BACKPROPAGATION ILLUSTRATED

D7046E - LTU.SE
NEURAL NETWORKS
AND LEARNING MACHINES

- MATRIX CALCULUS RULE

$$\frac{\partial(u \cdot v)}{\partial x} = u^T \frac{\partial v}{\partial x} + v^T \frac{\partial u}{\partial x}$$

- WEIGHT MATRICES w_i
- (HIDDEN) STATES h_i
- (ACTIVATION) FUNCTIONS f_i

- MINIBATCH OF SIZE M

DIMENSION	NETWORK	RELATION	DERIVATIVE
$M \times n_3$	LOSS L	$L(\hat{y}, y)$	
$M \times n_3$	OUTPUT $\hat{y}$	$\hat{y} = f_3(h_2 w_3)$	$\frac{\partial \hat{y}}{\partial w_3} = h_2^T f_3'(h_2 w_3)$ NOTATION: f_3'
$n_2 \times n_3$	w_3		
$M \times n_2$	HIDDEN h_2	$h_2 = f_2(h_1 w_2)$	$\frac{\partial \hat{y}}{\partial w_2} = \frac{\partial h_2}{\partial w_2}\frac{\partial \hat{y}}{\partial h_2} = h_1^T f_2' f_3' w_3^T$
$n_1 \times n_2$	w_2		
$M \times n_1$	HIDDEN h_1	$h_1 = f_1(x w_1)$	$\frac{\partial \hat{y}}{\partial w_1} = \frac{\partial h_1}{\partial w_1}\frac{\partial h_2}{\partial h_1}\frac{\partial \hat{y}}{\partial h_2} = x^T f_1' f_2' f_3' w_3^T w_2^T$
$n_0 \times n_1$	w_1		
$M \times n_0$	INPUT x		

DERIVATIVES OF LOSS WITH RESPECT TO THE WEIGHTS

$$\frac{\partial L}{\partial \hat{y}} = L' \qquad M \times n_3$$

$$\frac{\partial L}{\partial w_3} = \frac{\partial \hat{y}}{\partial w_3}\frac{\partial L}{\partial \hat{y}} = h_2^T f_3' L'$$

ELEMENT-WISE MULT.

$$(n_2 \times M)\left((M \times n_3) \odot (M \times n_3)\right) = n_2 \times n_3$$

SUMMING OVER MINIBATCH, AVERAGING OF GRADIENTS

BP

$$\frac{\partial L}{\partial w_2} = \frac{\partial \hat{y}}{\partial w_2}\frac{\partial L}{\partial \hat{y}} = h_1^T f_2' f_3' L' w_3^T$$

$$(n_1 \times M)\left(M \times n_2 \odot (M \times n_3)(n_3 \times n_2)\right) = n_1 \times n_2$$

BP

$$\frac{\partial L}{\partial w_1} = \frac{\partial \hat{y}}{\partial w_1}\frac{\partial L}{\partial \hat{y}} = x^T f_1' f_2' f_3' L' w_3^T w_2^T$$

$$(n_0 \times M)\left((M \times n_1) \odot \left((M \times n_2)(n_2 \times n_1)\right)\right) = n_0 \times n_1$$

BP: BACKPROPAGATION, THIS FACTOR CAN BE REUSED, DON'T COMPUTE IT AGAIN

STOCHASTIC GRADIENT DESCENT

$$w_i \leftarrow w_i - \gamma \cdot \frac{\partial L}{\partial w_i}$$

CONSISTENT DIMENSIONS

STEP SIZE (LEARNING RATE)

CALLED "STOCHASTIC" BECAUSE THE TRUE GRADIENT IS UNKNOWN: THE GRADIENT CALCULATED HERE IS VALID FOR THE DATA IN THE MINIBATCH