Backpropagation
&
Neural Network

2023-2 KUBIG 방학세션
DL

Category

1. Backpropagation
2. Neural Network
3. Normalization
4. Q&A

0. Revised Syllabus

	Assignment	Paper Reading	Toy project
Linear Algebra	Numpy Problems	-	-
Loss & Optimization	Design Loss & Optimizer	Adam/ RMSprop	Pytorch loss&Optimizer module with basic tensor
Backpropagation	Implementation naïve affine layer & Backprop	Batchnorm, Layernorm	Pytorch Dataset
CNN	Naïve convolution layer & backprop	ResNet/GoogLeNet	Pytorch NN design
RNN/LSTM	Naïve Recurrent layer & backprop + LSTM	LSTM / GRU	Pytorch Trainer
Training Techniques	Batchnorm, Layernorm, Dropout, Activation Function	BERT / Attention is all you need	KUBIG CONTEST
Yolo & Transformer	Transfer Learning of Package		KUBIG CONTEST
Kubig Contest: Dacon Contest + Paper Implementation(Transformer)			

*Most of Assignment, paper and even lecture slides might be written in English

*Familiarity with English is somewhat important in this field (as I have heard it from many professors)

*This syllabus could be revised, considering the level of club attendees
(Harder or Easier ?)

Q&A
*Paper Review 발표는 자유 진행
(1명 또는 여러명 발표 가능)
부담없이 진행해주세요 ☺

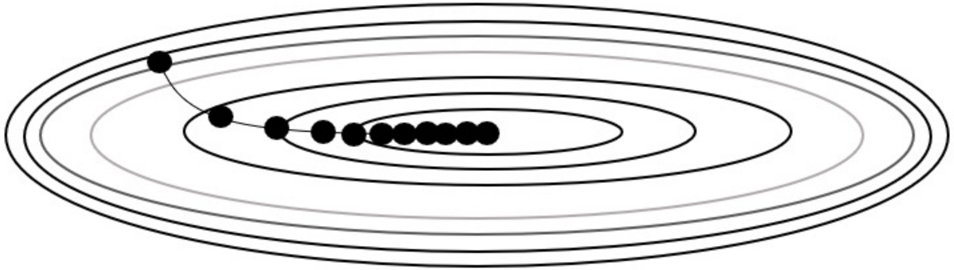
*티스토리는 팀 당 1편만 업로드 가능(하지만 웬만하면 모두 다 해보시는 것을 추천드립니다 ㅠ)

*Pytorch Toy Project도 github에 제출 바랍니다

0. Paper Review

0. Paper Review

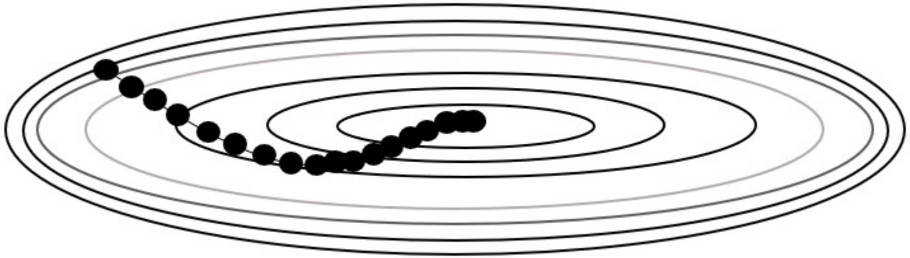
1) RMSprop



$$h_i = ph_{i-1} + (1-p)\frac{\partial L_i}{\partial W} \odot \frac{\partial L_i}{\partial W}$$

$$W = W - \eta \frac{1}{\sqrt{h}} \frac{\partial L}{\partial W}$$

2) Adam



1. Backpropagation

1. Backpropagation

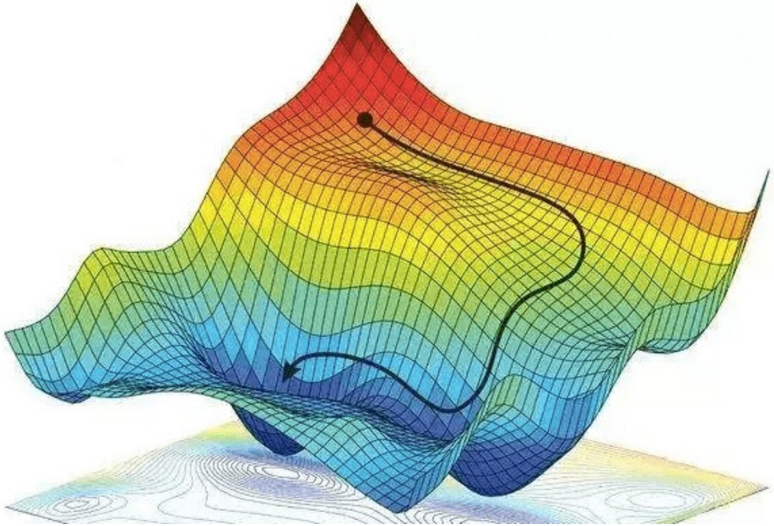
Where we are,,

$$L = \frac{1}{N} \sum_{i=1}^{N} L_i + \sum_k W_k^2$$

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

$$s = f(x; W) = Wx$$

want $\nabla_W L$



1. Backpropagation

Where we are,,

$$\frac{df(x)}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

Use Chain Rule & Analytic Gradient to update

Backpropagation

1. Backpropagation

Where we are,,

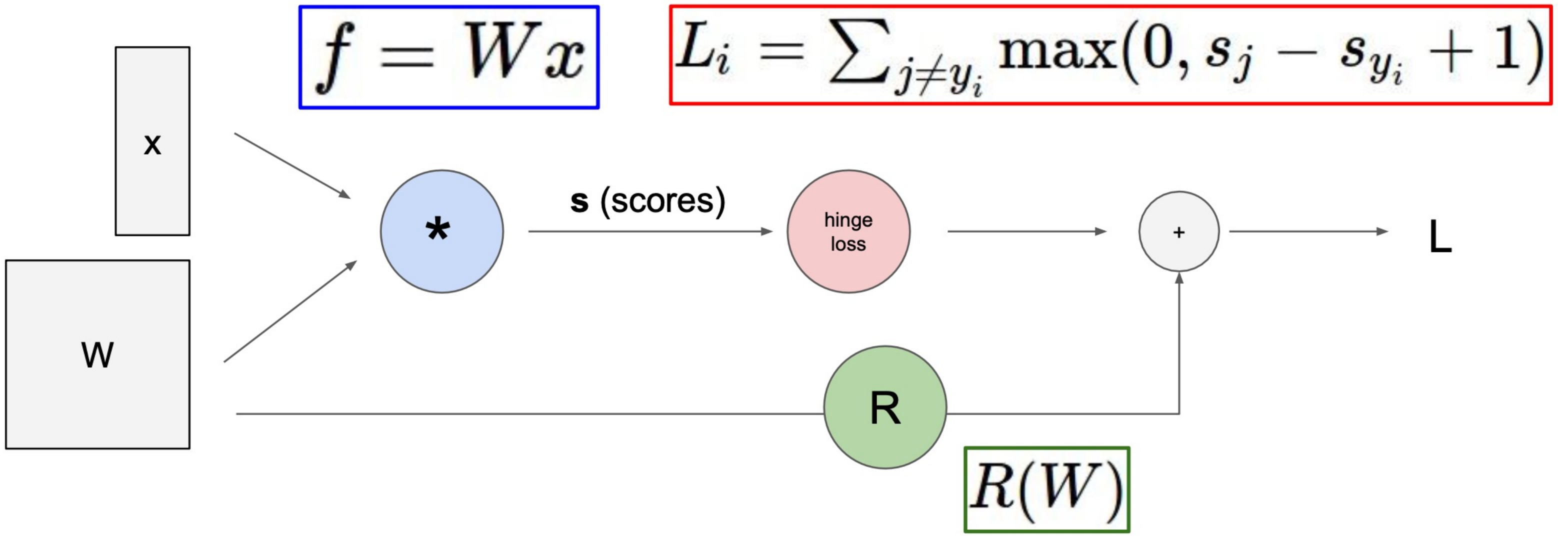
$$\frac{df(x)}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

