Лекция 1.
Word embeddings

О курсе «Искусственный интеллект»

- Yandex School of Data Analysis
- NLP Course For You

Word embeddings

- Зачем нужно векторное представление слов?
- Работа с неизвестными токенами
- Векторы с одним активным компонентом (One-Hot Vectors)
- Методы, основанные на подсчёте (Count-Based Methods)
- Word2Vec (Word to Vector)
- GloVe
- Внутренняя и внешняя оценки качества
- О Лабораторной работе 1 «Базовое использование LLM»

Зачем нужно векторное представление слов?

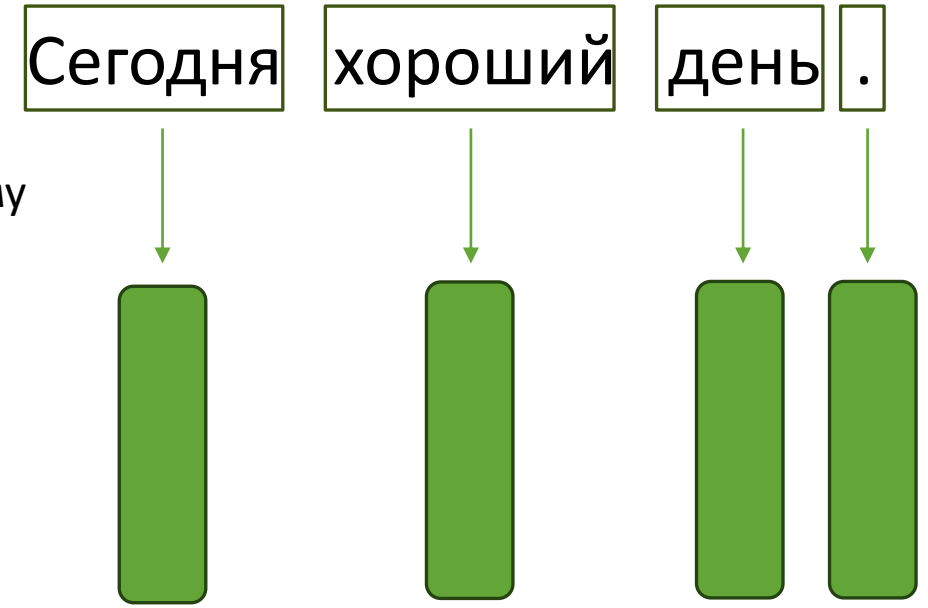
Векторное представление слов (векторное вложение слов, эмбеддинг) — это способ представления слов в виде числовых векторов, который используется компьютерными моделями для обработки естественного языка (NLP - Natural Language Processing).

Чтобы получить возможность представить семантическую близость, было предложено использовать embedding, то есть сопоставить слову некий вектор, отображающий его значение в «пространстве смыслов».

Зачем нужно векторное представление слов?

Предложение: Сегодня хороший день.

Разбиваем предложение на последовательность токенов:

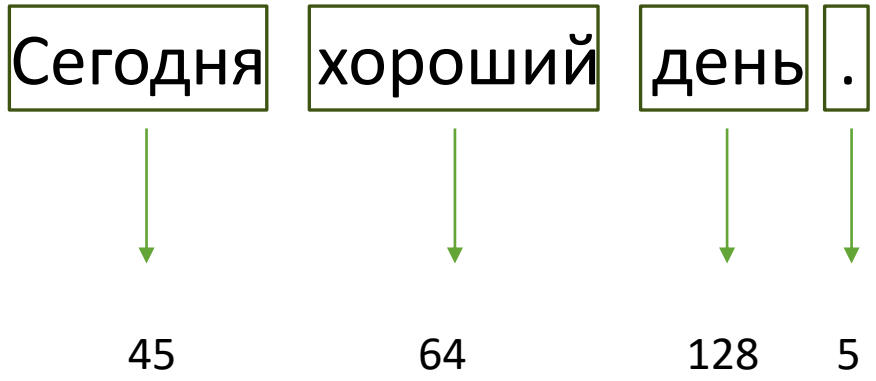


Ставим в соответствие каждому слову вектор

Зачем нужно векторное представление слов?

Каждый токен имеет свой номер в словаре токенов:

Каждый токен имеет свой номер в словаре токенов

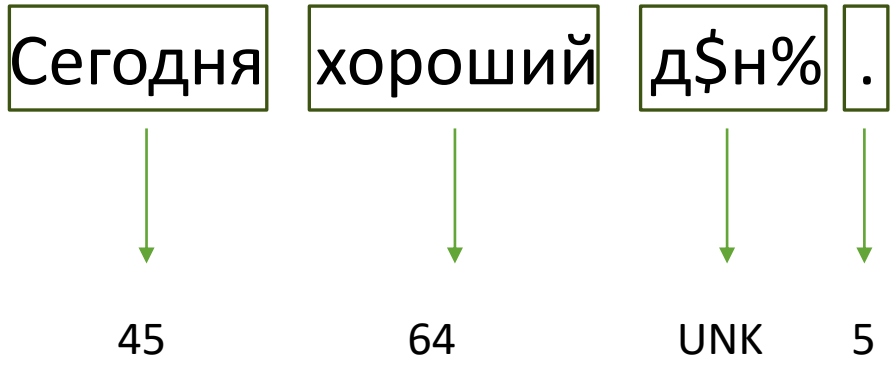


Словарь токенов

Неизвестные токены: Слова/единицы вне словаря

Некоторые токены могут не содержаться в словаре:

Каждый токен имеет свой номер в словаре токенов



Векторы с одним активным компонентом (One-Hot Vectors)

одна 1, остальные - 0

сегодня	0…0…0010…0…0….0…0
хороший	0…0…0000…0…1….0…0
день	0…0…1000…0…0….0…0

Размерность векторного представления = размеру словаря

Какие могут возникать проблемы при таком представлении слов?

Что такое смысл слова?

Например, испанское слово «***estante***».

Что такое смысл слова?

На всю стену стоит большой ***estante***.
В этом ***estante*** есть специальное место для верхней одежды.
Ребёнок не может достать до верхней полки ***estante***.
Мы купили этот ***estante*** в мебельном магазине.

Вы можете понять смысл слова с помощью контекста.

Что такое смысл слова?

На всю стену стоит большой _____.
В этом _____ есть специальное место для верхней одежды.
Ребёнок не может достать до верхней полки _____.
Мы купили этот _____ в мебельном магазине.

Вы можете понять смысл слова с помощью контекста.

Что такое смысл слова?

На всю стену стоит большой _____.
В этом _____ есть специальное место для верхней одежды.
Ребёнок не может достать до верхней полки _____.
Мы купили этот _____ в мебельном магазине.

Шкаф	1	1	1	1
Комод	1(?)	0	1	1

Похожие контексты, значит слова имеют похожие значения.

Дистрибутивная гипотеза

Слова, которые часто встречаются в похожих контекстах, имеют похожий смысл (Harris 1954).

Таким образом, основная идея представления слов:

