академия больших данных mail.ru group

Полносвязные сети. Метод обратного распространения ошибки

Фёдор Киташов

Программист-исследователь в команде компьютерного зрения

План лекции

- Повторение: логистическая регрессия через матричные перемножения
- Полносвязаные сети
- Обучение полносвязных сетей через метод обратного распространения ошибки (backpropagation)

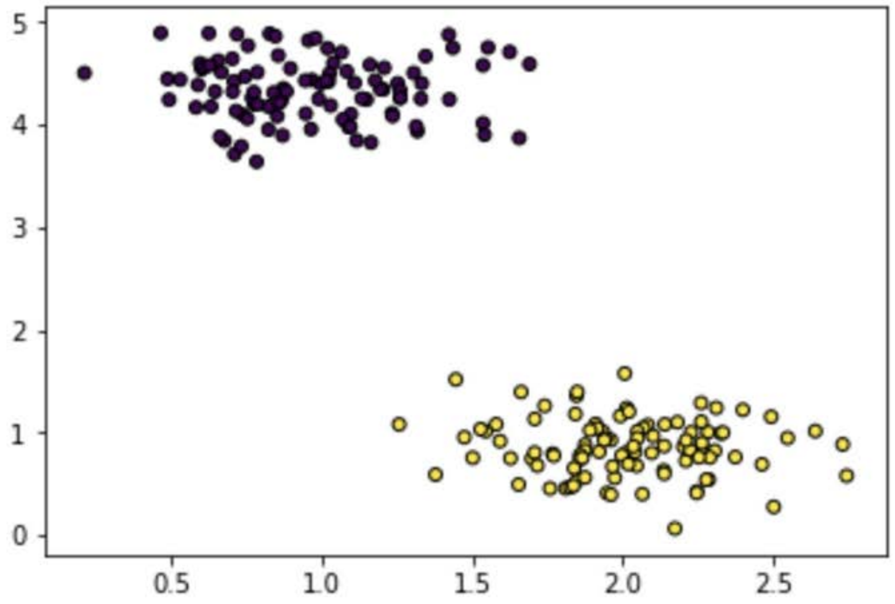
Повторение

```
import sklearn
from sklearn.datasets import make_blobs

X, y = make_blobs(n_samples=200, centers=2, n_features=2,
                  random_state=0, cluster_std=0.3)

plt.scatter(X[:, 0], X[:, 1], marker='o', c=y,
    s=25, edgecolor='k')
```

```
<matplotlib.collections.PathCollection at 0x1a2969dac8>
```



Повторение

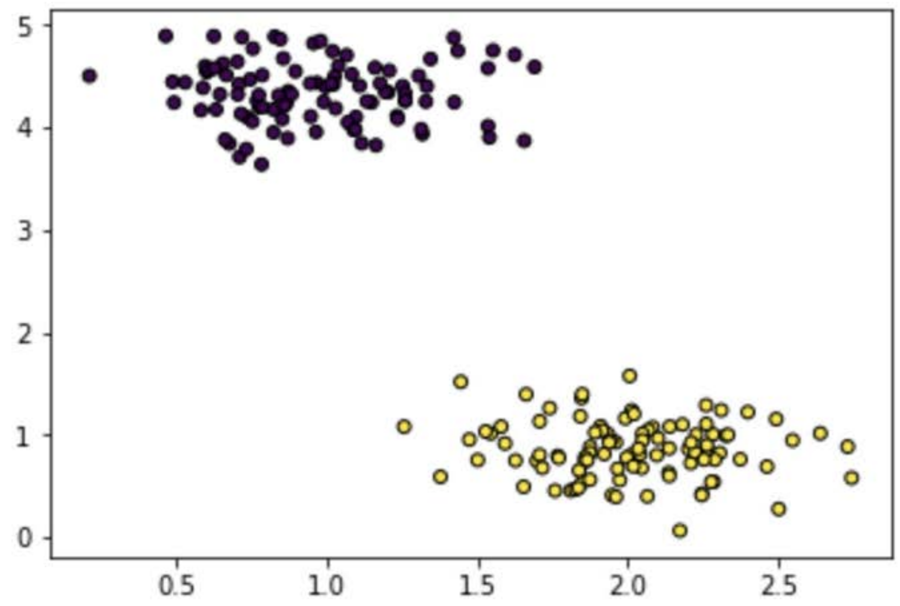
- Какие параметры модели?
- Какая размерность параметров модели?

```
import sklearn
from sklearn.datasets import make_blobs

X, y = make_blobs(n_samples=200, centers=2, n_features=2,
                  random_state=0, cluster_std=0.3)

plt.scatter(X[:, 0], X[:, 1], marker='o', c=y,
    s=25, edgecolor='k')
```

```
<matplotlib.collections.PathCollection at 0x1a2969dac8>
```



Повторение

- Какие параметры модели?
- Какая размерность параметров модели?

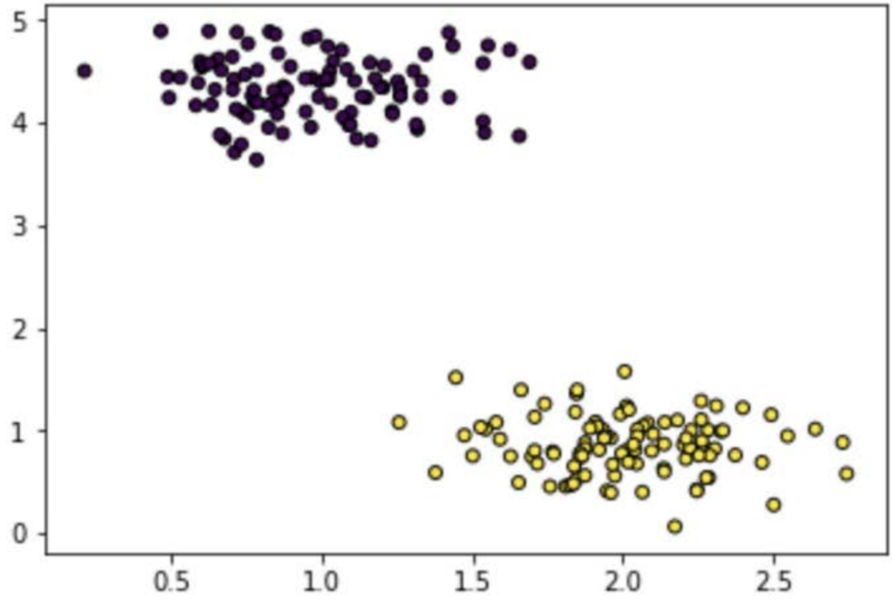
$$\begin{bmatrix} x \\ y \end{bmatrix}$$

```
import sklearn
from sklearn.datasets import make_blobs

X, y = make_blobs(n_samples=200, centers=2, n_features=2,
                  random_state=0, cluster_std=0.3)

plt.scatter(X[:, 0], X[:, 1], marker='o', c=y,
     s=25, edgecolor='k')
```

```
<matplotlib.collections.PathCollection at 0x1a2969dac8>
```



Повторение

- Какие параметры модели?
- Какая размерность параметров модели?

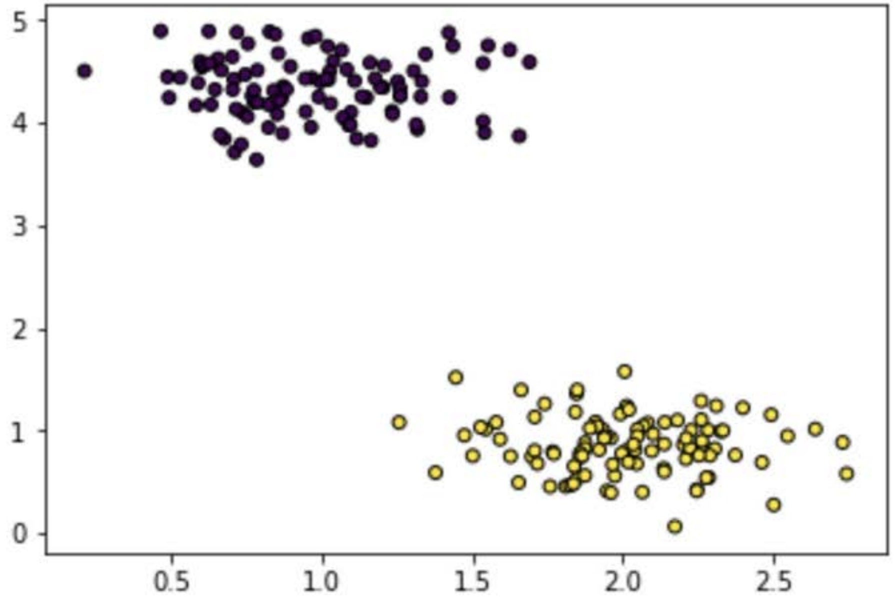
```python
import sklearn
from sklearn.datasets import make_blobs

X, y = make_blobs(n_samples=200, centers=2, n_features=2,
                  random_state=0, cluster_std=0.3)

plt.scatter(X[:, 0], X[:, 1], marker='o', c=y,
    s=25, edgecolor='k')
```

```
<matplotlib.collections.PathCollection at 0x1a2969dac8>
```

$$\begin{bmatrix} w0 & w1 \end{bmatrix} \times \begin{bmatrix} x \\ y \end{bmatrix}$$



Повторение

- Какие параметры модели?
- Какая размерность параметров модели?

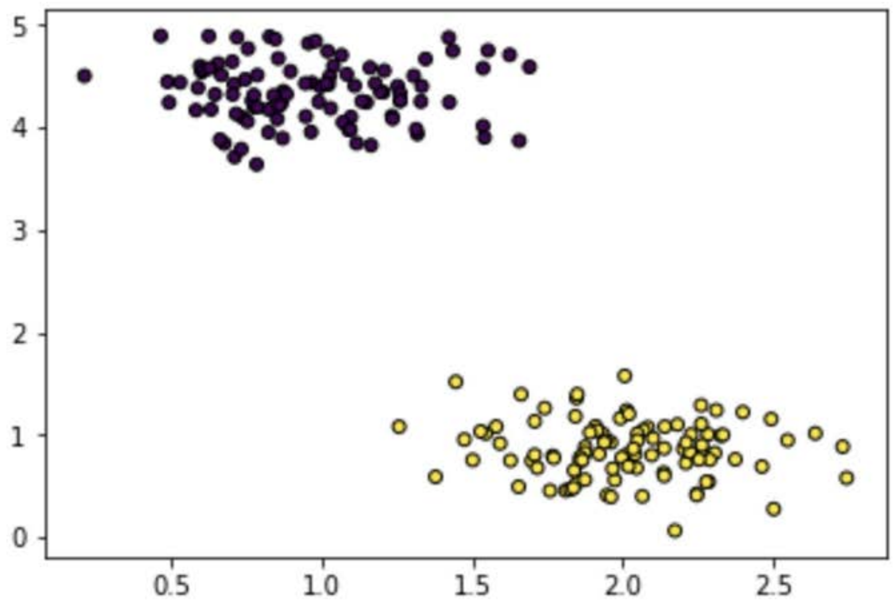
$$\begin{bmatrix} w0 & w1 \end{bmatrix} \times \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} b1 \end{bmatrix}$$

```python
import sklearn
from sklearn.datasets import make_blobs

X, y = make_blobs(n_samples=200, centers=2, n_features=2,
                  random_state=0, cluster_std=0.3)

plt.scatter(X[:, 0], X[:, 1], marker='o', c=y,
    s=25, edgecolor='k')
```

```
<matplotlib.collections.PathCollection at 0x1a2969dac8>
```



Повторение

- Какие параметры модели?
- Какая размерность параметров модели?

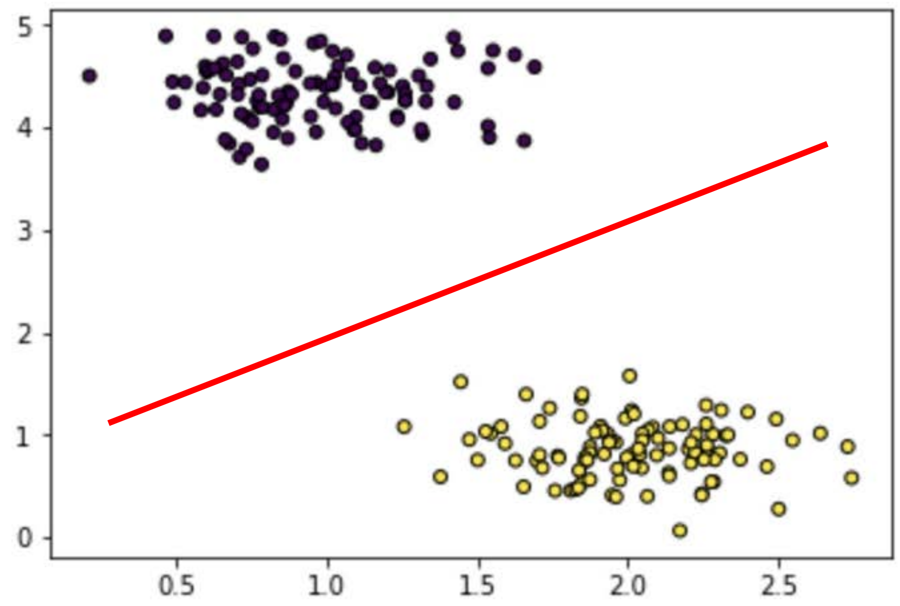
$$\mathrm{sign}\left(\begin{bmatrix} w0 & w1 \end{bmatrix} \times \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} b1 \end{bmatrix} \right)$$

```python
import sklearn
from sklearn.datasets import make_blobs

X, y = make_blobs(n_samples=200, centers=2, n_features=2,
                  random_state=0, cluster_std=0.3)

plt.scatter(X[:, 0], X[:, 1], marker='o', c=y,
    s=25, edgecolor='k')
```

```
<matplotlib.collections.PathCollection at 0x1a2969dac8>
```



Повторение: логистическая регрессия

- Какие параметры модели?
- Какая размерность параметров модели?

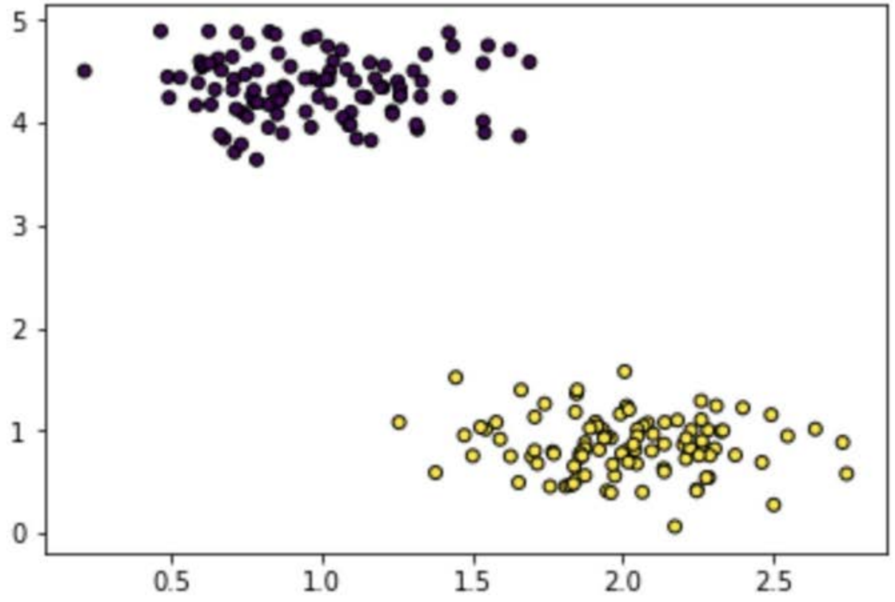
$$p = \text{sigmoid}\left(\begin{bmatrix} w0 & w1 \end{bmatrix} \times \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} b1 \end{bmatrix} \right)$$

```python
import sklearn
from sklearn.datasets import make_blobs

X, y = make_blobs(n_samples=200, centers=2, n_features=2,
                  random_state=0, cluster_std=0.3)

plt.scatter(X[:, 0], X[:, 1], marker='o', c=y,
    s=25, edgecolor='k')
```

```
<matplotlib.collections.PathCollection at 0x1a2969dac8>
```



Повторение: логистическая регрессия

- Какие параметры модели?
- Какая размерность параметров модели?

$$p = \mathrm{sigmoid}\left(\begin{bmatrix} w0 & w1 \end{bmatrix} \times \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} b1 \end{bmatrix} \right)$$

