

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	3
1 Постановка задачи	4
2 Используемые инструменты	6
3 Искусственные нейронные сети	7
3.1 Перцептрон	7
3.2 Функции активации	7
3.3 Простая полносвязная сеть	8
3.4 Функция потерь	9
3.5 Метод обратного распространения ошибки	10
3.6 Сврточные нейронные сети	10
3.7 Используемые сврточные архитектуры	10
4 Работа с данными	11
4.1 База семантических связей WordNet	11
4.2 Датасет SUN	11
4.3 Адоптация датасета SUN	12
5 Численные эксперименты	16
5.1 Архитектура программы	16
5.2 Особенности реализации	17
5.3 Полученные результаты	17
5.4 Обсуждение результатов	17
6 Валидация результатов	18
Выводы	19
Список литературы	20

Введение

С каждым днем пользователи социальных сетей создают и потребляют все большие и большие 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 *диэз* 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 офлайн приложениях, например, просмоторщиках фотографий.

Можно выделить две разновидности популярных хештегов 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 сети. 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. Большая 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 *#architecture*. 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* → *#sport*

- *ice shelf, forest* → $\#\text{nature}$
- *wind farm* → ?

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 менеджер

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 Искусственные нейронные сети

Настоящая часть работы предназначена для читателя, не знакомого с искусственными нейронными сетями и глубоким обучением. Здесь будут приведены теоретические основы глубокого обучения и рассмотрены сверточные архитектуры, используемые в дальнейшей работе.

3.1 Перцептрон

Начнем с рассмотрения одиночного нейрона — перцептрана Розенблатта — базового элемента, содержащегося в большинстве современных нейросетевых архитектур. Перцептрон имеет несколько входов и один выход, значение на котором вычисляется как взвешенная сумма значений входов (рисунок 1). Кроме того, обычно к выходному значению применяется сдвиг и некоторая нелинейная функция, называющаяся функцией активации нейрона. Ее предназначение мы обсудим позже.

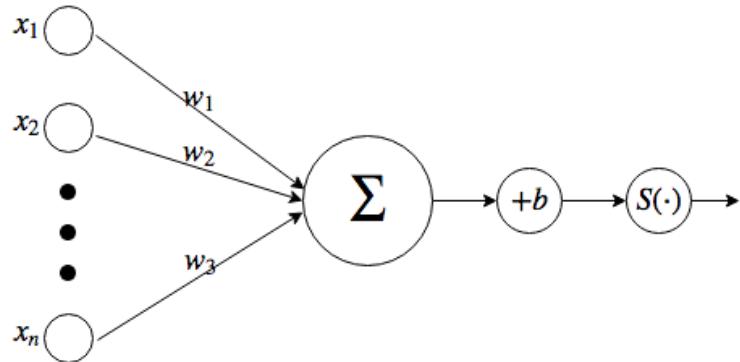


Рис. 1: Схематическое изображение работы одного отдельного нейрона.

Таким образом, значение на выходе нейрона задается выражением 1.

$$f(\vec{x}) = S\left(\sum_{i=1}^n x_i w_i + b\right) \quad (1)$$

где $f(\vec{x})$ — выходное значение нейрона, посчитанное для входов x_i , w_i — весовые коэффициенты для каждого входа, b — параметр смещения, а S — нелинейная функция активации. Далее для упрощения повествования положим $b \equiv 0$.

3.2 Функции активации

Существуют множество различных функций активации, например, гиперболический тангенс, логистическая сигмоида или $ReLU$, рисунок 2.

$$\begin{aligned}
S(x) &= th(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \quad th'(x) = 1 - th(x)^2 \\
S(x) &= \sigma(x) = \frac{1}{1 + e^{-2x}}, \quad \sigma'(x) = \sigma(x)(1 - \sigma(x)) \\
S(x) &= ReLU = \max(0, x), \quad ReLU'(x) = \theta(x)
\end{aligned} \tag{2}$$

где $\theta(x)$ – функция Хэвисайда.

