

Лекция 3.

Языковые модели

Языковые модели

- Что такое языковые модели?
- Общая идея работы
- Классический подход к работе с языковыми моделями
- Нейросетевой подход к работе с языковыми моделями

Что такое языковая модель?

Рассмотрим, например, модели поездов.

- имеют некоторые свойства поездов
(выглядят как поезда)
- могут вести себя подобно поездам
- хорошие модели больше похожи на настоящие поезда



Что такое языковая модель?

Рассмотрим, например, физические модели.

- помогут понять, какие события лучше согласуются с миром, какие более вероятны
- может предсказать, что произойдет, учитывая некоторый «контекст»



Will it fall?



In which direction?



Different masses



Complex scenes



Infer the mass



Predict fluids

Что такое языковая модель?

Отличается понятие события: для языка событие — это языковая единица (текст, предложение, токен, символ).

Языковые модели (LM) оценивают вероятность различных языковых единиц: символов, токенов, последовательностей токенов.

Примеры языковых моделей

Мы имеем дело с языковыми моделями каждый день.

Например:

- поисковая система

I saw a cat
I saw a cat on the chair
I saw a cat running after a dog
I saw a cat in my dream
I saw a cat book

- переводчик, электронная почта

I saw a ca
car ←

Примеры языковых моделей

Мы имеем дело с языковыми моделями каждый день.

Например:

- поиск опечаток

I saw a catt

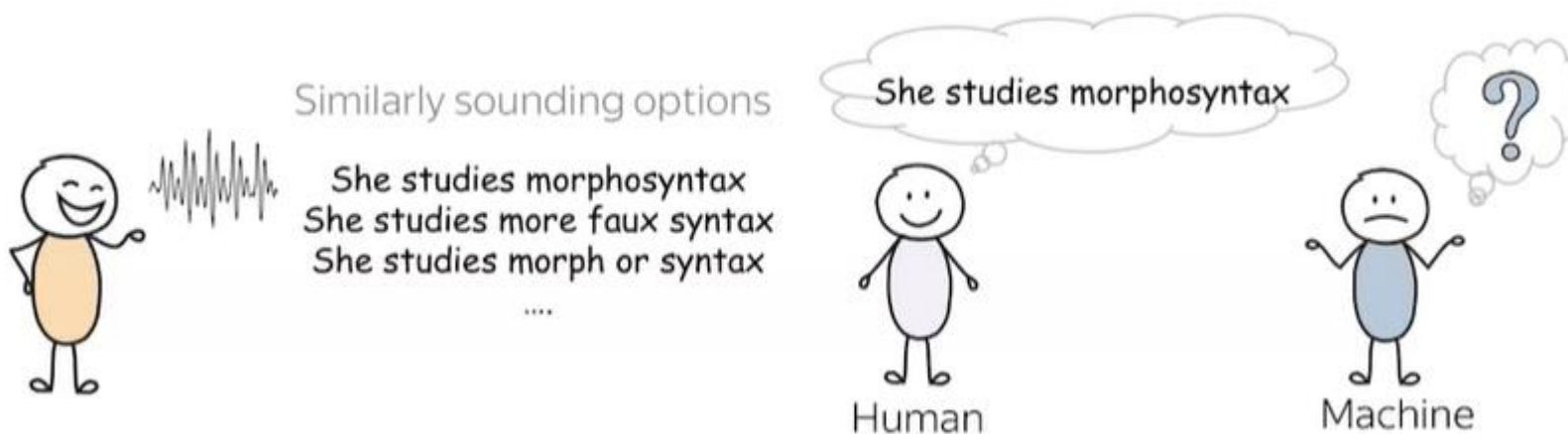
Probably you meant I saw a cat

I saw a catt

cat

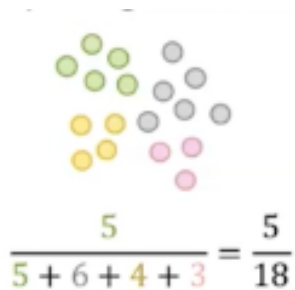
car

Сложности машинного восприятия

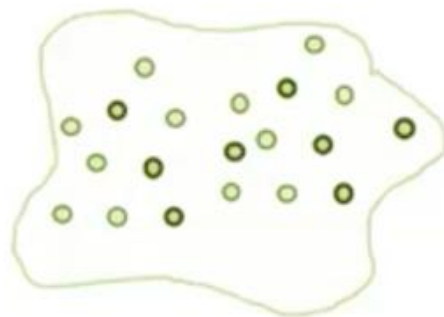


Какова вероятность появления предложения в языке?

Какова вероятность вытащить зелёный шар?



Можно ли рассуждать аналогично для предложений?



Text corpus

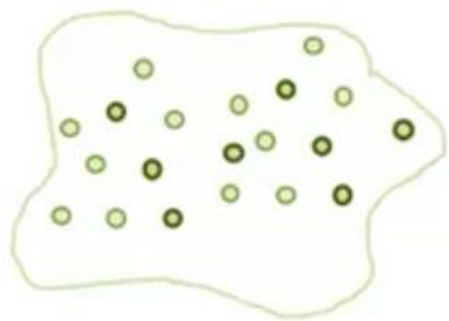
$$P(\text{the mut is tinming the tebn}) = \frac{0}{|\text{corpus}|} = 0$$

$$P(\text{mut the tinming tebn is the}) = \frac{0}{|\text{corpus}|} = 0$$

При таком подходе предложения, которые никогда не встречались в корпусе, получат нулевую вероятность.

Какова вероятность появления предложения в языке?

Можно ли рассуждать аналогично для предложений?



Text corpus

$$P(\text{the mut is tinming the tebn}) = \frac{0}{|\text{corpus}|} = 0$$

$$P(\text{mut the tinming tebn is the}) = \frac{0}{|\text{corpus}|} = 0$$

При таком подходе предложения, которые никогда не встречались в корпусе, получают нулевую вероятность.

Мы не сможем надежно оценить вероятности предложений, если будем рассматривать их как атомарные единицы!

Идея: разбивать предложения на меньшие части

- читаем предложение слово за словом
- обновляем вероятность каждый раз, когда встречается новый токен

$$P(I \text{ saw}) =$$

$$\underbrace{P(I)} \cdot \underbrace{P(\text{saw}|I)}$$

Probability of I Probability of saw given I

$$P(I \text{ saw a}) =$$

$$\underbrace{P(I) \cdot P(\text{saw}|I)} \cdot \underbrace{?}$$

Probability of I saw

$$P(I \text{ saw a}) =$$

$$\underbrace{P(I) \cdot P(\text{saw}|I)} \cdot \underbrace{P(a|I \text{ saw})}$$

Probability of I saw Probability of a given I saw

$$P(I \text{ saw a cat}) =$$

$$\underbrace{P(I) \cdot P(\text{saw}|I) \cdot P(a|I \text{ saw})} \cdot \underbrace{?}$$

Probability of I saw a

$$P(I \text{ saw a cat}) =$$

$$\underbrace{P(I) \cdot P(\text{saw}|I) \cdot P(a|I \text{ saw})} \cdot \underbrace{P(\text{cat}|I \text{ saw a})}$$

Probability of I saw a Probability of cat given I saw a

$$P(I \text{ saw a cat on}) =$$

$$\underbrace{P(I) \cdot P(\text{saw}|I) \cdot P(a|I \text{ saw}) \cdot P(\text{cat}|I \text{ saw a})} \cdot \underbrace{?}$$

Probability of I saw a cat

Идея: разбивать предложения на меньшие части

- $(y_1, y_2, \dots, y_n)$ — последовательность токенов
- $P(y_1, y_2, \dots, y_n)$ — вероятность увидеть эти токены в данной последовательности

$$P(y_1, y_2, \dots, y_n) = P(y_1) \cdot P(y_2|y_1) \cdot P(y_3|y_1, y_2) \cdot \dots \cdot P(y_n|y_1, \dots, y_{n-1}) = \prod_{t=1}^n P(y_t|y_{<t})$$

Структура моделирования языка с написанием слева направо

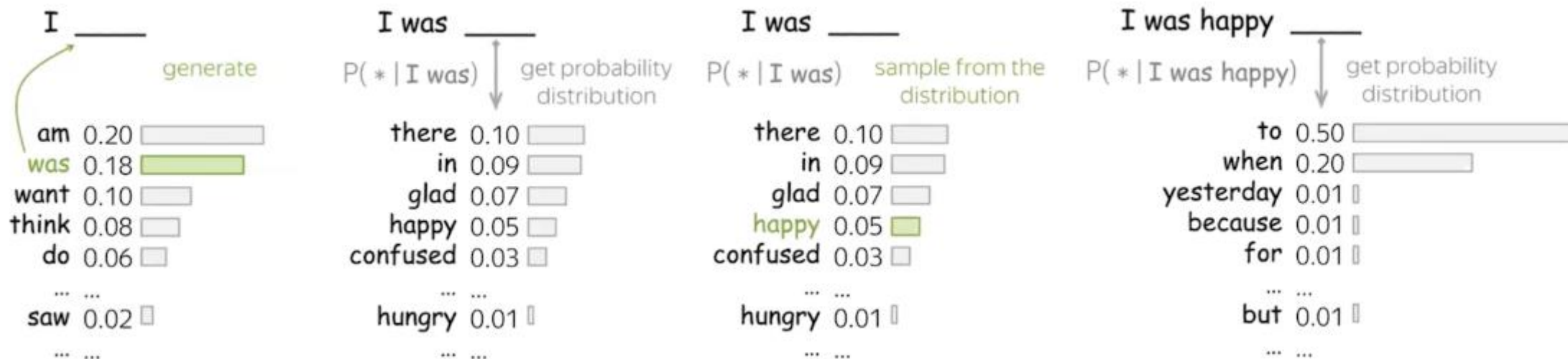
Модели различаются по способу вычисления вероятностей.

