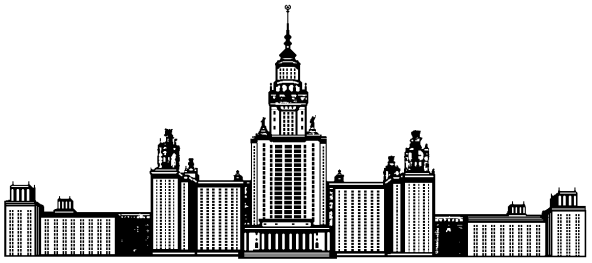


Московский государственный университет имени М.В. Ломоносова
Факультет Вычислительной математики и кибернетики
Кафедра Математических методов прогнозирования

ДИПЛОМНАЯ РАБОТА

Оценка релевантности изображения текстовому запросу на основе визуального контента

Выполнил:
студент 517 группы
Найдин Олег Павлович

Научный руководитель:
д.т.н., профессор
Местецкий Леонид Моисеевич

Москва, 2015

Содержание

Аннотация

В данной работе представлена глубокая нейронная сеть для оценки релевантности изображения текстовому запросу. Рассматриваемая архитектура состоит из трех основных блоков: сверточная нейронная сеть для извлечения дескрипторов изображений, полносвязная нейронная сеть для преобразования заранее обученного векторного описания текста и еще одна полносвязная нейронная сеть для вычисления близости полученных представлений. В рамках исследования изучены различные архитектуры и методы обучения вышеописанных блоков, векторные представления запросов, а также функции потерь. Приведены результаты экспериментов, подтверждающих возможность использования предсказанной релевантности в качестве фактора ранжирования в поисковых системах, а также оценена статистическая значимость полученных результатов.

1 Введение

С учетом быстрого роста объема мультимедийного контента в последние годы часто встает вопрос эффективного и качественного поиска изображений, музыки или видео в больших коллекциях. К сожалению, не все хранилища данных позволяют производить достаточно подробное аннотирование ресурсов для дальнейшего использования текстовых поисковых систем, к тому же ручное аннотирование требует больших затрат человеческого времени. В частности, если рассматривать системы интернет-поиска мультимедийных ресурсов, ручная разметка просто невозможна в силу колоссального количества различного контента на просторах всемирной паутины, а автоматическое получение каких либо аннотаций затруднительно для большинства ресурсов.

В ситуации отсутствия аннотации необходимо анализировать непосредственно контент ресурса, чему и посвящена данная работа. Основная сложность рассматриваемой проблемы заключается в гетерогенности исходных данных — необходимо оценивать близость объектов из множеств различной природы (текст и изображение/видеозапись/аудиозапись/...).

1.1 Основные понятия и определения

Ниже коротко перечислены основные понятия, рассматриваемые в данной работе, а также приведены ссылки на литературу, в которой можно найти их подробное описание. Более частные понятия будут вводиться по ходу изложения материала.

Мультимедийный ресурс — объект коллекции данных, в котором основная информация представлена в виде мультимедиа (текст/изображение/аудио/видео/...). Под представлением или дескриптором ресурса понимается его векторное описание в евклидовом пространстве: $r \in \mathbb{R}^m$. Процесс получения этого описания будем называть извлечением или обучением дескриптора.

Семантическая близость — мера схожести двух объектов, возможно, гетерогенных, основанная на их смысловом соответствии. Будем говорить, что

пространство обладает семантической близостью, если введенная в нем стандартная мера схожести, например, косинусное расстояние, является адекватной моделью оценки семантической близости объектов в этом пространстве.

Релевантность — семантическое соответствие поискового запроса и ресурса.

Ранжирование — сортировка ресурсов поисковой выдачи. Для сравнения различных методов сортировки вводят целевой функционал ранжирования, например, точность, полнота, DCG/nDCG[1] [20] и многие другие.

Генеративная модель — алгоритм решения задачи машинного обучения с помощью моделирования совместного распределения на множестве наблдаемых событий X и параметров модели Θ: $p(X, \Theta)$ [6].

Дискриминативная модель — алгоритм решения задачи машинного обучения с помощью моделирования условного распределения на множестве параметров модели Θ при условии наблюдаемых событий X: $p(\Theta|X)$ [6].

Аннотирование — составление краткого текстового описания чего-либо, в частности, ниже рассмотрены методы автоматического аннотирования мультимедийных ресурсов.

Асессор — человек, оценивающий результаты поисковой выдачи. Обычно выставляет каждому документу выдачи некоторую оценку по шкале релевантности.

Пользовательское поведение — информация, собираемая поисковыми системами о действиях пользователей, например, кликах на странице поисковой выдачи [45].

Нейронная сеть — математическая модель, построенная по принципу организации и функционирования сетей нервных клеток живого организма. Существует множество видов нейронных сетей, например, полносвязные[2] [36], сверточные [26] и другие. Глубина нейронной сети соответствует количеству

[1]DCG — discounted cumulative gain, nDCG — normalized DCG.

[2]Их также называют многослойными персептронами. Проекция вектора исходного пространства $x \in \mathbb{R}^n$ в результирующее пространство $\mathbb{R}^m$ осуществляется по правилу $\sigma(Wx + b)$, где $W \in \mathbb{R}^{m \times n}$ — матрица проекции, $b \in \mathbb{R}^m$ — вектор сдвига, а σ — нелинейная функция активации. Такие проекции могут следовать друг за другом, образуя последовательность преобразований вектора исходного пространства.

проекций исходных данных, глубокими часто называют сети более чем с двумя проекциями [8, 40].

Обучение пространства — построение проекции (отображения) из одного пространства в другое $f : X \to Y$. Обычно обучаемое пространство Y должно обладать некоторым свойством, например, семантической близостью.

