Лекция 2.
Классификация текстов

Классификация текстов

- Примеры классификационных задач
- Общий вид: Особенности + Классификатор
- Модели: генеративные и дискриминативные
- Классические методы
- Нейросетевые методы
- Классификация с несколькими метками
- Практические советы
- Анализ и интерпретируемость

Классификация текстов

- Многоклассовая классификация (Multi-class) (много меток, правильной является только одна из них)
- Бинарная классификация (Binary) (две метки, правильной является только одна из них)
- Многометочная классификация (Multi-label) (много меток, один текст может иметь несколько правильных меток)

Датасеты для классификации текстов

Датасеты сильно различаются по:

- Типу
- Количеству меток
- Размеру
- Средней длине в токенах (числу, которое показывает, сколько в среднем токенов содержится в одном тексте датасета)

Датасеты для классификации текстов

Dataset	Type	Number of labels	Size (train/test)	Avg. length (tokens)
SST	sentiment	5 or 2	8.5k / 1.1k	19
IMDb Review	sentiment	2	25k / 25k	271
Yelp Review	sentiment	5 or 2	650k / 50k	179
Amazon Review	sentiment	5 or 2	3m / 650k	79
TREC	question	6	5.5k / 0.5k	10
Yahoo! Answers	question	10	1.4m / 60k	131
AG's News	topic	4	120k / 7.6k	44
Sogou News	topic	6	54k / 6k	737
DBPedia	topic	14	560k / 70k	67

Датасеты для классификации текстов

- **SST**

Stanford Sentiment Treebank — это классический и очень известный датасет для оценки тональности текста (sentiment analysis). Предложения взяты из кинокритики (сайт Rotten Tomatoes).
Исходная разметка использует 5 классов:

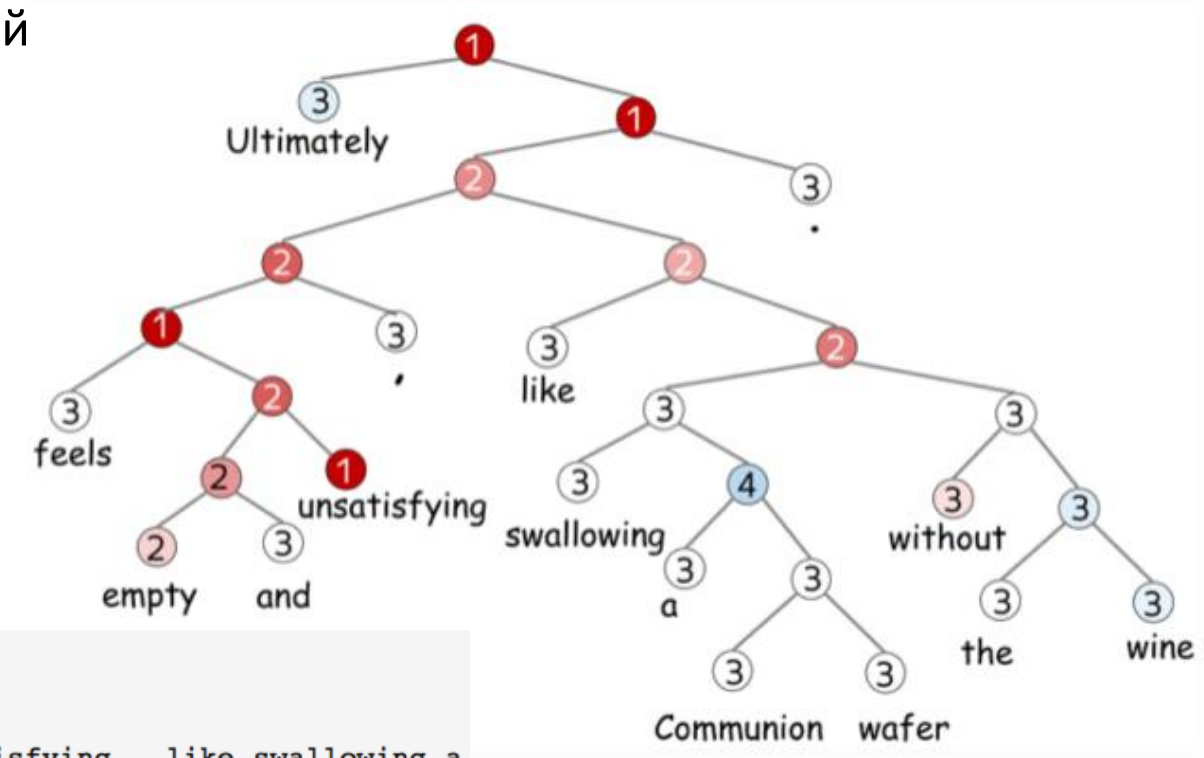
0 — Very Negative

1 — Negative

2 — Neutral

3 — Positive

4 — Very Positive




```
Label: 1
Review:
Ultimately feels empty and unsatisfying , like swallowing a
Communion wafer without the wine .
```

Датасеты для классификации текстов

- **IMDb Review**

IMDb (Internet Movie Database) — это огромная онлайн-база данных о фильмах, телесериалах и видео играх. Датасет **IMDb Review** представляет собой коллекцию отзывов зрителей на фильмы, размещенных на этом сайте

Задача: Бинарная классификация тональности (Sentiment Analysis)
Исходная разметка использует 2 класса:

Negative (0) — отрицательный отзыв (оценка пользователя ≤ 4 из 10)

Positive (1) — положительный отзыв (оценка пользователя ≥ 7 из 10)

Датасеты для классификации текстов

- **IMDb Review**

Label: negative

Review
Hobgoblins .... Hobgoblins .... where do I begin?!?

This film gives Manos - The Hands of Fate and Future War a run for their money as the worst film ever made . This one is fun to laugh at , where as Manos was just painful to watch . Hobgoblins will end up in a time capsule somewhere as the perfect movie to describe the term : " 80 's cheeze " . The acting (and I am using this term loosely) is atrocious , the Hobgoblins are some of the worst puppets you will ever see , and the garden tool fight has to be seen to be believed . The movie was the perfect vehicle for MST3 K , and that version is the only way to watch this mess . This movie gives Mike and the bots lots of ammunition to pull some of the funniest one - liners they have ever done . If you try to watch this without the help of Mike and the bots ..... God help you ! !

Датасеты для классификации текстов

- **Yelp Review**

Основан на реальных отзывах с платформы Yelp.
Yelp — это платформа, где пользователи оставляют отзывы о ресторанах, кафе, магазинах, салонах красоты и т.д. Датасет **Yelp Review** представляет собой огромную выборку этих отзывов, публично предоставленную Yelp для исследовательских целей.
Задача: Многоглассовая классификация

Классы: Отзывы размечены количеством звезд (1-5), что естественным образом соответствует тональности:

- **1 звезда** — Very Negative (очень негативный)
- **2 звезды** — Negative (негативный)
- **3 звезды** — Neutral (нейтральный)
- **4 звезды** — Positive (позитивный)
- **5 звезд** — Very Positive (очень позитивный)

Датасеты для классификации текстов

- **Yelp Review**

Label: 4

Review
I had a serious craving for Roti. So glad I found this place. A very small menu selection but it had exactly what I wanted. The serving for $8.20 after tax is enough for 2 meals. I know where to go from now on for a great meal with leftovers. This is a noteworthy place to bring my Uncle T.J. who's a Trini when he comes to visit.

Датасеты для классификации текстов

- **Amazon Review**

Датасет Amazon Review — это один из самых масштабных и значимых ресурсов в мире NLP и Data Science. Он представляет собой колоссальную коллекцию отзывов пользователей на товары, продаваемые на Amazon.com.

Задача: Многоклассовая классификация

Классы: Отзывы размечены количеством звезд (1-5), что естественным образом соответствует тональности:

- **1 звезда** — Very Negative (очень негативный)
- **2 звезды** — Negative (негативный)
- **3 звезды** — Neutral (нейтральный)
- **4 звезды** — Positive (позитивный)
- **5 звезд** — Very Positive (очень позитивный)

Датасеты для классификации текстов

- **Amazon Review**

```
Label: 3

Review Title: Simple

Review Content:
This book was not anything special. Although I love
romances, it was too simple. The symbolism was spelled out
to the readers in a blunt manner. The less educated readers
may appreciate it. The wording was quite beautiful at times
and the plot was enchanting (perfect for a movie) but it is
not heart wrenching like the movie Titantic (which was a
must see!) ;)
```

Датасеты для классификации текстов

- **TREC QA Track**

Text REtrieval Conference — классификация вопросов по типам.
Задача: Определить тип ожидаемого ответа на вопрос

Классы: Иерархическая система меток (6 грубых классов, 50 тонких).

- **ABBR** (abbreviation) — аббревиатуры и расшифровки.
- **DESC** (description) — описания, определения, объяснения.
- **ENTY** (entity) — сущности (лица, организации, продукты и т.д.).
- **HUM** (human) — люди.
- **LOC** (location) — местоположения.
- **NUM** (numeric) — числовые значения (цена, население, даты).

Датасеты для классификации текстов

- **Yahoo! Answers**

Датасет Yahoo! Answers — коллекция вопросов пользователей, ответов и лучших ответов на эти вопросы .

Задача: Тематическая классификация (Topic Classification). Вопросы и ответы относятся к одной из множества тематических категорий

Классы: 10 крупнейших основных категорий:

Society & Culture (Общество и культура), **Science & Mathematics** (Наука и математика), **Health** (Здоровье), **Education & Reference** (Образование и справочные материалы), **Computers & Internet** (Компьютеры и интернет), **Sports** (Спорт), **Business & Finance** (Бизнес и финансы), **Entertainment & Music** (Развлечения и музыка), **Family & Relationships** (Семья и отношения), **Politics & Government** (Политика и правительство).

Датасеты для классификации текстов

Label: Society & Culture

Question Title: Why do people have the bird, turkey for thanksgiving?

Question Content: Why this bird? Any Significance?

Best Answer
It is believed that the pilgrims and indians shared wild turkey and venison on the original Thanksgiving.

Turkey's "Americanness" was established by Benjamin Franklin, who had advocated for the turkey, not the bald eagle, becoming the national bird.

Датасеты для классификации текстов

- **DBpedia**

Датасет DBpedia — это глобальный общедоступный проект, целью которого является извлечение структурированной информации из Wikipedia и предоставление к ней доступа в виде связанных данных (Linked Data)..
Источник: Wikipedia (инфобоксы, категории, геокоординаты, ссылки и т.д.)
Для ML-сообщества наиболее известен именно **датасет для классификации** (DBpedia Classification).

Число классов: 14

Данные для каждого примера:

- **Название сущности** (название статьи Wikipedia).
- **Абстракт (аннотация)** — короткое текстовое описание сущности, извлеченное из первого абзаца соответствующей статьи Wikipedia.
- **Класс** — одна из 14 категорий DBpedia.

Датасеты для классификации текстов

Label: Artist

Title: Esfandiar Monfaredzadeh

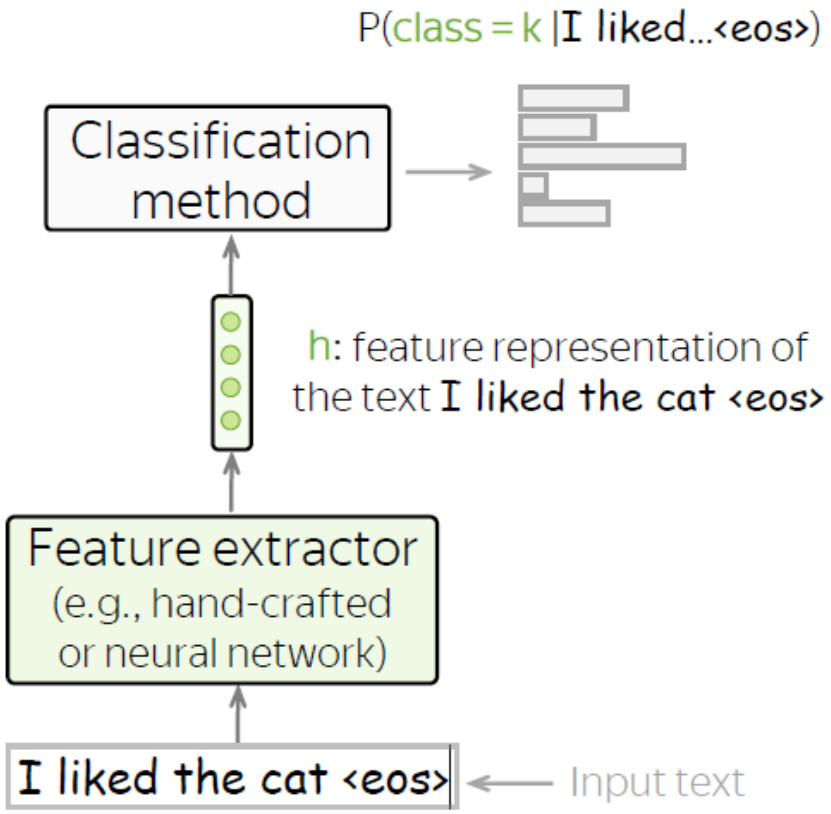
Abstract

Esfandiar Monfaredzadeh (Persian : اسفندیار منفردزاده) is an Iranian composer and director. He was born in 1941 in Tehran His major works are Gheisar Dash Akol Tangna Gavaznha. He has 2 daughters Bibinaz Monfaredzadeh and Sanam Monfaredzadeh Woods (by marriage).