Цель: понять, как вычислять условную вероятность $P(y_t | y_1, y_2, \dots, y_{t-1})$

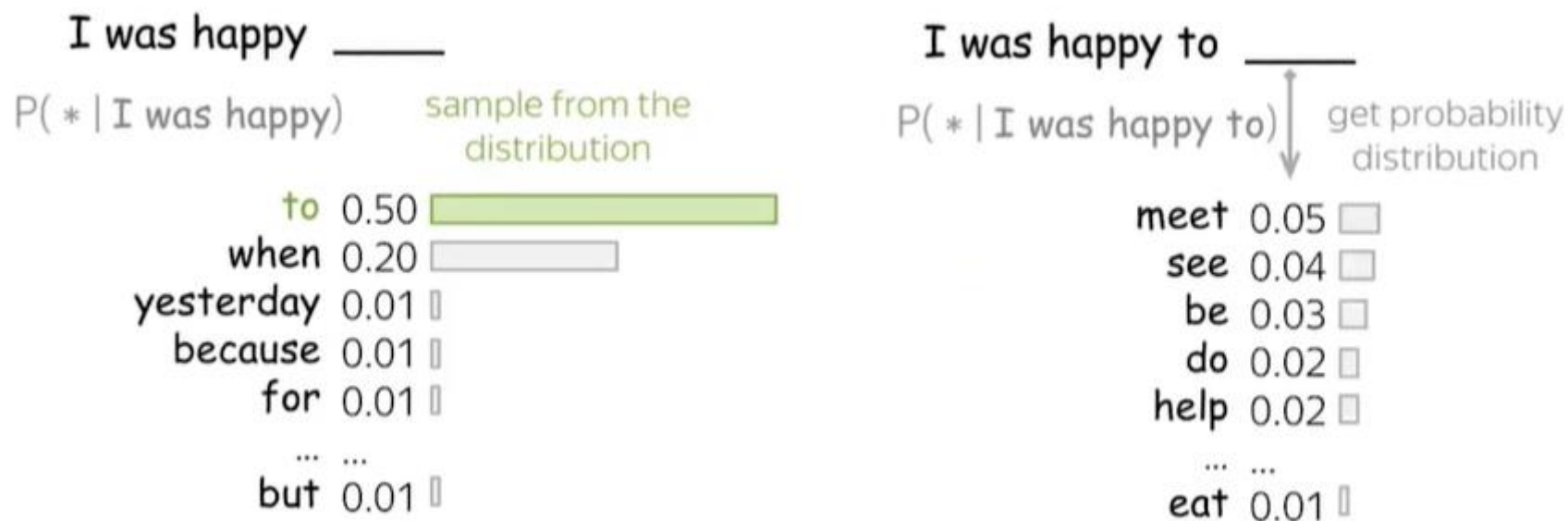
1 вариант: использование n-грамм (вычисления «вручную»)

2 вариант: нейросети

Генерация текста с помощью языковой модели



Генерация текста с помощью языковой модели



Традиционный подход к работе с языковыми моделями

Цель: необходимо вычислить условную вероятность в формуле

$$P(y_1, y_2, \dots, y_n) = P(y_1) \cdot P(y_2|y_1) \cdot P(y_3|y_1, y_2) \cdot \dots \cdot P(y_n|y_1, \dots, y_{n-1}) = \prod_{t=1}^n P(y_t|y_{<t})$$

Оценка на основе глобальной статистики текстового корпуса, т. е. подсчета.

Как вычислить условные вероятности?

Простой способ:

$$P(y_t | y_1, y_2, \dots, y_{t-1}) = \frac{N(y_1, y_2, \dots, y_t)}{N(y_1, y_2, \dots, y_{t-1})},$$

где $N(y_1, y_2, \dots, y_t)$ — количество раз, которое мы встречаем $(y_1, y_2, \dots, y_t)$ в обучающем корпусе

Марковское предположение

Вероятность слова зависит только от фиксированного числа предыдущих слов.

$$P(y_t | \underbrace{y_1, y_2, \dots, y_{t-1}}_{\text{all previous tokens}}) = P(y_t | \underbrace{y_{t-n+1}, \dots, y_{t-1}}_{\text{n-1 previous token}})$$

- Например:
- $n=3$ (trigram model): $P(y_t | y_1, y_2, \dots, y_{t-1}) = P(y_t | y_{t-2}, y_{t-1})$
 - $n=2$ (bigram model): $P(y_t | y_1, y_2, \dots, y_{t-1}) = P(y_t | y_{t-1})$
 - $n=1$ (unigram model): $P(y_t | y_1, y_2, \dots, y_{t-1}) = P(y_t)$

Марковское предположение. Пример

$P(\text{I saw a cat on a mat}) = ?$

Before	After (3-gram)
$P(\text{I saw a cat on a mat}) =$	$P(\text{I saw a cat on a mat}) =$
$P(\text{I})$	$P(\text{I})$
• $P(\text{saw} \text{I})$	• $P(\text{saw} \text{I})$
• $P(\text{a} \text{I saw})$	• $P(\text{a} \text{I saw})$
• $P(\text{cat} \text{I saw a})$	• $P(\text{cat} \text{I saw a})$
• $P(\text{on} \text{I saw a cat})$	• $P(\text{on} \text{I saw a cat})$
• $P(\text{a} \text{I saw a cat on})$	• $P(\text{a} \text{I saw a cat on})$
• $P(\text{mat} \text{I saw a cat on a})$	• $P(\text{mat} \text{I saw a cat on a})$



Возможная проблема

$$P(\text{mat} \mid \text{I saw a cat on a}) = P(\text{mat} \mid \text{cat on a}) = \frac{N(\text{cat on a mat})}{N(\text{cat on a})}$$

$$P(\text{mat} \mid \text{cat on a}) = \frac{N(\text{cat on a mat})}{\boxed{N(\text{cat on a})}} = ?$$

zero

not good: can not compute the probability

$$P(\text{mat} \mid \text{cat on a}) = \frac{\overset{\text{zero}}{\boxed{N(\text{cat on a mat})}}}{N(\text{cat on a})} = ?$$

not good: zeros out probability of the sentence

Решение проблемы (Back-off)

Использовать меньше контекста в ситуациях, о которых мы мало что знаем:

- если можно, использовать триграмму
- если нет, то биграмму
- если нет, то униграмму

$$P(\text{mat} \mid \text{cat on a}) = \frac{N(\text{cat on a mat})}{\boxed{N(\text{cat on a})}} = ?$$

zero

not good: can not compute the probability

$N(\text{cat on a}) = 0 \longrightarrow \text{try "on a"}$

$P(\text{mat} \mid \text{cat on a}) \approx P(\text{mat} \mid \text{on a})$

$N(\text{on a}) = 0 \longrightarrow \text{try "a"}$

$P(\text{mat} \mid \text{on a}) \approx P(\text{mat} \mid \text{a})$

$N(\text{a}) = 0 \longrightarrow \text{try unigram}$

$P(\text{mat} \mid \text{a}) \approx P(\text{mat})$

Лучшее решение: интерполяция

$$P(\text{mat} \mid \text{cat on a}) = \frac{N(\text{cat on a mat})}{N(\text{cat on a})} = ?$$

zero

not good: can not compute the probability

$$\begin{aligned}\hat{P}(\text{mat} \mid \text{cat on a}) \approx & \lambda_3 P(\text{mat} \mid \text{cat on a}) + \\ & \lambda_2 P(\text{mat} \mid \text{on a}) + \\ & \lambda_1 P(\text{mat} \mid \text{a}) + \\ & \lambda_0 P(\text{mat})\end{aligned}$$

$$\sum_i \lambda_i = 1$$

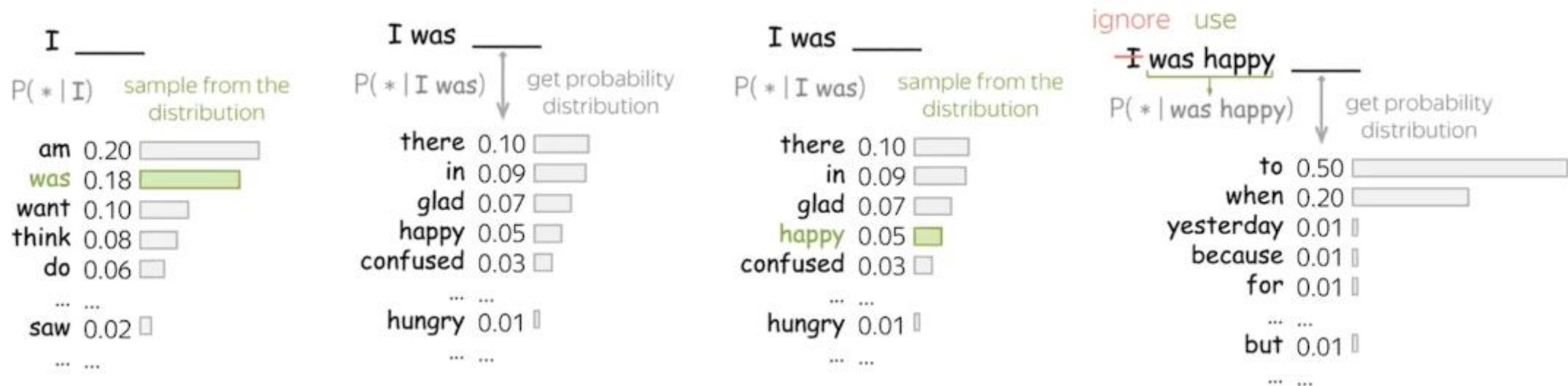
Сглаживание Лапласа

$$P(\text{mat} \mid \text{cat on a}) = \frac{\overset{\text{zero}}{N(\text{cat on a mat})}}{N(\text{cat on a})} = ?$$

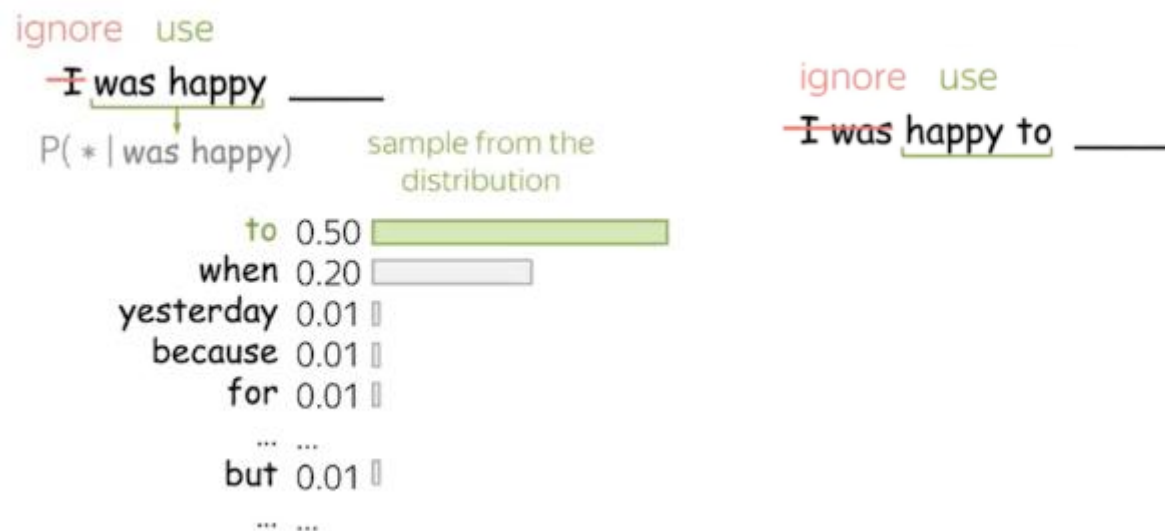
not good: zeros out probability of the sentence

$$\hat{P}(\text{mat} \mid \text{cat on a}) = \frac{\delta + N(\text{cat on a mat})}{\delta \cdot |V| + N(\text{cat on a})}$$

Генерация текста с помощью n-граммной языковой модели (триграммная модель)



Генерация текста с помощью n-граммной языковой модели (триграммная модель)



Пример сгенерированного текста

when this option may be the worst day of amnesty
international delegations visited israel , and felt that
his sisters , that they are reserved for zyryanovsk
concentrating factory there is a member of the shire ,"
given as to damage the expansion of a meeting over a large
health maintenance organization , smoking ,
airconditioning , designated smoking area . _eos_

so even when i talk a bit short , there was no easy thing
to do different buffer flushing strategies in the future ,
due to huge list of number - one just has started
production of frits in the process and has free wi - fi "
operation _eos_

Нейросетевой подход

Цель: необходимо вычислить условную вероятность в формуле:

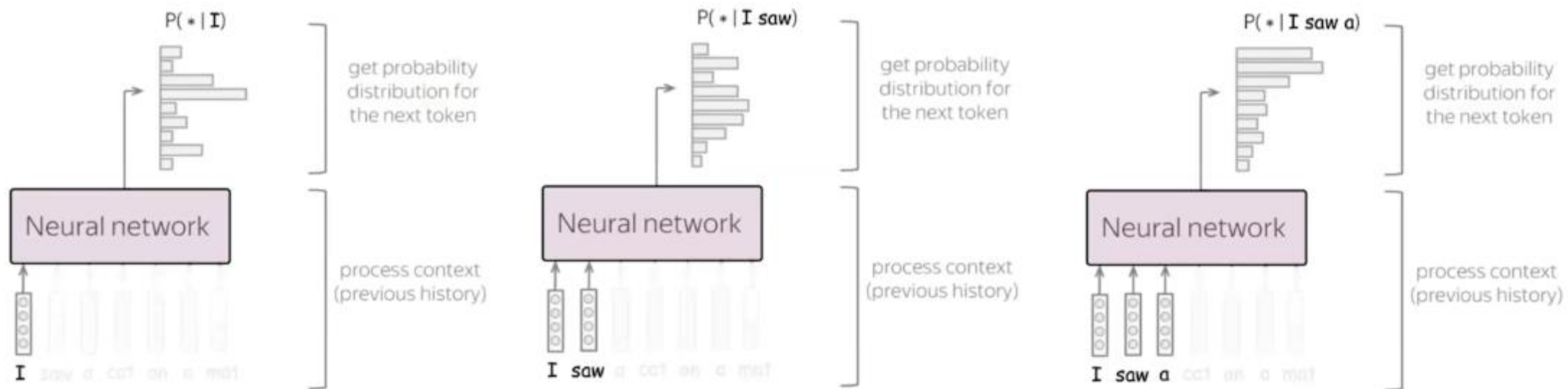
$$P(y_1, y_2, \dots, y_n) = P(y_1) \cdot P(y_2|y_1) \cdot P(y_3|y_1, y_2) \cdot \dots \cdot P(y_n|y_1, \dots, y_{n-1}) = \prod_{t=1}^n P(y_t|y_{<t})$$

В нейронных сетях нужно обучить нейронную сеть прогнозировать вероятности.

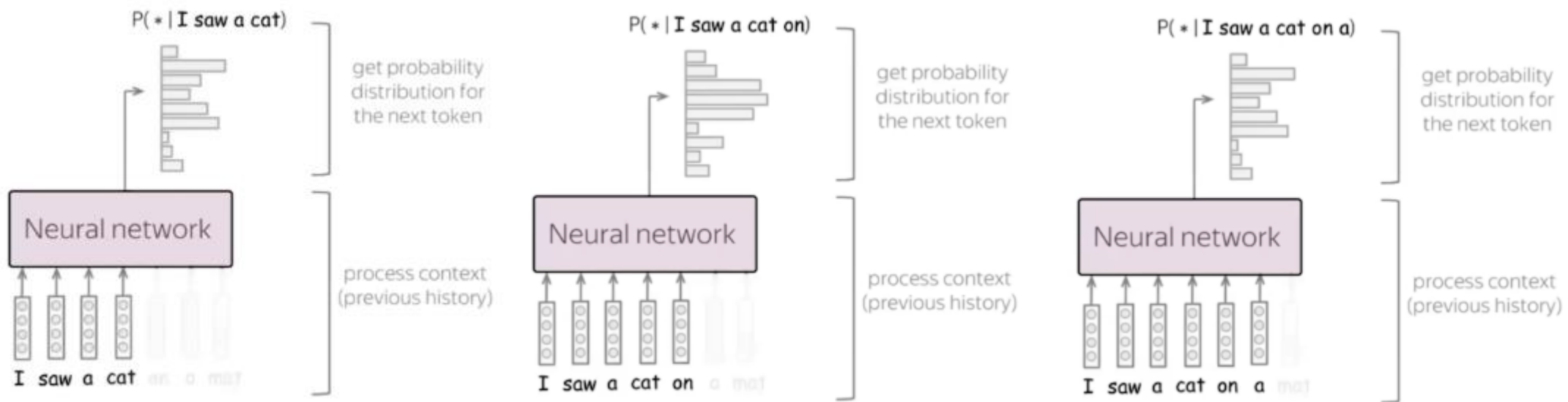
Общий подход

- получить векторное представление контекста
- оценить вероятности (прогноз распределения вероятностей для следующего токена)