разместить в векторы информацию про контексты

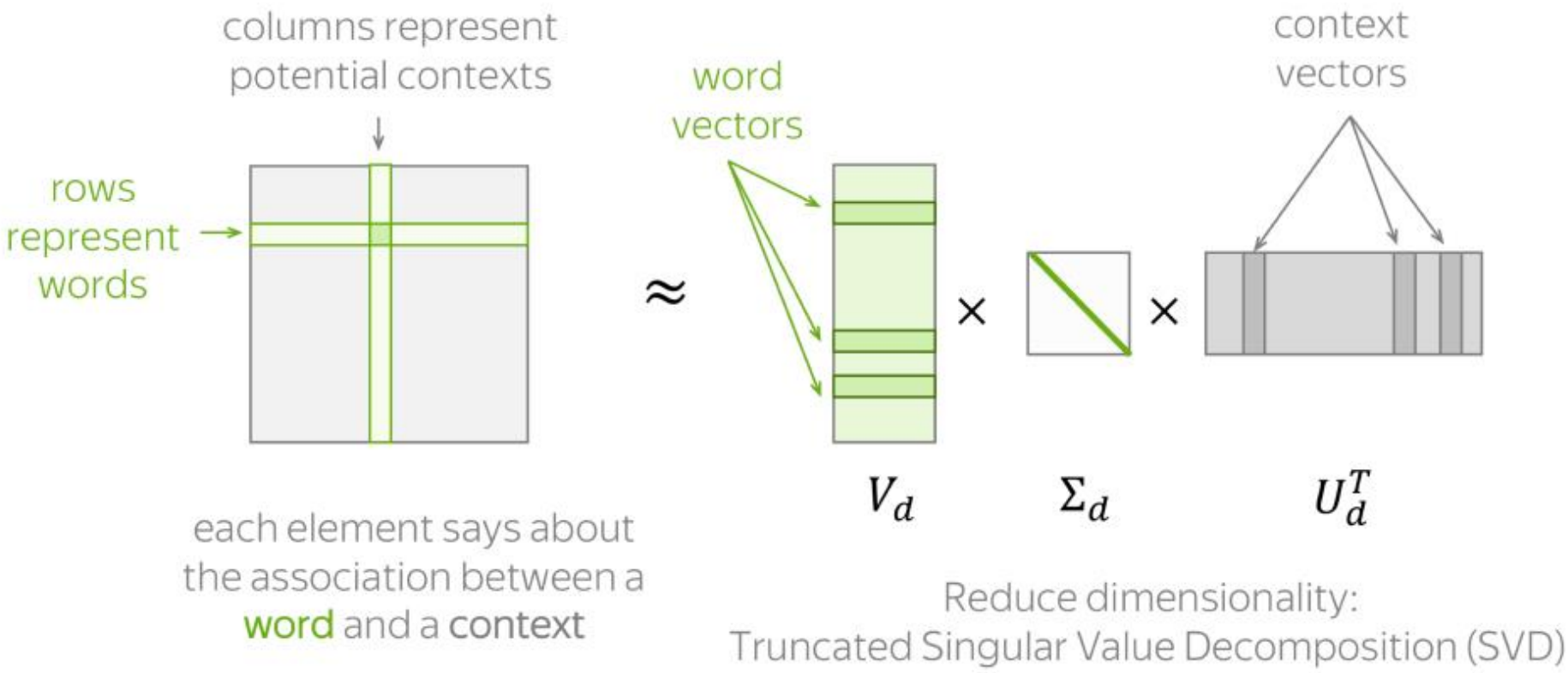
Идея методов, основанных на подсчёте (Count-Based Methods)

Основная идея:

разместить в векторы информацию про контексты

Данные методы подразумевают «вручную» задать эту информацию на основе глобальной статистики текстов.

Идея методов, основанных на подсчёте (Count-Based Methods)



Co-Occurrence Counts (как часто два слова стоят рядом)

Сегодня замечательный солнечный день, идеальный для лекции

Контекст для слова «день»: окружающие слова в окне размера L.

$N(w, c)$ — элемент матрицы, количество раз, которое слово w встречается в контексте c.

PPMI (Positive Pointwise Mutual Information)

Это мера, используемая для оценки того, насколько чаще два события (например, два слова) встречаются вместе по сравнению с тем, если бы они были статистически независимы.

- $PPMI(w, c) = \max(0, PMI(w, c))$,

where

$$PMI(w, c) = \log \frac{P(w, c)}{P(w)P(c)} = \log \frac{N(w, c)|(w,c)|}{N(w)N(c)}$$

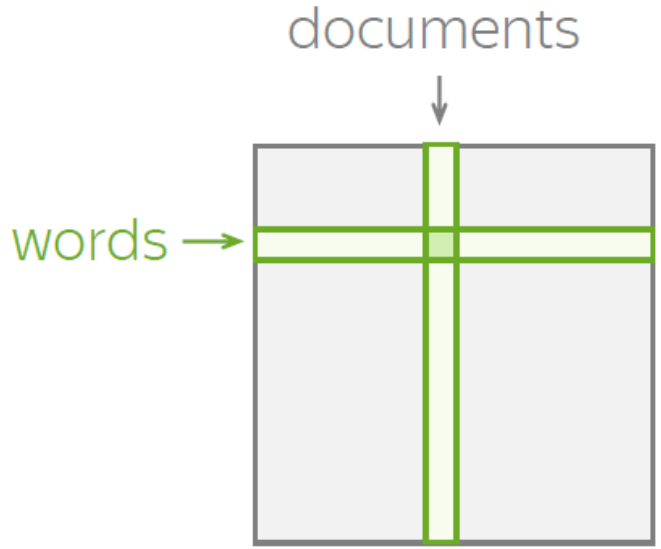
Анализ скрытых семантических структур (Latent Semantic Analysis (LSA))

Контекст: документ d из коллекции документов D

- $\text{tf-idf}(w, d, D) = \text{tf}(w, d) \cdot \text{idf}(w, D)$

$\text{tf}(w, d) = N(w, d)$

$$\text{idf}(w, D) = \log \frac{|D|}{|\{d \in D: w \in D\}|}$$



Count-Based Methods. Заключение

Основная идея:

- Положить в векторы информацию о контекстах.
- Заполнить эту информацию вручную на основе глобальной статистики корпуса.

Word2Vec (Word to Vector) (Prediction-Based Method)

Основная идея:

- положить информацию о контекстах в векторы
- обучение векторов слов через предсказание контекста

Изучаемые параметры: векторы слов

Цель: заставить каждый вектор «знать» о контекстах своего слова

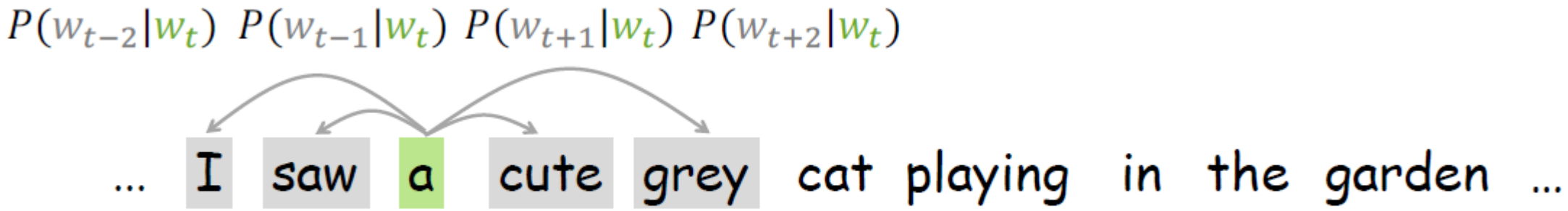
Способ: обучить векторы предсказывать возможные контексты по словам (или, альтернативно, по словам из контекстов)

Word2Vec: High-Level pipeline

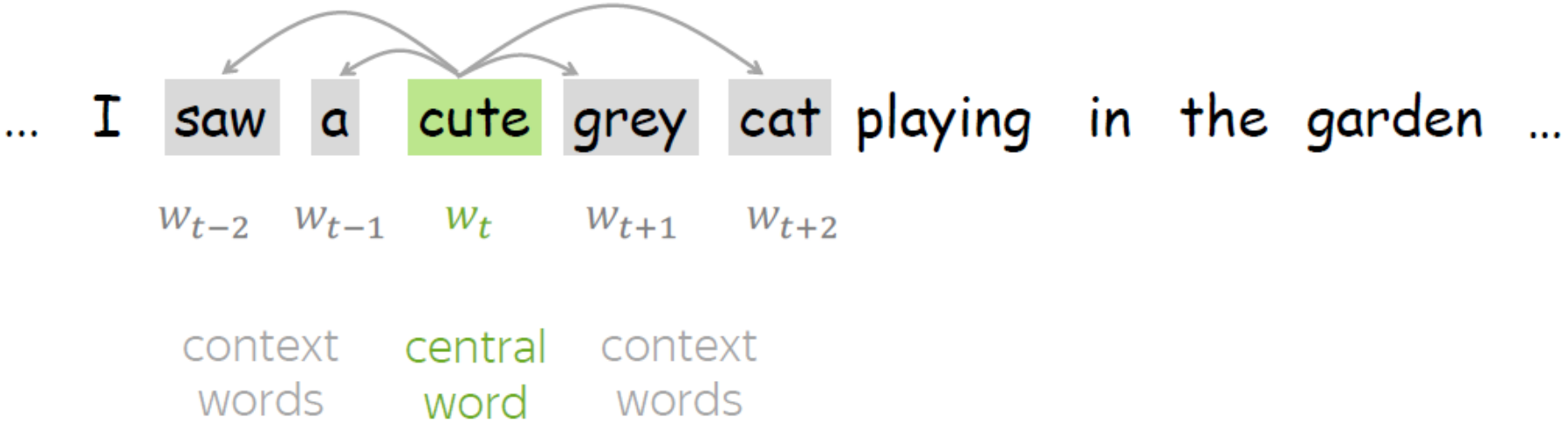
- взять большой текстовый корпус
- пройдитесь по тексту с помощью скользящего окна, перемещаясь по одному слову за раз
- для центрального слова вычислить вероятности контекстных слов
- скорректировать векторы, чтобы увеличить эти вероятности.