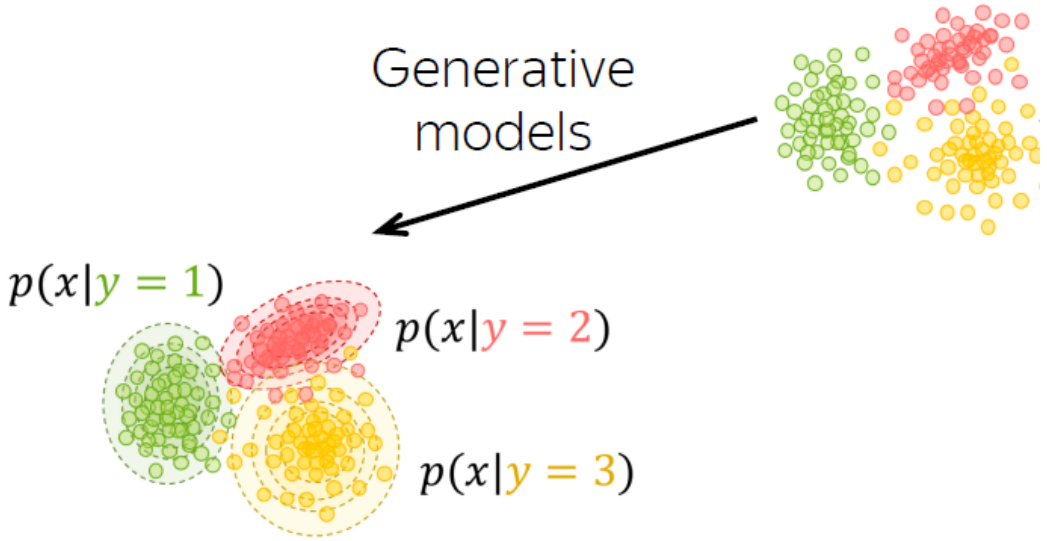
Получение представлений признаков и классификация

Извлечение признаков:
- Классические методы – признаки извлекаются вручную человеком
- Нейросетевые методы – признаки извлекаются сетью: сеть запоминает, что важно

Методы классификации:
- генеративные
- дискриминативные



Генеративные модели



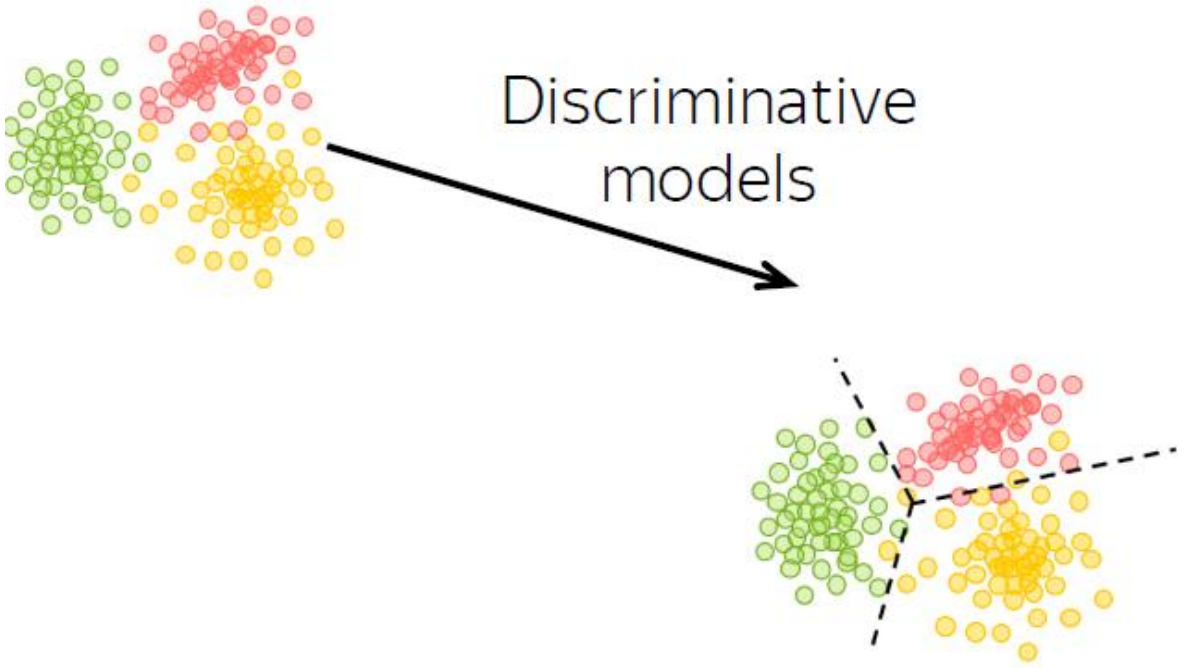
Learn: data distribution

$$p(x, y) = p(x|y) \cdot p(y)$$

How predict:

$$y = \arg\max_k p(x, y) = \arg\max_k p(x|y) \cdot p(y)$$

Дискриминативные модели



Learn: boundary between classes $p(y|x)$

How predict:

$$y = \arg\max_k p(y|x)$$

Классические методы

- Наивный Байес
- Логистическая регрессия
- SVM

Наивный Байес

$$y^* = \arg\max_k P(y = k|x) = \arg\max_k \frac{P(x|y = k) \cdot P(y = k)}{P(x)} =$$

$$= \arg\max_k P(x|y = k) \cdot P(y = k)$$

$$y^* = \arg\max_k P(y = k|x) = \arg\max_k P(x|y = k) \cdot P(y = k) = \arg\max_k P(x, y = k)$$

$$P(y = k) = \frac{N(y = k)}{\sum_{i=1}^{K} N(y = i)}$$

Вычисление $P(x|y=k)$

$$P(x|y=k) = P(x_1, x_2, \dots, x_n|y=k)$$

Наивные Байесовские предположения:

- Предположение о мешке слов (Bag of Words assumption) — порядок слов не имеет значения
- Предположение об условной независимости (Conditional Independence assumption) - признаки (слова) независимы для данного класса.

$$P(x|y=k) = P(x_1, x_2, \dots, x_n|y=k) = \prod_{i=1}^{n} P(x_i|y=k)$$

Вычисление $P(x|y=k)$

$$P(\text{This film is awesome !}|\, y=+) =$$

$$\cdot P(\text{This}|y=+)$$
$$\cdot P(\text{film}|y=+)$$
$$\cdot P(\text{is}|y=+)$$
$$\cdot P(\text{awesome}\,|y=+)$$
$$\cdot P(\text{!}|y=+)$$

Вычисление $\boldsymbol{P(x|y=k)}$

Используя наивные предположения, получили:

$$P(x|y=k) = P(x_1, x_2, \dots, x_n|y=k) = \prod_{i=1}^{n} P(x_i|y=k)$$

Вычисляем отдельные вероятности «вручную»:

$$P(x_i|y=k) = \frac{N(x_i, y=k)}{\sum_{t=1}^{|V|} N(x_t, y=k)}$$

Возможная проблема

Вычисляем отдельные вероятности «вручную»:

$$P(x_i|y=k) = \frac{N(x_i, y=k)}{\sum_{t=1}^{|V|} N(x_t, y=k)}$$

А что если числитель равен нулю? (то есть при обучении токена x_i не было в документах класса k)

$$P(x_i|y=k) = \frac{N(x_i, y=k)}{\sum_{t=1}^{|V|} N(x_t, y=k)} = 0$$

Решение проблемы

Сглаживание Лапласа (Laplace Smoothing).
Решает проблему нулевой вероятности

$$P(x_i|y=k) = \frac{\delta + N(x_i, y=k)}{\sum_{t=1}^{|V|}(\delta + N(x_t, y=k))} = \frac{\delta + N(x_i, y=k)}{\delta \cdot |V| + \sum_{t=1}^{|V|} N(x_t, y=k)}$$

Предсказание модели

Data: $x =$ This film is awesome !

x_1 x_2 x_3 x_4 x_5

$$y^* = \arg\max_k P(x, y = k) = \arg\max_k P(y = k) \cdot P(x|y = k)$$

Получение прогноза

Positive class

$$P(x, y = +) = P(y = +) \cdot P(x|y = +)$$

$$= P(y = +) \cdot P(\text{This}|y = +) \cdot P(\text{film}|y = +) \cdot P(\text{is}|y = +) \cdot P(\text{awesome}|y = +) \cdot P(!|y = +)$$

Negative class

$$P(x, y = -) = P(y = -) \cdot P(x|y = -)$$

$$= P(y = -) \cdot P(\text{This}|y = -) \cdot P(\text{film}|y = -) \cdot P(\text{is}|y = -) \cdot P(\text{awesome}|y = -) \cdot P(!|y = -)$$

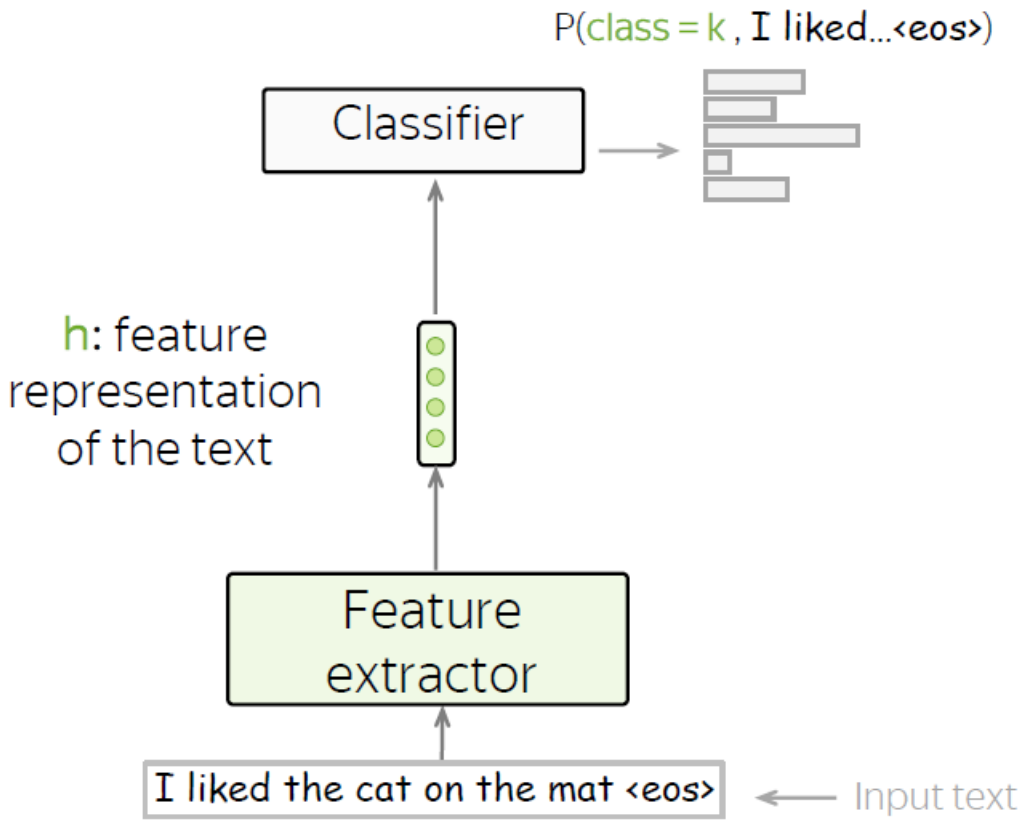
$0,5$

нейтральные слова

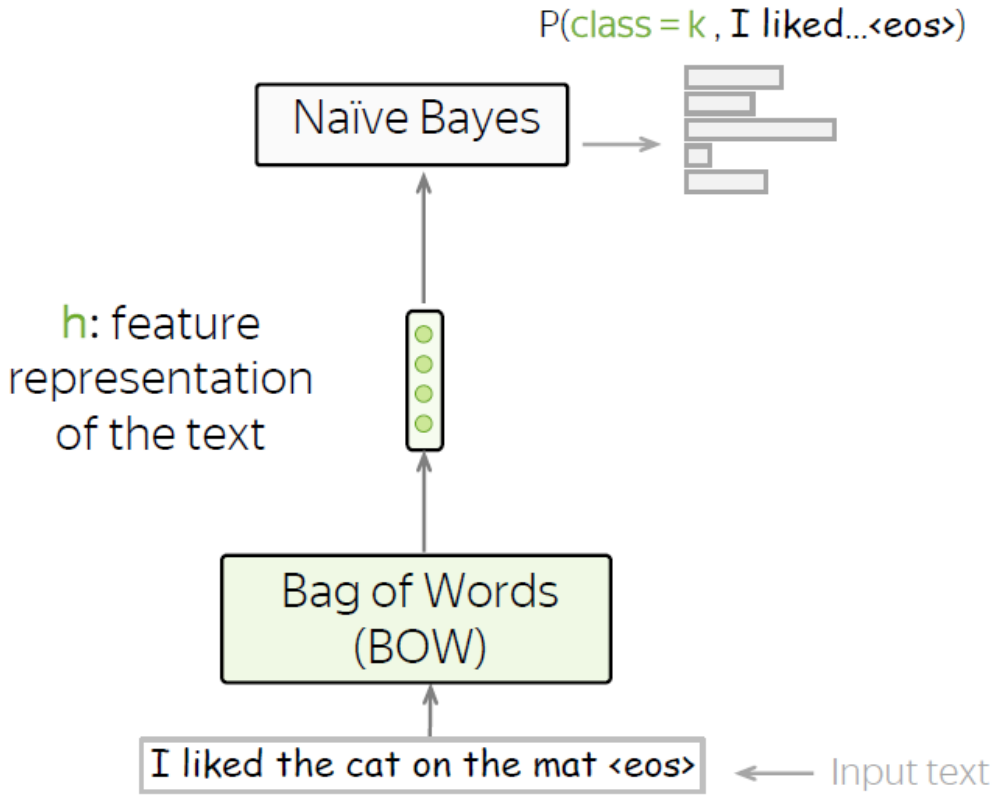
важное слово

$$P(\text{awesome}|y = +) \gg P(\text{awesome}|y = -)$$

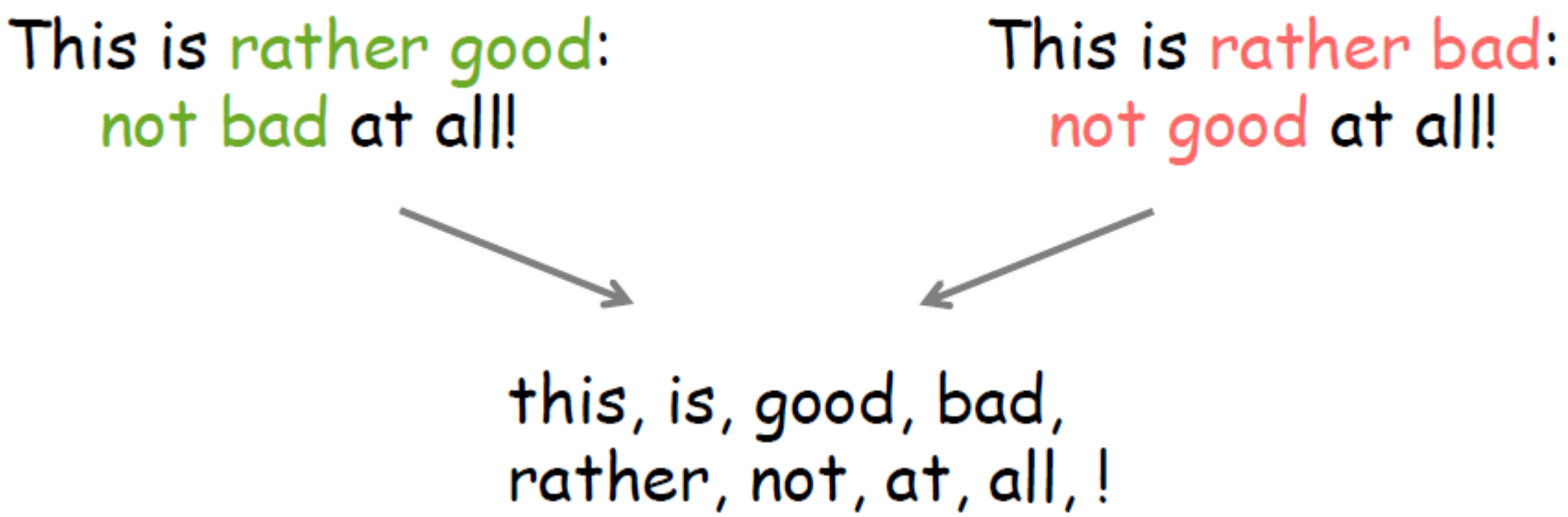
Наивный Байес: взгляд в рамках общей модели