Эти функции используются для добавления нелинейных зависимостей между слоями многослойной модели. Названные выше функции особенно популярны, так как значения их производных либо достаточно просты, либо легко выражаются через значения самих функций (выражение 2), что позволяет быстро вычислять значение производной.



Рис. 2: Функции активации

3.3 Полносвязная сеть

Одиночный нейрон не способен выразить сложные в наборе признаков \vec{x} , поэтому нейроны объединяют в слои, а их, в свою очередь, в многослойные сети. Рассмотрим сеть, состоящую из двух слоев нейронов. Пусть количество входных признаков равно N , количество нейронов скрытого слоя P , а размер выхода – M , рисунок 3. Такая архитектура, состоящая из простых линейных слоев, называется полносвязной.

Рассмотрев выражение 1 можно увидеть, что совокупность значений нейронов на 1 слое может быть получена простым матричным умножением входов \vec{x} на матрицу весов W^1 размера $P \times N$, с последующим поэлементным применением функции активации к получившимся значениям. Аналогично, значения нейронов 2 слоя получаются умножением предыдущих значений на весовую матрицу W^2 размером $M \times P$. Таким образом, применение нейросети ко входу \vec{x} можно задать выражением 3.

$$\vec{y} = W^2 S(W^1 \vec{x}) \tag{3}$$

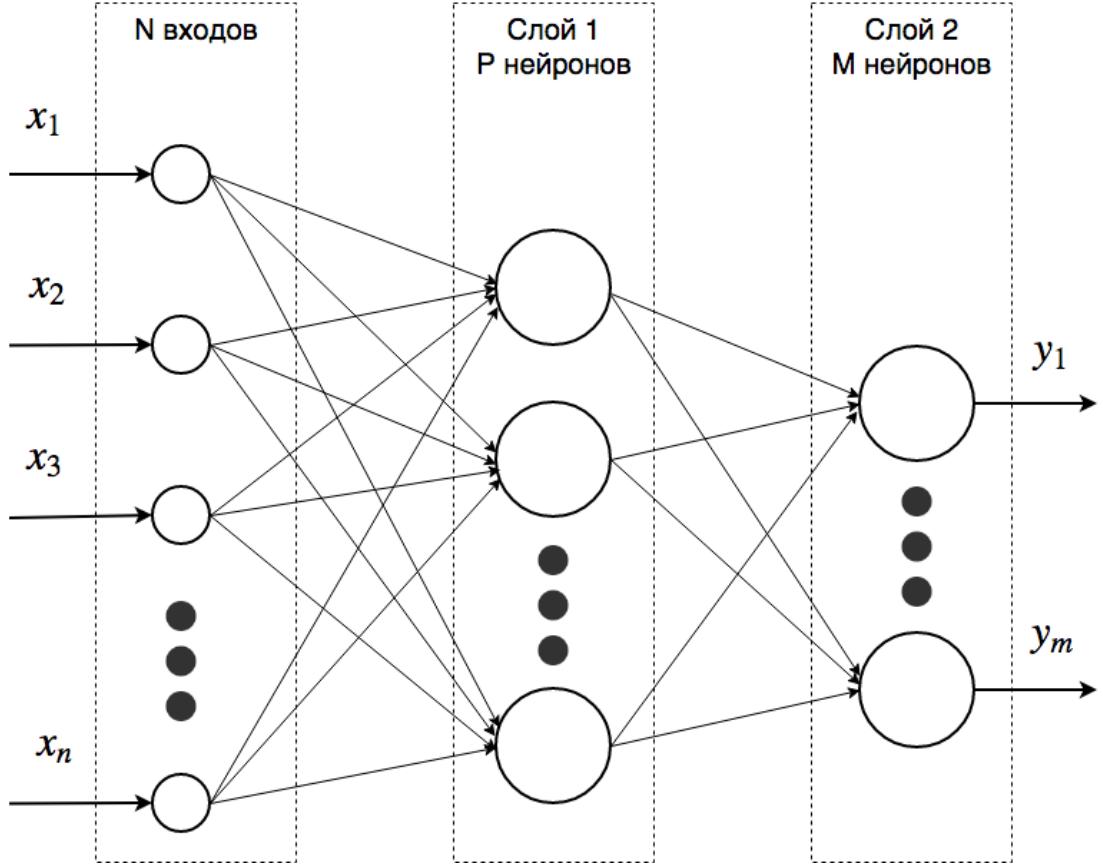


Рис. 3: Схематическое изображение полносвязной нейронной сети.

