X5 Tech

ТОКЕНИЗАЦИЯ В ЯЗЫКОВЫХ МОДЕЛЯХ

Андреева Дарья
Data Scientist, X5 Tech

токенизация

сейчас мы делим текст на слова — как быть с различными формами одного и того же слова? и как обрабатывать незнакомые слова?

деление по пробелам

мама мыла раму -> [‘мама’, ‘мыла’, ‘раму’]

деление по пробелам

мама мыла раму -> [‘мама’, ‘мыла’, ‘раму’]

минусы?

огромный словарь

на редкие слова плохие эмбеддинги

X5 Tech

КАК СЭМПЛИРОВАТЬ?

X5 Tech

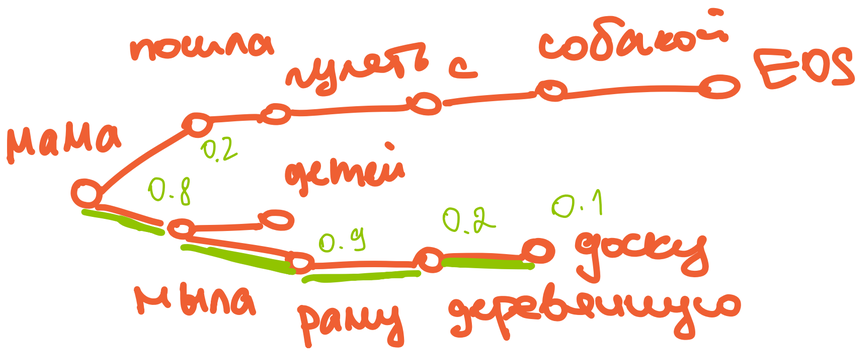
как сэмплировать?

хотим сгенерировать текст, похожий на человека

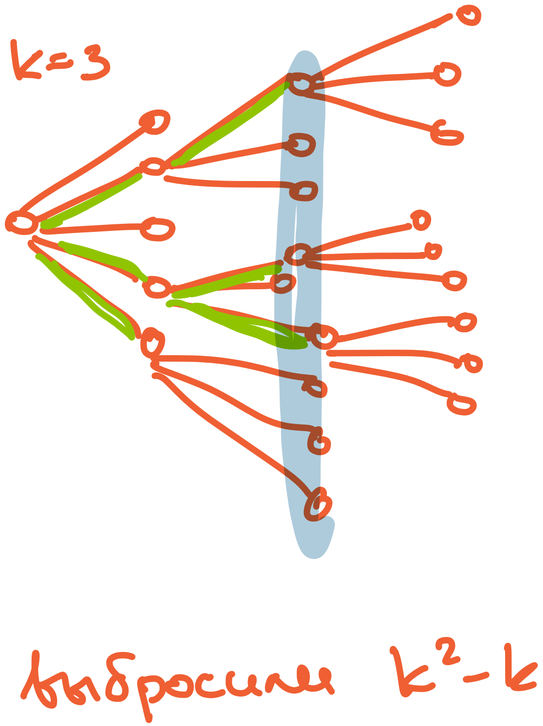
1. текст осознанный
2. текст разнообразный

жадный метод [greedy]

берём токены с наибольшей вероятностью



beam search



выбираем k наиболее вероятных токенов

для каждого из них выбираем свои k наиболее вероятных токенов

чтобы не порваться от экспоненциальности, отбрасываем k2-k вариантов, на каждом шаге оставляем только топ-k путей