Use Chain Rule & Analytic Gradient to update

Backpropagation

1. Backpropagation

Computational Graph

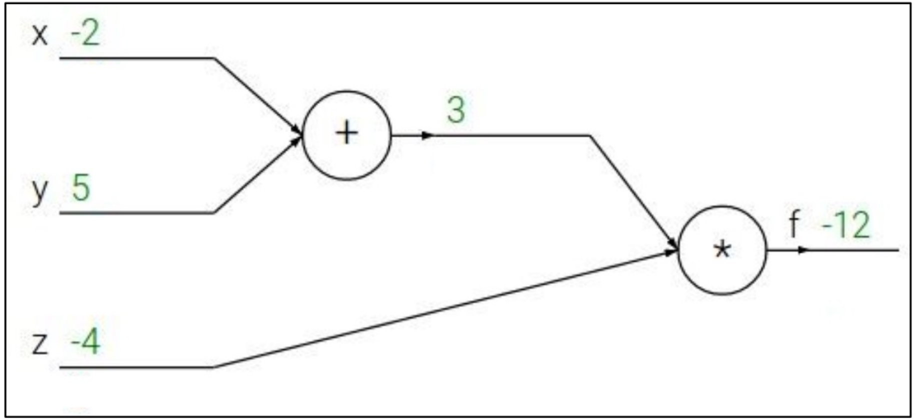


1. Backpropagation

Simple Example

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

x=-2
y=5
z=-4
x+y=3
(x+y)*z=-12



1. Backpropagation

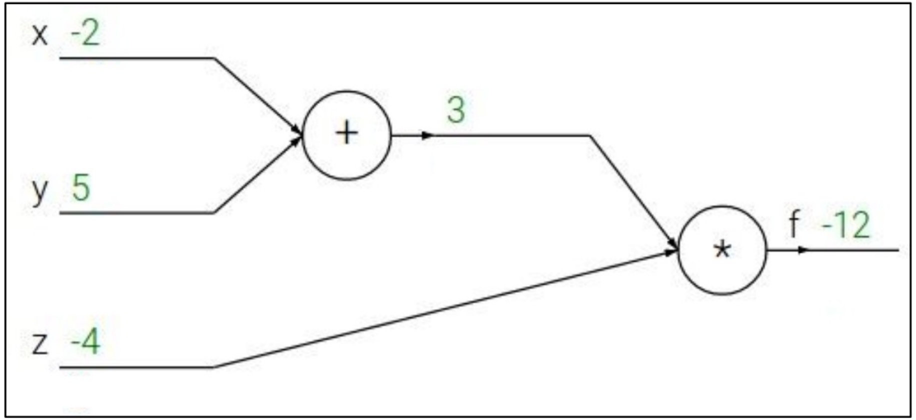
Simple Example

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

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

How to obtain $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



1. Backpropagation

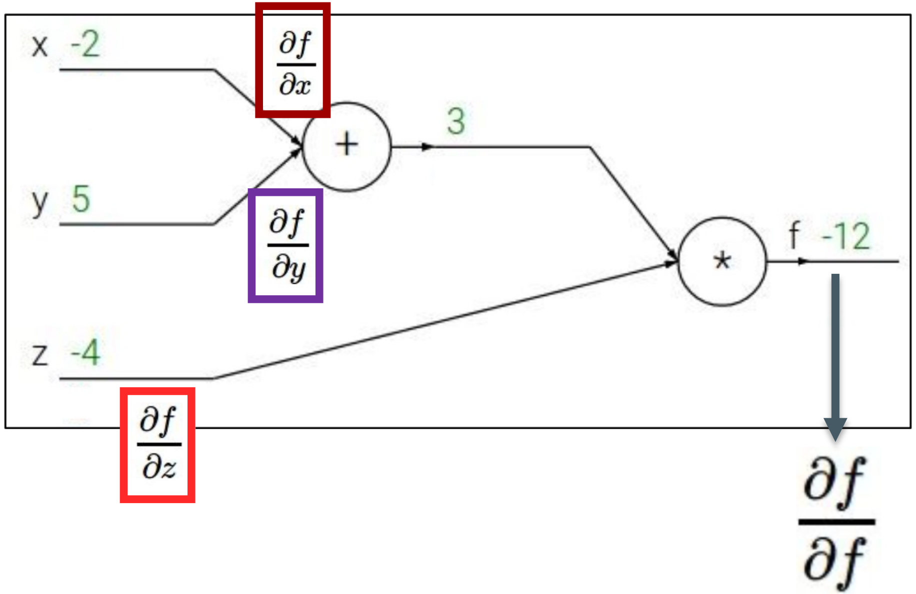
Simple Example

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

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

How to obtain $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



1. Backpropagation

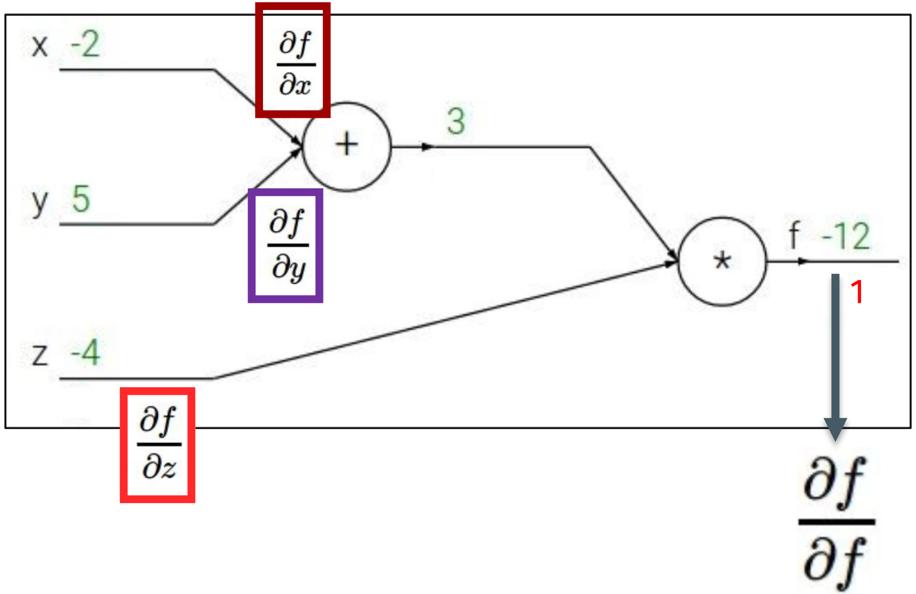
Simple Example

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

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

How to obtain $\dfrac{\partial f}{\partial x}, \dfrac{\partial f}{\partial y}, \dfrac{\partial f}{\partial z}$



1. Backpropagation

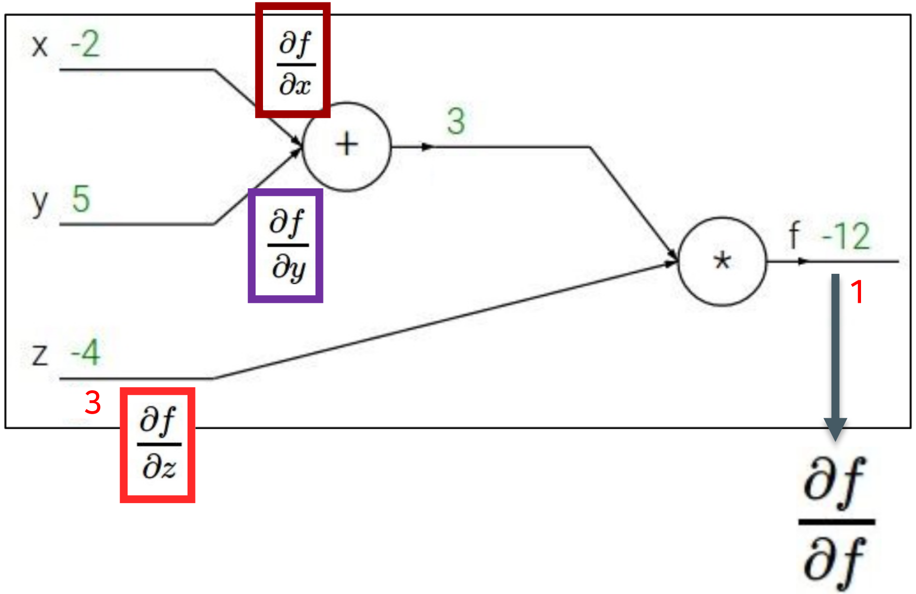
Simple Example

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

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

How to obtain $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



1. Backpropagation

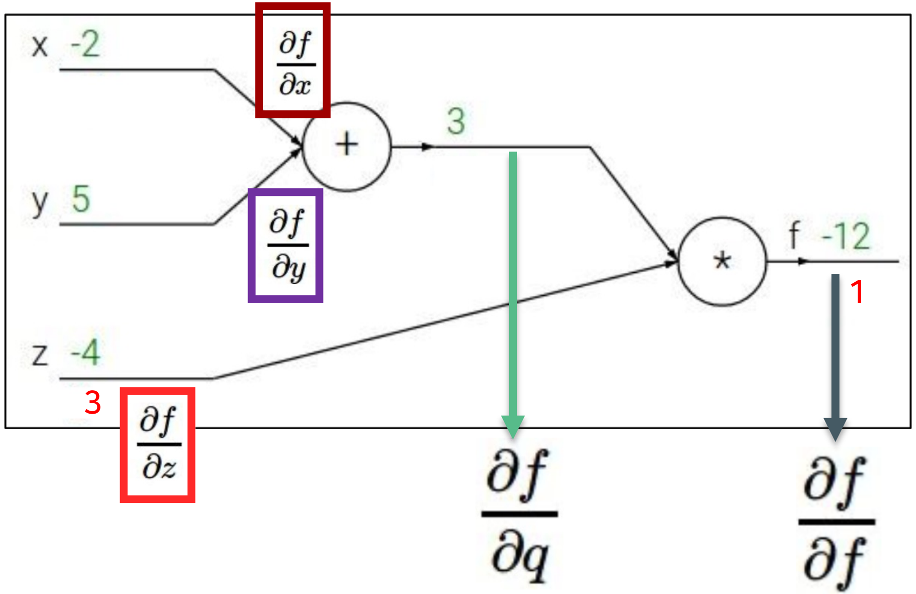
Simple Example

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

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

How to obtain $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



1. Backpropagation

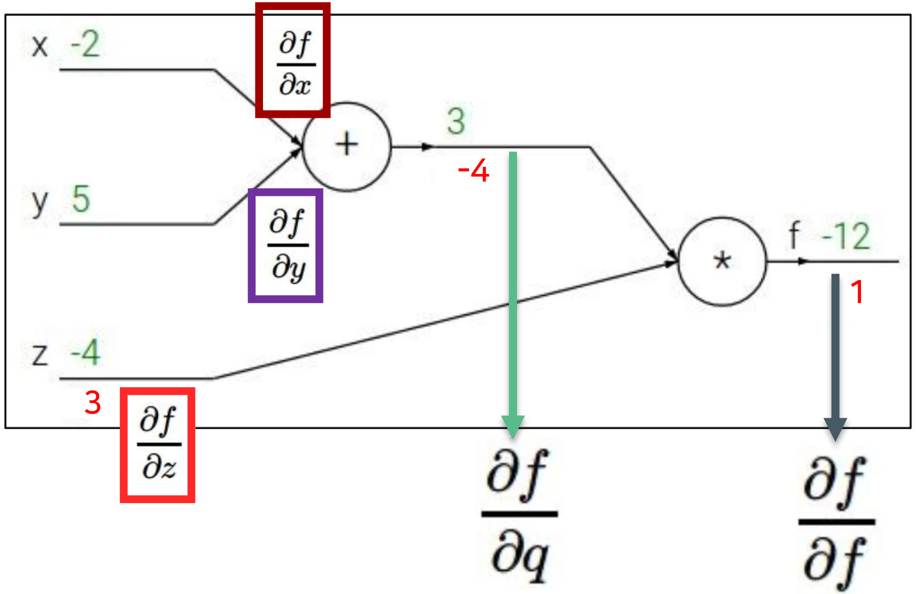
Simple Example

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

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

How to obtain $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



1. Backpropagation

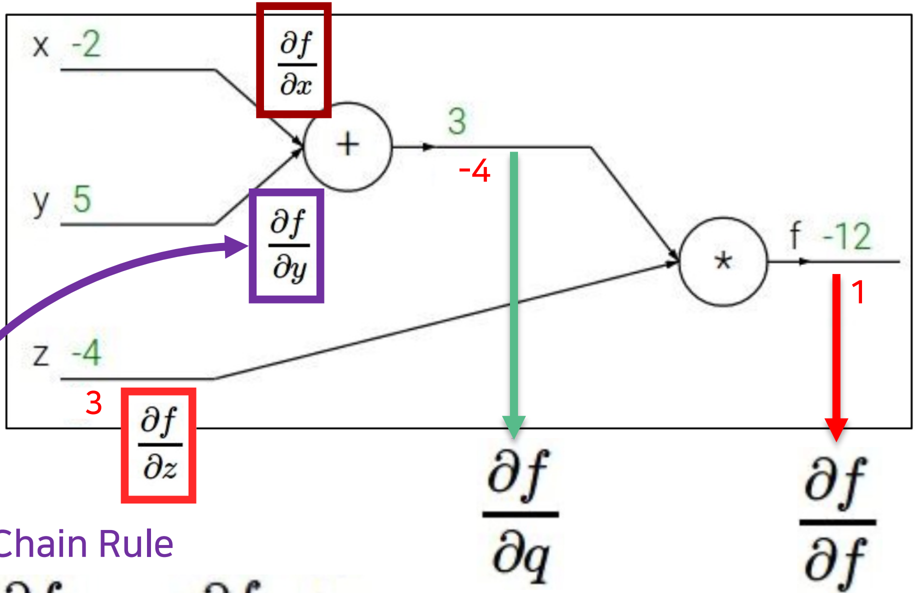
Simple Example

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

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

How to obtain $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



Chain Rule

$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial q} \frac{\partial q}{\partial y}$$

