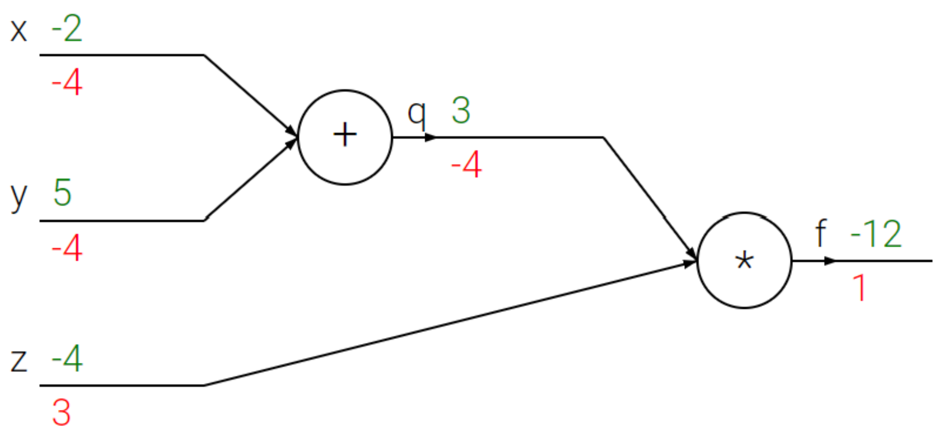
Введение

Backpropagation

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

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y}\frac{\partial y}{\partial x}$$

- Используем правило вычисления градиента сложной функции
- Если мы знаем вычислительный граф, то более «поздние» значения градиентов помогут вычислить более «ранние»!



Backprop

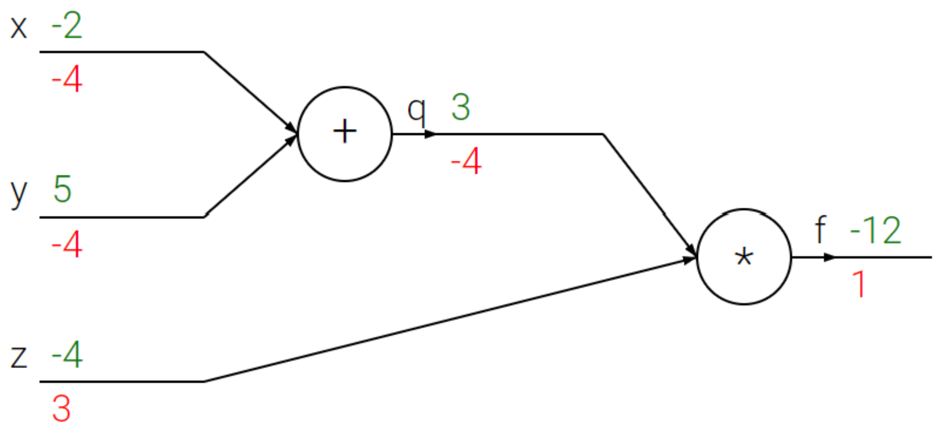
$$f(x, y, z) = (x + y)z$$

$$q = x + y \quad f = qz$$

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

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

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



Backprop

Проход вперед – посчитать значение

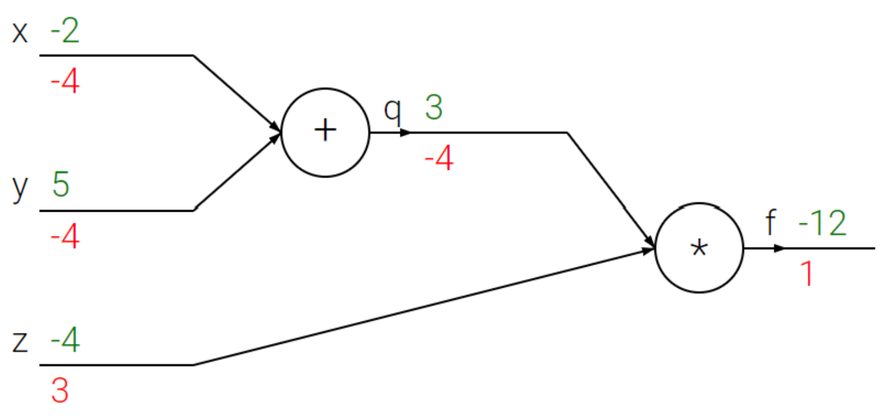
Проход назад – посчитать градиент

И этот процесс локален!

