Deep Learning

05 Backpropagation-1

Dr. Konda Reddy Mopuri
Dept. of AI, IIT Hyderabad
Jan-May 2025

Recap

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

Recap

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

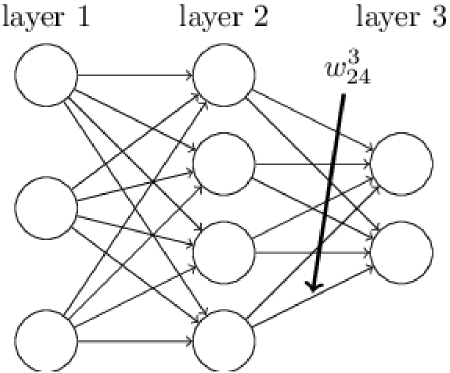
$$\mathbf{J} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla^{\mathrm{T}} f_1 \\ \vdots \\ \nabla^{\mathrm{T}} f_m \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

MLP: Some Notation

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



2.

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. σ 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

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

$\frac{\partial l_n}{\partial W_{jk}^{(l)}}$ and $\frac{\partial l_n}{\partial \mathbf{b}_j^{(l)}}$ 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, $x^{(0)} = x, f(x; W, \mathbf{b}) = x^{(L)}$

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

Chain rule of differential calculus

- Core concept of backpropagation

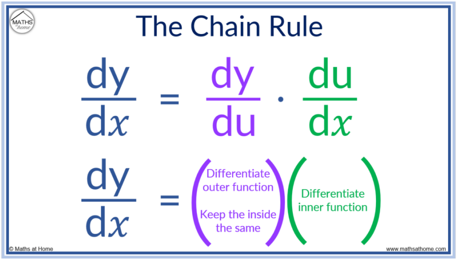
Chain rule of differential calculus

- Core concept of backpropagation
- $$(f \circ g)'(x) = f'(g(x)) \cdot g'(x)$$

Chain rule of differential calculus

- Core concept of backpropagation
- $$(f \circ g)'(x) = f'(g(x)) \cdot g'(x)$$
- $$\frac{\partial}{\partial x} f(g(x)) = \left.\frac{\partial f(a)}{\partial a}\right|_{a=g(x)} \cdot \frac{\partial g(x)}{\partial x}$$

Chain rule of differential calculus



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

1. $y = f(g_1(x), g_2(x), \ldots, g_M(x))$

Distributed Chain rule of differential calculus

1. $y = f(g_1(x), g_2(x), \ldots, g_M(x))$

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

Distributed Chain rule of differential calculus

1. $y = f(g_1(x), g_2(x), \ldots, g_M(x))$
2. $\frac{dy}{dx} = \frac{\partial f}{\partial g_1(x)} \frac{dg_1(x)}{dx} + \frac{\partial f}{\partial g_2(x)} \frac{dg_2(x)}{dx} + \ldots + \frac{\partial f}{\partial g_M(x)} \frac{dg_M(x)}{dx}$
3. Let $g_i(x) = z_i \rightarrow y = f(z_1, z_2, \ldots, z_M)$

Distributed Chain rule of differential calculus

1. $y = f(g_1(x), g_2(x), \ldots, g_M(x))$

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

3. Let $g_i(x) = z_i \rightarrow y = f(z_1, z_2, \ldots, z_M)$

4. $\Delta y = \frac{\partial f}{\partial z_1} \Delta z_1 + \frac{\partial f}{\partial z_2} \Delta z_2 + \ldots + \frac{\partial f}{\partial z_M} \Delta z_M$

Distributed Chain rule of differential calculus

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

1. $y = f(g_1(x), g_2(x), \ldots, g_M(x))$

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

3. Let $g_i(x) = z_i \rightarrow y = f(z_1, z_2, \ldots, z_M)$

4. $\Delta y = \frac{\partial f}{\partial z_1} \Delta z_1 + \frac{\partial f}{\partial z_2} \Delta z_2 + \ldots + \frac{\partial f}{\partial z_M} \Delta z_M$