1. Backpropagation

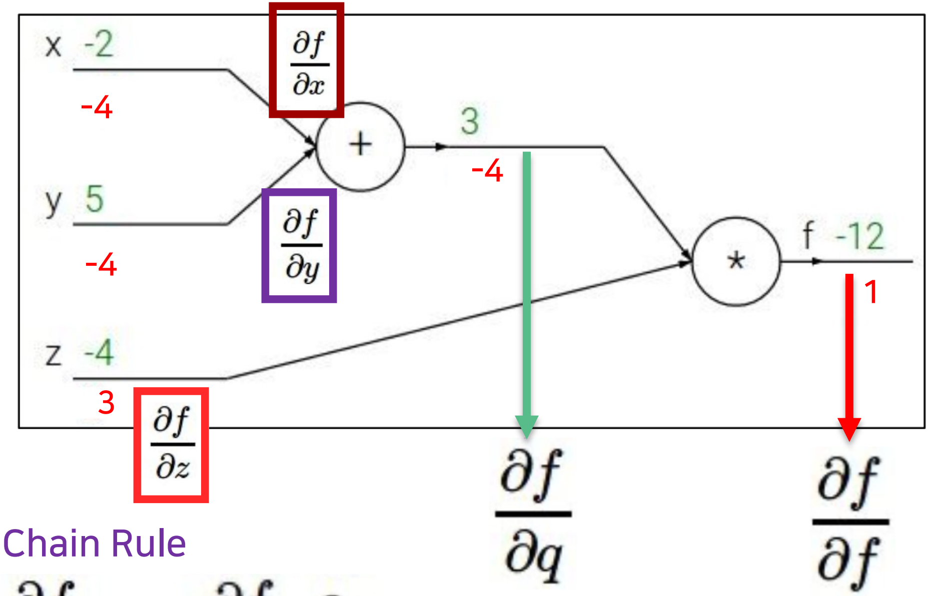
Simple Example

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

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

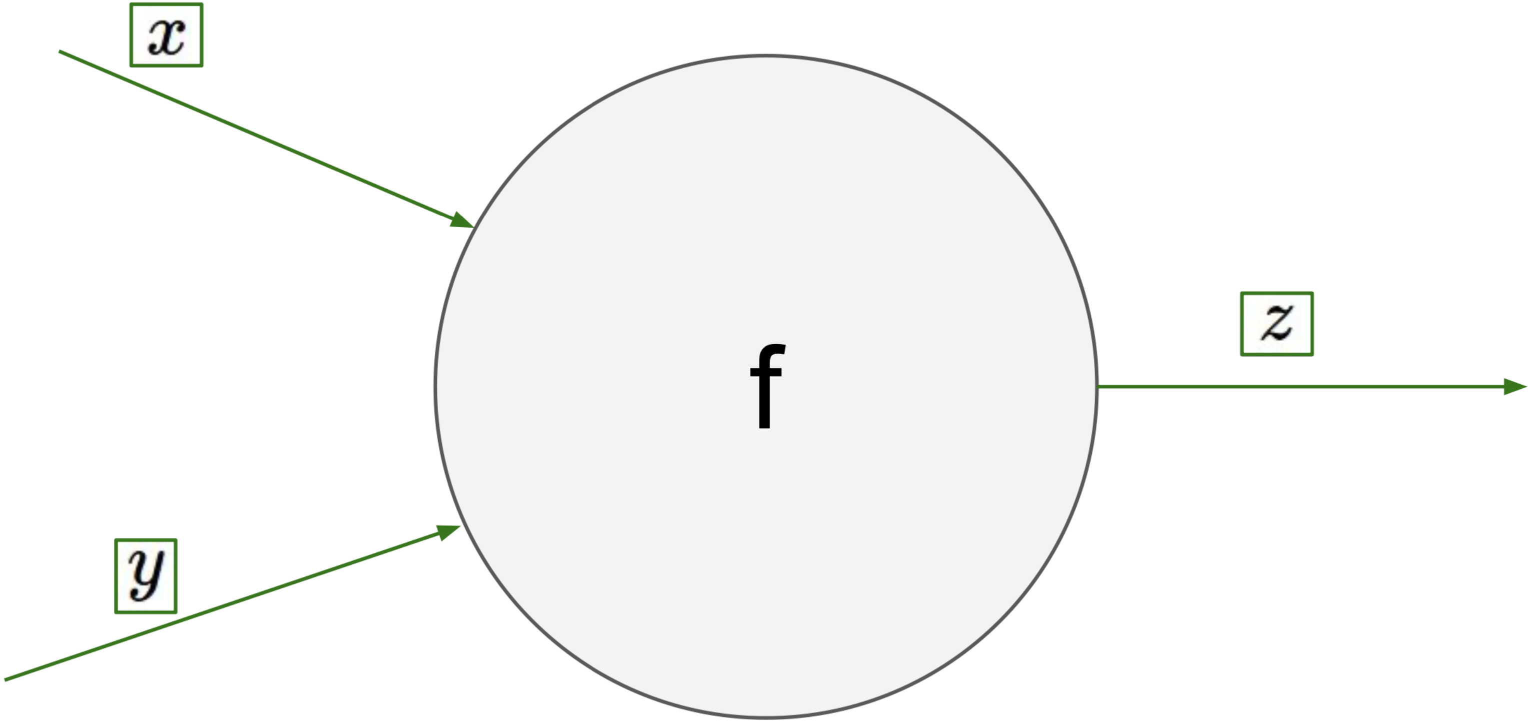
How to obtain $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



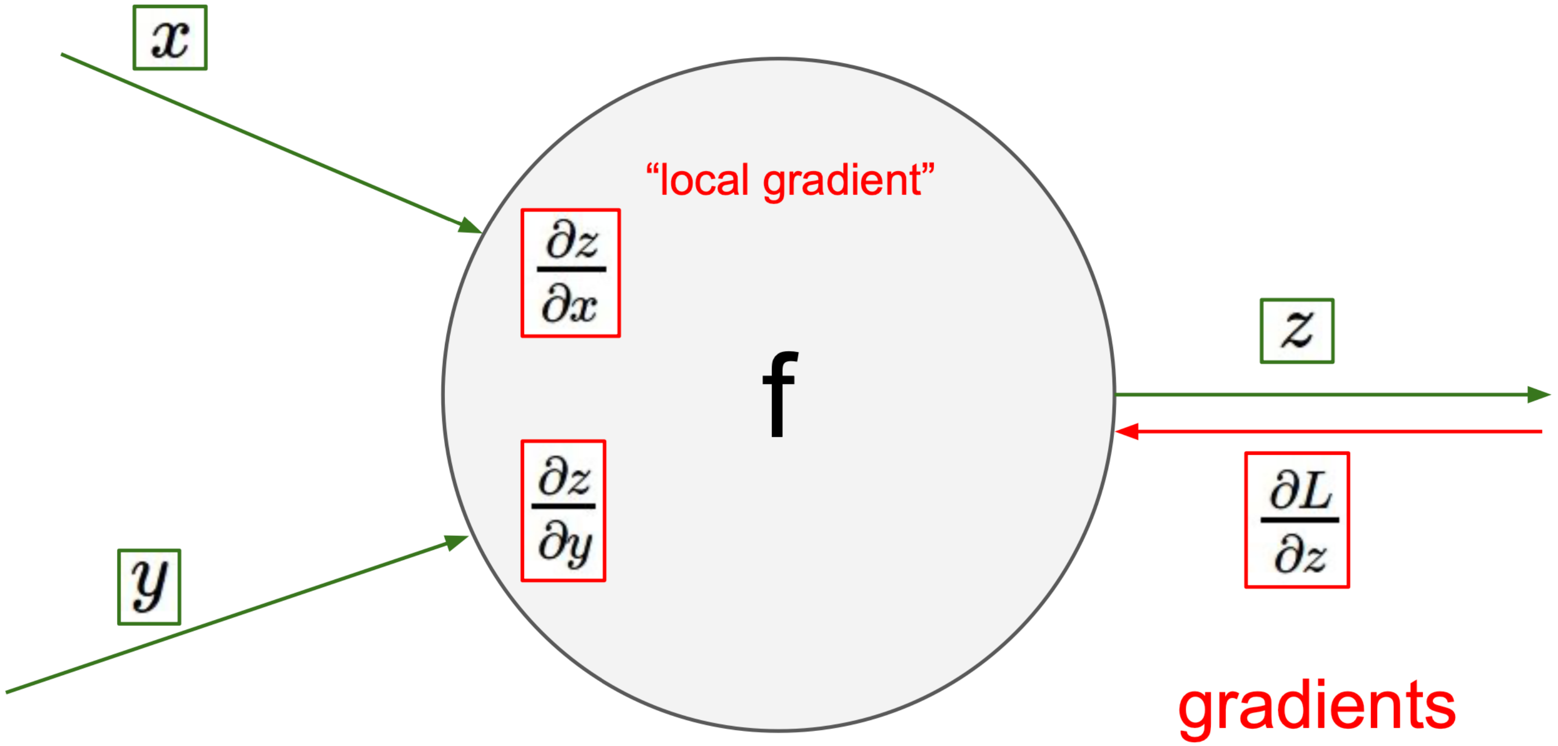
Chain Rule

$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial q} \frac{\partial q}{\partial y}$$

