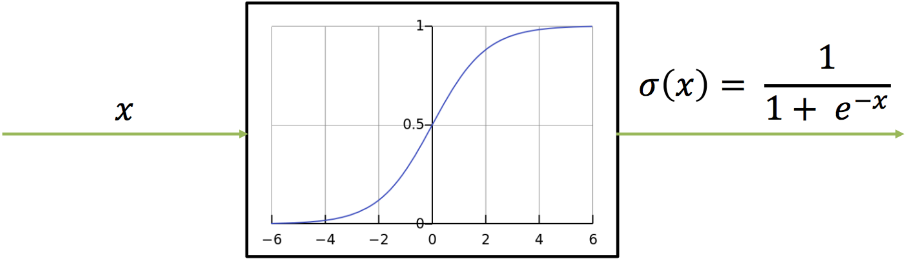
Глубокое обучение

Занятие 3: эвристики для обучения нейросеток

Какими бывают функции активации и как через них пробросить градиент

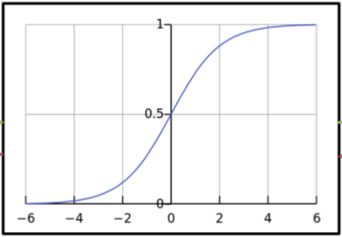
Sigmoid activation



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

Sigmoid activation

x



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

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial \sigma} \frac{\partial \sigma}{\partial x}$$

$$\frac{\partial L}{\partial \sigma}$$

$$\frac{\partial \sigma}{\partial x} = \sigma(x)(1 - \sigma(x))$$

Паралич сети

- В случае сигмоиды $\sigma'(x) = \sigma(x) \cdot (1 - \sigma(x))$
- Сигмоида принимает значения на отрезке $[0;1]$, значит максимальное значение её производной это $\frac{1}{4}$
- Если сеть очень глубокая, происходит затухание градиента
- Градиент затухает экспоненциально $\Rightarrow$ сходимость замедляется, более ранние веса обновляются дольше, более глубокие веса быстрее $\Rightarrow$ значение градиента становится ещё меньше $\Rightarrow$ наступает паралич сети
- В сетях с небольшим числом слоёв этот эффект незаметен

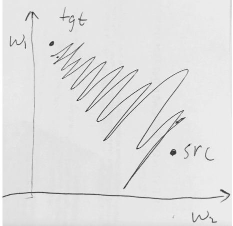
Центрирование

- Сигмоида не центрирована относительно нуля
- Выход слоя мы обычно находим как $o_i = \sigma(h_i)$, он всегда положительный, значит вектор градиента по весам, идущим на вход в текущий нейрон, тоже положительный $\Rightarrow$ они веса обновляются в одинаковом направлении (либо все уменьшаются, либо все увеличиваются)
- Сходимость идёт медленнее и зигзагообразно, но идёт

Центрирование

$$f = \sum w_i x_i + b$$
$$\frac{df}{dw_i} = x_i$$
$$\frac{dL}{dw_i} = \frac{dL}{df}\frac{df}{dw_i} = \frac{dL}{df}x_i$$

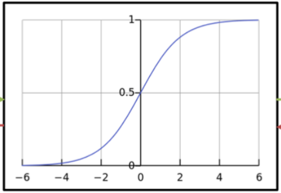
because $x_i > 0,$ the gradient $\frac{dL}{dw_i}$ always has the same sign as $\frac{dL}{df}$ (all positive or all negative). For every w_i the sign is the same



Источник: Cross Validated

Sigmoid activation

x



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

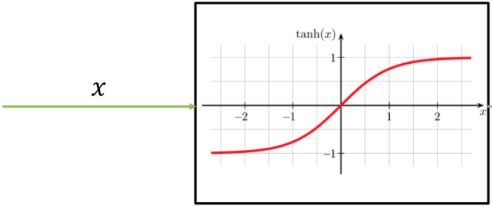
$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial \sigma}\frac{\partial \sigma}{\partial x}$$

$$\frac{\partial L}{\partial \sigma}$$

$$\frac{\partial \sigma}{\partial x} = \sigma(x)(1 - \sigma(x))$$

- Способствует затуханию градиента
- Не центрирована относительно нуля
- Вычислять e^x дорого

