Neural Networks: From Theory to Implementation in 1 Hour

December 29, 2024

Why This Video?

When I was learning machine learning, I was relying on online material and courses. The most famous one being Andrew Ng's Machine Learning Specialization and Deep Learning Specialization programs. A lot of them have two characteristics:

- Start from very basics: linear regression - curve fitting with a simple line
- Gradually and very slowly build up towards neural networks.

Here you learn

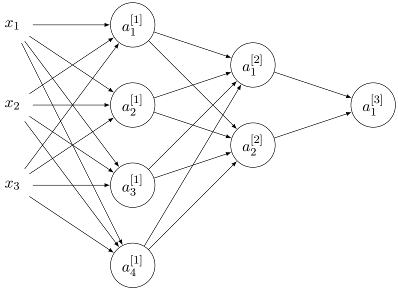
- the theory of simple neural networks and
- implement them in Python.

Save Time with This Video

If you are comfortable with multiplying two matrices and applying the chain rule to get the derivative of a function, then you can deep dive into neural networks right from the beginning. This video is for you.

Neural Networks

In the video, we build and train a neural network to determine whether a picture contains an image of a cat or not.
This is what a neural network looks like:



Input Representation

$$X = \begin{pmatrix} x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(i)} & \dots & x_1^{(m)} \\ x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(i)} & \dots & x_2^{(m)} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_j^{(1)} & x_j^{(2)} & \dots & x_j^{(i)} & \dots & x_j^{(m)} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n_x}^{(1)} & x_{n_x}^{(2)} & \dots & x_{n_x}^{(i)} & \dots & x_{n_x}^{(m)} \end{pmatrix}$$

Forward Propagation Equations

$$z_j^{[l](i)} = \sum_{j'=1}^{n_{l-1}} w_{jj'}^{[l]} a_{j'}^{[l-1](i)} + b_j^{[l]}$$

$$a_j^{[l](i)} = g^{[l]}(z_j^{[l](i)}) \qquad \text{Why?}$$

$$Z^{[l]} = W^{[l]} \cdot A^{[l-1]} + b^{[l]}$$

$$A^{[l]} = g^{[l]}(Z^{[l]})$$

$$(n_l, m) = (n_l, n_{l-1}) \quad \text{Mat. Mul.} \quad (n_{l-1}, m) \quad \text{Mat. Add.} \quad (n_l, m)$$

Logistic Regression - Single Unit Classifier - Last Layer

$$J = \sum_{i=1}^{m} L^{(i)}$$

$$L^{(i)} = -y^{(i)} \log(a^{[L](i)}) - (1 - y^{(i)}) \log(1 - a^{[L](i)})$$

$$a^{[L](i)} = \sigma(z)$$

$$z_j^{[L](i)} = \sum w_{1,j'}^{[L]} a_j^{[L-1](i)} + b_j^{[L](i)}$$

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Gradients - Last Layer

$$\frac{\partial J}{\partial w_j} = \frac{\partial J}{\partial a} \cdot \frac{\partial a}{\partial z} \cdot \frac{\partial z}{\partial w_j}$$

$$\frac{\partial J}{\partial a} = \sum_{i=1}^{m} \left(-\frac{y^{(i)}}{a} + \frac{1 - y^{(i)}}{1 - a} \right)$$

$$\frac{\partial a}{\partial z} = a(1 - a)$$

$$\frac{\partial z}{\partial w_{1,j'}^{[L]}} = a_{j'}^{[L-1](i)}$$

Gradients - Last Layer

$$\frac{\partial J}{\partial w_{1,j'}^{[L]}} = \sum_{i=1}^{m} \left(a_j^{[L](i)} - y^{(i)} \right) a_j^{[L-1](i)}$$

So for the last layer, the vectorization provides:

$$\mathrm{d}Z^{[L]} = A^{[L]} - Y$$

$$\mathrm{d}W^{[L]} = A^{[L-1]} \cdot \left(A^{[L]} - Y\right)^T$$

As for db, we have

$$\frac{\partial J}{\partial b^{[L]}} = \sum_{i=1}^{m} \frac{\partial L^{(i)}}{\partial a^{[L]}} \frac{\partial a^{[L]}}{\partial z^{[L]}} \frac{\partial z^{[L]}}{\partial b^{[L]}}$$

$$\mathrm{d}b^{[L]} = \sum_{\text{axis}=1} \mathrm{d}Z^{L} = \sum_{\text{axis}=1} \left(\mathrm{d}A^{L} - Y\right)$$

Gradients in Other Layers

$$\frac{\partial J}{\partial w_{jj'}^{[l]}} = \sum_{i=1}^{m} \frac{\partial L^{(i)}}{\partial a_j^{[l]}} \frac{\partial a_j^{[l]}}{\partial z_j^{[l]}} \frac{\partial z_j^{[l]}}{\partial w_{jj'}^{[l]}}$$

$$z_j^{[l](i)} = \sum_{j'=1}^{n_{l-1}} w_{jj'}^{[l]} a_{j'}^{[l-1](i)} + b_j^{[l]}$$

$$\frac{\partial z_j^{[l]}}{\partial w_{jj'}^{[l]}} = a_{j'}^{[l-1](i)} \qquad \frac{\partial a_j^{[l]}}{\partial z_j^{[l]}} = g'^{[l]}(z_j^{[l]})$$

$$\frac{\partial L^{(i)}}{\partial a_j^{[l]}} = \sum_{k=1}^{n_{l+1}} \frac{\partial L^{(i)}}{\partial z_k^{[l+1](i)}} \cdot \frac{\partial z_k^{[l+1](i)}}{\partial a_j^{[l](i)}}$$

$$\frac{\partial z_k^{[l+1](i)}}{\partial a_j^{[l](i)}} = w_{kj}^{[l+1]}$$

Gradients in Other Layers

$$\frac{\partial L^{(i)}}{\partial a_j^{[l]}} = \sum_{k=1}^{n_{l+1}} \frac{\partial L^{(i)}}{\partial z_k^{[l+1](i)}} w_{kj}^{[l+1]}$$

