

# Deep Learning

## 05 Backpropagation-1

Dr. Konda Reddy Mopuri  
Dept. of AI, IIT Hyderabad  
Jan-May 2024

# Recap

- Gradient of a scalar valued function  $f(\mathbf{x}): \mathbf{x} \rightarrow \left( \frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_D} \right)^T$

# Recap

- Gradient of a scalar valued function  $f(\mathbf{x})$ :  $\mathbf{x} \rightarrow \left( \frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_D} \right)^T$
- Gradient of a vector valued function  $\mathbf{f}(\mathbf{x})$  is called Jacobian:

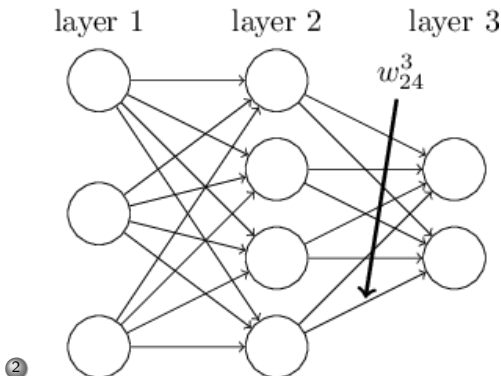
$$\mathbf{J} = \left[ \frac{\partial \mathbf{f}}{\partial x_1} \quad \dots \quad \frac{\partial \mathbf{f}}{\partial x_n} \right] = \begin{bmatrix} \nabla^T f_1 \\ \vdots \\ \nabla^T f_m \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \dots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

# MLP: Some Notation

- ①  $w_{jk}^l$  is the weight connecting  $j^{th}$  neuron in  $l^{th}$  layer and  $k^{th}$  neuron in  $(l - 1)^{st}$  layer

# MLP: Some Notation

- ①  $w_{jk}^l$  is the weight connecting  $j^{th}$  neuron in  $l^{th}$  layer and  $k^{th}$  neuron in  $(l - 1)^{st}$  layer



# MLP: Some Notation

- ①  $b_j^l$  is the bias of  $j^{th}$  neuron in  $l^{th}$  layer

# MLP: Some Notation

- ①  $b_j^l$  is the bias of  $j^{th}$  neuron in  $l^{th}$  layer
- ②  $x_j^l$  is the activation (output) of  $j^{th}$  neuron in  $l^{th}$  layer

# MLP: Some Notation

- ①  $b_j^l$  is the bias of  $j^{th}$  neuron in  $l^{th}$  layer
- ②  $x_j^l$  is the activation (output) of  $j^{th}$  neuron in  $l^{th}$  layer
- ③

$$x_j^l = \sigma\left(\sum_k w_{jk}^l x_k^{l-1} + b_j^l\right)$$



# MLP: Some Notation

- ①  $b_j^l$  is the bias of  $j^{th}$  neuron in  $l^{th}$  layer
- ②  $x_j^l$  is the activation (output) of  $j^{th}$  neuron in  $l^{th}$  layer

③

$$x_j^l = \sigma\left(\sum_k w_{jk}^l x_k^{l-1} + b_j^l\right)$$

- ④ Vector of activations (or, biases) at a layer  $l$  is denoted by a bold-faced  $\mathbf{x}^l$  ( or  $\mathbf{b}^l$ ) and  $W^l$  is the matrix of weights into layer  $l$

# MLP: Some Notation

①  $s_j^l$  is the weighted input to  $j^{th}$  neuron in  $l^{th}$  layer

# MLP: Some Notation

- ①  $s_j^l$  is the weighted input to  $j^{th}$  neuron in  $l^{th}$  layer
- ②  $s_j^l = \sum_k w_{jk}^l x_k^{l-1} + b_j^l$

# MLP: Some Notation

- ①  $s_j^l$  is the weighted input to  $j^{th}$  neuron in  $l^{th}$  layer
- ②  $s_j^l = \sum_k w_{jk}^l x_k^{l-1} + b_j^l$
- ③  $\mathbf{s}^l = W^l \mathbf{x}^{l-1} + \mathbf{b}^l$

# MLP: Some Notation

- ①  $s_j^l$  is the weighted input to  $j^{th}$  neuron in  $l^{th}$  layer
- ②  $s_j^l = \sum_k w_{jk}^l x_k^{l-1} + b_j^l$
- ③  $\mathbf{s}^l = W^l \mathbf{x}^{l-1} + \mathbf{b}^l$
- ④  $\sigma$  is the activation function that applies element-wise