Word2Vec: High-Level pipeline

1.

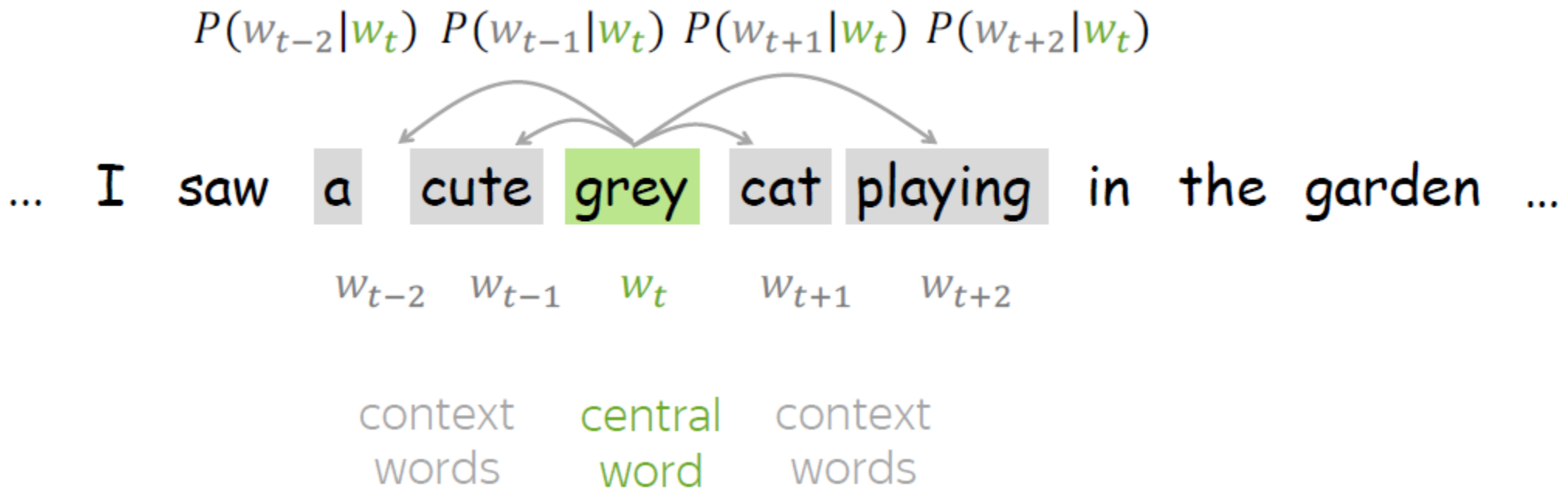


2.



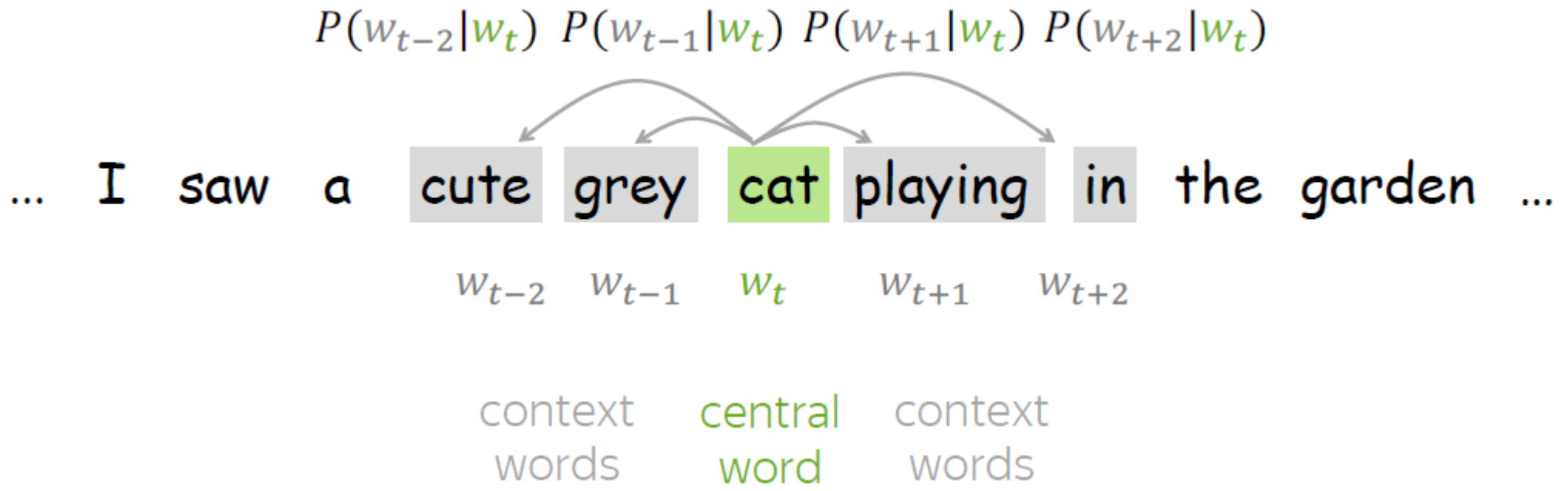
Word2Vec: High-Level pipeline

3.



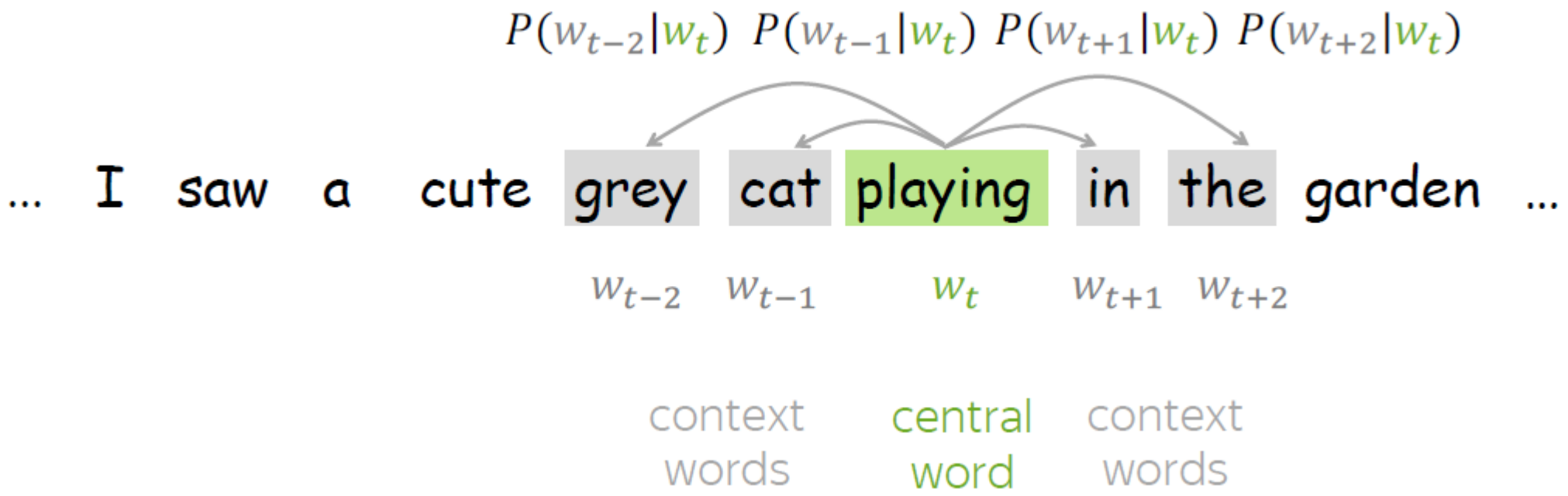
Word2Vec: High-Level pipeline

4.



