Deep Learning

5 Backpropagation-1

Dr. Konda Reddy Mopuri
Dept. of Artificial Intelligence
IIT Hyderabad
Jan-May 2023

Recap

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

Recap

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

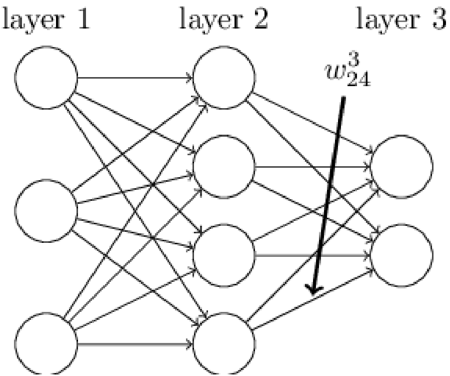
$$\mathbf{J} = \begin{bmatrix} \dfrac{\partial \mathbf{f}}{\partial x_1} & \cdots & \dfrac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla^{\mathrm{T}} f_1 \\ \vdots \\ \nabla^{\mathrm{T}} f_m \end{bmatrix} = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \cdots & \dfrac{\partial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

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

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^l_j 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^l_j is the weighted input to j^{th} neuron in l^{th} layer
2. $s^l_j = \sum_k w^l_{jk} x^{l-1}_k + b^l_j$
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, $x^{(0)} = x, f(x; W, \mathbf{b}) = x^{(L)}$

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

Chain rule of differential calculus

- Core concept of backpropagation

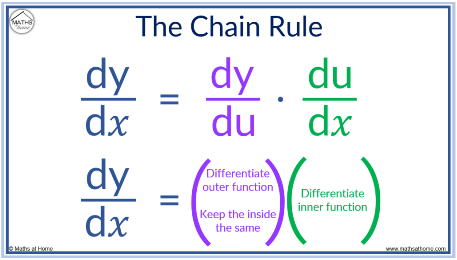
Chain rule of differential calculus

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

Chain rule of differential calculus

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

Chain rule of differential calculus

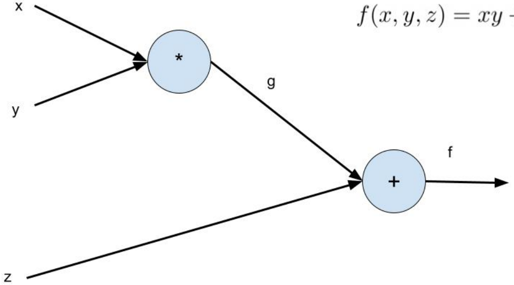


Chain rule of differential calculus

1. $f(x) = e^{sin(x^2)}$, let's find $\frac{\partial f}{\partial x}$ (work it out on the board)