greedy vs beam search

greedy

детерминированный

быстрый

можем выбрать лучшее на каждом шаге

не гарантирует разнообразия реплик

beam search

детерминированный

k = 1 -- жадный

вычислительно сложный

можно генерировать более разнообразные реплики

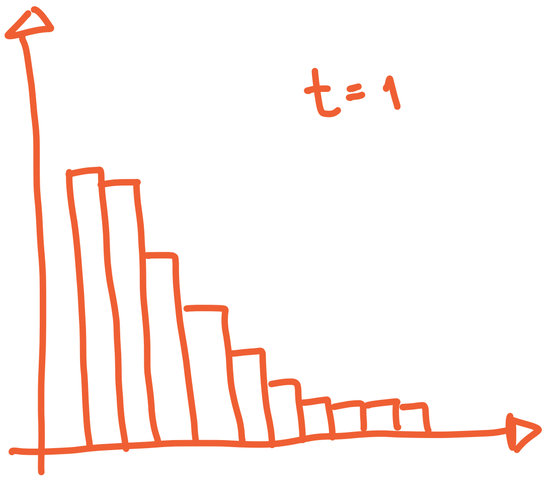
реплики очень общие

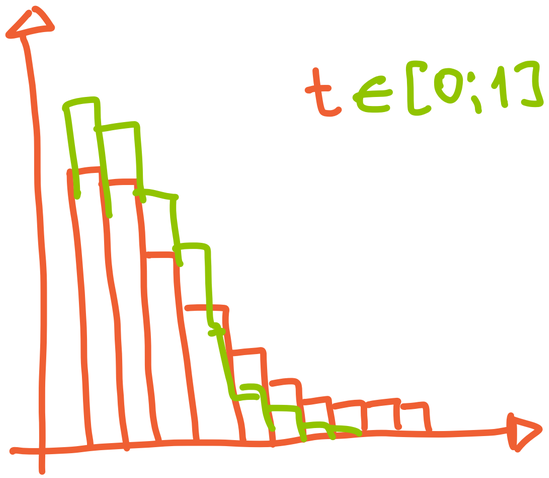
temperature samplings

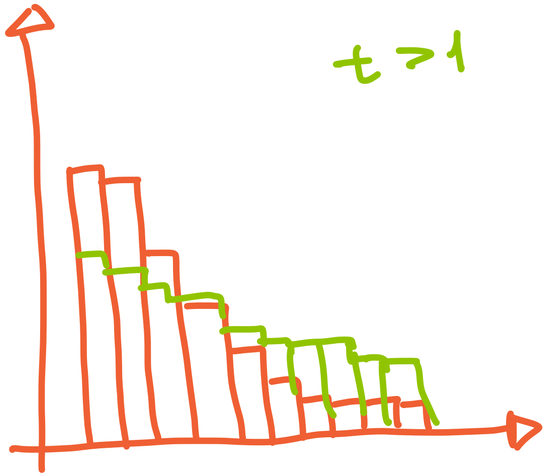
$$p(X_i) = \frac{e^{x_i/\tau}}{\sum_{j=1}^{|V|} e^{x_j/\tau}}$$

перевзвешиваем веса

параметр τ назовём температурой





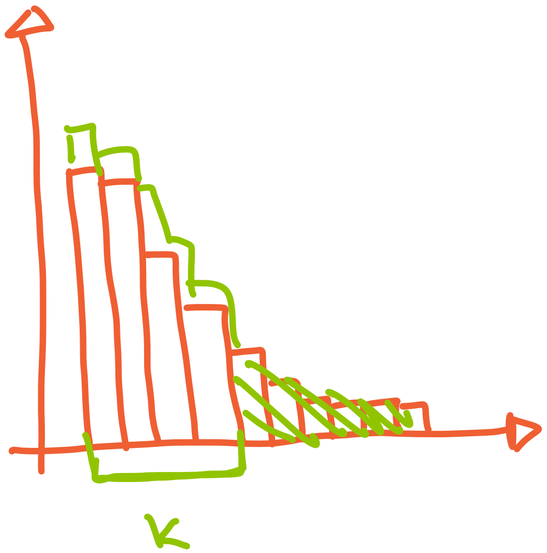


<s>
[UNK]