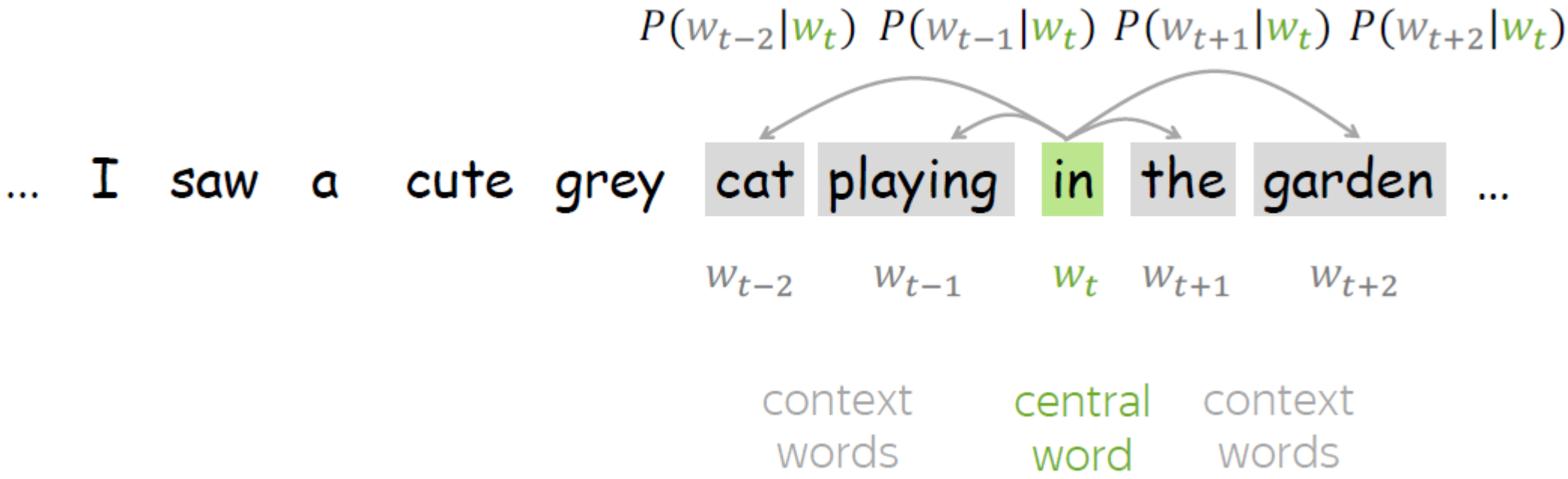
Word2Vec: High-Level pipeline

5.



Word2Vec: High-Level pipeline

6.



Целевая функция: отрицательный логарифм правдоподобия

Word2Vec пытается найти параметры, которые максимизируют вероятность данных:

Функция правдоподобия: $$L(\theta) = \prod_{t=1}^{T} \prod_{\substack{-m \le j \le m, \\ j \neq 0}} P(w_{t+j} | w_t, \theta)$$

где θ — оптимизируемые векторы

Целевая функция: отрицательный логарифм правдоподобия

Используем отрицательное (логарифмическое) правдоподобие в качестве функции потерь:

$$\text{Loss} = J(\theta) = -\frac{1}{T}\log L(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{-m\le j\le m,\\ j\neq 0}} \log P(w_{t+j}|w_t,\theta)$$

Вычисление вероятности в логарифмической функции правдоподобия

$P(w_{t+j}|w_t, \theta)$

Для каждого слова w существует два вектора:

- v_w для центральной позиции слова
- u_w для контекстной позиции слова

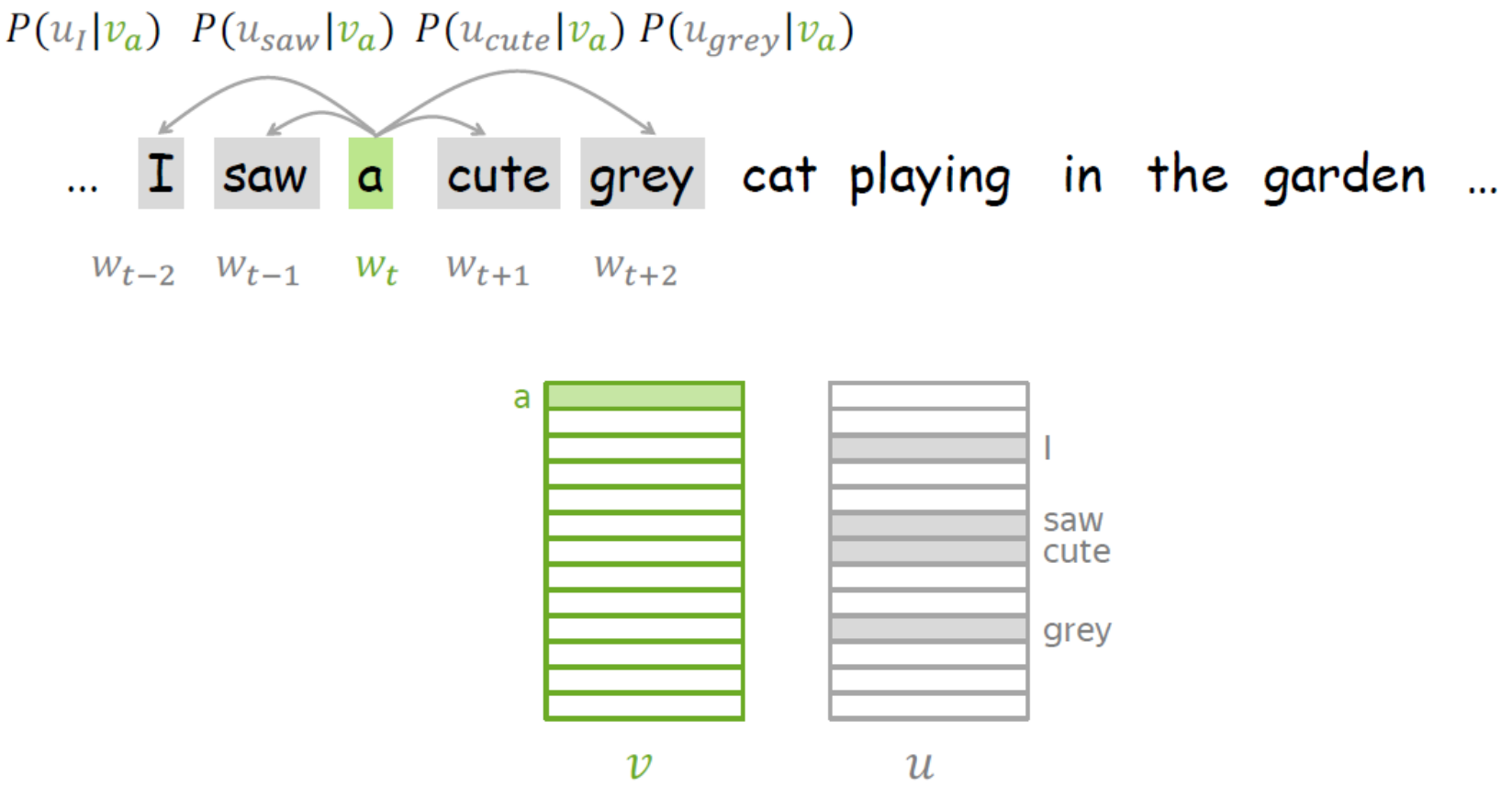
Для центрального слова c и контекстного слова o:

$$P(o|c) = \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}$$

Цель: скорректировать векторы для увеличения вероятностей

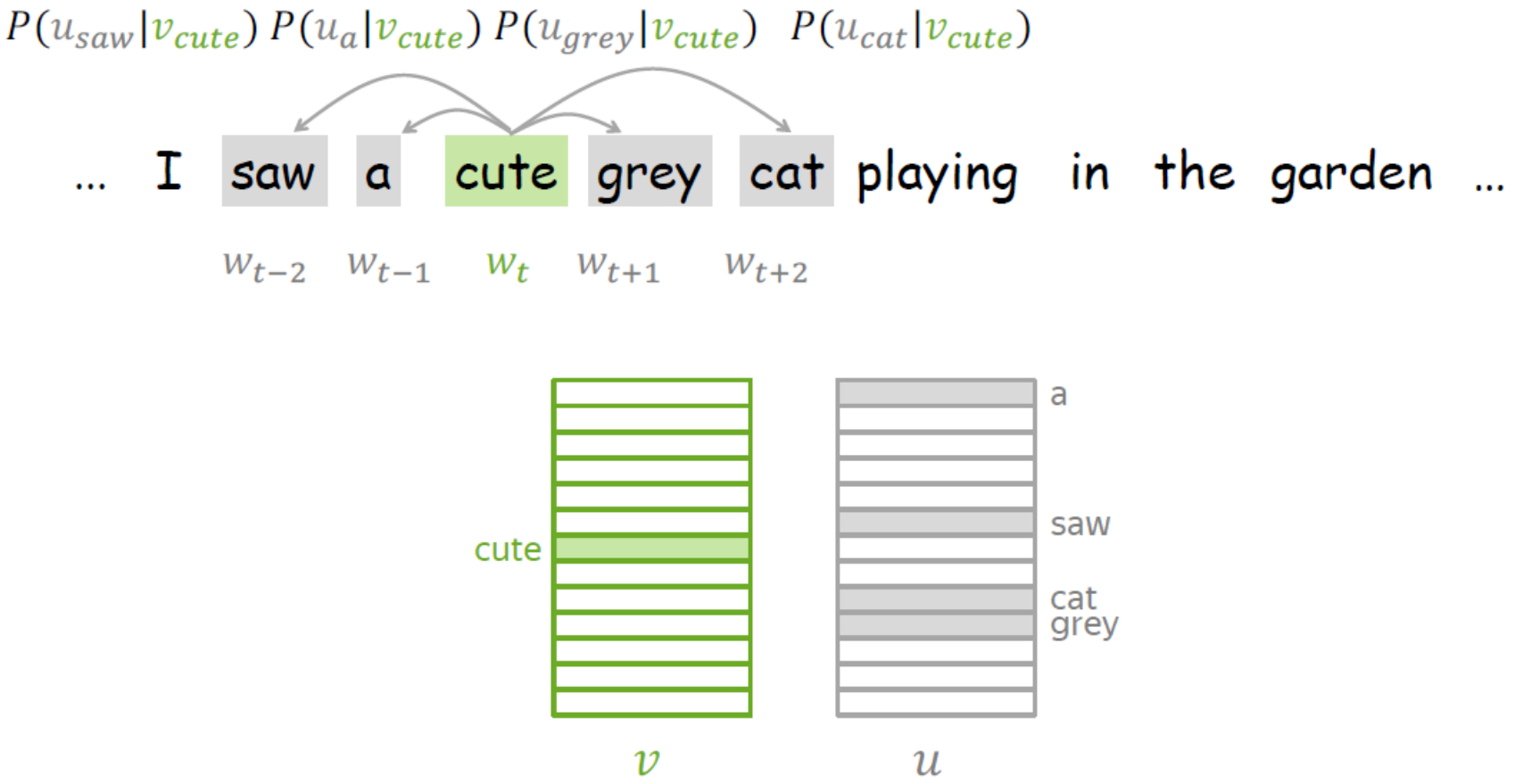
Два вектора для каждого слова. Пример

1.



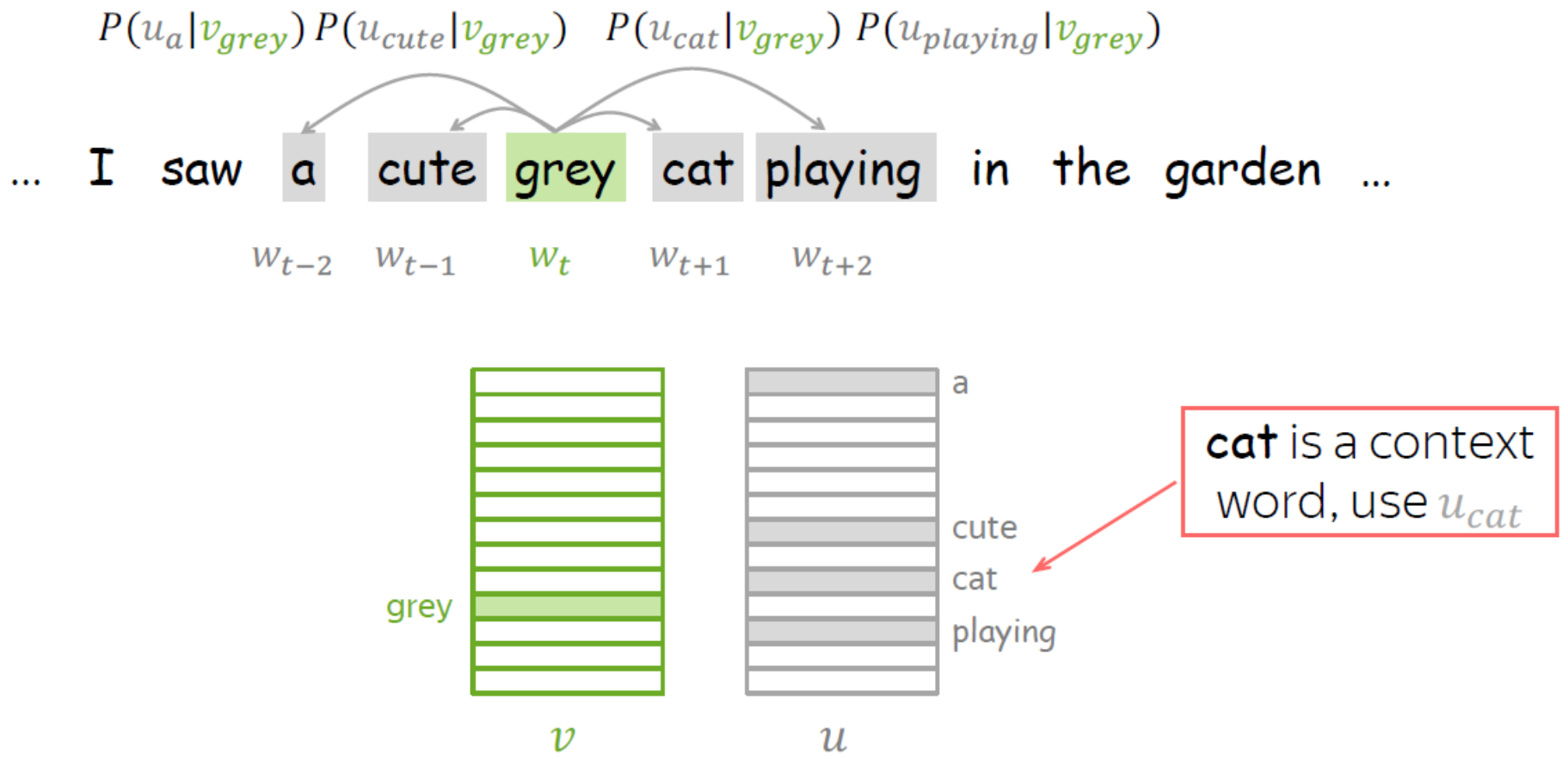
Два вектора для каждого слова. Пример

2.