где $S(\cdot)$ – применение нелинейности к каждому элементу входного вектора.

3.4 Функция потерь

Близость предсказания сети к правильным ответам оценивается с помощью функции ошибки, так же называемой функцией потерь (*loss function*). Например, для задачи регрессии в качестве функции потерь может применяться среднеквадратичное отклонение (*MSE*, выражение 5), или, в более простом случае – средняя разность между выходами модели и правильными ответами (выражение 4); для задачи классификации обычно используют перекрестную энтропию (*cross entropy*, выражение 6).

$$L_1(\vec{y}_{gt}, \vec{y}) = \frac{1}{M} \sum_{i=1}^M |y_i^{gt} - y_i| \quad (4)$$

$$MSE(\vec{y}_{gt}, \vec{y}) = \frac{1}{M} \sum_{i=1}^M (y_i^{gt} - y_i)^2 \quad (5)$$

$$CE(\vec{y}_{gt}, \vec{y}) = - \sum_{i=1}^M y_i \log y_i^{gt} \quad (6)$$

Исходя из постановки задачи, можно составить и другие функции ошибок, но они все должны удовлетворять главному требованию — быть дифференцируемыми. Это необходимо для того, чтобы пользоваться методом обратного распространения ошибки для обучения модели.

3.5 Метод обратного распространения ошибки

Поняв, как конструируются нейронные сети, что представляют собой функции активации и функции потерь, обсудим, как происходит обучение моделей.

3.6 Сверточные нейронные сети

Сверточная нейронная сеть (*Convolution neural network*) — это специальная сеть, сконструированная для обработки изображений, хотя в настоящее время спектр их применения значительно расширился. Как следует из названия, основной таких сетей являются сверточные слои *convolutional layers*. Кроме того, обычно вместе с ними используются такие слои как *BatchNormalization*, *Pooling*, *Softmax* и *Dropout*.

Рассмотрим, как устроено применение оператора свертки к изображению.

3.7 Используемые сверточные архитектуры

В данное работе используются следующие архитектуры нейронных сетей:

- *Residual netwrotk (ResNet)* — одна из самых популярных в настоящее время архитектур, предложенная в статье [3] 2015 года. Основная идея состоит в добавлении конкатенации значений нейронов на i -м слое с $i-2$ слоем (такая процедура получила название *skip connection*). Таким образом авторы успешно решают распространенную проблему обучения глубоких сетей — затухание градиентов.
- Inception — todo
- DenseNet — todo
- VGG — относительно старая, ставшая классической, архитектура, предложенная в 20xx году todo.

4 Работа с данными

4.1 База семантических связей WordNet

Обсуждение работы с данными наиболее логично начать с описания базы знаний *WordNet'a*, который использовался и авторами датасета *SUN*, и автором настоящей работы. Составители датасета *SUN* использовали *WordNet* для создания иерархии названий сцен (локаций). А в настоящей работе он используется для объединения названий локаций в обобщающий домен и для поиска синонимов к предлагаемым пользователю тегам.

WordNet — это электронный словарь/семантическая сеть для английского языка. Он содержит 4 подсети: для глаголов, существительных, прилагательных и наречий. Узлами сети являются не отдельные слова, а синсеты (*synset*), объединяющие слова со схожим значением. Таким образом, слова, имеющие несколько значений могут быть включены в несколько синсетов.

Синсеты связаны между собой различными отношениями. Например, один синсет может выступать по отношению к другому в роли гиперонима, гипонима, меронима, антонима и т.д. todo

Кроме того, между синсетами можно вычислять различные меры близости различными способами:

- *Path similarity* — показывает, насколько близки пути до синсетов в общем графе *WordNet'a*.
-

4.2 Датасет SUN

Набор данных *SUN* (*Scenes Understanding Dataset*) впервые был 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.