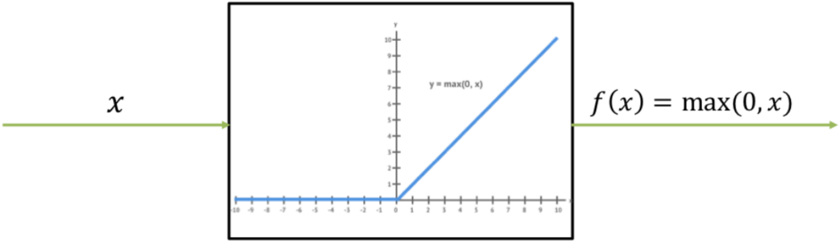
Tanh activation



$$\tanh(x) = \frac{2}{1 + e^{-2x}} - 1$$

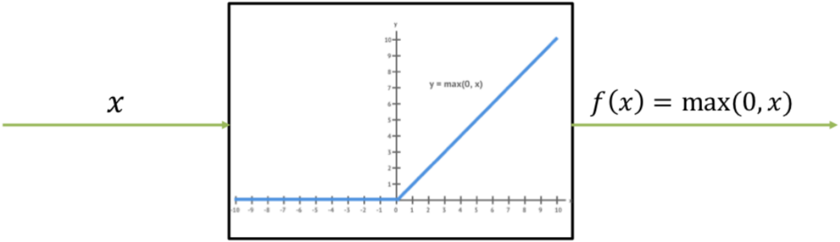
- Центрирован относительно нуля
- Всё ещё похож на сигмоиду
- $f'(x) = 1 - f(x)^2 \Rightarrow$ затухание градиента

ReLU activation



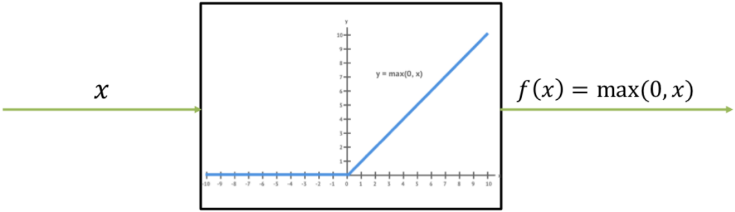
- Быстро вычисляется
- Градиент не затухает
- Сходимость сеток ускоряется

ReLU activation



- Сетка может умереть, если активация занулится на всех нейронах
- Не центрирован относительно нуля

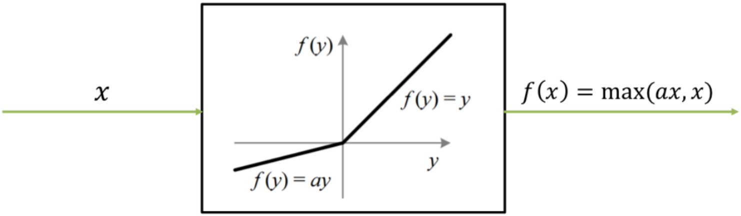
Зануление ReLU



- $f(x) = \text{max}(0, w_0 + w_1 \cdot h_1 + ... + w_k \cdot h_k)$

- Если w_0 инициализировано большим отрицательным числом, нейрон сразу умирает $\Rightarrow$ надо аккуратно инициализировать веса

Leaky ReLU activation



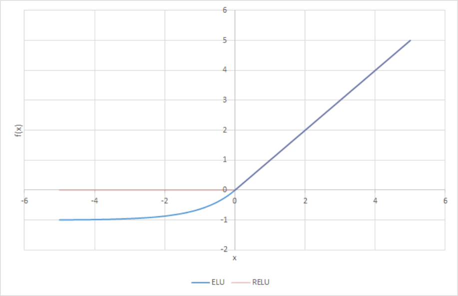
- Как ReLU, но не умирает, всё ещё легко считается
- Производная может быть любого знака
- Важно, чтобы $a \neq 1$, иначе линейность

Что же выбрать

- Обычно начинают с $ReLU$, если сетка умирает, берут $LeakyReLU$
- $ReLU$ — стандартный выбор для свёрточных сетей
- В рекурентных сетках чаще всего предпочитается $tanh$
- На самом деле это не очень важно, нужно держать в голове свойства функций, о которых выше шла речь и понимать, что от перебора функций обычно выигрыш в качестве довольно низкий
- Но есть и исключения ...