Каждый узел в сети может посчитать:

1) свой выход

2) локальные градиенты



Что происходит при обучении

$$Loss = f(x, y; \theta)$$

$$Loss = ((\sigma(xW_1 + b_1)W_2 + b_2) - y)^2$$

- 1. Определяем промежуточные функции
- 2. Считаем локальные градиенты
- 3. Добавляем ошибку, чтобы посчитатть полный градиент

$$h_1 = xW_1 + b_1$$

$$z_1 = \sigma(h_1)$$

$$z_2 = z_1 W_2 + b_2$$

$$Loss = (z_2 - y)^2$$

$$\frac{\partial h_1}{\partial x} = W_1^T$$

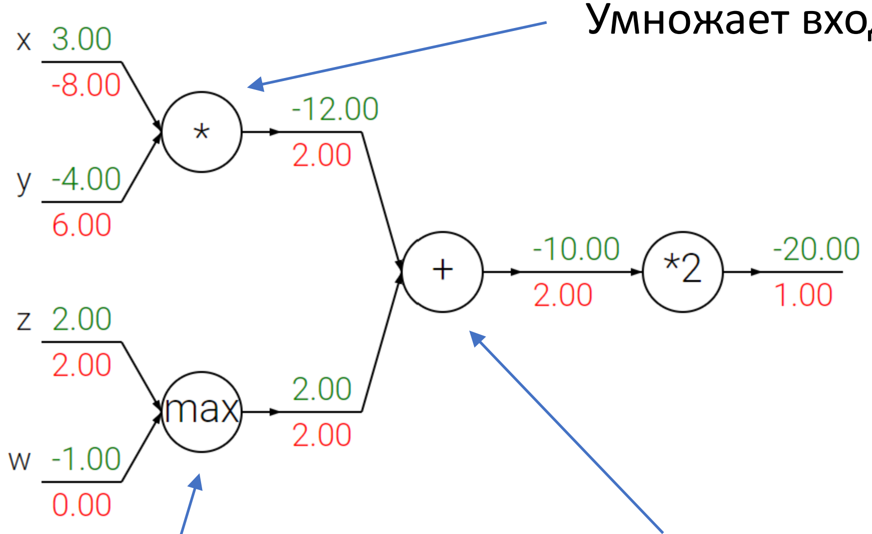
$$\frac{\partial z_1}{\partial h_1} = \sigma'(h_1) = z_1 \circ (1 - z_1)$$

$$\frac{\partial z_2}{\partial z_1} = W_2^\top$$

$$\frac{\partial Loss}{\partial z_2} = 2(z_2 - y)$$

А что с backprop для типичных блоков?

Умножает вход на другой выход



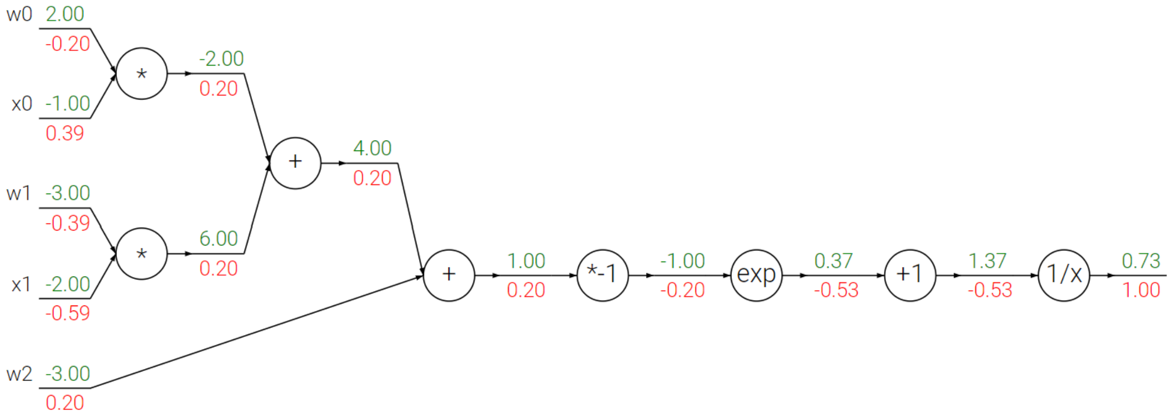
Распределяет градиент по входам

Направляет градиент!

