1 Title Page

Chikiryakina_Karina

Использование моделей машинного обучения для поиска идентичных продуктов в электронной коммерции

Выполнил: студент 4 курса Чикирякина Карина Александровна
Научный руководитель: Прошунин Александр Иванович

Воронежский государственный университет, 2024 г.

Цель

Создание модели машинного обучения поиска идентичных продуктов для сайта сравнения цен.

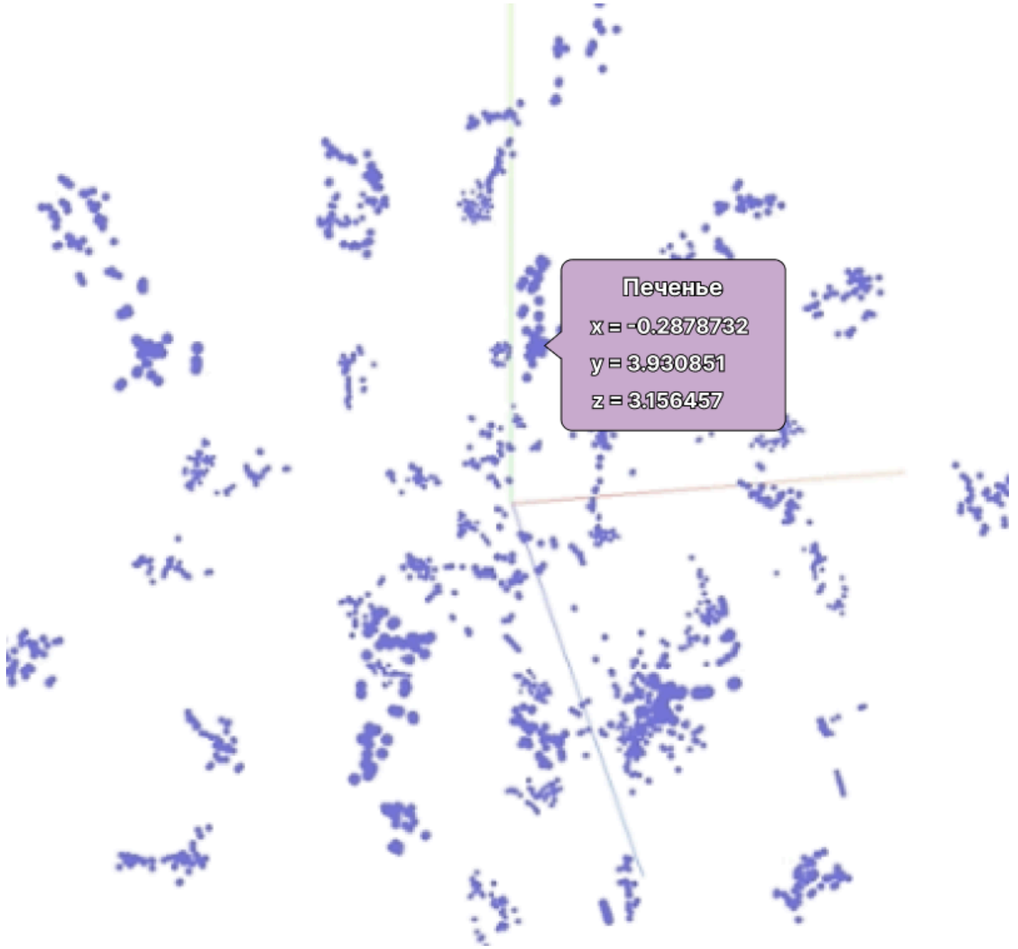
Задачи

1. Исследовать существующие разработки в данной сфере, выбрать подходящую модель на основе предоставленных компанией возможностей.
2. Подготовить датасет для обучения.
3. Реализовать модель машинного обучения, используя текст и изображения товаров.
4. Протестировать программу на новых данных и проанализировать получившийся результат.

Содержание

Матчинг продуктов

Сравнение описаний товаров, полученных из разных источников, чтобы собирать одинаковые товары от разных продавцов — и попадать в ожидания покупателей.



