Matrix Calculus: Essential Rules for Neural Networks

Vector Derivatives

$\frac{\partial}{\partial x}(Ax)$	$= A$	*Matrix-vector product*
$\frac{\partial}{\partial x}(x^T A)$	$= A^T$	*Vector-matrix product*
$\frac{\partial}{\partial x}(x^T x)$	$= 2x$	*Squared norm*
$\frac{\partial}{\partial x}(x^T A x)$	$= (A + A^T)x$	*Quadratic form*

Chain Rule

$\frac{\partial L}{\partial W}$	$= \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial W}$	*Scalar through matrix*
$\frac{\partial L}{\partial x}$	$= W^T \frac{\partial L}{\partial y}$	*Through linear layer*
$\frac{\partial L}{\partial W}$	$= \frac{\partial L}{\partial y} x^T$	*Weight gradient*

Common Layer Gradients

Linear: $y = Wx + b$	$\frac{\partial L}{\partial W} = \delta x^T$	$\frac{\partial L}{\partial b} = \delta$
Sigmoid: $y = \sigma(z)$	$\frac{\partial L}{\partial z} = \delta \odot \sigma(z)(1 - \sigma(z))$	
ReLU: $y = \max(0, z)$	$\frac{\partial L}{\partial z} = \delta \odot \mathbf{1}_{z>0}$	

Key Conventions

$\delta = \frac{\partial L}{\partial y}$ (upstream gradient) | $\odot$ = element-wise multiplication | All vectors are column vectors by default