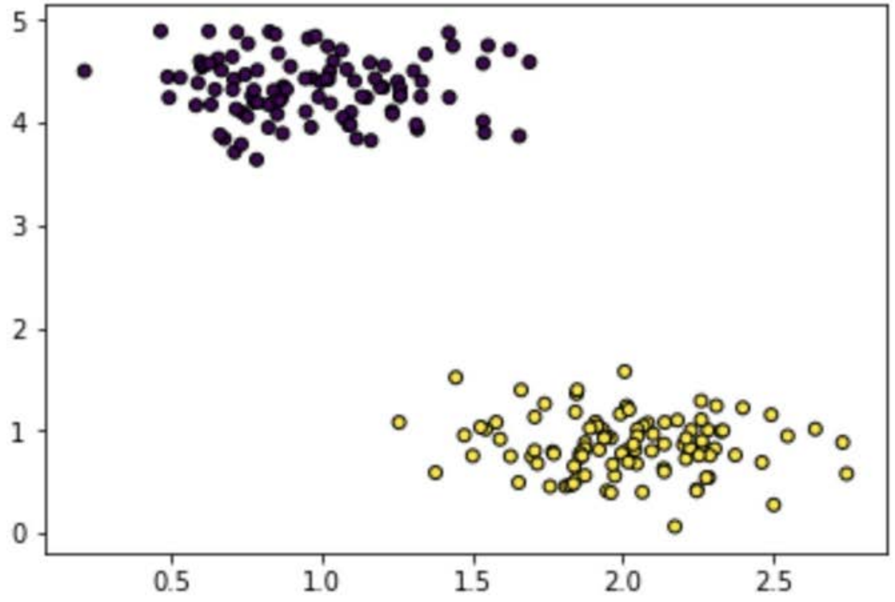
$$J = -(y \log(p) + (1-y) \log(1-p))$$

```
import sklearn
from sklearn.datasets import make_blobs

X, y = make_blobs(n_samples=200, centers=2, n_features=2,
                  random_state=0, cluster_std=0.3)

plt.scatter(X[:, 0], X[:, 1], marker='o', c=y,
    s=25, edgecolor='k')
```

```
<matplotlib.collections.PathCollection at 0x1a2969dac8>
```



Повторение: логистическая регрессия

- Какие параметры модели?
- Какая размерность параметров модели?
- **Как обновлять веса?**

$$p = \mathrm{sigmoid}\left(\begin{bmatrix} w0 & w1 \end{bmatrix} \times \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} b1 \end{bmatrix} \right)$$

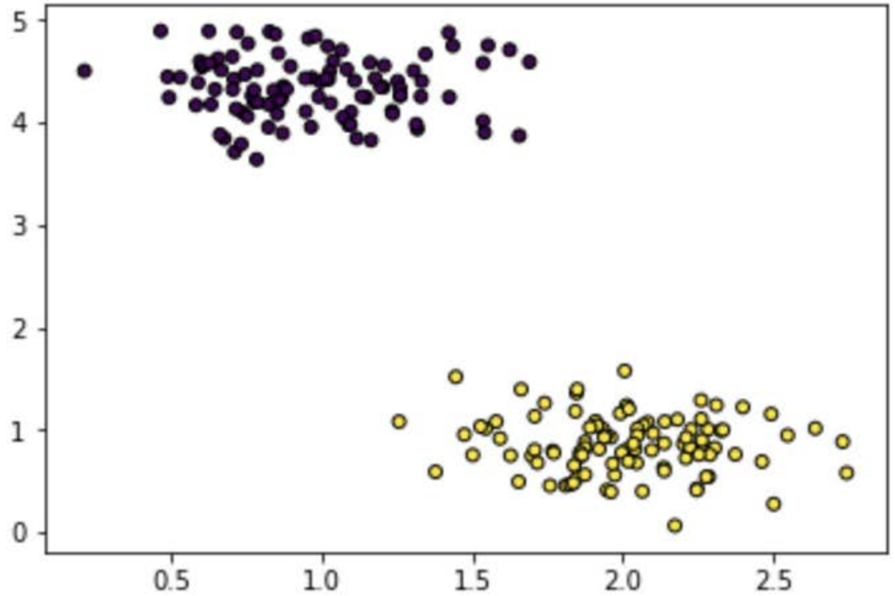
$$J = -(y\log(p) + (1-y)\log(1-p))$$

```python
import sklearn
from sklearn.datasets import make_blobs

X, y = make_blobs(n_samples=200, centers=2, n_features=2,
                  random_state=0, cluster_std=0.3)

plt.scatter(X[:, 0], X[:, 1], marker='o', c=y,
    s=25, edgecolor='k')
```

```
<matplotlib.collections.PathCollection at 0x1a2969dac8>
```



Повторение: логистическая регрессия

- Какие параметры модели?
- Какая размерность параметров модели?
- **Как обновлять веса?**

$$p = \mathrm{sigmoid}\left(\begin{bmatrix} w0 & w1 \end{bmatrix} \times \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} b1 \end{bmatrix} \right)$$

$$J = -(y\log(p) + (1-y)\log(1-p))$$

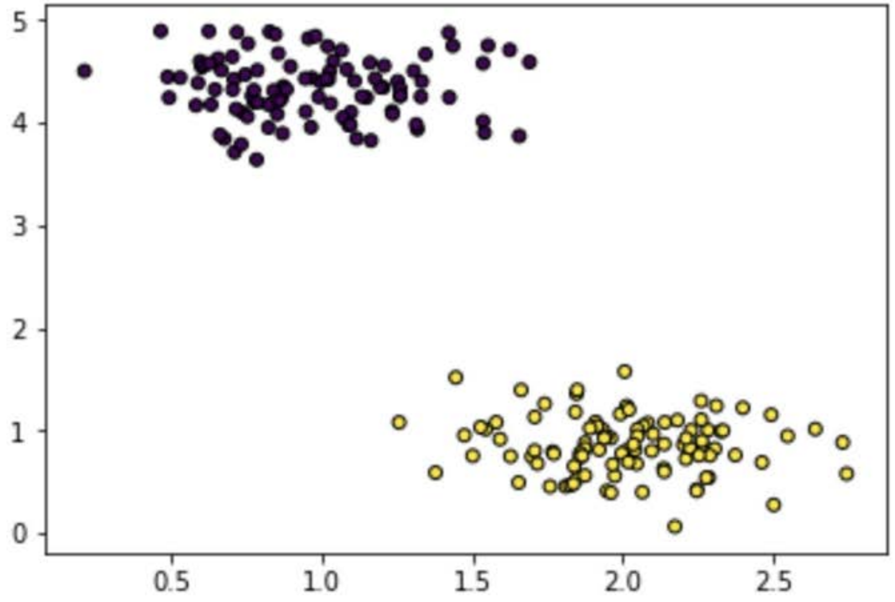
$$W^{t+1} = W^t - \eta \times \nabla Loss_W$$

```python
import sklearn
from sklearn.datasets import make_blobs

X, y = make_blobs(n_samples=200, centers=2, n_features=2,
                  random_state=0, cluster_std=0.3)

plt.scatter(X[:, 0], X[:, 1], marker='o', c=y,
    s=25, edgecolor='k')
```

```
<matplotlib.collections.PathCollection at 0x1a2969dac8>
```



Проблемы

- Как быть, если классов больше, чем 2?
- Как быть, если проблема нелинейная?

Проблемы

- **Как быть, если классов больше, чем 2?**
- Как быть, если проблема нелинейная?

Многоклассовая классификация

$$p(y = 1 \mid x) = \frac{1}{1 + \exp(-\langle w, x \rangle)}$$

Многоклассовая классификация

$$p(y = 1 \mid x) = \frac{1}{1 + \exp(-\langle w, x \rangle)}$$

- Как быть, если классов больше, чем два?

Многоклассовая классификация

$$p(y = 1 \mid x) = \frac{1}{1 + \exp(-\langle w, x \rangle)}$$

- Как быть, если классов больше, чем два? Предположим, что классов 3

Многоклассовая классификация

$$p(y = 1 \mid x) = \frac{1}{1 + \exp(-\langle w, x \rangle)}$$

- Как быть, если классов больше, чем два? Предположим, что классов 3

$$\begin{bmatrix} x \\ y \end{bmatrix}$$

Многоклассовая классификация

$$p(y=1\,|\,x)=\frac{1}{1+\exp(-\langle w,x\rangle)}$$

- Как быть, если классов больше, чем два? Предположим, что классов 3

$$\begin{bmatrix} w00 & w01 \\ w10 & w11 \\ w20 & w21 \end{bmatrix} \times \begin{bmatrix} x \\ y \end{bmatrix}$$

Многоклассовая классификация

$$p(y = 1 \mid x) = \frac{1}{1 + \exp(-\langle w, x \rangle)}$$

- Как быть, если классов больше, чем два? Предположим, что классов 3

$$\begin{bmatrix} w00 & w01 \\ w10 & w11 \\ w20 & w21 \end{bmatrix} \times \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} z0 \\ z1 \\ z2 \end{bmatrix}$$

Многоклассовая классификация

$$p(y=1\,|\,x) = \frac{1}{1+\exp(-\langle w, x\rangle)}$$

- Как быть, если классов больше, чем два? Предположим, что классов 3

$$\begin{bmatrix} w00 & w01 \\ w10 & w11 \\ w20 & w21 \end{bmatrix} \times \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} b0 \\ b1 \\ b2 \end{bmatrix} = \begin{bmatrix} z0 \\ z1 \\ z2 \end{bmatrix}$$

Многоклассовая классификация

$$p(y=1\,|\,x) = \frac{1}{1+\exp(-\langle w, x\rangle)}$$

- Как быть, если классов больше, чем два? Предположим, что классов 3

$$\text{softmax}\left(\begin{bmatrix} w00 & w01 \\ w10 & w11 \\ w20 & w21 \end{bmatrix} \times \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} b0 \\ b1 \\ b2 \end{bmatrix}\right) = \begin{bmatrix} z0 \\ z1 \\ z2 \end{bmatrix}$$

Многоклассовая классификация

$$p(y=1\,|\,x)=\frac{1}{1+\exp(-\langle w,x\rangle)}$$

- Как быть, если классов больше, чем два? Предположим, что классов 3

$$\sigma(\vec{z})_i=\frac{e^{z_i}}{\sum_{j=1}^{K}e^{z_j}}$$

σ = softmax

$\vec{z}$ = input vector

e^{z_i} = standard exponential function for input vector

K = number of classes in the multi-class classifier

e^{z_j} = standard exponential function for output vector

$$\text{softmax}\left(\begin{bmatrix} w00 & w01 \\ w10 & w11 \\ w20 & w21 \end{bmatrix} \times \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} b0 \\ b1 \\ b2 \end{bmatrix}\right) = \begin{bmatrix} z0 \\ z1 \\ z2 \end{bmatrix}$$

Softmax

```python
import numpy as np

def softmax(xs):
    return np.exp(xs) / sum(np.exp(xs))

xs = np.array([-1, 0, 3, 5])
print(softmax(xs)) # [0.0021657, 0.00588697, 0.11824302, 0.87370431]
```

https://victorzhou.com/blog/softmax/

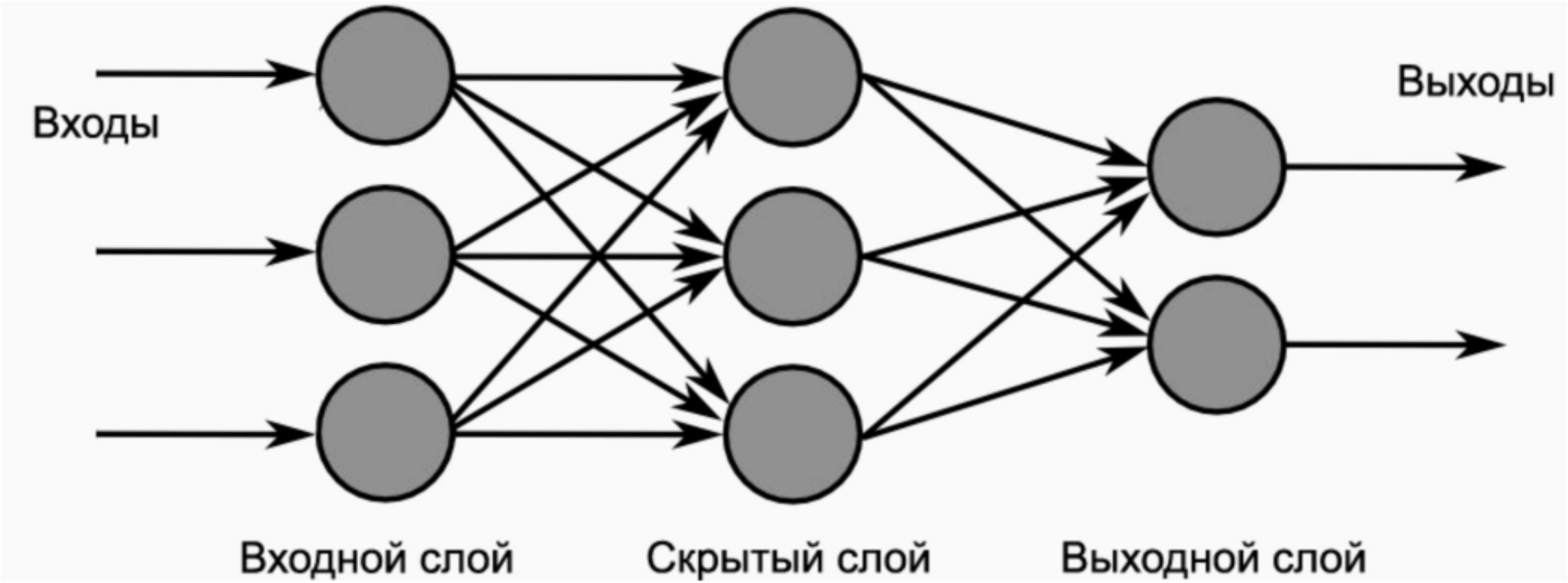
Проблемы

- Как быть, если классов больше, чем 2?
- **Как быть, если проблема нелинейная?**

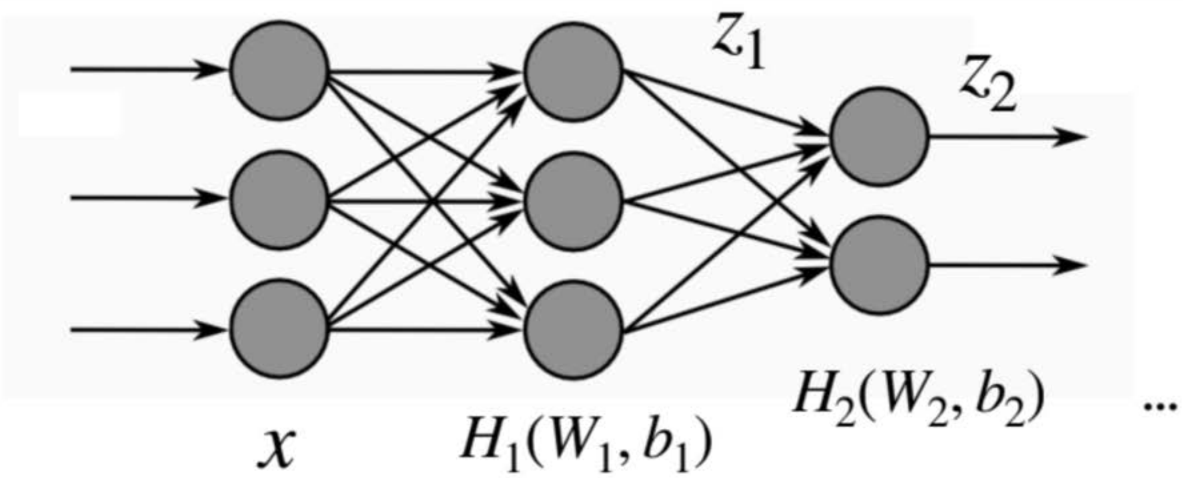
Как быть, если проблема нелинейная?