Векторное представление товаров

attention_is_all_you_need

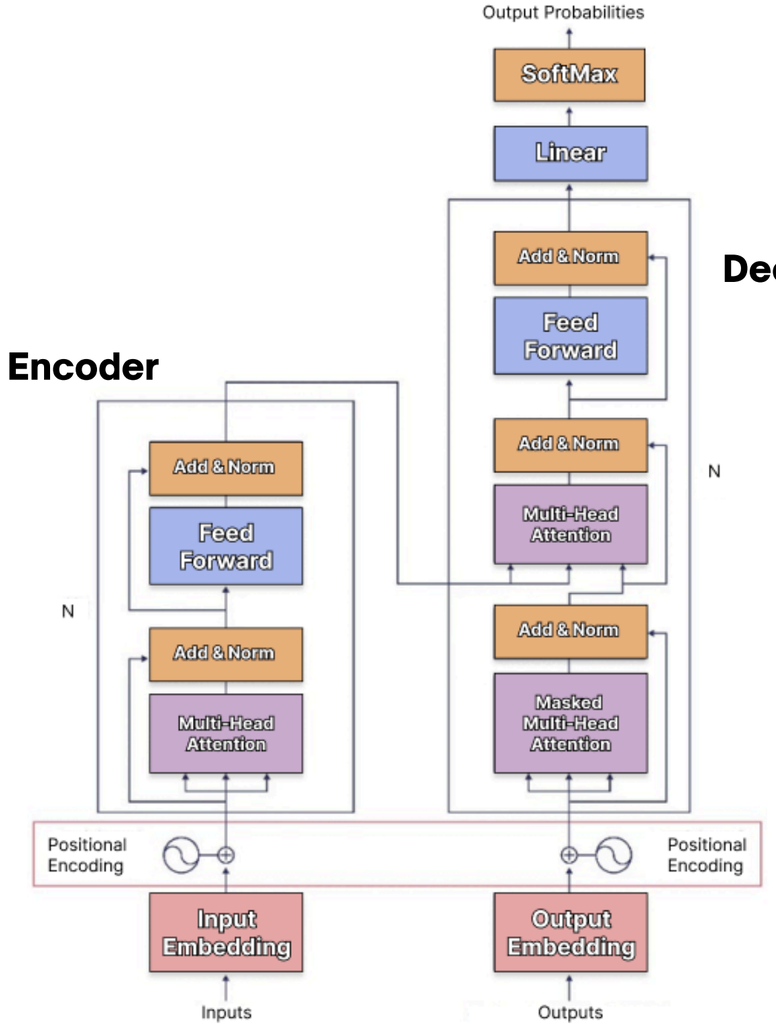
Трансформер

Энкодер

Синтезирует векторы, которые можно использовать в последующих задачах. К популярным моделям-энкодерам относится семейство BERT.

Декодер

Из полученных векторов снова генерирует последовательность токенов. Главное отличие от энкодера — маскирование. К популярным моделям-декодерам относится семейство GPT.



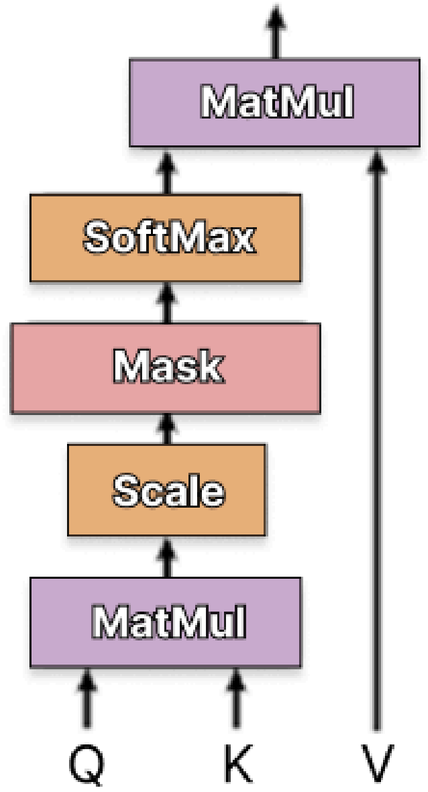
Модель трансформер представленная компанией Google в 2017г