1.2 Обзор решаемой задачи

В данной работе решается задача оптимизации функционала ранжирования для результатов поиска мультимедийных ресурсов в большой коллекции без использования аннотаций этих ресурсов. Как и в большинстве задач машинного обучения, существует два основных подхода к решению задачи такого рода: генеративный и дискриминативный.

Генеративные модели автоматически порождают описания ресурсов для дальнейшего использования текстовых поисковых систем. Такой подход нацелен на хорошее аннотирование, а не на ранжирование мультимедийных ресурсов, то есть не решает поставленную выше задачу, однако хорошее аннотирование часто способствует повышению качества ранжирования текстовых поисковых систем. Большинству генеративных моделей необходима достаточно большая аннотированная коллекция ресурсов для обучения [4, 24], получение которой невозможно или требует огромных затрат на ручную разметку. Существуют методы, не нуждающиеся во вручную аннотированных ресурсах для обучения [44], часто они основаны на анализе текстового контекста, в котором ресурс встречается в интернете, однако использование подобных систем затруднено необходимостью очистки собранной выборки от нерелевантных аннотаций.

Дискриминативные модели нацелены непосредственно на минимизацию целевого функционала ранжирования. Таким моделям также нужна обучающая выборка большого объема, однако в этом случае это должны быть не аннотации ресурсов, а триплеты вида <текстовый запрос, ресурс, степень релевантности>. Получение такой выборки с помощью асессоров требует гораздо меньших затрат по сравнению с ручным аннотированием. При таком подходе, обычно, текстовые запросы и мультимедийные ресурсы проецируются в общее пространство с некоторой мерой сходства, которая аппроксимирует семантическую близость и, как

следствие, релевантность [11, 12]. Как и в случае генеративных моделей, существуют методы, не нуждающиеся в размеченных ресурсах для обучения [48, 50], большинство из которых основано на анализе поведения пользователей интернет-поисковых систем.

Анализ пользовательского поведения может давать лучшие результаты по сравнению с асессорской разметкой за счет нескольких особенностей:

- асессор не знает, что пользователь ищет на самом деле, поэтому может ошибиться при выставлении оценки релевантности[3],
- получить информацию о поведении пользователей можно из логов поисковых систем, причем объем этой информации определяется количеством пользователей. Таким образом, крупные поисковые системы имеют возможность получать выборки достаточного объема для использования любых алгоритмов машинного обучения.

Однако фильтрация сигнала, которым является пользовательское поведение для поисковых систем, — отдельная трудоемкая задача[4].

Как показано в исследовании [33], дискриминативные модели обычно достигают лучших результатов в задачах как классификации, так и регрессии, а также им необходима обучающая выборка, получить которую с помощью асессоров гораздо проще и дешевле по сравнению с генеративным подходом. Таким образом, использование дискриминативного метода оценки релевантности мультимедийного ресурса текстовому запросу предпочтительно, поэтому в данной работе представлена именно дискриминативная модель.

Существует множество дескрипторов как текстов, так и мультимедийных ресурсов, однако задача построения гетерогенной меры схожести этих признаковых описаний нетривиальна. Рассмотрим множество текстовых запросов $T = \{t_i\}_i$ и множество ресурсов $R = \{r_j\}_j$. Обычно строится некоторое отображение $f : T \rightarrow$

[3]На этапе разметки поисковой выдачи асессор часто видит только текстовый запрос пользователя и набор ресурсов, для которых он должен выставить оценки релевантности. В связи с неумением подавляющего большинства пользователей формулировать поисковый запрос с достаточной степенью конкретности, асессор может неправильно его интерпретировать, и, как следствие, ошибочно выставить оценки релевантности.

[4]Если, например, рассматривать в качестве сигнала пользовательские клики, то возникает проблема фильтрации кликов, не коррелирующих с релевантностью: случайные клики или клики, вызванные высоким расположением ресурса в выдаче, а не его релевантностью. Попытки извлечения данных о релевантности из пользовательского поведения рассмотрены в [19, 22]

$\mathbb{R}^n$, не зависящее от пространства R, и аналогичное отображение $g : R \to \mathbb{R}^m$, не зависящее от пространства T, затем определяется некоторая мера схожести между описаниями образов объектов исходных пространств $\mu : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$. Основная проблема такого подхода заключается в независимом подборе функций f, g и μ, в то время как для качественного решения задачи построения гетерогенной функции близости необходимо обучать непосредственно суперпозицию описанных преобразований: $\tilde{\mu}(t_i, r_j) = \mu(f(t_i), g(r_j))$. В данной работе предлагается подход совместного обучения целевого функционала $\tilde{\mu}$ с помощью блочной нейронной сети.

До этого момента рассматривалась задача оценки релевантности ресурсов из множества произвольной природы текстовому запросу, далее будет рассмотрен частный случай: поиск изображений в большой коллекции на основании текстового запроса. Также будут описаны архитектура предлагаемой блочной нейронной сети и метод ее обучения. Под изображением будет пониматься RGB-изображение, преобразованное к размеру 256×256 пикселей без сохранения соотношения сторон; под релевантностью — вещественное число из отрезка $[0, 1]$, где чем оно больше, тем релевантнее изображение запросу.