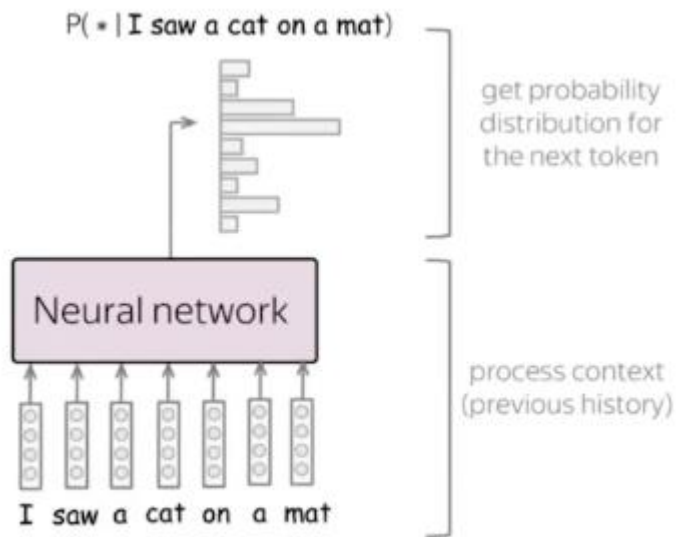
Общий подход



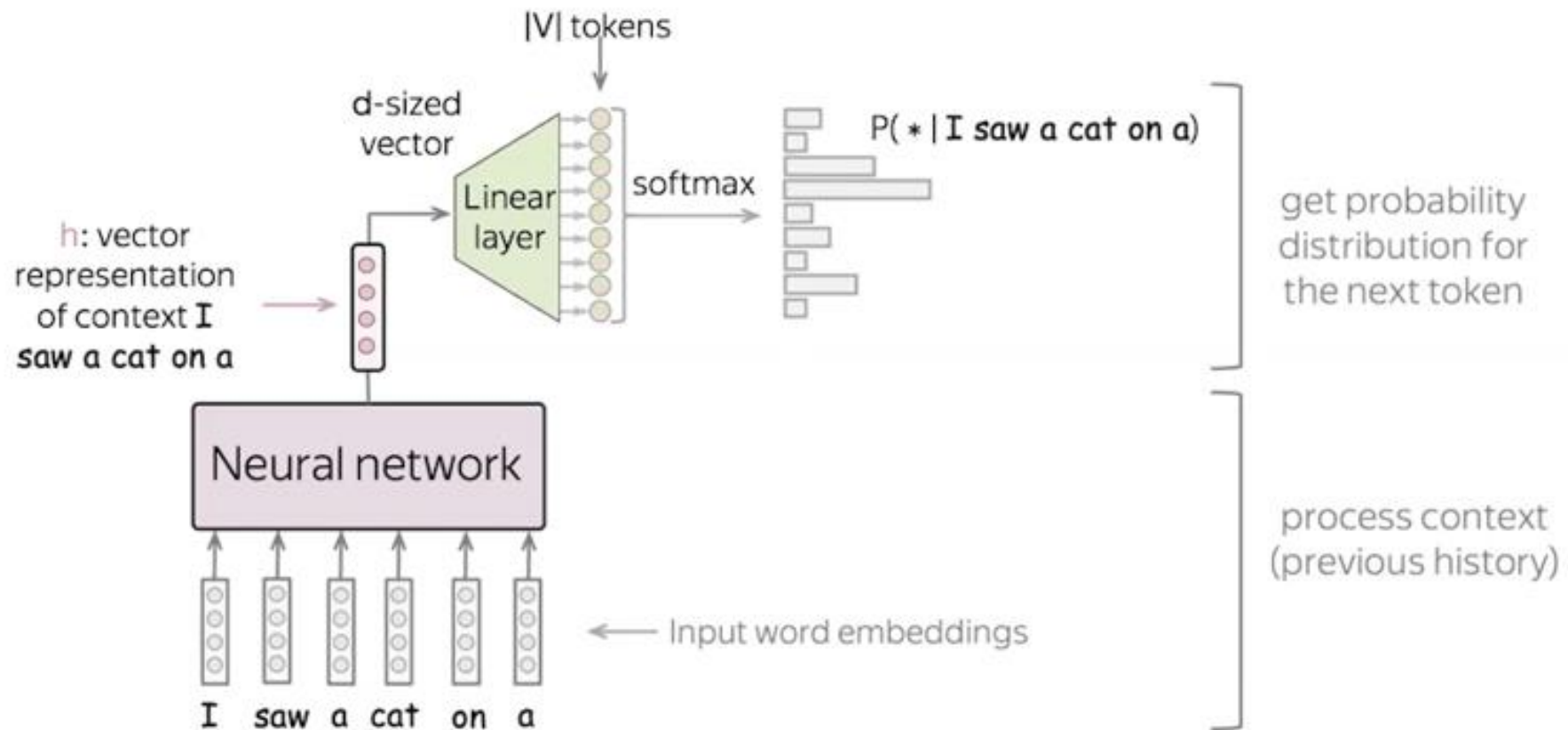
Общий подход



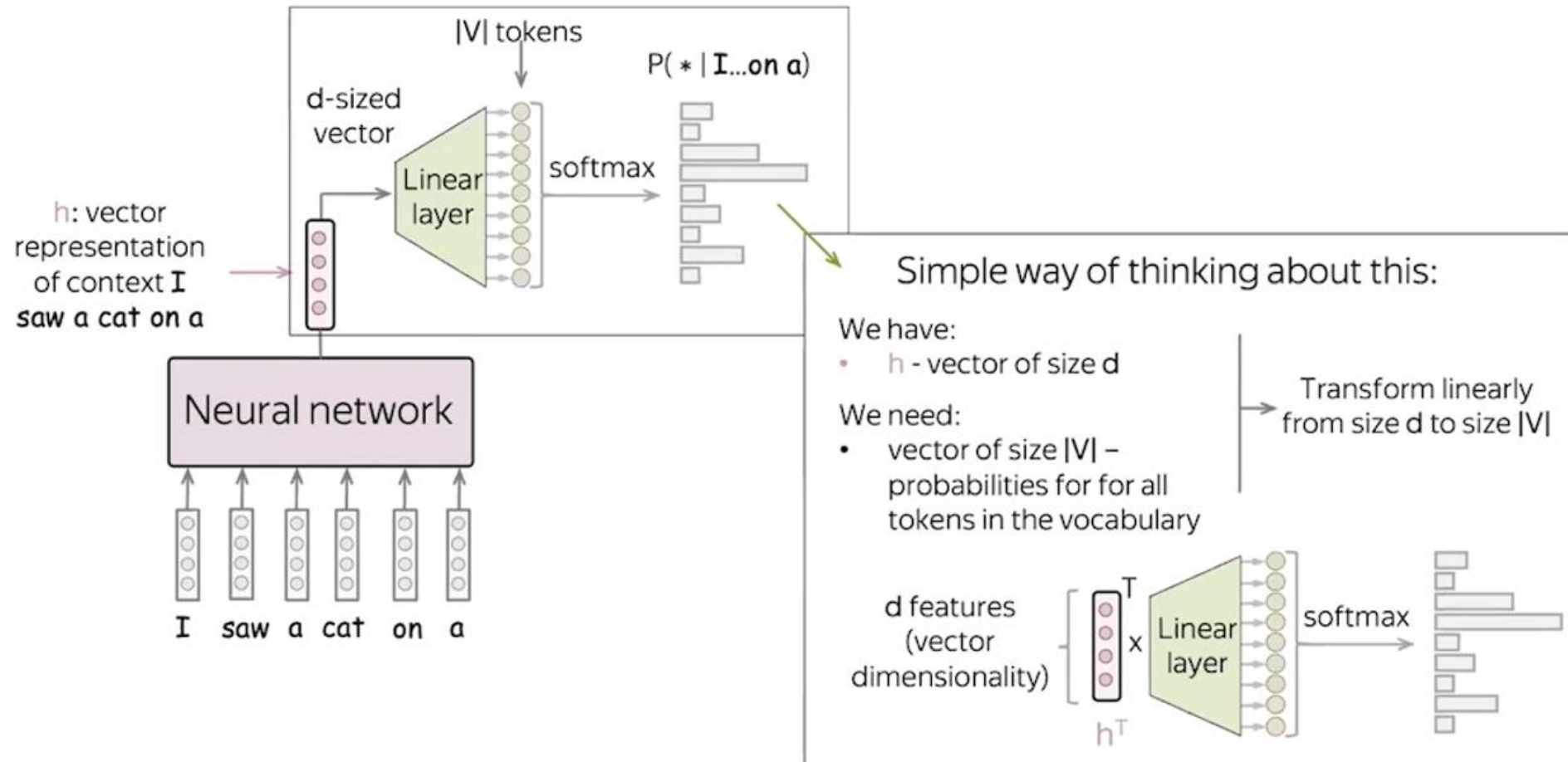
Общий подход



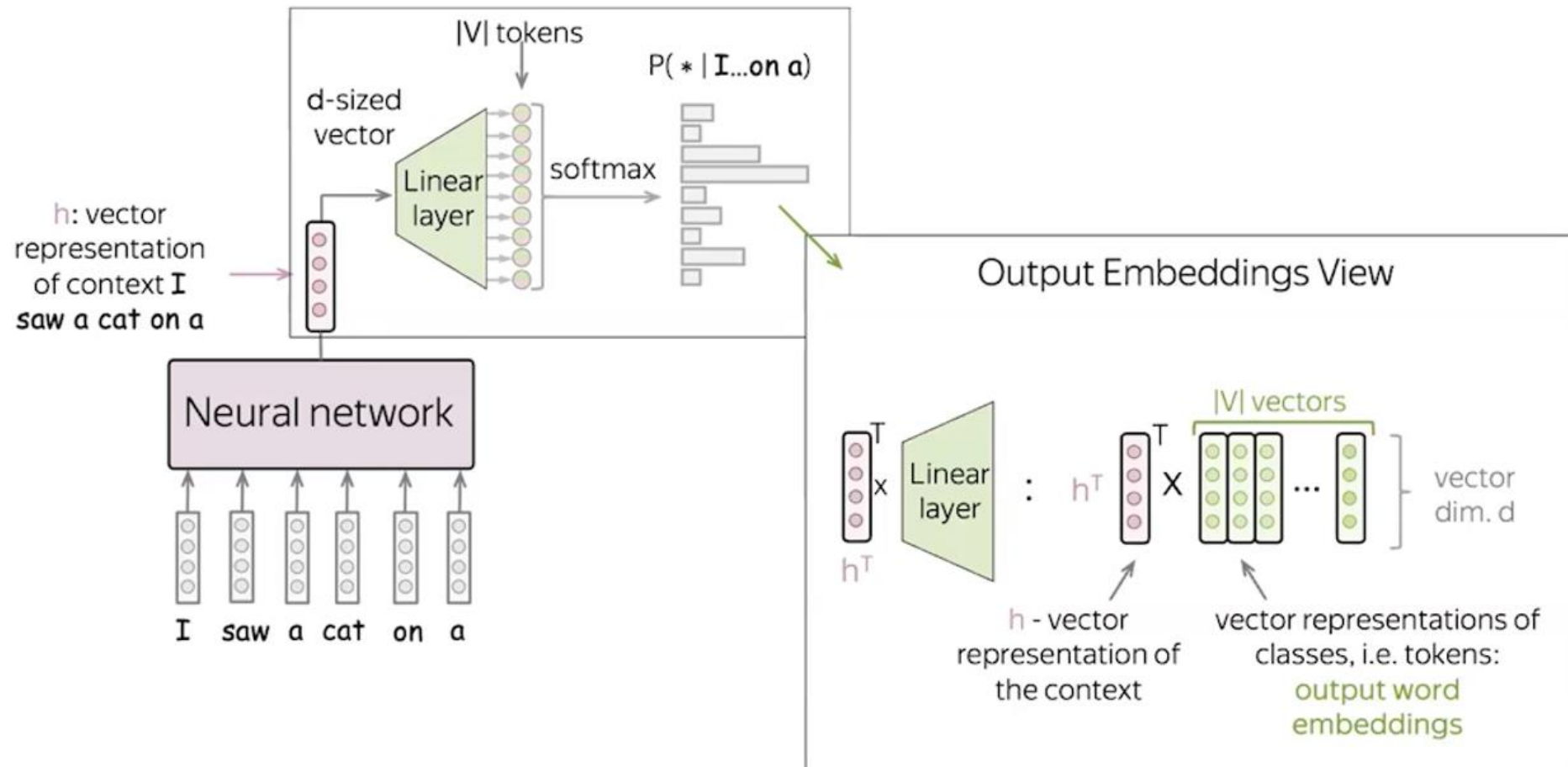
Общий подход



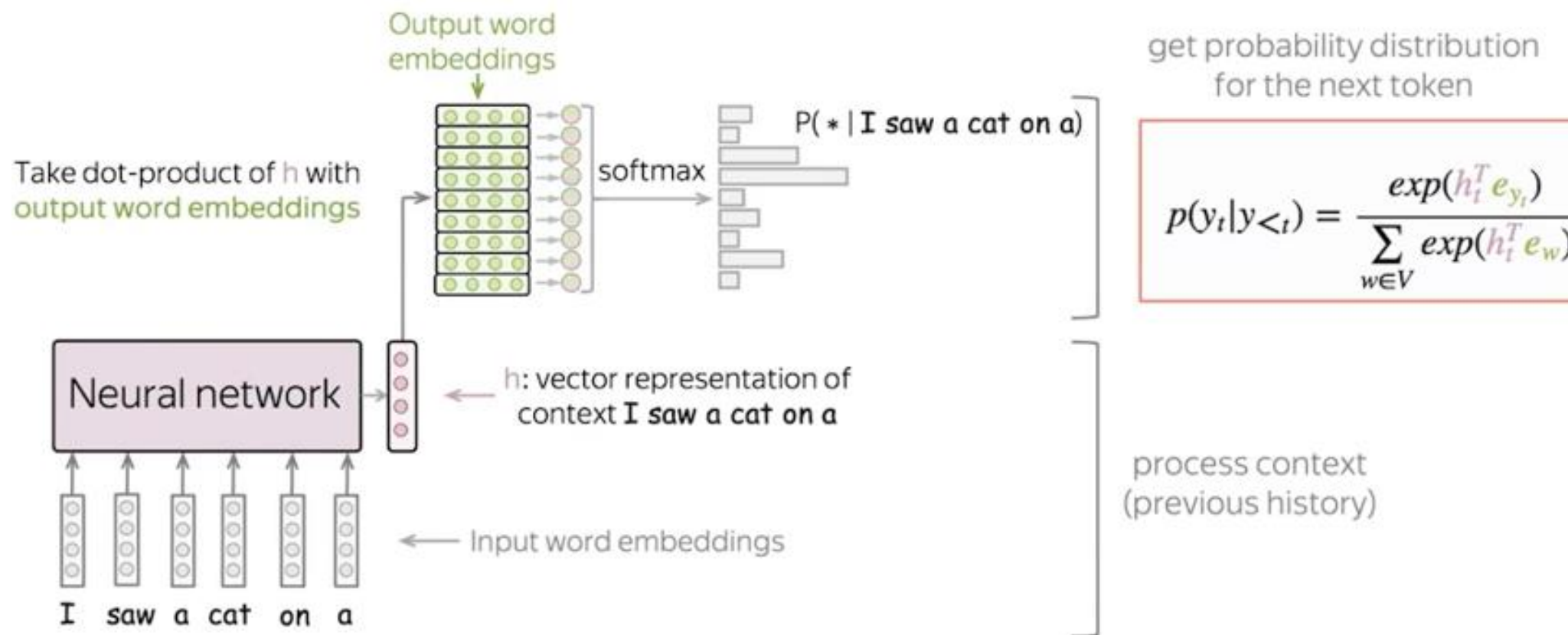
Общий подход



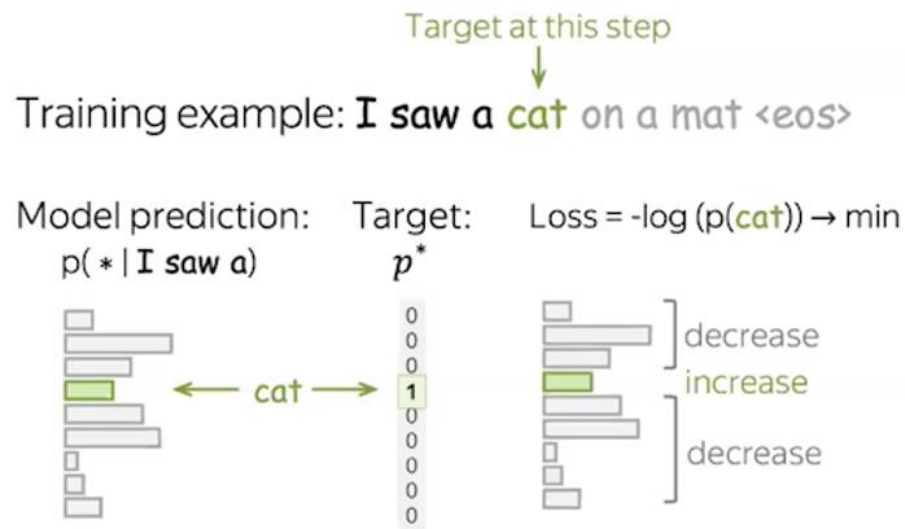
Общий подход



Представление вложений выходных данных



Обучение: кросс-энтропия



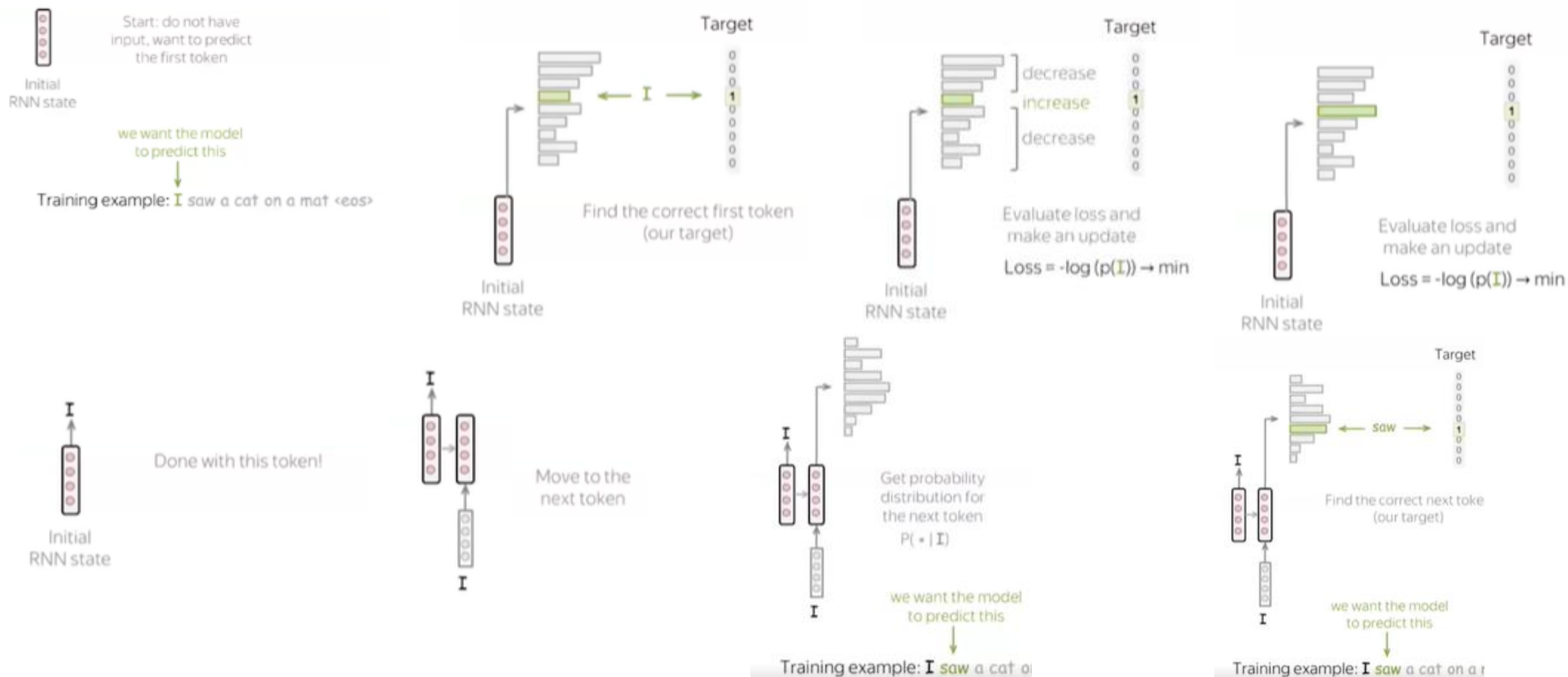
Cross-entropy loss:

$$-\sum_{i=1}^{|V|} p_i^* \cdot \log P(y_t = i | x) \rightarrow \min (p_k^* = 1, p_i^* = 0, i \neq k)$$

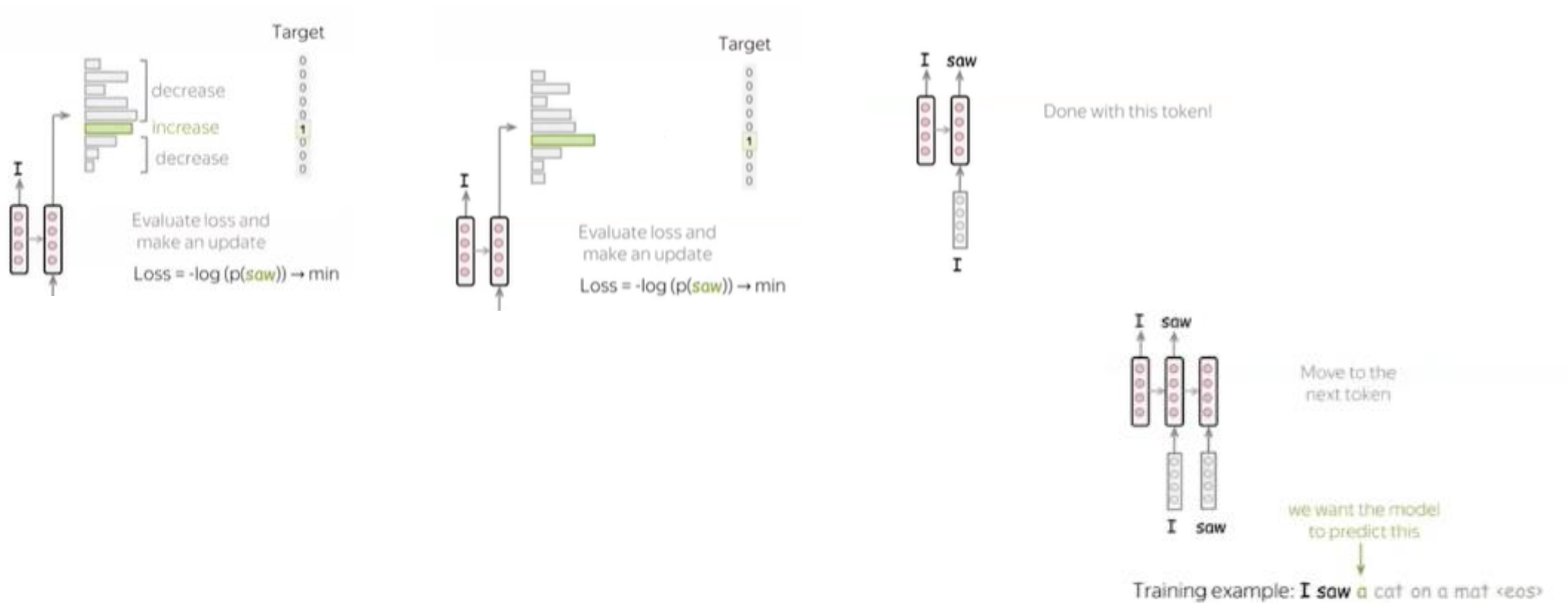
For one-hot targets, this is equivalent to

$$-\log P(y_t = \text{cat} | x) \rightarrow \min$$

