Семинар 5.04.21

Связь дисперсий последовательных слоев глубокой нейросети

Educated guess: зависит от (начальной инициализации) весов

Напоминание:

$x \sim \mathcal{U}[a, b]$ – равномерн. распред-е сл. в.

$\mathbb{E}[x] = \frac{a+b}{2}$

$Var(x) = \frac{(b-a)^2}{12}$

Рассмотрим кейс:

$w_i \sim \mathcal{U}\left[-\frac{1}{\sqrt{n_{out}}}, \frac{1}{\sqrt{n_{out}}}\right]$, где

$w_i \in W_{ij}$ – матрица весов

n_{out} – кол-во нейронов на выходном слое

$\mathbb{E}[w_i] = 0$, $Var[w_i] = \frac{1}{3 n_{out}}$

~~Следовательно~~

Из лекции мы помним, что

$$Var(y_i^{(l+1)}) = Var(w_i^{(l+1)}) Var(y_i^{(l)})$$

где l – номер слоя

$$\text{т.е.} = Var(w_i^{(l+1)}) Var(w_i^{(l)}) Var(y_i^{(l-1)})$$

$$= \ldots$$

$$= \prod_l Var(w_i^{(l)}) Var(y_i^{(0)})$$

– на каждом слое веса распределены одинаково и не зависят др. от др

$$= \prod_l \frac{1}{3 n_{out}^{(l)}} \cdot Var(y_i^{(0)}) \xrightarrow[l \text{ большая}]{} 0$$

~~Итак~~ То есть с увеличением глубины сети сигнал затухает: нулевые матожидание и дисперсия.

Это уже достаточно плохо, но дальше только хуже. Рассмотрим обратный проход:

Посчитаем градиент функции ошибки L на слое l. Вспомним, что слой $l+1$ – это

$$\underbrace{\begin{bmatrix} \vdots \end{bmatrix}}_{y^{(l+1)}} = \underbrace{f}_{\text{ф-ция активации}}\left(\underbrace{\begin{bmatrix} \ & \ \\ \hline \ & \ \end{bmatrix}}_{W^{(l+1)} \text{ – веса}} \begin{bmatrix} \ \end{bmatrix}_{y^{(l)}} \right) : \quad y_i^{(l+1)} = f\left(\sum_j w_{ij}^{(l+1)} y_i^{(l)}\right)$$

$$\frac{\partial L}{\partial y_i^{(l)}} = \sum_j \frac{\partial L}{\partial y_j^{(l+1)}} \cdot \frac{\partial y_j^{(l+1)}}{\partial y_i^{(l)}} = \sum_j \frac{\partial L}{\partial y_j^{(l+1)}} \underbrace{\frac{\partial f}{\partial y_i^{(l)}}}_{\text{не зависит от } j} \frac{\partial (\sum_j w_{ij} y_j^{(l)})}{\partial y_i} =$$

$$= \frac{\partial f}{\partial y_i^{(l)}} \sum_j \frac{\partial L}{\partial y_j^{(l+1)}} \cdot w_{ij}^{(l+1)}$$

Для симметричной относительно начала координат ф-ции активации f (напр., tanh)

$\frac{\partial f}{\partial y_i^{(l)}} \approx 1$ в окрестности нуля
(рассм. именно эту область, потому что у нас нулевое матожидание)

$$Var\left(\frac{\partial L}{\partial y_i^{(l)}}\right) = Var\left(\sum_j \frac{\partial L}{\partial y_j^{(l+1)}} w_{ij}^{(l+1)}\right) = \sum_j Var\left(\frac{\partial L}{\partial y_j^{(l+1)}}\right) \cdot Var\, w_{ij}^{(l+1)}$$

$$\approx \frac{1}{3 n_{in}} \cdot \sum Var \frac{\partial L}{\partial y_j^{(l+1)}}$$

n_{in} – число входящих нейронов

И здесь дисперсия стремится к нулю.

Вывод: такая инициализация весов $w_i \sim U\left[-\frac{1}{\sqrt{n_{out}}}, \frac{1}{\sqrt{n_{out}}}\right]$ не подходит для глубоких сетей.

Решение

- Xavier Initialization для симметричной ф-ции активации
- He initialization для ReLU