Датасет представляет собой набор фотографий, на каждой из которых запечатлена одна из 908 локаций, примеры приведены на рисунке [?]. Причем часть локаций представлена в двух вида: снаружи и изнутри. Чтобы отличить эти два случая к названиям сцен добавляются слова *outdoor* или *exterier* и *indoor* или *interier* соответственно. Кроме того, в 2012 году авторы для части изображений представили разметку на уровне объектов: были предоставлены маски для 300 тыс. объектов, относящихся к одной из 5 тыс. категорий.

Наконец, авторы оставили только хорошо интерпретирующиеся классы сцен, содержащие хотя бы 100 примеров, после чего организовали на этом наборе данных соревнование по машинному обучению. Итоговый датасет для решения задачи классификации, который

и будет использоваться в данной работе, содержит 108754 изображений (около 40 ГБ на жестком диске), каждое из которых отнесено к одному из 397 классов.



Рис. 4: Примеры размеченных фотографий из датасета *SUN*.

Посмотреть больше изображений из датасета *SUN* можно через интерактивный веб-обозреватель¹, которым можно воспользоваться для просмотра упорядоченных изображений как по сценам, так и по объектам.

4.3 Адаптация 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*², ориентированная на обмен фото и видео контентом между пользователями.

Итак, предварительно необходимо очистить названия классов *SUN* от служебных слов

¹groups.csail.mit.edu/vision/SUN/

²www.instagram.com

и символов, таких как *indoor*, *outdoor*, *exterier*, *interier*, знаков "/" и однобуквенных алфавитных указателей. Затем для полученных слов или словосочетаний подбирается соответствующий 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.

№	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.

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, 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 соответствие 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 5.

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.