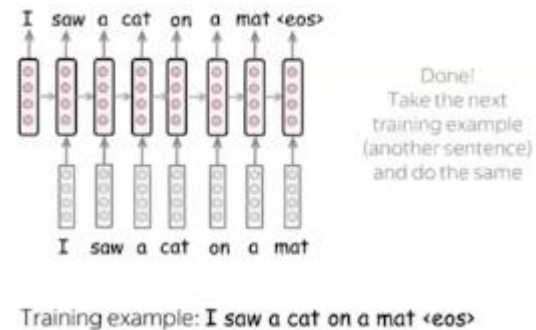
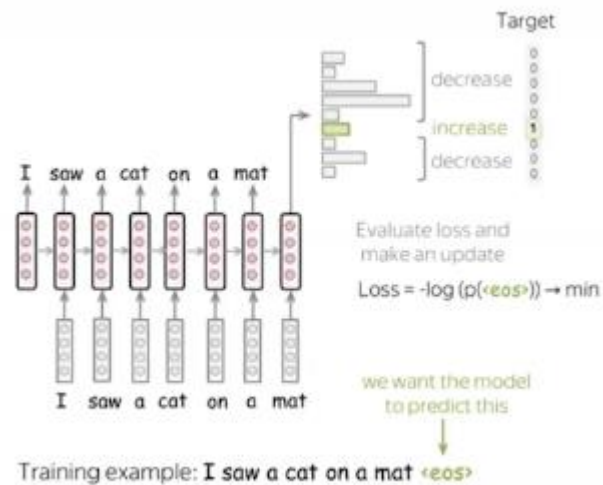
Обучение: кросс-энтропия



Обучение: кросс-энтропия

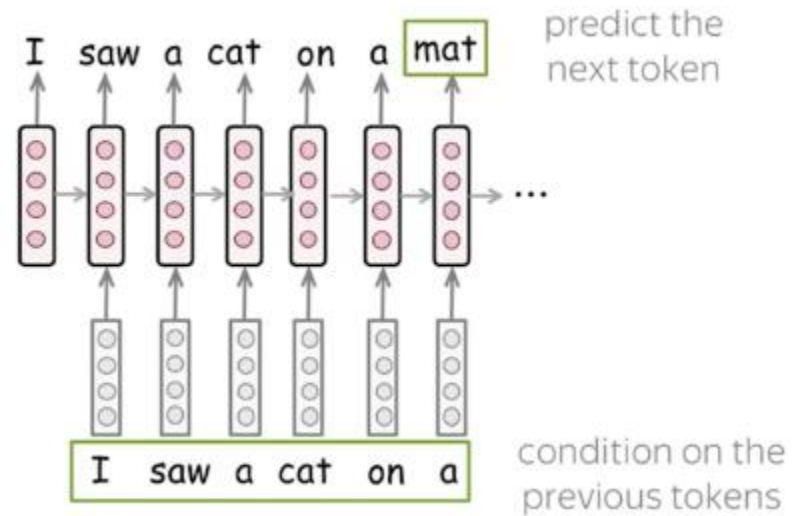


Обучение: кросс-энтропия



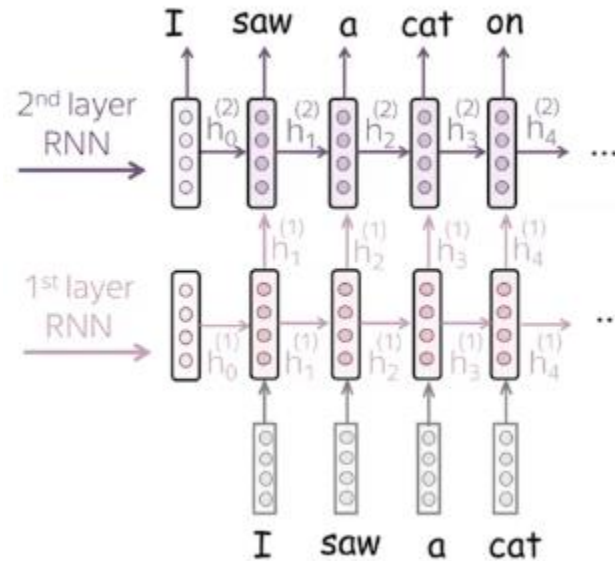
Рекуррентные модели

Читаем текст, на каждом шаге предсказываем следующий токен



Рекуррентные модели

Многослойность: передача состояний из одной RNN в другую



Связность и разнообразие

Сгенерированные тексты должны быть:

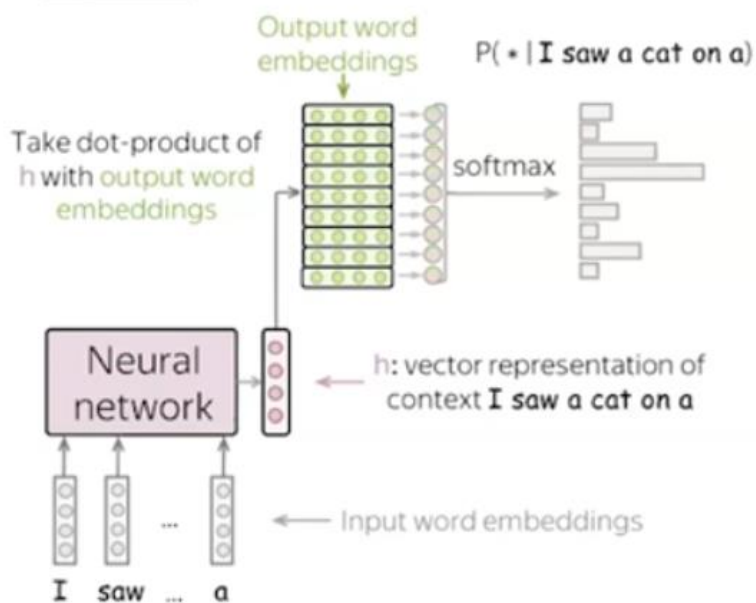
- связными и осмысленными— сгенерированный текст должен иметь смысл;
- разнообразными — модель должна быть способна создавать очень разные образцы.