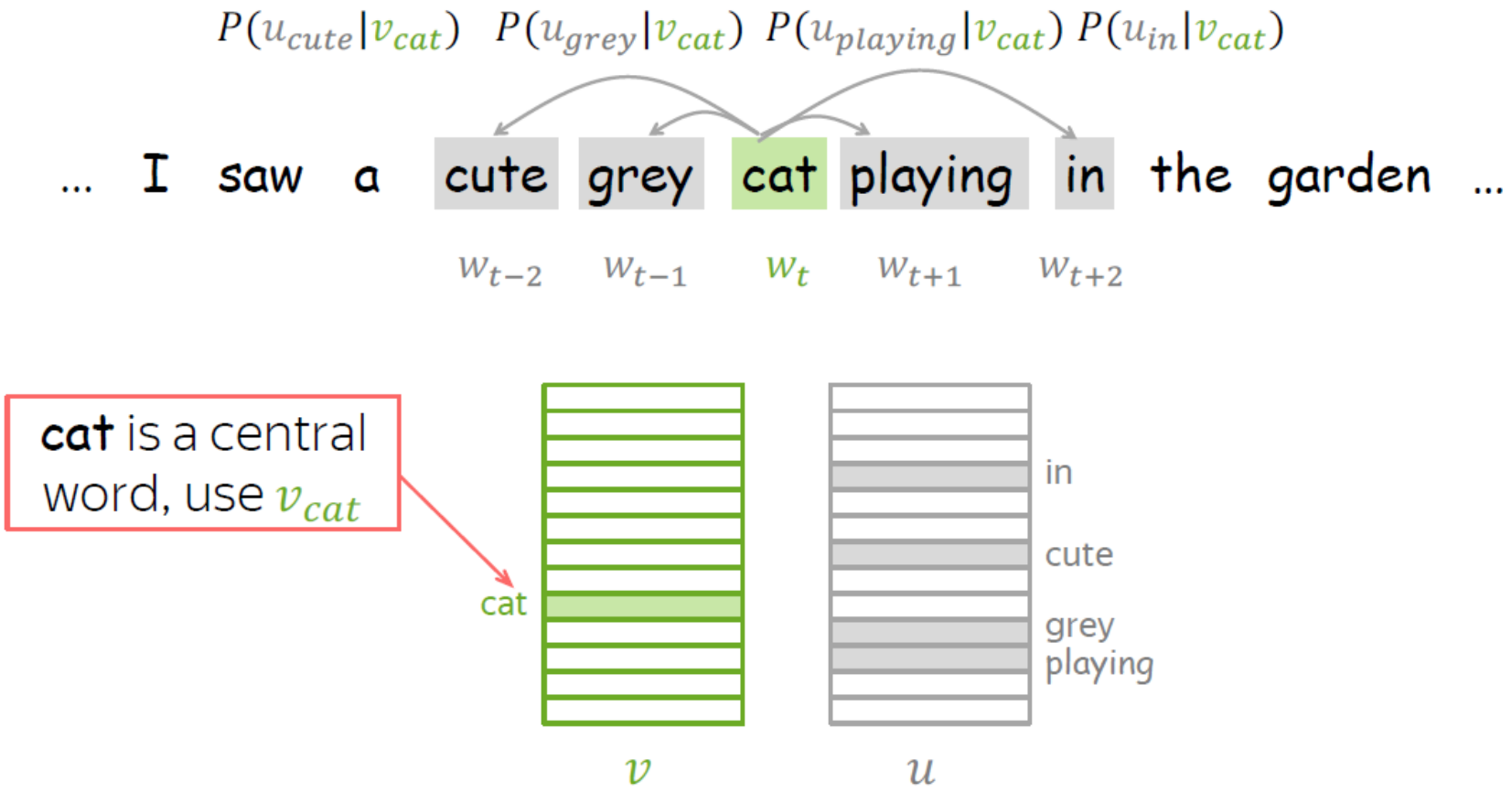
Два вектора для каждого слова. Пример

3.



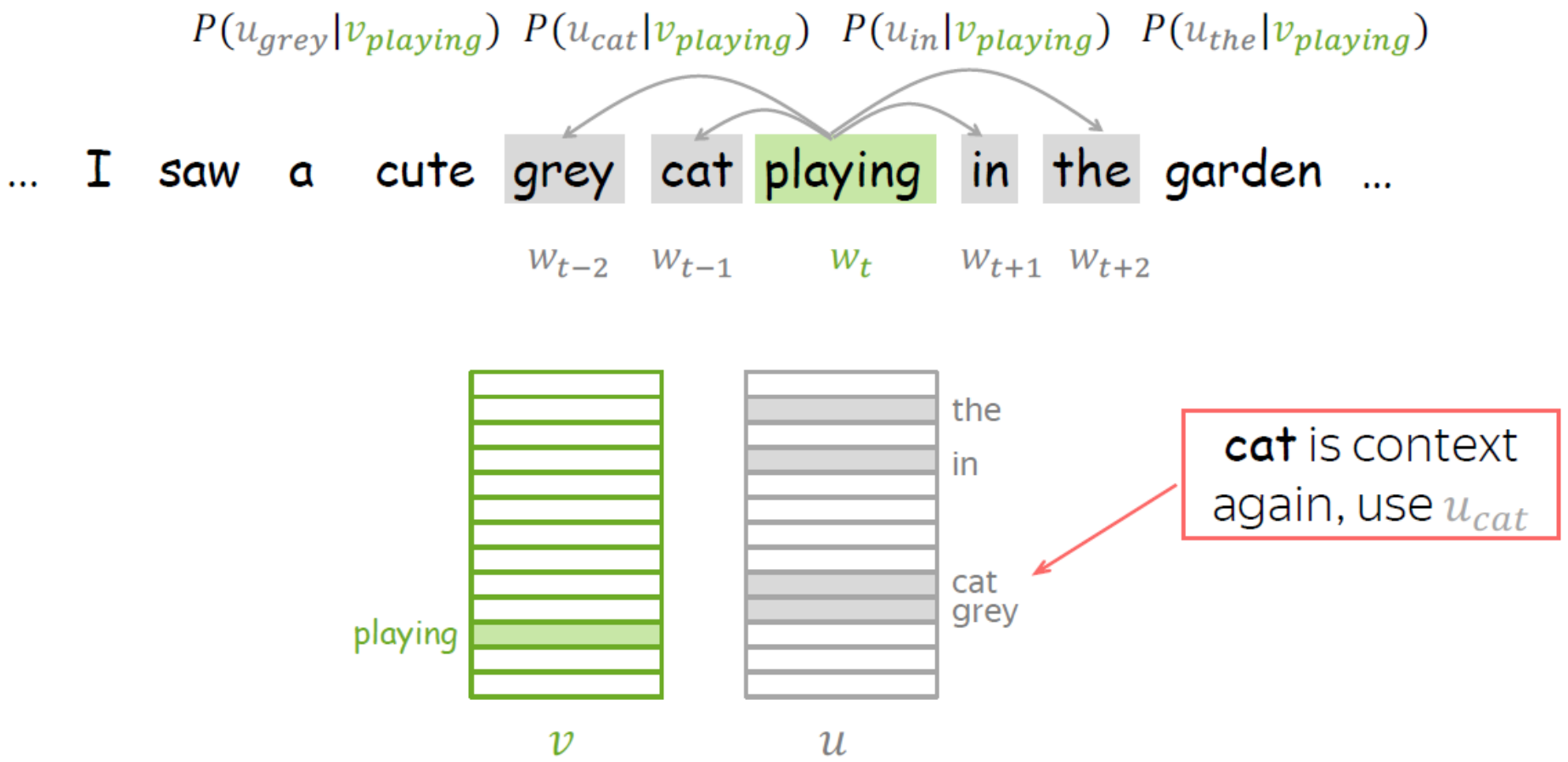
Два вектора для каждого слова. Пример

4.



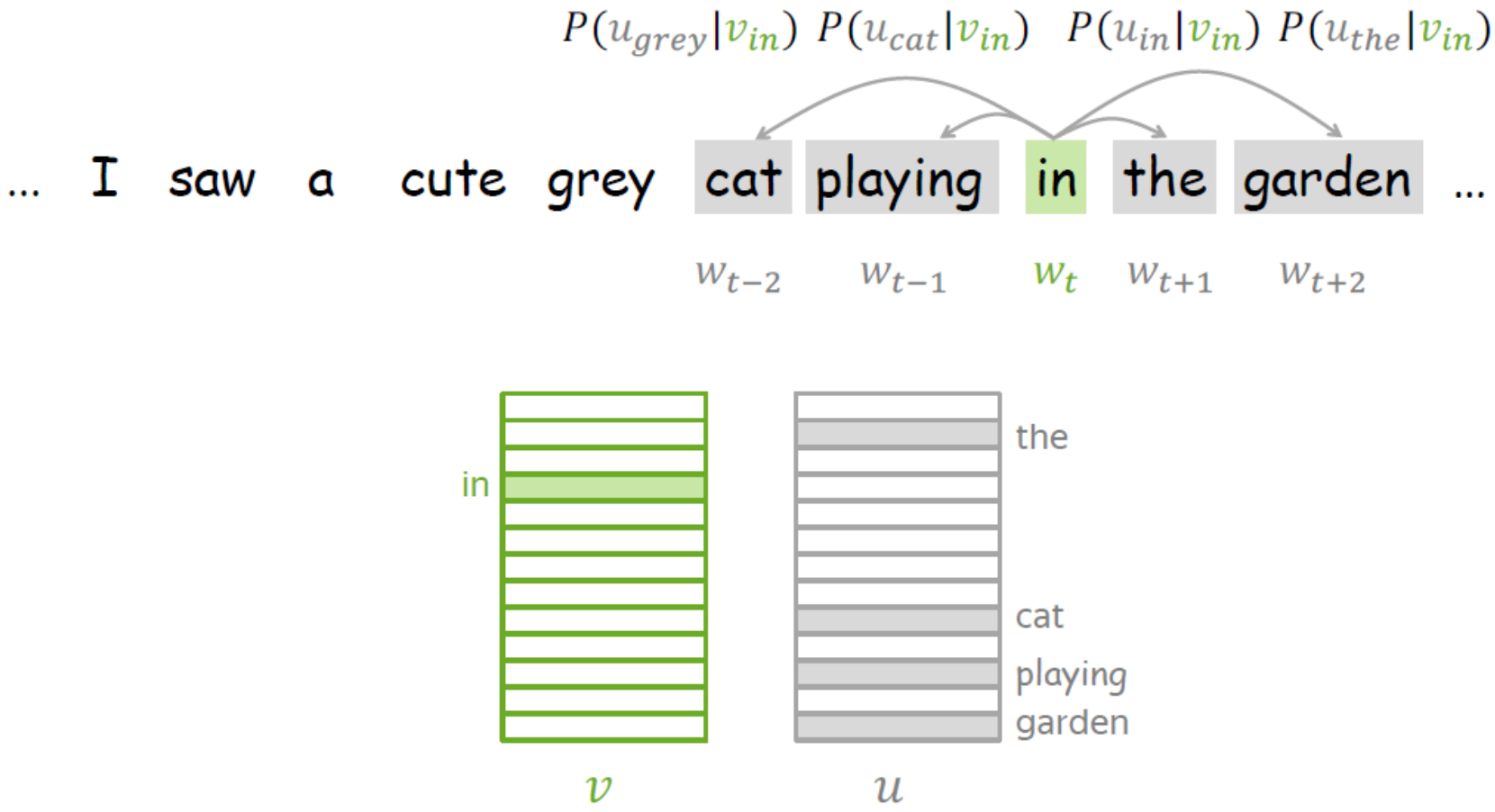
Два вектора для каждого слова. Пример

5.



