Разработка алгоритмов аннотирования информации в социальных сетях

И.Д. Смирнов[1], И.А. Рыцарев[1,2], А.В. Куприянов[1,2], Д.В. Кирш[1,2]

[1]Самарский национальный исследовательский университет им. академика С.П. Королева, Московское шоссе 34А, Самара, Россия, 443086
[2]Институт систем обработки изображений РАН - филиал ФНИЦ «Кристаллография и фотоника» РАН, Молодогвардейская 151, Самара, Россия, 443001

Аннотация Работа посвящена исследованию алгоритмов аннотирования информации в социальных сетях. В качестве объекта исследования была выбрана социальная сеть Instagram. Для решения проблемы получения необходимой информации были проведены исследования в области сбора данных для социальной сети Instagram. Разработаны программные средства, обеспечивающие сбор необходимых данных и аннотирование информации. Существующие алгоритмы аннотирования информации были исследованы и применены на изображениях, как основных данных хранимых социальной сетью Instagram.

1. Введение

На данный момент времени одними из самых перспективных данных являются данные социальных сетей. Ежедневно пользователи социальных сетей генерируют миллиарды данных, на основе которых проводятся различные исследования в области экономики, маркетинга и других областях науки [1]. Но для того, чтобы проводить эти исследования необходимо получить краткую аннотацию или описание этих данных. Поэтому подобные задачи являются очень актуальными в наше время [1, 2].

Методы машинного обучения являются наиболее эффективными для решения подобных задач, позволяющие анализировать информацию. Существует множество алгоритмов для решения, данных задач: распознавание объектов, сцен, образов и т.д. [3]. Но самыми удобными из них являются алгоритмы, не просто позволяющие определить объект, а получить краткую аннотацию с помощью описания их на естественном языке (ЕЯ) [2]. Подобные методы позволяют эффективно производить аннотацию видео контента, изображений и текстов.

Предложенная в данной статье технология позволяет производить сбор медиаконтента из социальных сетей, а также производить его краткую аннотацию на ЕЯ.

2. Сбор и получение доступа к данным в социальной сети.

В данной работе в качестве объекты исследования выбраны изображения из социальной сети Instagram. Данный выбор был обусловлен следующими причинами:

1. Основой любого поста является изображение или видео.
2. Является одной из самых популярный социальных сетей в мире, причем её популярность не перестает расти из-за преобладания визуальной информации.

Но также возникает проблема по сбору информации. Социальная сеть Instagram не предоставляет открытый программный интерфейс (Application Programming Interface - API), в отличии от Facebook, Twitter, Google+ и др [1]. Этот факт усложняет задачу по сбору данных из-за отсутствия автоматизации. Чтобы решить данную проблему необходимо разработать модуль не только получающий данные из социальной сети Instagram, но и модуль, который будет получать доступ к этим данным.

Программный комплекс для аннотирования данных будет состоять из двух основных блоков, представленных на рисунке 1. В1 – включает в себя модули для авторизации, доступа и хранения данных. В2 – включает в себя модули для аннотирования изображений, как основного носителя информации в социальной сети Instagram.

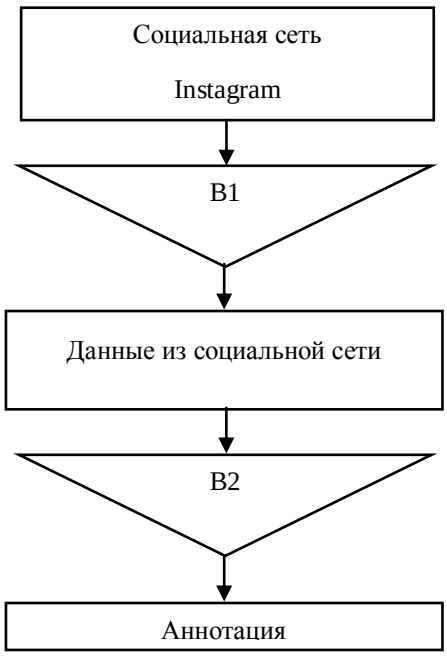

Рисунок 1. Схема аннотирования данных в социальной сети Instagram.

Данные в социальной сети Instagram хранят в себе большое количество информации: текст, изображения, видео, геометки, количество лайков, комментарии, дата выхода поста и т.п. Все посты можно получить в виде JSON объектов, основная структура которого предоставлена в таблице 1.

Таблица 1. Структура JSON – объекта.