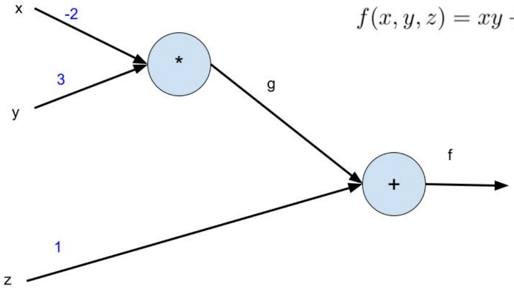
Chain rule of differential calculus

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



Chain rule of differential calculus

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

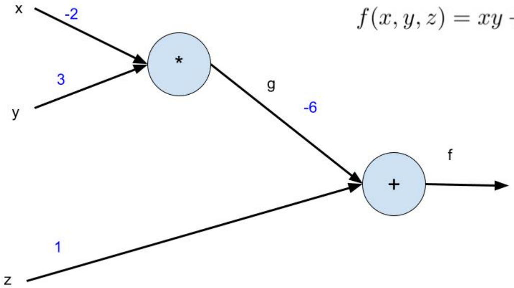


Chain rule of differential calculus

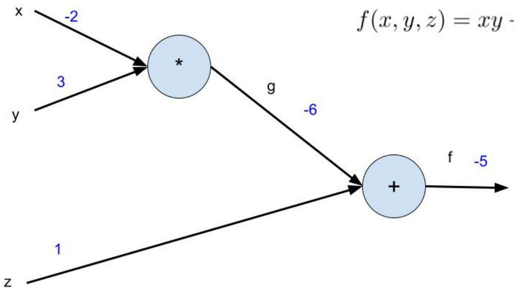
$$g = xy$$

$$f = g + z$$

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



Chain rule of differential calculus

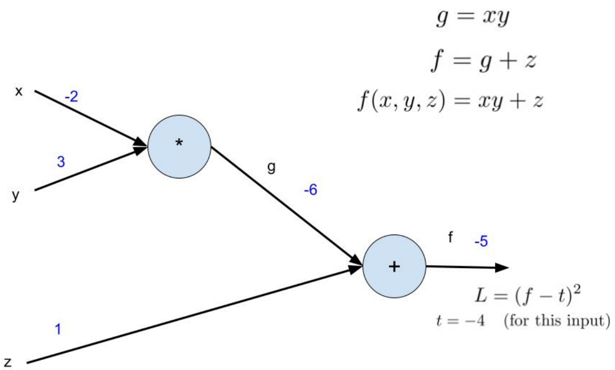


$$g = xy$$

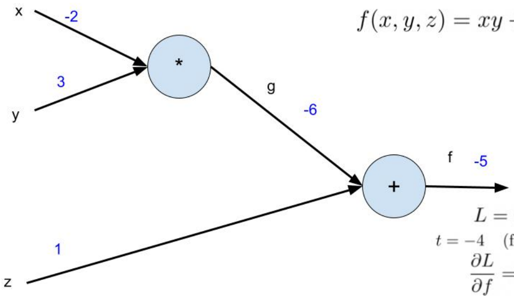
$$f = g + z$$

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

Chain rule of differential calculus



Chain rule of differential calculus



$$g = xy$$

$$f = g + z$$

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

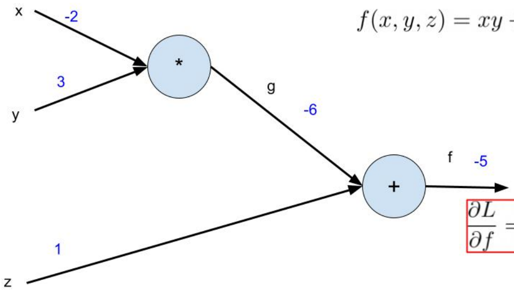
$$L = (f - t)^2$$

$t = -4$ (for this input)