1.3 Описание обучающей выборки

В качестве обучающей выборки используются данные о релевантности, полученные с помощью асессоров. База оценок состоит из 532855 записей, каждая из которых является триплетом следующего вида: <текстовый запрос, изображение, метка релевантности>. В разметке присутствует три типа меток, проставленных асессорами:

RELEVANT_PLUS изображение полностью релевантно запросу;

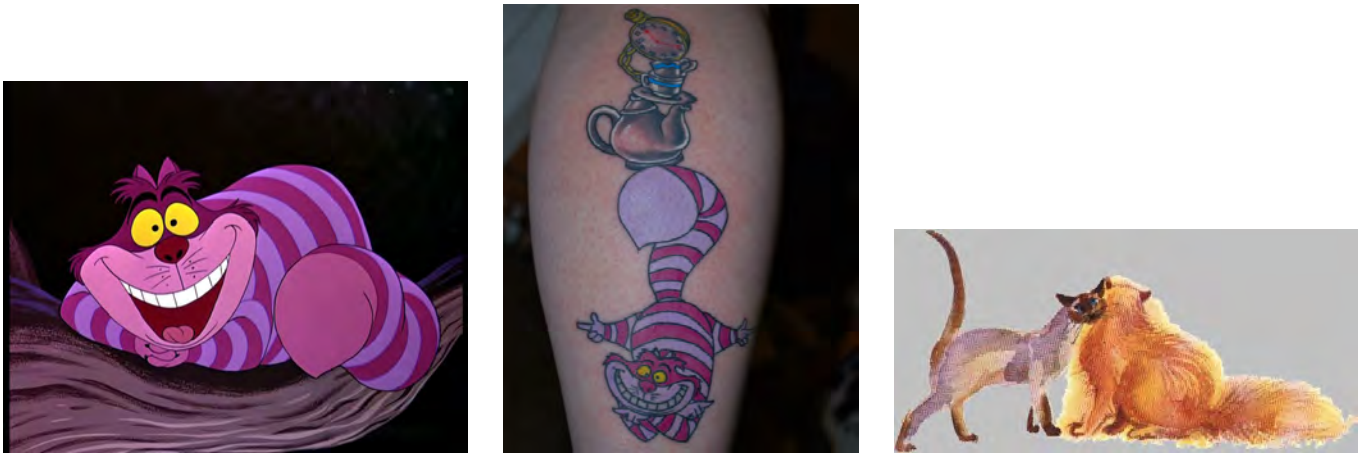
RELEVANT_MINUS изображение частично релевантно запросу, либо релевантно, но имеет плохое качество или какие-либо артефакты;

IRRELEVANT изображение нерелевантно запросу.

В случае противоречивости асессорских оценок использовалась последняя по времени. Распределение объектов по классам представлено в таблице 1, примеры

Метка класса	Количество объектов	Доля от общего числа объектов
RELEVANT_PLUS	260819	0.49
RELEVANT_MINUS	74456	0.14
IRRELEVANT	197580	0.37

Таблица 1: Распределение размеченных триплетов по классам



(a) RELEVANT_PLUS (b) RELEVANT_MINUS (c) IRRELEVANT

Рис. 1: Примеры размеченных изображений для запроса [Диснеевский Чеширский кот]

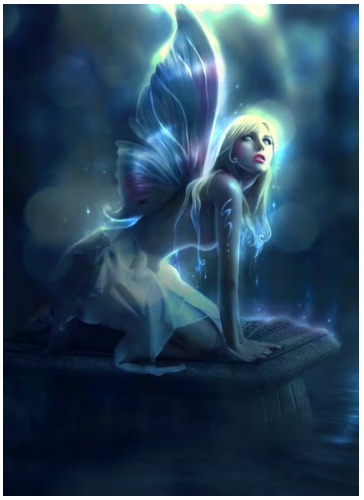
запросов[5] с изображениями разных классов приведены на рисунках 1 и 2. Также отметим, что класс RELEVANT_MINUS, объекты которого можно отнести как к классу RELEVANT_PLUS, так и к IRRELEVANT, достаточно противоречив, поэтому возможны различные варианты его использования в процессе обучения.

Обучающая выборка была случайно разделена на обучение и валидацию в отношении 9:1. Также использовалась заранее отложенная тестовая корзина, не пересекающаяся с обучением и валидацией ни по текстовым запросам, ни по изображениям, ее описание представлено в разделе 4.3.

Анализ пользовательского поведения выходит за рамки данной работы, однако рассматриваемая дискриминативная модель допускает использование любой обучающей выборки вида <текстовый запрос, изображение, степень релевантности>, то есть возможно использование данных, полученных с помощью анализа логов действий пользователей, которые собирают поисковые системы, достаточно лишь

[5]Здесь и далее текст пользовательского запроса обособляется квадратными скобками.

(a) RELEVANT_PLUS



(b) RELEVANT_MINUS

(c) IRRELEVANT

Рис. 2: Примеры размеченных изображений для запроса [картинки фея воды]

заменить асессорскую метку на, например, CTR[6]. Таким образом, наличие описанной выше асессорской разметки не является обязательным для использования рассматриваемого подхода, что существенно расширяет область его применимости.

2 Обзор существующих методов генерации признаковых описаний

Алгоритм дискриминативного решения задачи оценки релевантности изображения текстовому запросу состоит из трех основных этапов, которые могут выполняться как последовательно, так и одновременно:

1. Генерация описания текстового запроса;
2. Генерация описания изображения;
3. Оценка близости полученных на первых двух этапах описаний.

В описанном в данной работе методе на каждом этапе используется нейронная сеть, причем генерация описаний и оценка их близости происходят одновременно. В общем случае возможны и другие подходы как к извлечению признаковых

[6]CTR (Click-Through Rate) — отношение числа кликов ресурса к числу его показов в поисковой выдаче. Является одной из характеристик пользовательского поведения.

описаний, так и к оцениванию релевантности. Ниже представлены различные дескрипторы текстовых запросов и изображений, функции потерь, а также приведены работы, в которых они описаны и успешно применялись.