X5 Tech

top-k sampling

$$p' = \sum_{x \in \mathrm{Vocab}} p(x_i | x_{1:i-1})$$



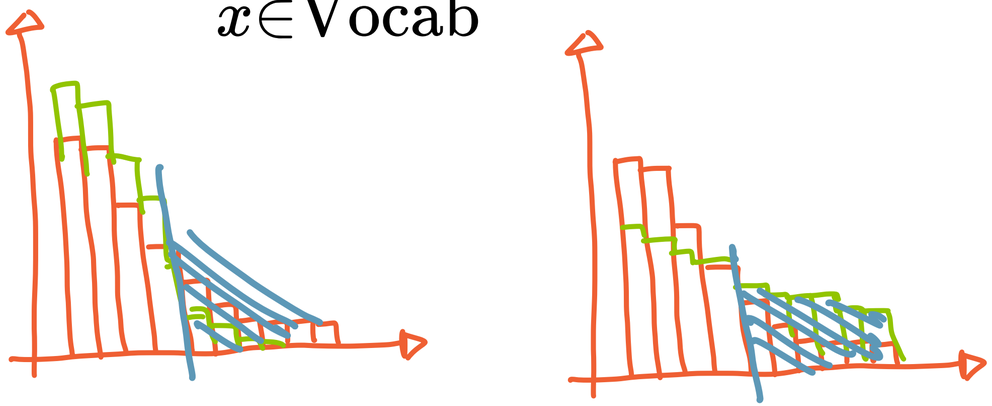
на каждом шаге отбираем k самых вероятных слов

введём нормировочный коэффициент

шкалируем выходы софтмакса

top-k sampling

$$p' = \sum_{x \in \mathrm{Vocab}} p(x_i | x_{1:i-1})$$



на каждом шаге отбираем k самых вероятных слов

введём нормировочный коэффициент

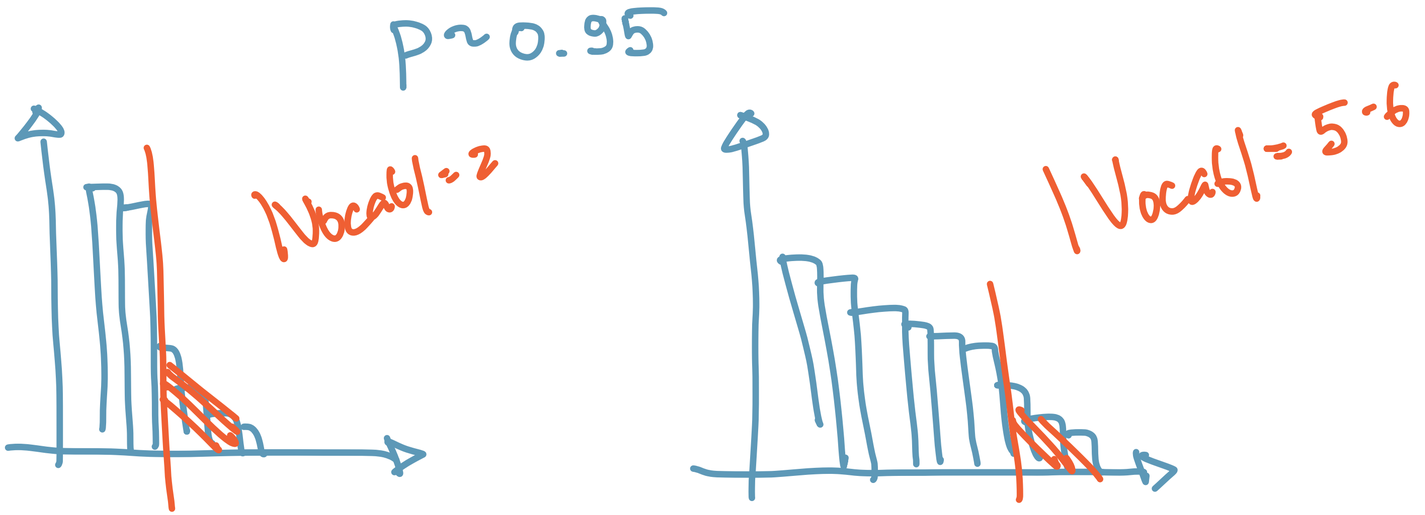
шкалируем выходы софтмакса

из-за разности распределений хотим иметь разный k на каждом шаге

nucleus [top-p] sampling [2020]

