РАЗРАБОТКА АГРЕГАТОРА ТЕМАТИЧЕСКИХ ЭЛЕКТРОННЫХ РЕСУРСОВ

С. С. Кацуба

stas.katsuba@gmail.com;
Научный руководитель — И. Д. Лукьянов, старший преподаватель

С развитием технологий модернизировались методы получения информации, а с появлением сети Интернет возникла необходимость в нахождении методов классификации и группировки информации. Решение этих задач требовалось как при создании первых поисковых машин, так и сейчас для работы индексаторов, агрегаторов и более продвинутых поисковых машин. Но если в начале 1990-х годов использовали более простые алгоритмы решения задачи поиска и классификации, которые позволяли эффективно обрабатывать малый объем данных того времени, то с увеличением объема данных в сети Интернет алгоритмы пришлось модернизировать.

В работе были разработаны алгоритмы решения задач сбора данных из различных электронных ресурсов, извлечения информации из полученных данных, а также задачи кластеризации и классификации объектов. С помощью разработанных алгоритмов были получены агрегатор новостных сюжетов и агрегатор товаров по его названию и изображению.

Ключевые слова: агрегирование; кластеризация; нейронные сети; парсинг; извлечение информации; классификация изображений; обработка текстов; машинное обучение; векторное представление текстов.

ПОСТАНОВКА ЗАДАЧИ

Под задачей агрегации информации будем подразумевать совокупность задач сбора, извлечения признаков и объединения информации. В контексте обработки данных, агрегация информации относится к процессу комбинирования данных из различных источников с целью получения обобщенного представления.

СБОР ДАННЫХ

Под задачей сбора данных будем подразумевать задачу получения содержимого электронных ресурсов, которые отвечают определенным требованиям. Решение задачи сбора данных можно свести к автоматизации GET-запросов к электронным ресурсам, с последующей фильтрацией и обработкой полученных HTML-объектов [1].

Во время автоматизации запросов возникает необходимость обработки статических и динамических электронных ресурсов. Статические электронные ресурсы сохраняют структуру документа во время работы с ним, тогда как динамические могут изменять ее с

помощью различных скриптов. Для обработки динамических необходимо эмулировать работу пользователя с электронными ресурсами.

Полученные данные проверяются на соответствие необходимой структуре, после чего сохраняются и фильтруются. Хранение данных осуществляется в реляционной базе данных.

ИЗВЛЕЧЕНИЕ ПРИЗНАКОВ И ОБЪЕДИНЕНИЕ ОБЪЕКТОВ

Для работы с текстовыми и графическими данными будем использовать их представление в некотором латентном векторном пространстве R^N. Близость векторных представлений объектов в этом пространстве будет в некоторой степени характеризовать их похожесть. В таком случае, под сюжетом новостей можно понимать кластер векторных представлений новостных артиклей в латентном пространстве.

