}{m} \underset{5\times 4}{\begin{bmatrix} \Delta_{10}^{(2)} & \Delta_{11}^{(2)} & \Delta_{12}^{(2)} & \Delta_{13}^{(2)} \\ \Delta_{20}^{(2)} & \Delta_{21}^{(2)} & \Delta_{22}^{(2)} & \Delta_{23}^{(2)} \\ \Delta_{30}^{(2)} & \Delta_{31}^{(2)} & \Delta_{32}^{(2)} & \Delta_{33}^{(2)} \\ \Delta_{40}^{(2)} & \Delta_{41}^{(2)} & \Delta_{42}^{(2)} & \Delta_{43}^{(2)} \\ \Delta_{50}^{(2)} & \Delta_{51}^{(2)} & \Delta_{52}^{(2)} & \Delta_{53}^{(2)} \end{bmatrix}}^{\Delta^{(2)}} + \lambda \overbrace{\underset{5\times 4}{\begin{bmatrix} 0 & \theta_{11}^{(2)} & \theta_{12}^{(2)} & \theta_{13}^{(2)} \\ 0 & \theta_{21}^{(2)} & \theta_{22}^{(2)} & \theta_{23}^{(2)} \\ 0 & \theta_{31}^{(2)} & \theta_{32}^{(2)} & \theta_{33}^{(2)} \\ 0 & \theta_{41}^{(2)} & \theta_{42}^{(2)} & \theta_{43}^{(2)} \\ 0 & \theta_{51}^{(2)} & \theta_{52}^{(2)} & \theta_{53}^{(2)} \end{bmatrix}}^{\begin{bmatrix} 0 & \Theta^{(2)}[1:5;2:4] \end{bmatrix}}}^{\text{Regularization}}$$

$$\underset{3\times 5 = size\ of\ [\Theta^{(1)}]}{\begin{bmatrix} D_{10}^{(1)} & D_{11}^{(1)} & D_{12}^{(1)} & D_{13}^{(1)} & D_{14}^{(1)} \\ D_{20}^{(1)} & D_{21}^{(1)} & D_{22}^{(1)} & D_{23}^{(1)} & D_{24}^{(1)} \\ D_{30}^{(1)} & D_{31}^{(1)} & D_{32}^{(1)} & D_{33}^{(1)} & D_{34}^{(1)} \end{bmatrix}}^{\substack{D^{(1)} \\ (\Theta^{(1)}\ \text{gradient})}} = \frac{1}{m} \underset{3\times 5}{\begin{bmatrix} \Delta_{10}^{(1)} & \Delta_{11}^{(1)} & \Delta_{12}^{(1)} & \Delta_{13}^{(1)} & \Delta_{14}^{(1)} \\ \Delta_{20}^{(1)} & \Delta_{21}^{(1)} & \Delta_{22}^{(1)} & \Delta_{23}^{(1)} & \Delta_{24}^{(1)} \\ \Delta_{30}^{(1)} & \Delta_{31}^{(1)} & \Delta_{32}^{(1)} & \Delta_{33}^{(1)} & \Delta_{34}^{(1)} \end{bmatrix}}^{\Delta^{(1)}} + \lambda \overbrace{\underset{3\times 5}{\begin{bmatrix} 0 & \theta_{11}^{(1)} & \theta_{12}^{(1)} & \theta_{13}^{(1)} & \theta_{14}^{(1)} \\ 0 & \theta_{21}^{(1)} & \theta_{22}^{(1)} & \theta_{23}^{(1)} & \theta_{24}^{(1)} \\ 0 & \theta_{31}^{(1)} & \theta_{32}^{(1)} & \theta_{33}^{(1)} & \theta_{34}^{(1)} \end{bmatrix}}^{\begin{bmatrix} 0 & \Theta^{(1)}[1:3;2:5] \end{bmatrix}}}^{\text{Regularization}}$$

Backpropagation Algorithm