2.1 Представления текстового запроса

В настоящие время существует несколько широко распространенных методов генерации признакового описания текста. Некоторые модели можно обучать непосредственно на запросах, остальные требуют некоторый текстовый корпус[7] для обучения. Первый вариант предпочтительней, так как распределение слов, синтаксическое и семантическое строения запросов отличаются от соответствующих характеристик текстов на естественном языке, однако некоторым методам необходимо большое число слов в документах корпуса, что делает использование запросов затруднительным. Часто в качестве обучающего текстового корпуса используется Википедия или другие схожие интернет-ресурсы.

2.1.1 Мешок слов

Один из самых распространенных и простых в реализации алгоритмов представления текста, подробно описан в [23].

Данный подход[8] основан на предварительном построении словаря на некотором обучающем корпусе $C = \{d_k\}_{k=1}^{M}$ (состоящем из документов d_k). Затем каждому слову с номером i текста T ставится в соответствие количество его вхождений в этот текст n_i. Тогда, если словарь состоит из N слов, произвольный текст можно описать вектором $t \in \mathbb{R}^N$, где i-ая координата описывается следующим соотношением:

$$t_i = \frac{n_i}{\sum_{l=1}^{N} n_l} \times \log \frac{M}{|i \in d_k|}, \tag{1}$$

где $|i \in d_k|$ — количество документов, в которые входит слово с номером i. Таким образом, наибольший вес в тексте получают слова, которые часто встречаются

[7]Набор документов, длина которых превышает несколько предложений на естественном языке. Длина же большинства пользовательских запросов не превышает пяти–шести слов.

[8]В данной работе описан так называемый TF-IDF (TF — term frequency, IDF — inverse document frequency) подход к построению мешка слов.

в нем, но редко в документах обучающего корпуса (повышается вес слов, характерных для текста, и понижается вес частотных слов — предлогов, союзов, ...). Отметим, что в качестве обучающего корпуса может быть использовано и множество запросов.

К основным достоинствам метода можно отнести простоту реализации. В то же время данный подход обладает целым рядом недостатков:

- ограниченность словаря не позволяет получить представления для всех текстовых запросов, в частности, обрабатывать запросы с опечатками;
- построенный на большом корпусе словарь даже с усечением может содержать сотни тысяч словоформ, что вызывает необходимость работать с разреженными векторами;
- полученное векторное пространство не обладает достаточной семантической близостью даже при сильной морфологической нормализации запросов[9], то есть слова [баран] и [овца] могут находиться друг к другу не ближе, чем [баран] и [автомобиль].

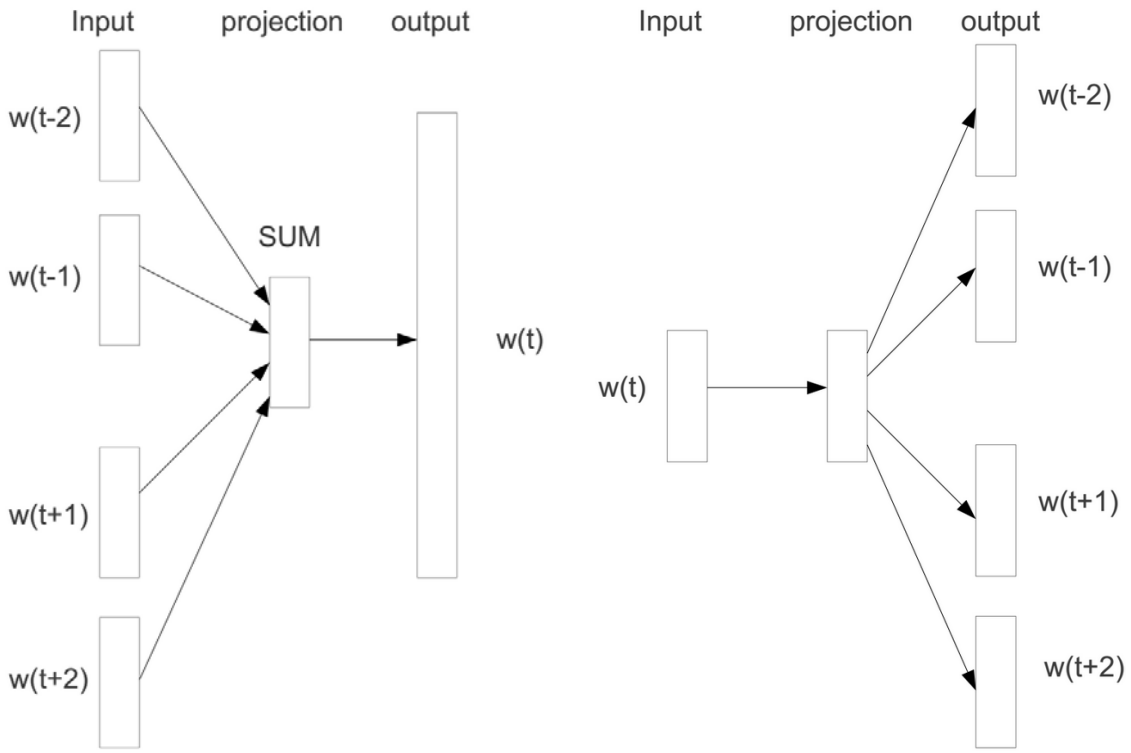
Мешок слов используется, например, в статьях [11, 12]. Отметим, что размер словаря в этих работах составляет всего 179 слов, то есть, по сути, решается задача поиска изображений по тегам, что не совпадает с целью представляемого исследования.

2.1.2 Word2vec

Один из методов, позволяющих обучить проекцию в векторное пространство с семантической близостью, описан в работах [29, 30].