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

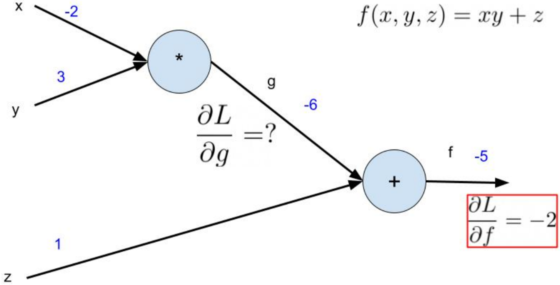
Chain rule of differential calculus

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



$$\boxed{\frac{\partial L}{\partial f} = -2}$$

Chain rule of differential calculus



$$g = xy$$

$$f = g + z$$

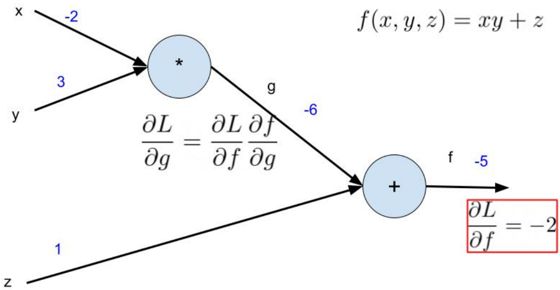
$$f(x, y, z) = xy + z$$

Chain rule of differential calculus

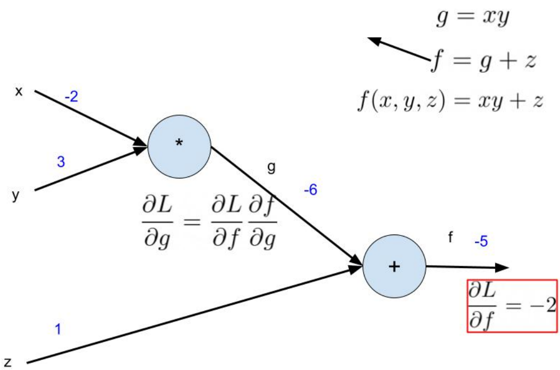
$$g = xy$$

$$f = g + z$$

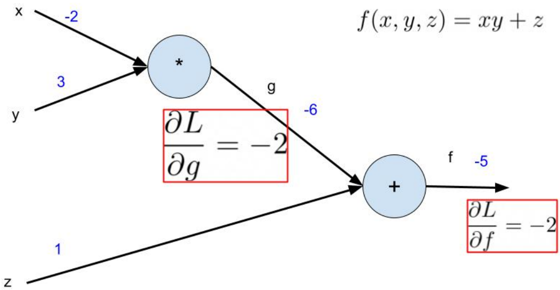
$$f(x, y, z) = xy + z$$



Chain rule of differential calculus



Chain rule of differential calculus

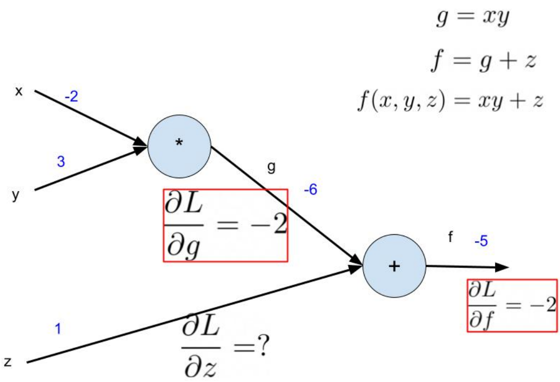


$$g = xy$$

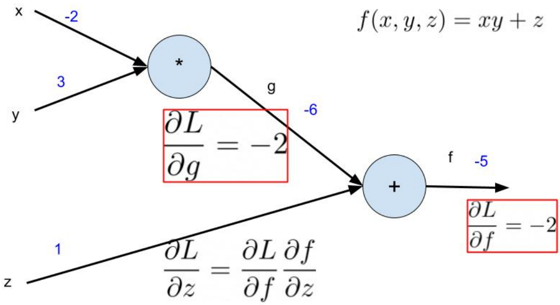
$$f = g + z$$

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

Chain rule of differential calculus



Chain rule of differential calculus



$$g = xy$$

$$f = g + z$$

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

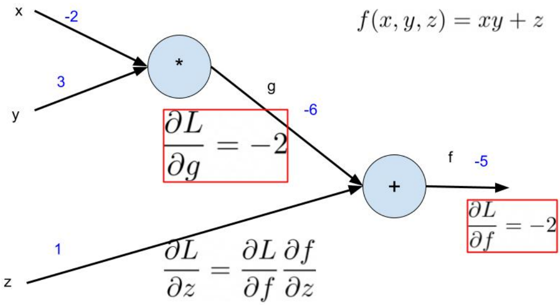
$$\boxed{\frac{\partial L}{\partial g} = -2}$$

$$\boxed{\frac{\partial L}{\partial f} = -2}$$

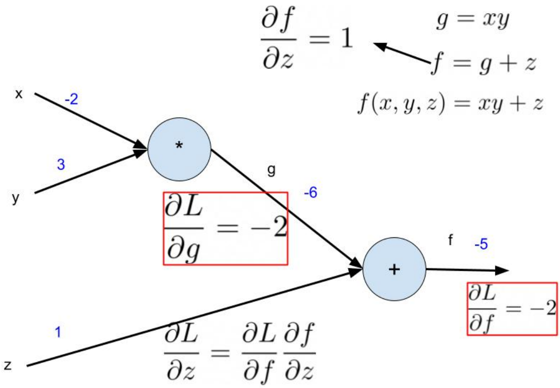
$$\frac{\partial L}{\partial z} = \frac{\partial L}{\partial f}\frac{\partial f}{\partial z}$$

Chain rule of differential calculus

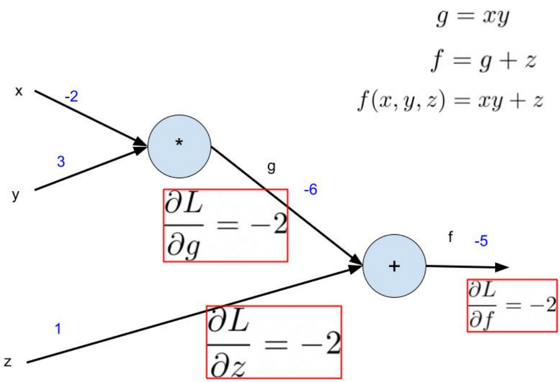
$$g = xy$$
$$f = g + z$$
$$f(x, y, z) = xy + z$$



Chain rule of differential calculus



Chain rule of differential calculus



Chain rule of differential calculus

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

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