$$\sum_{x \in \text{Vocab}} p(x_i | x_{1:i-1}) \geq p$$

берем минимальное число токенов, сумма вероятностей которых >= p



X5 Tech

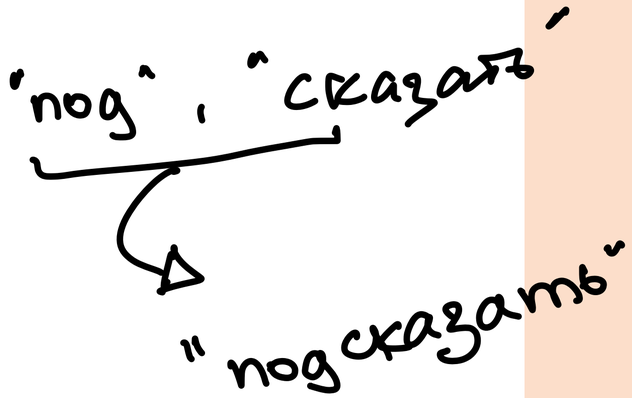
что хотим?

словарь ограниченного размера

содержит в себе токены, которые достаточно часто встречаются

BPE — BYTE PAIR ENCODING

Neural Machine Translation of Rare Words with Subword Units: https://arxiv.org/abs/1508.07909

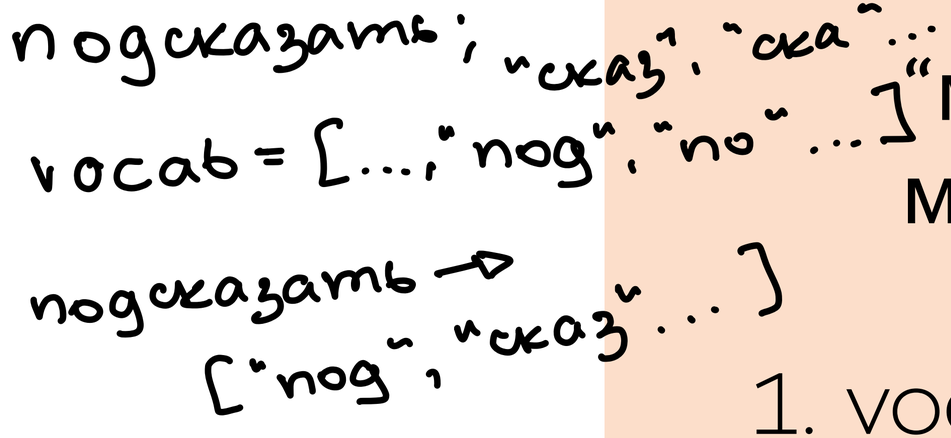


BPE — BYTE PAIR ENCODING

“мама мыла раму,
маму мыла рама”

1. vocab = [“м”, “а”, “р”, “у”…]
2. vocab += [“ма”, “му”,…]
3. vocab += [“мам”, “мыл”…]
4. vocab.size == n: stop

+ рама



BPE — BYTE PAIR ENCODING

“мама мыла раму,
маму мыла рама”

1. vocab = [“м”, “а”, “р”, “у”…]
2. vocab += [“ма”, “му”,…]
3. vocab += [“мам”, “мыл”…]
4. vocab.size == n: stop

при токенизации берем
самую большую подстроку

X5 Tech

WORD PIECE —

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation: https://arxiv.org/abs/1609.08144

WORD PIECE —

$$\text{score} = \frac{\text{freq\_of\_pair}[w_1, w_2]}{\text{freq}[w_1] * \text{freq}[w_2]}$$

“мама мыла раму, маму мыла рама”