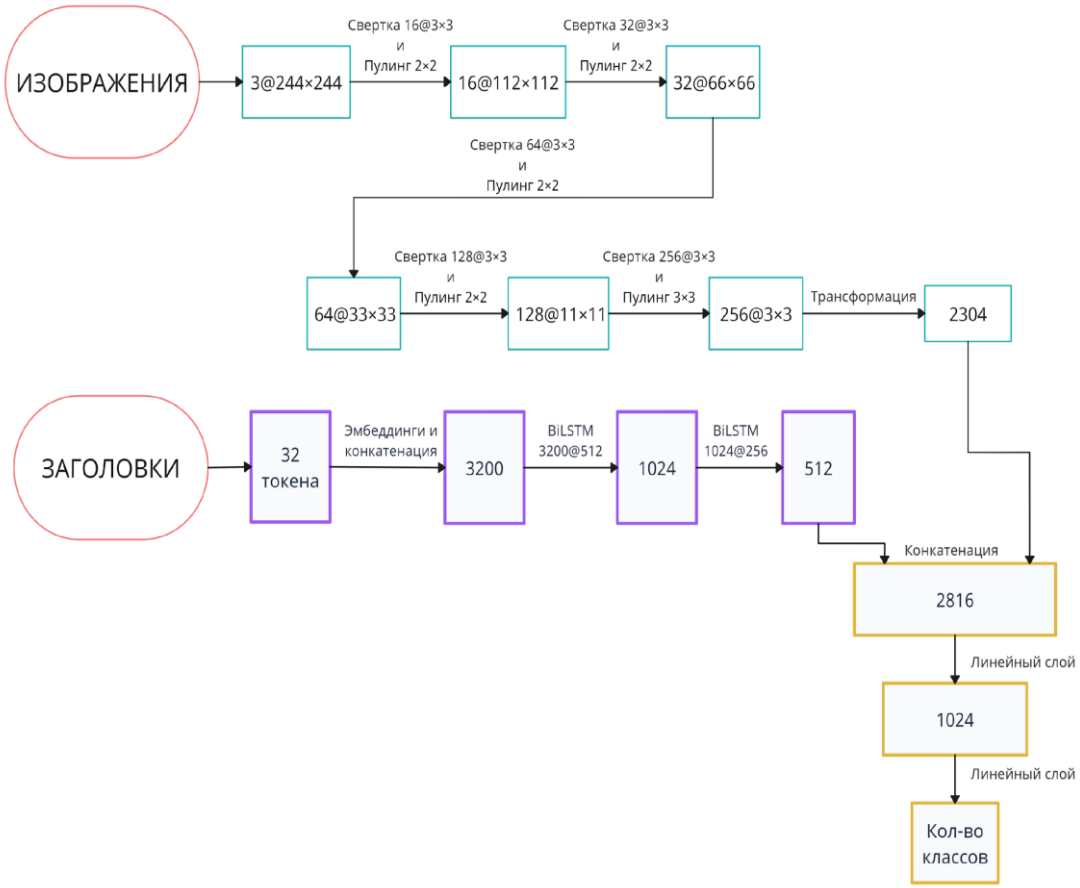

Рис.1. Пример комбинированной модели кодировщика-декодировщика для классификации объектов

Для получения векторного представления объектов будет обучаться модель кодировщика-декодировщика [2] как для текстовых, так и для графических данных. После обучения текущей модели можно использовать часть кодировщика для генерации векторных

представлений, а часть декодировщика для объединения товаров в классы или же кластеризации новостей.

Для заголовков и изображений товаров будем обучать собственные кодировщики-декодировщики. В качестве кодировщиков для текстовых данных будем использовать два двунаправленных слоя LSTM, а для графических – предобученные ResNet и VGG, а также стэк из нескольких слоев пулинга и свертки.

Помимо этого, возможно комбинировать уже обученные кодировщики различных моделей, к примеру кодировщики графических и текстовых данных. Разные кодировщики хорошо различают различные классы, а их объединение может помочь увеличить качество итоговой модели.

В общем случае мы не можем сами обучить модель кодировщика-декодировщика для решения задачи кластеризации новостных артиклей, т.к. неизвестно точное количество классов сюжетов новостей. В общем случае количество таких сюжетов является бесконечным. В таком случае мы можем получить только векторное представление документа, которое в дальнейшем мы будем использовать в алгоритмах кластеризации. В работе использовались предобученные на корпусе русскоязычных текстов модели архитектуры BERT [3]. Помимо этого, по всему набору данных новостей высчитывались TF-IDF вектора. Данное векторное представление будет считаться базовым, т.к. не учитывает структуру документа, а только общие слова.

Для кластеризации использовались алгоритмы агломеративной кластеризации и DBSCAN с метриками косинусного и евклидового расстояний. Данные алгоритмы позволяют установить ограничение на расстояние между кластерами, а не оперировать их количеством, как, к примеру, это делает алгоритм K-средних.

АНАЛИЗ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ

По результатам работы наиболее высокие результаты классификации товаров показала модель, которая комбинируют кодировщики для текстовой и графической информации. Полученный результат метрики F1 составляет 97,3%. Точность предсказаний класса составляет 97,1%. Такие высокие значения говорят об успешном обучении кодировщика-декодировщика.

Качество кластеризации новостей будем оценивать с помощью индекса Фоулкса-Мэллова. Наибольшее значение индекса равное 0,82 было получено при использовании модели MPNET-Base-v2, обученной на мультиязычном наборе данных. Использовалась агломеративная

кластеризация полной связности с метрикой косинусного расстояния. Такое значение индекса говорит об успешном решении задачи.

ЗАКЛЮЧЕНИЕ

В рамках работы была поставлена задача агрегации тематически схожих электронных ресурсов. Были разработаны и проанализированы алгоритмы сбора данных, алгоритмы извлечения признаков из полученных данных, а также алгоритмы классификации и кластеризации объектов. Метрики работы разработанных алгоритмов высоки, что свидетельствует об успехе в решении поставленной задачи.

В результате исследования была получена программа, которую можно использовать для агрегации новостных артиклей и специализированных товаров, создания собственных агрегаторов новостей, а также для поддержки онлайн-каталогов товаров.

В дальнейшем планируется обучить более оптимальные модели для получения векторного представления новостных артиклей. Помимо этого, планируется исследовать возможность создания алгоритмов для агрегации мультиязычных электронных ресурсов. Архитектуры алгоритмов, которые были представлены в работе, пока не позволяют сделать это.

Библиографические ссылки

1. *Ryan M.* Web Scraping with Python: Collecting More Data from the Modern Web / M. Ryan – Изд. 4-е. – М. : O‘Reilly Media, 2018. 308 с
2. An Exploration of Encoder-Decoder Approaches to Multi-Label Classification for Legal and Biomedical Text [Электронный ресурс] / Yova Kementchedjhieva, Ilias Chalkidis // arXiv - 2023 – Режим доступа: https://arxiv.org/abs/2305.05627 – Дата доступа: 16.03.2023
3. *Kenton J. D. M. W. C., Toutanova L. K.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding //Proceedings of NAACL-HLT. 2019. С. 4171-4186.
4. *Song K.* et al. MPNet: masked and permuted pre-training for language understanding // Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020. С. 16857-16867.