ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
(СПбГУ)

Направление: Прикладные математика и физика

Полуавтоматическое структурирование изображений в социальной сети с помощью методов машинного обучения

Санкт - Петербург
2019

Содержание

Введение

С каждым днем пользователи социальных сетей создают и потребляют все бóльшие и бóльшие объемы информации, в том числе огромное количество фотографий. Построение системы для быстрой и точной навигации в миллионах изображений — не тривиальная задача. Одно из самых распространенных решений этой проблемы заключается в использовании тегов. Частный случай такого подхода — использование хештегов в социальных сетях.

Хештег — это любое слово или фраза без пробелов, перед которой стоит символ #, который называется *диез* или *решетка*, а в англоязычном варианте – *hash*, отсюда и название. Приведем несколько примеров: #*masterwork*, #*spbu*, #htaglovesport. Обычно в браузере или приложении хештег отображается как гипертекст, кликнув по которому можно получить список публикаций, снабженных таким же тегом.

Кроме простоты и удобства теги обладают еще одним полезным свойством – они позволяют не думать об иерархии структурируемой информации. Например, набор изображений можно разложить по папкам, создав иерархию по датам, геолокациям или авторству. Причем в отдельных случаях подобрать наиболее подходящую иерархию бывает затруднительно. Проблему можно решить так: достаточно поставить несколько тегов для всех изображений, а сами они могут храниться в плоской системе файлов. Благодаря этому свойству тегирование используются для рубрикации контента не только только онлайн, но и в оффлайн приложениях, например, просмоторщиках фотографий.

Можно выделить две разновидности популярных хэштегов в социальных сетях. Первые, используемые недолго и посвященные каким-то социальным явлениям или событиям, например: #*elections2018*, #*metoo*. И вторые, широкораспространенные, но не связанные с новостной повесткой, например: #*sport*, #*cafe*; они и будут нас интересовать. Данная работа посвящена разработке интеллектуальной системы, подсказывающей пользователю релевантные хештеги к загружаемым фотографиям. Кроме того, с помощью такой системы можно решать и “обратную”задачу – определять, уместно ли поставлены те или иные теги к заданным изображениям. Способность системы давать ответ на такой вопрос можно использовать для выявления злоупотреблений со стороны пользователей. Например, зачастую в рекламных целях продвигаемые публикации снабжают множеством популярных тегов, не имеющих никакого отношения к публикуемой информации.

1 Постановка задачи

Целью настоящей работы является построение интеллектуальной системы для структурирования изображений в социальных сетях. А именно, предлагается автоматизировать процесс добавления пользовательских хештегов к загружаемым изображениям. С точки зрения машинного обучения решается задача классификации изображений.

В настоящее время наилучшие результаты в задаче классификации изображений удаётся получить, используя глубокие сверточные нейронные сетей. В качестве примера можно привести один из самых известных и больших конкурсов по классификации изображений датасета *ImageNet*[1], который проводится ежегодно с 2010 года. В настоящее время решения всех призеров так или иначе базируются на сверточных нейронных сетях. TODO links

Идея применить машинное обучение к тегированию изображений в социальных сетях не нова, можно привести в пример исследования s1, s2 и s3 TODO links.

В данной работе предлагается способ улучшить точность предсказания популярных тегов, сузив информационный домен, к которому эти теги относятся. Больша́я часть из наиболее употребимых тегов в социальных сетях описывает эмоции, чувства или другие абстрактные понятия, не имеющие прямого выражения в объектах реального мира. Например, в 2018 году одними из самых популярных хэштегов в сети *Instagram* стали *#love*, *#happy*, *#beautiful*. Понятно, что модели компьютерного зрения наоборот будут точнее работать для тегов, связанных с наличием в кадре тех или иных сущностей, например *#beach*, *#sky* или *#arhcitecture*. Для таких случаев и будет строиться модель, описанная в настоящей работе.