1. vocab = [“м”, “а”, “р”, “у”…]
2. vocab += [“ма”, “му”,…] пары с наивысшим скором
3. vocab.size == n: stop

X5 Tech

WORD PIECE —

$$\text{score} = \frac{\text{freq\_of\_pair}[w_1, w_2]}{\text{freq}[w_1] * \text{freq}[w_2]}$$

“мама мыла раму, маму мыла рама”

[“рам”, “а”] = $\frac{1}{2 \cdot 8} = \frac{1}{16}$

[“ра”, “ма”] = $\frac{1}{2 \cdot 4} = \frac{1}{8}$

X5 Tech

WORD PIECE —

$$\text{score} = \frac{\text{freq\_of\_pair}[w_1, w_2]}{\text{freq}[w_1] * \text{freq}[w_2]}$$

“мама мыла раму, маму мыла рама”

[“рам”, “а”] = 1/16
[“ра”, “ма”] = 1/8

хотим добавлять токены, которые отдельно встречаются редко, но вместе часто

X5 Tech

UNIGRAM —

X5 Tech

Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates: https://arxiv.org/abs/1804.10959

["бан", "бум", "бах"]
– корпус

"а"(2), "б"(3), "м"(2), "у"(1), "ба"(2), "ам"(1), "бу"(1), "ум"(1), "ах"(1)

$P(\text{"ба"}, \text{"м"}) = {}^{2}/_{5} \cdot {}^{2}/_{5} = {}^{4}/_{5^2}$

$P(\text{"б"}, \text{"а"}, \text{"м"}) = {}^{3}/_{5} \cdot {}^{2}/_{5} \cdot {}^{2}/_{5} = {}^{12}/_{5^3}$

$P(\text{"б"}, \text{"ам"}) = {}^{3}/_{5} \cdot {}^{1}/_{5} = {}^{3}/_{5^2}$

UNIGRAM –

словарь – BPE с максимальным словарём

для каждого токена считаем, насколько уменьшится вероятность при его удалении

X5 Tech

ЯЗЫКОВЫЕ МОДЕЛИ

X5 Tech

языковые модели

$$P(w_1, \ldots, w_T) = P(w_1) \prod_{i=2}^{T} P(w_i | w_1, \ldots, w_{i-1})$$

n-gram language model

$$P(w_1, \ldots, w_T) = P(w_1) \prod_{i=2}^{T} P(w_i | w_1, \ldots, w_{i-1})$$

предположим, что вероятность слова не очень зависит от слов, которые были далеко в начале

$$P(w_i | w_{i-1}, \ldots, w_1) \sim P(w_i | w_{i-1}, \ldots, w_{i-n})$$

n-gram language model

$$P(w_i|w_{i-1},\ldots,w_{i-n}) = \frac{P(w_{i-n},\ldots,w_i)}{P(w_{i-n},\ldots,w_{i-1})}$$

n-gram language model

$$P(w_i|w_{i-1},\ldots,w_{i-n}) = \frac{P(w_{i-n},\ldots,w_i)}{P(w_{i-n},\ldots,w_{i-1})}$$