The goal is to vectorize. So

$$\mathsf{d}A^{[l]} = {W^{[l+1]}}^T \cdot \mathsf{d}Z^{[l+1]}$$

Remember

$$\frac{\partial J}{\partial w_{jj'}^{[l]}} = \sum_{i=1}^{m} \sum_{k=1}^{n_{l+1}} \frac{\partial L^{(i)}}{\partial z_k^{[l+1](i)}} w_{kj}^{[l+1]} \cdot g'^{[l]}(z_j^{[l]}) \cdot a_{j'}^{[l-1](i)}$$

$$\mathsf{d}W^{[l]}(j, j') = \sum_{i=1}^{m} \mathsf{d}A_j^{[l]} g'^{[l]}(z_j^{[l]}) a_{j'}^{[l-1](i)}$$

Gradients in Other Layers

Assume

$$\mathsf{d}Z^{[l]} = \mathsf{d}A^{[l]} \circ g'^{[l]}(Z^{[l]})$$

So

$$\mathsf{d}W^{[l]}(j, j') = \sum_{i=1}^{m} \mathsf{d}Z_j^{[l]} a_{j'}^{[l-1](i)}$$

The goal is to vectorize. So

$$\mathsf{d}W^{[l]} = \mathsf{d}Z^{[l]} \cdot {A^{[l-1]}}^T$$

Gradients in Other Layers

$$\frac{\partial J}{\partial w_{jj'}^{[l]}} = \sum_{i=1}^{m} \frac{\partial L^{(i)}}{\partial a_j^{[l]}} \frac{\partial a_j^{[l]}}{\partial z_j^{[l]}} \frac{\partial z_j^{[l]}}{\partial w_{jj'}^{[l]}}$$

To summarize, so far we have had

$$\mathsf{d}W^{[l]} = (\mathsf{d}Z^{[l]} \cdot A^{[l-1]T})$$

$$\mathsf{d}A^{[l]} = W^{[l+1]T} \cdot \mathsf{d}Z^{[l+1]}$$

$$\mathsf{d}Z^{[l]} = \mathsf{d}A^{[l]} \circ g'^{[l]}(Z^{[l]})$$

Gradients in Other Layers

$$\frac{\partial L^{(i)}}{\partial b_j^{[l]}} = \frac{\partial L^{(i)}}{\partial z_j^{[l](i)}} \frac{\partial z_j^{[l](i)}}{\partial b_j^{[l]}} = \mathsf{d}z_j^{[l](i)}$$

$$\frac{\partial J}{\partial b_j^{[l]}} = \sum dZ_j^{[l]} = db^{[l]}$$

$$\frac{\partial L^{(i)}}{\partial b_j^{[l]}} = \frac{\partial L^{(i)}}{\partial z_j^{[l](i)}} \frac{\partial z_j^{[l](i)}}{\partial b_j^{[l]}} = \mathsf{d}Z_j^{[l](i)}; \qquad \frac{\partial z_j^{[l](i)}}{\partial b_j^{[l]}} = 1$$

$$J = \sum L^{(i)}$$

$$\frac{\partial J}{\partial b_j^{[l]}} = \sum \mathsf{d}z_j^{[l]} = db_j^{[l]}$$

$$\mathsf{d}b^{[l]} = \sum_{\mathsf{axis}=1} \mathsf{d}Z^{l}$$

Forward & Backward Prop Summary

$$Z^{[l]} = W^{[l]} \cdot A^{[l-1]} + b^{[l]}$$

$$A^{[l]} = g^{[l]}(Z^{[l]})$$

$$\mathsf{d}W^{[l]} = (\mathsf{d}Z^{[l]} \cdot A^{[l-1]T}); \qquad \mathsf{d}W^{[L]} = A^{[L-1]} \cdot (A^{[L]} - Y)^T$$

$$\mathsf{d}A^{[l]} = W^{[l+1]T} \cdot \mathsf{d}Z^{[l+1]}$$

$$\mathsf{d}Z^{[l]} = \mathsf{d}A^{[l]} \circ g'^{[l]}(Z^{[l]})$$

$$\mathsf{d}b^{[l]} = \sum_{\text{axis}=1} \mathsf{d}Z^{l}; \qquad \mathsf{d}b^{[L]} = \sum_{\text{axis}=1} (\mathsf{d}A^{L} - Y)$$

Update Parameters

$$W^{[l]} = W^{[l]} - \alpha \, \mathsf{d}W^{[l]}$$
$$b^{[l]} = b^{[l]} - \alpha \, \mathsf{d}b^{[l]}$$