Наивный Байес: взгляд в рамках общей модели



Проблемы наивного Байеса



Логистическая регрессия

$h = (f_1, f_2, \dots, f_n)$ – вектор признаков входного текста

$w^{(k)} = \left(w_1^{(k)}, w_2^{(k)}, \dots, w_n^{(k)}\right)$ – веса признаков

$$w^{(k)} h = w_1^{(k)} \cdot f_1 + \cdots + w_n^{(k)} \cdot f_n, \qquad k = 1, \dots, K$$

$$P(class = k|h) = \frac{\exp(w^{(k)} h)}{\sum_{i=1}^{K} \exp(w^{(i)} h)}$$

Обучение: максимизация вероятности

Есть некоторый текст. Текст имеет метку k.

- Максимизация логарифмического правдоподобия правильного класса $\log P(y = k|x) \to max$
- Минимизация отрицательного логарифмического правдоподобия правильного класса $-\log P(y = k|x) \to min$
- Минимизация потери кросс-энтропии:

$$-\sum_{i=1}^{K} p_i^* \cdot \log P(y = i|x) \to min$$

$$(p_k^* = 1, p_i^* = 0, i \neq k)$$

Сравнение наивного Байеса и логистической регрессии

Наивный Байес:

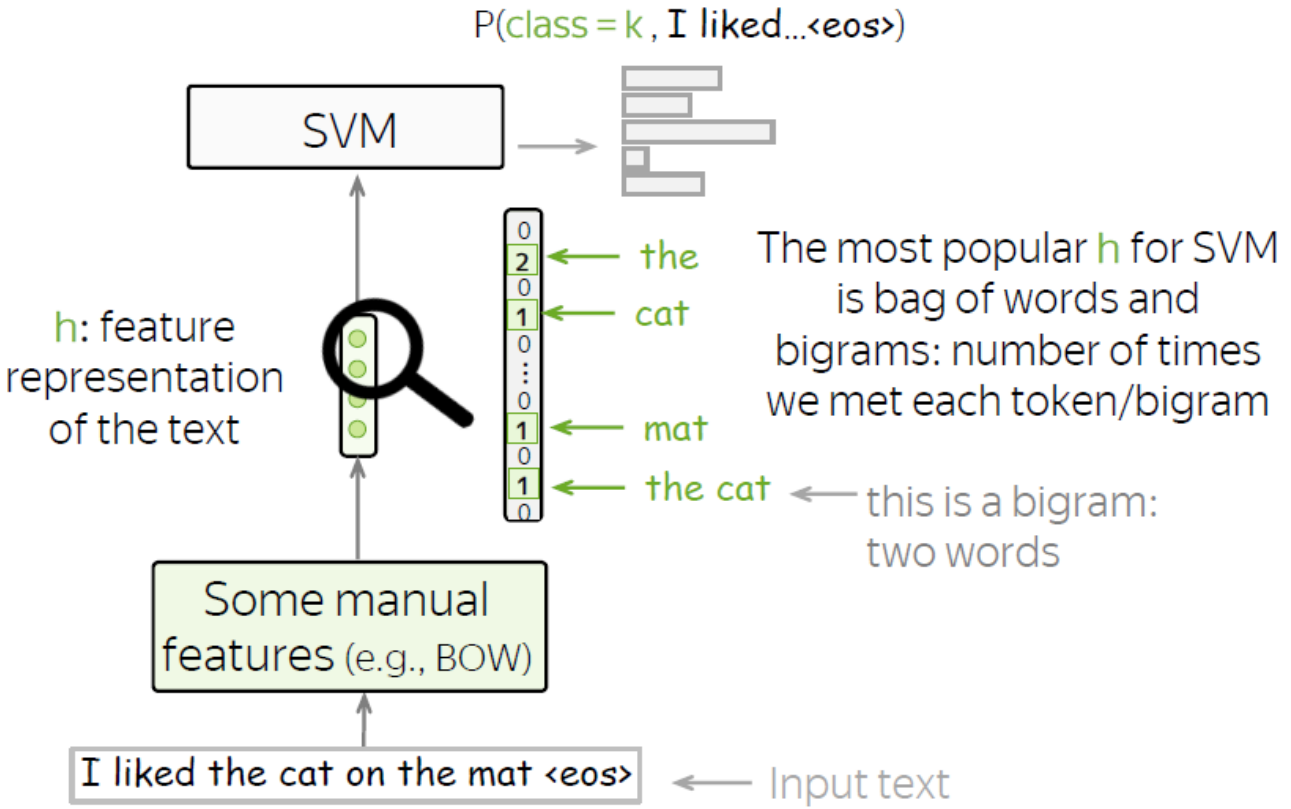
- очень простой
- очень быстрый
- интерпретируемый
- предполагает, что признаки условно независимы
- текстовое представление: определяется вручную (и часто слишком просто, например, BOW)

Сравнение наивного Байеса и логистической регрессии

Логистическая регрессия:

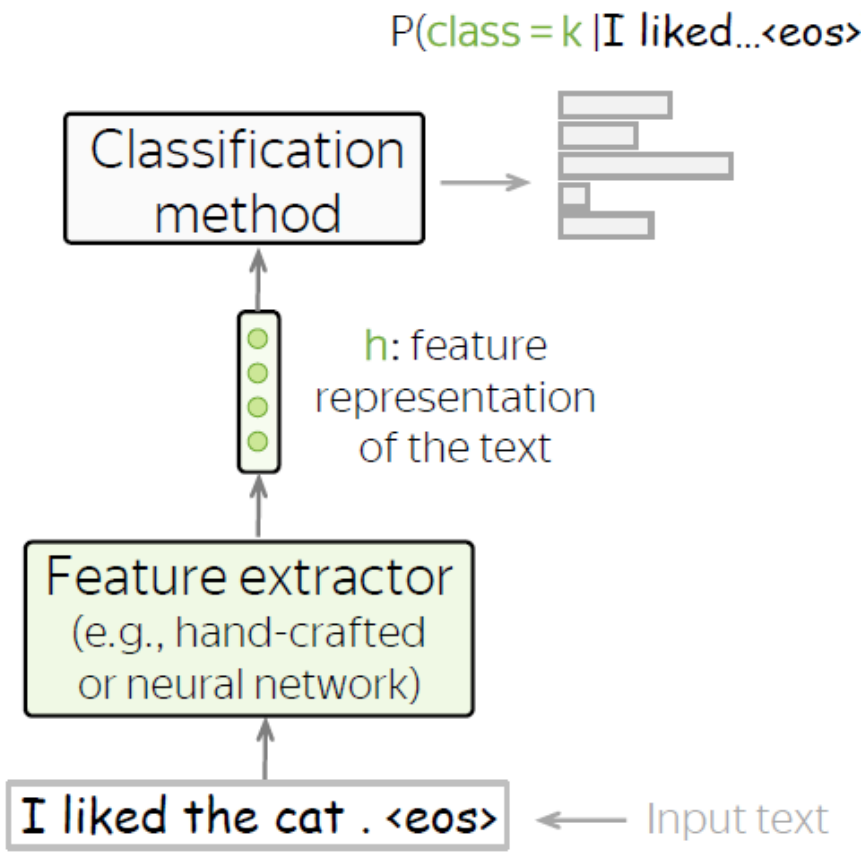
- очень простая
- интерпретируемая
- не предполагает, что признаки условно независимы
- не такая быстрая (множественные итерации градиентного спуска)
- текстовое представление: определяется вручную

SVM


P(class = k , I liked...<eos>)
SVM
h: feature representation of the text
0
2
0
1
0
0
1
0
1
0
the
cat
mat
the cat
this is a bigram: two words
The most popular h for SVM is bag of words and bigrams: number of times we met each token/bigram
Some manual features (e.g., BOW)
I liked the cat on the mat <eos>
Input text

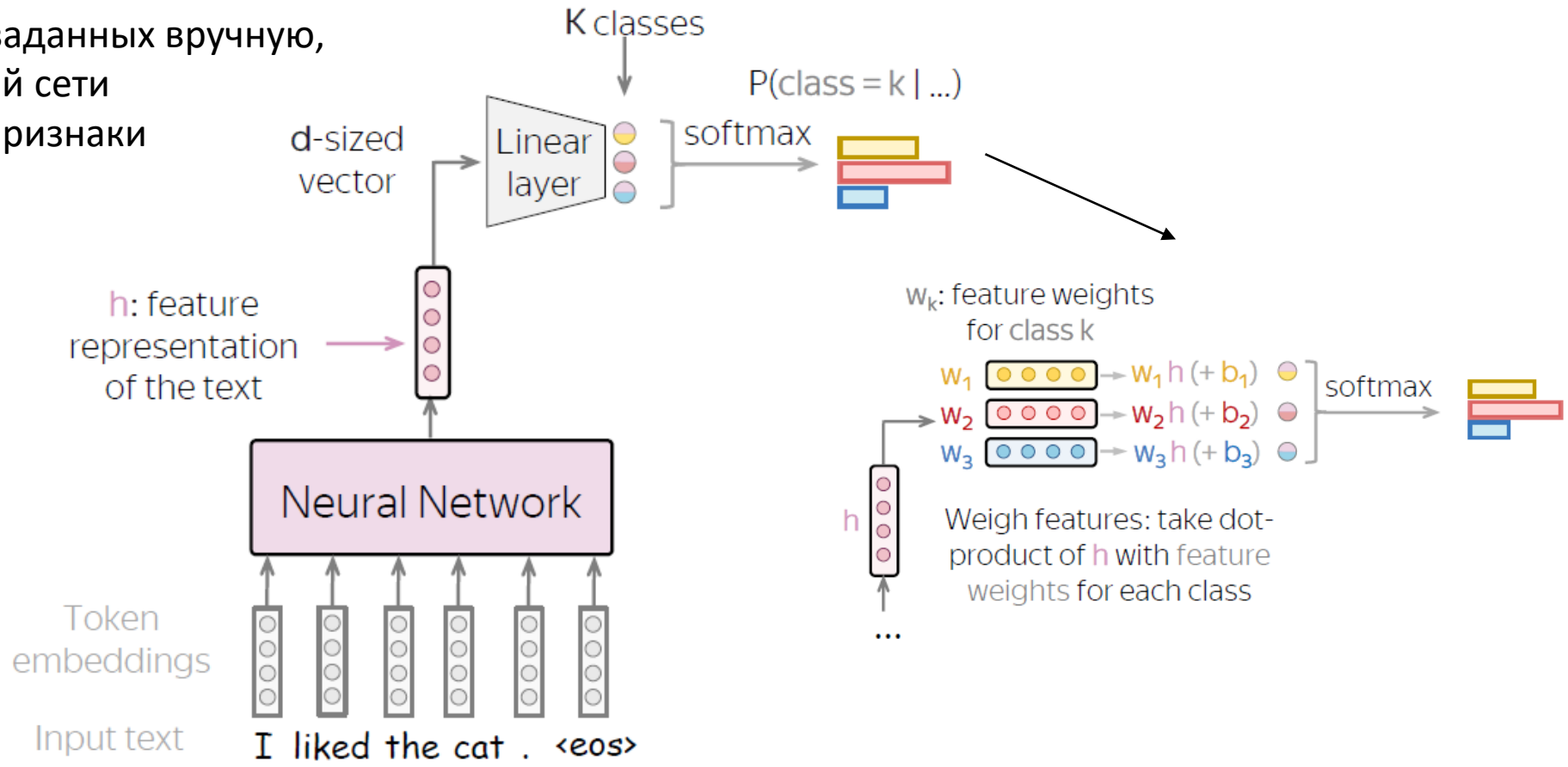
Классификация с помощью классических методов

Общая схема:

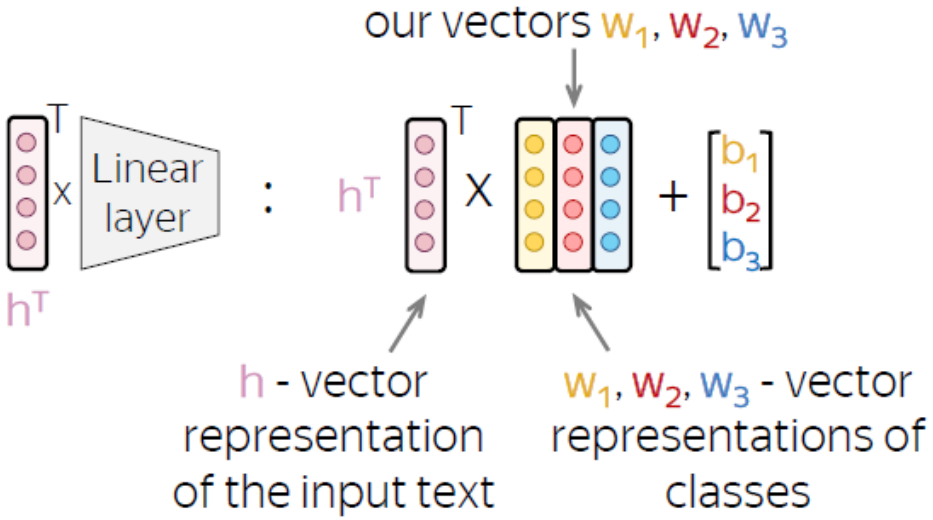


Классификация с помощью нейронных сетей

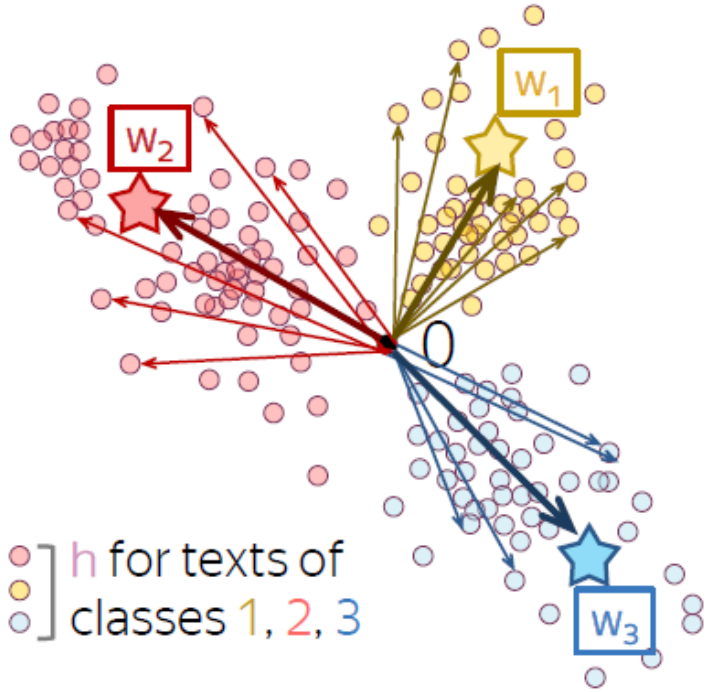
Вместо признаков, заданных вручную, позволим нейронной сети выявить полезные признаки



Представление текста и представление классов



Векторы текста указывают в направлении соответствующих векторов класса:

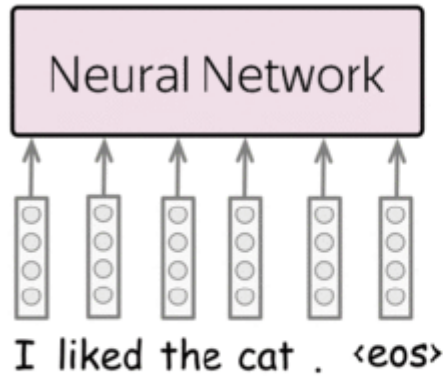


Обучение: кросс-энтропия

Есть некоторый текст. Текст имеет метку k.

Model prediction: P(class = i |I liked...<eos>)

Target: p^*



Feed a text to the network

Cross-entropy loss:

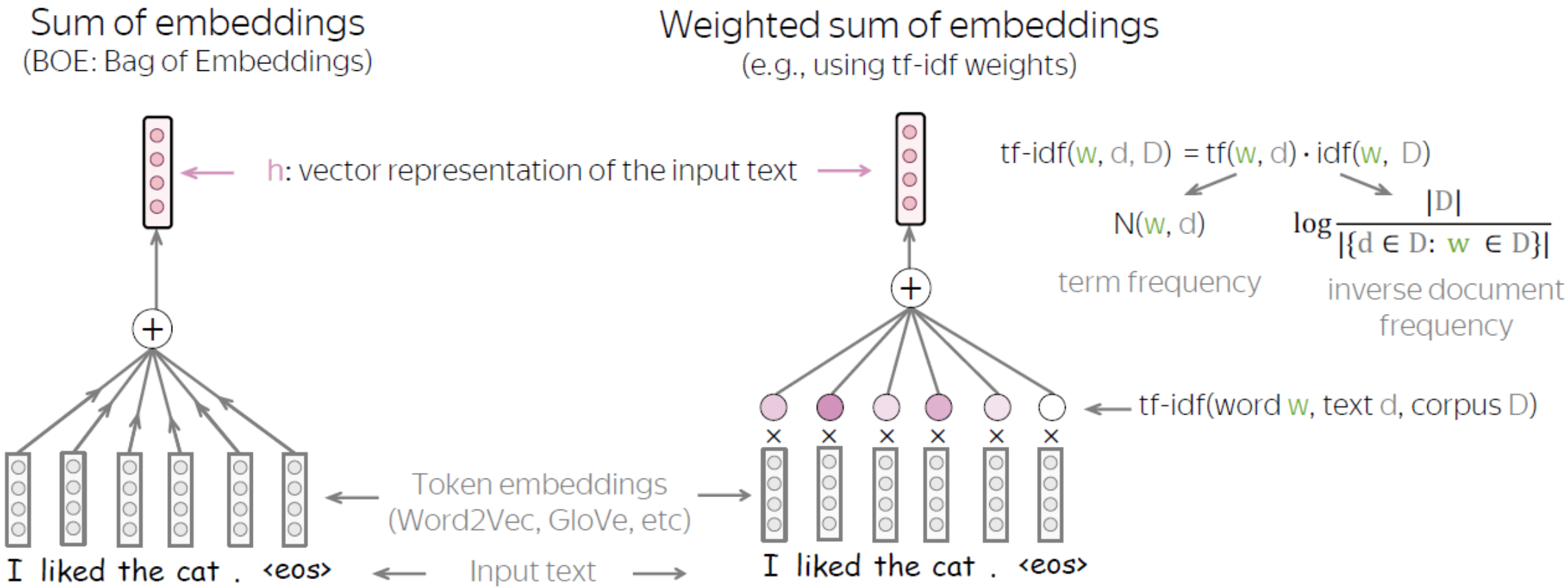
$$-\sum_{i=1}^{K} p_i^* \cdot \log P(y=i|x) \to min \quad (p_k^* = 1, p_i^* = 0, i \neq k)$$

Correct label: 4 ← we want the model to predict this

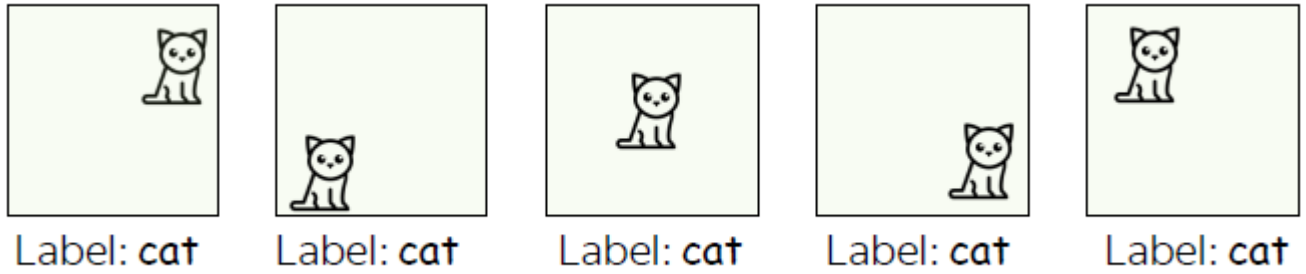
For one-hot targets, this is equivalent to

$$-\log P(y=k|x) \to min$$

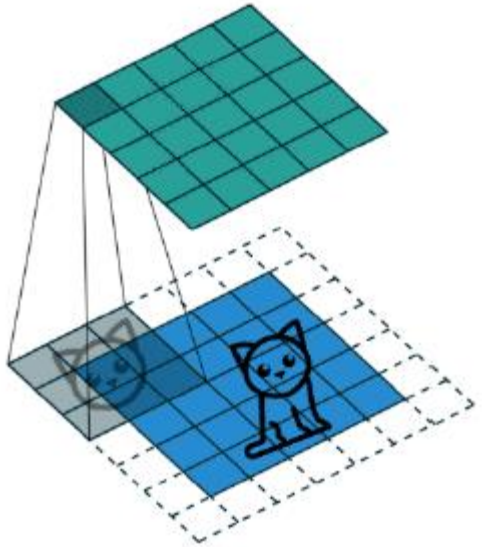
Простейшие модели: BOE и взвешенный BOE



Свёрточные нейросети для обработки изображений и Translation Invariance



- Применить ту же операцию к небольшим частям входных данных
- найти «совпадения» с шаблонами
- сеть узнает, какие шаблоны полезны
- шаблоны развиваются от простого к сложному



Как это использовать в текстах?

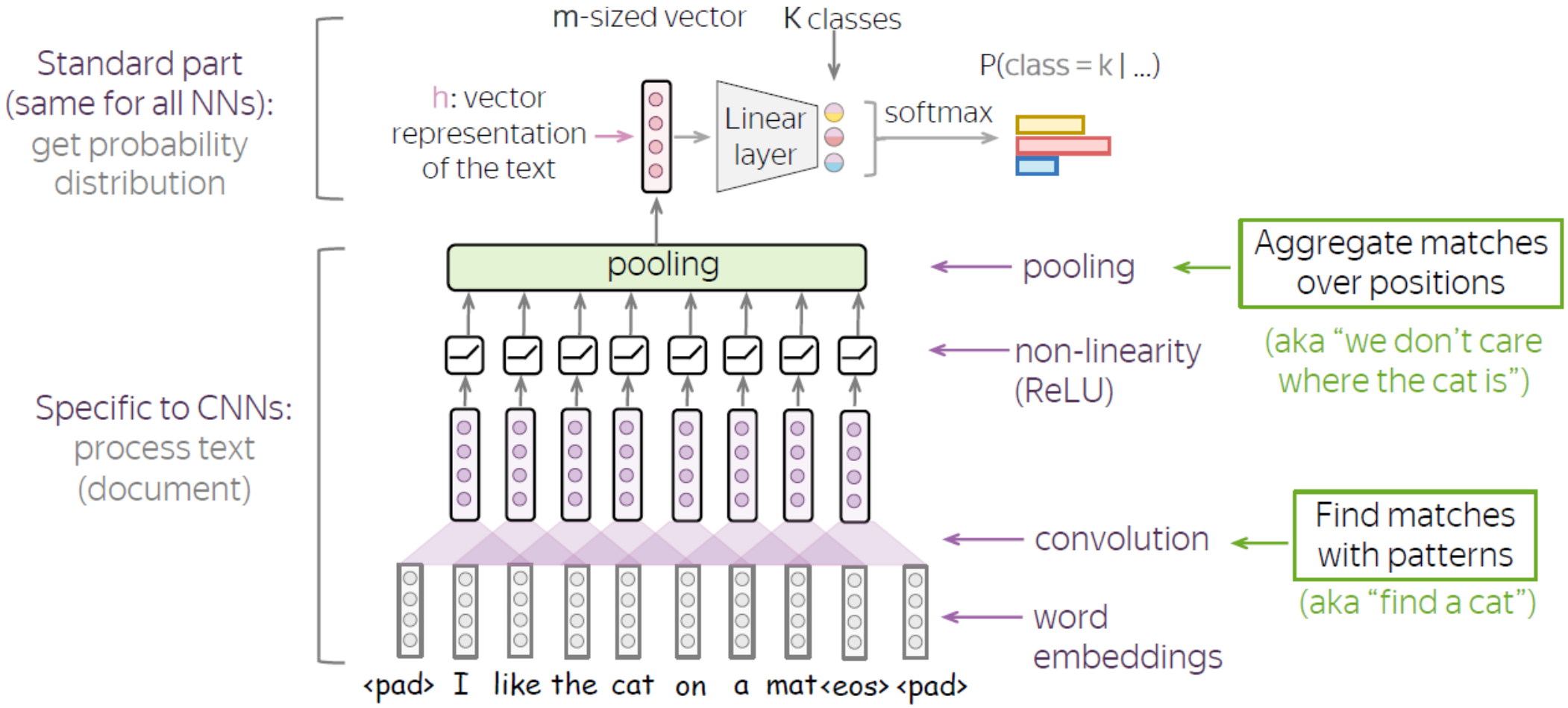
Если в тексте есть некоторая фраза, и она очень важная, то, возможно, нам не очень важно, где именно она находится.

An absolutely great movie! I watched the premiere with my friends.

The movie about cats was absolutely great, and the cats were cute.

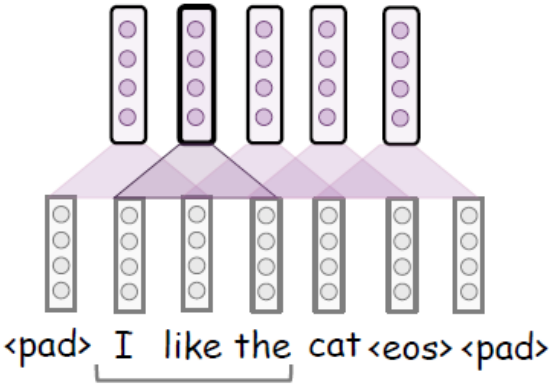
The movie is about cats running around, and it is absolutely great.