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

Distributed Chain rule of differential calculus

1. $y = f(g_1(x), g_2(x), \ldots, g_M(x))$
2. $\frac{dy}{dx} = \frac{\partial f}{\partial g_1(x)} \frac{dg_1(x)}{dx} + \frac{\partial f}{\partial g_2(x)} \frac{dg_2(x)}{dx} + \ldots + \frac{\partial f}{\partial g_M(x)} \frac{dg_M(x)}{dx}$
3. Let $g_i(x) = z_i \rightarrow y = f(z_1, z_2, \ldots, z_M)$
4. $\Delta y = \frac{\partial f}{\partial z_1} \Delta z_1 + \frac{\partial f}{\partial z_2} \Delta z_2 + \ldots + \frac{\partial f}{\partial z_M} \Delta z_M$
5. $\Delta y = \frac{\partial f}{\partial z_1} \frac{dz_1}{dx} \Delta x + \frac{\partial f}{\partial z_2} \frac{dz_2}{dx} \Delta x + \ldots + \frac{\partial f}{\partial z_M} \frac{dz_M}{dx} \Delta x$
6. $\Delta y = \frac{\partial f}{\partial g_1(x)} \frac{dg_1(x)}{dx} \Delta x + \frac{\partial f}{\partial g_2(x)} \frac{dg_2(x)}{dx} \Delta x + \ldots + \frac{\partial f}{\partial g_M(x)} \frac{dg_M(x)}{dx} \Delta x$
7. $\Delta y = \left(\frac{\partial f}{\partial g_1(x)} \frac{dg_1(x)}{dx} + \frac{\partial f}{\partial g_2(x)} \frac{dg_2(x)}{dx} + \ldots + \frac{\partial f}{\partial g_M(x)} \frac{dg_M(x)}{dx} \right) \Delta x$

Chain rule of differential calculus

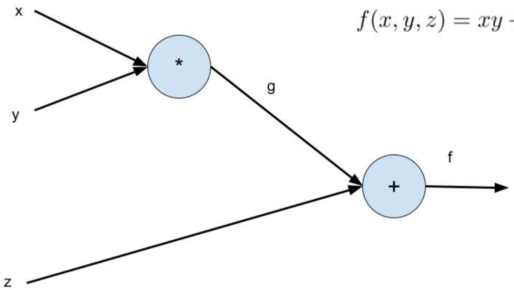
1. $f(x) = e^{sin(x^2)}$, let's find $\frac{\partial f}{\partial x}$