В данном подходе обучающий текстовый корпус просматривается окном ширины $2h + 1$ слов, и для каждого окна однослойная нейронная сеть предсказывает центральное слово окна $w(t)$ по окружающим $w(t + i)$, $i \in [-h, h]\backslash\{0\}$ или наоборот. Эти архитектуры называются Continuous Bag-of-words (рис. 3a) и Skip-gram (рис. 3b) соответственно. Минимизируя ошибку предсказания, нейронная

[9]Например, усечение слов до корня.



(a) Continuous Bag-of-words (b) Skip-gram

Рис. 3: Архитектуры нейронных сетей, представленные в [29, 30] для окна ширины $h = 5$

сеть строит проекцию слов в векторное пространство заранее определенной размерности. При достижении заданной точности предсказания или определенного числа эпох, алгоритм генерирует словарь с векторными представлениями для слов из обучающего корпуса. В качестве представления текста предлагается использовать векторную сумму представлений входящих в него слов. Основным преимуществом данного подхода является хорошая аппроксимация семантической близости слов с помощью косинусной меры сходства между проекциями этих слов в полученном пространстве. Примеры ближайших соседей в обученном пространстве представлены в таблице 2.

Помимо семантической близости, обученное векторное пространство обладает также и лингвистической регулярностью, то есть разность векторных представлений слов [баран] и [овца] близка к разности [король] и [королева] или [он] и [она]. Иллюстрирующая это визуализация, полученная с помощью PCA-проекции [35] векторного пространства на две главные компоненты, представлена на рисунке 4a. Аналогичная визуализация относительно временных форм неправильных английских глаголов представлена на рисунке 4b. Однако данный подход не лишен недо-

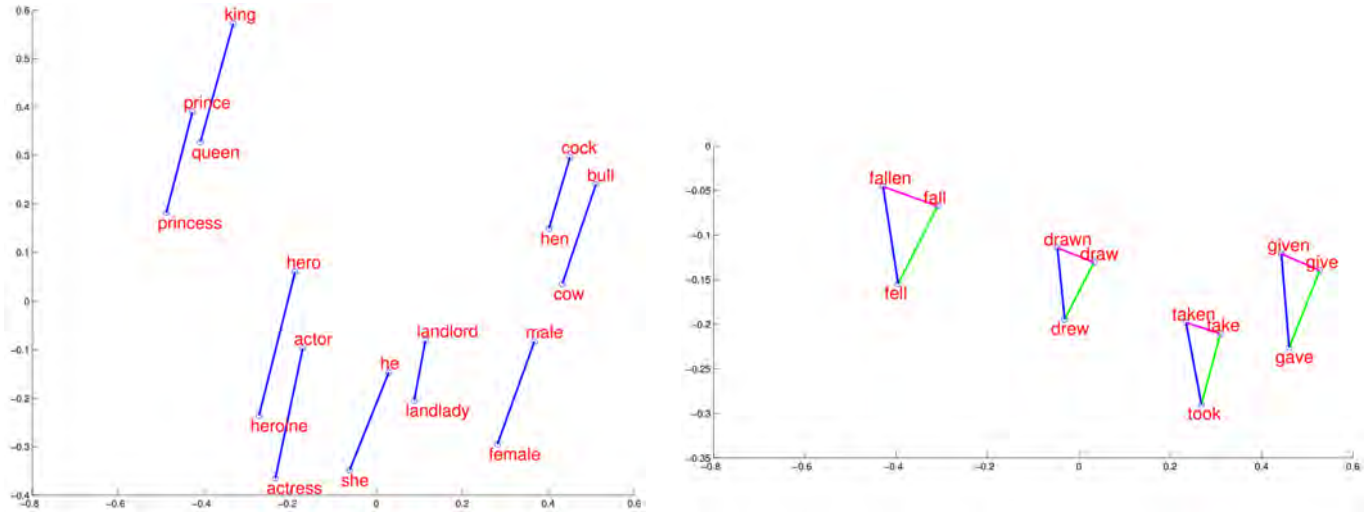
Запрос	Пять ближайших запросов с расстояниями
[рфб баскетбол]	[баскетбол нба] — 0.218208 [рфб] — 0.24761 [баскетбол нба плей офф] — 0.377312 [цска химки баскетбол пятый матч] — 0.406221 [баскетбол локомотив жальгирис счет] — 0.428
[куклы винкс ты настоящая фея]	[винкс кто ты из винкс] — 0.189255 [волшебницы винкс] — 0.204697 [новые куклы винкс] — 0.227931 [принцесса клубничка кукла] — 0.230699 [какой секрет открыла черепаха тортилла в сказке буратино] — 0.242411
[декан вмк мгу]	[ректор мгу] — 0.12568 [мгу] — 0.210726 [декан] — 0.241007 [мгу факультет журналистики] — 0.241822 [иппк мгу] — 0.248235
[поиск экстремумов с помощью нейронной сети]	[расчет параметров сети по ip и маске] — 0.404547 [site-to-site соединение пингуется только с локального хоста] — 0.408153 [Восприятие сигналов среды нервной системой осуществляется с помощью] — 0.409178 [построение цветовых профилей самостоятельно с помощью сканера] — 0.440118 [сетевое моделирование. Метод критического пути] — 0.448421

Таблица 2: Пять ближайших соседей в обученном Word2vec пространстве

статков:

- ограниченность словаря не позволяет получить представления для всех текстовых запросов, в частности, обрабатывать запросы с опечатками;
- семантическая близость присутствует лишь для слов, при их суммировании для получения представления текста близость может теряться, причем чем больше слов суммируется, тем заметнее этот эффект;
- в силу краткости запросов невозможно использовать окно достаточной ширины[10], что сказывается на качестве аппроксимации семантической близости

[10]Рекомендуемые значения ширины окна: $h = 5$ для Continuous Bag-of-words и $h = 11$ для Skip-gram.