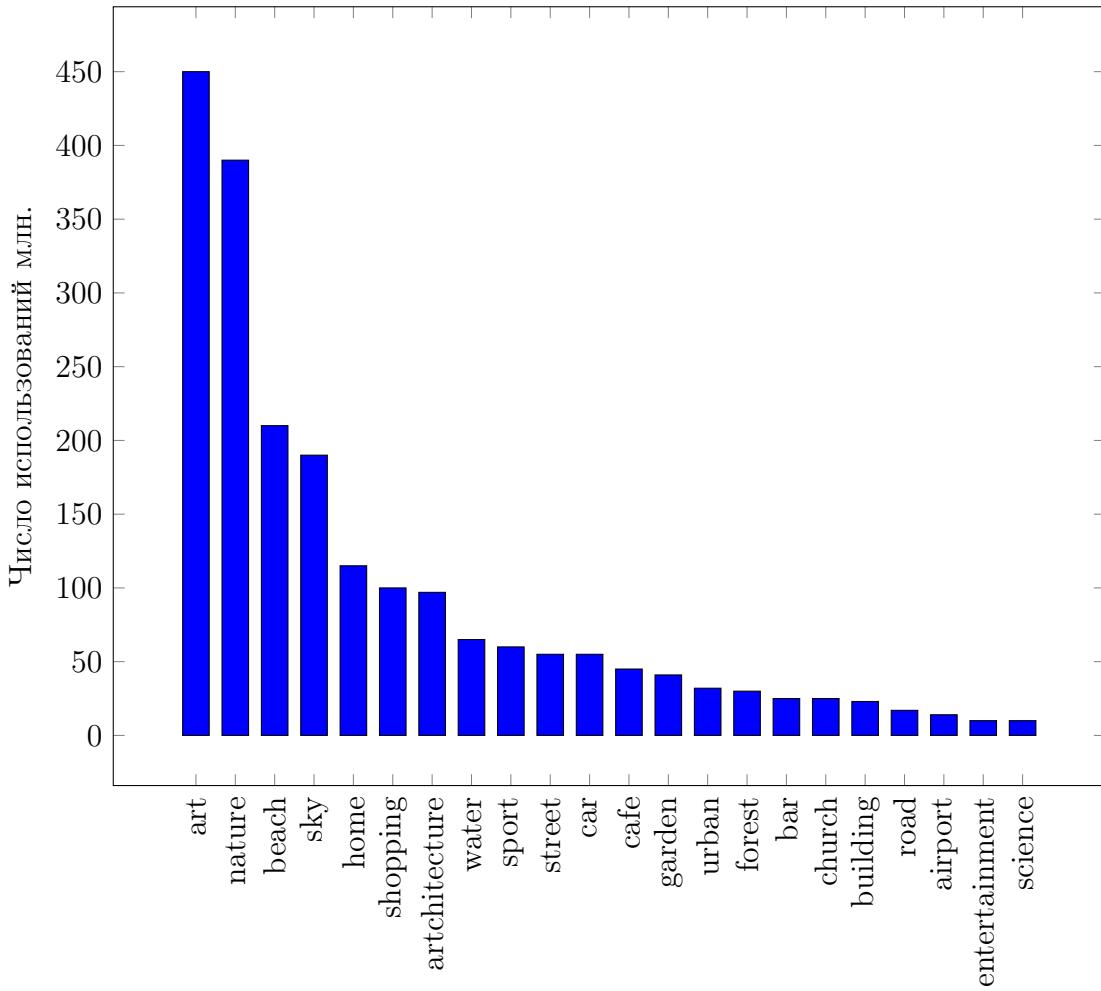


Рис. 5: Встречаемость предсказываемых хэштегов в социальной сети *Instagram*.

Вторая попытка аналогичная первой, но использовалась сторонняя база знаний *WordNet Domains*³. 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, класторизация по заданному максимальному расстоянию 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.

³<http://wndomains.fbk.eu/>

Приведем пример такого сопоставления. В таблице 2 перечислены названия локаций, которым поставлен в соответствие 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`.

Файлы с полной информацией об итоговом сопоставлении можно найти в репозитории автора ⁴.

⁴github.com/AlekseySh/scenesTODO

5 Численные эксперименты

В данной главе мы обсудим экспериментальную часть работы, начиная от архитектуры программы и заканчивая обсуждением полученных результатов.

5.1 Архитектура программы

Основная часть программы, выполняющая обучение и тестирование модели состоит из стандартного для фреймворка *pytorch* набора взаимосвязанных компонент (классов). Перечислим их:

- *Module* — описывает непосредственно вычислительный граф нейросети. Здесь указаны параметры и количество всех слоев, описаны связи между ними.
- *Dataset* — позволяет итерироваться по набору данных и объединять их в батчи для подачи на вход нейросети.
- *Loss* — вычисляет функцию ошибки/потери между предсказанными моделью и правильными значениями целевой переменной.
- *Optimizer* — совершает шаг градиентного спуска на заданное расстояние, которое определяется скоростью обучения (*learning rate*). А именно, изменяет веса модели так, чтобы уменьшить среднюю ошибку для очередного поданного на вход модели батча данных.
- *Scheduler* — изменяет скорость обучения модели (*learning rate*) с течением времени по заданному правилу.
- *Stopper* — останавливает тренировку при выполнении заданного условия, например, если в течение последних n эпох не произошло увеличения точности модели хотя бы на ϵ .
- *MetricsCalculator* — оценивает точность модели на некоторой размеченной подвыборке данных по заданным метрикам.
- *TensorboardX* — система для визуального логирования обучения модели; позволяет строить графики изменения функции ошибки, метрик и выводить любые другие пользовательские изображения.
- *Trainer* — объединяет воедино компоненты, названные выше. Обучает модель эпоху за эпохой, с заданной частотой проверяет текущую точность на тестовом подмножестве данных. Останавливает тренировку по достижению некоторого критерия. Сохраняет промежуточные веса модели. Визуализирует процесс обучения.

Взаимосвязь между компонентами программы можно проследить на рисунке 6.