attention

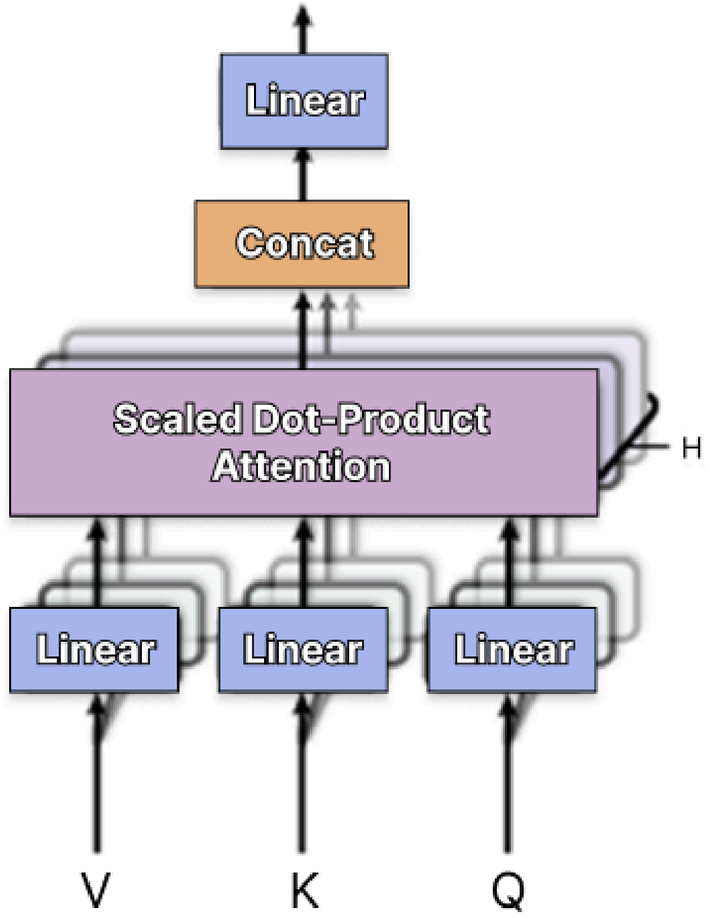
Слой внимания

Multi-Head Attention

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$$



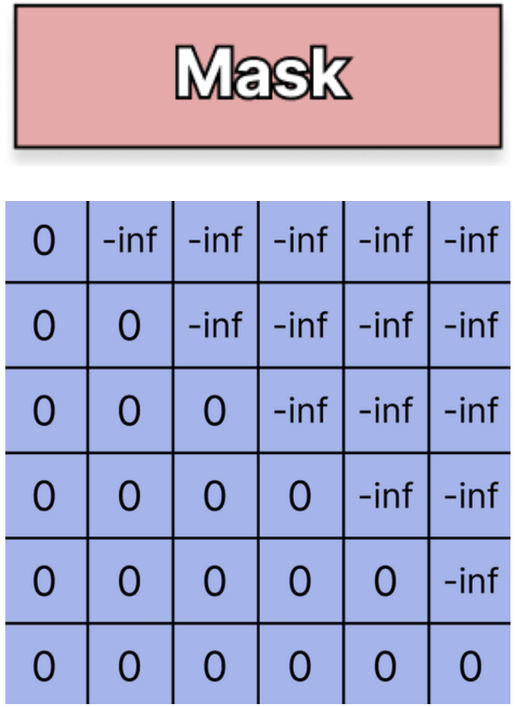
Scaled dot-product attention
(слой внимания)



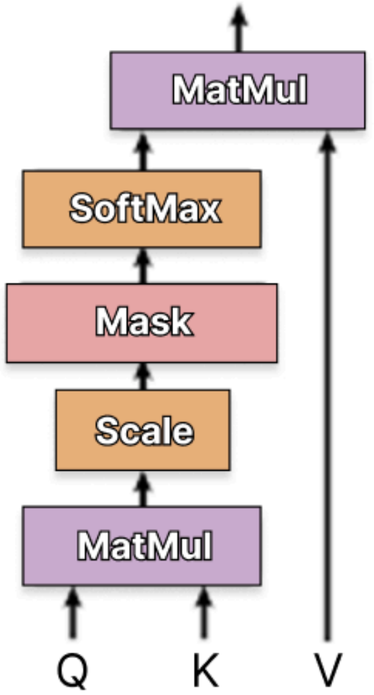
Multi-head attention
(многоголовое внимание)

Маска

Маска используется в декодере на этапе обучения. Если мы сгенерировали m слов, то все последующие будут замаскированы путем прибавления -inf. Это скрывает от модели будущие токены, давая возможность ей самостоятельно предсказать значения.



Маска декодера



Scaled dot-product attention (слой внимания)

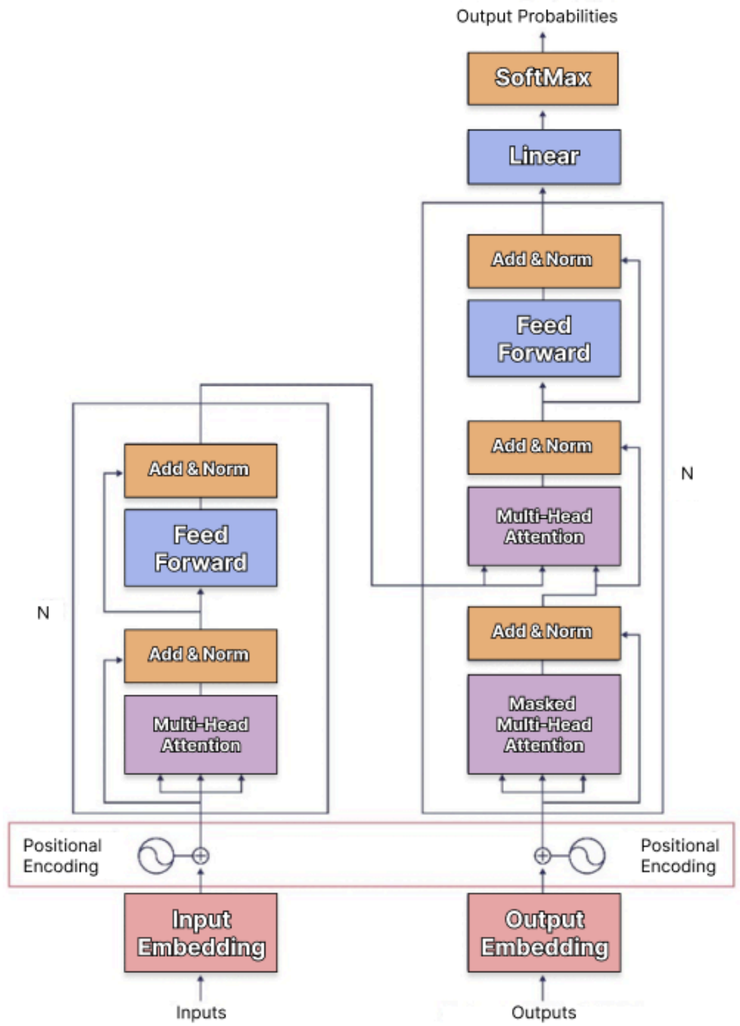
normalization

Нормализация

Складываем матрицу на входе и на выходе многоголового внимания. Выполняем нормализацию:

$$y = \frac{x - E[x]}{\sqrt{Var[x] + e}} \cdot \alpha + \beta$$

Add & Norm



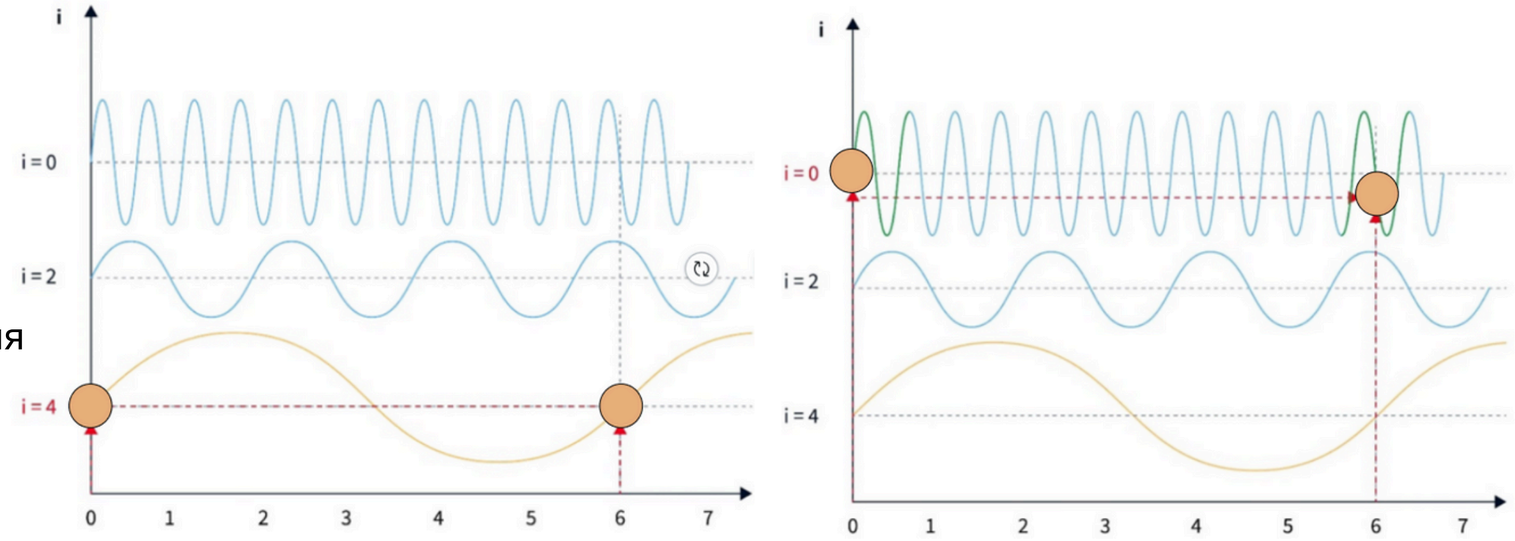
Модель трансформер

Позиционное кодирование

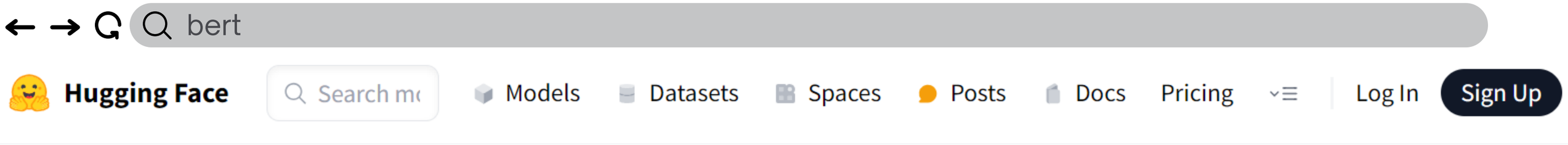
Метод, используемый в обработке естественного языка (NLP) для представления порядка слов в тексте. Это важно, поскольку значение слов может зависеть от их положения в предложении.

$$PE_{(pos,\ 2i)} = \sin\left(\frac{pos}{10000^{2i/d}}\right)$$

$$PE_{(pos,\ 2i+1)} = \cos\left(\frac{pos}{10000^{2i/d}}\right)$$



Суть позиционного кодирования, предложенного компанией Google



Rubert-Tiny

BERT Bidirectional Encoder Representation Transformers ("двунаправленная нейронная сеть-кодировщик").

BERT представляет собой лишь часть исходного трансформера (энкодер).

На выходе модели генерируются векторы новых последовательностей, которые могут использоваться для широкого спектра задач. Это позволяет настраивать предварительно обученную модель BERT с помощью лишь одного дополнительного выходного слоя.

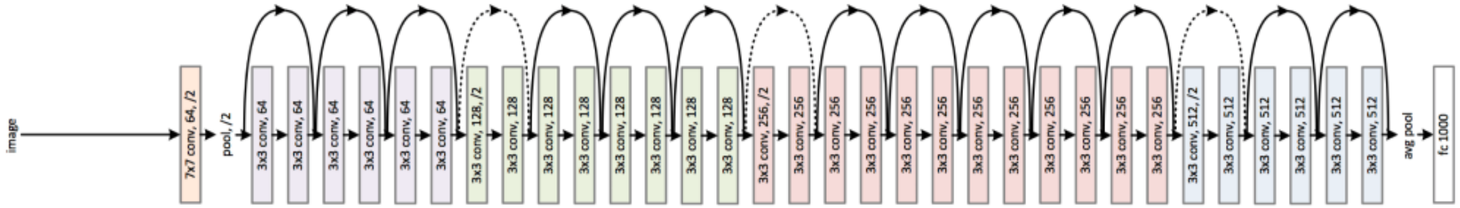
Rubert-Tiny

Очень маленькая версия модели bert-base-multilingual-case для русского и английского языков.

