Neural Networks

Notations

$W_{ij}^{(l)}$: $1 \le l \le n_l$ layers

$1 \le j \le s_l$ inputs

$1 \le i \le s_{l+1}$ outputs

$a^{(l)}$: the activation value of the layer l

$a_i^{(l)}$: the ith activation value of the layer l

Computing Boolean Functions

$z = (w_1 x_1 + w_2 x_2 + b)$, $f(z) = \frac{1}{1+e^{-z}}$

AND: $w_1 = 20$, $w_2 = 20$, $b = -30$

OR: $w_1 = 20$, $w_2 = 20$, $b = -10$

NOT: $w_1 = -20$, $b = 10$

Forward Propagation

$a^{(2)} = f(W^{(1)} a^{(1)} + b^{(1)})$

$a^{(3)} = f(W^{(2)} a^{(2)} + b^{(2)}) = f(W^{(2)} f(W^{(1)} a^{(1)} + b^{(1)}) + b^{(2)})$

Algorithm:

$a^{(1)} = x$

for l = 1 to $n_l - 1$:

  $z^{(l+1)} = W^{(l)} a^{(l)} + b^{(l)}$

  $a^{(l+1)} = f(z^{(l+1)})$

$\hat{y} = a^{(n_l)}$

Batch gradient descent algorithm

$J(w, b, x, y) = \frac{1}{2}(y - \hat{y})^2 = \frac{1}{2}(y - a^{(n_l)})^2$

Activation: $a_i^{(l)} = f(z_i^{(l)})$

Algorithm:

Randomly initialize the weights for each layer: $W^{(l)}$

While iterations < iteration limit:

$\Delta W^{(l)} = 0$

$\Delta b^{(l)} = 0$

For i = 1 to N:

Run forward propagation and save for each level, the value $a^{(l)}$, $z^{(l)}$

Run backward propagation to calculate δ_k for each level 2-n

Update $\Delta W^{(l)} = 0$, $\Delta b^{(l)} = 0$ for each layer

Perform a gradient descent:

$W^{(l)} = W^{(l)} - \alpha \cdot \frac{1}{N} \cdot \Delta W^{(l)}$

$b^{(l)} = b^{(l)} - \alpha \cdot \frac{1}{N} \cdot \Delta b^{(l)}$

$\delta_j^{(l)} = \frac{\partial J}{\partial z_j^{(l)}} = \sum_{i=1}^{s_{l+1}} \delta_i^{(l+1)} W_{ij}^{(l)} f'(z_j^{(l)}) = \sum_{i=1}^{s_{l+1}} \delta_i^{(l+1)} W_{ij}^{(l)} a_j^{(l)} (1 - a_j^{(l)})$

$\frac{\partial J}{\partial W_{ij}^{(l)}} = \frac{\partial J}{\partial z_j^{(l)}} \frac{\partial z_j^{(l)}}{\partial W_{ij}^{(l)}} = \delta_i^{(l+1)} a_j^{(l)T}$

$\frac{\partial J}{\partial b_i^{(l)}} = \frac{\partial J}{\partial z_j^{(l)}} \frac{\partial z_j^{(l)}}{\partial b_i^{(l)}} = \delta_i^{(l+1)}$

CNN

Convolution layer (filter)

Convolution: $(W - F + 1)^2$

Padding when using convolution: $(W + 2P - F + 1)^2$

Striding when using convolution: $((W + 2P - S)/F + 1)^2$

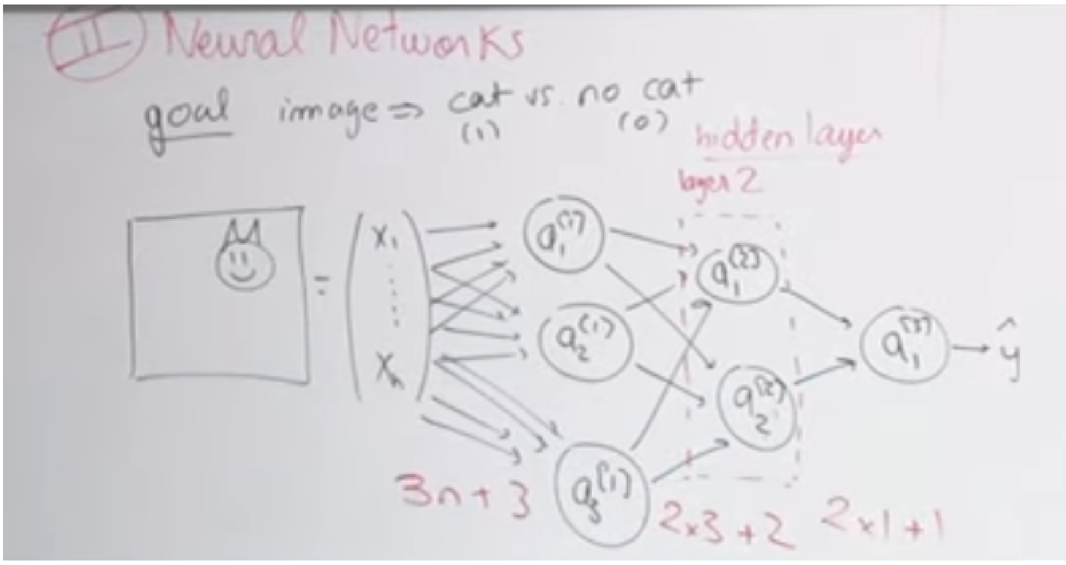
Same convolution: the size of the matrix s the same after a convolution