[Интерактивная визуализация]

Решение: многослойная сеть



Решение: многослойная сеть



$$z_1 = Activation(H_1(x)) = Activation(W_1 \times x + b_1)$$

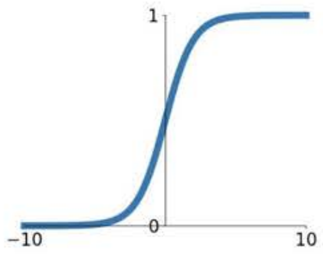
$$z_2 = Activation(H_2(z_1)) = Activation(W_2 \times z_1 + b_2)$$

$$W^{t+1} = W^t - \eta \times \nabla Loss_W$$

Активации a.k.a нелинейности

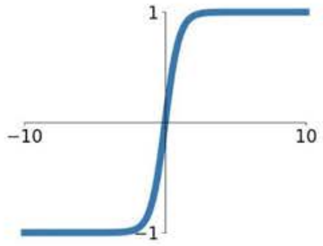
Sigmoid

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



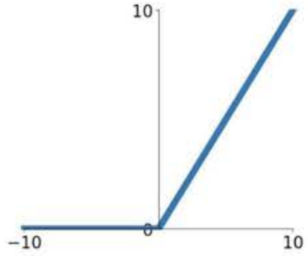
tanh

$\tanh(x)$



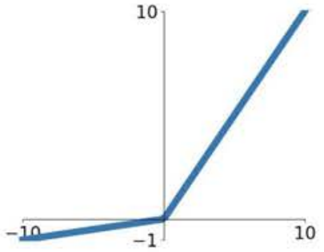
ReLU

$\max(0, x)$



Leaky ReLU

$\max(0.1x, x)$

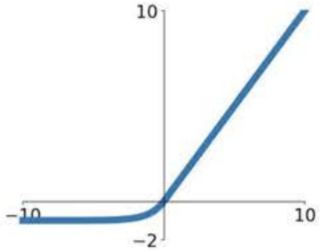


Maxout

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

ELU

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



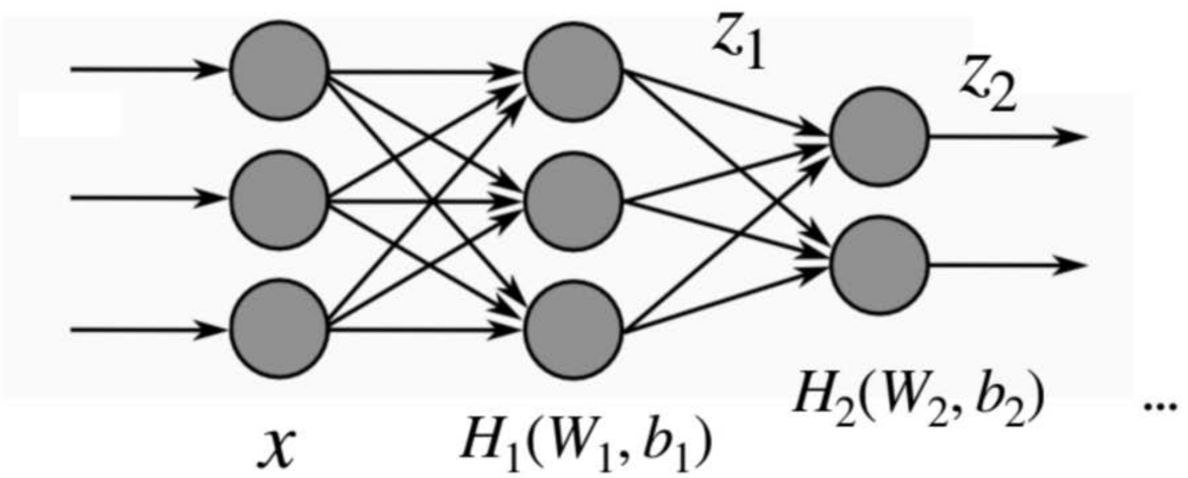
http://cs231n.stanford.edu/slides/2020/lecture_4.pdf

Какой лосс используется при классификации?

Какой лосс используется при классификации?

$$H(p, q) = -\sum_{x} p(x) \log q(x)$$

Обучение многослойной сети



$$z_1 = Activation(H_1(x)) = Activation(W_1 \times x + b_1)$$

$$z_2 = Activation(H_2(z_1)) = Activation(W_2 \times z_1 + b_2)$$

$$W^{t+1} = W^t - \eta \times \nabla Loss_W$$

Метод обратного распространения ошибки

$$\boldsymbol{x} = \text{input}$$

$$\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}_1$$

$$\boldsymbol{h} = \text{ReLU}(\boldsymbol{z})$$

$$\boldsymbol{\theta} = \boldsymbol{U}\boldsymbol{h} + \boldsymbol{b}_2$$

$$\hat{\boldsymbol{y}} = \text{softmax}(\boldsymbol{\theta})$$

$$J = CE(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

Метод обратного распространения ошибки

$$\boldsymbol{x} = \text{input}$$

$$\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}_1$$

$$\boldsymbol{h} = \text{ReLU}(\boldsymbol{z})$$

$$\boldsymbol{\theta} = \boldsymbol{U}\boldsymbol{h} + \boldsymbol{b}_2$$

$$\hat{\boldsymbol{y}} = \text{softmax}(\boldsymbol{\theta})$$

$$J = CE(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

Какие градиенты нам нужно посчитать, чтобы обновить веса слоёв?

Метод обратного распространения ошибки

$$\boldsymbol{x} = \text{input}$$

$$\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}_1$$

$$\boldsymbol{h} = \text{ReLU}(\boldsymbol{z})$$

$$\boldsymbol{\theta} = \boldsymbol{U}\boldsymbol{h} + \boldsymbol{b}_2$$

$$\hat{\boldsymbol{y}} = \text{softmax}(\boldsymbol{\theta})$$

$$J = CE(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

Какие градиенты нам нужно посчитать, чтобы обновить веса слоёв?

$$\frac{\partial J}{\partial \boldsymbol{U}} \qquad \frac{\partial J}{\partial \boldsymbol{b}_2} \qquad \frac{\partial J}{\partial \boldsymbol{W}} \qquad \frac{\partial J}{\partial \boldsymbol{b}_1}$$

Метод обратного распространения ошибки

$$\boldsymbol{x} = \text{input}$$

$$\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}_1$$

$$\boldsymbol{h} = \text{ReLU}(\boldsymbol{z})$$

$$\boldsymbol{\theta} = \boldsymbol{U}\boldsymbol{h} + \boldsymbol{b}_2$$

$$\hat{\boldsymbol{y}} = \text{softmax}(\boldsymbol{\theta})$$

$$J = CE(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

Какие градиенты нам нужно посчитать, чтобы обновить веса слоёв?

$$\frac{\partial J}{\partial \boldsymbol{U}} \quad \frac{\partial J}{\partial \boldsymbol{b}_2} \quad \frac{\partial J}{\partial \boldsymbol{W}} \quad \frac{\partial J}{\partial \boldsymbol{b}_1}$$

Какие эффективно посчитать эти градиенты?

Ответ: chain rule

Метод обратного распространения ошибки

$$\boldsymbol{x} = \text{input}$$

$$\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}_1$$

$$\boldsymbol{h} = \text{ReLU}(\boldsymbol{z})$$

$$\boldsymbol{\theta} = \boldsymbol{U}\boldsymbol{h} + \boldsymbol{b}_2$$

$$\hat{\boldsymbol{y}} = \text{softmax}(\boldsymbol{\theta})$$

$$J = CE(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

Какие градиенты нам нужно посчитать, чтобы обновить веса слоёв?

$$\frac{\partial J}{\partial \boldsymbol{U}} \qquad \frac{\partial J}{\partial \boldsymbol{b}_2} \qquad \frac{\partial J}{\partial \boldsymbol{W}} \qquad \frac{\partial J}{\partial \boldsymbol{b}_1}$$

Какие эффективно посчитать эти градиенты?

Ответ: chain rule

$$\frac{\partial J}{\partial \boldsymbol{U}} = \frac{\partial J}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{\theta}} \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{U}}$$

$$\frac{\partial J}{\partial \boldsymbol{b}_2} = \frac{\partial J}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{\theta}} \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{b}_2}$$

Chain rule: $$\frac{d}{dx}\left[f\Big(g(x)\Big)\right] = f'\Big(g(x)\Big)g'(x)$$

https://www.khanacademy.org/math/ap-calculus-ab/ab-differentiation-2-new/ab-3-1a/a/chain-rule-review

Chain rule: $\frac{d}{dx}\left[f\Big(g(x)\Big)\right] = f'\Big(g(x)\Big)g'(x)$

$$h(x) = \underbrace{(\overbrace{5-6x}^{\text{inner}})^5}_{\text{outer}}$$

$g(x) = 5 - 6x$ inner function

$f(x) = x^5$ outer function

https://www.khanacademy.org/math/ap-calculus-ab/ab-differentiation-2-new/ab-3-1a/a/chain-rule-review

Chain rule: $\frac{d}{dx}\left[f\left(g(x)\right)\right] = f'\left(g(x)\right) \cdot g'(x)$

$$h(x) = \underbrace{(\overbrace{5-6x}^{\text{inner}})^5}_{\text{outer}}$$

$g(x) = 5 - 6x$ inner function

$f(x) = x^5$ outer function

https://www.khanacademy.org/math/ap-calculus-ab/ab-differentiation-2-new/ab-3-1a/a/chain-rule-review

Chain rule: $\frac{d}{dx}\left[f\left(g(x)\right)\right] = f'\left(g(x)\right) \cdot g'(x)$

$$h(x) = \underbrace{(\overbrace{5-6x}^{\text{inner}})^5}_{\text{outer}}$$

$g(x) = 5 - 6x$ inner function

$f(x) = x^5$ outer function

$g'(x) = -6$

$f'(x) = 5x^4$

https://www.khanacademy.org/math/ap-calculus-ab/ab-differentiation-2-new/ab-3-1a/a/chain-rule-review

Chain rule: $$\frac{d}{dx}\left[f\Big(g(x)\Big)\right] = f'\Big(g(x)\Big)\cdot g'(x)$$

$$h(x) = \underbrace{(\overbrace{5-6x}^{\text{inner}})^5}_{\text{outer}}$$

$g(x) = 5 - 6x$ inner function

$f(x) = x^5$ outer function

$g'(x) = -6$

$f'(x) = 5x^4$

$$\frac{d}{dx}\left[f\Big(g(x)\Big)\right]$$

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

$$= 5(5-6x)^4 \cdot -6$$

$$= -30(5-6x)^4$$

https://www.khanacademy.org/math/ap-calculus-ab/ab-differentiation-2-new/ab-3-1a/a/chain-rule-review

Метод обратного распространения ошибки

$$\boldsymbol{x} = \text{input}$$

$$\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}_1$$

$$\boldsymbol{h} = \text{ReLU}(\boldsymbol{z})$$

$$\boldsymbol{\theta} = \boldsymbol{U}\boldsymbol{h} + \boldsymbol{b}_2$$

$$\hat{\boldsymbol{y}} = \text{softmax}(\boldsymbol{\theta})$$

$$J = CE(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

Какие градиенты нам нужно посчитать, чтобы обновить веса слоёв?

$$\frac{\partial J}{\partial \boldsymbol{U}} \qquad \frac{\partial J}{\partial \boldsymbol{b}_2} \qquad \frac{\partial J}{\partial \boldsymbol{W}} \qquad \frac{\partial J}{\partial \boldsymbol{b}_1}$$

Какие эффективно посчитать эти градиенты?

Ответ: chain rule

$$\frac{\partial J}{\partial \boldsymbol{U}} = \frac{\partial J}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{\theta}} \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{U}}$$

$$\frac{\partial J}{\partial \boldsymbol{b}_2} = \frac{\partial J}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{\theta}} \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{b}_2}$$

Метод обратного распространения ошибки: одномерный случай

Backpropagation: a simple example

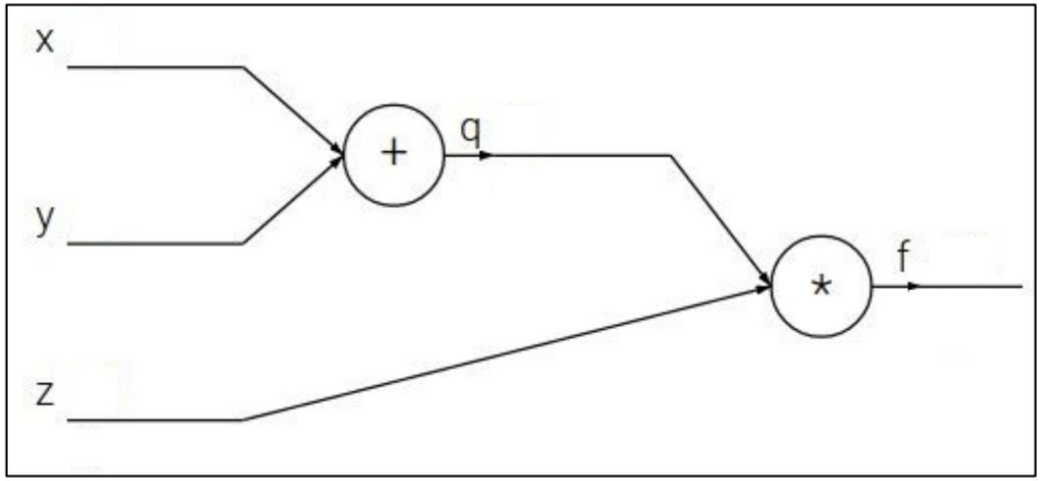
$$f(x, y, z) = (x + y)z$$

http://cs231n.stanford.edu/slides/2020/lecture_4.pdf

Метод обратного распространения ошибки: одномерный случай

Backpropagation: a simple example

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



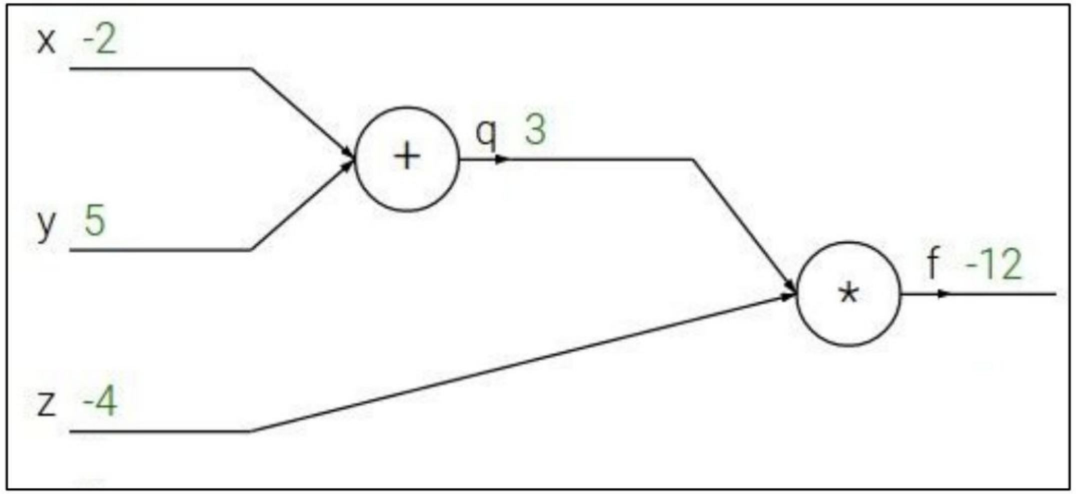
http://cs231n.stanford.edu/slides/2020/lecture_4.pdf

Метод обратного распространения ошибки: одномерный случай

Backpropagation: a simple example

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

e.g. x = -2, y = 5, z = -4



http://cs231n.stanford.edu/slides/2020/lecture_4.pdf

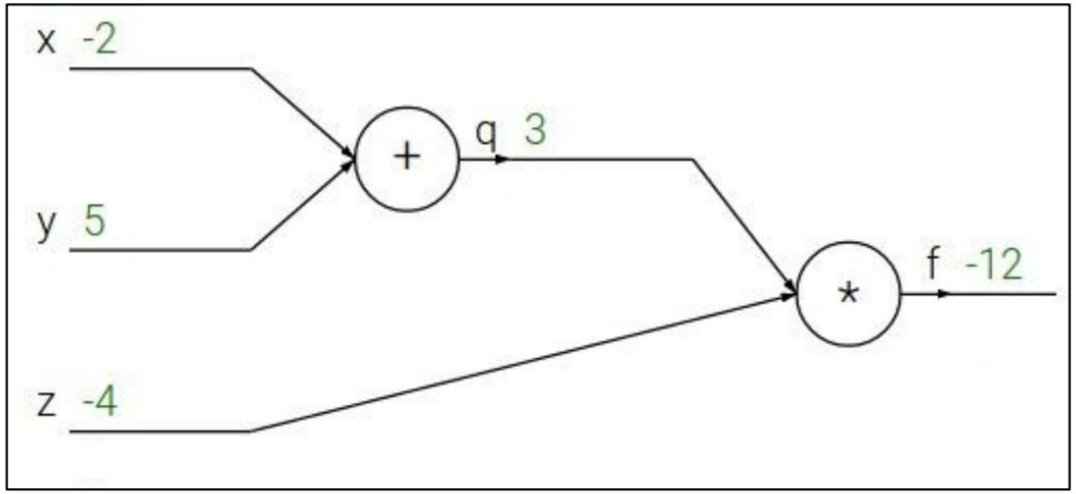
Метод обратного распространения ошибки: одномерный случай