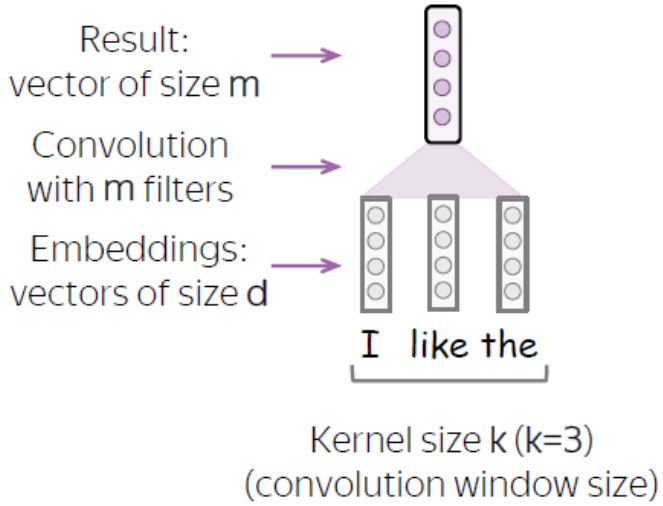
Типичная модель: свертка + объединение

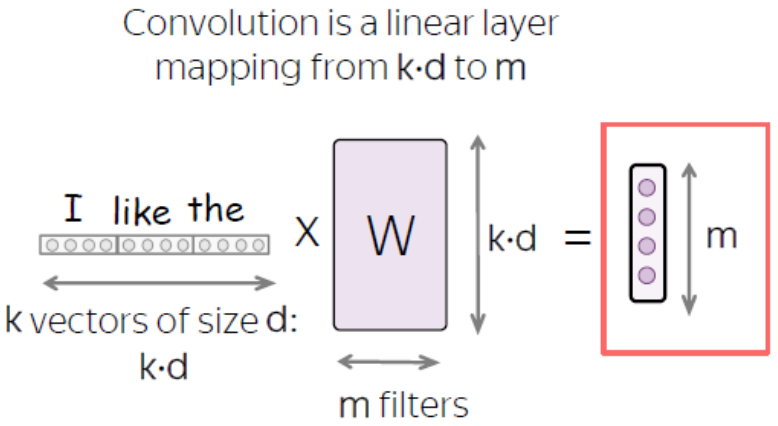


Фильтр свертки для текста

Свертка — это линейная операция, применяемая к каждому окну







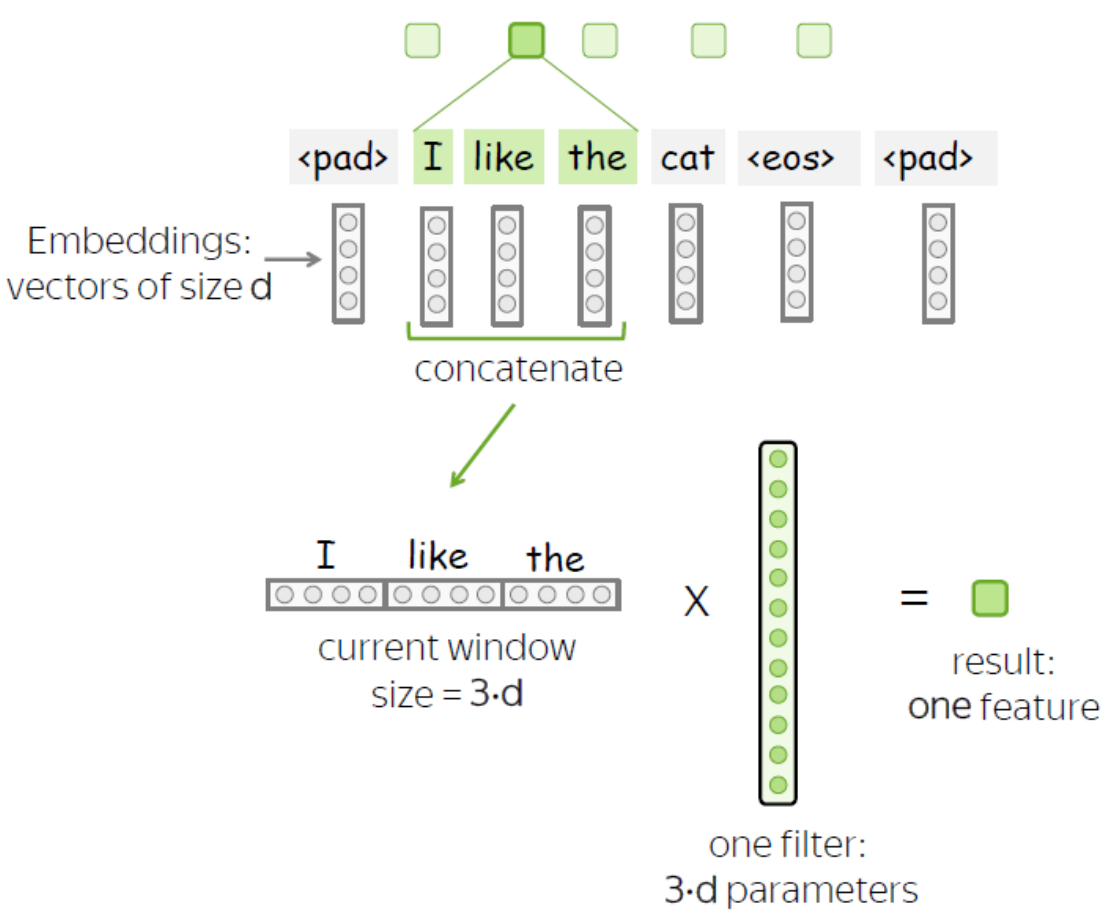
- $x_1, x_2, \ldots, x_n$ - input vectors (e.g., word embeddings)
- d (input channels) – input vector size
- k (kernel size) – conv. window length
- m (output channels) – number of filters

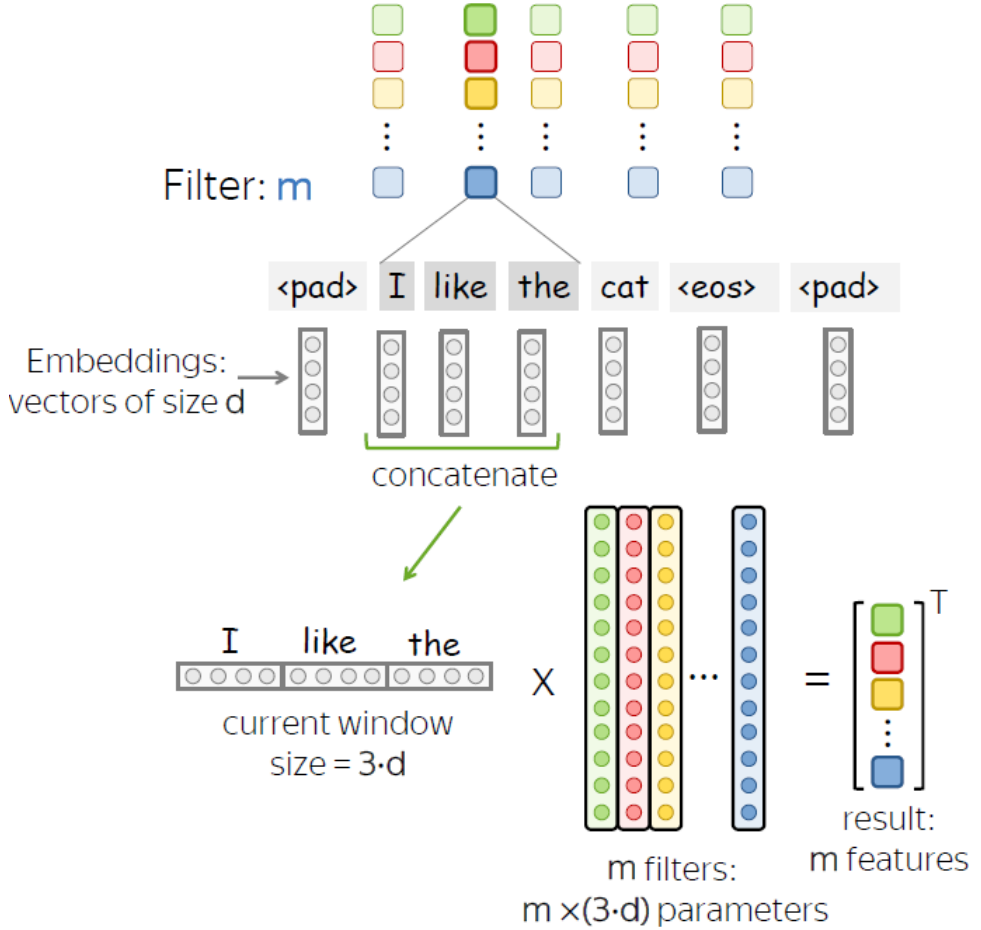
$u_i = [x_i, \ldots, x_{i+k-1}] \in \mathbb{R}^{k \cdot d}$ - concatenate representations in the i-th window

$W \in \mathbb{R}^{(k \cdot d) \times m}$ - convolution

$F_i = u_i \times W$ – convolution applied to the i-th window

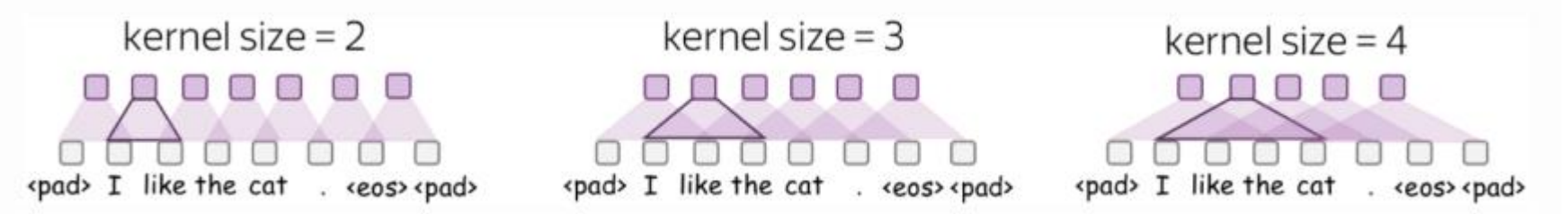
Фильтр


<pad> I like the cat <eos> <pad>
Embeddings:
vectors of size d
concatenate
I like the
current window
size = 3·d
X
one filter:
3·d parameters
=
result:
one feature

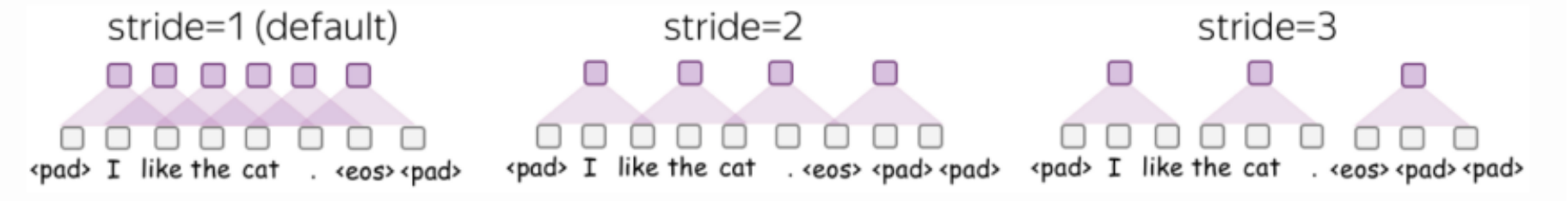

Filter: m
<pad> I like the cat <eos> <pad>
Embeddings:
vectors of size d
concatenate
I like the
current window
size = 3·d
X
m filters:
m ×(3·d) parameters
=
T
result:
m features

Свёртка: параметры

Размер ядра: сколько токенов за один раз вы смотрите



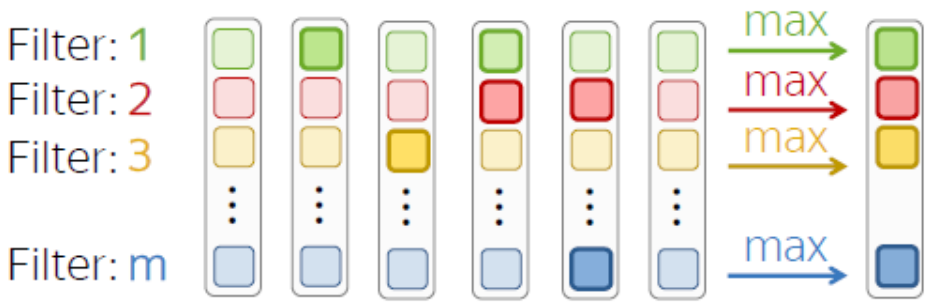
Шаг: насколько нужно переместить фильтр на каждом шаге



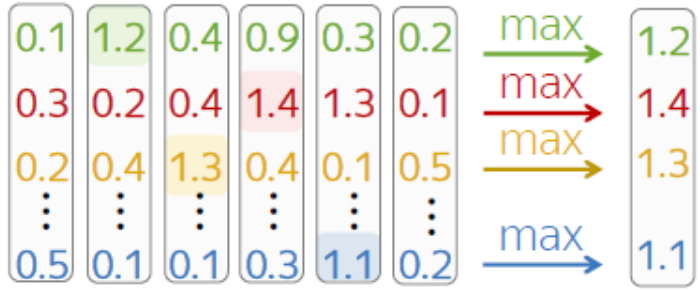
Pooling (max, mean, etc) (объединение)

Max pooling: максимальное значение для каждого измерения (признака)

Mean pooling: среднее значение для каждого измерения (признака)



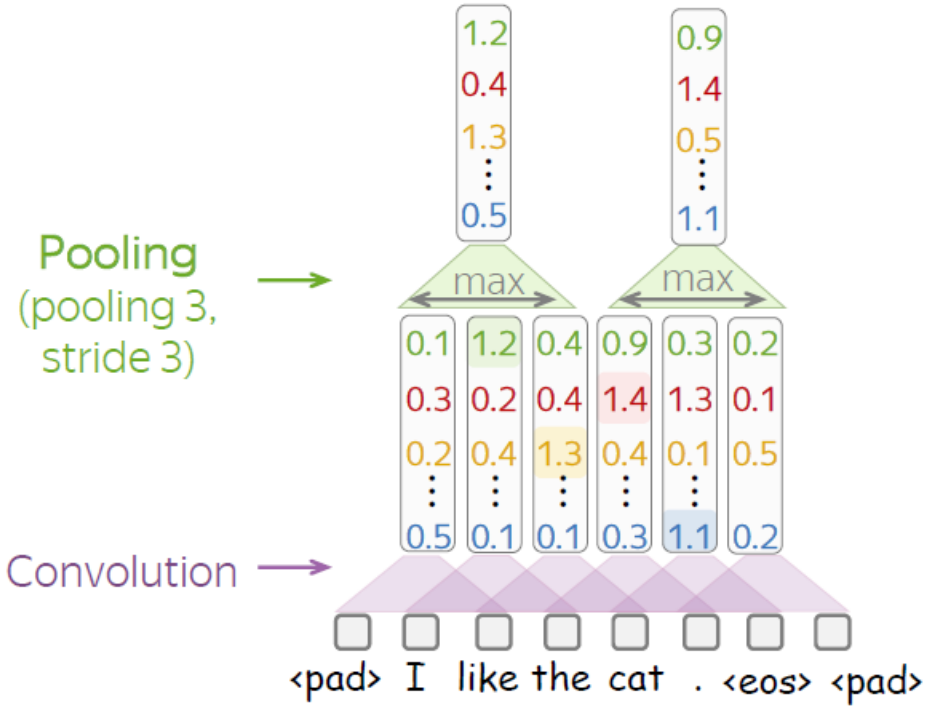
Max pooling:
maximum for each
dimension (feature)