Использование softmax temperature

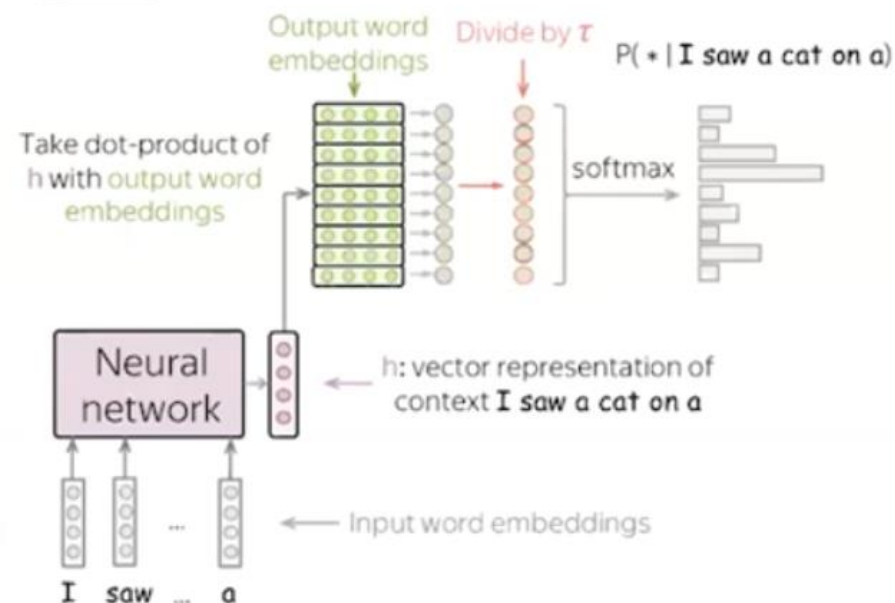
$$\frac{\exp(h^T w)}{\sum_{w_i \in V} \exp(h^T w_i)} \rightarrow \frac{\exp\left(\frac{h^T w}{\tau}\right)}{\sum_{w_i \in V} \exp\left(\frac{h^T w_i}{\tau}\right)}$$

τ - softmax temperature

Before



After



Пример: temperature = 2

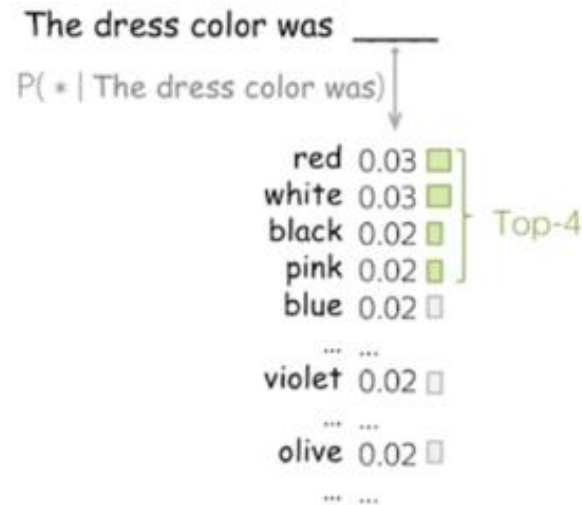
```
paradise sits farms started paint hollow almost  
unprecedented decisions, care using withdrawal from  
rebel cis ( , saying graphics mongolia official line,  
greeted agenda victor is exploring anger :) draw  
testify liberalization decay productive 2 went  
exchanges of marketing drawing enabling challenging  
systematic crisis influencing the executive arrangement  
performs designs
```

Пример: temperature = 0,2

the first time the two - year - old - old girl with a
new version of the new version of the new version of
the new version of the new version of the new version
of the new version of the new version of the new
version of the

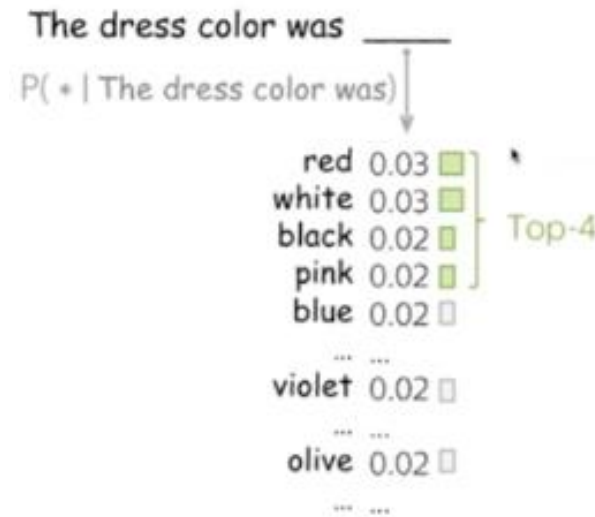
Связность и разнообразие

- выбирать из первых K наиболее вероятных токенов (обычно K составляет около 10-50)

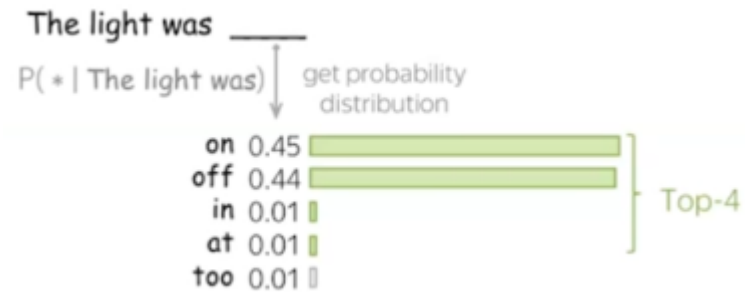


Как выбрать K?

Top-K: недостаточно

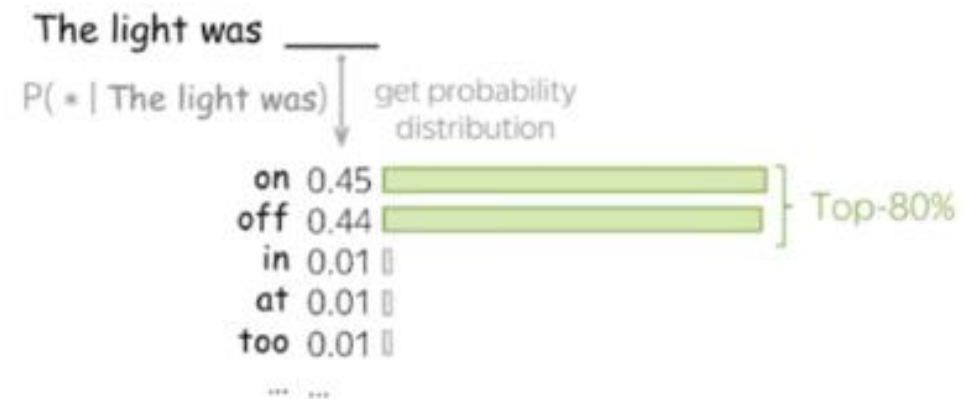
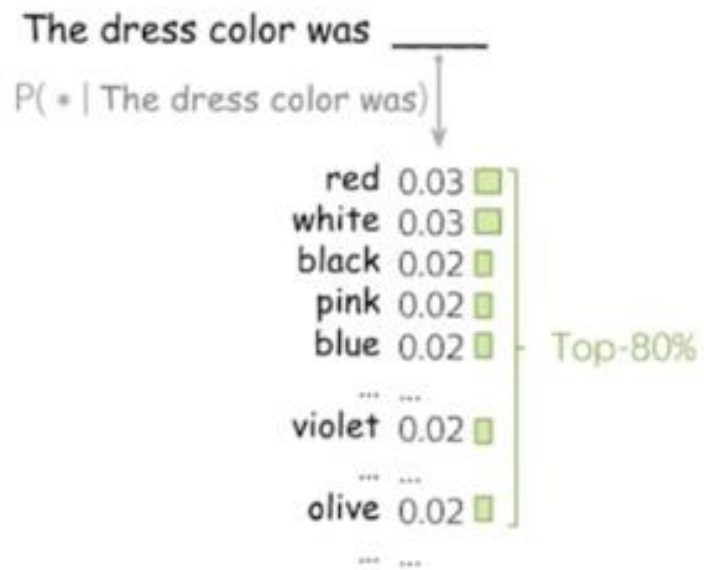


Top-K: слишком много



Top-p sampling

- на каждом шаге выбирать столько верхних токенов, чтобы покрыть $p\%$ вероятностной массы



Оценка моделей. Перплексия

Насколько модель «удивляется» при чтении нового текста?

$$L(y_{1:M}) = L(y_1, y_2, \dots, y_M) = \sum_{t=1}^M \log_2 p(y_t | y_{<t}) \quad \text{Loss}(y_{1:M}) = - \sum_{t=1}^M \log p(y_t | y_{<t})$$

Log-likelihood of the text

Note: cross-entropy (our loss)
is negative log-likelihood

$$\text{Perplexity}(y_{1:M}) = 2^{-\frac{1}{M}L(y_{1:M})}$$

Какие значения может принимать перплексия?