Chain rule of differential calculus

$$g = xy$$

$$f = g + z$$

$$f(x, y, z) = xy + z$$

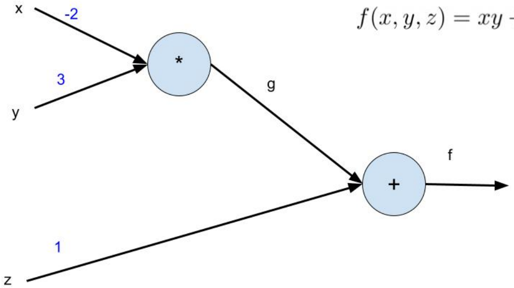


Chain rule of differential calculus

$$g = xy$$

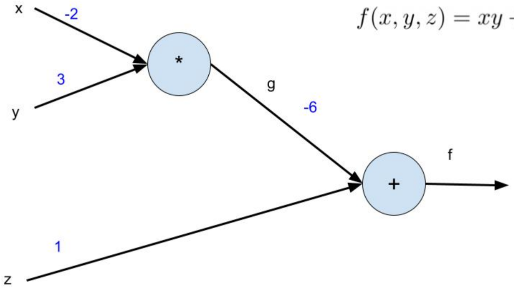
$$f = g + z$$

$$f(x, y, z) = xy + z$$



Chain rule of differential calculus

$$g = xy$$
$$f = g + z$$
$$f(x, y, z) = xy + z$$

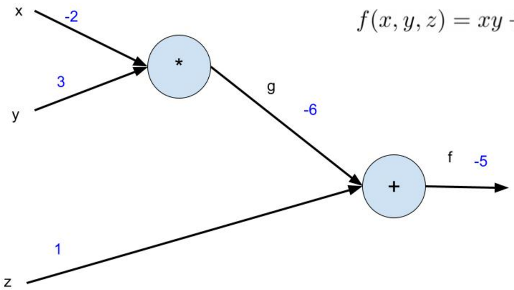


Chain rule of differential calculus

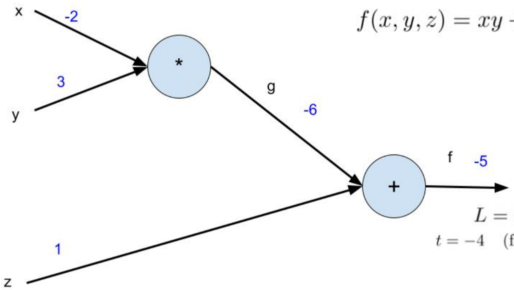
$$g = xy$$

$$f = g + z$$

$$f(x, y, z) = xy + z$$



Chain rule of differential calculus



$$g = xy$$

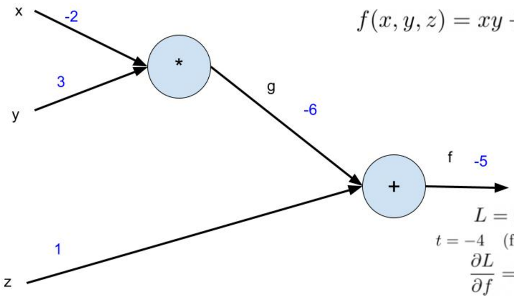
$$f = g + z$$

$$f(x, y, z) = xy + z$$

$$L = (f - t)^2$$

$t = -4$ (for this input)

Chain rule of differential calculus



$$g = xy$$

$$f = g + z$$

$$f(x, y, z) = xy + z$$

$$L = (f - t)^2$$

$t = -4$ (for this input)

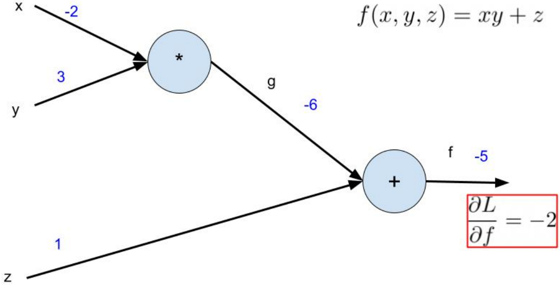
$$\frac{\partial L}{\partial f} = 2(f - t)$$