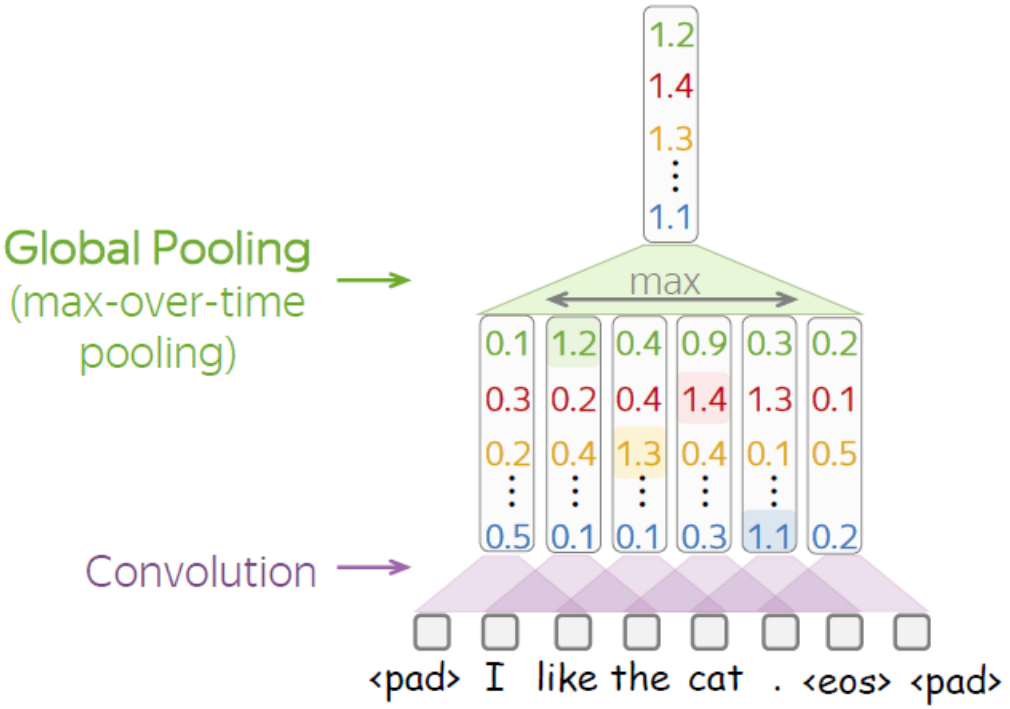


Pooling and Global Pooling

Pooling: агрегация признаков в некоторой области

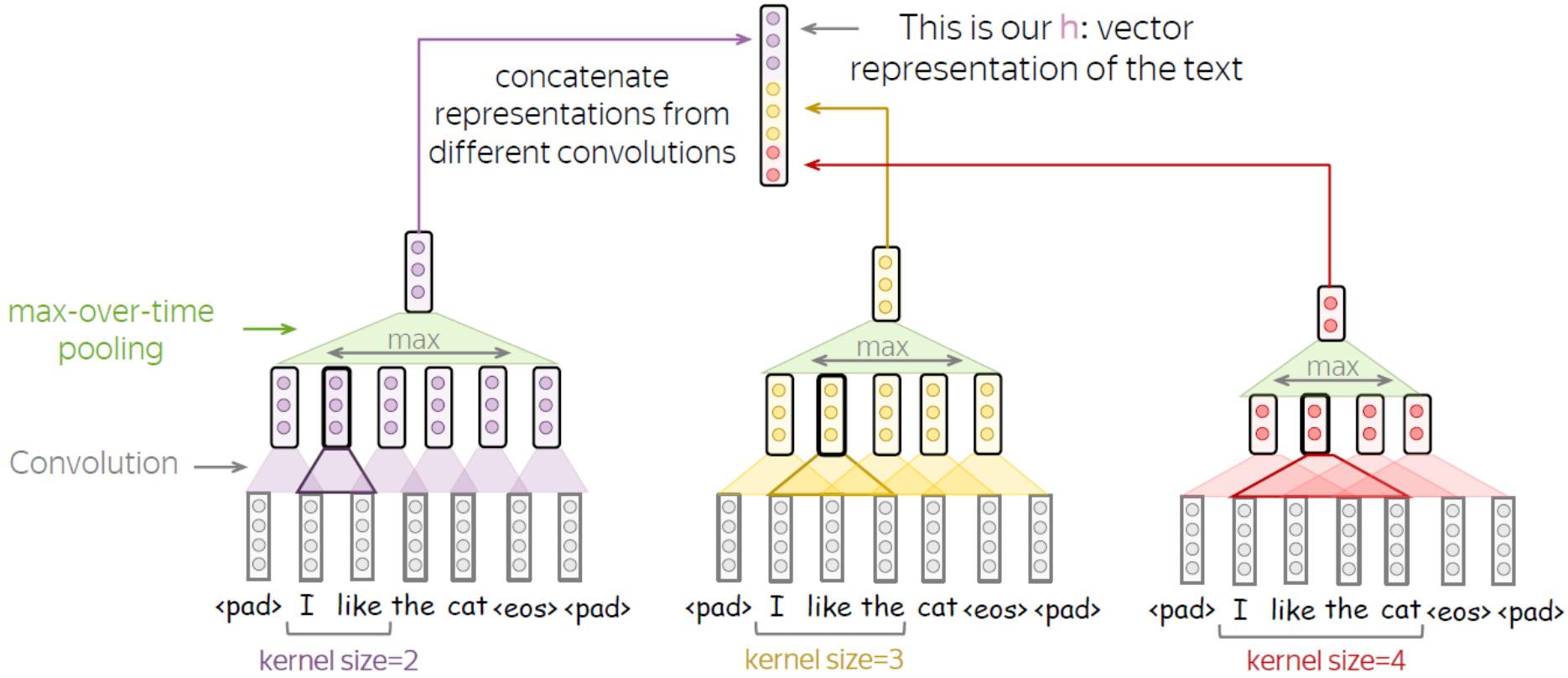
Global pooling: агрегация признаков по всем входным данным





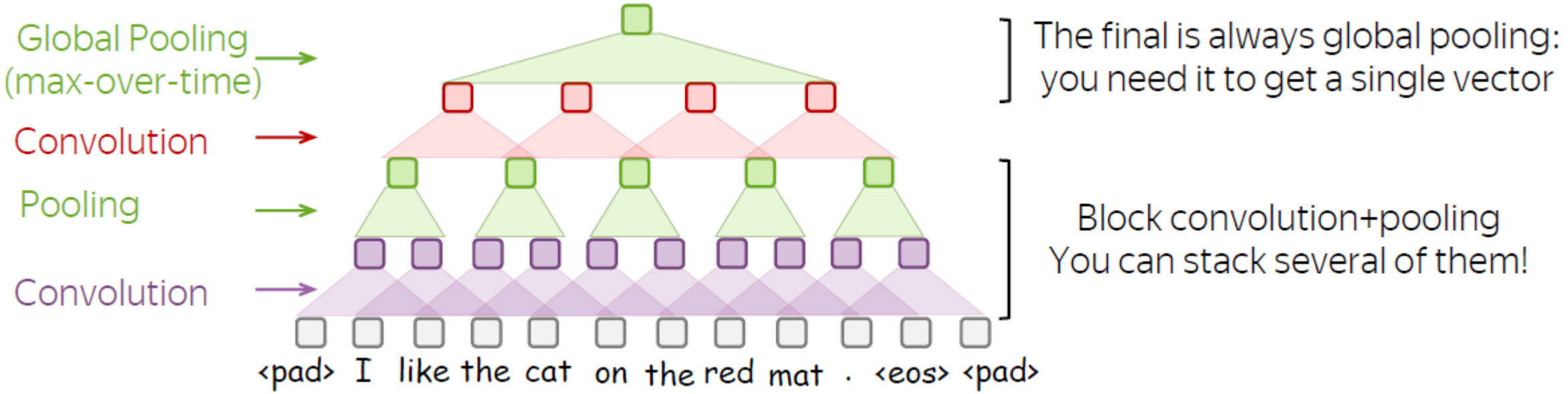
Сверточные модели для классификации текстов

Несколько сверток с разными размерами ядра

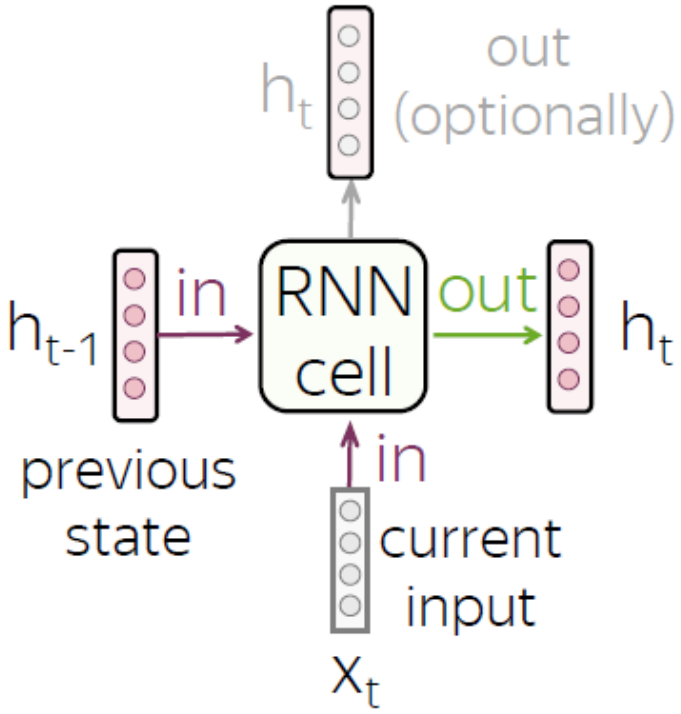


Сверточные модели для классификации текстов

Стек из нескольких блоков: свертка+пулинг



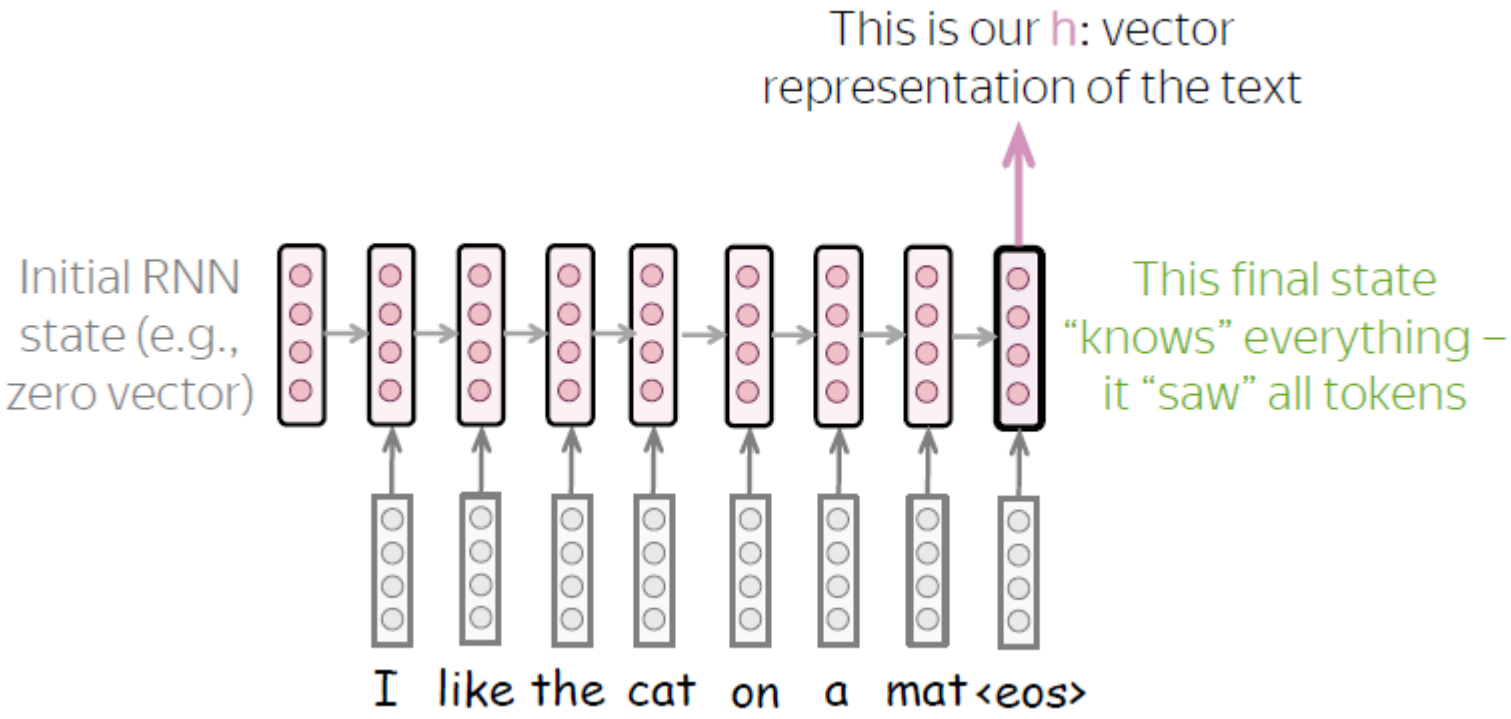
Рекуррентные нейронные сети



На каждом этапе рекуррентная нейронная сеть получает новый входной вектор (например, эмбеддинг) и предыдущее состояние сети (которое, будем надеяться, содержит всю предыдущую информацию). Используя эти входные данные, ячейка рекуррентной нейронной сети вычисляет новое состояние, которое она выдаёт в качестве результата. Это новое состояние теперь содержит информацию как о текущих входных данных, так и об информации, полученной на предыдущих этапах.