# Gradient descent on MLP

- Loss is  $\mathcal{L}(W, \mathbf{b}) = \sum_n l(f(x_n; W, \mathbf{b}), y_n) = \sum_n l(\mathbf{x}^L, y_n)$  ( $L$  is the number of layers in the MLP)

# Gradient descent on MLP

- Loss is  $\mathcal{L}(W, \mathbf{b}) = \sum_n l(f(x_n; W, \mathbf{b}), y_n) = \sum_n l(\mathbf{x}^L, y_n)$  ( $L$  is the number of layers in the MLP)
- For applying Gradient descent, we need gradient of individual sample loss with respect to all the model parameters

$$l_n = l(f(x_n; W, \mathbf{b}), y_n)$$

$$\frac{\partial l_n}{\partial W_{jk}^{(l)}} \text{ and } \frac{\partial l_n}{\partial \mathbf{b}_j^{(l)}} \text{ for all layers } l$$

# Forward pass operation

$$x^{(0)} = x \xrightarrow{W^{(1)}, \mathbf{b}^{(1)}} s^{(1)} \xrightarrow{\sigma} x^{(1)} \xrightarrow{W^{(2)}, \mathbf{b}^{(2)}} s^{(2)} \dots x^{(L-1)} \xrightarrow{W^{(L)}, \mathbf{b}^{(L)}} s^{(L)} \xrightarrow{\sigma} x^{(L)} = f(x; W, \mathbf{b})$$

Formally,  $x^{(0)} = x, f(x; W, \mathbf{b}) = x^{(L)}$

$$\forall l = 1, \dots, L \quad \begin{cases} s^{(l)} & = W^{(l)}x^{(l-1)} + \mathbf{b}^{(l)} \\ x^{(l)} & = \sigma(s^{(l)}) \end{cases}$$



# Chain rule of differential calculus

- Core concept of backpropagation

# Chain rule of differential calculus

- Core concept of backpropagation



$$(f \circ g)'(x) = f'(g(x)) \cdot g'(x)$$

# Chain rule of differential calculus

- Core concept of backpropagation




$$(f \circ g)'(x) = f'(g(x)) \cdot g'(x)$$



$$\frac{\partial}{\partial x} f(g(x)) = \frac{\partial f(a)}{\partial a} \Big|_{a=g(x)} \cdot \frac{\partial g(x)}{\partial x}$$

# Chain rule of differential calculus



The Chain Rule

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$$
$$\frac{dy}{dx} = \left( \begin{array}{l} \text{Differentiate} \\ \text{outer function} \\ \text{Keep the inside} \\ \text{the same} \end{array} \right) \left( \begin{array}{l} \text{Differentiate} \\ \text{inner function} \end{array} \right)$$

© Maths at Home [www.mathsathome.com](http://www.mathsathome.com)

# Chain rule of differential calculus

- For any nested function  $y = f(g(x))$

# Chain rule of differential calculus

- For any nested function  $y = f(g(x))$
- $\frac{dy}{dx} = \frac{\partial f}{\partial g(x)} \frac{dg(x)}{dx}$

# Chain rule of differential calculus

- For any nested function  $y = f(g(x))$
- $\frac{dy}{dx} = \frac{\partial f}{\partial g(x)} \frac{dg(x)}{dx}$
- $\Delta y = \frac{dy}{dx} \Delta x$

# Chain rule of differential calculus

- For any nested function  $y = f(g(x))$

- $\frac{dy}{dx} = \frac{\partial f}{\partial g(x)} \frac{dg(x)}{dx}$

- $\Delta y = \frac{dy}{dx} \Delta x$

- $z = g(x) \rightarrow \Delta z = \frac{dg(x)}{dx} \Delta x$



# Chain rule of differential calculus

- For any nested function  $y = f(g(x))$
- $\frac{dy}{dx} = \frac{\partial f}{\partial g(x)} \frac{dg(x)}{dx}$
- $\Delta y = \frac{dy}{dx} \Delta x$
- $z = g(x) \rightarrow \Delta z = \frac{dg(x)}{dx} \Delta x$
- $y = f(z) \rightarrow \Delta y = \frac{df}{dz} \Delta z = \frac{df}{dz} \frac{dg(x)}{dx} \Delta x = \frac{df}{dg(x)} \frac{dg(x)}{dx} \Delta x$