(a) Регулярность относительно пола объекта (b) Регулярность относительно времени глагола

Рис. 4: Лингвистическая регулярность в обученном векторном пространстве

в обученном пространстве.

Благодаря лингвистической регулярности, данный алгоритм применяется при машинном переводе [47]. Реализация алгоритма доступна на сайте разработчиков: `https://code.google.com/p/word2vec/`.

2.2 Представления изображения

Аналогично представлению текстовых данных, существует множество дескрипторов изображений. Ниже рассмотрены некоторые из них.

2.2.1 Мешок визуальных слов

Адаптированный для изображений аналог описанного выше локального метода представления текстов, подробное описание которого можно найти в [9].

Данный подход основан на выделении областей изображений обучающего корпуса с последующей кластеризацией их дескрипторов на основе Евклидова расстояния. Разбиение на области и извлечение дескрипторов производится с помощью регулярной сетки или особых точек, которые можно получить с помощью различных детекторов особых точек [5, 28, 37]. Центроид каждого кластера становится словом в визуальном словаре, а описание изображения строится аналогично классическому мешку слов.

Существуют различные вариации этого метода, например, в [12] предлагается использовать словарь для построения дескриптора цвета, то есть для получения описания изображения кластеризуются все его пиксели согласно их цвету.

2.2.2 Сверточная нейронная сеть

Метод построения глобальных дескрипторов изображений впервые описан в [26]. Исходя из результатов проводимого ежегодно соревнования по распознаванию изображений ImageNet Large Scale Visual Recognition Challenge [39] можно заключить, что наилучшие признаковые описания изображений получены именно с помощью сверточных нейронных сетей. Извлекаемые ими дескрипторы достаточно информативны и могут быть использованы не только для классификации в ILSVRC, но и в других задачах [3, 49].

Сверточные нейронные сети, по сути, являются расширением полносвязных сетей слоями двух видов:

Сверточный слой осуществляет свертку изображения с фильтром[11], поэтому следует либо за слоем-входом, либо за другим сверточным слоем. Каждый такой слой состоит из нескольких фильтров одного размера, то есть на выходе получается несколько так называемых независимых карт признаков — по одной карте на фильтр. Веса фильтров настраиваются во время обучения сети. За счет локальности сверточные слои помогают как сократить набор настраиваемых параметров по сравнению с полносвязными слоями, так и увеличить обобщающую способность сети.

Субдискретизирующий слой осуществляет субдискретизацию карты признаков, то есть понижает размерность. Существует несколько подходов к субдискретизации: взятие каждого n-ого элемента строки/столбца карты, выбор среднего или максимального элемента в некоторой области карты признаков. Наличие субдискретизирующих слоев делает сеть инвариантной к изменениям масштаба, а также ускоряет вычисления за счет уменьшения вычислительной сложности операции свертки.

[11] Под фильтром понимается матрица весов, с которой сворачивается изображение.

Обычно сверточные и субдискретизирующие слои чередуются, таким образом на выходе получается несколько тысяч независимых карт признаков, которые связываются с полносвязными слоями для дальнейших преобразований.

2.3 Функции потерь

В этой секции рассмотрены различные функции потерь, которые может оптимизировать (минимизировать) на выходе глубокая нейронная сеть для оценки близости изображения и текстового запроса. Введем следующие обозначения: $X = \{x_i\}_{i=1}^m$ — обучающая выборка, y_i — предсказание блочной нейронной сети на объекте x_i, а $\hat{y}_i$ — наблюдаемый отклик на объекте x_i. Также рассмотрим $\mathcal{L}(y_i, \hat{y}_i)$ — функцию потерь, характеризующую отклонение ответа алгоритма y_i от правильного ответа $\hat{y}_i$ на объекте x_i. Таким образом, эмпирический риск (функционал качества, характеризующий среднюю ошибку на обучающей выборке) запишется следующим образом:

$$Q(\{y_i\}_{i=1}^m, \{\hat{y}_i\}_{i=1}^m) = \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}(y_i, \hat{y}_i). \tag{2}$$

В процессе обучения блочная нейронная сеть оптимизирует именно эмпирический риск, конкретный вид которого зависит от выбранной функции потерь. Ниже представлены наиболее распространенные виды таких функций.

2.3.1 Среднеквадратическая ошибка

При решении задачи регрессии на выходе нейронной сети необходимо получить вещественное число, например, $y_i, \hat{y}_i \in [0, 1]$. Классической функцией потерь в данном случае является $\mathcal{L}(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$, то есть нейронная сеть минимизирует среднеквадратическую ошибку на обучающей выборке.

Так как исходная разметка в рассматриваемой задаче состоит из трех классов, было использовано следующие преобразование меток в числа из интервала [0,1]:

$$\hat{y}_i = \begin{cases} 1 & : \text{RELEVANT\_PLUS} \\ 0.66 & : \text{RELEVANT\_MINUS} \\ 0 & : \text{IRRELEVANT} \end{cases}$$

2.3.2 Кросс-энтропия

При решении задачи K-классовой классификации на выходе нейронной сети необходимо получить вероятность принадлежности рассматриваемого объекта каждому из классов, то есть в данном случае $y_i, \hat{y_i} \in [0,1]^K$, где $\sum_{j=1}^{K} y_{ij} = 1$, $\sum_{j=1}^{K} \hat{y_{ij}} = 1$ и $\forall i \ \exists j : \hat{y_{ij}} = 1$. Классической функцией потерь в данном случае является $\mathcal{L}(y_i, \hat{y_i}) = -\sum_{j=1}^{K} \hat{y_{ij}} \ln y_{ij}$[12], то есть нейронная сеть минимизирует среднюю кросс-энтропию на обучающей выборке.