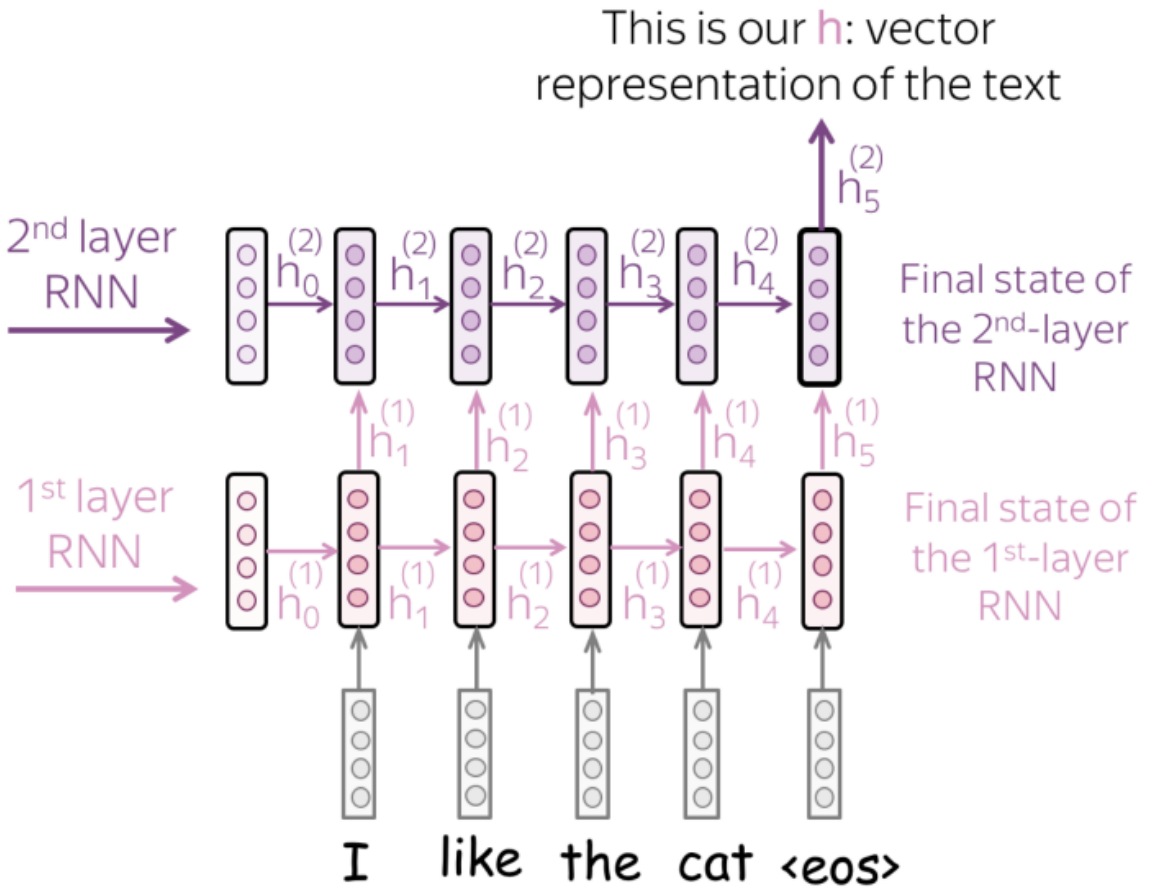
Рекуррентные модели для классификации текстов

просто: прочитайте текст,
определите конечное состояние



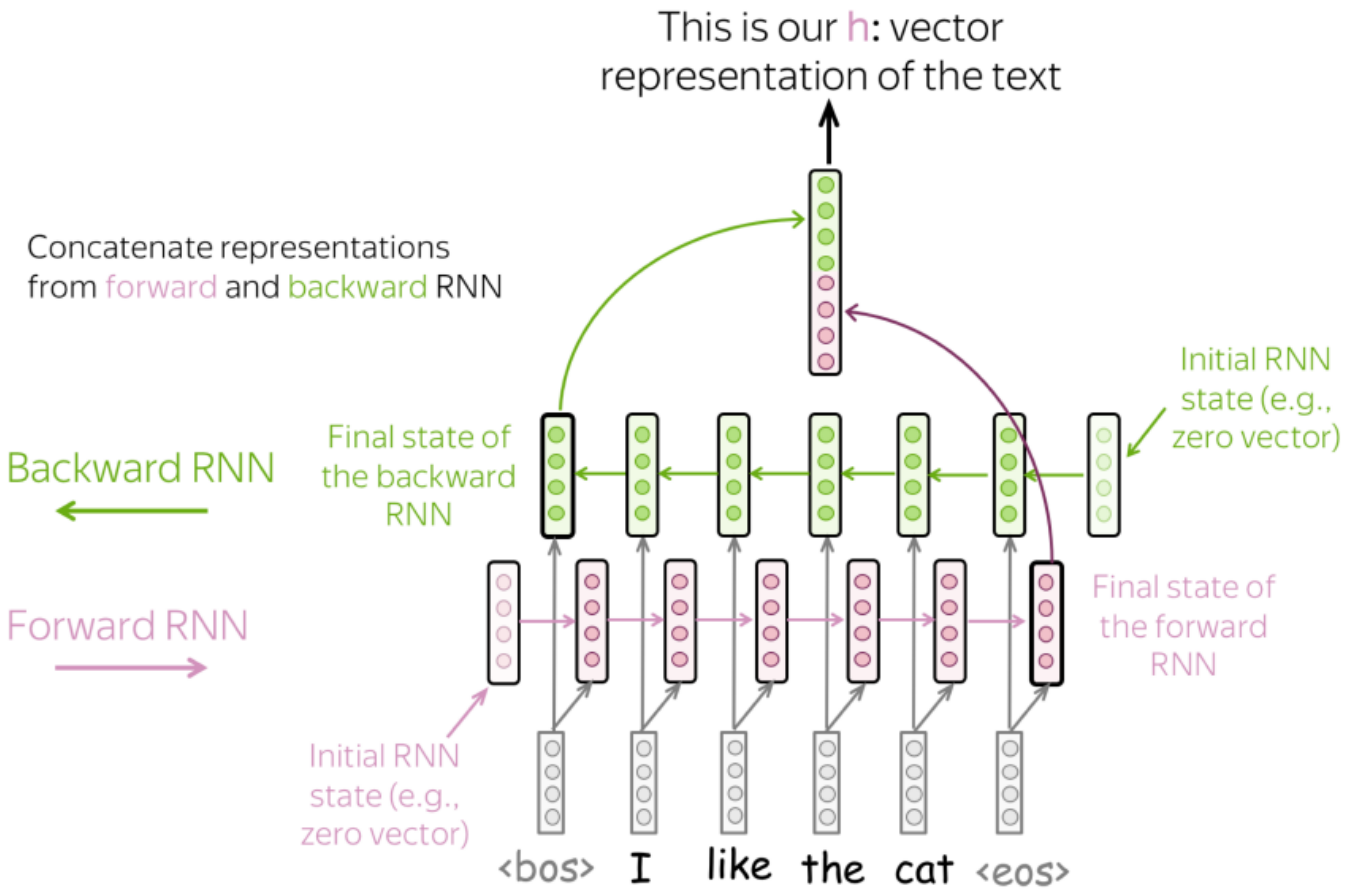
Рекуррентные модели для классификации текстов

Несколько слоев: передача состояний из одной RNN в другую



Рекуррентные модели для классификации текстов

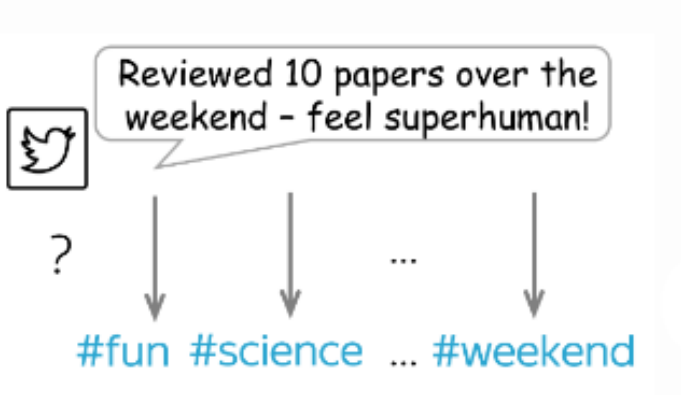
Два направления: использовать конечные состояния из прямых и обратных RNN



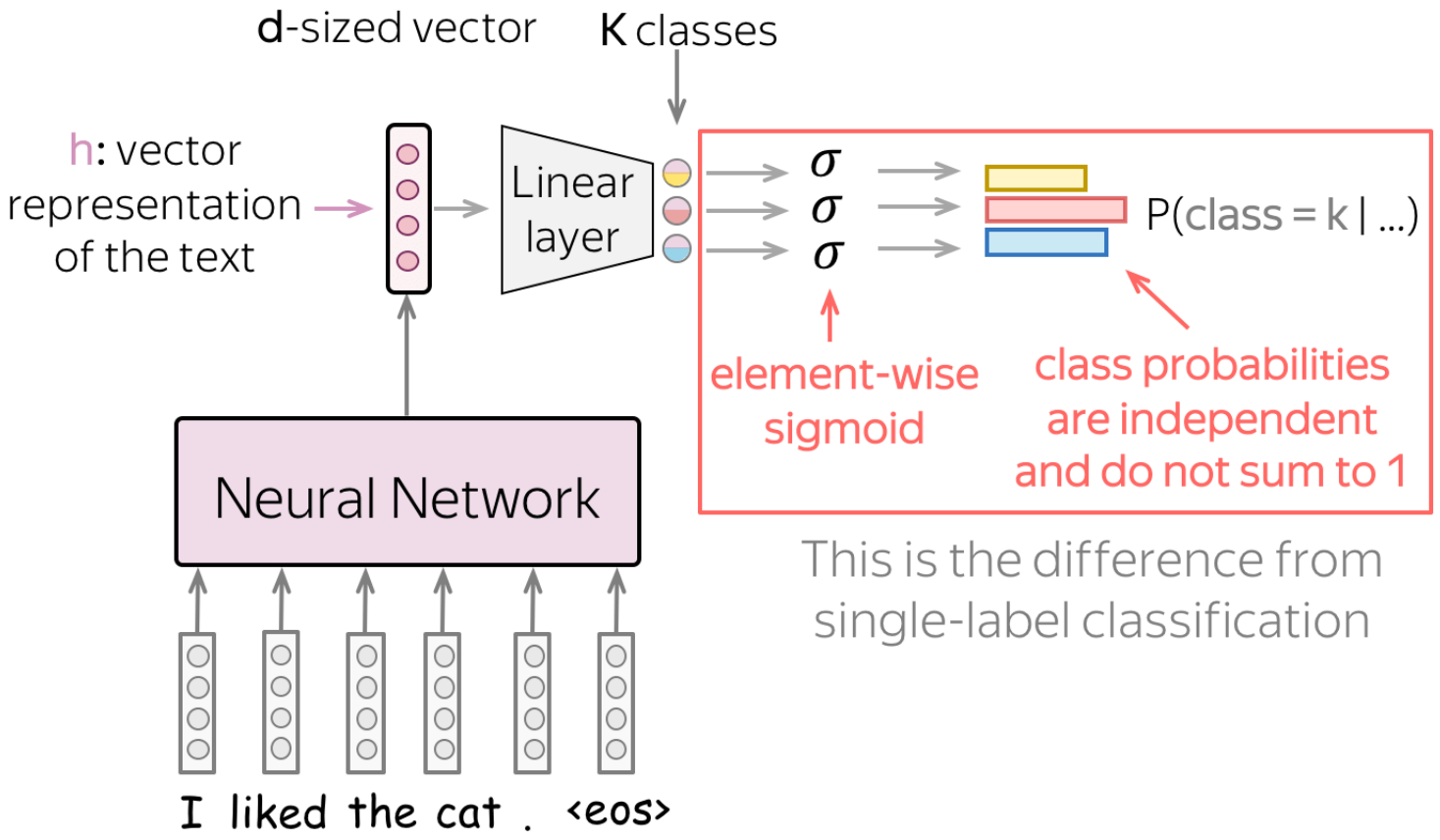
Многометочная классификация

Для задачи с несколькими метками нам нужно изменить два элемента в схеме с одной меткой:

- модель (способ оценки вероятностей классов);
- функцию потерь



Многометочная классификация



Функция потерь: бинарная кросс-энтропия для каждого класса

Training example: I liked the cat on the mat <eos>

Labels: k, t

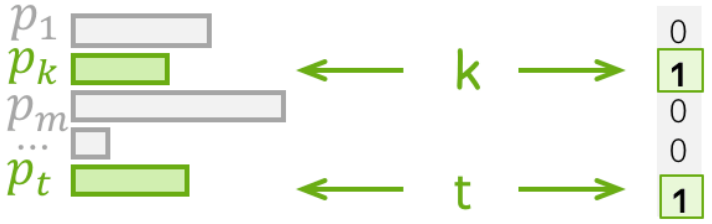
target

Model prediction:

$P(y_i = 1 \mid \ldots\text{<eos>}), i = 1..K$

Target:

$p_i^*, i = 1..K$



Binary cross-entropy loss for each class:

$$-\sum_{i=1}^{K}[p_i^* \cdot \log P(y_i = 1|x) + (1 - p_i^*) \cdot \log\ P(y_i = 0|x)] \rightarrow min,$$

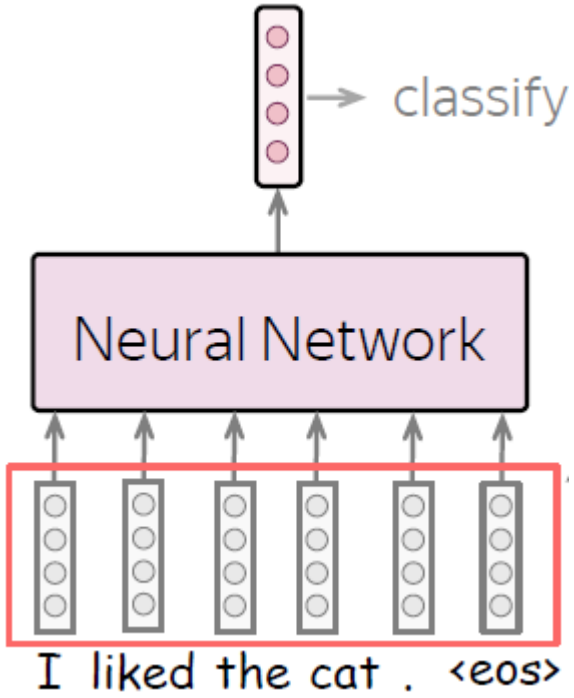
where $P(y_i = 0|x) = 1 - P(y_i = 1|x)$

binary classifier loss for class i

Практические советы. Что делать с эмбеддингами?