Имя	Значение
“graphql”.“shortcode_media”.“id”	Уникальный id поста
“graphql”.“shortcode_media”.“display_url”	Ссылка на изображение
“graphql”.“shortcode_media”.“accessibility_caption”	Текстовая аннотация изображения от Instagra,
“graphql”.“shortcode_media”.“is_video”	Метка определения типа контента
“graphql”.“shortcode_media”.“location”	Место публикации поста

Для решения проблемы сбора информации в социальной сети Instagram был разработан программный модуль на языке программирования Python, который работает непосредственно с HTML-кодом [4]. В данный программный модуль входит: авторизация, поиск данных, преобразование постов в JSON-объекты [5], получение JSON-объектов.

В качестве СУБД для хранения данных было принято решение использовать PostgeSQL, так как данная СУБД поддерживает прямое хранение JSON-объектов, что позволяет осуществлять

быстрый доступ данным хранящемся внутри JSON-объектов, не извлекая их из базы данных [6].

С помощью данного программного модуля было получено более 50 000 постов Instagram для дальнейших исследований.

3. Аннотирование данных социальной сети

Для разработки программного модуля для аннотирования данных социальной сети Instagram будем использовать язык программирования Python.

В качестве инструмента для аннотирования данных, а именно для аннотирования изображений, как главного носителя информации в социальной сети Instagram было принято решение использовать модель Show and Tell [7].

Модель Show and Tell это глубокая нейронная сеть, способная описывать содержимое картинок на ЕЯ (естественном языке) [2, 7]. Визуализация основных частей модели представлена на рисунке 2.

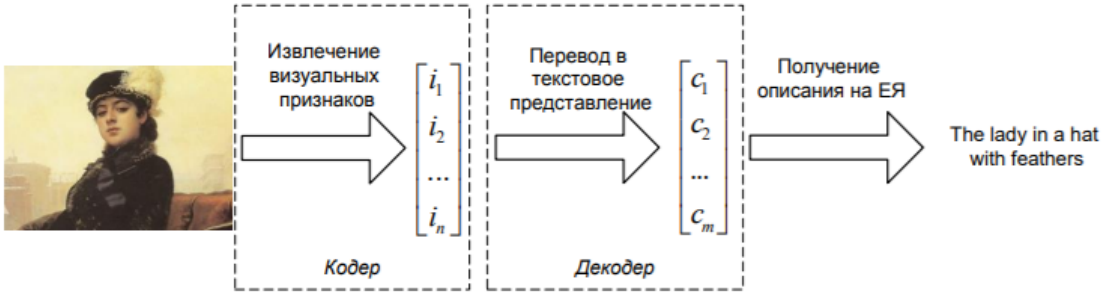

Рисунок 2. Основные части модели.

Кодировщик изображений представляет собой глубокую сверточную нейронную сеть (convolutional neural network – CNN). Этот тип сети широко используется для решения задач распознавания и обнаружение объектов на изображении [8].

Декодер представляет собой сеть с длительной кратковременной памятью (Long short-term memory – LSTM). Этот тип сети обычно используется для задач последовательного моделирования, таких как моделирование языка и машины перевод. В данной модели, сеть LSTM натренирована как языковая модель, которая обусловлена кодировкой изображения. Архитектура данной модели представлена на рисунке 3.

На рисунке 3, показана диаграмма, где $\{s_0, s_1, ..., s_{N-1}\}$ являются словами заголовка, а $\{W_e s_0, W_e s_1, ..., W_e s_{N-1}\}$ - их соответствующими векторами вложения слов. Выходные данные $\{p_1, p_2, ..., p_n\}$ LSTM блоков - это распределения вероятностей, сгенерированные моделью для следующего слова в предложении [7]. Термины $\{\log p_1(s_1), \log p_2(s_2), ..., \log p_N(s_N)\}$ являются логарифмическими правдоподобиями правильного слова на каждом шаге [7].

В качестве кодера авторы статьи используют модель опознавания изображения Inception v3 [7, 9], которая является устаревшей на данный момент. Поэтому нами было принято решение в качестве кодера использовать более новую модель Inception ResNet v2, описание которой приведено в статье [10], которая способна повысить точность аннотирования изображений. Для обучения модели Show and Tell, был использован датасет MSCOCO.

После обучения, программный модуль использующий модель Show and Tell, был включён в программный комплекс для аннотирования данных социальной сети Instagram.

4. Результаты работы

Результатами работы программного комплекса стали текстовые аннотации изображений содержащихся в постах пользователей на ЕЯ. Примеры изображений для аннотации представлен на рисунке 4, 5.

Для изображения, расположенного на рисунке 4 было получено 3 ответа с наибольшей вероятностью правильности: “a red and blue train traveling down train tracks”, “a red and black

train traveling down train tracks", "a red and yellow train traveling down train tracks". Аннотация данная социальной сетью Instagram этому же является: "train and on the street".