В качестве датасета, который может быть использован для наших целей и разметка которого напрямую связана с объектами, находящимися в кадре, был выбран *Scene Understanding Dataset (SUN)*. Это набор изображений для каждого из которых выбрана одна из четырехсот локаций (сцен), вот несколько примеров:

- *baseball field*
- *basketball court*
- *ice shelf*
- *forest*
- *wind farm*

Большинство названий локаций сами по себе не являются популярными тегами из социальных сетей. Поэтому необходимо сопоставить их широко распространенным хэштегам (если это возможно):

- *baseball field, basketball court* $\rightarrow$ *#sport*

- *ice shelf, forest* $\rightarrow$ *#nature*
- *wind farm* $\rightarrow$ *?*

Выполнив сопоставление, можно присутпить к написанию и обучению сверточной нейронной сети. В итоге будет получена модель, которая по окружению, обнаруженном на пользовательском фото, сможет подсказывать подходящий хэштег.

Ясно, что полученная модель будет корректно работать лишь для ограниченного (пусть и большого) домена фотографий. Следовательно, необходимо обучить её распознавать не входящие в этот домен изображения и не пытаться определить их категорию. Кроме того, в случае низкой уверенности в правильности предсказания так же лучше ничего не делать. По мнению автора, гораздо предпочтительнее не предложить пользователю подходящий хэштег, чем многократно предлагать нерелевантные варианты.

Наконец, точность предсказаний обученной модели будет проверена вручную группой пользователей на выборке реальных фотографий из социальных сетей.

Научная новизна работы определяется:

- Адаптацией датасета *SUN* для решения задачи о структурировании изображений в социальной сети;
- Исследованием новых схем обучения неронных сетей, предложенных автором.

2 Используемые инструменты

Программные средства

В качестве языка программирования использовался *Python 3.6.7* (сборка *Anaconda*), в качестве среды разработки — *PyCharm Professional 2018.1*, операционная система *Ubuntu 16.04.4 LT*. Использованы следующие сторонние библиотеки для *python*:

- *pytorch, torchvision* — построение и обучение нейронных сетей
- *tensorboardX* — визуальное логирование процесса обучения
- *PIL, opencv, scikit-image, scipy* — обработка изображений
- *matplotlib* — отрисовка графиков
- *numpy* — матричные вычисления
- *pandas* — работа с таблицами
- *nltk* — работа с текстом, в том числе с базой *WordNet*
- *scikit-learn* — библиотека машинного обучения общего плана
- *pip* — пакетный мененджер

Вычислительные мощности

Обучение моделей производилось на удаленном сервере со следующей конфигурацией:

- видеокарта *GEFORCE GTX 1080 Ti*, (*11 ГБ* видеопамяти)
- процессор *AMD Ryzen Threadripper 1920X 12-Core*
- оперативная память объемом *62 ГБ*

3 Сверточные нейронные сети

3.1 Введение в теорею сверточных нейросетей

3.2 Обзор используемых архитектур

4 Работа с данными

4.1 База семантических связей WordNet

Обсуждение подготовки данных наиболее логично начать с описания *WordNet'a*, который использовался и авторами датасета *SUN*, и автором настоящей работы. Составители датасета *SUN* использовали *WordNet* для отыскания иерархии названий сцен. А в настоящей работе он используется для обобщения названий сцен и для поиска синонимов к предлагаемым пользователю тегам.

...Описание... гиперонимы синсеты

4.2 Датасет SUN

Набор данных *SUN (Scenes Understunding Dataset)* был впервые представлен исследовательскому сообществу в 2010 году на конференции CVPR, посвященной компьютерному зрению. Одновременно авторы опубликовали статью [2], в которой приводят различные статистики по датасету; описывают процесс сбора и разметки данных; применяют к задаче распознавания сцен лучшией из имеющихся на тот момент методов.

4.3 Адоптация датасета SUN

Как было сказано во введении, названия локаций (сцен) из датасета *SUN* не являются сами по себе популярными тегами из социальных сетей. Поэтому прежде всего необходимо выполнить сопоставление. Условно можно разбить процедуру сопоставления на 2 части: объединение исходных классов датасета *SUN* в семантические домены и сопоставление полученных доменов с популярными хэштегами. В качестве источника хэштегов была выбрана социальная сеть *Instagram* [1], ориентированная на обмен фото и видео контентом между пользователями.

Итак, предварительно необходимо очистить названия классов *SUN* от служебных слов и символов, таких как *indoor, outdoor, exterier, interier*, знаков "/"и однобуквенных алфавитных указателей. Затем для полученных слов или словосочетаний подбирается соответсвующий синсет из базы знаний *Wordnet*. Далее для синсетов находились гиперонимы, которые либо уже были достаточно абстрактны, чтобы представлять собой часто встречающийся тег, либо автор работы находил для синсетов обобщающее понятие вручную. Несколько примеров приведено в таблице 1.