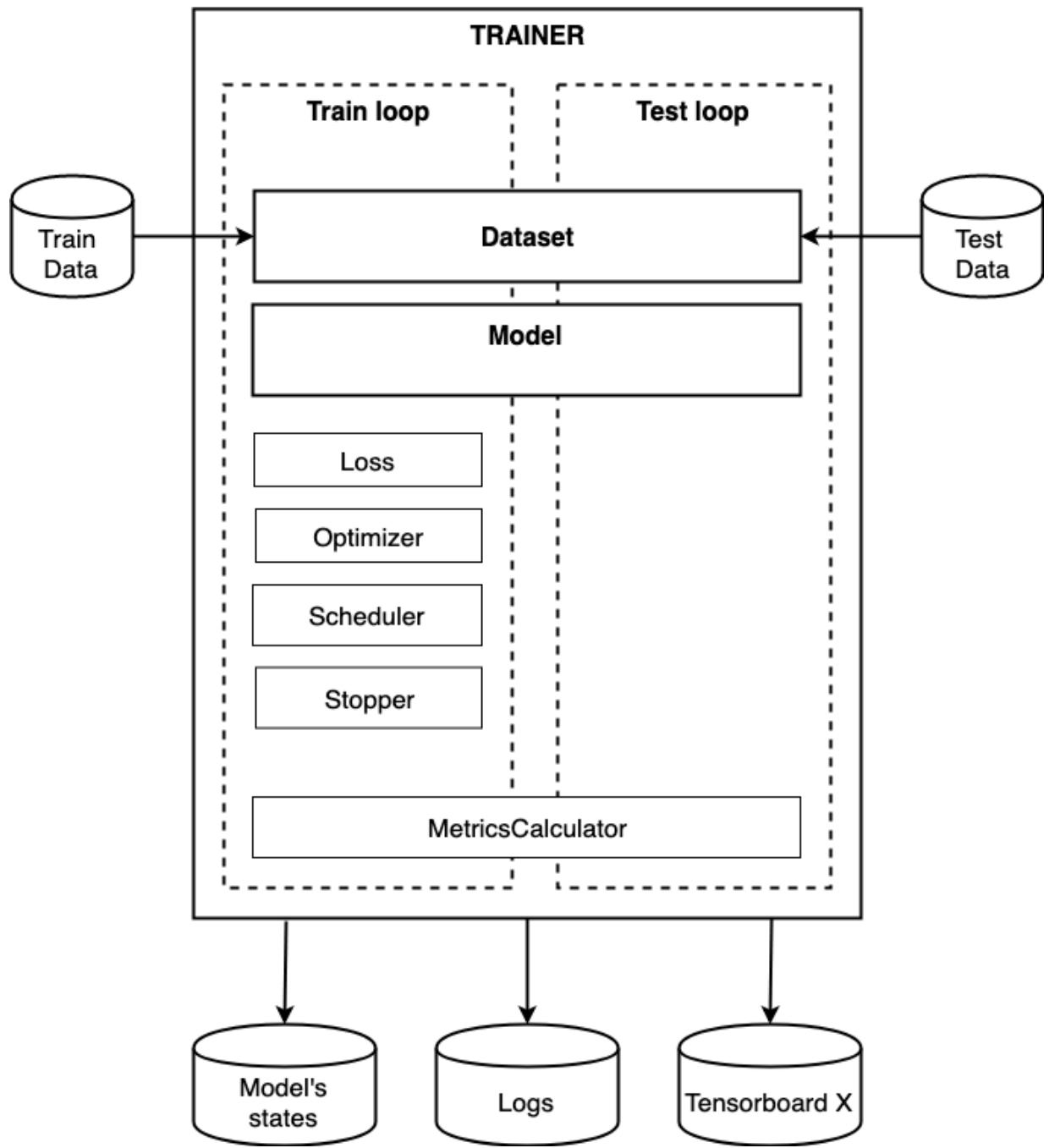


Рис. 6: Структура программы для тренировки и обучения модели.

5.2 Особенности реализации

todo TTA

5.3 Полученные результаты

Картинка, графики метрик, confusion matrix

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).
- [3] K. He, X. Zhang, S. Ren, J. Sun. [2015] *Deep Residual Learning for Image Recognition*. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).