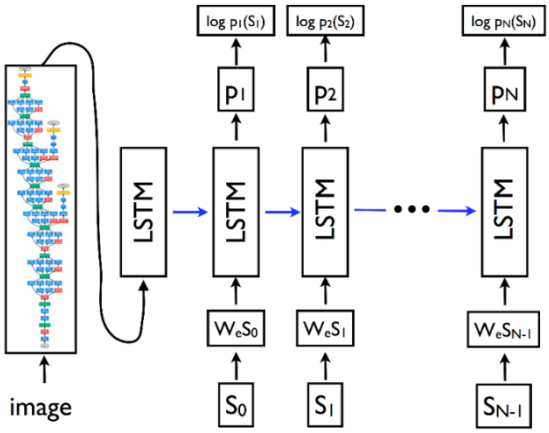

Рисунок 3. Архитектура модели Show and Tell.

Рисунок 4. Пример изображения 1.

Рисунок 5. Пример изображения 2.

Для изображения, расположенного на рисунке 5 также было получено 3 ответа с наибольшей вероятностью правильности: "a woman sitting in the snow", "a woman sitting in a blue jacket", " a woman sitting in the winter". Аннотация данная социальной сетью Instagram этому же является: " one or more people on the street and nature".

Входе эксперимента с помощью реализованного программного комплекса были собраны и проаннотированы изображения. Мы получили более детальную и явную аннотацию по сравнению с аннотацией социальной сети Instagram.

5. Заключение

В результате исследований был разработан программный комплекс для сбора и аннотирования данных в социальных сетях, основанный на алгоритмах Show and Tell и Inception ResNet v2. Разработанный программный комплекс показал многообещающие результаты и позволил получить текстовое описание изображений, содержащихся в постах. Результаты позволили произвести сравнение полученных аннотаций с аннотациями, которые производит социальная сеть Instagram, в результате мы получили более полные и качественные описания изображений на ЕЯ, это позволит нам проводить дальнейшие исследования с более детальным анализом медиаконтента в данной социальной сети, а также применить данный алгоритм в других социальных сетях.

6. Благодарности

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (№ 18-37-00418, № 19-29-01135, № 19-31-90160) и Министерства науки и

высшего образования Российской Федерации в рамках выполнения государственного задания Самарского университета и ФНИЦ «Кристаллография и фотоника» РАН.

7. Литература

[1] Рыцарев, И.А. Кластеризация медиа-контента из социальных сетей с использованием технологии BigData / И.А. Рыцарев, Д.В. Кирш, А.В. Куприянов // Компьютерная оптика. – 2018. – Т. 42, № 5. – С. 921-927. DOI: 10.18287/2412-6179-2018-42-5- 921-927.
[2] Коршунова К.П. Задачи и методы автоматического описания изображений // Системы управления, связи и безопасности. – 2018. – № 1.
[3] Schulze, M. Social Network Analysis / M. Schulze, F. Ries //Palgrave Handbook of Inter-Organizational Relations in World Politics – Palgrave Macmillan, London, 2017. – P. 113-134.
[4] Lawson, R. Web scraping with Python – Packt Publishing Ltd, 2015.
[5] Pezoa, F. Foundations of JSON schema // Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2016. – P. 263-273.
[6] Obe, R.O. PostgreSQL: Up and Running: a Practical Guide to the Advanced Open Source Database / R.O. Obe, L.S. Hsu – O'Reilly Media, Inc., 2017.
[7] Vinyals, O. Show and tell: A neural image caption generator // Proceedings of the IEEE conference on computer vision and pattern recognition, 2015. – P. 3156-3164.
[8] Сикорский, О.С. Обзор свёрточных нейронных сетей для задачи классификации изображений // Новые информационные технологии в автоматизированных системах. – 2017. – № 20.
[9] Szegedy, C. Rethinking the inception architecture for computer vision // Proceedings of the IEEE conference on computer vision and pattern recognition, 2016. – P. 2818-2826.
[10] Szegedy, C. Inception-v4, inception-resnet and the impact of residual connections on learning // Thirty-First AAAI Conference on Artificial Intelligence, 2017.

Development of algorithms for annotating information in social networks

I.D. Smirnov[1], I.A. Rytsarev[1,2], A.V. Kupriyanov[1,2], D.V. Kirsh[1,2]

[1]Samara National Research University, Moskovskoe Shosse 34A, Samara, Russia, 443086
[2]Image Processing Systems Institute of RAS - Branch of the FSRC "Crystallography and Photonics" RAS, Molodogvardejskaya street 151, Samara, Russia, 4430011

Abstract. The work is devoted to the research of algorithms for annotating information in social networks. As the object of research, the social network Instagram was selected. To solve the problem of obtaining the necessary information, studies in the field the data collection of the social network Instagram were carried out. A software tools that provides the collection of necessary data and the annotates information have been developed. The existing algorithms for annotating information have been investigated and mined on images as basic data from an Instagram social network.