Backpropagation: a simple example

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

e.g. x = -2, y = 5, z = -4

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



http://cs231n.stanford.edu/slides/2020/lecture_4.pdf

Метод обратного распространения ошибки: одномерный случай

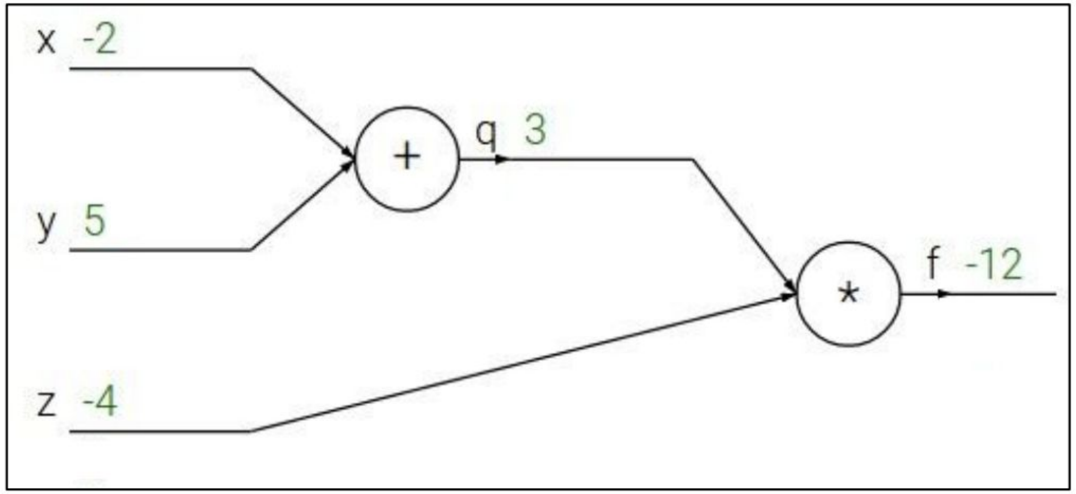
Backpropagation: a simple example

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

e.g. x = -2, y = 5, z = -4

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

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



http://cs231n.stanford.edu/slides/2020/lecture_4.pdf

Метод обратного распространения ошибки: одномерный случай

Backpropagation: a simple example

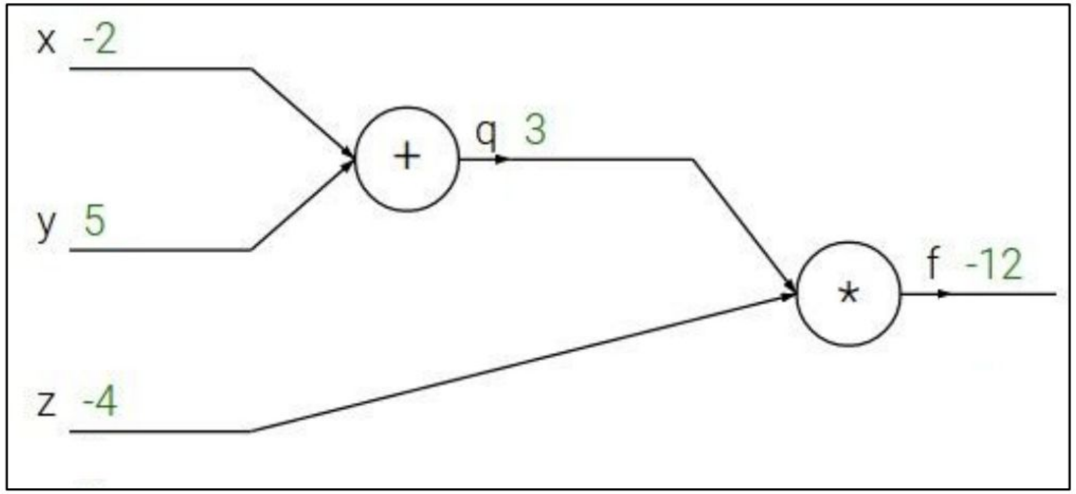
$$f(x, y, z) = (x + y)z$$

e.g. x = -2, y = 5, z = -4

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

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

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



http://cs231n.stanford.edu/slides/2020/lecture_4.pdf

Метод обратного распространения ошибки: одномерный случай

Backpropagation: a simple example

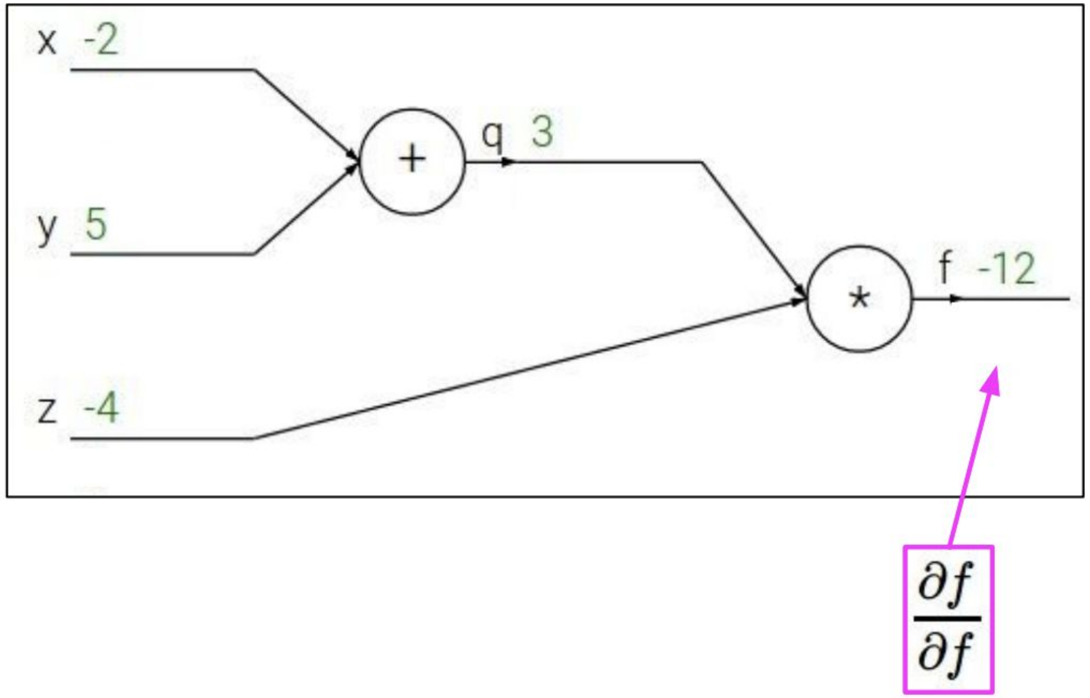
$$f(x, y, z) = (x + y)z$$

e.g. x = -2, y = 5, z = -4

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

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

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



http://cs231n.stanford.edu/slides/2020/lecture_4.pdf

Метод обратного распространения ошибки: одномерный случай

Backpropagation: a simple example

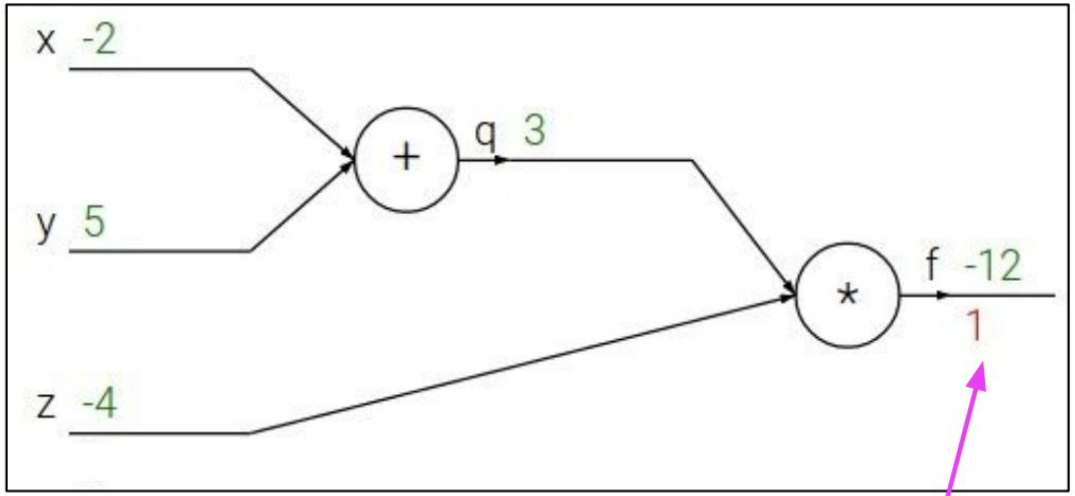
$$f(x, y, z) = (x + y)z$$

e.g. x = -2, y = 5, z = -4

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

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

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



$$\frac{\partial f}{\partial f}$$

http://cs231n.stanford.edu/slides/2020/lecture_4.pdf

Метод обратного распространения ошибки: одномерный случай

Backpropagation: a simple example

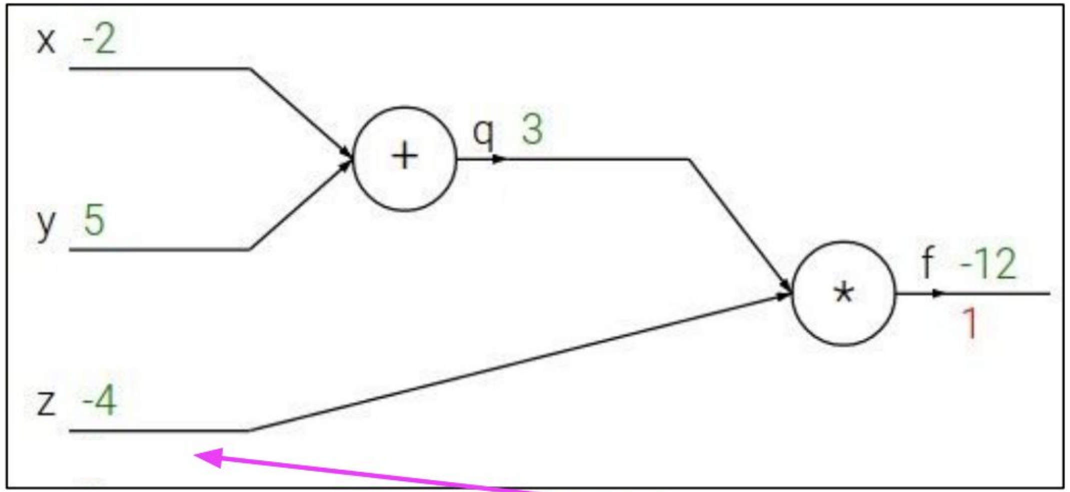
$$f(x, y, z) = (x + y)z$$

e.g. x = -2, y = 5, z = -4

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

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

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



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

http://cs231n.stanford.edu/slides/2020/lecture_4.pdf

Метод обратного распространения ошибки: одномерный случай

Backpropagation: a simple example

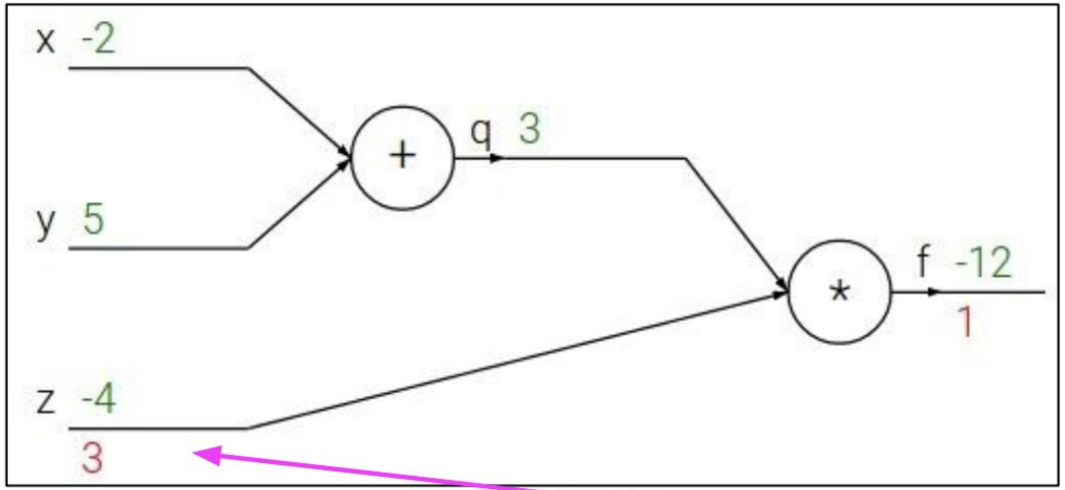
$$f(x, y, z) = (x + y)z$$

e.g. x = -2, y = 5, z = -4

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

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

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



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

http://cs231n.stanford.edu/slides/2020/lecture_4.pdf

Метод обратного распространения ошибки: одномерный случай

Backpropagation: a simple example

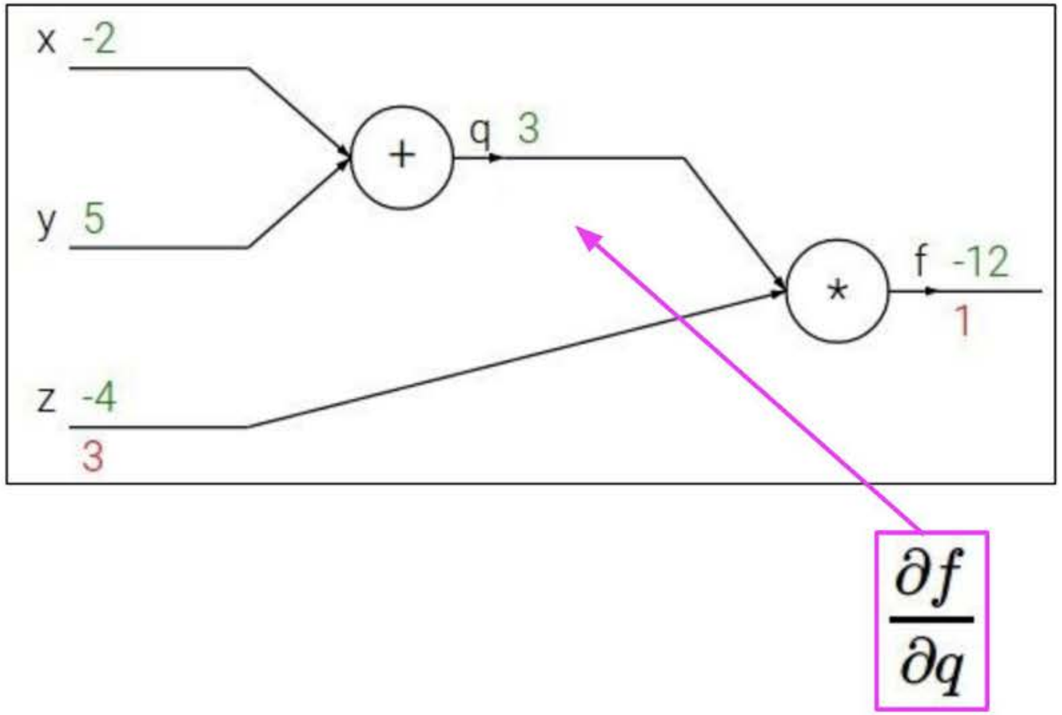
$$f(x, y, z) = (x + y)z$$

e.g. x = -2, y = 5, z = -4

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

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

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



http://cs231n.stanford.edu/slides/2020/lecture_4.pdf