Краткий обзор функций активаций: https://arxiv.org/pdf/1804.02763.pdf

ELU activation



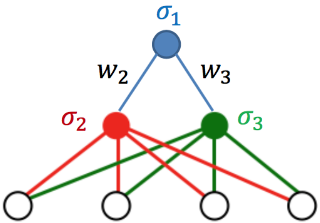
- ELU улучшает сходимость для глубоких сеток

$$f(x) = \begin{cases} x, x \geq 0 \\ \alpha \cdot (e^x - 1), x < 0 \end{cases}$$

https://arxiv.org/pdf/1511.07289.pdf

Инициализация весов

Инициализация весов

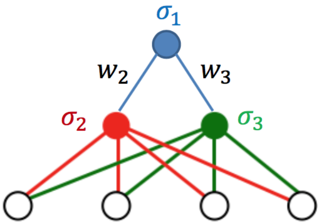


$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \sigma_1}\sigma_1(1-\sigma_1)\sigma_2$$

$$\frac{\partial L}{\partial w_3} = \frac{\partial L}{\partial \sigma_1}\sigma_1(1-\sigma_1)\sigma_3$$

- Что будет, если инициализировать веса нулями?

Инициализация весов



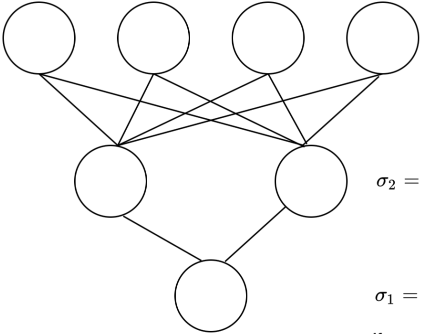
$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \sigma_1}\sigma_1(1-\sigma_1)\sigma_2$$

$$\frac{\partial L}{\partial w_3} = \frac{\partial L}{\partial \sigma_1}\sigma_1(1-\sigma_1)\sigma_3$$

- Что будет, если инициализировать веса нулями?

- W_2 и W_3 будут обновляться одинаково

Инициализация весов



$$\sigma_2 = \sigma(W_1 X + b_1)$$

$$\sigma_1 = f(W_2 \sigma_2 + b_2)$$

$$\frac{\partial L}{\partial W_2} = \frac{\partial L}{\partial \sigma_1} \sigma_1 (1 - \sigma_1) \sigma_2$$

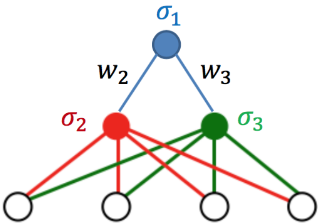
$$W_2 = W_2 - \alpha \frac{\partial L}{\partial W_2}$$

- If you have sigmoid activation $g(0) \neq 0$ g(0)≠0 then it will cause weights to move "together", limiting the power of back-propagation to search the entire space

- If you have tanhtanh or ReLu activation $g(0) = 0$ g(0)=0 then all the outputs will be 0, and the gradients for the weights will always be 0.

Если веса инициализируются нулями, то выходы слоев σ_1, σ_2 либо нулевые, либо константные, что ведет к такому же обновлению весов

Инициализация весов



$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \sigma_1}\sigma_1(1-\sigma_1)\sigma_2$$

$$\frac{\partial L}{\partial w_3} = \frac{\partial L}{\partial \sigma_1}\sigma_1(1-\sigma_1)\sigma_3$$

- Хочется уничтожить симметрию
- Обычно инициализируют маленькими рандомными числами из какого-то распределения (нормальное, равномерное)

Инициализация весов

- Наши признаки X пришли к нам из какого-то распределения
- Выход слоя $f(XW)$ будет принадлежать другому распределению
- Если инициализировать веса неправильно, дисперсия распределения може от слоя к слою затухать (сигнал будет теряться) либо наоброт, возрастать (сигнал будет рассеиваться)
- Эмпирически было выяснено, что это может портить сходимость для глубоких сеток
- Хочется контролировать дисперсию

Инициализация весов

- Посмотрим на выход нейрона перед активацией:

$$h_i = w_0 + \sum_{i=1}^{n_{in}} w_i x_i$$

- Дисперсия h_i выражается через дисперсии x и w
- Она не зависит от константы w_0
- Будем считать, что веса $w_1, \ldots, w_k \sim iid$, наблюдения $x_1, \ldots, x_n \sim iid$, а ещё x_i и w_i независимы между собой