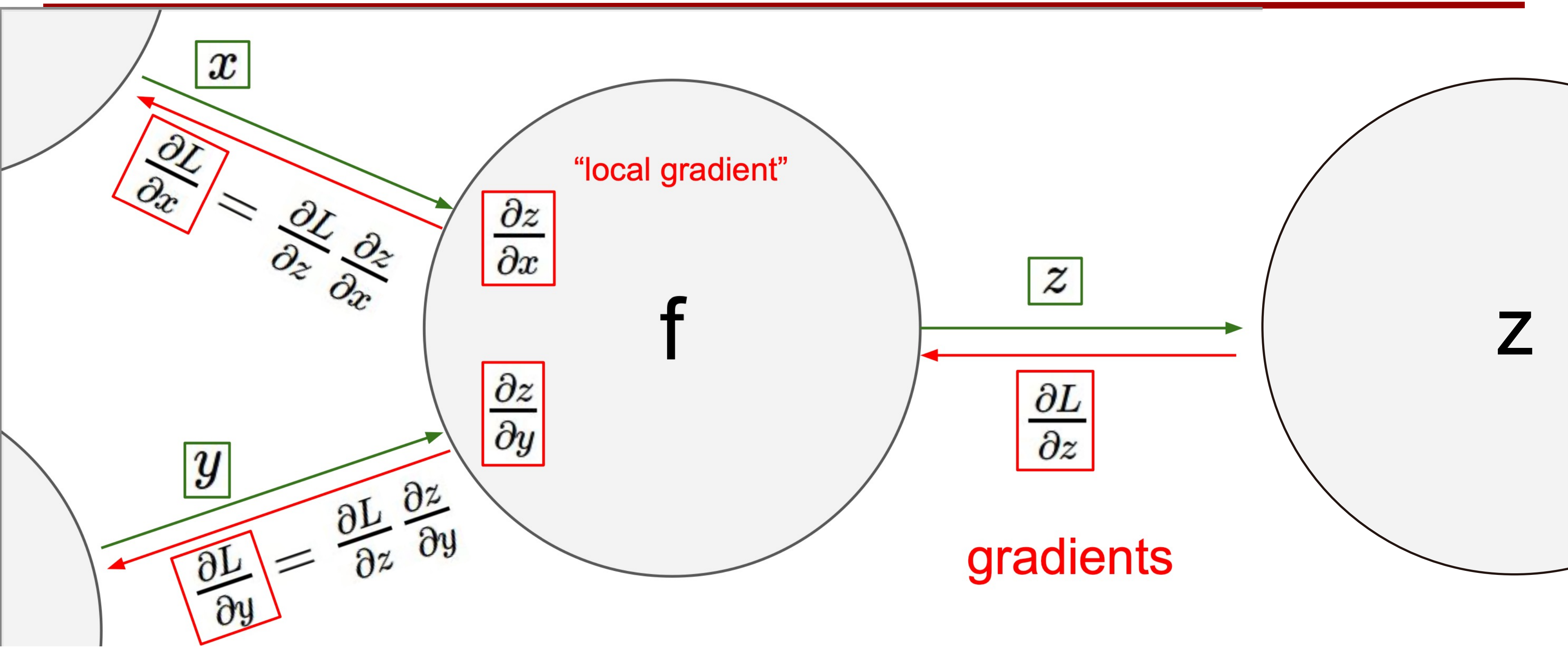
1. Backpropagation



1. Backpropagation



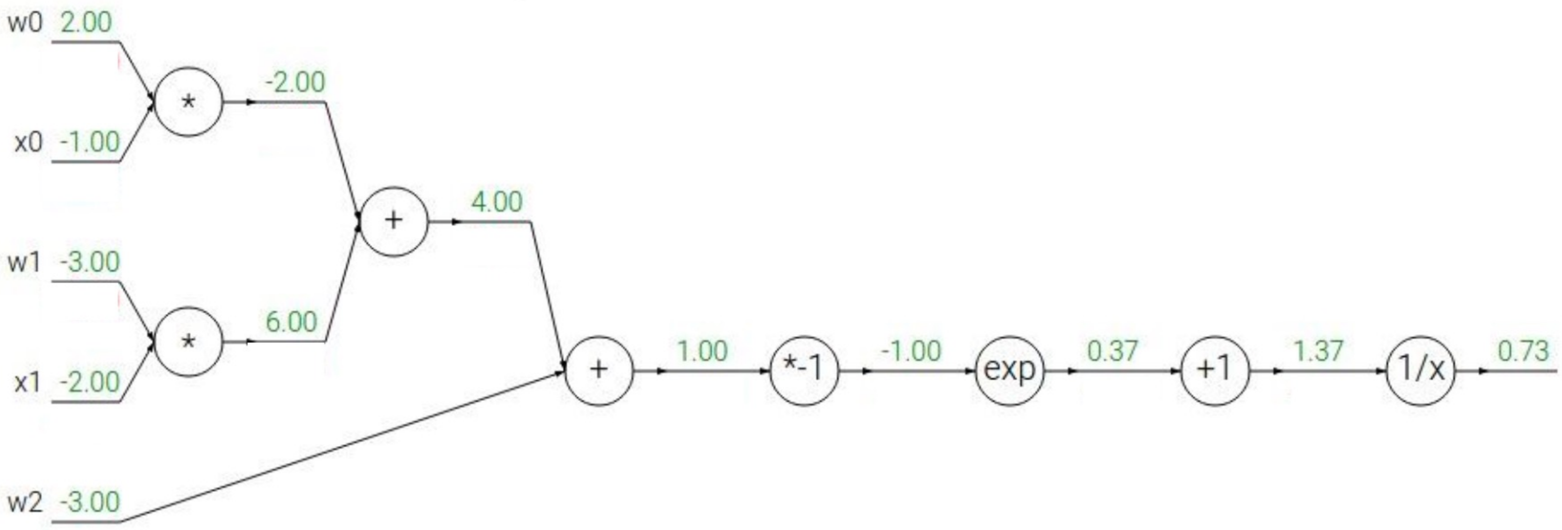
1. Backpropagation



1. Backpropagation

Second Example

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$



1. Backpropagation

Second Example

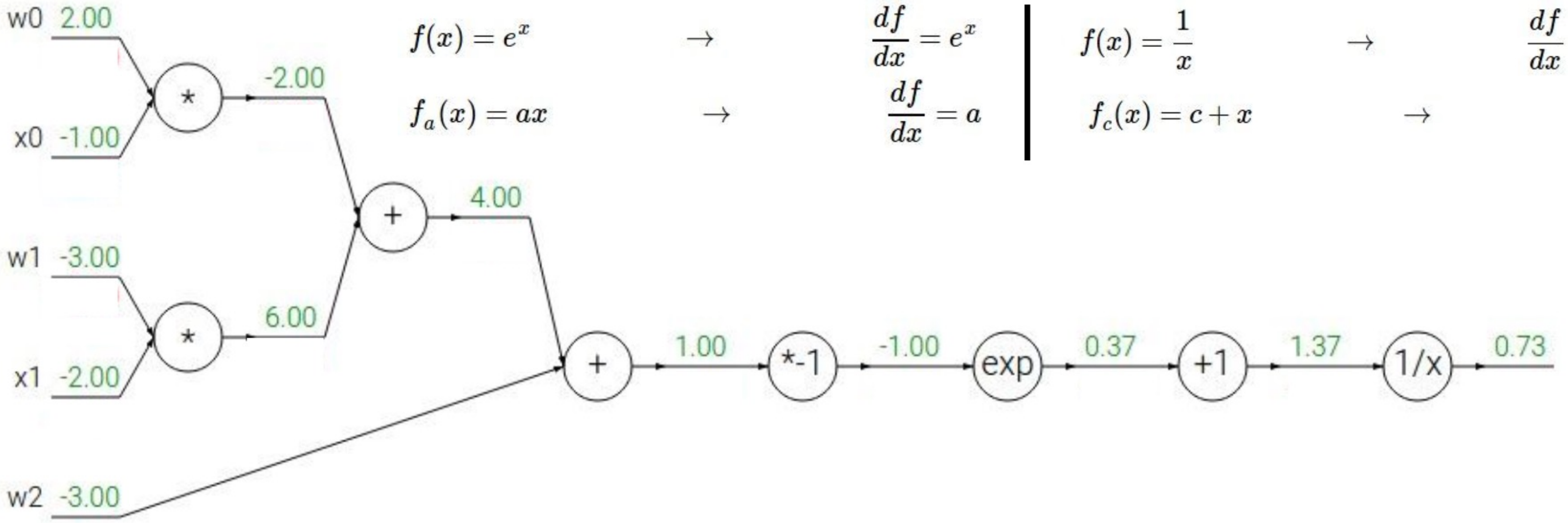
$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$

$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$

$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$

$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$



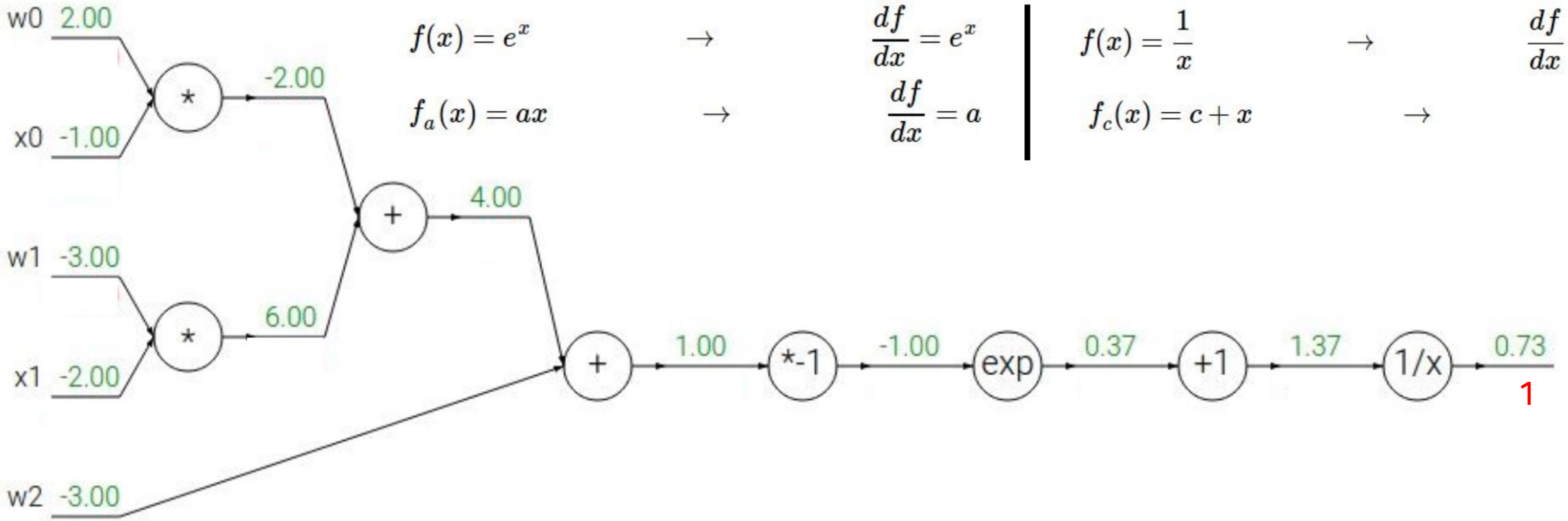
1. Backpropagation

Second Example

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

$$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x \qquad\qquad f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$$

$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a \qquad\qquad f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$$



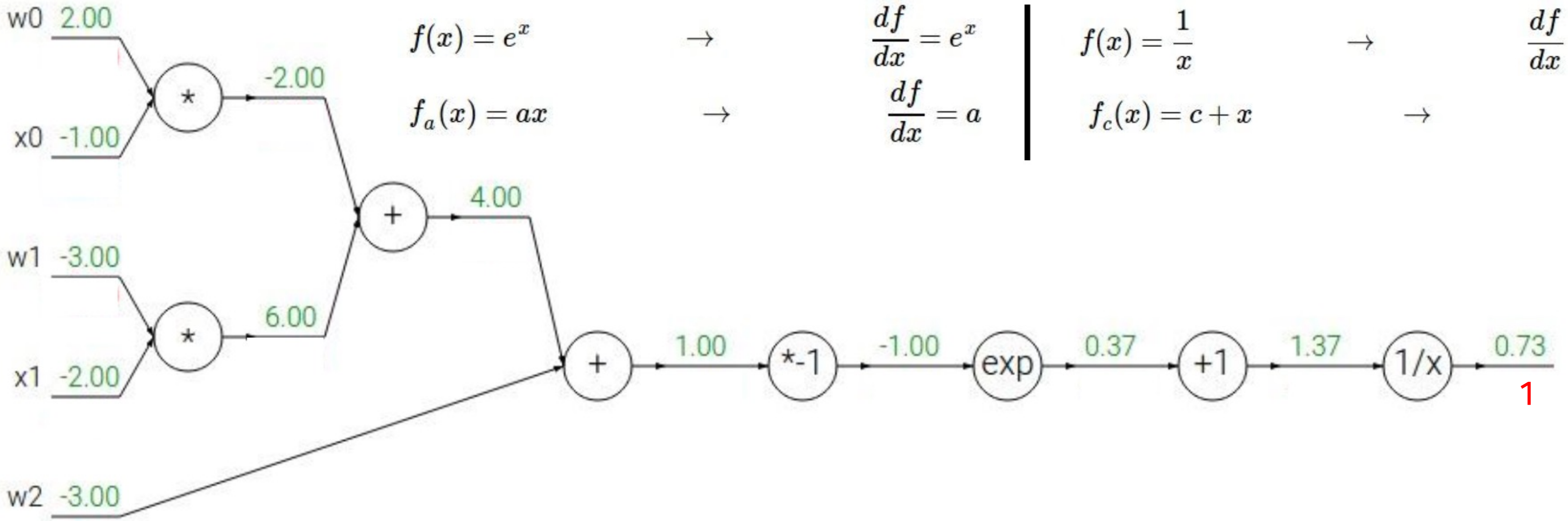
1. Backpropagation

Second Example

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

$$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x \qquad\Big|\qquad f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$$

$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a \qquad\Big|\qquad f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$$



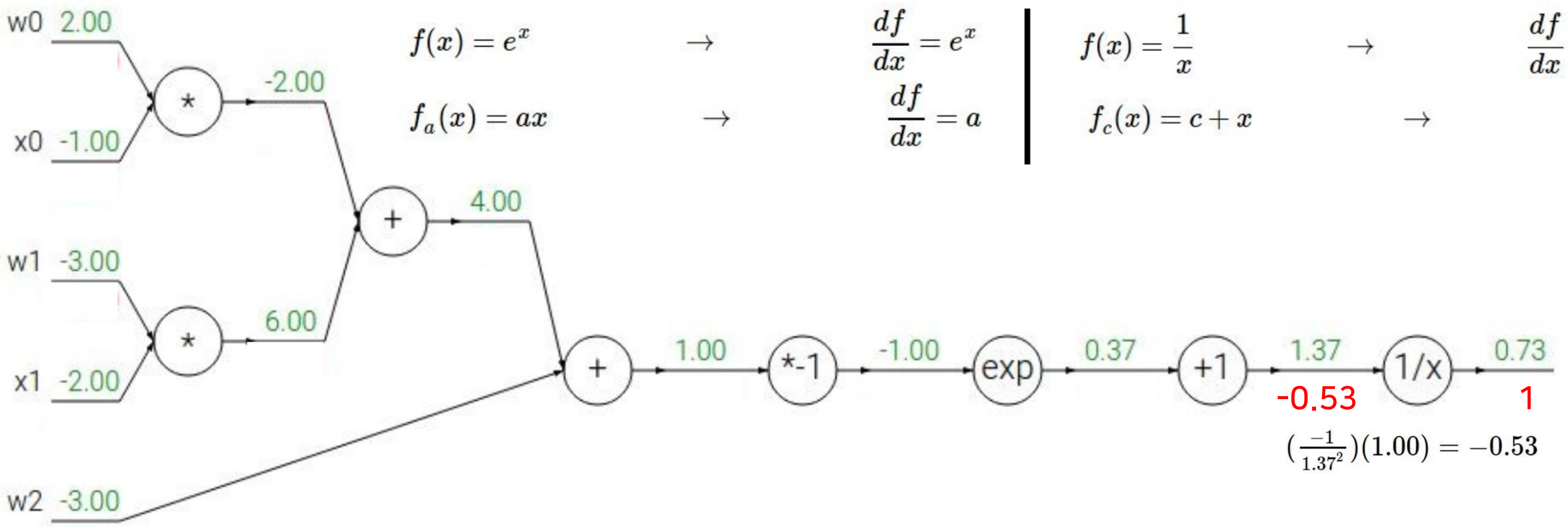
1. Backpropagation

Second Example

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

$$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x \qquad\Big|\qquad f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$$