Входные векторные представления слов:

- Обучение с нуля
- Использование предварительно обученных данных (Word2Vec, GloVe)
- Инициализация с использованием предварительно обученных данных, затем тонкая настройка



Какие данные у нас есть?

- обучающие данные для классификации текста (с меткой): небольшой объём; тема, специфичная для конкретной задачи
- обучающие данные для векторных представлений слов (без меток): огромные и разнообразные данные; нет специфики тем

Что делать с эмбеддингами?

- обучение с нуля (данных может быть недостаточно, чтобы изучить взаимосвязи между словами)
- предобученные данные (Word2Vec, GloVe) (знают взаимосвязи между словами, но не знают специфику задачи)
- инициализация с помощью предобученных данных, затем тонкая настройка (знают взаимосвязи между словами и адаптируются к задаче)

Получение дополнительных данных (Data Augmentation)

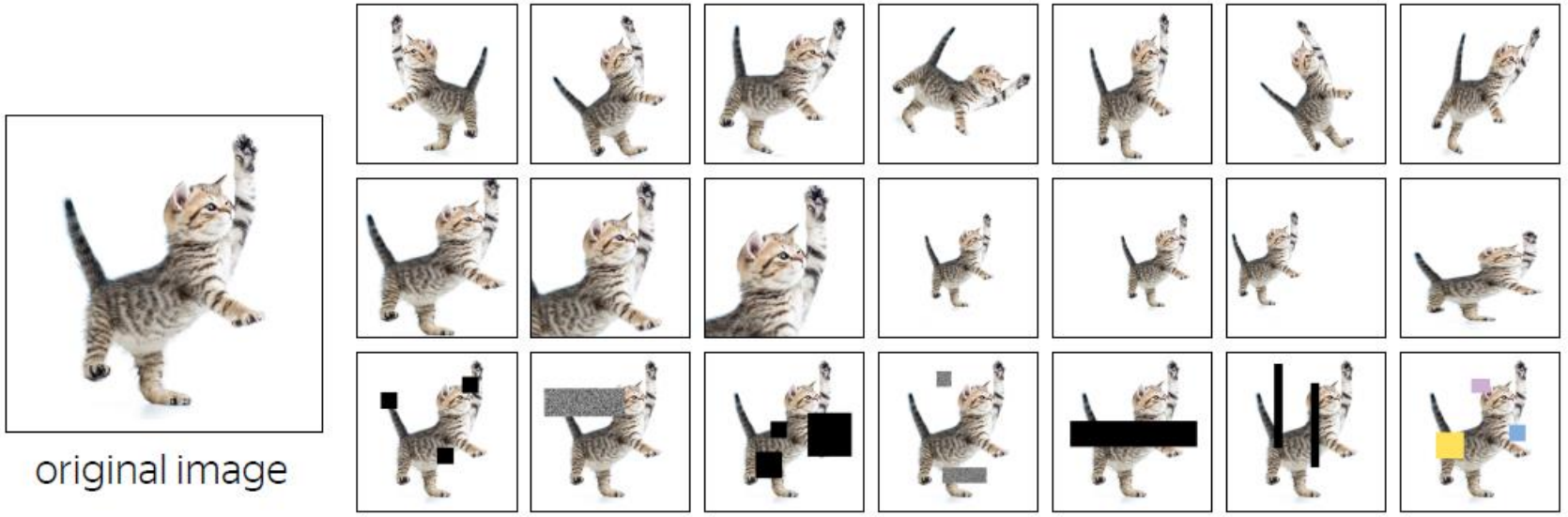
Аугментация данных изменяет ваш набор данных различными способами, чтобы получить альтернативные версии одного и того же обучающего примера.

Она может увеличить:

- объем данных
- разнообразие данных

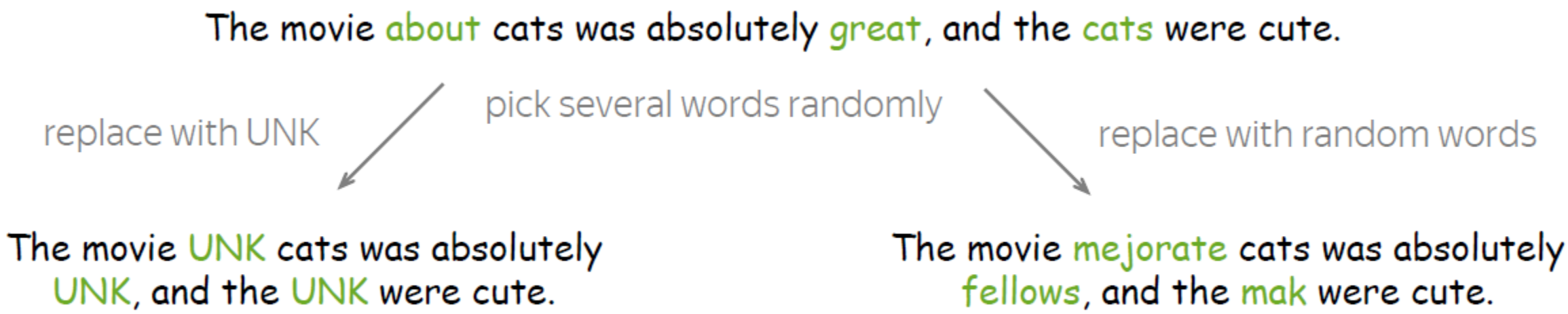
Data Augmentation для изображений

- переворот изображения
- геометрические преобразования (вращение, растяжение, увеличение/уменьшение масштаба)
- скрытие отдельных участков



Data Augmentation для текста

- word dropout (исключение слов)



Data Augmentation для текста

- использование внешних ресурсов (например, тезауруса)

The movie about cats was absolutely great, and the cats were cute.

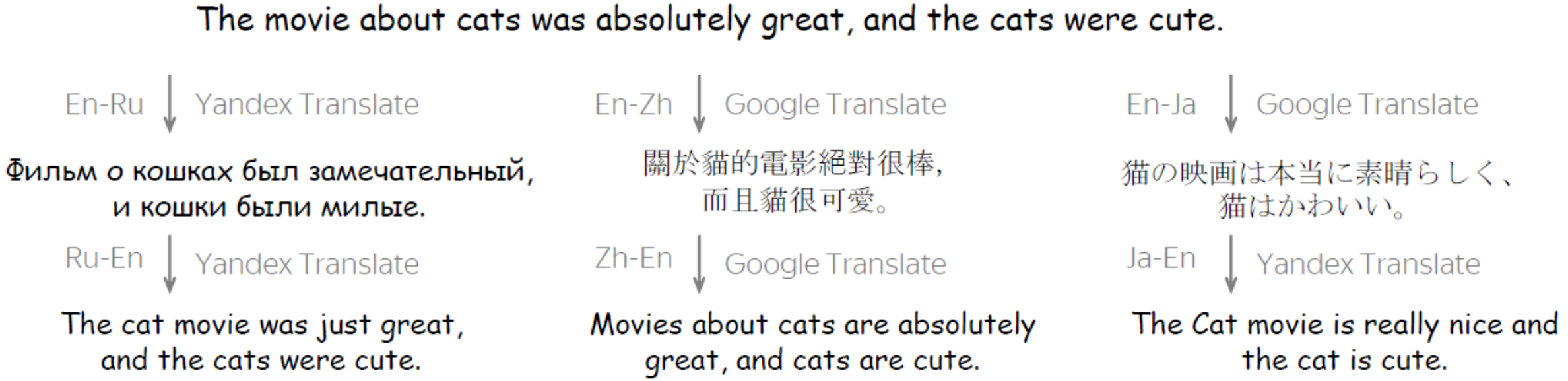
pick words where you have synonyms and use thesaurus

The film about cats was certainly great, and the cats were nice.

The video about cats was completely great, and the cats were charming.

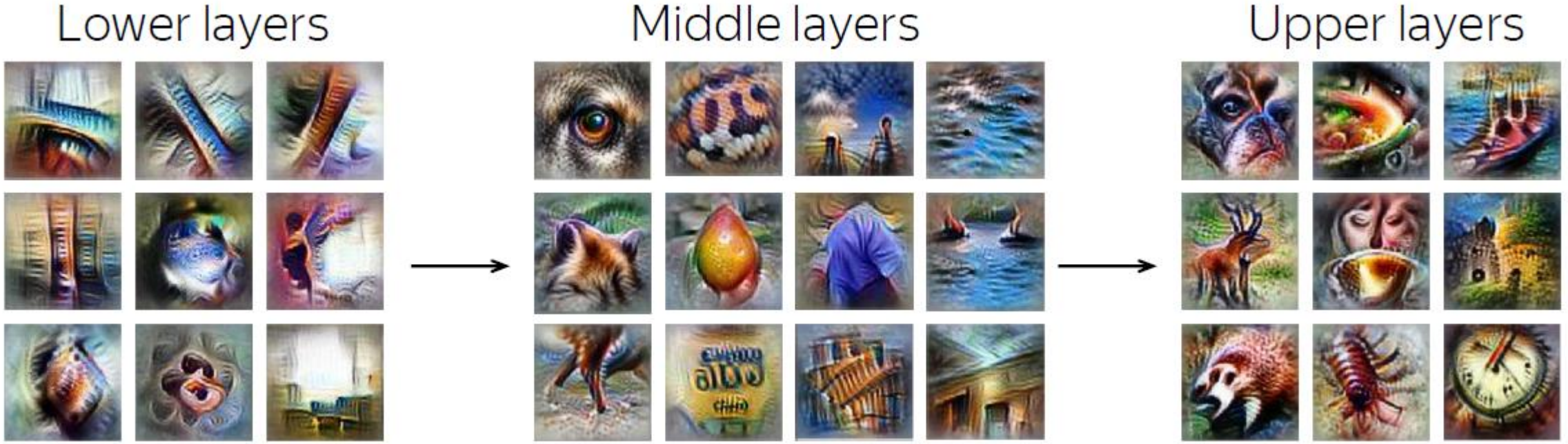
Data Augmentation для текста

- использование отдельных моделей для перефразирования



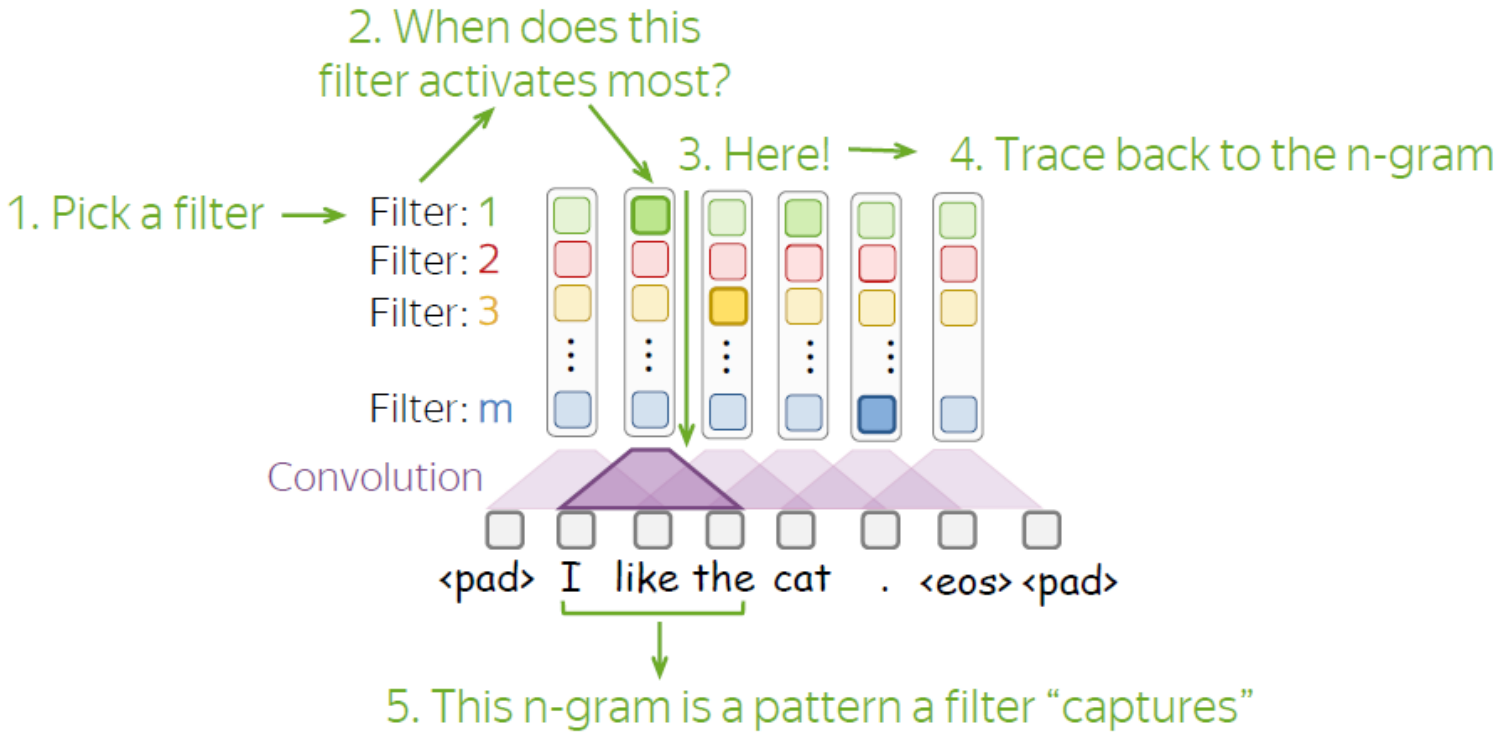
Анализ сверточных фильтров

Примеры паттернов, полученных с помощью фильтров свёртки изображений:



Анализ сверточных фильтров

Примеры паттернов, полученных с помощью фильтров свёртки изображений:



Анализ сверточных фильтров

filter	Top n-gram	Score
1	poorly designed junk	7.31
2	simply would not	5.75
3	a minor drawback	6.11
4	still working perfect	6.42
5	absolutely gorgeous .	5.36
6	one little hitch	5.72
7	utterly useless .	6.33
8	deserves four stars	5.56
9	a mediocre product	6.91

Top n-grams for filter 4		Score
1	still working perfect	6.42
2	works - perfect	5.78
3	isolation proves invaluable	5.61
4	still near perfect	5.6
5	still working great	5.45
6	works as good	5.44
7	still holding strong	5.37

A filter activates for a family of n-grams with similar meaning