Инициализация весов (симметричный случай)

$$\mathsf{Var}(h_i) = \mathsf{Var}(\sum_{i=1}^{n_{in}} w_i x_i) = \sum_{i=1}^{n_{in}} \mathsf{Var}(w_i x_i) =$$

$$= \sum_{i=1}^{n_{in}} \Big([\mathsf{E}(x_i)]^2 \cdot \mathsf{Var}(w_i) + [\mathsf{E}(w_i)]^2 \cdot \mathsf{Var}(x_i) + \mathsf{Var}(x_i) \cdot \mathsf{Var}(w_i) \Big) =$$

Инициализация весов (симметричный случай)

$$\mathsf{Var}(h_i) = \mathsf{Var}(\sum_{i=1}^{n_{in}} w_i x_i) = \sum_{i=1}^{n_{in}} \mathsf{Var}(w_i x_i) =$$

$$= \sum_{i=1}^{n_{in}} \Big([\mathsf{E}(x_i)]^2 \cdot \mathsf{Var}(w_i) + [\mathsf{E}(w_i)]^2 \cdot \mathsf{Var}(x_i) + \mathsf{Var}(x_i) \cdot \mathsf{Var}(w_i) \Big) =$$

- Если функция активации симметричная ($f(z) + f(-z) = 1$), тогда $E(x_i) = 0$.
- Будем инициализировать веса с нулевым средним, тогда $E(w_i) = 0$.

Инициализация весов (симметричный случай)

$$\mathsf{Var}(h_i) = \mathsf{Var}(\sum_{i=1}^{n_{in}} w_i x_i) = \sum_{i=1}^{n_{in}} \mathsf{Var}(w_i x_i) =$$

$$= \sum_{i=1}^{n_{in}} \Big([\mathsf{E}(x_i)]^2 \cdot \mathsf{Var}(w_i) + [\mathsf{E}(w_i)]^2 \cdot \mathsf{Var}(x_i) + \mathsf{Var}(x_i) \cdot \mathsf{Var}(w_i) \Big) =$$

$$= \sum_{i=1}^{n_{in}} \mathsf{Var}(x_i) \cdot \mathsf{Var}(w_i)$$

- Если функция активации симметричная ($f(z) + f(-z) = 1$), тогда $E(x_i) = 0$.
- Будем инициализировать веса с нулевым средним, тогда $E(w_i) = 0$.

Инициализация весов (симметричный случай)

$$\text{Var}(h_i) = \text{Var}(\sum_{i=1}^{n_{in}} w_i x_i) = \sum_{i=1}^{n_{in}} \text{Var}(w_i x_i) =$$

$$= \sum_{i=1}^{n_{in}} \Big([\text{E}(x_i)]^2 \cdot \text{Var}(w_i) + [\text{E}(w_i)]^2 \cdot \text{Var}(x_i) + \text{Var}(x_i) \cdot \text{Var}(w_i) \Big) =$$

$$= \sum_{i=1}^{n_{in}} \text{Var}(x_i) \cdot \text{Var}(w_i) = \text{Var}(x) \cdot [n_{in} \cdot \text{Var}(w)]$$

Инициализация весов (симметричный случай)

$$\text{Var}(h_i) = \text{Var}(\sum_{i=1}^{n_{in}} w_i x_i) = \sum_{i=1}^{n_{in}} \text{Var}(w_i x_i) =$$

$$= \sum_{i=1}^{n_{in}} \Big([\text{E}(x_i)]^2 \cdot \text{Var}(w_i) + [\text{E}(w_i)]^2 \cdot \text{Var}(x_i) + \text{Var}(x_i) \cdot \text{Var}(w_i) \Big) =$$

$$= \sum_{i=1}^{n_{in}} \text{Var}(x_i) \cdot \text{Var}(w_i) = \text{Var}(x) \cdot \underbrace{[n_{in} \cdot \text{Var}(w)]}_{=1}$$

Плохая инициализация весов

Пущай

$$w_i \sim U\left[-\frac{1}{\sqrt{n_{in}}}; \frac{1}{\sqrt{n_{in}}}\right],$$

тогда

$$\text{Var}(w_i) = \frac{1}{12} \cdot \left(\frac{1}{\sqrt{n_{in}}} + \frac{1}{\sqrt{n_{in}}}\right)^2 = \frac{1}{3n_{in}} \Rightarrow Var(h_i) = \frac{\text{Var}(x)}{3}$$

Получаем затухание!