Для текстов использовалось два Rubert-Tiny — для заголовков и атрибутов

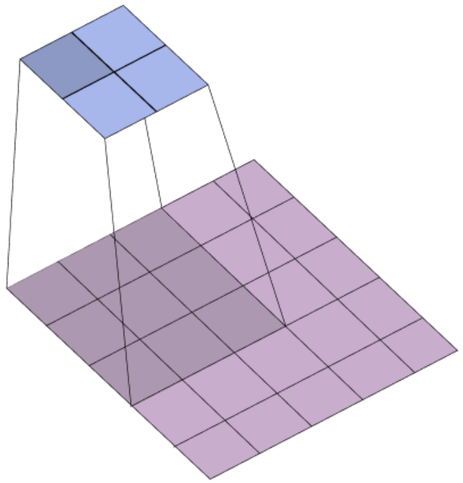
resnet34

ResNet34



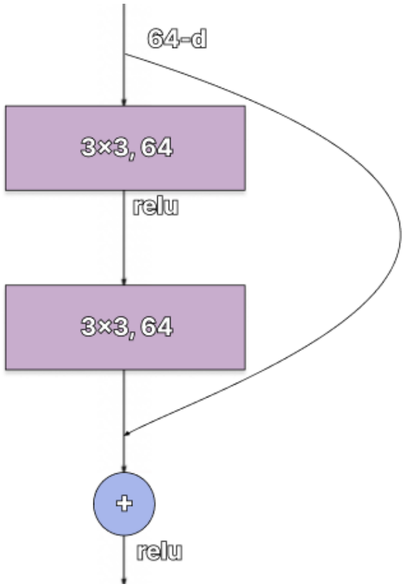
Сверточные сети

Способна захватывать пространственные зависимости в изображении с помощью соответствующих фильтров.



Остаточные связи

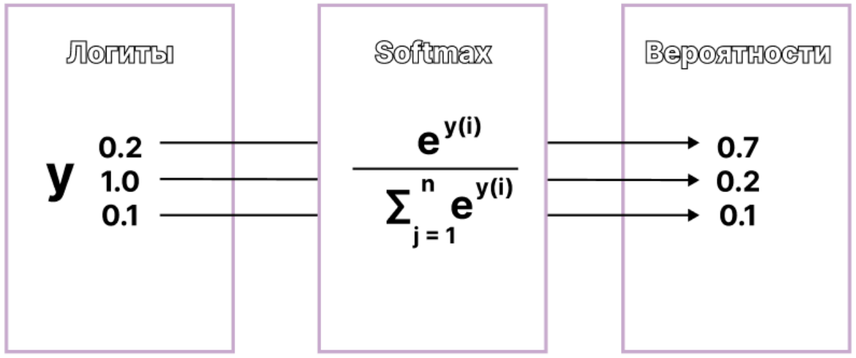
Позволили смягчить эффект исчезающих градиентов.



ArcFace

Softmax

Функция потерь softmax не обеспечивает более высокого сходства для выборок внутри класса и разнообразия для выборок между классами.



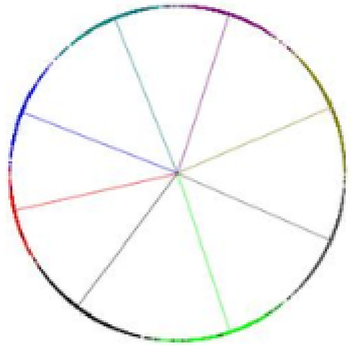
Функция Softmax

ArcFace

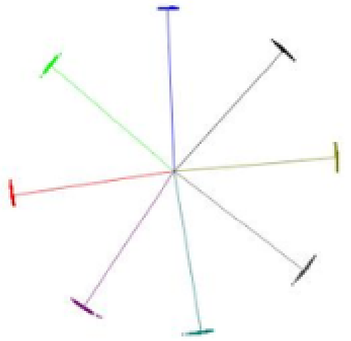
$$\frac{e^{s\cos(\theta_{y(i)}+m)}}{e^{s\cos(\theta_{y(i)}+m)}+\sum_{j=1,j\neq y(i)}^{n} e^{s\cos(\theta_{y(i)})}}$$

Функция ArcFace

Softmax



ArcFace



Расположение объектов разных классов на окружности после применения softmax и arcface