# Distributed Chain rule of differential calculus



$$\textcircled{1} \quad y = f(g_1(x), g_2(x), \dots, g_M(x))$$

# Distributed Chain rule of differential calculus

$$\textcircled{1} \quad y = f(g_1(x), g_2(x), \dots, g_M(x))$$

$$\textcircled{2} \quad \frac{dy}{dx} = \frac{\partial f}{\partial g_1(x)} \frac{dg_1(x)}{dx} + \frac{\partial f}{\partial g_2(x)} \frac{dg_2(x)}{dx} + \dots + \frac{\partial f}{\partial g_M(x)} \frac{dg_M(x)}{dx}$$

# Distributed Chain rule of differential calculus

$$\textcircled{1} \quad y = f(g_1(x), g_2(x), \dots, g_M(x))$$

$$\textcircled{2} \quad \frac{dy}{dx} = \frac{\partial f}{\partial g_1(x)} \frac{dg_1(x)}{dx} + \frac{\partial f}{\partial g_2(x)} \frac{dg_2(x)}{dx} + \dots + \frac{\partial f}{\partial g_M(x)} \frac{dg_M(x)}{dx}$$

$$\textcircled{3} \quad \text{Let } g_i(x) = z_i \rightarrow y = f(z_1, z_2, \dots, z_M)$$

# Distributed Chain rule of differential calculus

$$\textcircled{1} \quad y = f(g_1(x), g_2(x), \dots, g_M(x))$$

$$\textcircled{2} \quad \frac{dy}{dx} = \frac{\partial f}{\partial g_1(x)} \frac{dg_1(x)}{dx} + \frac{\partial f}{\partial g_2(x)} \frac{dg_2(x)}{dx} + \dots + \frac{\partial f}{\partial g_M(x)} \frac{dg_M(x)}{dx}$$

$$\textcircled{3} \quad \text{Let } g_i(x) = z_i \rightarrow y = f(z_1, z_2, \dots, z_M)$$

$$\textcircled{4} \quad \Delta y = \frac{\partial f}{\partial z_1} \Delta z_1 + \frac{\partial f}{\partial z_2} \Delta z_2 + \dots + \frac{\partial f}{\partial z_M} \Delta z_M$$

# Distributed Chain rule of differential calculus

$$\textcircled{1} \quad y = f(g_1(x), g_2(x), \dots, g_M(x))$$

$$\textcircled{2} \quad \frac{dy}{dx} = \frac{\partial f}{\partial g_1(x)} \frac{dg_1(x)}{dx} + \frac{\partial f}{\partial g_2(x)} \frac{dg_2(x)}{dx} + \dots + \frac{\partial f}{\partial g_M(x)} \frac{dg_M(x)}{dx}$$

$$\textcircled{3} \quad \text{Let } g_i(x) = z_i \rightarrow y = f(z_1, z_2, \dots, z_M)$$

$$\textcircled{4} \quad \Delta y = \frac{\partial f}{\partial z_1} \Delta z_1 + \frac{\partial f}{\partial z_2} \Delta z_2 + \dots + \frac{\partial f}{\partial z_M} \Delta z_M$$

$$\textcircled{5} \quad \Delta y = \frac{\partial f}{\partial z_1} \frac{dz_1}{dx} \Delta x + \frac{\partial f}{\partial z_2} \frac{dz_2}{dx} \Delta x + \dots + \frac{\partial f}{\partial z_M} \frac{dz_M}{dx} \Delta x$$

# Distributed Chain rule of differential calculus

$$\textcircled{1} \quad y = f(g_1(x), g_2(x), \dots, g_M(x))$$

$$\textcircled{2} \quad \frac{dy}{dx} = \frac{\partial f}{\partial g_1(x)} \frac{dg_1(x)}{dx} + \frac{\partial f}{\partial g_2(x)} \frac{dg_2(x)}{dx} + \dots + \frac{\partial f}{\partial g_M(x)} \frac{dg_M(x)}{dx}$$

$$\textcircled{3} \quad \text{Let } g_i(x) = z_i \rightarrow y = f(z_1, z_2, \dots, z_M)$$

$$\textcircled{4} \quad \Delta y = \frac{\partial f}{\partial z_1} \Delta z_1 + \frac{\partial f}{\partial z_2} \Delta z_2 + \dots + \frac{\partial f}{\partial z_M} \Delta z_M$$