Из таблицы 1 видно, что некоторые синсеты имеют общие гиперонимы. Кроме того, некоторые гиперонимы без каких-либо дополнительных изменений могли быть использованы пользователями в качестве тегов. Таким образом, использование гиперонимов позволило немного уменьшить количество ручной работы. Так же в таблице 1 приведен пример,

[1] www.instagram.com

№	Исходное название	Синсет	Гипероним	Хэштег
1	/s/shoe_shop	shoe shop	shop	#*shopping*
2	/t/toyshop	toyshop	shop	#*shopping*
3	/v/volleyball_court/indoor	volleyball court	court	#*sport*
4	/w/wrestling_ring/indoor	wrestling ring	ring	#*sport*
5	/r/rainforest	rain forest	forest	#*forest*
6	/p/pantry	pantry	storeroom	?

Таблица 1: Сопоставление искомых классов и хэштегов.

когда для локации сложно подобрать какой-то подходящий и широко распространенный тег. В итоге использовалось около половины из *397* искомых классов датасета *SUN*, каждому из которых удалось поставить в соответсвие один из *20* популярных хэштегов. В данной работе популярными считаются теги, использованные в сети *Instagram* более *10 млн.* раз. При этом среднее число упоминаний отобранных тегов составило *100 млн.* раз, а максимальное — *450 млн.*. Полная информация о частоте встречаемости для всех хэштегов приведена на рисунке 1.

Отметим, что автор предпринял несколько попыток произвести процедуру адаптации разметки датасета *SUN* полностью автоматически, но они оказались неудачными.

Первая попытка — обобщить искомые классы, используя метод *topic_domains*, который доступен для синсетов в *python* реализации API к базе *WordNet*. Например, для синсета *basketball_court.n.01*, который определяется как *the court on which basketball is played*, вызов данного метода возвращает *basketball.n.02*, который определяется так: *a game played on a court by two opposing teams of 5 players; points are scored by throwing the ball through an elevated horizontal hoop*. К сожалению, проблема заключалась в том, что для подавляющего числа названий локаций *topic_domains* возвращал пустое значение, т.е. для этих синсетов авторами базы знаний не было назначено доменов.

Вторая попытка аналогичная первой, но использовалась сторонняя база знаний *WordNet Domains* [2]. К сожалению, такое расширение базы доменов не позволило решить проблему, описанную выше.

Третья идея заключась в использовании информации о семантической близости синсетов, в частности, API *WordNet'a* позволяет для любых двух синсетов вычислить степень похожести несколькими способами: *jcn_similarity, lch_similarity, res_similarity, wup_similarity*. Для разных видов измерения расстояния были вычислены матрицы попарных дистанций, на основе которых производилась иерархическая кластеризация с различными гиперпараметрами (например, кластеризация по заданному максимальному рас-