Сигмоида

Логистическая регрессия:

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



Градиенты:

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

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

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

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

Градиент сигмоиды

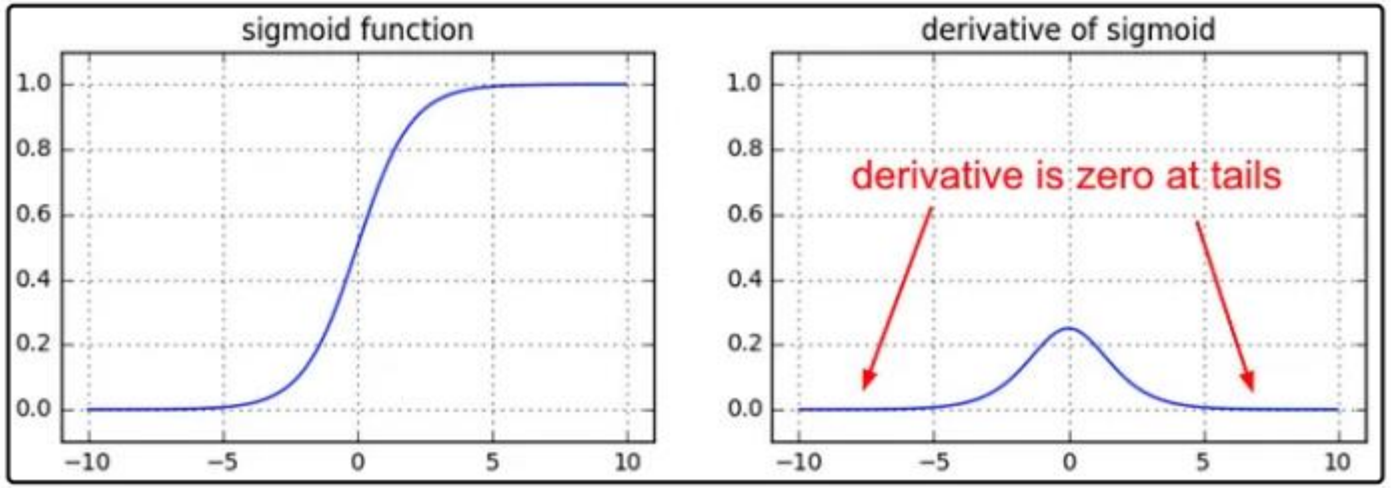
- Сигмоида от 1.0 – около 0.73
- Тогда локальный градиент (1- 0.73) * 0.73 ~= 0.2

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

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

Сигмоида не так проста!

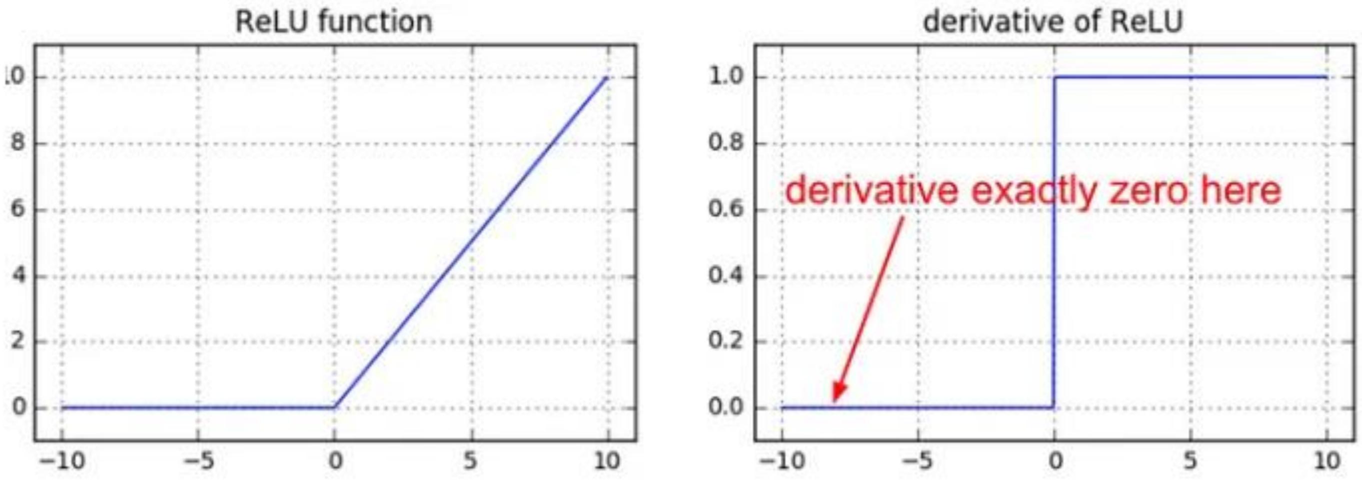
- Если веса слишком большие, то значение сигмоиды будет около 1
- Тогда градиент будет практически нулевым!