$$\textcircled{5} \quad \Delta y = \frac{\partial f}{\partial z_1} \frac{dz_1}{dx} \Delta x + \frac{\partial f}{\partial z_2} \frac{dz_2}{dx} \Delta x + \dots + \frac{\partial f}{\partial z_M} \frac{dz_M}{dx} \Delta x$$

$$\textcircled{6} \quad \Delta y = \frac{\partial f}{\partial g_1(x)} \frac{dg_1(x)}{dx} \Delta x + \frac{\partial f}{\partial g_2(x)} \frac{dg_2(x)}{dx} \Delta x + \dots + \frac{\partial f}{\partial g_M(x)} \frac{dg_M(x)}{dx} \Delta x$$

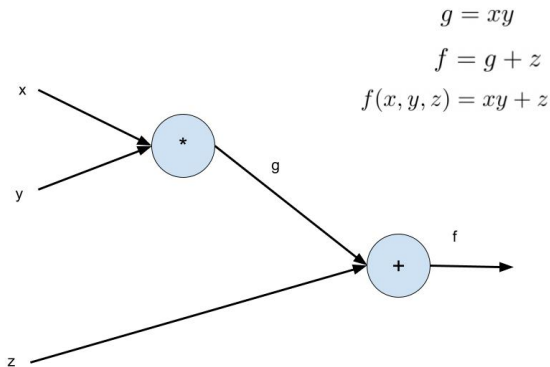
$$\textcircled{7} \quad \Delta y = \left( \frac{\partial f}{\partial g_1(x)} \frac{dg_1(x)}{dx} + \frac{\partial f}{\partial g_2(x)} \frac{dg_2(x)}{dx} + \dots + \frac{\partial f}{\partial g_M(x)} \frac{dg_M(x)}{dx} \right) \Delta x$$

# Chain rule of differential calculus

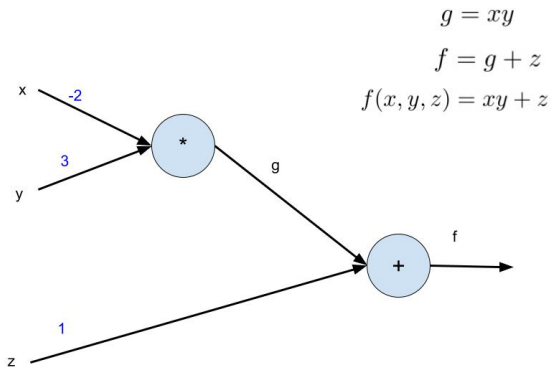
①  $f(x) = e^{\sin(x^2)}$ , let's find  $\frac{\partial f}{\partial x}$