data

jupyter Data Last Checkpoint: last month

File Edit View Run Kernel Settings Help

Trusted

Программная реализация

Данные

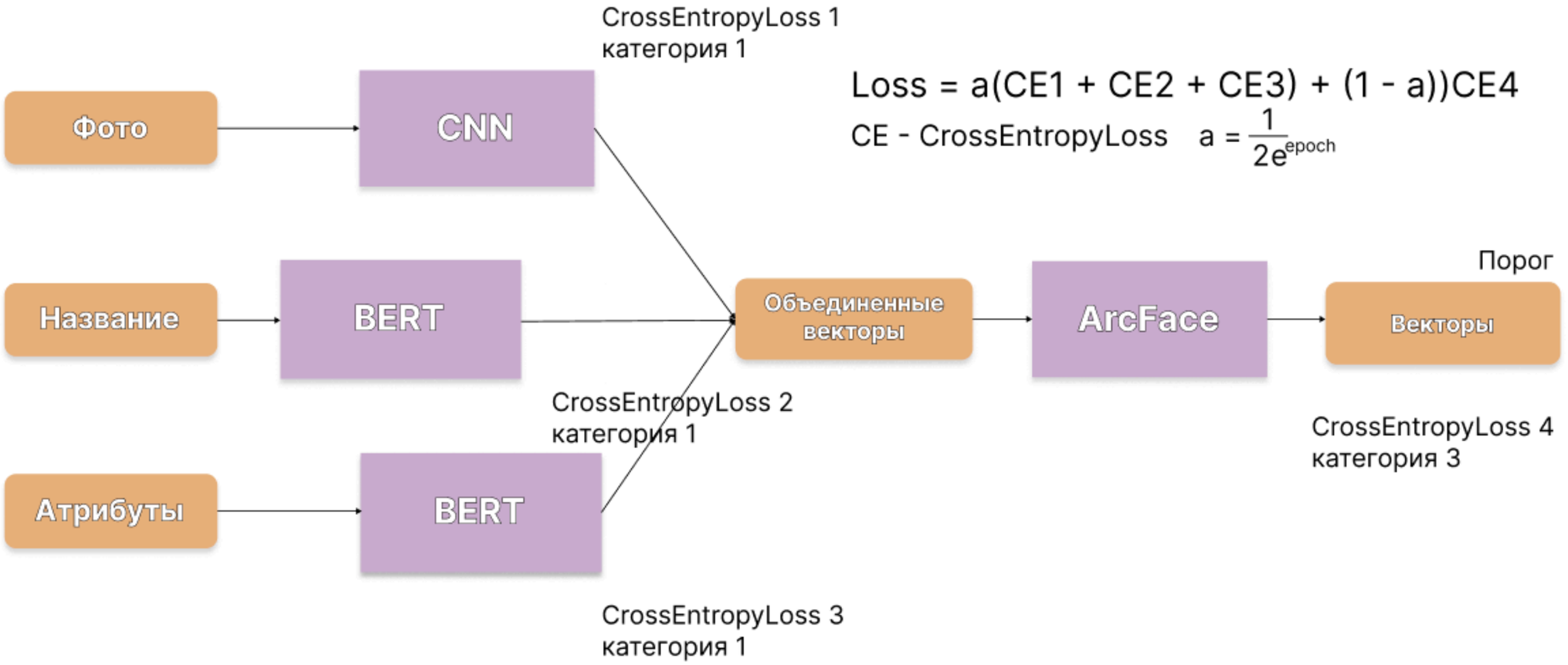
[32]:

	название	изображение	атрибуты	кат. 1	кат. 2	кат. 3
0	Торт Kristof клубника со сливками, 450 г	small_1000d62a27c1704e71336b046d4e4f3838e85f75...	{'Бренд': 'Kristof', 'Вид упаковки': 'Пластик'...	Сладости	Торты	None
1	Набор одноразовых стаканов Actuel пластик, 20х...	small_4ed8d12281a9102b9777887b51b814195a190945...	{'Бренд': 'Actuel', 'Масса нетто, кг': '0.2', ...	Кухня	Стаканы, бокалы	None
2	Сироп «Баринофф» десертный Вишня, 1 л	small_4abc7e4195076a6538d2ef5e84c89bcca4b73702...	{'Бренд': 'Баринофф', 'Упаковка': 'Стеклянная ...	Сладости	Сиропы	None
3	Пирожное «Медвежонок Барни» бисквитное с молоч...	small_7c4033e6867cdcfbdb3a847dc97d2401f06ce2d5...	{'Бренд': 'Барни', 'Белки на 100 г, г': '6', '...	Сладости	Пирожные, десерты	None
4	Пирожное «Медвежонок Барни» бисквитное со сгущ...	small_bdbe36876182b6fd32dc9e0989ecfa0fef23baa1...	{'Бренд': 'Барни', 'Белки на 100 г, г': '6', '...	Сладости	Пирожные, десерты	None