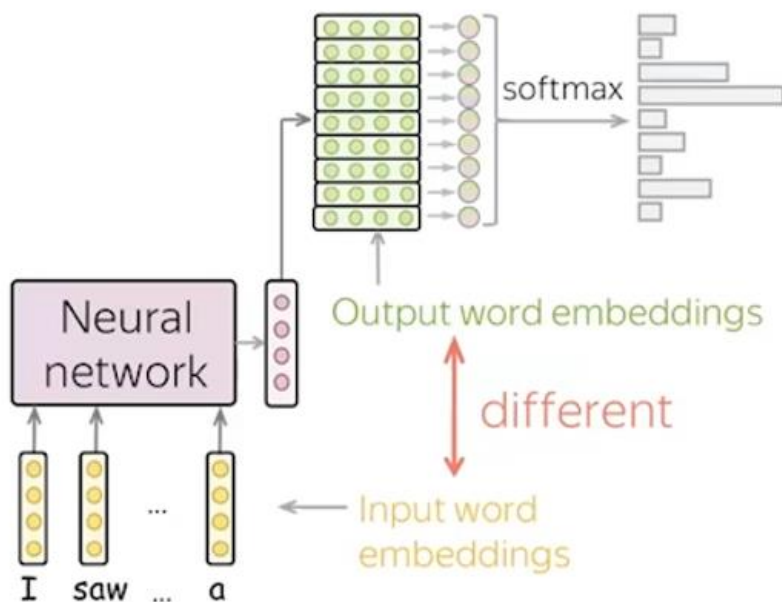
- лучшее значение – это 1 (если модель идеальна и присваивает вероятность 1 правильным токенам, то логарифмические вероятности равны нулю)
- худшее значение — это $|V|$ (если модель ничего не знает о данных, она присваивает всем токенам вероятность $1/|V|$, независимо от контекста).

$$\text{Perplexity}(y_{1:M}) = 2^{-\frac{1}{M}L(y_{1:M})}$$

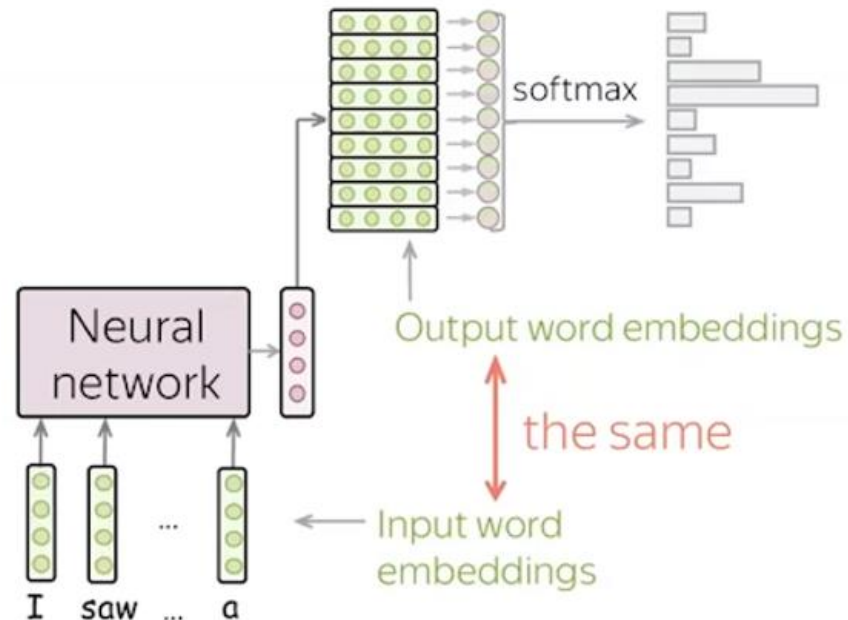
Привязка веса (Weight Tying)

Большая часть параметров модели берется из этих матриц — вы можете уменьшить размер модели!

Default (no weight tying)



Weight tying



Примеры сгенерированного текста

Proof. Omitted. □

Lemma 0.1. *Let $\mathcal{C}$ be a set of the construction.*

Let $\mathcal{C}$ be a gerber covering. Let $\mathcal{F}$ be a quasi-coherent sheaves of $\mathcal{O}$ -modules. We have to show that

$$\mathcal{O}_{\mathcal{O}_X} = \mathcal{O}_X(\mathcal{L})$$

Proof. This is an algebraic space with the composition of sheaves $\mathcal{F}$ on $X_{\text{étale}}$ we have

$$\mathcal{O}_X(\mathcal{F}) = \{\text{morph}_1 \times_{\mathcal{O}_X} (\mathcal{G}, \mathcal{F})\}$$

where $\mathcal{G}$ defines an isomorphism $\mathcal{F} \rightarrow \mathcal{F}$ of $\mathcal{O}$ -modules. □

Lemma 0.2. *This is an integer Z is injective.*

Proof. See Spaces, Lemma ?? □

Lemma 0.3. *Let S be a scheme. Let X be a scheme and X is an affine open covering. Let $U \subset X$ be a canonical and locally of finite type. Let X be a scheme. Let X be a scheme which is equal to the formal complex.*

The following to the construction of the lemma follows.

Let X be a scheme. Let X be a scheme covering. Let

$$b : X \rightarrow Y' \rightarrow Y \rightarrow Y' \times_X Y \rightarrow X.$$

be a morphism of algebraic spaces over S and Y .

Proof. Let X be a nonzero scheme of X . Let X be an algebraic space. Let $\mathcal{F}$ be a quasi-coherent sheaf of $\mathcal{O}_X$ -modules. The following are equivalent

- (1) $\mathcal{F}$ is an algebraic space over S .
- (2) If X is an affine open covering.

Consider a common structure on X and X the functor $\mathcal{O}_X(U)$ which is locally of finite type. □

This since $\mathcal{F} \in \mathcal{F}$ and $x \in \mathcal{G}$ the diagram

$$\begin{array}{ccc} S & \xrightarrow{\quad} & \\ \downarrow & & \downarrow \\ \xi & \xrightarrow{\quad} & \mathcal{O}_{X'} \\ \downarrow g \circ \pi_x & & \downarrow \\ & \xrightarrow{\quad} & \alpha' \\ & \downarrow & \downarrow \\ & \xrightarrow{\quad} & \alpha \end{array} \quad \begin{array}{ccc} & & X \\ & & \downarrow \\ & & \text{Mor}_{\text{sets}} \\ & & d(\mathcal{O}_{X_{X/\mathcal{K}}}, \mathcal{G}) \end{array}$$

is a limit. Then $\mathcal{G}$ is a finite type and assume S is a flat and $\mathcal{F}$ and $\mathcal{G}$ is a finite type f_* . This is of finite type diagrams, and

- the composition of $\mathcal{G}$ is a regular sequence,
- $\mathcal{O}_{X'}$ is a sheaf of rings.

□

Proof. We have see that $X = \text{Spec}(R)$ and $\mathcal{F}$ is a finite type representable by algebraic space. The property $\mathcal{F}$ is a finite morphism of algebraic stacks. Then the cohomology of X is an open neighbourhood of U . □

Proof. This is clear that $\mathcal{G}$ is a finite presentation, see Lemmas ??.

A reduced above we conclude that U is an open covering of $\mathcal{C}$. The functor $\mathcal{F}$ is a "field"

$$\mathcal{O}_{X,S} \rightarrow \mathcal{F}_Y \rightarrow I(\mathcal{O}_{X_{\text{étale}}}) \rightarrow \mathcal{O}_{X_1}^{-1} \mathcal{O}_{X_1}(\mathcal{O}_{X_1}^{\mathcal{F}})$$

is an isomorphism of covering of $\mathcal{O}_{X_1}$. If $\mathcal{F}$ is the unique element of $\mathcal{F}$ such that X is an isomorphism.

The property $\mathcal{F}$ is a disjoint union of Proposition ?? and we can filtered set of presentations of a scheme $\mathcal{O}_X$ -algebra with $\mathcal{F}$ are opens of finite type over S .

If $\mathcal{F}$ is a scheme theoretic image points. □

If $\mathcal{F}$ is a finite direct sum $\mathcal{O}_{X_1}$ is a closed immersion, see Lemma ?? This is a sequence of $\mathcal{F}$ is a similar morphism.

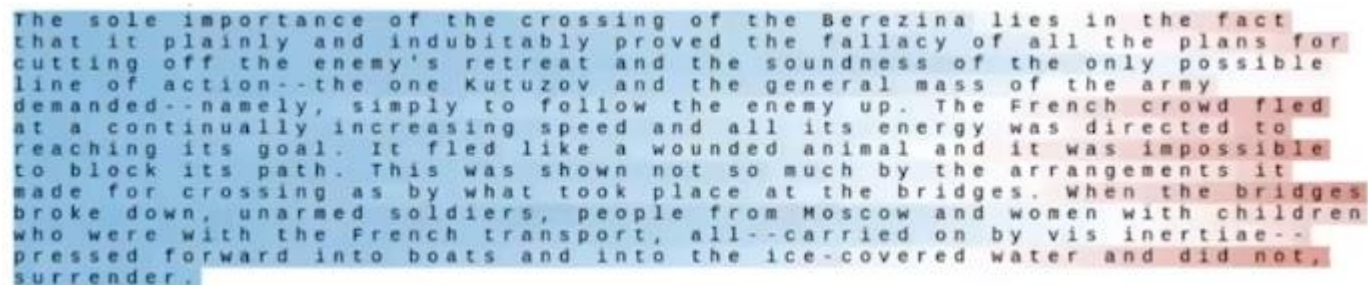
Примеры сгенерированного текста

```
/*
 * Increment the size file of the new incorrect BT_FILTER group information
 * of the size generatively.
 */
static int indicate_policy(void)
{
    int error;
    if (fd == KERN_EPT) {
        /*
         * The kernel blank will could it to use space.
         */
        if (ss->segment < mem_total)
            unblock_graph_and_set_blocked();
        else
            ret = 1;
        goto bail;
    }
    seqaddr = in_SB(in.saddr);
    selector = seq / 16;
    setup_works = true;
    for (i = 0; i < blocks; i++) {
        seq = buf[i++];
        bpf = bd->bd.next + i * search;
        if (fd) {
            current = blocked;
        }
    }
    rw->name = "Getjbbregs";
    bprm_self_clear(&iv->version);
    regs->new = blocks[(BPF_STATS << info->historical) | PFMR_CLOUATHING_SECONDS << 12];
    return seqtable;
}
```

Визуализация активации нейронов

LSTM-сети на уровне символов, обученные на ядре Linux и «Войне и мире»

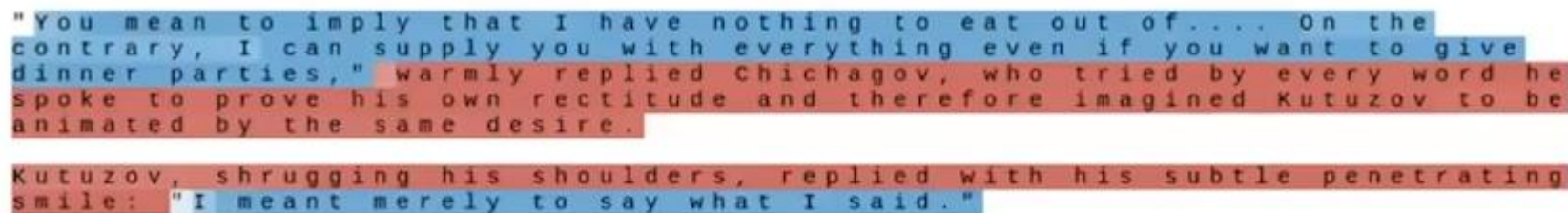
- Нейрон чувствителен к положению в строке



The sole importance of the crossing of the Berezina lies in the fact that it plainly and indubitably proved the fallacy of all the plans for cutting off the enemy's retreat and the soundness of the only possible line of action--the one Kutuzov and the general mass of the army demanded--namely, simply to follow the enemy up. The French crowd fled at a continually increasing speed and all its energy was directed to reaching its goal. It fled like a wounded animal and it was impossible to block its path. This was shown not so much by the arrangements it made for crossing as by what took place at the bridges. When the bridges broke down, unarmed soldiers, people from Moscow and women with children who were with the French transport, all--carried on by vis inertiae--pressed forward into boats and into the ice-covered water and did not, surrender.

This visualization shows a heatmap where the background color of each word indicates its activation level. The text is divided into two main horizontal bands: a blue band at the top and a red band at the bottom. The blue band covers the first part of the text, while the red band covers the second part. The transition between the two bands occurs around the middle of the sentence, highlighting the model's sensitivity to the position of words within the overall structure of the paragraph.