Метод обратного распространения ошибки: одномерный случай

Backpropagation: a simple example

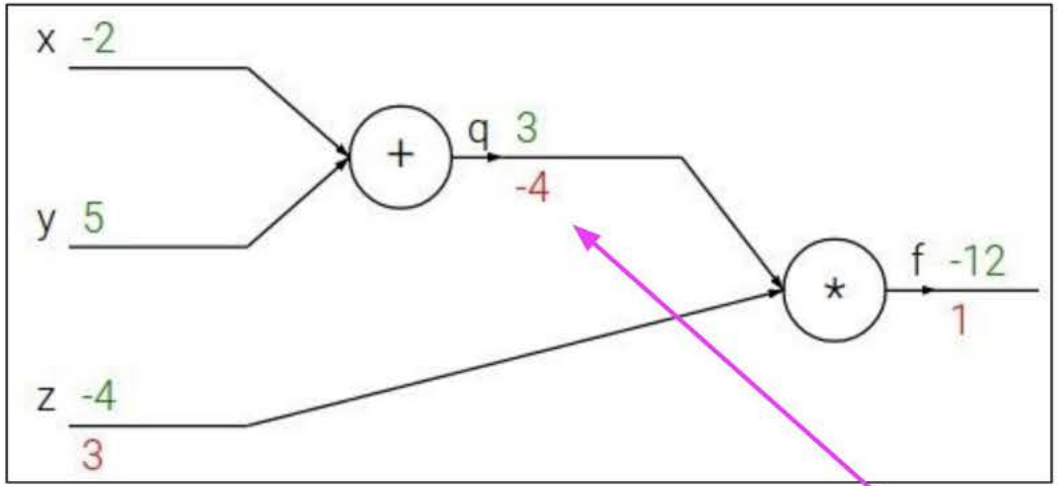
$$f(x, y, z) = (x + y)z$$

e.g. x = -2, y = 5, z = -4

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

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

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



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

http://cs231n.stanford.edu/slides/2020/lecture_4.pdf

Метод обратного распространения ошибки: одномерный случай

Backpropagation: a simple example

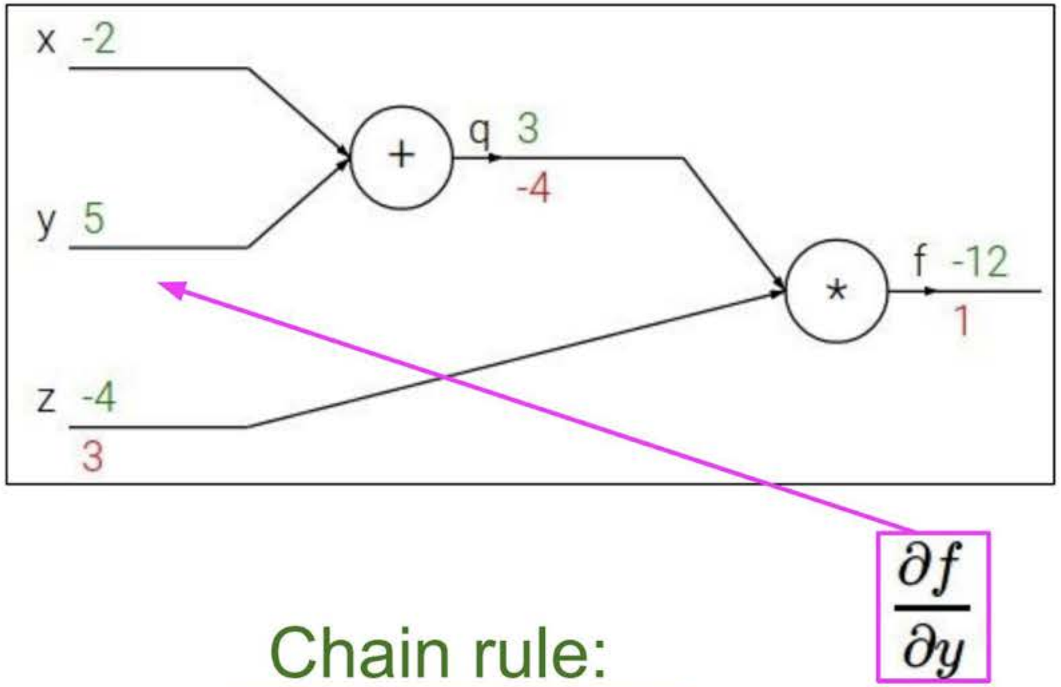
$$f(x, y, z) = (x + y)z$$

e.g. x = -2, y = 5, z = -4

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

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

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



Chain rule:

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

Upstream gradient

Local gradient

http://cs231n.stanford.edu/slides/2020/lecture_4.pdf

Метод обратного распространения ошибки: одномерный случай

Backpropagation: a simple example

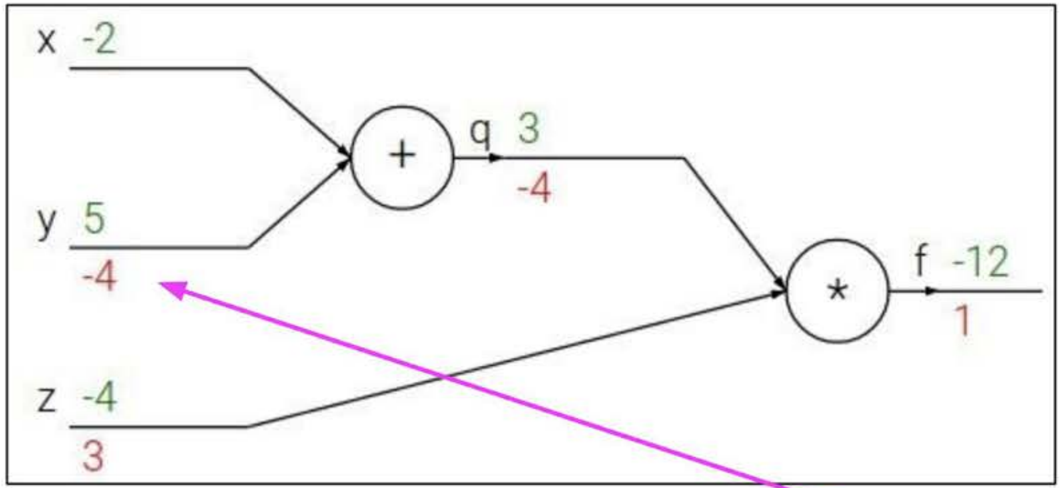
$$f(x, y, z) = (x + y)z$$

e.g. x = -2, y = 5, z = -4

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

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

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



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

Chain rule:

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

Upstream gradient

Local gradient

http://cs231n.stanford.edu/slides/2020/lecture_4.pdf

Метод обратного распространения ошибки: одномерный случай

Backpropagation: a simple example

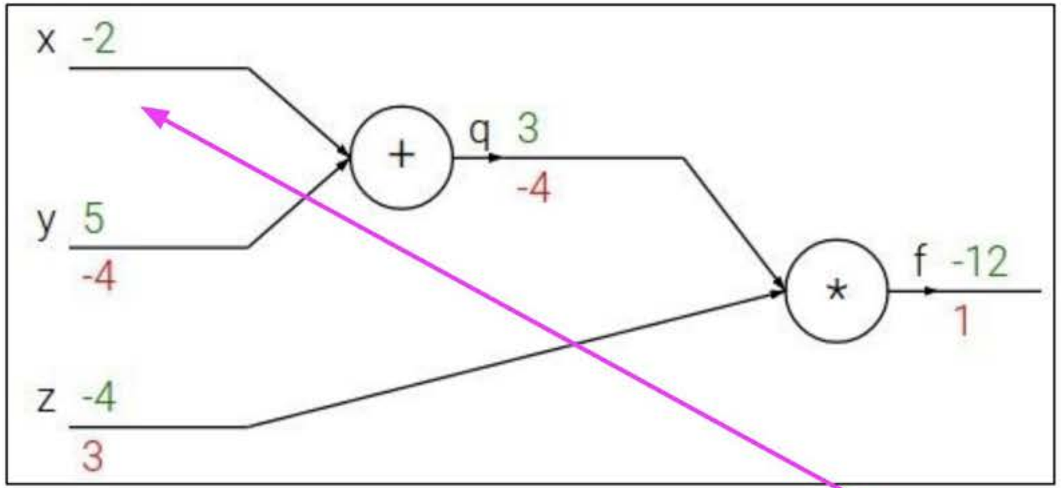
$$f(x, y, z) = (x + y)z$$

e.g. x = -2, y = 5, z = -4

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

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

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



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

Chain rule:

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

Upstream gradient

Local gradient

http://cs231n.stanford.edu/slides/2020/lecture_4.pdf

Метод обратного распространения ошибки: одномерный случай

Backpropagation: a simple example

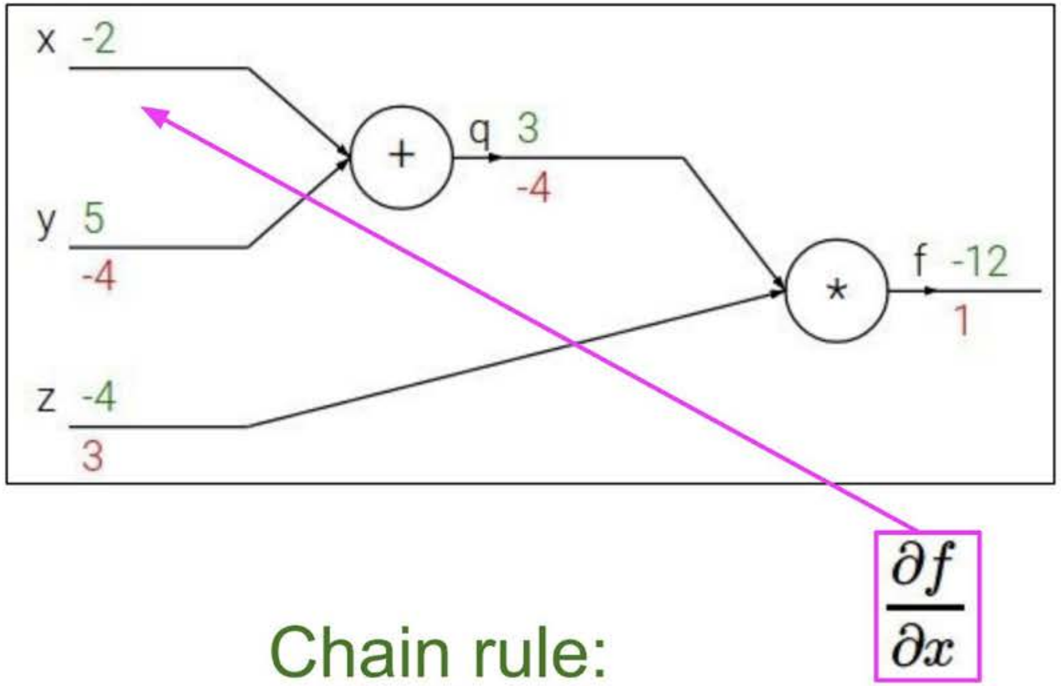
$$f(x, y, z) = (x + y)z$$

e.g. x = -2, y = 5, z = -4

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

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

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



Chain rule:

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

Upstream gradient — $\frac{\partial f}{\partial q}$

Local gradient — $\frac{\partial q}{\partial x}$

http://cs231n.stanford.edu/slides/2020/lecture_4.pdf

Метод обратного распространения ошибки: одномерный случай

Backpropagation: a simple example

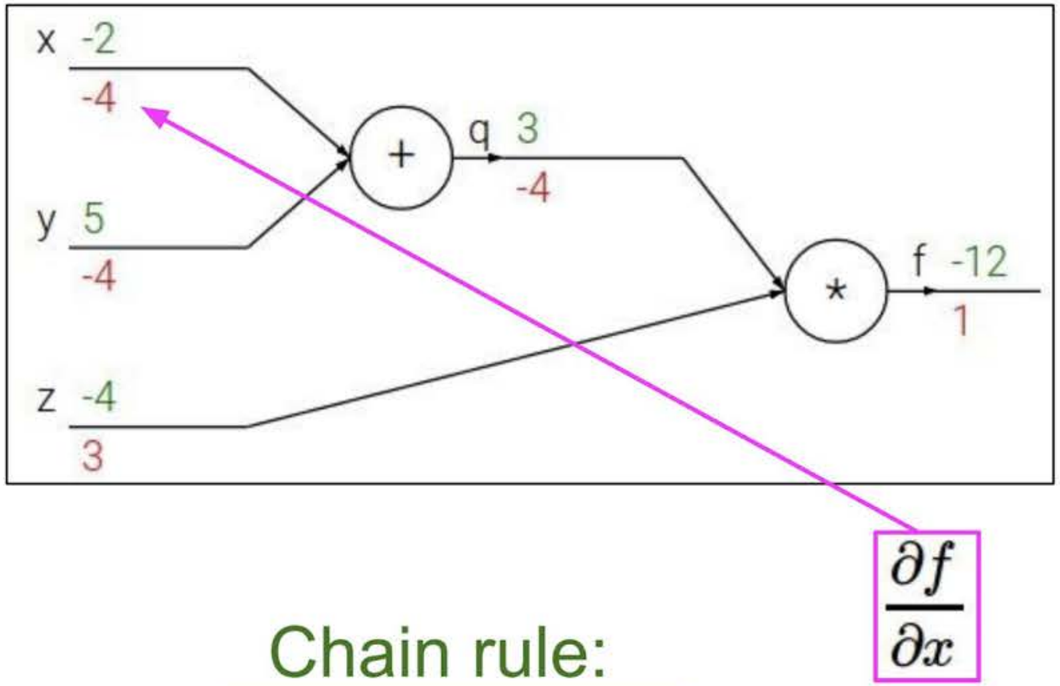
$$f(x, y, z) = (x + y)z$$

e.g. x = -2, y = 5, z = -4

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

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

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



Chain rule:

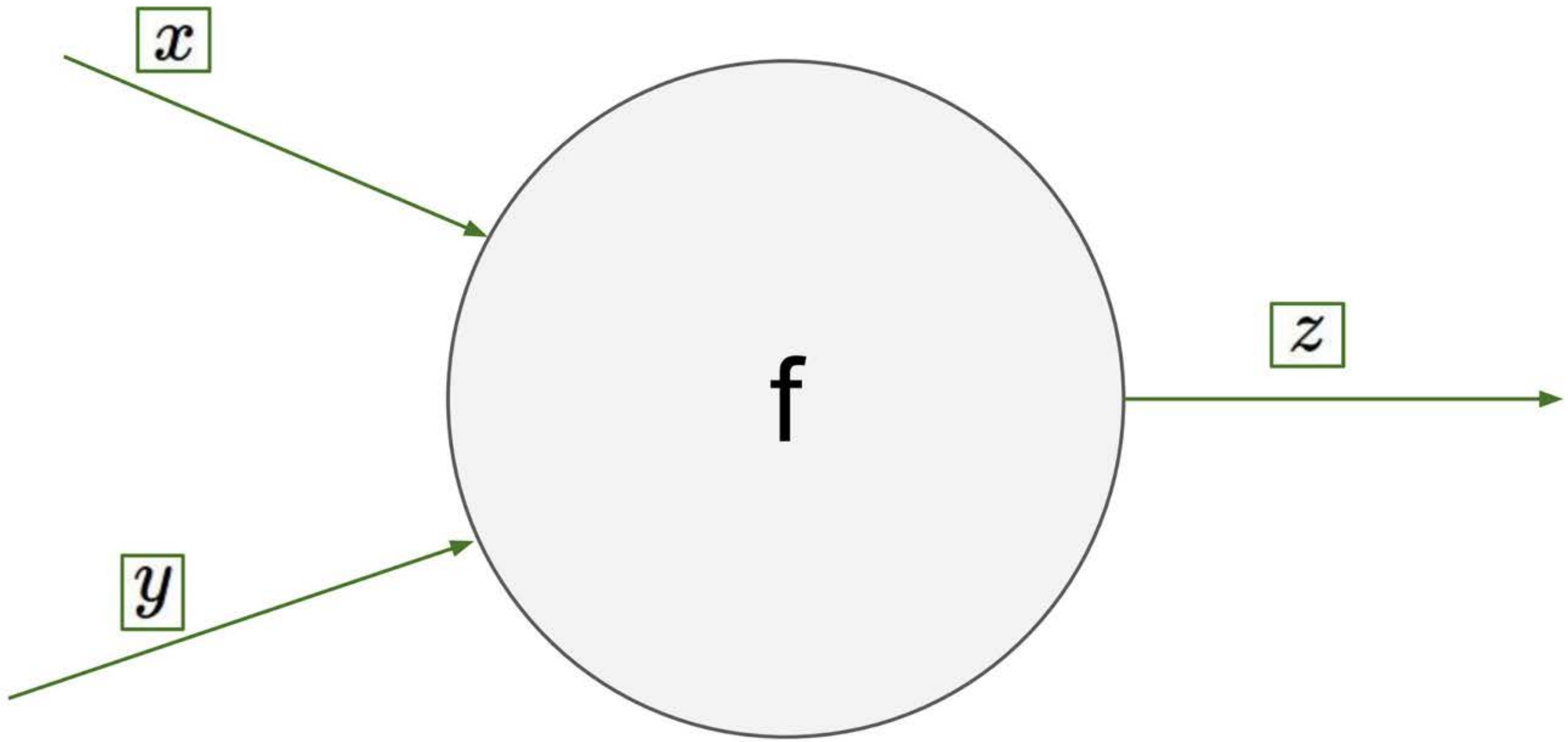
$$\boxed{\frac{\partial f}{\partial x} = \frac{\partial f}{\partial q} \frac{\partial q}{\partial x}}$$