Chain rule of differential calculus

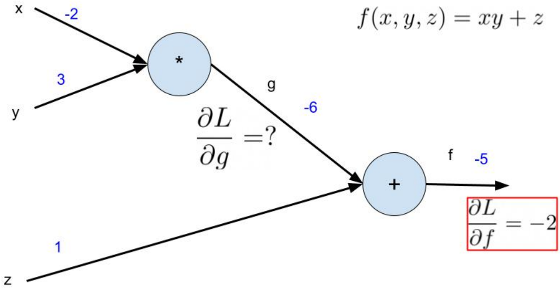
$$g = xy$$

$$f = g + z$$

$$f(x, y, z) = xy + z$$



Chain rule of differential calculus

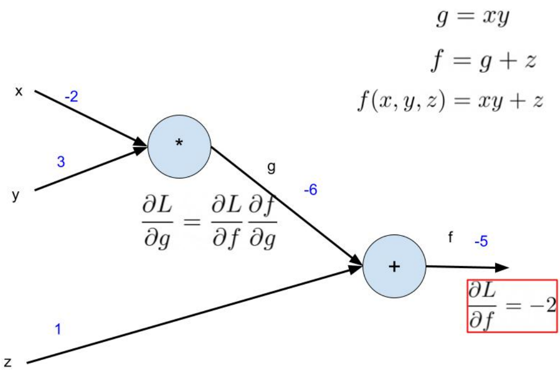


$$g = xy$$

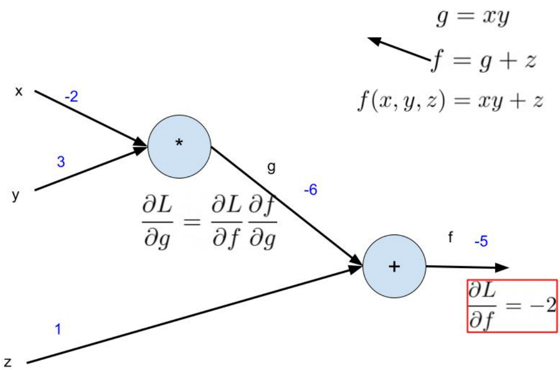
$$f = g + z$$

$$f(x, y, z) = xy + z$$

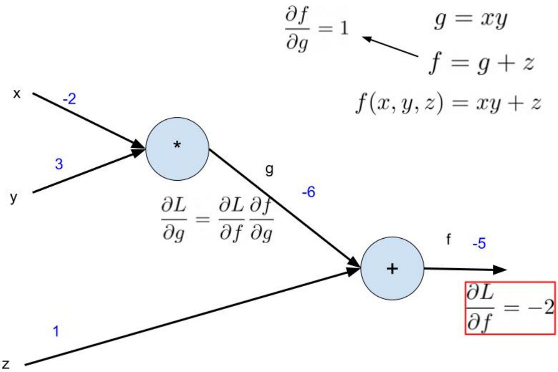
Chain rule of differential calculus



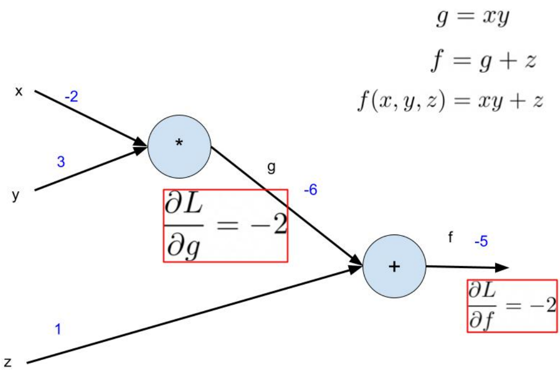
Chain rule of differential calculus



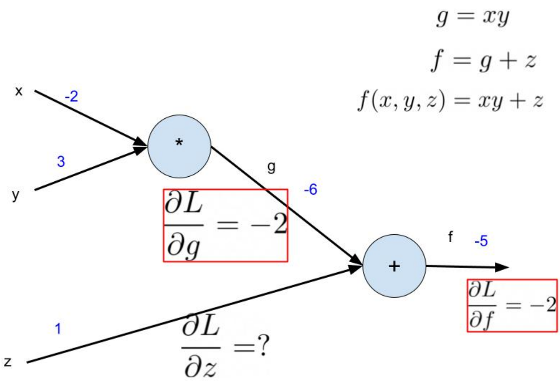
Chain rule of differential calculus



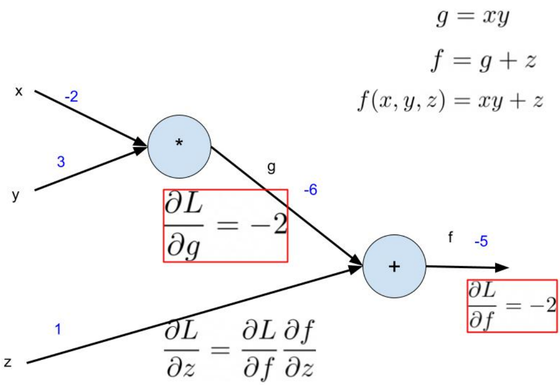
Chain rule of differential calculus



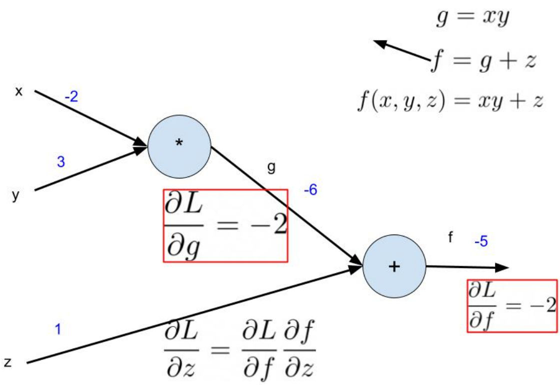
Chain rule of differential calculus