$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a \qquad\Big|\qquad f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$$



$$\left(\frac{-1}{1.37^2}\right)(1.00) = -0.53$$

1. Backpropagation

Second Example

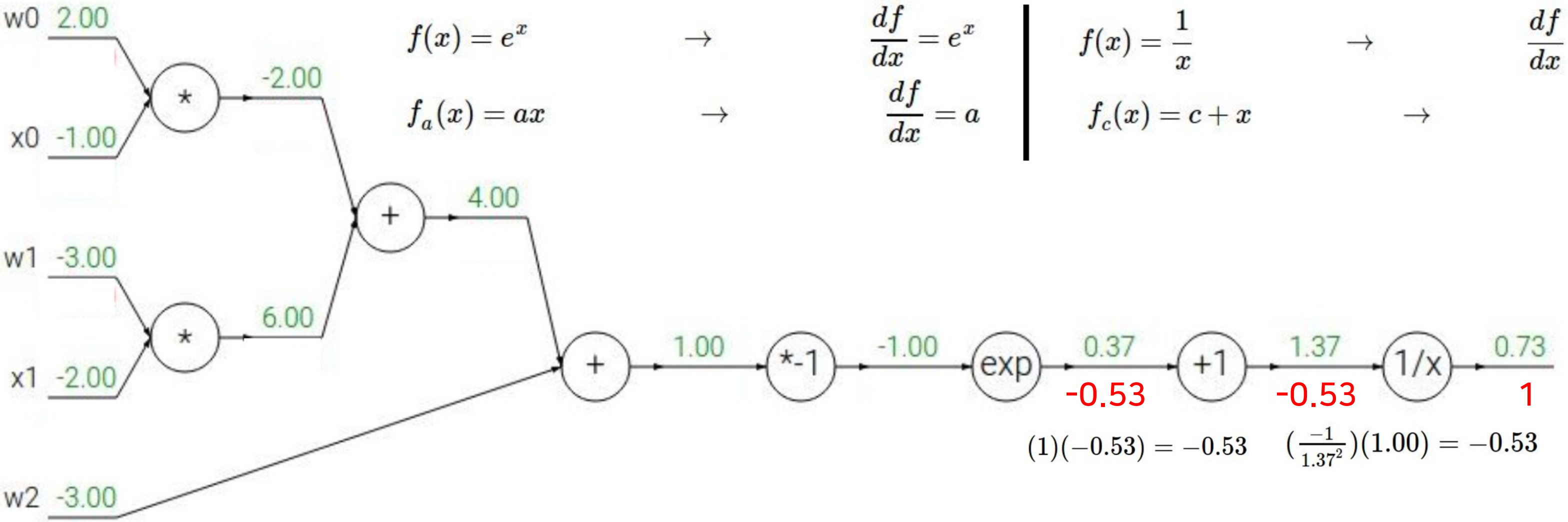
$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$

$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$

$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$

$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$



$(1)(-0.53) = -0.53$ $\quad (\frac{-1}{1.37^2})(1.00) = -0.53$

1. Backpropagation

Second Example

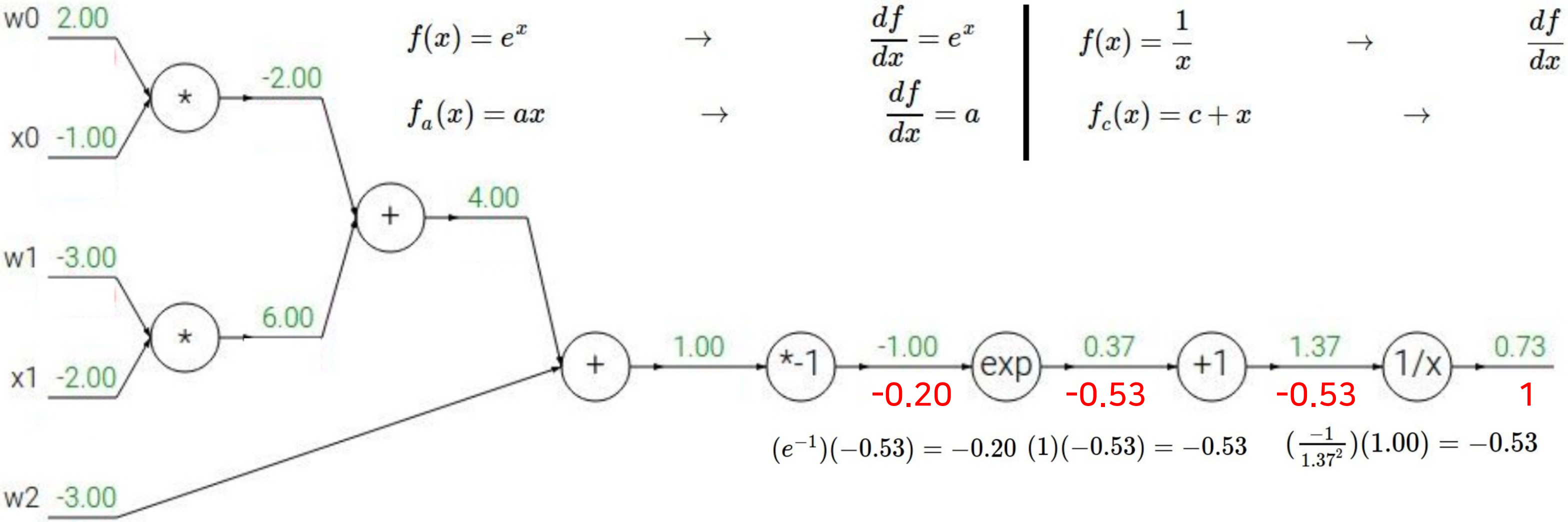
$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$

$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$

$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$

$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$



$(e^{-1})(-0.53) = -0.20$ $(1)(-0.53) = -0.53$ $(\frac{-1}{1.37^2})(1.00) = -0.53$

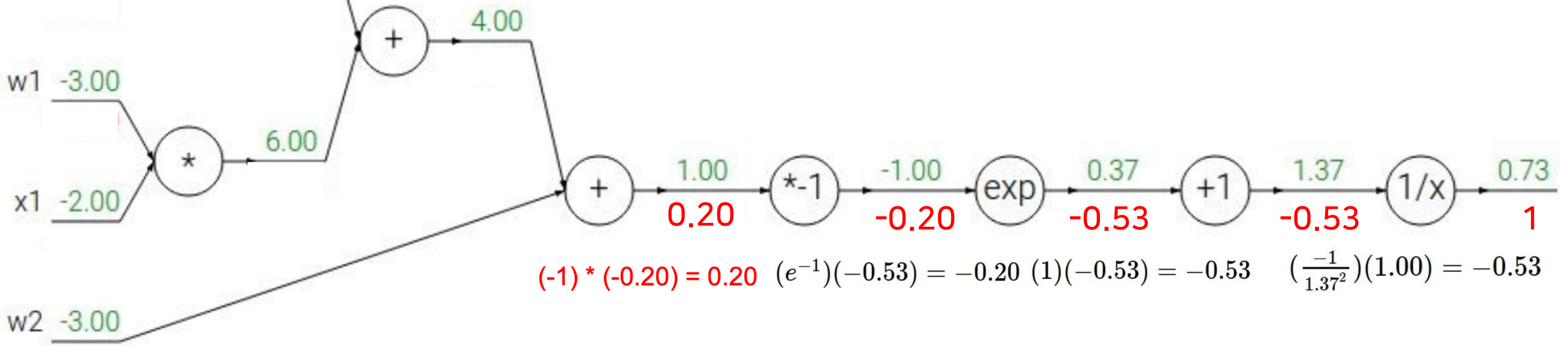
1. Backpropagation

Second Example

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$ | $f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$

$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$ | $f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$



1. Backpropagation

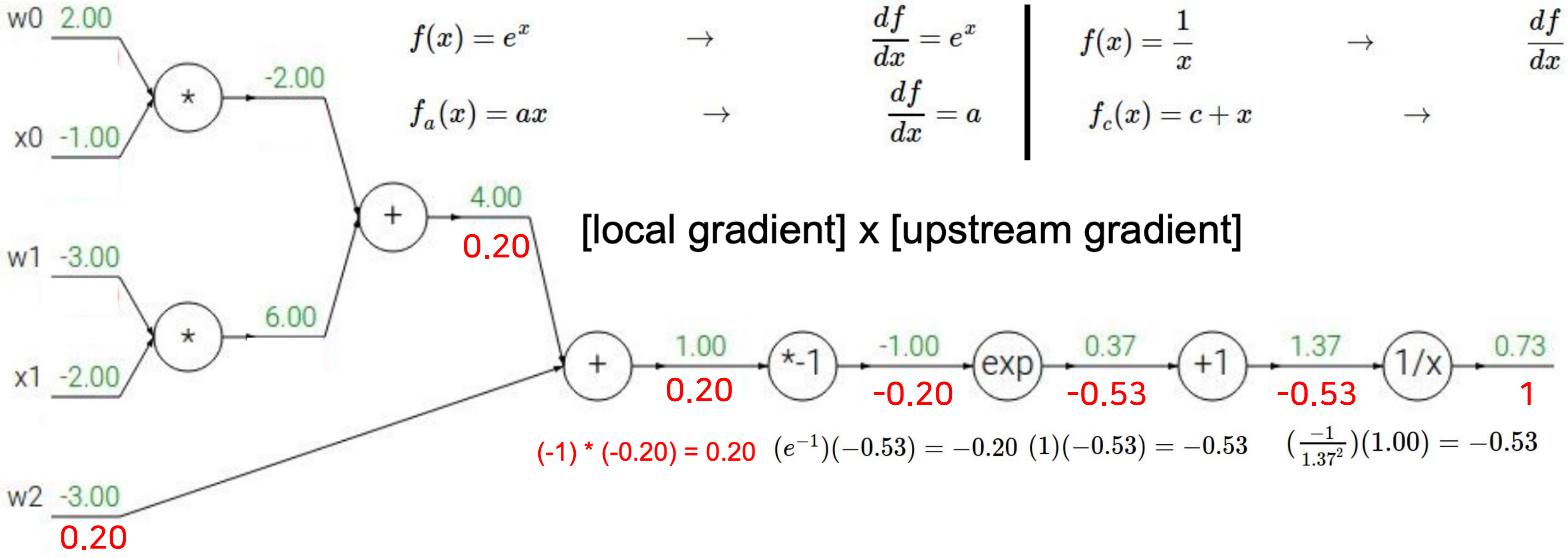
Second Example

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$ | $f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$

$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$ | $f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$

[local gradient] x [upstream gradient]



(-1) * (-0.20) = 0.20 $\quad (e^{-1})(-0.53) = -0.20 \quad (1)(-0.53) = -0.53 \quad (\frac{-1}{1.37^2})(1.00) = -0.53$