- Нейрон активируется на внутренние кавычки



"You mean to imply that I have nothing to eat out of.... On the contrary, I can supply you with everything even if you want to give dinner parties," warmly replied Chichagov, who tried by every word he spoke to prove his own rectitude and therefore imagined Kutuzov to be animated by the same desire.

Kutuzov, shrugging his shoulders, replied with his subtle penetrating smile: "I meant merely to say what I said."

This visualization shows a heatmap where the background color of each word indicates its activation level. The text is divided into two main horizontal bands: a blue band at the top and a red band at the bottom. The blue band covers the first part of the text, while the red band covers the second part. The transition between the two bands occurs around the middle of the sentence, highlighting the model's sensitivity to the position of words within the overall structure of the paragraph.

Визуализация активации нейронов

- Нейрон активируется внутри операторов if

```
static int __dequeue_signal(struct sigpending *pending, sigset_t *mask,
                           siginfo_t *info)
{
    int sig = next_signal(pending, mask);
    if (sig) {
        if (current->notifier) {
            if (sigismember(current->notifier_mask, sig)) {
                if (!(current->notifier)(current->notifier_data)) {
                    clear_thread_flag(TIF_SIGPENDING);
                    return 0;
                }
            }
        }
        collect_signal(sig, pending, info);
    }
    return sig;
}
```


Визуализация активации нейронов

- Нейрон активируется на внутренние комментарии и цитаты

```
/* Duplicate LSM field information. The lsm_rule is opaque, so
 * re-initialized. */
static inline int audit_dupe_lsm_field(struct audit_field *df,
                                     struct audit_field *sf)
{
    int ret = 0;
    char *lsm_str;
    /* our own copy of lsm_str */
    lsm_str = kstrdup(sf->lsm_str, GFP_KERNEL);
    if (unlikely(!lsm_str))
        return -ENOMEM;
    df->lsm_str = lsm_str;
    /* our own (refreshed) copy of lsm_rule */
    ret = security_audit_rule_init(df->type, df->op, df->lsm_str,
                                   (void **)&df->lsm_rule);
    /* Keep currently invalid fields around in case they
     * become valid after a policy reload. */
    if (ret == -EINVAL) {
        pr_warn("audit rule for LSM '%s' is invalid\n",
                df->lsm_str);
        ret = 0;
    }
}
```


Визуализация активации нейронов

- Нейрон чувствителен к глубине кода

```
#ifdef CONFIG_AUDITSYSCALL
static inline int audit_match_class_bits(int class, u32 *mask)
{
    int i;
    if (classes[class]) {
        for (i = 0; i < AUDIT_BITMASK_SIZE; i++)
            if (mask[i] & classes[class][i])
                return 0;
    }
    return 1;
}
```

Визуализация активации нейронов

- Нейрон активируется перед началом новой строки

```
char *audit_unpack_string(void **bufp, size_t *remain, si
{
    char *str;
    if (!*bufp || (len == 0) || (len > *remain))
        return ERR_PTR(-EINVAL);
    /* Of the currently implemented string fields, PATH_MAX
     * defines the longest valid length.
     */
    if (len > PATH_MAX)
        return ERR_PTR(-ENAMETOOLONG);
    str = kmalloc(len + 1, GFP_KERNEL);
    if (unlikely(!str))
        return ERR_PTR(-ENOMEM);
    memcpy(str, *bufp, len);
    str[len] = 0;
    *bufp += len;
    *remain -= len;
    return str;
}
```

Нейроны, следящие за настроением

LSTM-сеть на уровне символов, обученная на отзывах Amazon

This is one of Crichton's best books. The characters of Karen Ross, Peter Elliot, Munro, and Amy are beautifully developed and their interactions are exciting, complex, and fast-paced throughout this impressive novel. And about 99.8 percent of that got lost in the film. Seriously, the screenplay AND the directing were horrendous and clearly done by people who could not fathom what was good about the novel. I can't fault the actors because frankly, they never had a chance to make this turkey live up to Crichton's original work. I know good novels, especially those with a science fiction edge, are hard to bring to the screen in a way that lives up to the original. But this may be the absolute worst disparity in quality between novel and screen adaptation ever. The book is really, really good. The movie is just dreadful.

Интерпретируемые нейроны для управления генерируемыми текстами

SENTIMENT FIXED TO POSITIVE

I couldn't figure out the shape at first but it definitely does what it's meant to do. It's a great product and I recommend it highly

I couldn't figure out why this movie had been discontinued! Now I can enjoy it anytime I like. So glad to have found it again.

I couldn't figure out how to use the video or the book that goes along with it, but it is such a fantastic book on how to put it into practice!

I couldn't figure out how to use just one and my favorite running app. I use it all the time. Good quality, You cant beat the price.

I couldn't figure out how to attach these balls to my little portable drums, but these fit the bill and were well worth every penny.

SENTIMENT FIXED TO NEGATIVE

I couldn't figure out how to use the product. It did not work. At least there was no quality control; this tablet does not work. I would have given it zero stars, but that was not an option.

I couldn't figure out how to set it up being that there was no warning on the box. I wouldn't recommend this to anyone.

I couldn't figure out how to use the gizmo. What a waste of time and money. Might as well through away this junk.

I couldn't figure out how to stop this drivel. At worst, it was going absolutely nowhere, no matter what I did. Needless to say, I skim-read the entire book. Don't waste your time.

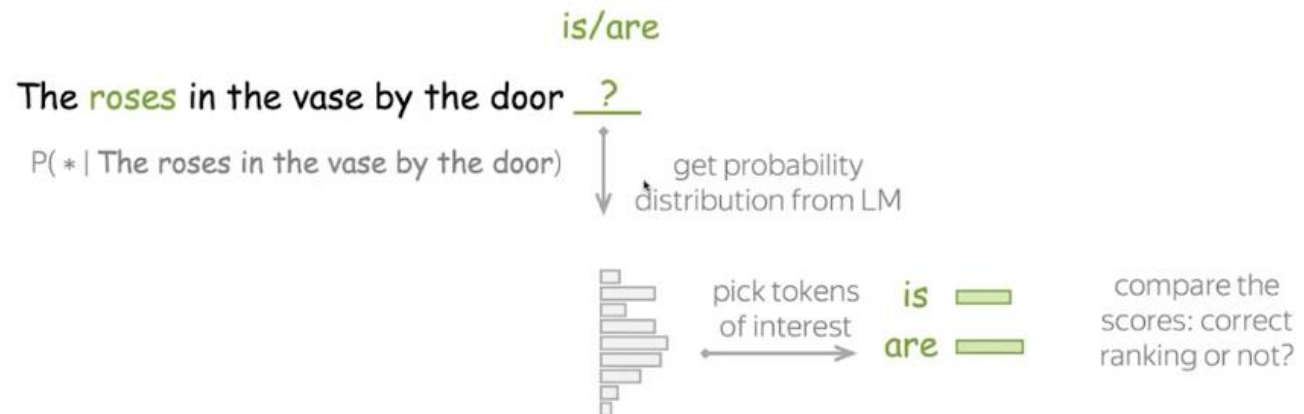
I couldn't figure out how to play it.

Контрастная оценка: специфический текст

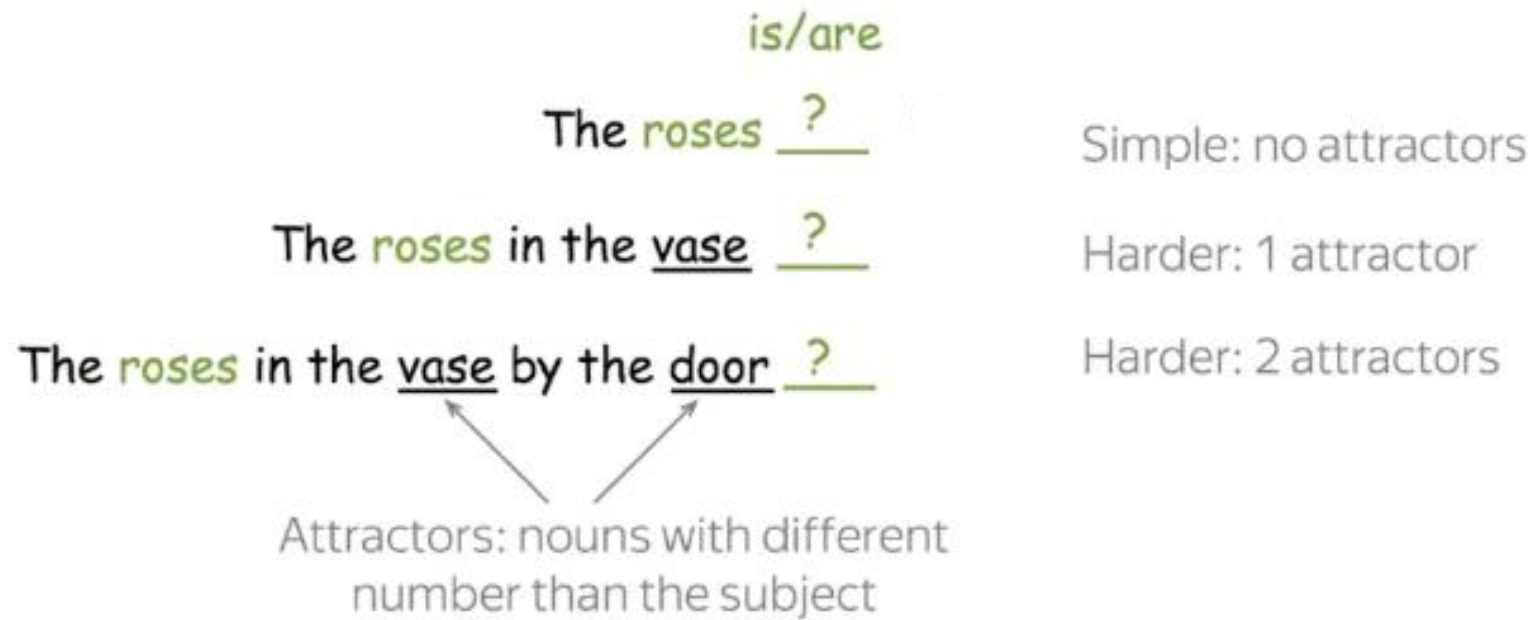
The **roses** in the vase by the door ? Competing answers: **is**, **are**

$P(\text{The } \text{roses} \text{ in the vase by the door } \text{are})$
 $P(\text{The } \text{roses} \text{ in the vase by the door } \text{is})$

Is the correct answer ranked higher?
 $P(\dots \text{are}) > P(\dots \text{is})?$



Контрастная оценка: специфический текст



ЛР 2

Квантизация - это метод снижения вычислительных затрат и затрат памяти при выполнении логического вывода за счет представления весов и активаций с использованием типов данных низкой точности.

Quantization_config — это параметр (аргумент) метода `from_pretrained` в библиотеке Hugging Face transformers, который используется для передачи конфигурации квантования в модель.

BitsAndBytesConfig — это класс из библиотеки transformers от Hugging Face, который предназначен для настройки и управления квантованием моделей с помощью библиотеки bitsandbytes.

ЛР 3

Pydantic играет ключевую роль при работе со структурированными данными в LLM.

Самое главное применение — заставить LLM возвращать данные в строго определённом формате.

JSON string — это строка текста, которая содержит данные в формате JSON (JavaScript Object Notation).

JSON string — это **текстовое представление структурированных данных**, которое выглядит как JavaScript-объект, но является обычной строкой.