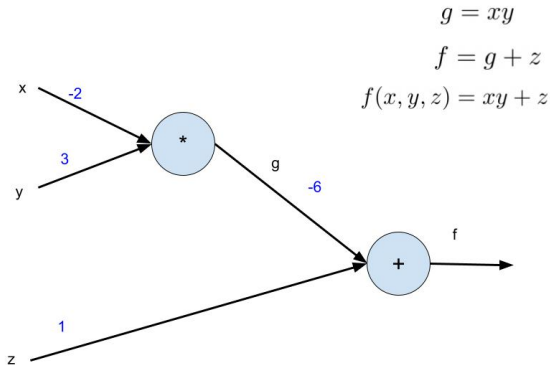
# Chain rule of differential calculus



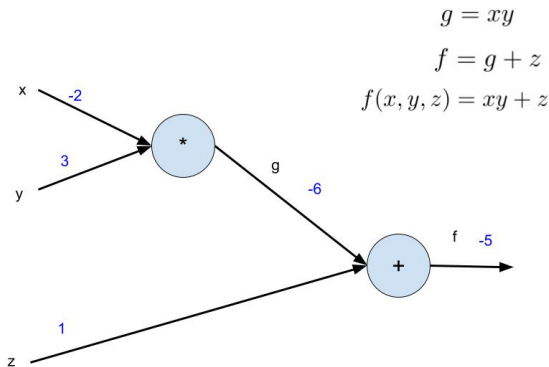
# Chain rule of differential calculus



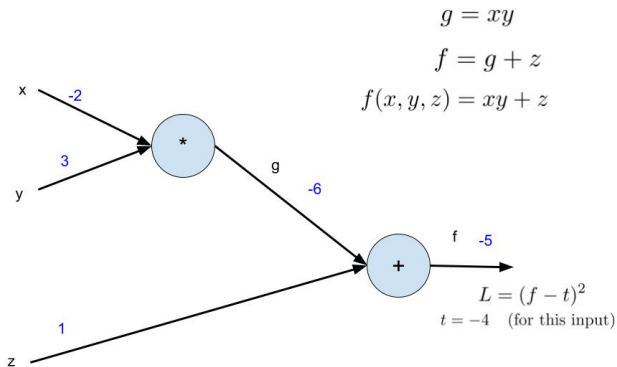
# Chain rule of differential calculus



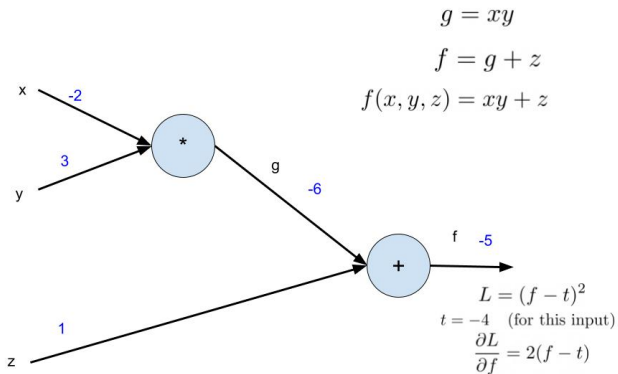
# Chain rule of differential calculus



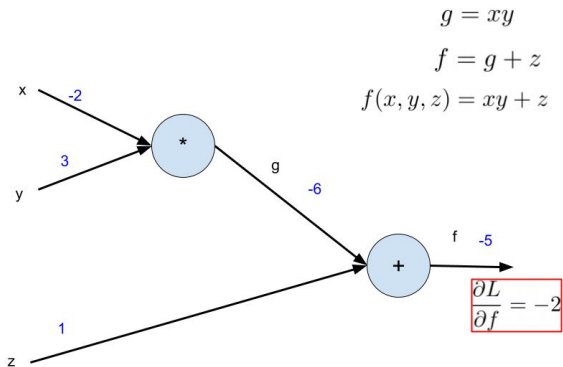
# Chain rule of differential calculus



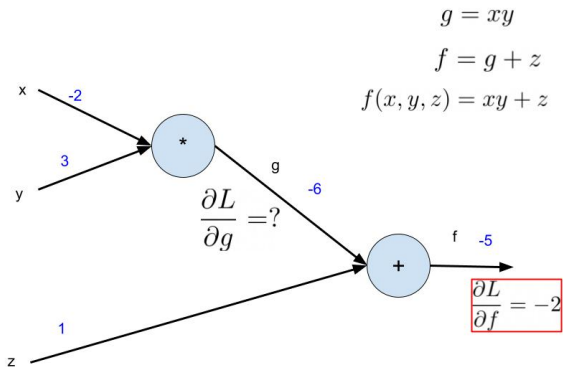
# Chain rule of differential calculus



# Chain rule of differential calculus

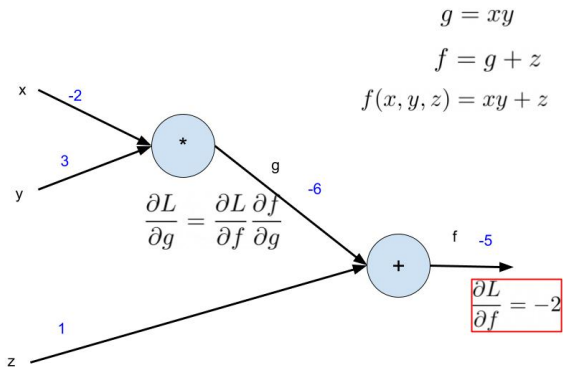


# Chain rule of differential calculus

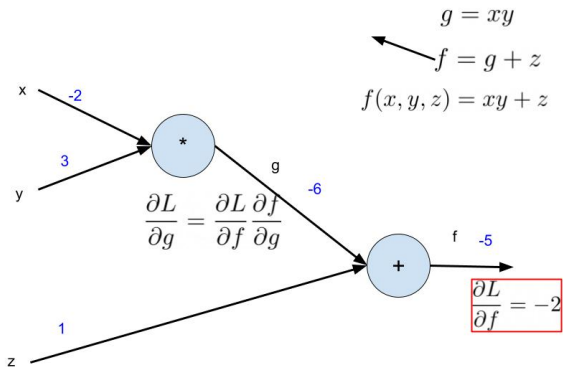