1. Backpropagation

Second Example

x0: [2] x [0.2] = 0.4
w0: [-1] x [0.2] = -0.2

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

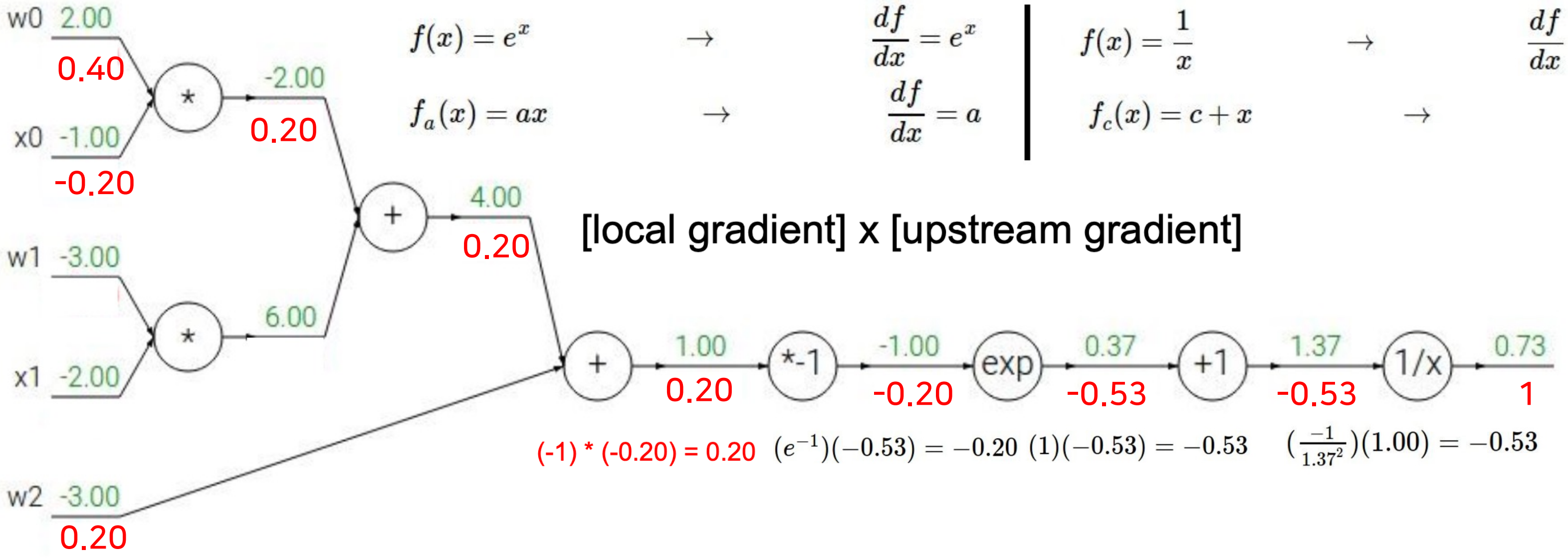
$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$

$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$

$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$

$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$

[local gradient] x [upstream gradient]



(-1) * (-0.20) = 0.20 $\quad (e^{-1})(-0.53) = -0.20 \quad (1)(-0.53) = -0.53 \quad (\frac{-1}{1.37^2})(1.00) = -0.53$

1. Backpropagation

Second Example

x0: [2] x [0.2] = 0.4
w0: [-1] x [0.2] = -0.2

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

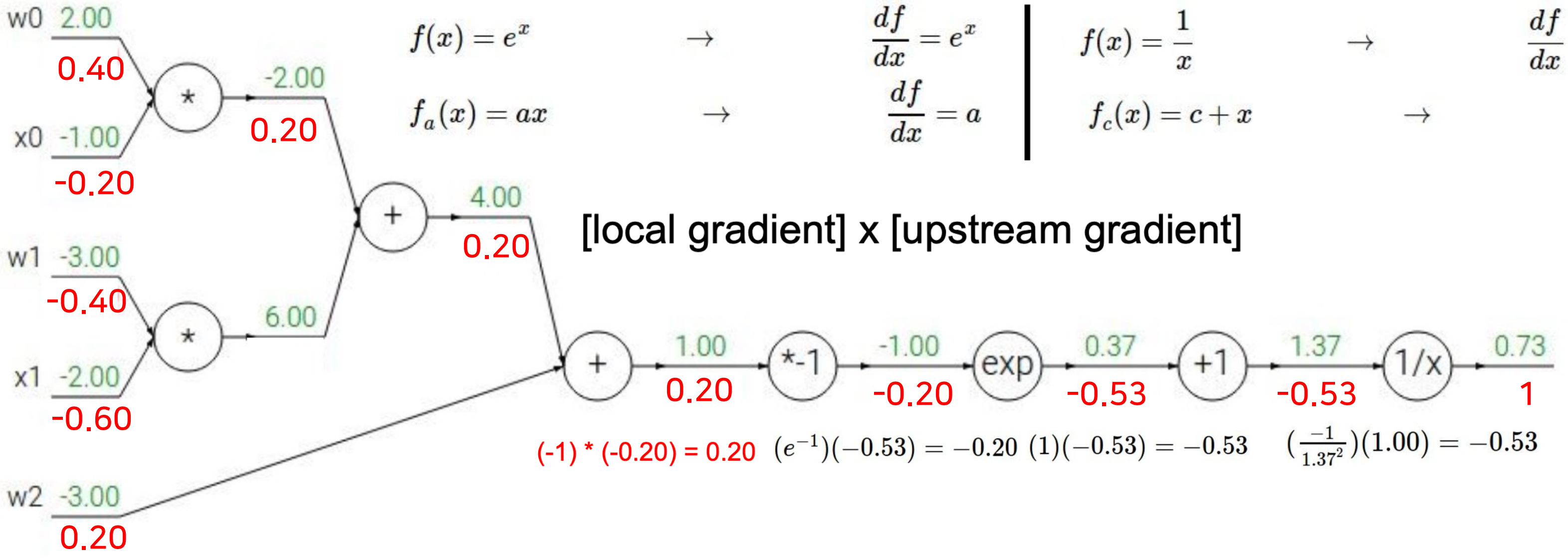
$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$

$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$

$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$

$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$

[local gradient] x [upstream gradient]



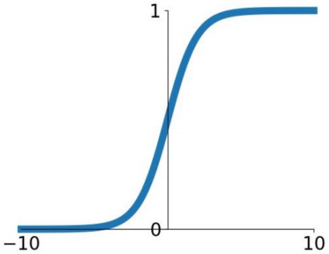
(-1) * (-0.20) = 0.20 $\quad (e^{-1})(-0.53) = -0.20 \quad (1)(-0.53) = -0.53 \quad (\frac{-1}{1.37^2})(1.00) = -0.53$

1. Backpropagation

Second Example

Sigmoid Function ☺

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}} \Rightarrow \sigma(x) = \frac{1}{1+e^{-x}}$$



Sigmoid

$$\frac{d\sigma(x)}{dx} = \frac{e^{-x}}{(1+e^{-x})^2} = \left(\frac{1+e^{-x}-1}{1+e^{-x}}\right)\left(\frac{1}{1+e^{-x}}\right) = (1-\sigma(x))\,\sigma(x)$$

We can compute the gradient at once if we know how to obtain the derivative !

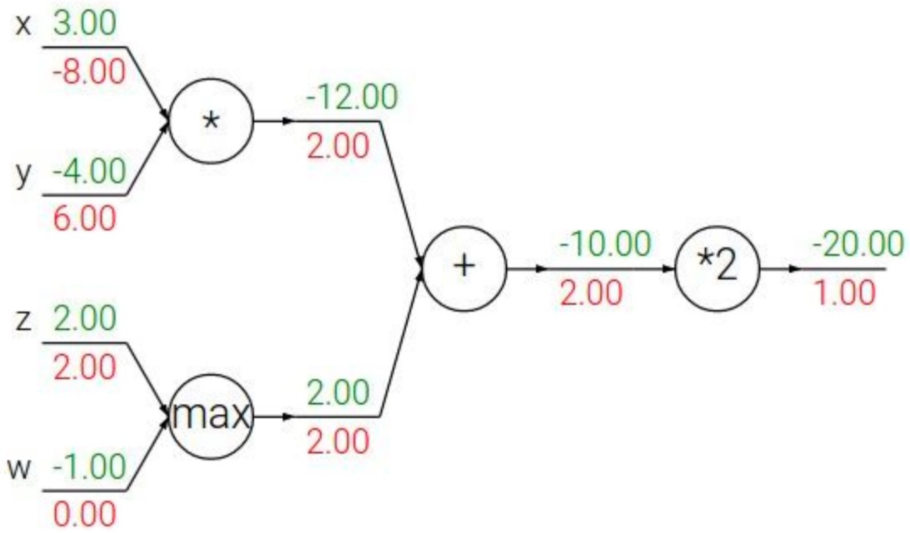
$$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x \qquad\qquad f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$$

$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a \qquad\qquad f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$$

1. Backpropagation

Rule of Backward()

add gate: gradient distributor

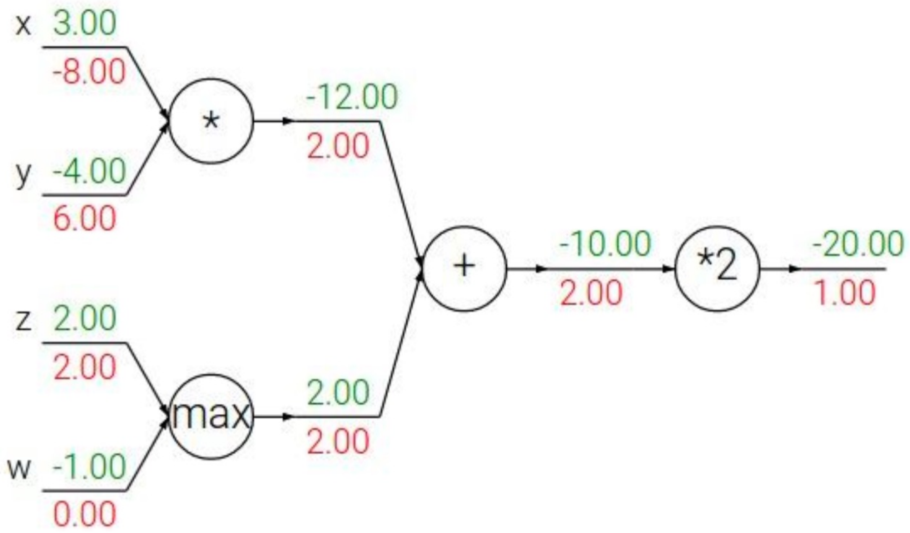


1. Backpropagation

Rule of Backward()

add gate: gradient distributor

Q: What is a max gate?



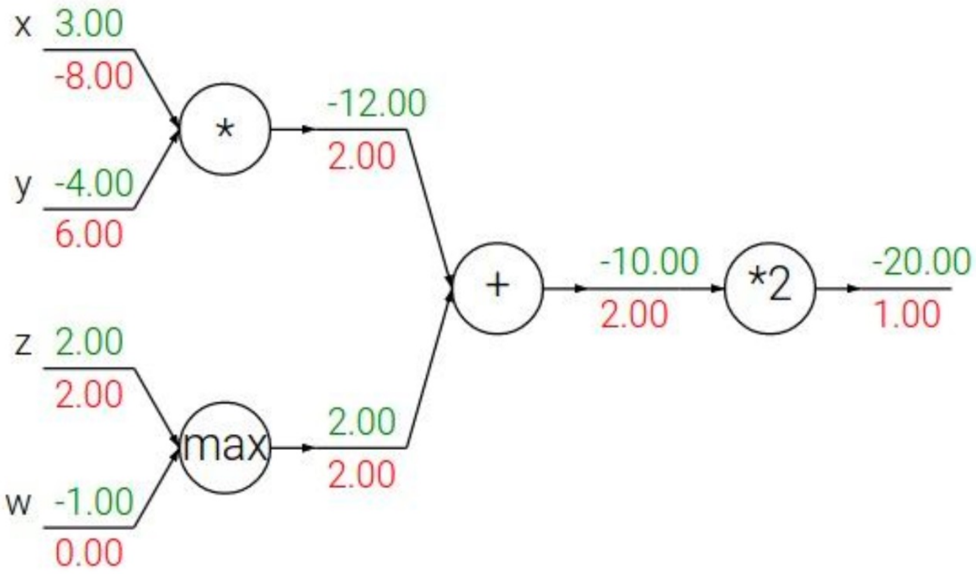
1. Backpropagation

Rule of Backward()

add gate: gradient distributor

max gate: gradient router

Q: What is a **mul** gate?