Два вектора для каждого слова. Пример

6.



Процедура обучения

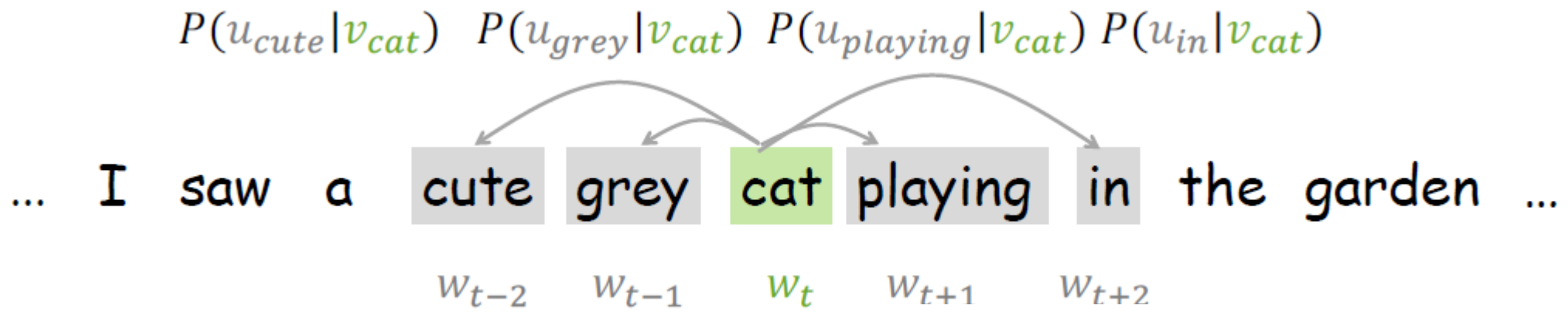
Есть функция потерь:

$$\text{Loss} = J(\theta) = -\frac{1}{T}\log L(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{-m\le j\le m,\\ j\neq 0}} \log P\left(w_{t+j}\middle|w_t,\theta\right)$$

Перепишем в виде: $-\frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{-m\le j\le m,\\ j\neq 0}} J_{t,j}(\theta)$

потеря для слова j в контексте t

Процедура обучения. Пример



Выбираем одно из контекстных слов:

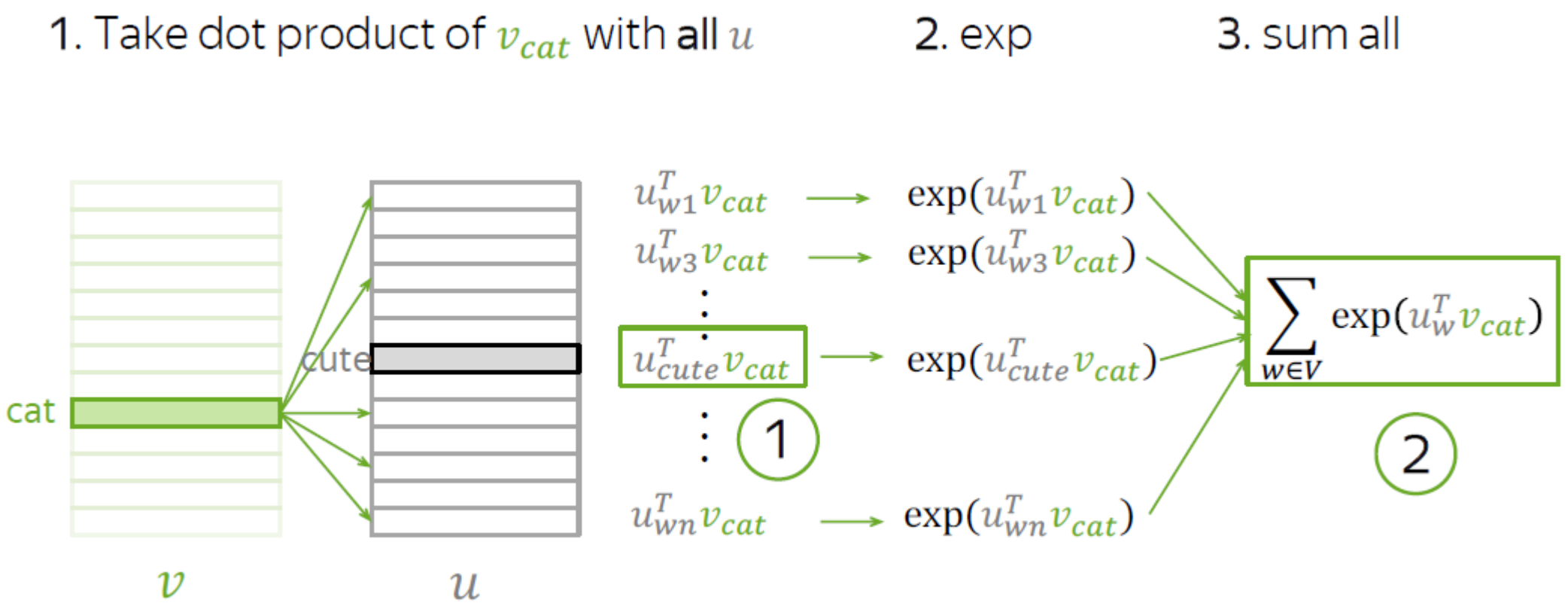
Процедура обучения. Пример

Рассмотрим функцию потерь для этого шага:

$$-\log P(cute|cat) = -\log \frac{\exp(u_{cute}^T v_{cat})}{\sum_{w \in V} \exp(u_w^T v_{cat})}$$

$$= -u_{cute}^T v_{cat} + \log \sum_{w \in V} \exp(u_w^T v_{cat})$$

Процедура обучения. Пример



Процедура обучения. Пример

4. get loss (for this one step)

$$J_{t,j}(\theta) = -\underbrace{u_{cute}^T v_{cat}}_{1} + \underbrace{\log \sum_{w \in V} \exp(u_w^T v_{cat})}_{2}$$

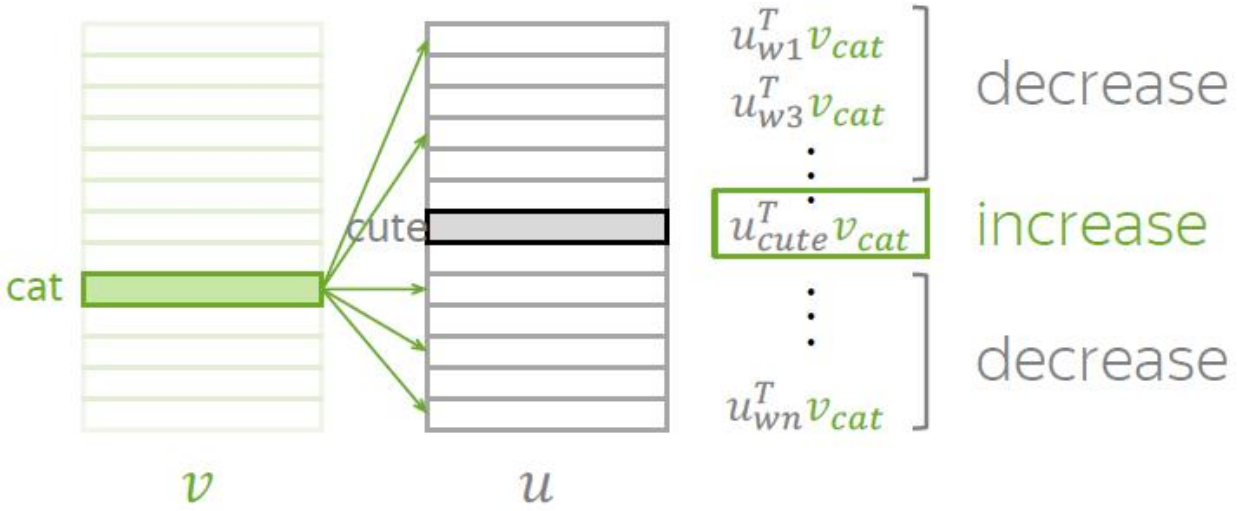
5. evaluate the gradient, make an update

$$v_{cat} := v_{cat} - \alpha \frac{\partial J_{t,j}(\theta)}{\partial v_{cat}}$$

$$u_w := u_w - \alpha \frac{\partial J_{t,j}(\theta)}{\partial u_w} \quad \forall w \in V$$