В данной работе для упрощения рассматривалась классификация на 2 класса: RELEVANT_PLUS и IRRELEVANT[13].

3 Описание предлагаемой модели

В этой части рассматриваются итоговая архитектура представляемой блочной нейронной сети, методы расширения обучающей выборки, а также метод оптимизации параметров сети.

3.1 Архитектура глубокой нейронной сети

В данной работе предлагается использовать блочную нейронную сеть для оценки релевантности изображения текстовому запросу. Для извлечения дескрипторов изображений предлагается использовать сверточную нейронную сеть, этот подход отлично зарекомендовал себя в последние годы в области компьютерного зрения. Дескрипторы текстовых запросов обучаются отдельно[14], поэтому они подаются на вход полносвязной нейронной сети для получения лучших результатов за счет совместного обучения со сверточным блоком. Выходные нейроны сверточного и текстового блоков объединяются на входе еще одного полносвязного блока, который и вычисляет близость изображения текстовому запросу. Таким образом,

[12]Наличие $\ln y_{ij}$ может приводить к вычислительным проблемам при $y_{ij} = 0$. В таком случае можно принять $\ln y_{ij} = -100$.

[13]Можно рассматривать и классификацию с тремя классами, но в таком случае необходимо ввести матрицу стоимости ошибок, то есть штрафовать за отнесение, например, объекта класса RELEVANT_PLUS к классу IRRELEVANT сильнее, чем за отнесение объекта класса RELEVANT_MINUS к классу RELEVANT_PLUS.

[14]Полностью уйти от использования предподсчитанных векторных представлений запросов сложно, так как нейросетевые подходы для работы непосредственно с текстовыми данными слабо развиты.

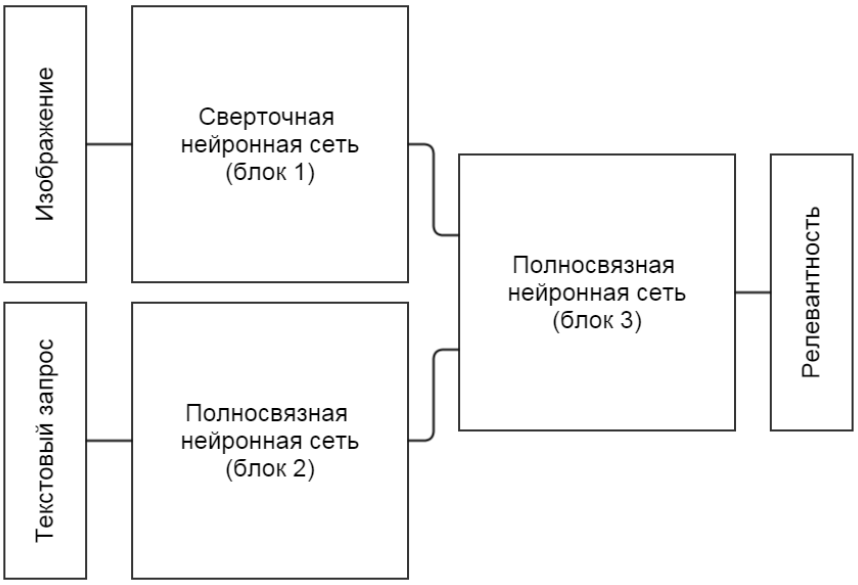


Рис. 5: Блочная архитектура нейронной сети

представляемая модель не генерирует аннотацию изображения, а непосредственно оценивает релевантность, то есть является дискриминативной.

К преимуществам рассматриваемой архитектуры глубокой нейронной сети можно отнести совместное обучение проекций во всех трех блоках, то есть построение меры схожести $\tilde{\mu}$ между объектами исходных пространств T и R. Также стоит отметить независимость блоков, что дает возможность предобучать или вносить изменения в архитектуры каждого из них отдельно. Блочная структура нейронной сети представлена на рисунке 5.

Ниже рассмотрены архитектуры каждого из трех блоков.

3.1.1 Блок 1 (сверточная нейронная сеть)

В данной работе в качестве сверточного блока нейронной сети рассматривается архитектура AlexNet [25], представленная на рисунке 6, — глубокая сверточная нейронная сеть, выигравшая соревнование ILSVRC'12. Выходы нейронов первого полносвязного слоя[15] этой модели рассматриваются как признаковое описание изображений. Аналогично [25], сверточный блок был предобучен на базе ImageNet, состоящей из 1000 классов, на каждый из которых приходится по ≈ 1000 изображений. Веса данного блока фиксированы на протяжении всего процесса обучения блочной нейронной сети. Часто после обучения сети с фиксирован-

[15]Третий справа на рисунке 6.