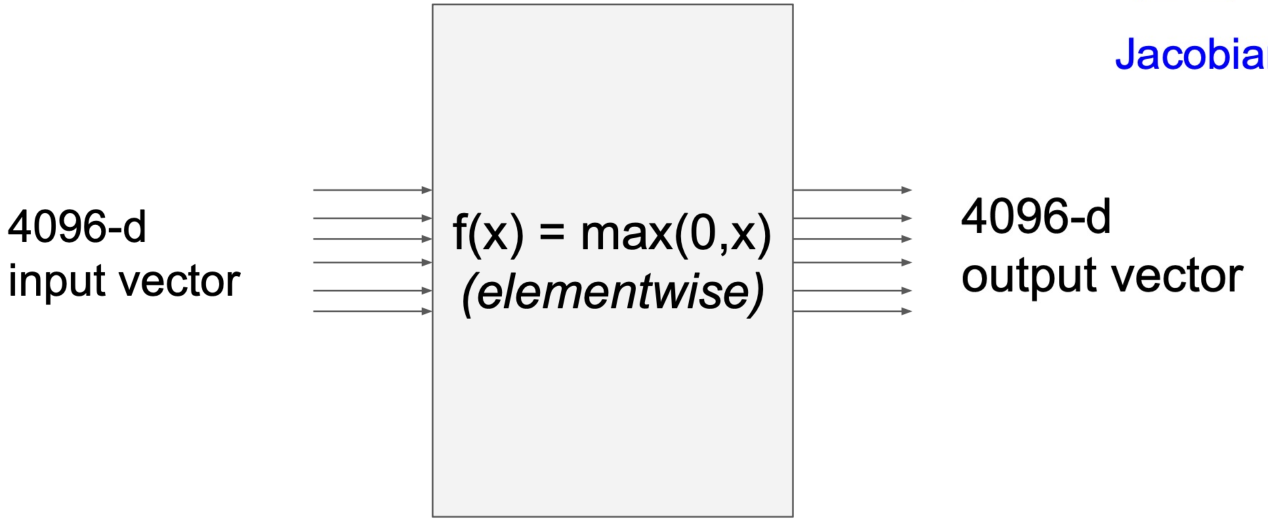


1. Backpropagation

Vectorized Operation

$$\frac{\partial L}{\partial x} = \boxed{\frac{\partial f}{\partial x}} \frac{\partial L}{\partial f}$$

Jacobian matrix



1. Backpropagation

Vectorized Operation

$$\frac{\partial L}{\partial x} = \boxed{\frac{\partial f}{\partial x}} \frac{\partial L}{\partial f}$$

Jacobian matrix

4096-d
input vector

$$\mathbf{J} = \frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}} = \left[\frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_1} \cdots \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_u}\right] = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_u} \\ \vdots & & \vdots \\ \frac{\partial f_v(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_v(\mathbf{x})}{\partial x_u} \end{bmatrix}$$

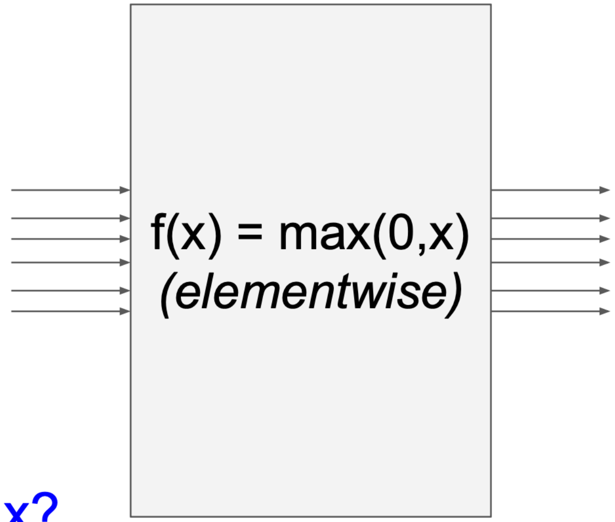
1. Backpropagation

Vectorized Operation

$$\frac{\partial L}{\partial x} = \boxed{\frac{\partial f}{\partial x}} \frac{\partial L}{\partial f}$$

Jacobian matrix

4096-d
input vector

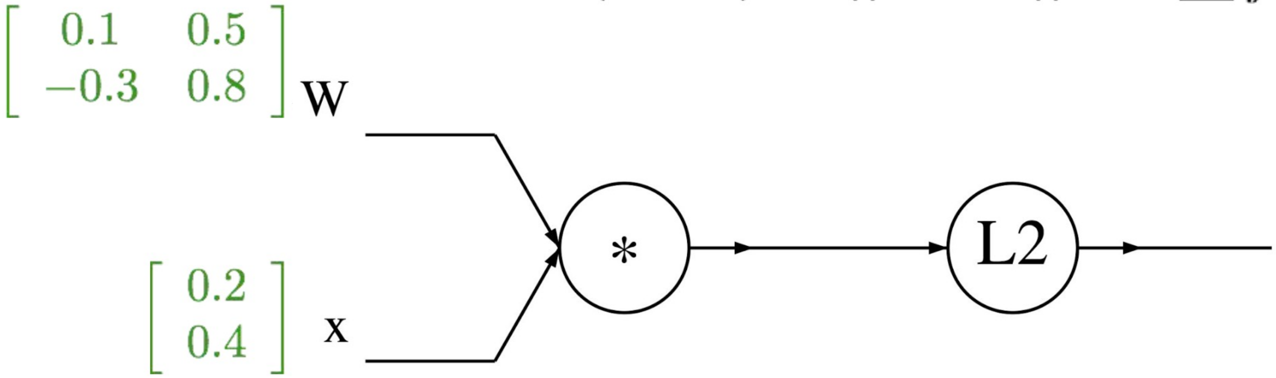


4096-d
output vector

Q: what is the size of the Jacobian matrix?

1. Backpropagation

A vectorized example: $f(x, W) = ||W \cdot x||^2 = \sum_{i=1}^{n} (W \cdot x)_i^2$



1. Backpropagation

PyTorch

backward() is all you need

```
for batch, (X, y) in enumerate(dataloader):
    X, y = X.to(device), y.to(device)

    # 예측 오류 계산
    pred = model(X)
    loss = loss_fn(pred, y)

    # 역전파
    optimizer.zero_grad()
    loss.backward()
    optimizer.step()
```

This one line makes whole operations at once, <u>automatically</u>

2. Neural Network

2. Neural Network

Nerual Network: without brain stuff

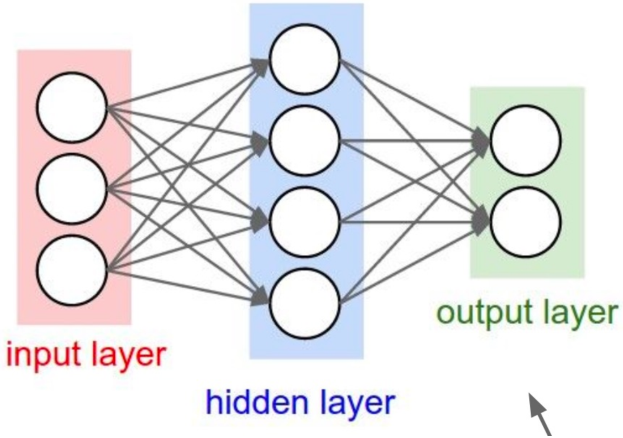
(Before) Linear Function $$f = Wx$$

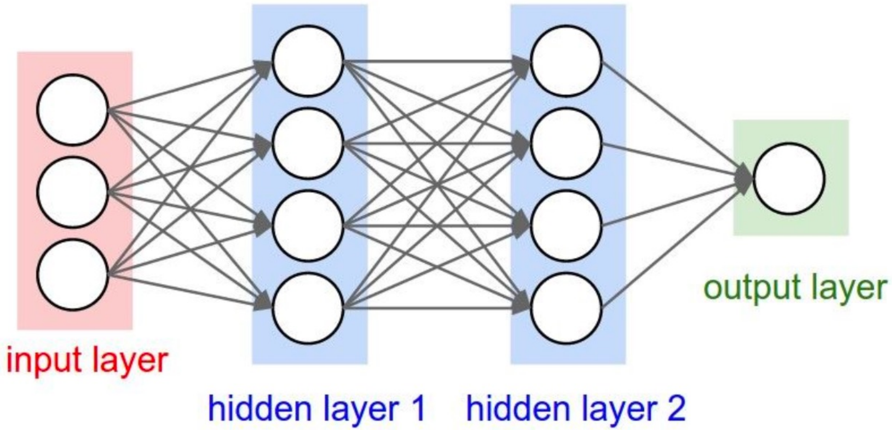
(Now) 2-layer Neural Network $$f = W_2 \max(0, W_1 x)$$

Or 3-layer Neural Network $$f = W_3 \max(0, W_2 \max(0, W_1 x))$$

2. Neural Network

Nerual Network: without brain stuff

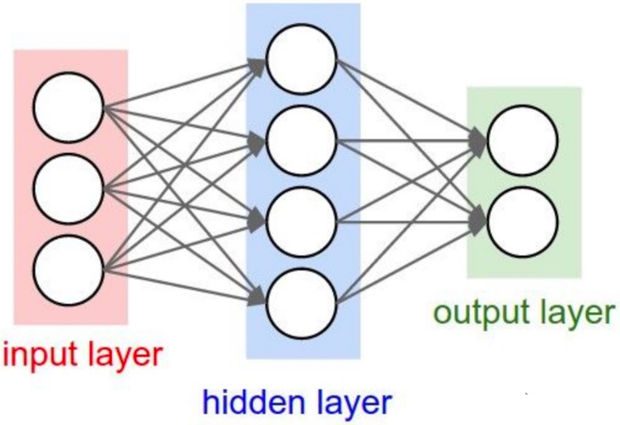


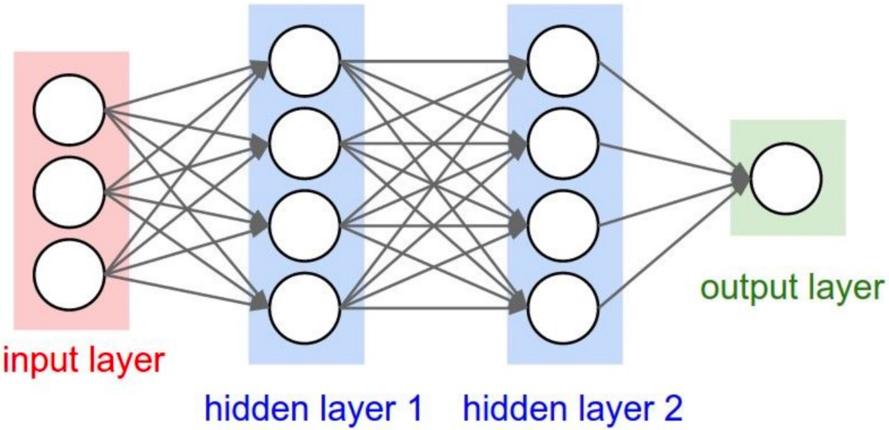


Fully Connected Layer !
a.k.a. FC Layer

2. Neural Network

Nerual Network: without brain stuff






```
f=lambda x:1.0/(1.0+np.exp(-x))
x=np.random.randn(3,1)
h1=f(np.dot(W1,x)+b1)
h2=f(np.dot(W2,h1)+b2)
output=np.dot(W3, h2)+b3
```

2. Neural Network

Nerual Network Full implementation with numpy

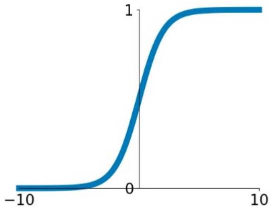
```
import numpy as np

N, D_in, H, D_out=64, 1000,100,10
x,y=np.random.randn(N,D_in),np.random.randn(N,D_out)
w1, w2=np.random.randn(D_in, H), np.random.randn(H, D_out)
learning_rate=1e-4

for t in range(2000):
  h=1/(1+np(exp(-x.dot(w1))))
  y_pred=h.dot(w2)
  loss=np.square(y_pred-y).sum()
  print(t, loss)

  grad_y_pred=2.0*(y_pred-y)
  grad_w2=h.T.dot(grad_y_pred)
  grad_h=grad_y_pred.dot(w2.T)
  grad_w1=x.T.dot(grad_h*h*(1-h))

  w1-=learning_rate*grad_w1
  w2-=learning_rate*grad_w2
```

2. Neural Network

Activation Function : That's why deep learning is called a nonlinear model

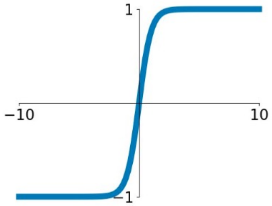
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



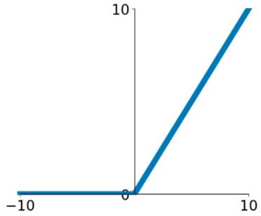
tanh

$$\tanh(x)$$



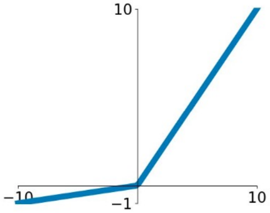
ReLU

$$\max(0, x)$$



Leaky ReLU

$$\max(0.1x, x)$$

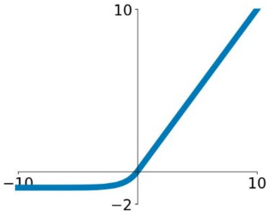


Maxout

$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

ELU

$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



2. Neural Network

How about Pytorch?