https://karpathy.medium.com/yes-you-should-understand-backprop-e2f06eab496b

A ReLU?

- Если в начале обучения вес попадет в область ниже нуля, то градиент для него всегда будет ноль
- Это может возникнуть и при слишком аггресивном обучении



- Скалярный выход, векторный вход

$$\frac{\partial y}{\partial \mathbf{x}} = \left[\frac{\partial y}{\partial x_1} \quad \frac{\partial y}{\partial x_2} \cdots \frac{\partial y}{\partial x_n}\right]$$

- Векторный выход, векторный вход

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

J

- Скалярный выход, матричный вход

$$\frac{\partial y}{\partial A} = \begin{bmatrix} \frac{\partial y}{\partial A_{11}} & \frac{\partial y}{\partial A_{12}} & \cdots & \frac{\partial y}{\partial A_{1n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial A_{m1}} & \frac{\partial y}{\partial A_{m2}} & \cdots & \frac{\partial y}{\partial A_{mn}} \end{bmatrix}$$

- Векторный выход ,матричный вход

$$\frac{\partial y}{\partial A_{ij}} = \frac{\partial y}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial A_{ij}}$$

$$f(x,y) = \frac{x + \sigma(y)}{\sigma(x) + (x+y)^2}$$

```
x = 3 # example values
y = -4

# forward pass
sigy = 1.0 / (1 + math.exp(-y)) # sigmoid in numerator   #(1)
num = x + sigy # numerator                                #(2)
sigx = 1.0 / (1 + math.exp(-x)) # sigmoid in denominator #(3)
xpy = x + y                                               #(4)
xpysqr = xpy**2                                           #(5)
den = sigx + xpysqr # denominator                         #(6)
invden = 1.0 / den                                        #(7)
f = num * invden # done!                                  #(8)
```

$$f(x,y) = \frac{x + \sigma(y)}{\sigma(x) + (x+y)^2}$$

```
# backprop f = num * invden
dnum = invden # gradient on numerator                          #(8)
dinvden = num                                                  #(8)
# backprop invden = 1.0 / den
dden = (-1.0 / (den**2)) * dinvden                             #(7)
# backprop den = sigx + xpysqr
dsigx = (1) * dden                                             #(6)
dxpysqr = (1) * dden                                           #(6)
# backprop xpysqr = xpy**2
dxpy = (2 * xpy) * dxpysqr                                     #(5)
# backprop xpy = x + y
```

$$f(x,y) = \frac{x + \sigma(y)}{\sigma(x) + (x+y)^2}$$

```
dx = (1) * dxpy                                                      #(4)
dy = (1) * dxpy                                                      #(4)
# backprop sigx = 1.0 / (1 + math.exp(-x))
dx += ((1 - sigx) * sigx) * dsigx # Notice += !! See notes below  #(3)
# backprop num = x + sigy
dx += (1) * dnum                                                     #(2)
dsigy = (1) * dnum                                                   #(2)
# backprop sigy = 1.0 / (1 + math.exp(-y))
dy += ((1 - sigy) * sigy) * dsigy                                    #(1)
# done! phew
```