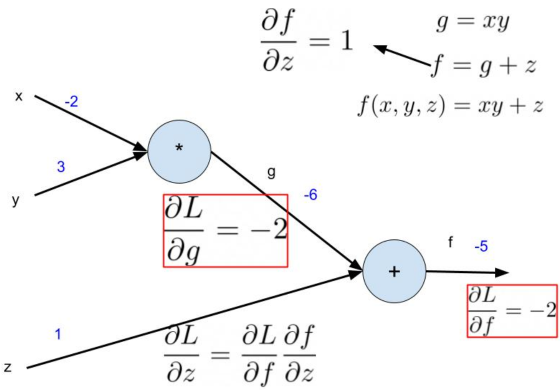
Chain rule of differential calculus



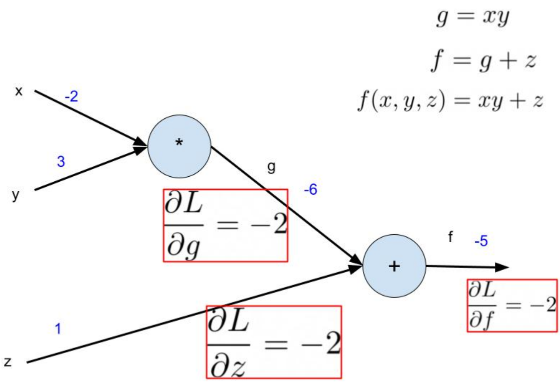
Chain rule of differential calculus



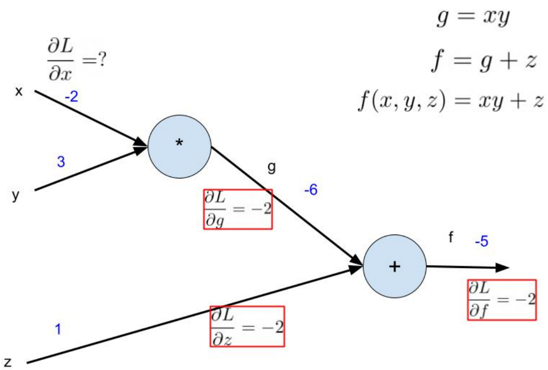
Chain rule of differential calculus



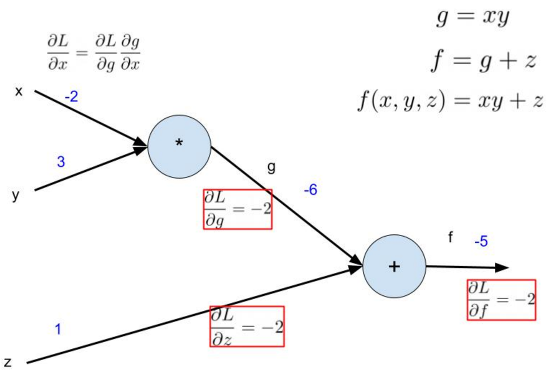
Chain rule of differential calculus



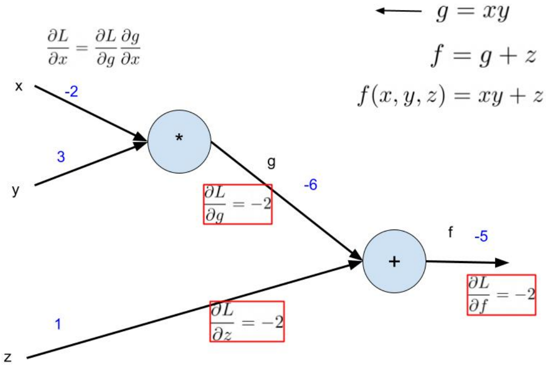
Chain rule of differential calculus



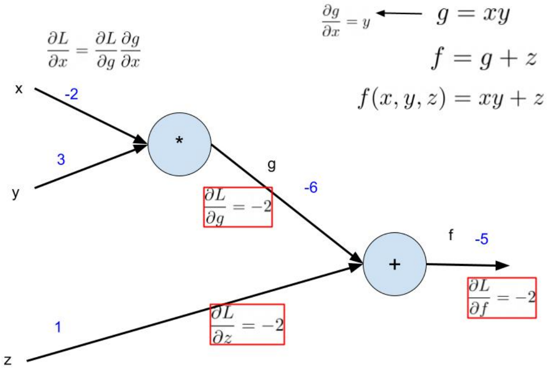
Chain rule of differential calculus



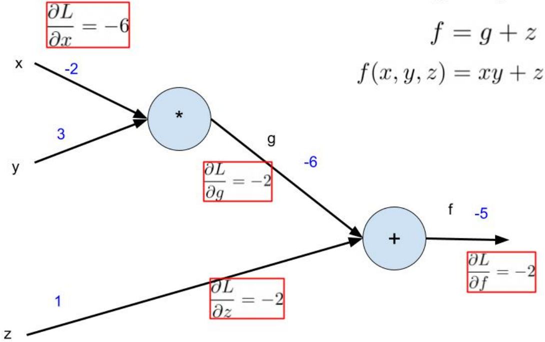
Chain rule of differential calculus



Chain rule of differential calculus



Chain rule of differential calculus



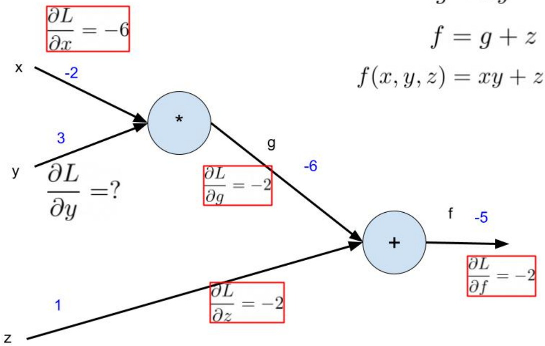
$$g = xy$$