Fully Connected Layer

nn.Linear(input_dim, output_dim, bias=True(default))

nn.functional.linear(input, weight, bias)

What is the difference between torch.nn and torch.nn.functional?

Activation Functions

nn.ReLU()

nn.Sigmoid()

nn.Tanh()

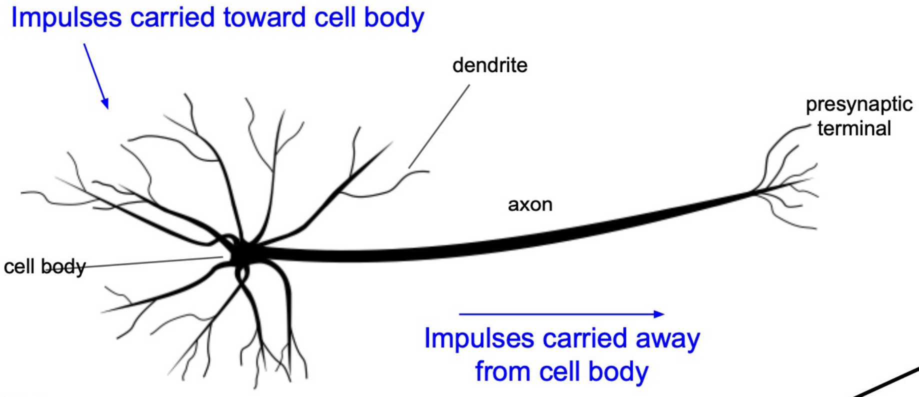
nn.LeakyReLU()

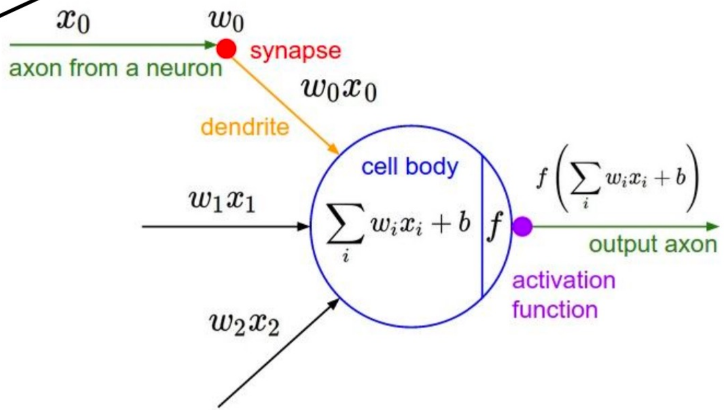
nn.ELU()

torch.max()

2. Neural Network

Neural Network: with brain stuff





3. Normalization

3. Normalization

*Think about Gradient
-> There are many gradients which are directly affected by the input
& We get data as ‘Mini-batch’

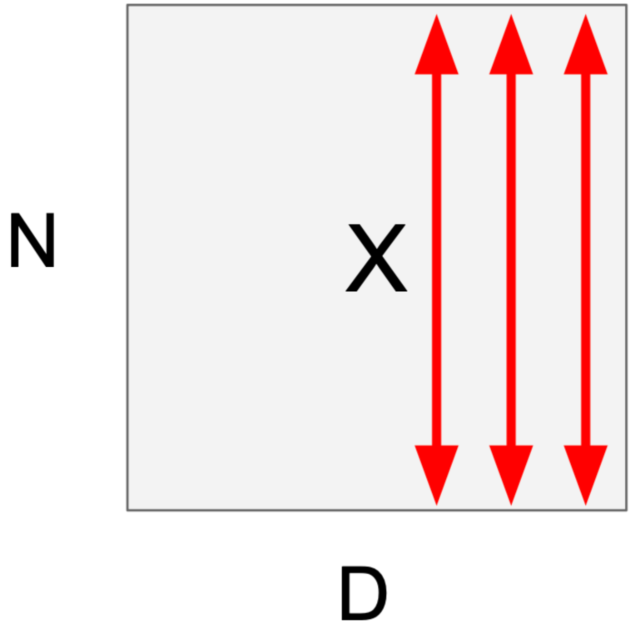
How if input’s scales are so different?

-> Normalization: makes each dimension unit gaussian, apply

$$\widehat{x}^{(k)} = \frac{x^{(k)} - \mathrm{E}[x^{(k)}]}{\sqrt{\mathrm{Var}[x^{(k)}]}}$$

3. Normalization

Batch Normalization

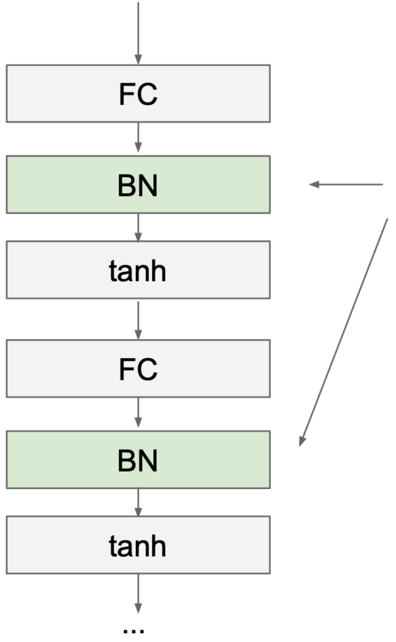


1. Compute the mean and variance independently
2. Normalize

$$\widehat{x}^{(k)} = \frac{x^{(k)} - \mathrm{E}[x^{(k)}]}{\sqrt{\mathrm{Var}[x^{(k)}]}}$$

3. Normalization

Batch Normalization: Frequently used in Computer Vision

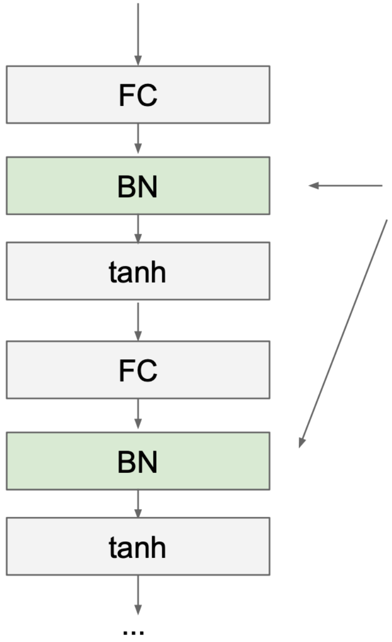


Usually used after Fully connected layer before nonlinearity

$$\widehat{x}^{(k)} = \frac{x^{(k)} - \mathrm{E}[x^{(k)}]}{\sqrt{\mathrm{Var}[x^{(k)}]}}$$

3. Normalization

Batch Normalization



Usually used after Fully connected layer before nonlinearity

$$\gamma^{(k)} = \sqrt{\mathrm{Var}[x^{(k)}]}$$

$$\beta^{(k)} = \mathrm{E}[x^{(k)}]$$

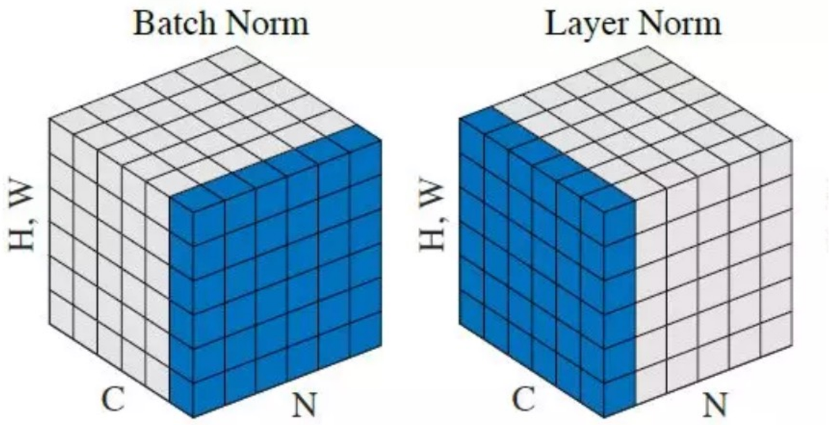
learning Mean & Variance

$$\widehat{x}^{(k)} = \frac{x^{(k)} - \mathrm{E}[x^{(k)}]}{\sqrt{\mathrm{Var}[x^{(k)}]}}$$

3. Normalization

Layer Normalization

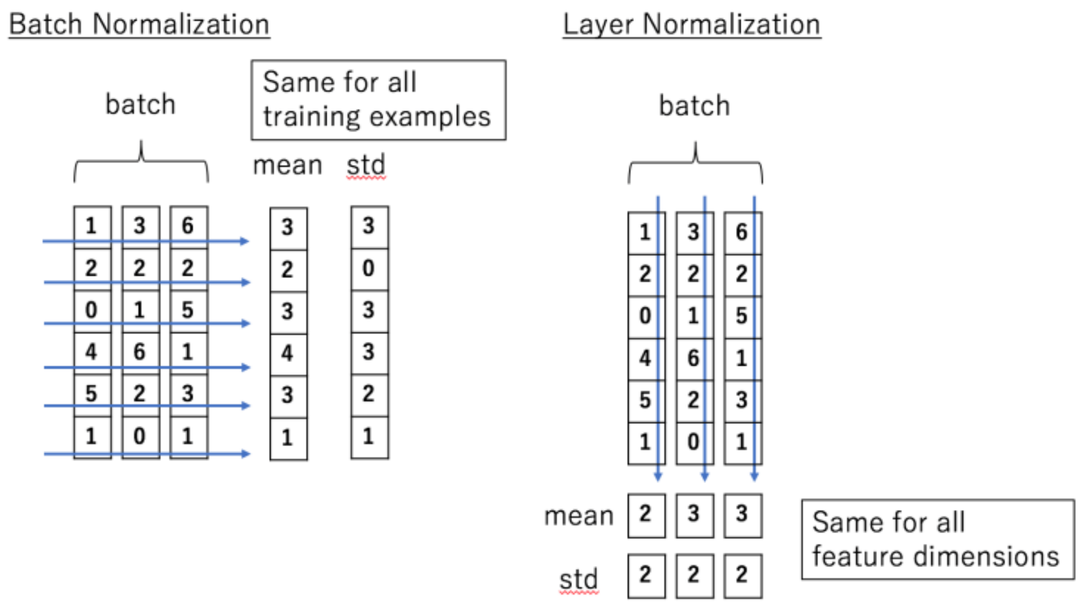
*What is the difference of LayerNorm & BatchNorm?



$$\tilde{z}^{(l)} = \frac{z^{(l)} - E(z^{(l)})}{\sqrt{var(z^{(l)} + \epsilon)}}$$

3. Normalization

Layer Normalization



1) Batch Normalization
- Normalize each "feature"
- Obtain mean & variance of each "feature"

2) Layer Normalization
- Normalize each "feature of input"
- Obtain mean & variance of feature of input

*More details would be explained on Next Paper Review ☺

Q&A

Have a nice week🥰