Немного лучше

Пущай

$$w_i \sim U\left[-\frac{\sqrt{3}}{\sqrt{n_{in}}}; \frac{\sqrt{3}}{\sqrt{n_{in}}}\right],$$

тогда

$$\mathsf{Var}(w_i) = \frac{1}{12} \cdot \left(\frac{\sqrt{3}}{\sqrt{n_{in}}} + \frac{\sqrt{3}}{\sqrt{n_{in}}}\right)^2 = \frac{1}{n_{in}} \Rightarrow Var(h_i) = \mathsf{Var}(x)$$

Немного лучше

Пущай

$$w_i \sim U\left[-\frac{\sqrt{3}}{\sqrt{n_{in}}}; \frac{\sqrt{3}}{\sqrt{n_{in}}}\right],$$

тогда

$$\text{Var}(w_i) = \frac{1}{12} \cdot \left(\frac{\sqrt{3}}{\sqrt{n_{in}}} + \frac{\sqrt{3}}{\sqrt{n_{in}}}\right)^2 = \frac{1}{n_{in}} \Rightarrow Var(h_i) = \text{Var}(x)$$

При forward pass на вход идёт n_{in} наблюдений, при backward pass на вход идёт n_{out} градиентов $\Rightarrow$ канал с дисперсией может быть непостоянным, если число весов от слоя к слою сильно колеблется

Инициализация Xavier Glorot

Для неодинаковых размеров слоёв невозможно удовлетворить обоим условиям, поэтому обычно усредняют:

$$w_i \sim U\left[-\frac{\sqrt{6}}{\sqrt{n_{out}+n_{in}}}; \frac{\sqrt{6}}{\sqrt{n_{out}+n_{in}}}\right],$$

Такая инициализация называется инициализацией Xavier Glorot

http://proceedings.mlr.press/v9/glorot10a/glorot10a.pdf

Инициализация Хе Kaiming

$$\mathsf{Var}(h_i) = \mathsf{Var}(\sum_{i=1}^{n_{in}} w_i x_i) = \sum_{i=1}^{n_{in}} \mathsf{Var}(w_i x_i) =$$

$$= \sum_{i=1}^{n_{in}} [\mathsf{E}(x_i)]^2 \cdot \mathsf{Var}(w_i) + [\mathsf{E}(w_i)]^2 \cdot \mathsf{Var}(x_i) + \mathsf{Var}(x_i) \cdot \mathsf{Var}(w_i)]$$

- Когда нет симметрии, можно занулить только второе слагаемое

Инициализация Хе Kaiming

$$\mathrm{Var}(h_i) = \mathrm{Var}(\sum_{i=1}^{n_{in}} w_i x_i) = \sum_{i=1}^{n_{in}} \mathrm{Var}(w_i x_i) =$$

$$= \sum_{i=1}^{n_{in}} [\mathrm{E}(x_i)]^2 \cdot \mathrm{Var}(w_i) + [\mathrm{E}(w_i)]^2 \cdot \mathrm{Var}(x_i) + \mathrm{Var}(x_i) \cdot \mathrm{Var}(w_i)] =$$

$$= \sum_{i=1}^{n_{in}} [\mathrm{E}(x_i)]^2 \cdot \mathrm{Var}(w_i) + \mathrm{Var}(x_i) \cdot \mathrm{Var}(w_i) = \sum_{i=1}^{n_{in}} \mathrm{Var}(w_i) \cdot E(x_i^2)$$

- Когда нет симметрии, можно занулить только второе слагаемое

Инициализация Xe Kaiming

$$\text{Var}(h_i) = \text{Var}(\sum_{i=1}^{n_{in}} w_i x_i) = \sum_{i=1}^{n_{in}} \text{Var}(w_i x_i) =$$

$$= \sum_{i=1}^{n_{in}} [\text{E}(x_i)]^2 \cdot \text{Var}(w_i) + [\text{E}(w_i)]^2 \cdot \text{Var}(x_i) + \text{Var}(x_i) \cdot \text{Var}(w_i)] =$$

$$= \sum_{i=1}^{n_{in}} [\text{E}(x_i)]^2 \cdot \text{Var}(w_i) + \text{Var}(x_i) \cdot \text{Var}(w_i) = \sum_{i=1}^{n_{in}} \text{Var}(w_i) \cdot E(x_i^2) =$$

$$= E(x^2) \cdot [n_{in} \cdot \text{Var}(w)]$$

Инициализация Хе Kaiming

$$\mathsf{Var}(h_i) = E(x_i^2) \cdot [n_{in} \cdot \mathsf{Var}(w)]$$
$$x_i = \mathsf{max}(0; h_{i-1})$$