Max Pooling layer — reduce the size

CS229

neuron = linear + activation

model = architecture + parameter



Propogation equations:

$z^{(1)} = w^{(1)}x + b^{(1)}$

$a^{(1)} = \sigma(z^{(1)})$ — layer 1

$z^{(2)} = w^{(2)}x + b^{(2)}$

$a^{(2)} = \sigma(z^{(2)})$ — layer 2

$z^{(3)} = w^{(3)}x + b^{(3)}$

$a^{(3)} = \sigma(z^{(3)})$ — layer 3

Optimizing $w^{(1)}, w^{(2)}, w^{(3)}, b^{(1)}, b^{(2)}, b^{(3)}$

Loss/cost function: $J(\hat{y}, y) = \frac{1}{m}\Sigma_{i=1}^{m} L^{(i)}(\hat{y}, y)$,

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

Backward Propagation:

$\forall l = 1, 2, 3 \quad w^{(l)} = w^{(l)} - \alpha\frac{\partial J}{\partial w^{(l)}}$

$b^{(l)} = b^{(l)} - \alpha\frac{\partial J}{\partial b^{(l)}}$

$\frac{\partial J}{\partial w^{(3)}} = \frac{\partial J}{\partial a^{(3)}} \cdot \frac{\partial a^{(3)}}{\partial z^{(3)}} \cdot \frac{\partial z^{(3)}}{\partial w^{(3)}}$

$\frac{\partial J}{\partial w^{(2)}} = \frac{\partial J}{\partial z^{(3)}} \cdot \frac{\partial z^{(3)}}{\partial a^{(2)}} \cdot \frac{\partial a^{(2)}}{\partial z^{(2)}} \cdot \frac{\partial z^{(2)}}{\partial w^{(2)}}$, where $\frac{\partial J}{\partial z^{(3)}} = \frac{\partial J}{\partial a^{(3)}} \cdot \frac{\partial a^{(3)}}{\partial z^{(3)}}$

$\frac{\partial J}{\partial w^{(1)}} = \frac{\partial J}{\partial z^{(2)}} \cdot \frac{\partial z^{(2)}}{\partial a^{(1)}} \cdot \frac{\partial a^{(1)}}{\partial z^{(1)}} \cdot \frac{\partial z^{(1)}}{\partial w^{(1)}}$, where $\frac{\partial J}{\partial z^{(2)}} = \frac{\partial J}{\partial z^{(3)}} \cdot \frac{\partial z^{(3)}}{\partial a^{(2)}} \cdot \frac{\partial a^{(2)}}{\partial z^{(2)}}$

$\frac{\partial L}{\partial w^{(3)}} = -[y^{(i)}\frac{\partial}{\partial w^{(3)}}log(\sigma(w^{(3)}a^{(2)} + b^{(3)})) + (1 - y^{(i)})\frac{\partial}{\partial w^{(3)}}log(1 - \sigma(w^{(3)}a^{(2)} + b^{(3)}))]$

$= -(y^{(i)} - a^{(3)})a^{(2)T}$

$\frac{\partial J}{\partial w^{(3)}} = -\frac{1}{m}\Sigma_{i=1}^{m}(y^{(i)} - a^{(3)})a^{(2)T}$

$\frac{\partial L}{\partial w^{(2)}} = (\frac{\partial J}{\partial z^{(3)}}) \cdot \frac{\partial z^{(3)}}{\partial a^{(2)}} \cdot \frac{\partial a^{(2)}}{\partial z^{(2)}} \cdot \frac{\partial z^{(2)}}{\partial w^{(2)}} = [-(y - a^{(3)})][w^{(3)T}][a^{(2)}(1 - a^{(2)})][a^{(1)T}]$

$= w^{(3)T}a^{(2)}(1 - a^{(2)})(a^{(3)} - y)a^{(i)T}$

$\frac{\partial J}{\partial w^{(3)}} = \frac{1}{m}\Sigma_{i=1}^{m}\frac{\partial L}{\partial w^{(i)}}$