31786 строк x 6 столбцов

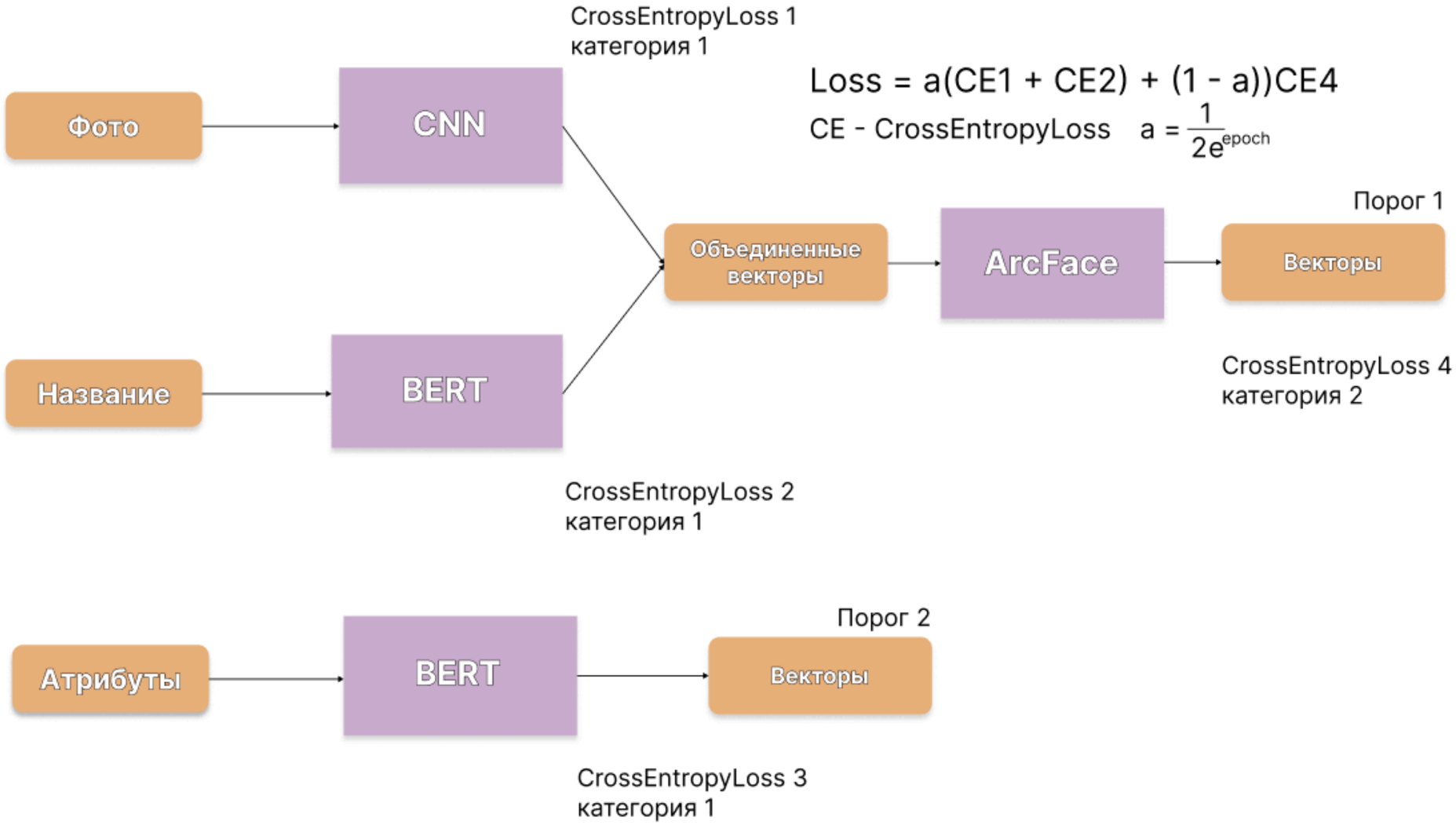
ozon.architecture

OZON.RU



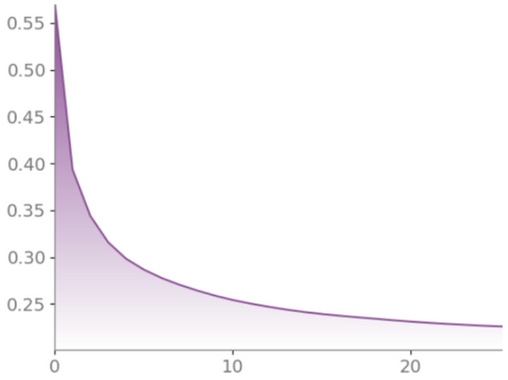
Архитектура модели OZON

architecture

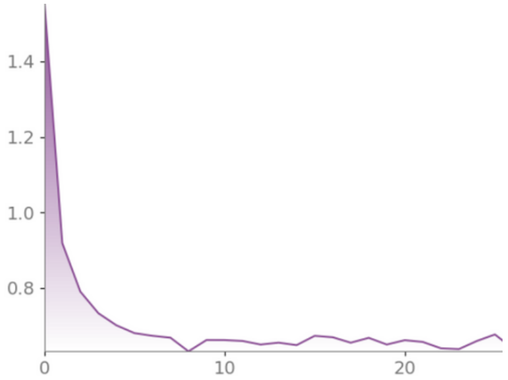


Новая архитектура

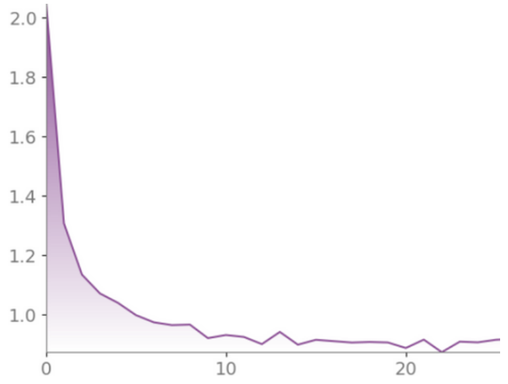
Обучение



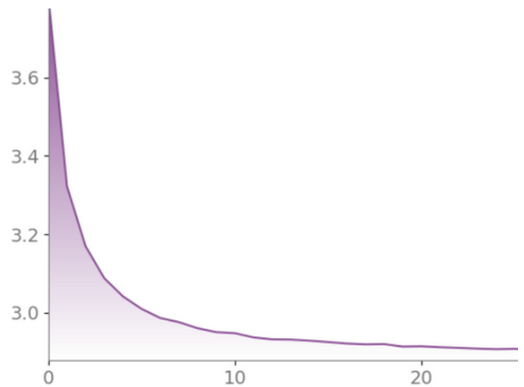
01 CrossEntropyLoss ResNet для изображений (25 эпох)