Инициализация Хе Kaiming

$$\mathsf{Var}(h_i) = E(x_i^2) \cdot [n_{in} \cdot \mathsf{Var}(w)]$$
$$x_i = \mathsf{max}(0; h_{i-1})$$

Пусть w_{i-1} симметрично распределён относительно нуля и $b_{i-1} = 0$. Тогда h_{i-1} также симметрично распределён относительно нуля в силу того, что $h_{i-1} = w_{i-1} \cdot \mathsf{ReLU}(h_{i-2})$. А если h_{i-1} также симметрично распределён относительно нуля, то

$$E(x_i^2) = \frac{1}{2} \cdot \mathsf{Var}(h_{i-1})$$

https://arxiv.org/pdf/1502.01852.pdf

Инициализация Хе Kaiming

Если

$$E(x_i^2) = \frac{1}{2} \cdot \mathsf{Var}(h_{i-1})$$

Тогда

$$\mathsf{Var}(h_i) = \frac{1}{2} \cdot \mathsf{Var}(h_{i-1}) \cdot [n_{in} \cdot \mathsf{Var}(w)]$$

$$\mathsf{Var}(w_i) = \frac{2}{n_{in}}$$

Итого: инициализируем веса нормальным распределением с нулевым мат. ожиданием и дисперсией $\frac{2}{n_{in}}$:

$$N(0, \frac{2}{n_{in}})$$

https://arxiv.org/pdf/1502.01852.pdf

Инициализация Xe - пояснения к формулам выше

h_{i-1} has zero mean and has a symmetrical distribution around zero.

$$\begin{aligned}\text{Var}\left(h_{i-1}\right) &= \mathbb{E}\left[h_{i-1}^2\right] \\ &= \mathbb{E}\left[h_{i-1}^2 \mid h_{i-1} > 0\right] \text{Pr}\left[h_{i-1} > 0\right] + \mathbb{E}\left[h_{i-1}^2 \mid h_{i-1} < 0\right] \text{Pr}\left[h_{i-1} < 0\right] \\ &= 2\mathbb{E}\left[h_{i-1}^2 \mid h_{i-1} > 0\right] \text{Pr}\left[h_{i-1} > 0\right]\end{aligned}$$

Since

$$x_i = \text{max}\left(0, h_{i-1}\right)$$

Hence

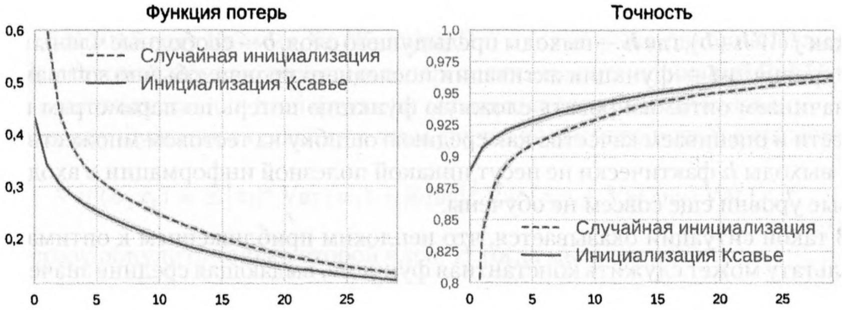
$$x_i^2 = \begin{cases} h_{i-1}^2 & , h_{i-1} > 0 \\ 0 & , h_{i-1} \leq 0 \end{cases}$$

$$\implies \mathbb{E}\left[x_i^2\right] = \mathbb{E}\left[h_{i-1}^2 \mid h_{i-1} > 0\right] \text{Pr}\left[h_{i-1} > 0\right] = \frac{1}{2} \text{Var}\left(h_{i-1}\right)$$

Краткие итоги

- Для симметричных функций с нулевым средним используйте инициализацию Ксавье init="glorot_uniform" или
- Для ReLU и им подобным инициализацию Хе init="he_uniform" или init="he_nomal"
- Эти две инициализации корректируют параметры распределений в зависимости от входа и выхода слоя так, чтобы поддерживать дисперсию равной единице

Эксперимент с MNIST



Источник: Николенко, страница 149