Upstream gradient

Local gradient

http://cs231n.stanford.edu/slides/2020/lecture_4.pdf

Кэширование градиентов при форвард пассе

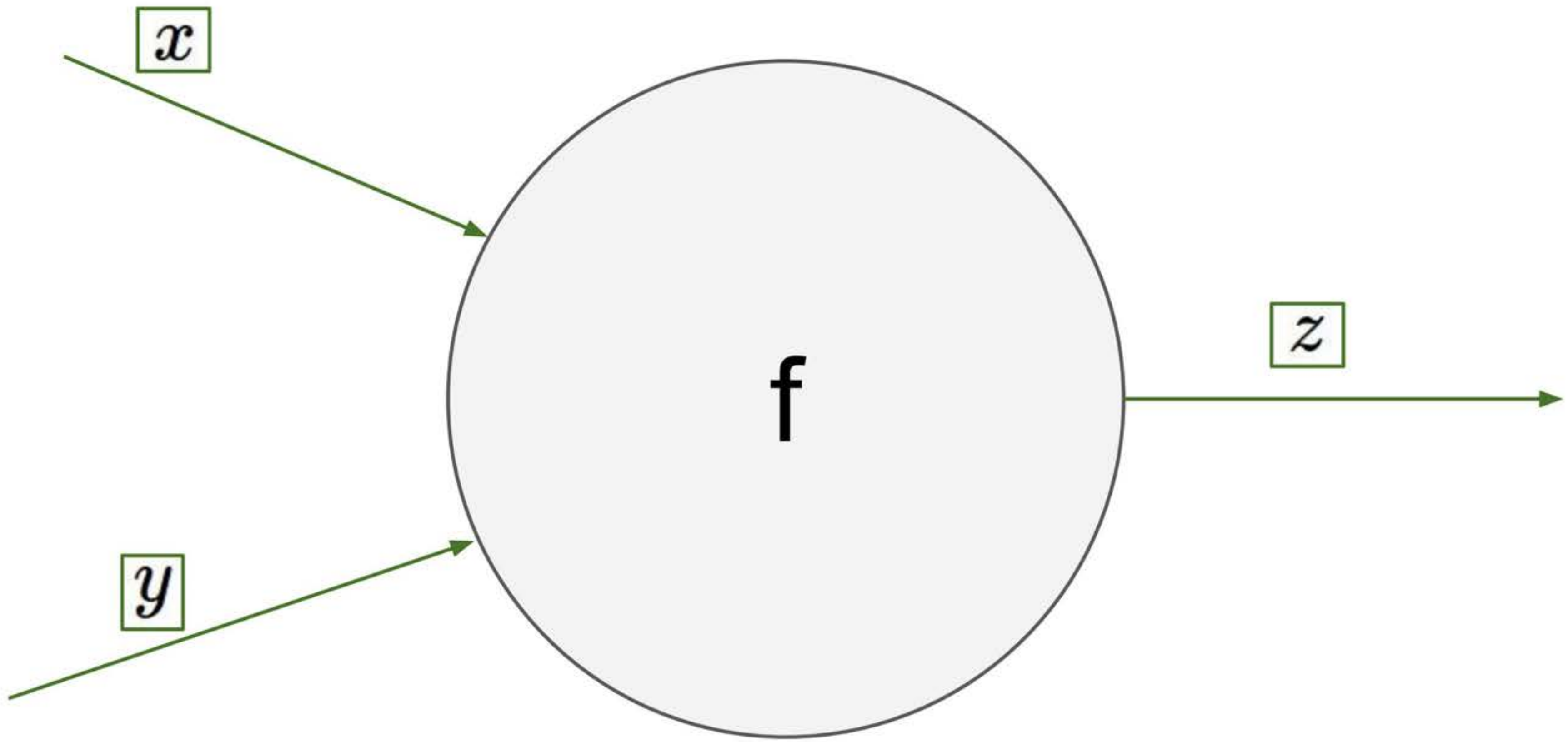


http://cs231n.stanford.edu/slides/2020/lecture_4.pdf

Кэширование градиентов при форвард пассе

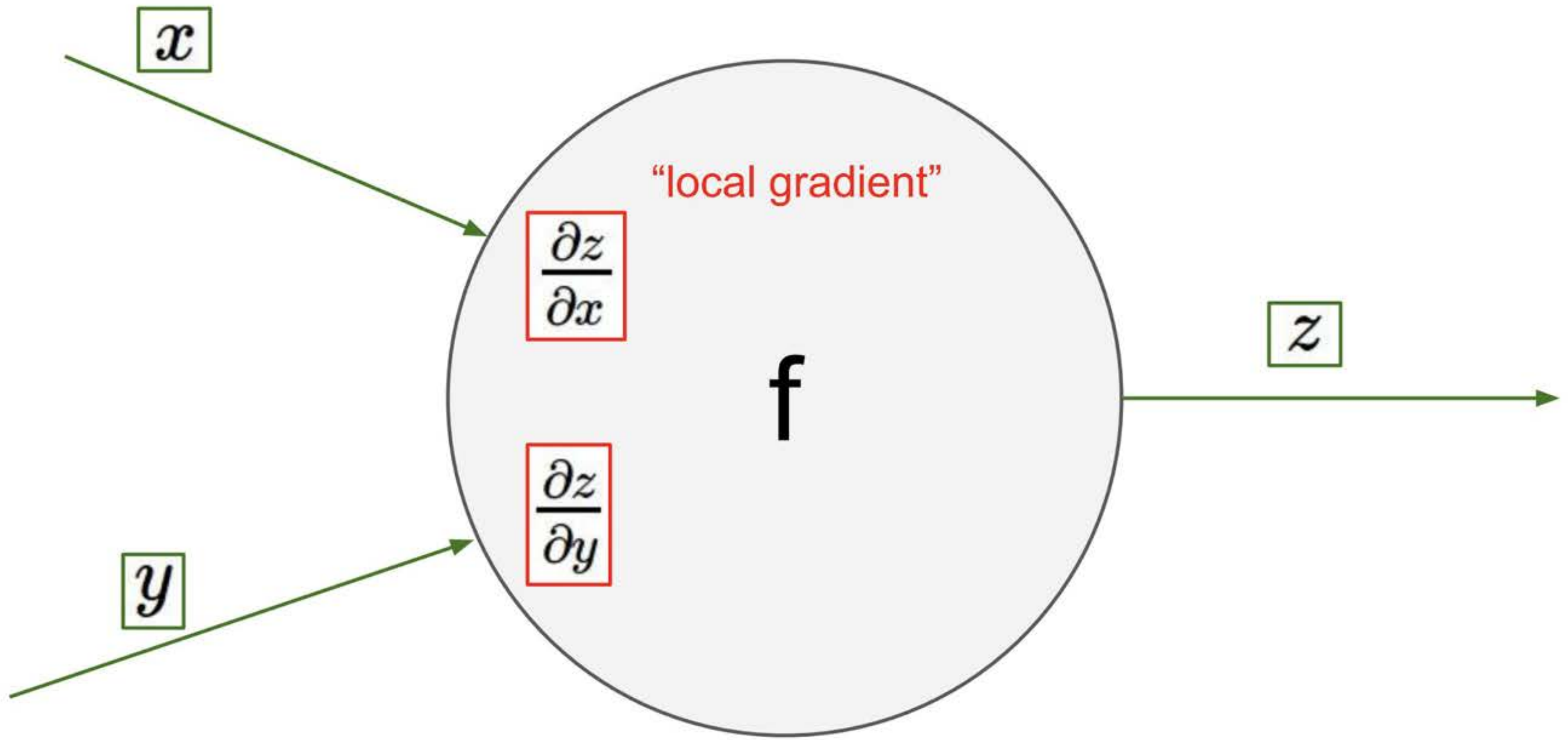
http://cs231n.stanford.edu/slides/2020/lecture_4.pdf

Кэширование градиентов при форвард пассе



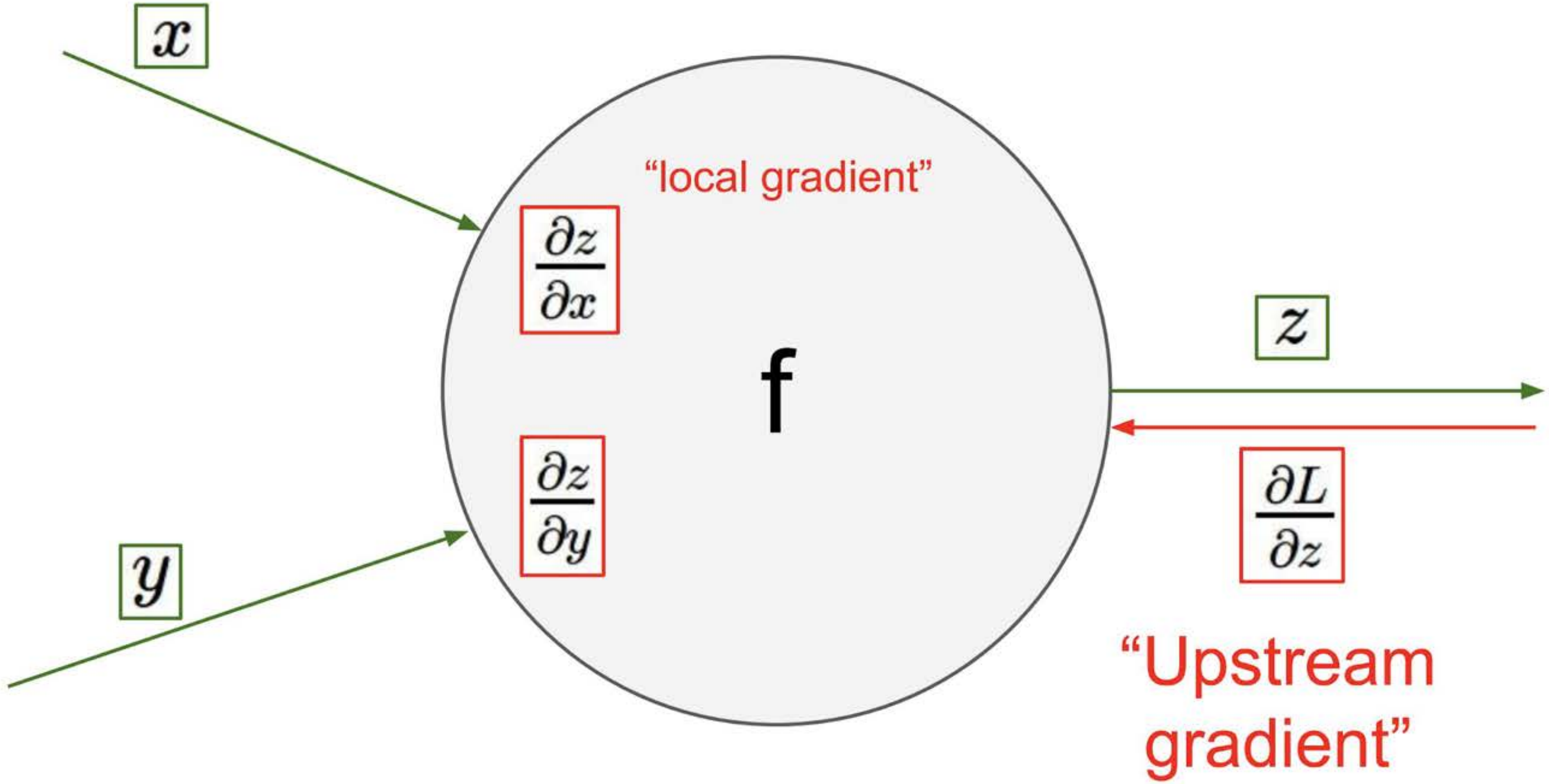
http://cs231n.stanford.edu/slides/2020/lecture_4.pdf

Кэширование градиентов при форвард пассе



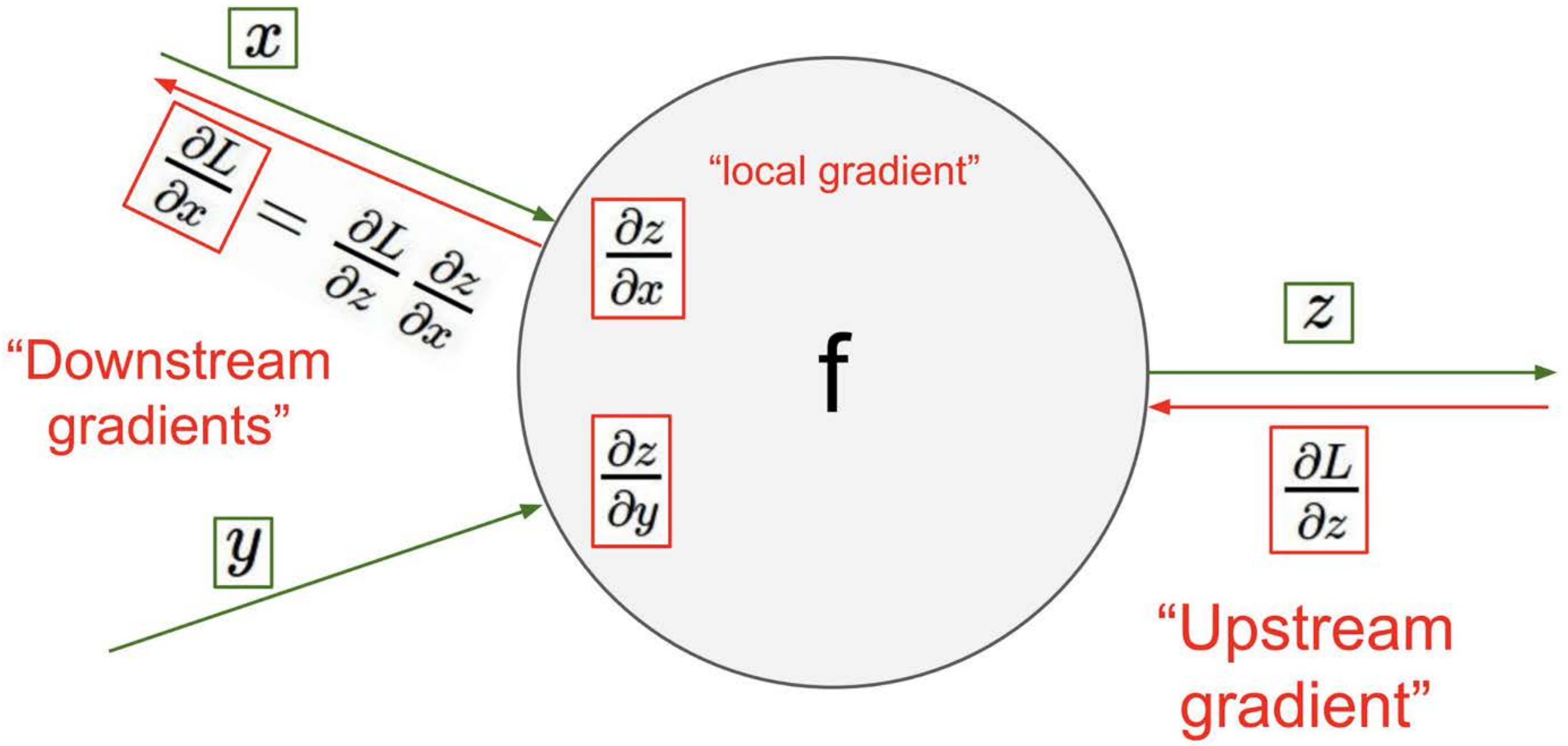
http://cs231n.stanford.edu/slides/2020/lecture_4.pdf