$$f = g + z$$
$$f(x, y, z) = xy + z$$

Chain rule of differential calculus

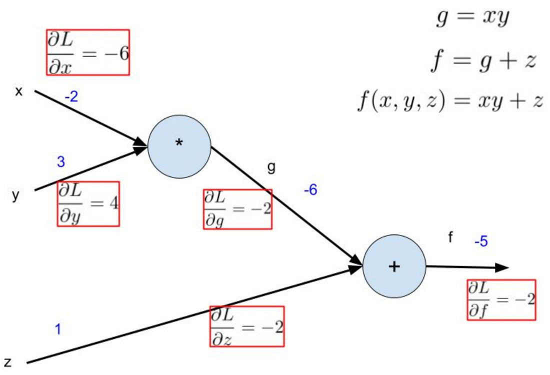
$$g = xy$$

$$f = g + z$$

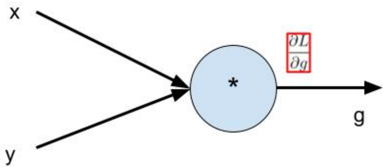
$$f(x, y, z) = xy + z$$



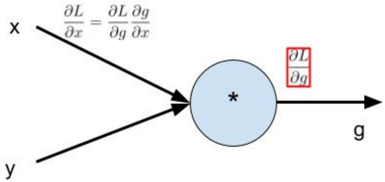
Chain rule of differential calculus



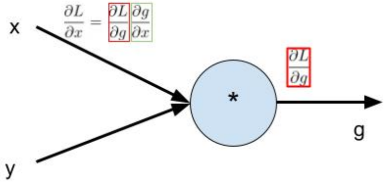
Gradient Flow



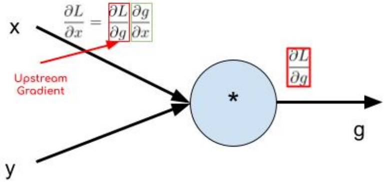
Gradient Flow



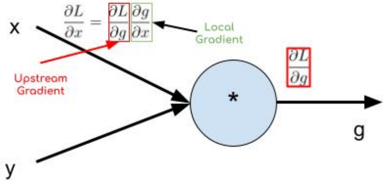
Gradient Flow



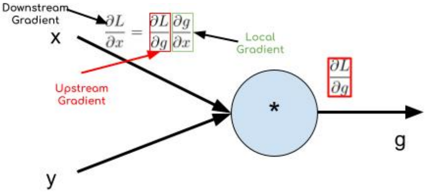
Gradient Flow



Gradient Flow



Gradient Flow



Chain rule of differential calculus for an MLP

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

  $J_{f(x)}$ is Jacobian of f computed at x.