w_i = одна буква

$$= \frac{count(w_{i-n},\ldots,w_i)}{count(w_{i-n},\ldots,w_{i-1})}$$

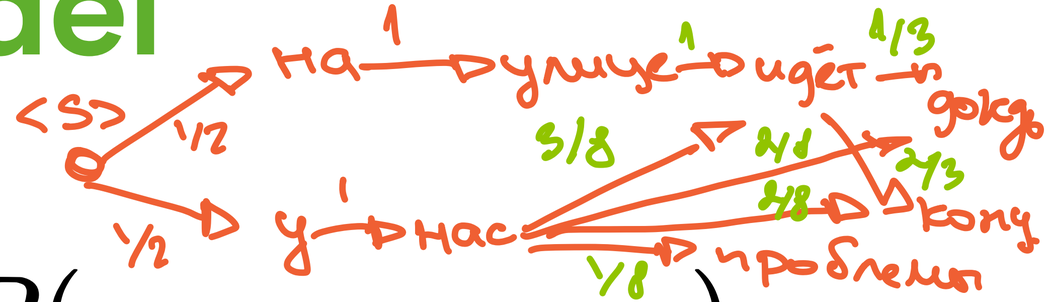
$\alpha \uparrow$ $\alpha|V|\uparrow$

$\alpha = 1, 3, 5, 7$ et

мама

мам

n-gram language model



$$P(w_i|w_{i-1},\ldots,w_{i-n}) = \frac{P(w_{i-n},\ldots,w_i)}{P(w_{i-n},\ldots,w_{i-1})}$$

$$= \frac{count(w_{i-n},\ldots,w_i)}{count(w_{i-n},\ldots,w_{i-1})}$$

n-gram language model

$$\frac{count(w_{i-n}, \ldots, w_i)}{count(w_{i-n}, \ldots, w_{i-1})}$$

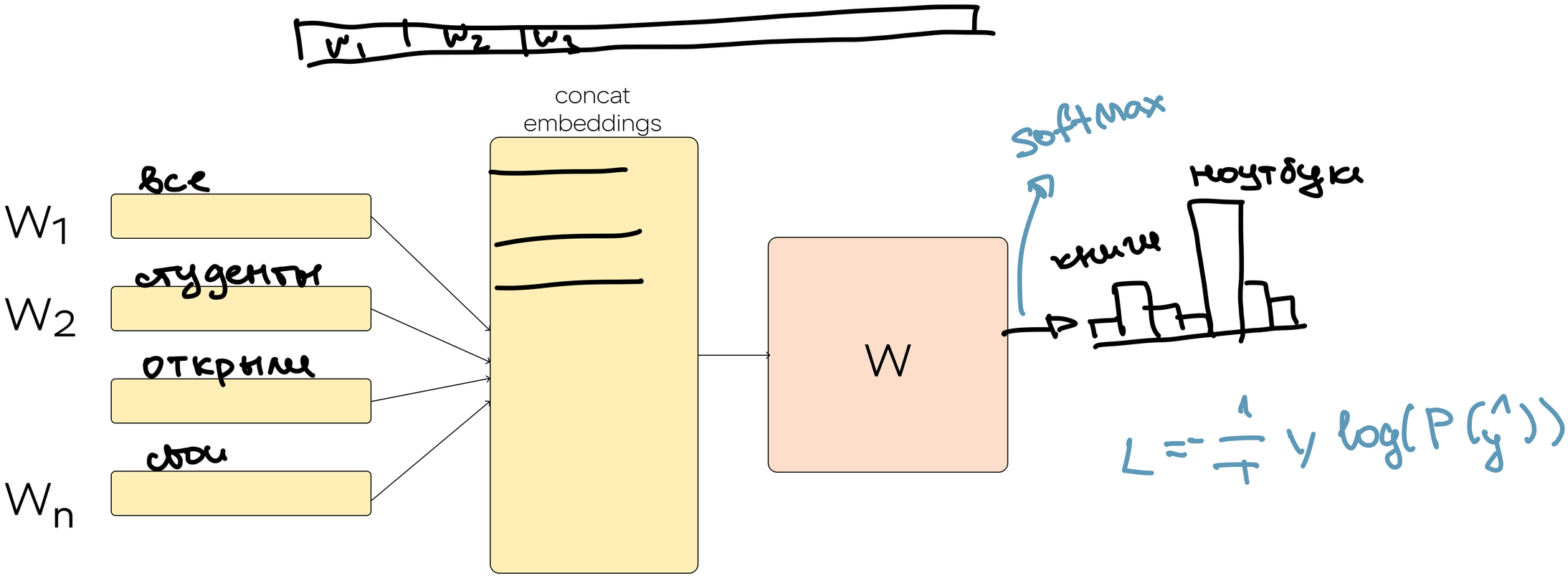
минусы?