[2] http://wndomains.fbk.eu/

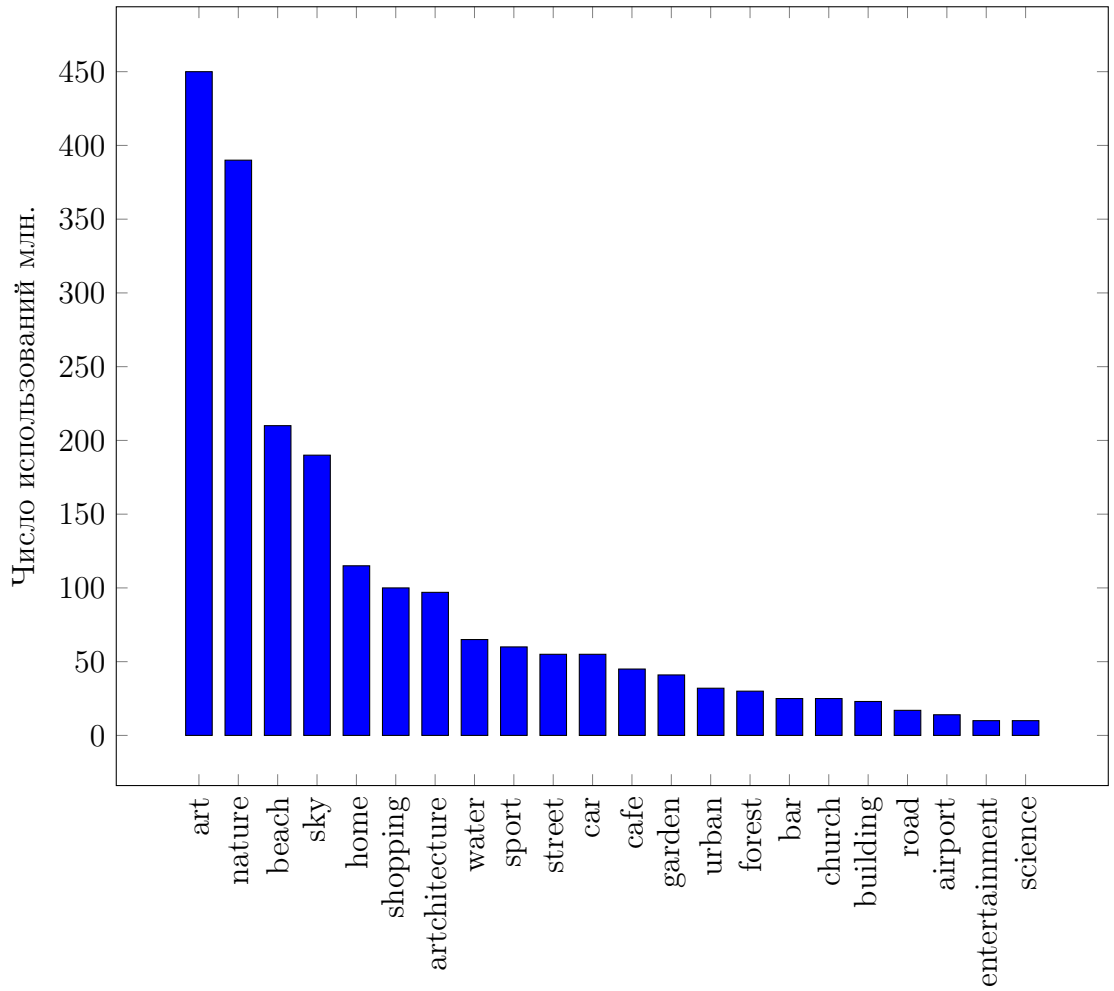

Рис. 1: Встречаемость предсказываемых хэштегов в социальной сети *Instagram*.

стоянию в кластере, по заданному количеству кластеров). К сожалению, автору не удалось получить кластеры, большую часть которых можно было бы без труда "озаглавить"каким-либо популярным хэштегом.

Таким образом, процедуру сопоставления всех *397* искомых классов датасета *SUN* и хэштегов из социальной сети *Instagram* пришлось произвести в полуручном режиме (опираясь только на гиперонимы), как это было описано выше в данном разделе.

Приведем пример такого сопастовления. В таблице 2 перечислены названия локаций, которым поставлен в соответсвие хэштег $\#sport$:

/w/wrestling_ring/indoor	/a/athletic_field/outdoor	/b/badminton_court/indoor
/b/ball_pit	/b/baseball_field	/b/basketball_court/outdoor
/b/boxing_ring	/b/bullring	/g/golf_course
/g/gymnasium/indoor	/m/martial_arts_gym	/r/racecourse
/r/riding_arena	/s/ski_lodge	/s/ski_resort
/s/ski_slope	/s/squash_court	/s/stadium/baseball
/s/stadium/football	/s/swimming_pool/indoor	/s/swimming_pool/outdoor
/t/tennis_court/indoor	/t/tennis_court/outdoor	/v/volleyball_court/indoor
/v/volleyball_court/outdoor		

Таблица 2: Список классов датасета *SUN*, которым сопоставлен хэштег #*sport*.

5 Численные эксперименты

5.1 Архитектура программы

5.2 Особенности реализации

5.3 Полученные результаты

5.4 Обсуждение результатов

6 Валидация результатов

Выводы

Список литературы

[1] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei. [2015] *ImageNet Large Scale Visual Recognition Challenge*. IJCV.

[2] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. [2010] *SUN Database: Large-scale Scene Recognition from Abbey to Zoo*. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).