Кэширование градиентов при форвард пассе



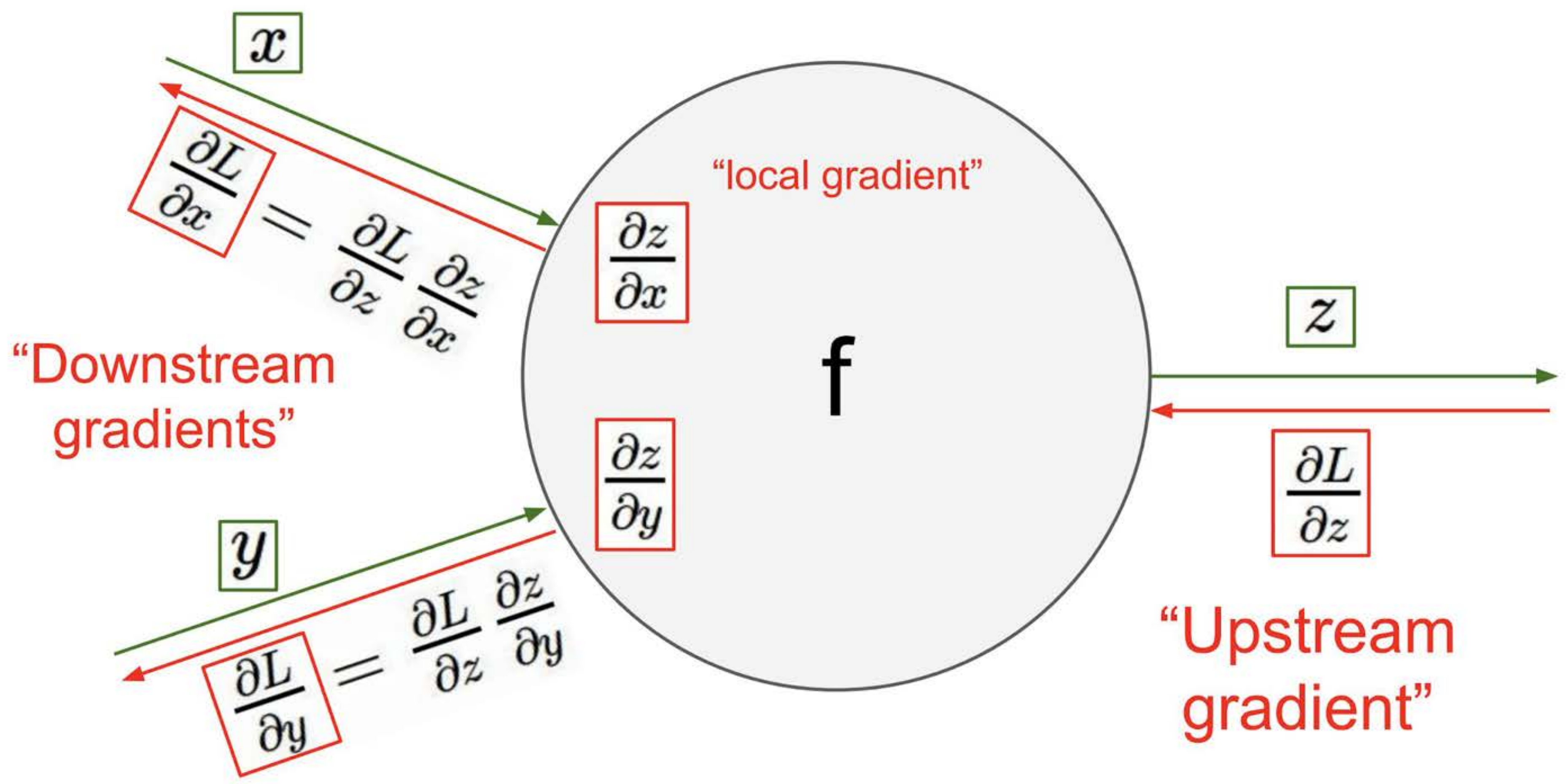
http://cs231n.stanford.edu/slides/2020/lecture_4.pdf

Кэширование градиентов при форвард пассе



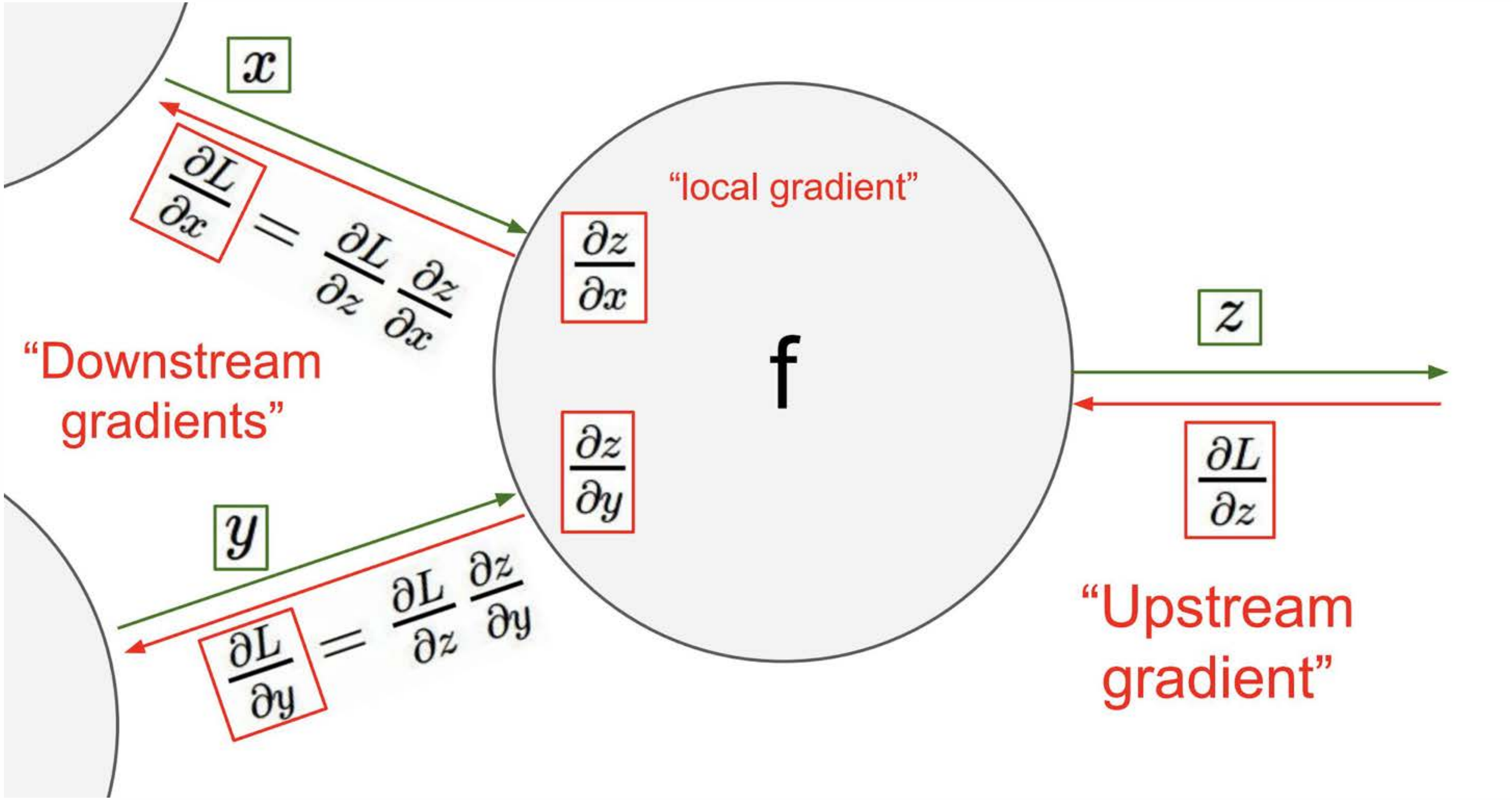
http://cs231n.stanford.edu/slides/2020/lecture_4.pdf

Кэширование градиентов при форвард пассе



http://cs231n.stanford.edu/slides/2020/lecture_4.pdf

Кэширование градиентов при форвард пассе



http://cs231n.stanford.edu/slides/2020/lecture_4.pdf

Forward pass. Входной вектор: ***h***

$$\boldsymbol{\theta} = \boldsymbol{U}\boldsymbol{h} + \boldsymbol{b}_2$$

$$\hat{\boldsymbol{y}} = \mathrm{softmax}(\boldsymbol{\theta})$$

$$J = CE(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Forward pass: step 1

$$\boxed{\boldsymbol{\theta} = \boldsymbol{U}\boldsymbol{h} + \boldsymbol{b}_2}$$

$$\hat{\boldsymbol{y}} = \mathrm{softmax}(\boldsymbol{\theta})$$

$$J = CE(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Forward pass: step 1. Beca

$$\boldsymbol{\theta} = \boxed{\boldsymbol{U}}\boldsymbol{h} + \boxed{\boldsymbol{b}_2}$$

$$\hat{\boldsymbol{y}} = \mathrm{softmax}(\boldsymbol{\theta})$$

$$J = CE(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Forward pass: логирование градиентов

$$\boldsymbol{\theta} = \boxed{\boldsymbol{U}}\boldsymbol{h} + \boxed{\boldsymbol{b}_2}$$
$$\hat{\boldsymbol{y}} = \operatorname{softmax}(\boldsymbol{\theta})$$
$$J = CE(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

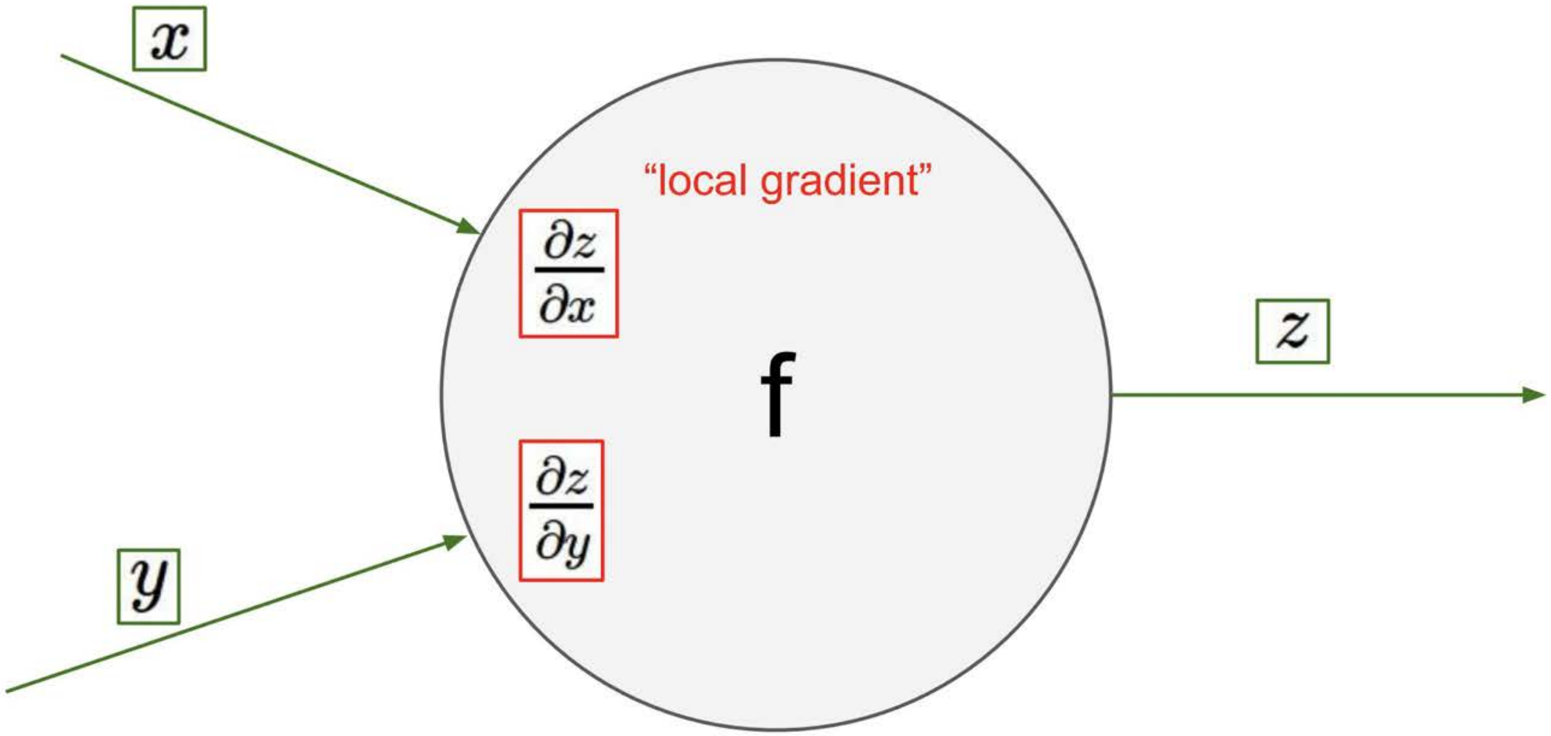
$$\boxed{\begin{array}{c}\dfrac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{U}} \\ \dfrac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{b}_2}\end{array}}$$

локальные градиенты

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Forward pass: логирование градиентов

$$\boldsymbol{\theta} = \boxed{\boldsymbol{U}}\boldsymbol{h} + \boxed{\boldsymbol{b}_2}$$



$$\boxed{\begin{matrix}\dfrac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{U}} \\ \dfrac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{b}_2}\end{matrix}}$$

локальные градиенты

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Forward pass: step 2

$$\boldsymbol{\theta} = \boldsymbol{U}\boldsymbol{h} + \boldsymbol{b}_2$$

$$\boxed{\hat{\boldsymbol{y}} = \text{softmax}(\boldsymbol{\theta})}$$

$$J = CE(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Forward pass: логирование градиентов

$$\boldsymbol{\theta} = \boldsymbol{U}\boldsymbol{h} + \boldsymbol{b}_2$$

$$\boxed{\hat{\boldsymbol{y}} = \mathrm{softmax}(\boldsymbol{\theta})}$$

$$J = CE(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

$$\boxed{\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{\theta}}}$$

локальные градиенты

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Forward pass: step 3

$$\boldsymbol{\theta} = \boldsymbol{U}\boldsymbol{h} + \boldsymbol{b}_2$$

$$\hat{\boldsymbol{y}} = \mathrm{softmax}(\boldsymbol{\theta})$$

$$\boxed{J = CE(\boldsymbol{y}, \hat{\boldsymbol{y}})}$$

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Forward pass: логирование градиентов

$$\boldsymbol{\theta} = \boldsymbol{U}\boldsymbol{h} + \boldsymbol{b}_2$$

$$\hat{\boldsymbol{y}} = \mathrm{softmax}(\boldsymbol{\theta})$$

$$\boxed{J = CE(\boldsymbol{y}, \hat{\boldsymbol{y}})}$$

$$\boxed{\frac{\partial J}{\partial \hat{\boldsymbol{y}}}}$$

локальные градиенты

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Backward pass

$$\boldsymbol{\theta} = \boldsymbol{U}\boldsymbol{h} + \boldsymbol{b}_2$$

$$\hat{\boldsymbol{y}} = \text{softmax}(\boldsymbol{\theta})$$

$$J = CE(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

$$\boxed{\frac{\partial J}{\partial \hat{\boldsymbol{y}}}}$$

локальные градиенты

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Backward pass

$$\boldsymbol{\theta} = \boldsymbol{U}\boldsymbol{h} + \boldsymbol{b}_2$$

$$\hat{\boldsymbol{y}} = \text{softmax}(\boldsymbol{\theta})$$

$$J = CE(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

$$\frac{\partial J}{\partial \boldsymbol{U}} =$$

$$\frac{\partial J}{\partial \boldsymbol{b}_2} =$$

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Backward pass

$$\boldsymbol{\theta} = \boldsymbol{U}\boldsymbol{h} + \boldsymbol{b}_2$$
$$\hat{\boldsymbol{y}} = \mathrm{softmax}(\boldsymbol{\theta})$$
$$J = CE(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

$$\frac{\partial J}{\partial \boldsymbol{U}} = \boxed{\frac{\partial J}{\partial \hat{\boldsymbol{y}}}}$$
$$\frac{\partial J}{\partial \boldsymbol{b}_2} = \boxed{\frac{\partial J}{\partial \hat{\boldsymbol{y}}}}$$

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Backward pass

$$\boldsymbol{\theta} = \boldsymbol{U}\boldsymbol{h} + \boldsymbol{b}_2$$

$$\hat{\boldsymbol{y}} = \mathrm{softmax}(\boldsymbol{\theta})$$

$$J = CE(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

$$\frac{\partial J}{\partial \boldsymbol{U}} = \frac{\partial J}{\partial \hat{\boldsymbol{y}}} \boxed{\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{\theta}}}$$

$$\frac{\partial J}{\partial \boldsymbol{b}_2} = \frac{\partial J}{\partial \hat{\boldsymbol{y}}} \boxed{\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{\theta}}}$$

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Backward pass

$$\boldsymbol{\theta} = \boldsymbol{U}\boldsymbol{h} + \boldsymbol{b}_2$$

$$\hat{\boldsymbol{y}} = \mathrm{softmax}(\boldsymbol{\theta})$$

$$J = CE(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

$$\frac{\partial J}{\partial \boldsymbol{U}} = \frac{\partial J}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{\theta}} \boxed{\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{U}}}$$

$$\frac{\partial J}{\partial \boldsymbol{b}_2} = \frac{\partial J}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{\theta}} \boxed{\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{b}_2}}$$

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Backward pass. Размерности

$$\boldsymbol{\theta} = \boldsymbol{U}\boldsymbol{h} + \boldsymbol{b}_2$$
$$\hat{\boldsymbol{y}} = \mathrm{softmax}(\boldsymbol{\theta})$$
$$J = CE(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

$$\frac{\partial J}{\partial \boldsymbol{U}} = \frac{\partial J}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{\theta}} \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{U}}$$
$$\frac{\partial J}{\partial \boldsymbol{b}_2} = \frac{\partial J}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{\theta}} \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{b}_2}$$