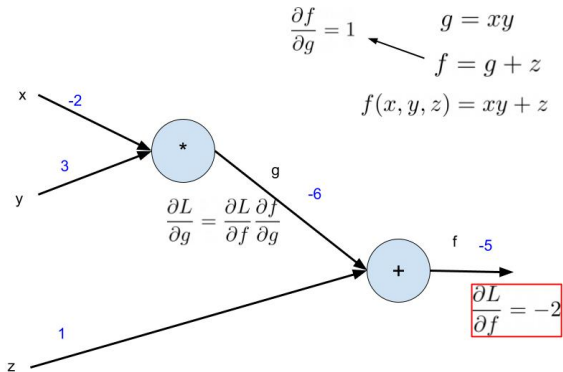
# Chain rule of differential calculus



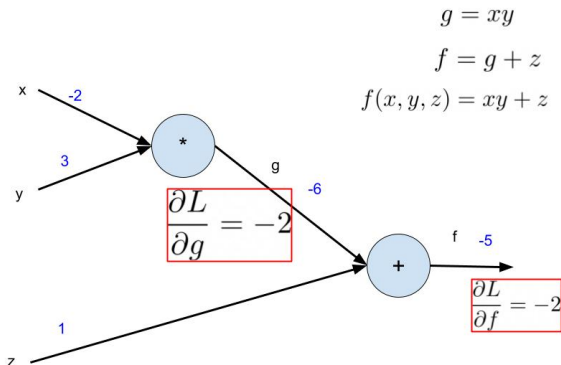
# Chain rule of differential calculus



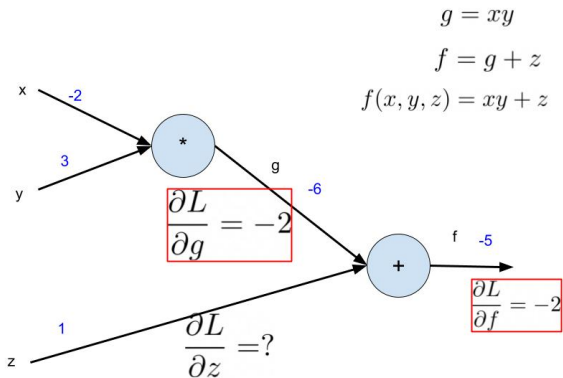
# Chain rule of differential calculus



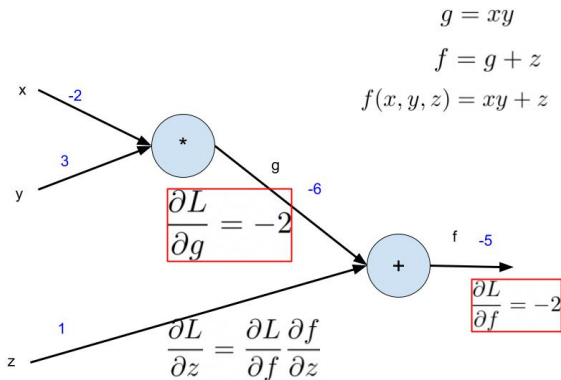
# Chain rule of differential calculus



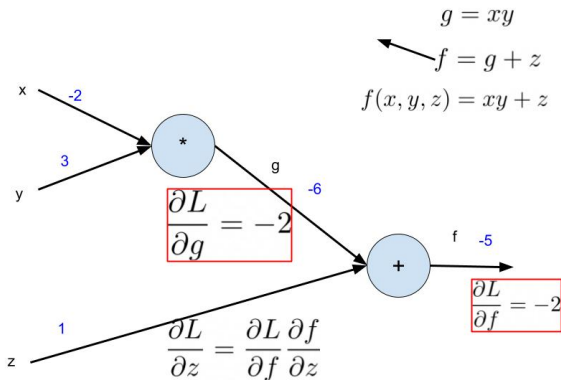
# Chain rule of differential calculus



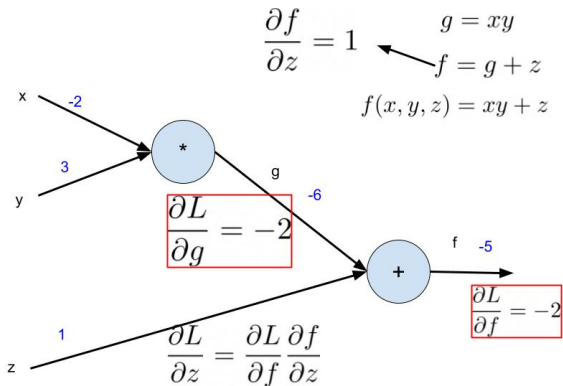
# Chain rule of differential calculus



# Chain rule of differential calculus

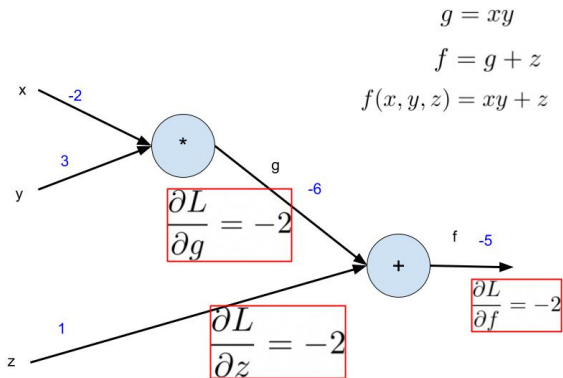