биграммы = $|V|^2$

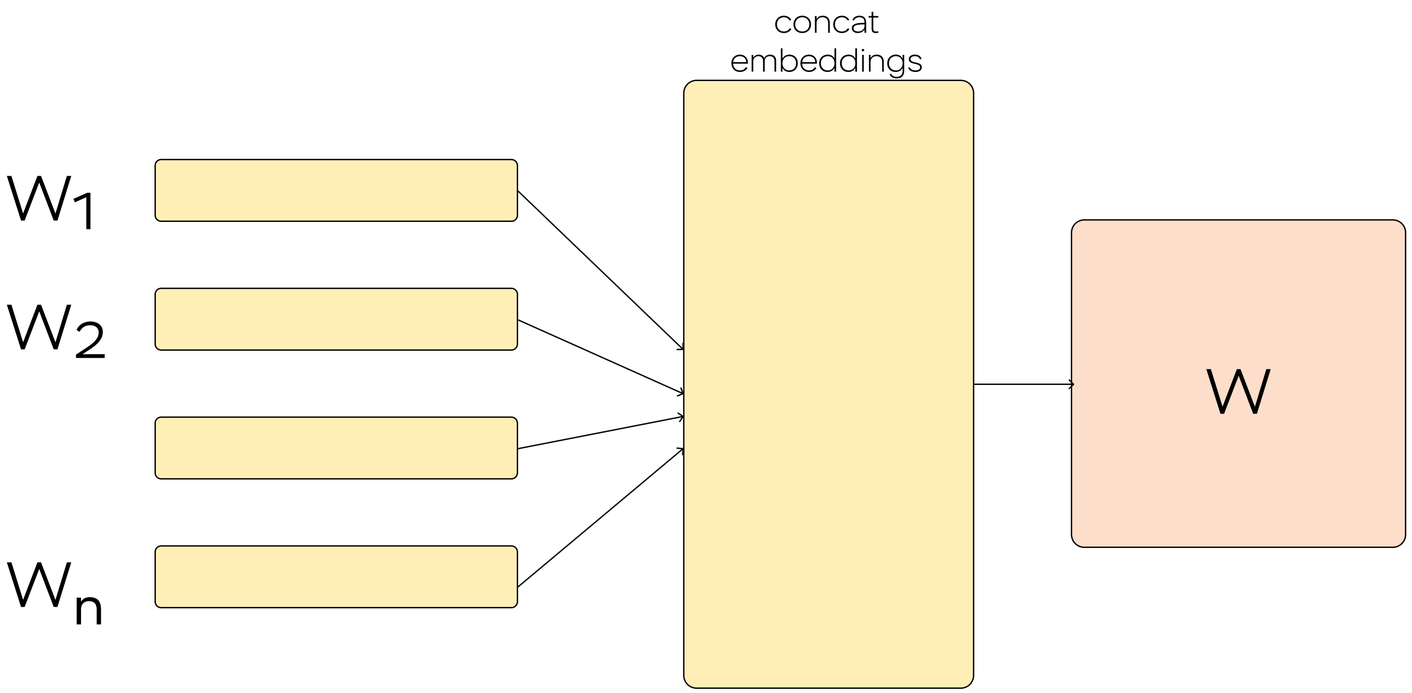
тяжело считать и хранить при больших n

при больших n будут низкие вероятности и большая дисперсия

нейросетевые модели



нейросетевые модели

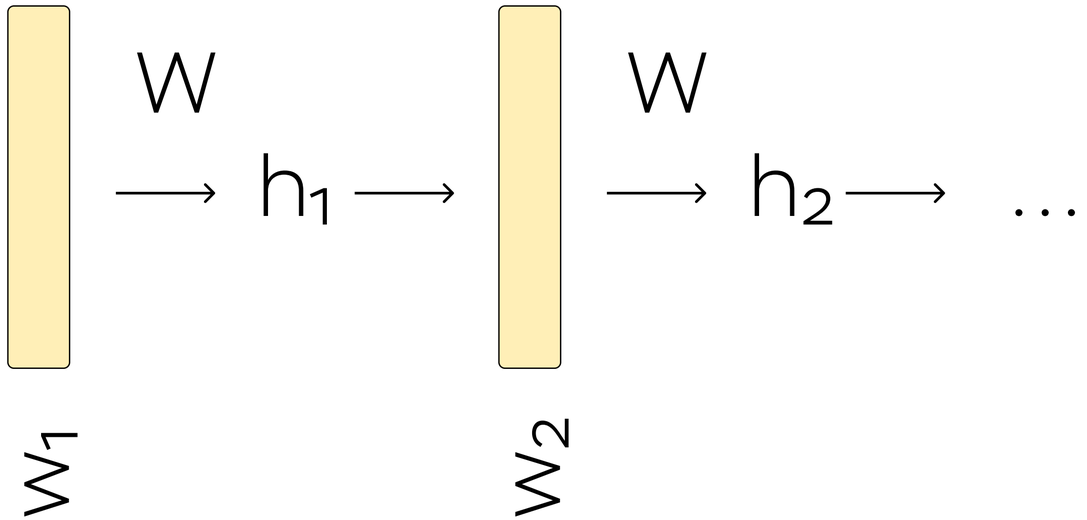


плюсы:
не нужно хранить n-граммы

минусы:
увеличение окна (n)
увеличивает w

рекуррентная языковая модель

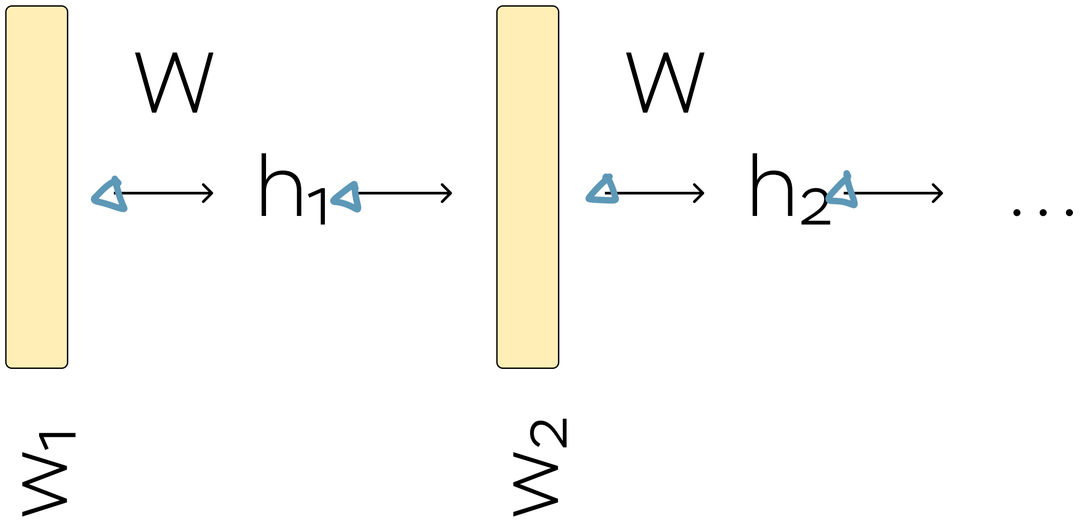
$$\frac{\partial L}{\partial h_1} = \frac{\partial L}{\partial h_n} \cdot \frac{\partial h_n}{\partial h_{n-1}} \cdots \frac{\partial h_2}{\partial h_1}$$



плюсы:
не нужно хранить n-граммы
учитываем порядок

минусы:
постепенно забываем, что было в начале
нельзя параллелить
затухание градиента

рекуррентная языковая модель



плюсы:
не нужно хранить n-граммы
учитываем порядок

минусы:
постепенно забываем, что было в начале
нельзя параллелить
затухание градиента