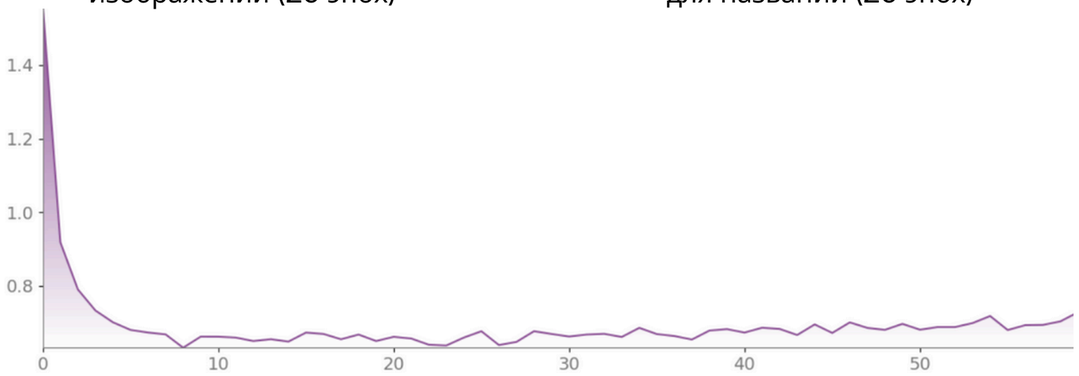
02 CrossEntropyLoss Rubert-Tiny для названий (25 эпох)



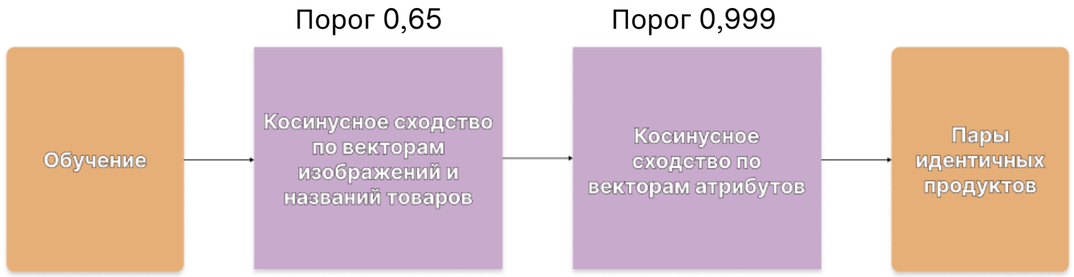
03 CrossEntropyLoss Rubert-Tiny для атрибутов (25 эпох)



04 CrossEntropyLoss ArcFace (25 эпох)



05 CrossEntropyLoss Rubert-Tiny для названий (60 эпох)



Отбор идентичных продуктов по установленным порогам сходства

Результаты

Обычный поиск сравнивания названий продуктов обнаружил 3953 пар.
Модель смогла обнаружить их все и еще представила 506 идентичных товаров.

Примеры 01 и 02: полностью идентичные товары, которые нашли алгоритм и модель.

Примеры 03: отличается буквой ё. Обычный алгоритм решил, что продукты разные, модель верно определила их идентичность.

Примеры 04: названия одинаковые, но отрибуты отличаются (ЗМЖ/БЗМЖ) Обычный алгоритм решил, что продукты идентичные, модель верно определила их различие.

01 Соус Uni Dan брусничный для вторых обеденных блюд, 270мл

02 Горбуша + скумбрия Европром рубленая с чесноком, 180г

03 Шейка свиная Черкизово категории Б охлаждённая

Шейка свиная Черкизово категории Б охлажденная

04 Мороженое ванильное «Время Летать» с черничным наполнителем, 450 г (ЗМЖ)

Мороженое ванильное «Время Летать» с черничным наполнителем, 450 г (БЗМЖ)

Вывод:

Создана модель машинного обучения поиска идентичных продуктов, которая успешно применяется на сайте сравнения цен.

Как можно улучшить?

01 Увеличить количество товаров

02 Восполнить пропущенную информацию в данных

03 Разметить данные парами (идентичные/разные)

04 Взять более сложные архитектуры BERT
(увеличит время поиска, но повысит качество результата)

Список использованной литературы

01 Ashish Vaswani, Attention Is All You Need / Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. arXiv: 1706.03762v7, 2023.

02 Kaiming He, Deep Residual Learning for Image Recognition / Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. arXiv: 1512.03385v1, 2015.

03 Jiankang Deng, ArcFace: Additive Angular Margin Loss for Deep Face Recognition / Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, Stefanos Zafeiriou. arXiv: 1801.07698v4, 2015.

04 Ajinkya More, Product Matching in eCommerce using deep learning. Walmart Global Tech Blog, 2017.

05 Petar Ristoski, A Machine Learning Approach for Product Matching and Categorization / Petar Petrovski, Peter Mika, Heiko Paulheim, 2016.

Спасибо за внимание!