# Chain rule of differential calculus

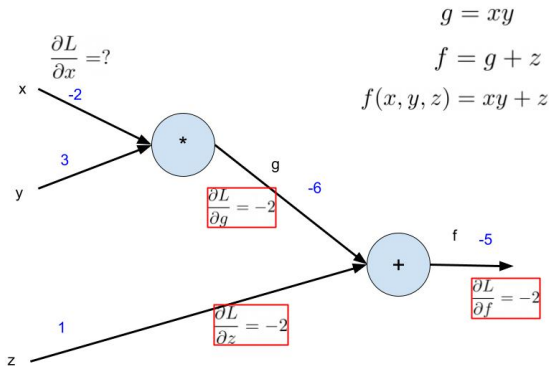




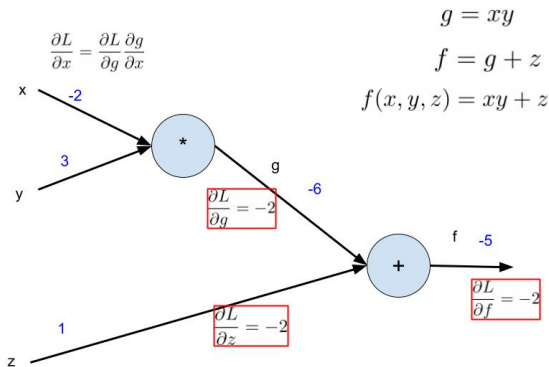
# Chain rule of differential calculus



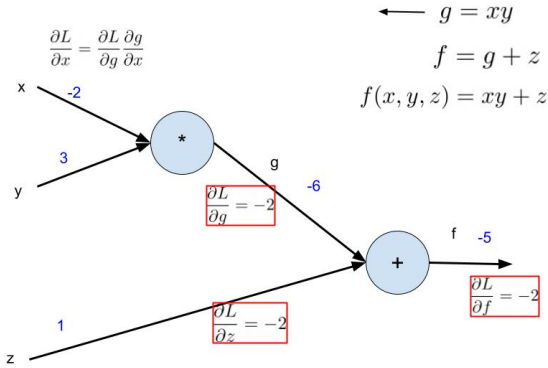
# Chain rule of differential calculus



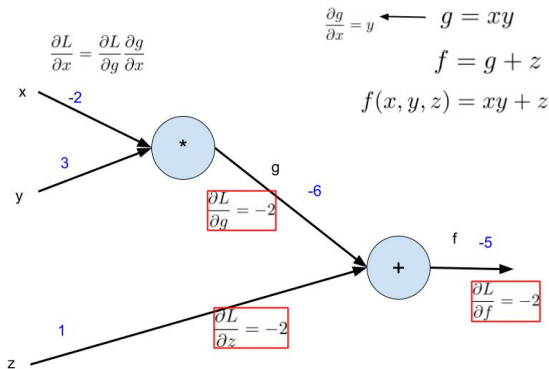
# Chain rule of differential calculus



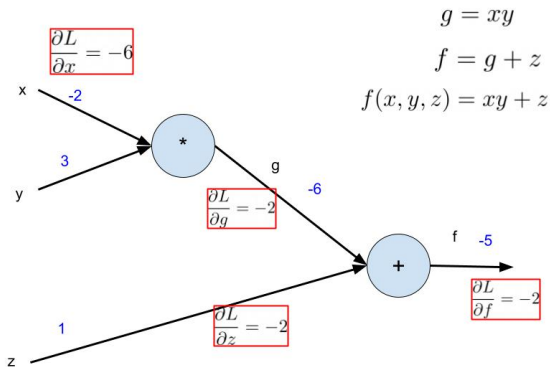
# Chain rule of differential calculus



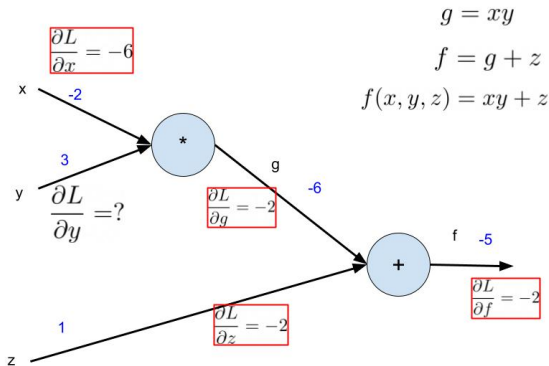
# Chain rule of differential calculus



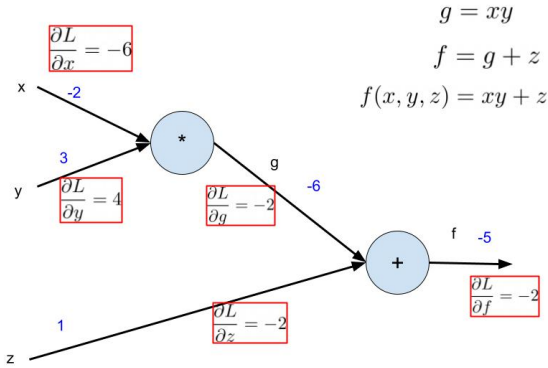
# Chain rule of differential calculus



# Chain rule of differential calculus

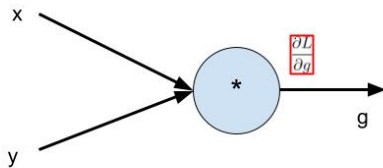