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Backward pass. Размерности

$$\boldsymbol{\theta} = \boldsymbol{U}\boldsymbol{h} + \boldsymbol{b}_2$$
$$\hat{\boldsymbol{y}} = \mathrm{softmax}(\boldsymbol{\theta})$$
$$J = CE(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

$$\frac{\partial J}{\partial \boldsymbol{U}} = \frac{\partial J}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{\theta}} \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{U}}$$
$$\frac{\partial J}{\partial \boldsymbol{b}_2} = \frac{\partial J}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{\theta}} \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{b}_2}$$

$$\boldsymbol{b}_2 \in \mathbb{R}^{N_c \times 1}$$

$$\boldsymbol{U} \in \mathbb{R}^{N_c \times D_h}$$

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Backward pass. Одномерный случай

Backpropagation: a simple example

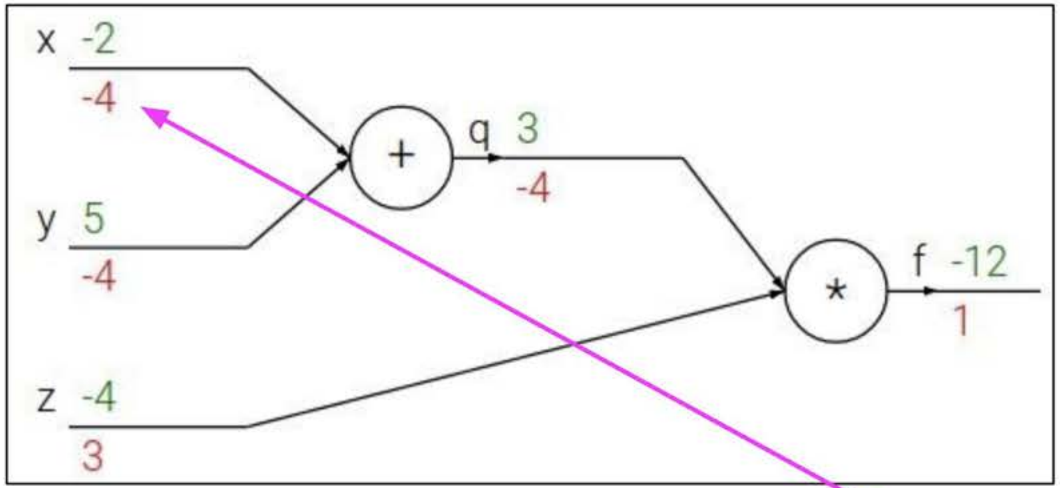
$$f(x,y,z) = (x+y)z$$

e.g. x = -2, y = 5, z = -4

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

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

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



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

Chain rule:

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

Upstream gradient

Local gradient

Backward pass. Двумерный случай

Backward pass. Двумерный случай

$$f : \mathbb{R}^n \to \mathbb{R}^m$$

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Backward pass. Двумерный случай

$$\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$$

$$\boldsymbol{f}(\boldsymbol{x}) = [f_1(x_1, ..., x_n), f_2(x_1, ..., x_n), ..., f_m(x_1, ..., x_n)]$$

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Backward pass. Двумерный случай

$$f(x) = \begin{bmatrix} w01 & w02 & w03 \\ w11 & w12 & w13 \\ w21 & w22 & w23 \\ w31 & w32 & w33 \end{bmatrix} \times \begin{bmatrix} x0 \\ x1 \\ x2 \end{bmatrix}$$

$$\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$$

$$\boldsymbol{f}(\boldsymbol{x}) = [f_1(x_1, ..., x_n), f_2(x_1, ..., x_n), ..., f_m(x_1, ..., x_n)]$$

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Backward pass. Двумерный случай

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} w01\ w02\ w03 \\ w11\ w12\ w13 \\ w21\ w22\ w23 \\ w31\ w32\ w33 \end{bmatrix} \times \begin{bmatrix} x0 \\ x1 \\ x2 \end{bmatrix}$$

$$\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$$

$$\boldsymbol{f}(\boldsymbol{x}) = [f_1(x_1, ..., x_n), f_2(x_1, ..., x_n), ..., f_m(x_1, ..., x_n)]$$

$$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Якобиан

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} w01 & w02 & w03 \\ w11 & w12 & w13 \\ w21 & w22 & w23 \\ w31 & w32 & w33 \end{bmatrix} \times \begin{bmatrix} x0 \\ x1 \\ x2 \end{bmatrix}$$

$$\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$$

$$\boldsymbol{f}(\boldsymbol{x}) = [f_1(x_1, \ldots, x_n), f_2(x_1, \ldots, x_n), \ldots, f_m(x_1, \ldots, x_n)]$$

$$\boxed{\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}}$$

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Якобиан

$$f(x) = \begin{bmatrix} w01 & w02 & w03 \\ w11 & w12 & w13 \\ w21 & w22 & w23 \\ w31 & w32 & w33 \end{bmatrix} \times \begin{bmatrix} x0 \\ x1 \\ x2 \end{bmatrix}$$

$$\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$$

$$\boldsymbol{f}(\boldsymbol{x}) = [f_1(x_1, \ldots, x_n), f_2(x_1, \ldots, x_n), \ldots, f_m(x_1, \ldots, x_n)]$$

$$\boxed{\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}}$$

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Backward pass. Двумерный случай. Chain rule

$$\boldsymbol{f}(x) = [f_1(x), f_2(x)]$$

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Backward pass. Двумерный случай. Chain rule

$$\boldsymbol{f}(x) = [f_1(x), f_2(x)]$$

$$\boldsymbol{g}(\boldsymbol{y}) = [g_1(y_1, y_2), g_2(y_1, y_2)]$$

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Backward pass. Двумерный случай. Chain rule

$$\boldsymbol{f}(x) = [f_1(x), f_2(x)]$$

$$\boldsymbol{g}(\boldsymbol{y}) = [g_1(y_1, y_2), g_2(y_1, y_2)]$$

$$\boldsymbol{g}(x) = [g_1(f_1(x), f_2(x)), g_2(f_1(x), f_2(x))]$$

$$h(x) = \underbrace{(\overbrace{5 - 6x}^{\text{inner}})^5}_{\text{outer}}$$

$g(x) = 5 - 6x$ inner function

$f(x) = x^5$ outer function

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Backward pass. Двумерный случай. Chain rule

$$\boldsymbol{f}(x) = [f_1(x), f_2(x)]$$

$$\boldsymbol{g}(\boldsymbol{y}) = [g_1(y_1, y_2), g_2(y_1, y_2)]$$

$$\boldsymbol{g}(x) = [g_1(f_1(x), f_2(x)), g_2(f_1(x), f_2(x))]$$

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Якобиан

$$\boldsymbol{f}(x) = [f_1(x), f_2(x)]$$

$$\boldsymbol{g}(\boldsymbol{y}) = [g_1(y_1, y_2), g_2(y_1, y_2)]$$

$$\boldsymbol{g}(x) = [g_1(f_1(x), f_2(x)), g_2(f_1(x), f_2(x))]$$

$$\frac{\partial \boldsymbol{g}}{\partial x} =$$

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Якобиан

$$\boldsymbol{f}(x) = [f_1(x), f_2(x)]$$

$$\boldsymbol{g}(\boldsymbol{y}) = [g_1(y_1, y_2), g_2(y_1, y_2)]$$

$$\boldsymbol{g}(x) = [g_1(f_1(x), f_2(x)), g_2(f_1(x), f_2(x))]$$

$$\frac{\partial \boldsymbol{g}}{\partial x} = \begin{bmatrix} \frac{\partial}{\partial x} g_1(f_1(x), f_2(x)) \\ \frac{\partial}{\partial x} g_2(f_1(x), f_2(x)) \end{bmatrix}$$

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Якобиан

$$\boldsymbol{f}(x) = [f_1(x), f_2(x)]$$

$$\boldsymbol{g}(\boldsymbol{y}) = [g_1(y_1, y_2), g_2(y_1, y_2)]$$

$$\boldsymbol{g}(x) = [g_1(f_1(x), f_2(x)), g_2(f_1(x), f_2(x))]$$

$$\frac{\partial \boldsymbol{g}}{\partial x} = \begin{bmatrix} \frac{\partial}{\partial x} g_1(f_1(x), f_2(x)) \\ \frac{\partial}{\partial x} g_2(f_1(x), f_2(x)) \end{bmatrix} = \begin{bmatrix} \frac{\partial g_1}{\partial f_1}\frac{\partial f_1}{\partial x} + \frac{\partial g_1}{\partial f_2}\frac{\partial f_2}{\partial x} \\ \frac{\partial g_2}{\partial f_1}\frac{\partial f_1}{\partial x} + \frac{\partial g_2}{\partial f_2}\frac{\partial f_2}{\partial x} \end{bmatrix}$$

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Повторение: Якобиан

$$\frac{\partial \boldsymbol{g}}{\partial x} = \begin{bmatrix} \frac{\partial}{\partial x} g_1(f_1(x), f_2(x)) \\ \frac{\partial}{\partial x} g_2(f_1(x), f_2(x)) \end{bmatrix} = \begin{bmatrix} \frac{\partial g_1}{\partial f_1}\frac{\partial f_1}{\partial x} + \frac{\partial g_1}{\partial f_2}\frac{\partial f_2}{\partial x} \\ \frac{\partial g_2}{\partial f_1}\frac{\partial f_1}{\partial x} + \frac{\partial g_2}{\partial f_2}\frac{\partial f_2}{\partial x} \end{bmatrix}$$

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Якобиан

$$\frac{\partial \boldsymbol{g}}{\partial x} = \begin{bmatrix} \frac{\partial}{\partial x} g_1(f_1(x), f_2(x)) \\ \frac{\partial}{\partial x} g_2(f_1(x), f_2(x)) \end{bmatrix} = \begin{bmatrix} \frac{\partial g_1}{\partial f_1}\frac{\partial f_1}{\partial x} + \frac{\partial g_1}{\partial f_2}\frac{\partial f_2}{\partial x} \\ \frac{\partial g_2}{\partial f_1}\frac{\partial f_1}{\partial x} + \frac{\partial g_2}{\partial f_2}\frac{\partial f_2}{\partial x} \end{bmatrix}$$

$$\frac{\partial \boldsymbol{g}}{\partial x} = \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{f}}\frac{\partial \boldsymbol{f}}{\partial x} = \begin{bmatrix} \frac{\partial g_1}{\partial f_1} & \frac{\partial g_1}{\partial f_2} \\ \frac{\partial g_2}{\partial f_1} & \frac{\partial g_2}{\partial f_2} \end{bmatrix} \begin{bmatrix} \frac{\partial f_1}{\partial x} \\ \frac{\partial f_2}{\partial x} \end{bmatrix}$$

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Якобиан

$$\frac{\partial \boldsymbol{g}}{\partial x} = \begin{bmatrix} \frac{\partial}{\partial x} g_1(f_1(x), f_2(x)) \\ \frac{\partial}{\partial x} g_2(f_1(x), f_2(x)) \end{bmatrix} = \begin{bmatrix} \frac{\partial g_1}{\partial f_1}\frac{\partial f_1}{\partial x} + \frac{\partial g_1}{\partial f_2}\frac{\partial f_2}{\partial x} \\ \frac{\partial g_2}{\partial f_1}\frac{\partial f_1}{\partial x} + \frac{\partial g_2}{\partial f_2}\frac{\partial f_2}{\partial x} \end{bmatrix}$$

$$\frac{\partial \boldsymbol{g}}{\partial x} = \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{f}}\frac{\partial \boldsymbol{f}}{\partial x} = \begin{bmatrix} \frac{\partial g_1}{\partial f_1} & \frac{\partial g_1}{\partial f_2} \\ \frac{\partial g_2}{\partial f_1} & \frac{\partial g_2}{\partial f_2} \end{bmatrix} \begin{bmatrix} \frac{\partial f_1}{\partial x} \\ \frac{\partial f_2}{\partial x} \end{bmatrix}$$

$$\frac{d}{dx}\left[f\Big(g(x)\Big)\right] = f'\Big(g(x)\Big) \cdot g'(x)$$

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Якобиан

$$\frac{\partial \boldsymbol{g}}{\partial x} = \begin{bmatrix} \frac{\partial}{\partial x} g_1(f_1(x), f_2(x)) \\ \frac{\partial}{\partial x} g_2(f_1(x), f_2(x)) \end{bmatrix} = \begin{bmatrix} \frac{\partial g_1}{\partial f_1}\frac{\partial f_1}{\partial x} + \frac{\partial g_1}{\partial f_2}\frac{\partial f_2}{\partial x} \\ \frac{\partial g_2}{\partial f_1}\frac{\partial f_1}{\partial x} + \frac{\partial g_2}{\partial f_2}\frac{\partial f_2}{\partial x} \end{bmatrix}$$

$$\frac{\partial \boldsymbol{g}}{\partial x} = \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{f}}\frac{\partial \boldsymbol{f}}{\partial x} = \begin{bmatrix} \frac{\partial g_1}{\partial f_1} & \frac{\partial g_1}{\partial f_2} \\ \frac{\partial g_2}{\partial f_1} & \frac{\partial g_2}{\partial f_2} \end{bmatrix} \begin{bmatrix} \frac{\partial f_1}{\partial x} \\ \frac{\partial f_2}{\partial x} \end{bmatrix}$$

$$\frac{d}{dx}\left[f\Big(g(x)\Big)\right] = f'\Big(g(x)\Big) \cdot g'(x)$$

$$\frac{d}{dx}\left[f\Big(g(x)\Big)\right]$$

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

$$= 5(5-6x)^4 \cdot -6$$

$$= -30(5-6x)^4$$

http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf

Backprop. Пример расчета градиента

$$\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x}$$

Backprop. Пример расчета градиента

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}}?$$

$$\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x}$$

Backprop. Пример расчета градиента

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}}?$$

$$\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x}$$

$$z_i = \sum_{k=1}^{m} W_{ik} x_k$$

Backprop. Пример расчета градиента

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}}?$$

$$\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x}$$

$$z_i = \sum_{k=1}^{m} W_{ik} x_k$$

$$\left(\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}}\right)_{ij} =$$

Backprop. Пример расчета градиента

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}}?$$

$$\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x}$$

$$z_i = \sum_{k=1}^{m} W_{ik} x_k$$

$$\left(\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}}\right)_{ij} = \frac{\partial z_i}{\partial x_j} =$$

Backprop. Пример расчета градиента $\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}}$?

$$\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x} \qquad z_i = \sum_{k=1}^{m} W_{ik} x_k$$

$$\left(\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}}\right)_{ij} = \frac{\partial z_i}{\partial x_j} = \frac{\partial}{\partial x_j} \sum_{k=1}^{m} W_{ik} x_k =$$

Backprop. Пример расчета градиента

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}}?$$

$$\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x} \qquad z_i = \sum_{k=1}^{m} W_{ik} x_k$$

$$\left(\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}}\right)_{ij} = \frac{\partial z_i}{\partial x_j} = \frac{\partial}{\partial x_j} \sum_{k=1}^{m} W_{ik} x_k = \sum_{k=1}^{m} W_{ik} \frac{\partial}{\partial x_j} x_k =$$

Backprop. Пример расчета градиента

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}}?$$

$$\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x}$$

$$z_i = \sum_{k=1}^{m} W_{ik} x_k$$

$$\left(\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}}\right)_{ij} = \frac{\partial z_i}{\partial x_j} = \frac{\partial}{\partial x_j} \sum_{k=1}^{m} W_{ik} x_k = \sum_{k=1}^{m} W_{ik} \frac{\partial}{\partial x_j} x_k = W_{ij}$$

Backprop. Пример расчета градиента

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}}?$$

$$\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x} \qquad z_i = \sum_{k=1}^{m} W_{ik} x_k$$

$$\left(\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}}\right)_{ij} = \frac{\partial z_i}{\partial x_j} = \frac{\partial}{\partial x_j} \sum_{k=1}^{m} W_{ik} x_k = \sum_{k=1}^{m} W_{ik} \frac{\partial}{\partial x_j} x_k = W_{ij}$$

$$\boxed{\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} = \boldsymbol{W}}$$

Резюме

- Полносвязные сети (Fully-connected neural networks) - сеть из N блоков, где каждый блок состоит из линейной операции (умножение на матрицу) и нелинейной функции активации
- Примеры нелинейных функций: tanh, sigmoid, softmax, relu
- Backpropagation - рекурсивный алгоритм подсчета градиентов в сети, использующий chain rule
- forward pass - подсчет выходов из каждого слоя сети
- backward pass - подсчет градиентов и обновление весов сети

В следующих сериях

- Лекция о разных алгоритмах оптимизации и регуляризации сетей
- Семинар: введение в Pytorch, полносвязная сеть на Pytorch

Полезные ссылки

- Лекция по backpropagation, CS231n, CV with Deep Learning: https://www.youtube.com/watch?v=i94OvYb6noo
- Лекция по backpropagation, CS224n, NLP with Deep Learning: https://www.youtube.com/watch?v=yLYHDSv-288
- Затеханный семинар по backprop из Стэнфорда: http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf