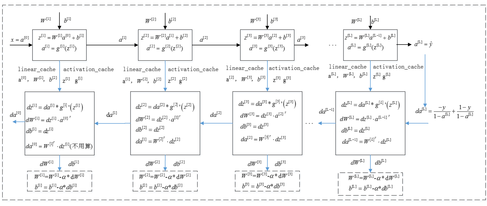
单样本输入时，L层神经网络参数更新图



Author：贺志尧
Email:2282815808@qq.com

单样本输入时，L层神经网络参数更新公式

Forward Propagation

$$Z^{[1]} = W^{[1]}x + b^{[1]}$$
$$A^{[1]} = g^{[1]}(Z^{[1]})$$
$$Z^{[2]} = W^{[2]}A^{[1]} + b^{[2]}$$
$$A^{[2]} = g^{[2]}(Z^{[2]})$$
$$\cdots$$
$$Z^{[L]} = W^{[L]}A^{[L-1]} + b^{[L]}$$
$$A^{[L]} = g^{[L]}(Z^{[L]})$$

$$A^{[L]} = W^{[L]}A^{[L-1]} + b^{[L]}$$
$$\hat{y} = g^{[L]}(Z^{[L]})$$

Backward Propagation

$$da^{[L]} = da^{[L]} * g^{[L]\prime}(Z^{[L]})$$
$$dW^{[L]} = da^{[L]} \cdot a^{[L-1]T}$$
$$db^{[L]} = da^{[L]}$$
$$da^{[L-1]} = W^{[L]T} \cdot da^{[L]}$$

$$\cdots$$

$$da^{[2]} = da^{[2]} * g^{[2]\prime}(Z^{[2]})$$
$$dW^{[2]} = da^{[2]} \cdot a^{[1]T}$$
$$db^{[2]} = da^{[2]}$$
$$da^{[1]} = W^{[2]T} \cdot da^{[2]}$$

$$da^{[1]} = da^{[1]} * g^{[1]\prime}(Z^{[1]})$$
$$dW^{[1]} = da^{[1]} \cdot x^{T}$$
$$db^{[1]} = da^{[1]}$$
$$dx = W^{[1]T} \cdot da^{[1]}\text{(不用算)}$$

注释：

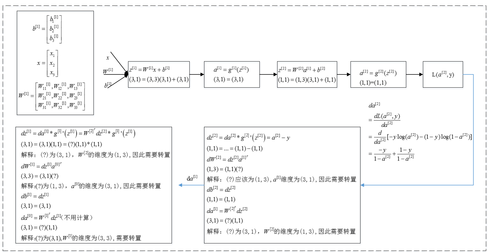
linear_cache: $W^{[1]}, b^{[1]}$

activation_cache: $Z^{[1]}$

$$da^{[L]} = \frac{dL(a^{[L]},Y)}{da^{[L]}} = \frac{d}{da^{[L]}}[-y\log(a^{[L]}) - (1-y)\log(1-a^{[L]})] = \frac{-y}{1-a^{[L]}} + \frac{1-y}{1-a^{[L]}}$$

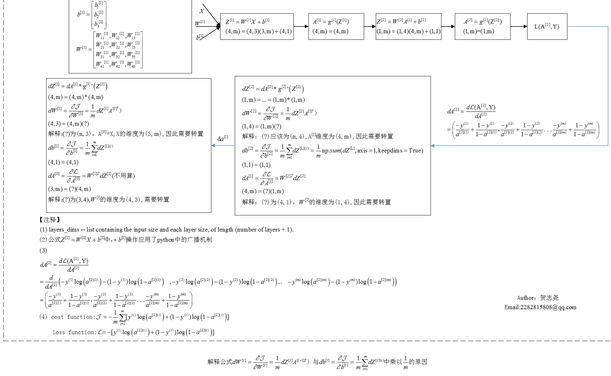
Author：贺志尧
Email:2282815808@qq.com

解释公式$dW^{[L]} = da^{[L]} \cdot a^{[L-1]T}$与$da^{[L-1]} = W^{[L]T} \cdot da^{[L]}$中，$a^{[L-1]T}$与$W^{[L]T}$转置的原因



Author：贺志尧
Email:2282815808@qq.com

m个样本输入时，L层神经网络参数更新详解(以2层为例)



【注释】

(1) layers_dims -- list containing the input size and each layer size, of length (number of layers + 1).

(2) 公式$Z^{[1]} = W^{[1]}X + b^{[1]}$中，$b^{[1]}$确实应用了python中的广播机制

(3)

$$dZ^{[2]} = \frac{dL(A^{[2]},Y)}{dZ^{[2]}}$$
$$= \frac{d}{dZ^{[2]}}(-y^{(m)}\log\left(a^{[2](m)}\right) - (1-y^{(1)})\log\left(1-a^{[2](1)}\right) \quad \cdots \quad -y^{(1)}\log\left(a^{[2](1)}\right) - (1-y^{(1)})\log\left(1-a^{[2](1)}\right) \quad \cdots \quad -y^{(m)}\log\left(a^{[2](m)}\right) - (1-y^{(m)})\log\left(1-a^{[2](m)}\right))$$
$$= \left(\frac{-y^{(1)}}{a^{[2](1)}} + \frac{1-y^{(1)}}{1-a^{[2](1)}} \quad \cdots \quad \frac{-y^{(m)}}{a^{[2](m)}} + \frac{1-y^{(m)}}{1-a^{[2](m)}}\right)$$

(4) cost function: $\mathcal{J} = -\frac{1}{m}\sum_{i=1}^{m}y^{(i)}\log\left(a^{[2](i)}\right) + (1-y^{(i)})\log\left(1-a^{[2](i)}\right)]$

loss function: $\mathcal{L} = -\left[y^{(i)}\log\left(a^{[2](i)}\right) + (1-y^{(i)})\log\left(1-a^{[2](i)}\right)\right]$

Author：贺志尧
Email:2282815808@qq.com

解释公式$dW^{[L]} = \frac{dJ}{dW^{[L]}} = \frac{1}{m}dZ^{[L]}A^{[L-1]T}$与$db^{[L]} = \frac{dJ}{dB^{[L]}} = \frac{1}{m}\sum_{i=1}^{m}dZ^{[L](i)}$中乘以$\frac{1}{m}$的原因

【证明1】

$$J = \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}^{(i)}$$
$$\frac{dJ}{dW^{[2]}} = \left[\frac{d\mathcal{L}^{(1)}}{dW^{[2]}} \quad \frac{d\mathcal{L}^{(2)}}{dW^{[2]}} \quad \cdots \quad \frac{d\mathcal{L}^{(m)}}{dW^{[2]}}\right] = \left[da^{[2](1)} \cdot a^{[1](1)T} \quad da^{[2](2)} \cdot a^{[1](2)T} \quad \cdots \quad da^{[2](m)} \cdot a^{[1](m)T}\right]$$
$$dW^{[2]}$$
$$= \frac{dJ}{dW^{[2]}}$$
$$= \frac{1}{m}\left(da^{[2](1)} \cdot a^{[1](1)T} + da^{[2](2)} \cdot a^{[1](2)T} + \ldots + da^{[2](m)} \cdot a^{[1](m)T}\right)$$
$$= \frac{1}{m}\left[da^{[2](1)} \quad da^{[2](2)} \quad \cdots \quad da^{[2](m)}\right]\begin{bmatrix} a^{[1](1)(1)} & a^{[1](2)(1)} & \cdots & a^{[1](m)(1)} \\ a^{[1](1)(2)} & a^{[1](2)(2)} & \cdots & a^{[1](m)(2)} \\ \vdots & \vdots & \ddots & \vdots \\ a^{[1](1)(n)} & a^{[1](2)(n)} & \cdots & a^{[1](m)(n)} \end{bmatrix}^{T}$$
$$= \frac{1}{m}dZ^{[2]}A^{[1]T}$$

【证明2】

$$J = \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}^{(i)}$$
$$\frac{dJ}{db^{[2]}} = \left[\frac{d\mathcal{L}^{(1)}}{db^{[2]}} \quad \frac{d\mathcal{L}^{(2)}}{db^{[2]}} \quad \cdots \quad \frac{d\mathcal{L}^{(m)}}{db^{[2]}}\right] = \left[da^{[2](1)} \quad da^{[2](2)} \quad \cdots \quad da^{[2](m)}\right]$$
$$db^{[2]}$$
$$= \frac{dJ}{db^{[2]}}$$
$$= \frac{1}{m}\left(da^{[2](1)} + da^{[2](2)} + \ldots + da^{[2](m)}\right)$$
$$= \frac{1}{m}\sum_{i=1}^{m}dZ^{[2]}$$

附录二：Backward Propagation

$$dZ^{[L]} = \frac{\partial\mathcal{L}}{\partial Z^{[L]}} = dA^{[L]} * g^{[L]\prime}(Z^{[L]})$$，若$g^{[L]}$为sigmoid函数，则$dZ^{[L]} = A^{[L]} - Y$

$$dW^{[L]} = \frac{\partial\mathcal{L}}{\partial W^{[L]}} = \frac{1}{m}dZ^{[L]} \cdot A^{[L-1]T}$$
$$db^{[L]} = \frac{\partial\mathcal{L}}{\partial b^{[L]}} = \frac{1}{m}\sum_{i=1}^{m}dZ^{[L](i)} = \frac{1}{m}\text{np.sum}(dZ^{[L]}, \text{axis} = 1, \text{keepdims} = \text{True})$$，其中axis = 1表示按行求和

$$dA^{[L-1]} = \frac{\partial\mathcal{L}}{\partial A^{[L-1]}} = W^{[L]T}dZ^{[L]}$$

$$\cdots$$

$$dZ^{[1]} = dA^{[1]} * g^{[1]\prime}(Z^{[1]})$$
$$dW^{[1]} = \frac{\partial\mathcal{J}}{\partial W^{[1]}} = \frac{1}{m}dZ^{[1]}A^{[0]T}$$
$$db^{[1]} = \frac{\partial\mathcal{J}}{\partial b^{[1]}} = \frac{1}{m}\sum_{i=1}^{m}dZ^{[1](i)} = \frac{1}{m}\text{np.sum}(dZ^{[1]}, \text{axis} = 1, \text{keepdims} = \text{True})$$
$$dA^{[0]} = \frac{\partial\mathcal{L}}{\partial A^{[0]}} = W^{[1]T}dZ^{[1]}\text{(不用算)}$$

附录1：注意比较下面两公式的区别

$$\frac{dJ(A^{[2]},Y)}{dA^{[2]}}$$
$$= \frac{d}{dA^{[2]}}\left[-\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}\log\left(a^{[2](i)}\right) + (1-y^{(i)})\log\left(1-a^{[2](i)}\right)\right)\right]$$
$$= \frac{1}{m}\left(\frac{-y^{(1)}}{a^{[2](1)}} + \frac{1-y^{(1)}}{1-a^{[2](1)}} \quad \cdots \quad \frac{-y^{(m)}}{a^{[2](m)}} + \frac{1-y^{(m)}}{1-a^{[2](m)}}\right)$$

$$\frac{dL(A^{[2]},Y)}{dA^{[2]}}$$
$$= \left(\frac{-y^{(1)}}{a^{[2](1)}} + \frac{1-y^{(1)}}{1-a^{[2](1)}} \quad \cdots \quad \frac{-y^{(m)}}{a^{[2](m)}} + \frac{1-y^{(m)}}{1-a^{[2](m)}}\right)$$

Author：贺志尧
Email:2282815808@qq.com

m个样本输入时，L层神经网络参数更新公式

Forward Propagation

$$X^{(1)} = X$$
$$Z^{[1](1)} = W^{[1]}X^{(1)} + b^{[1]}$$
$$A^{[1](1)} = g^{[1]}(Z^{[1](1)})$$
$$Z^{[2](1)} = W^{[2]}A^{[1](1)} + b^{[2]}$$
$$A^{[2](1)} = g^{[2]}(Z^{[2](1)})$$
$$\vdots$$
$$Z^{[L](1)} = W^{[L]}A^{[L-1](1)} + b^{[L]}$$
$$A^{[L](1)} = g^{[L]}(Z^{[L](1)}) = \hat{Y}$$

【注释】 cost function and loss function

cost function: $\mathcal{J} = -\frac{1}{m}\sum_{i=1}^{m}y^{(i)}\log\left(a^{[2](i)}\right) + (1-y^{(i)})\log\left(1-a^{[2](i)}\right)]$

loss function: $\mathcal{L} = -\left[y^{(i)}\log\left(a^{[2](i)}\right) + (1-y^{(i)})\log\left(1-a^{[2](i)}\right)\right]$

Backward Propagation

$$dZ^{[L]} = \frac{\partial\mathcal{L}}{\partial Z^{[L]}} = dA^{[L]} * g^{[L]\prime}(Z^{[L]})$$，若$g^{[L]}$为sigmoid函数，则$dZ^{[L]} = A^{[L]} - Y$

$$dW^{[L]} = \frac{\partial\mathcal{J}}{\partial W^{[L]}} = \frac{1}{m}dZ^{[L]} \cdot A^{[L-1]T})$$
$$db^{[L]} = \frac{\partial\mathcal{J}}{\partial b^{[L]}} = \frac{1}{m}\sum_{i=1}^{m}dZ^{[L](i)} = \frac{1}{m}\text{np.sum}(dZ^{[L]}, \text{axis} = 1, \text{keepdims} = \text{True})$$
$$dA^{[L-1]} = \frac{\partial\mathcal{L}}{\partial A^{[L-1]}} = W^{[L]T}dZ^{[L]}$$

$$\vdots$$

$$dZ^{[1]} = dA^{[1]} * g^{[1]\prime}(Z^{[1]})$$
$$dW^{[1]} = \frac{\partial\mathcal{J}}{\partial W^{[1]}} = \frac{1}{m}dZ^{[1]}A^{[0]T})$$
$$db^{[1]} = \frac{\partial\mathcal{J}}{\partial b^{[1]}} = \frac{1}{m}\sum_{i=1}^{m}dZ^{[1](i)} = \frac{1}{m}\text{np.sum}(dZ^{[1]}, \text{axis} = 1, \text{keepdims} = \text{True})$$
$$dA^{[0]} = \frac{\partial\mathcal{L}}{\partial A^{[0]}} = W^{[1]T}dZ^{[1]}\text{(不用算)}$$

Author：贺志尧
Email:2282815808@qq.com