# Chain rule of differential calculus

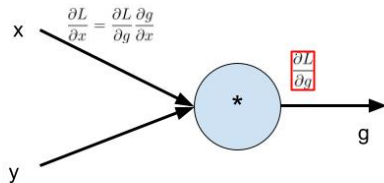




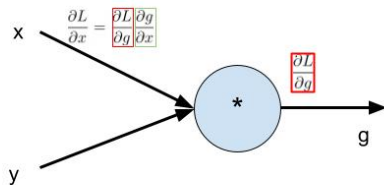
# Gradient Flow



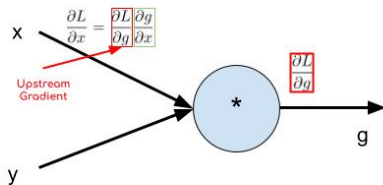
# Gradient Flow



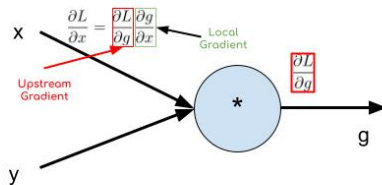
# Gradient Flow



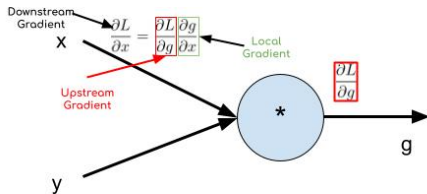
# Gradient Flow



# Gradient Flow



# Gradient Flow



# Chain rule of differential calculus for an MLP



$$J_{f_N \circ f_{N-1} \circ \dots \circ f_1}(x) = J_{f_N}(f_{N-1}(\dots f_1(x))) \cdot J_{f_{N-1}}(f_{N-2}(\dots f_1(x))) \cdot \dots \cdot J_{f_2}(f_1(x)) \cdot J_{f_1}(x)$$

$J_f(x)$  is Jacobian of  $f$  computed at  $x$ .