Что происходит за одно обновление?

$$-\log P(cute|cat) = -u_{cute}^T v_{cat} + \log \sum_{w \in V} \exp(u_w^T v_{cat})$$

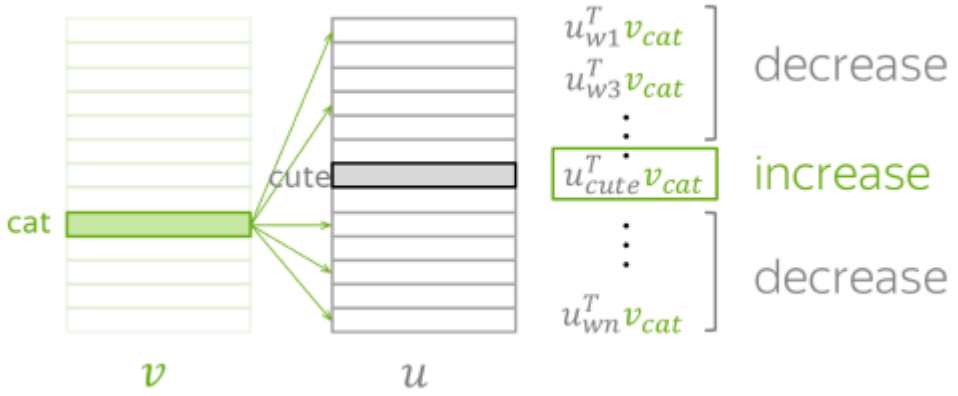


Результат каждого шага

Обычно увеличивается вероятность правильного контекста и уменьшаются остальные.

Необходимо обновить параметры:

- v_{cat}
- u_w для всех w в словаре



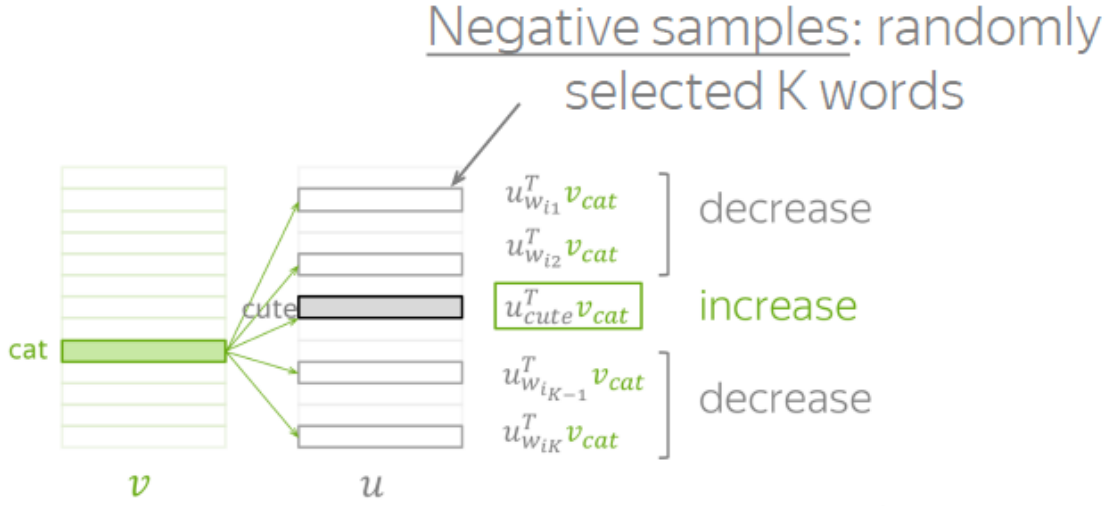
Проблема!

Negative Sampling (негативная выборка)

Идея: уменьшать вероятность какого-то подмножества слов (случайным образом выбирать K слов).

Необходимо обновить параметры:

- v_{cat}
- u_{cute} и u_w для всех w в подмножестве размера K



Word2Vec. Skip-Gram

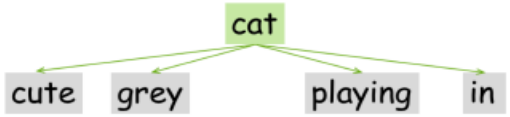
Главная гипотеза: Слова, которые встречаются в сходных контекстах, имеют сходные значения.

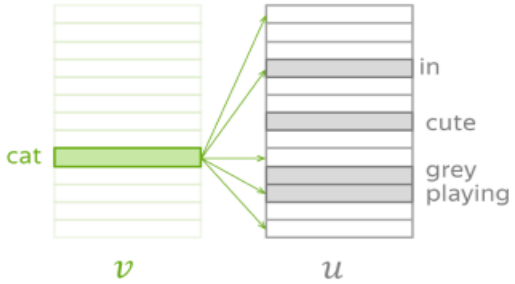
Задача Skip-Gram: Научить модель предсказывать окружающие слова (контекст) по заданному целевому слову.

Word2Vec. Skip-Gram. Пример

… I saw a cute grey cat playing in the garden …

Skip-Gram: from central predict context
(one at a time)





(this is what we did so far)

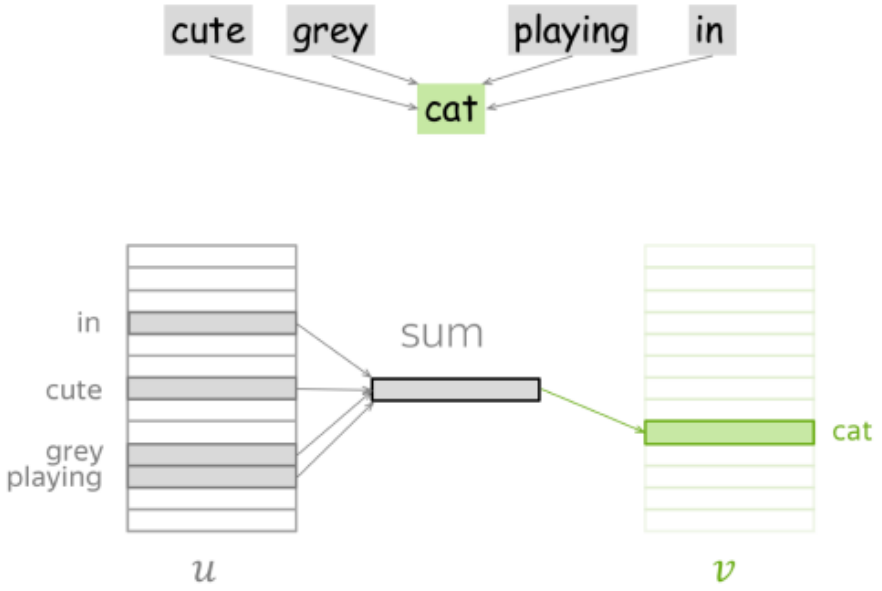
Word2Vec. CBOW

Главная гипотеза: Остается той же — слова в сходных контекстах имеют сходные значения. Но задача обратная.

Задача CBOW: Научить модель предсказывать **целевое (пропущенное) слово** по окружающим его словам (контексту).

Word2Vec. CBOW. Пример

CBOW: from sum of context predict central



(Continuous Bag of Words)

Стандартные гиперпараметры

- Модель: Skip-Gram с отрицательной выборкой;
- Количество отрицательных примеров: для небольших наборов данных 15–20; для больших наборов данных (которые обычно используются) может быть 2–5.
- Размерность встраивания: часто используемое значение — 300, но возможны и другие варианты (например, 100 или 50).
- Размер скользящего окна (контекста): 5–10.

GloVe (Global Vectors for Word Representation)

- В **Count-Based** методах информация приходит из глобальной статистики всего корпуса текстов, а векторы получены путём понижения размерности.
- В **Prediction-based** методах информация приходит из последовательного предсказания локальных контекстов, а векторы обучаются методом градиентного спуска.
- В **GloVe** информация приходит из глобальной статистики всего корпуса текстов, а векторы обучаются методом градиентного спуска.