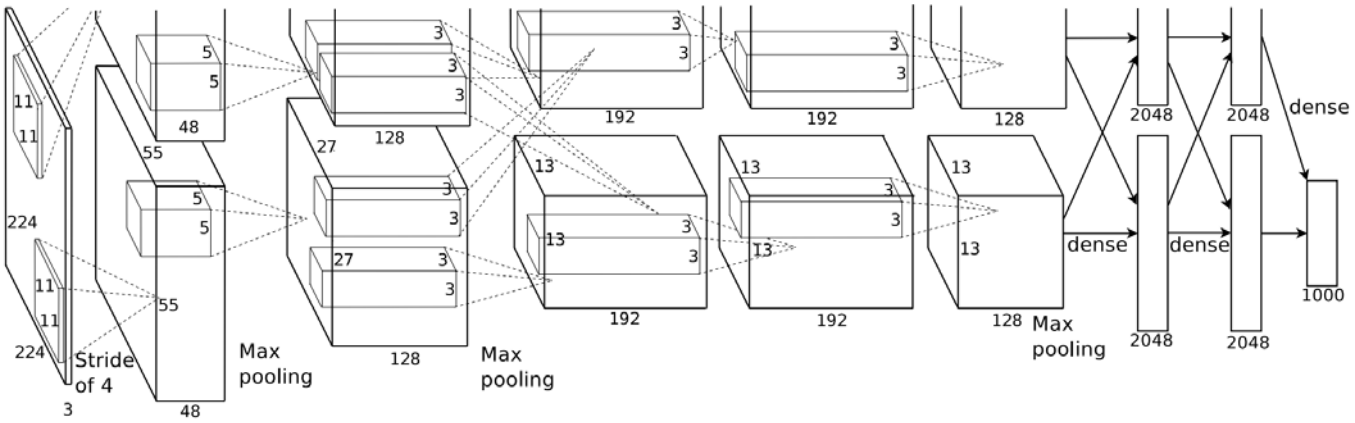


Рис. 6: Архитектура сверточной нейронной сети, представленной в [25]

ными предобученными сверточными слоями их дообучают вместе с остальными под решаемую задачу [3], однако в данном исследовании этот подход не был использован в силу ограниченности временных и вычислительных ресурсов.

3.1.2 Блок 2 (полносвязная нейронная сеть)

Произвольное векторное представление запроса проецируется в пространство $\mathbb{R}^{1024}$ полносвязным слоем, за которым следуют ReLU[16] [31, 25] в качестве функции активации и dropout [15, 25] во избежание переобучения.

3.1.3 Блок 3 (полносвязная нейронная сеть)

Полученные представления текстового запроса и изображения конкатенируются и проецируются в пространство $\mathbb{R}^{1024}$ полносвязным слоем с ReLU и dropout аналогично блоку 2.

Для оптимизации среднеквадратической ошибки этот слой соединяется со слоем из одного нейрона с сигмоидальной функцией активации $\sigma(x) = \frac{1}{1+e^{-x}}$, для оптимизации кросс-энтропии — со слоем из двух нейронов (для двухклассовой классификации) с softmax функцией активации $\sigma(x)_j = \frac{e^{x_j}}{\sum_{i=1}^{2} e^{x_i}}, j \in \{1, 2\}$. В качестве итоговой релевантности в задаче классификации используется значение softmax функции активации для класса RELEVANT_PLUS.

[16]ReLU — rectified linear unit.

3.2 Расширение обучающей выборки

Так как глубокие нейронные сети обладают сотнями миллионов настраиваемых параметров, для их обучения необходима большая обучающая выборка. Для ее расширения будем строить новые изображения, внося искажения в исходные данные. В данном исследовании применяются методы, аналогичные рассмотренным в [25]:

PCA Color Augmentation — преобразование яркостей каналов исходного изображения.

Кадрирование — извлечение изображения меньшего размера (224×224 пикселя) в произвольном месте исходного изображения.

Отражение — зеркалирование входного изображения относительно горизонтальной или вертикальной осей симметрии изображения с некоторой вероятностью.

Преобразования изображений производятся в режиме реального времени во время обучения сети, поэтому не приводят к увеличению объема информации, хранимой на диске. В силу того, что векторные представления запросов обучаются заранее, невозможно производить их преобразования в режиме реального времени, поэтому в данной работе расширение текстовых запросов не применяется[17].

3.3 Метод оптимизации

С развитием вычислительных систем стала возможна стахостическая градиентная оптимизация весов глубоких нейронных сетей с помощью метода обратного распространения ошибки [38]. В данном исследовании оптимизация проводится с помощью градиентного спуска Нестерова [32] с использованием моментум с константой 0.9 и L2-регуляризацией весов матриц проекций с константой 0.0005 [8, 25].

[17] Одним из вариантов внесения искажений в текстовые данные является замена синонимичных слов согласно тезаурусу [51].

Также использовалось обучение на минибатчах с 244 триплетами <текстовый запрос, изображение, метка релевантности> в батче [8]. Использование приведенных метода стохастической оптимизации и значений констант обусловлено большим количеством исследований, проведенных в этой области [8, 25, 42]. Константа скорости обучения для каждого эксперимента подбирается отдельно, а также уменьшается по достижении сходимости для более точной настройки весов. Обычно итерации уменьшения константы повторяются, пока ошибка на валидационной выборке не перестанет меняться или не начнет увеличиваться, однако из-за ограниченности временных и вычислительных ресурсов некоторые эксперименты прерывались до достижения минимального значения оптимизируемого функционала.

Как уже отмечалось, сверточный блок глубокой нейронной сети был предобучен, а его веса во время обучения не оптимизировались. Таким образом одновременно настраивались 3 полносвязных слоя: преобразование текстового представления, преобразование объединенных представлений текстового запроса и изображения и вспомогательный слой, выход которого оценивает релевантность.

4 Эксперименты

В этом разделе приведен обзор известных моделей, решающих схожие задачи, которые могут быть использованы для сравнения с предлагаемой блочной нейронной сетью. Также представлены результаты экспериментов с различными представлениями запросов и функциями потерь для выбора наилучшей блочной сети.

4.1 Известные модели

Одни из первых подходов посвященных оценке близости контента изображения текстовому запросу описаны в [11, 12]. В [12] рассматривается блочная нейронная сеть, архитектура которой представлена на рисунке 7. Блок, отвечающий за построение описания изображения (L1, A2 и T3), вычисляет инженерные де-

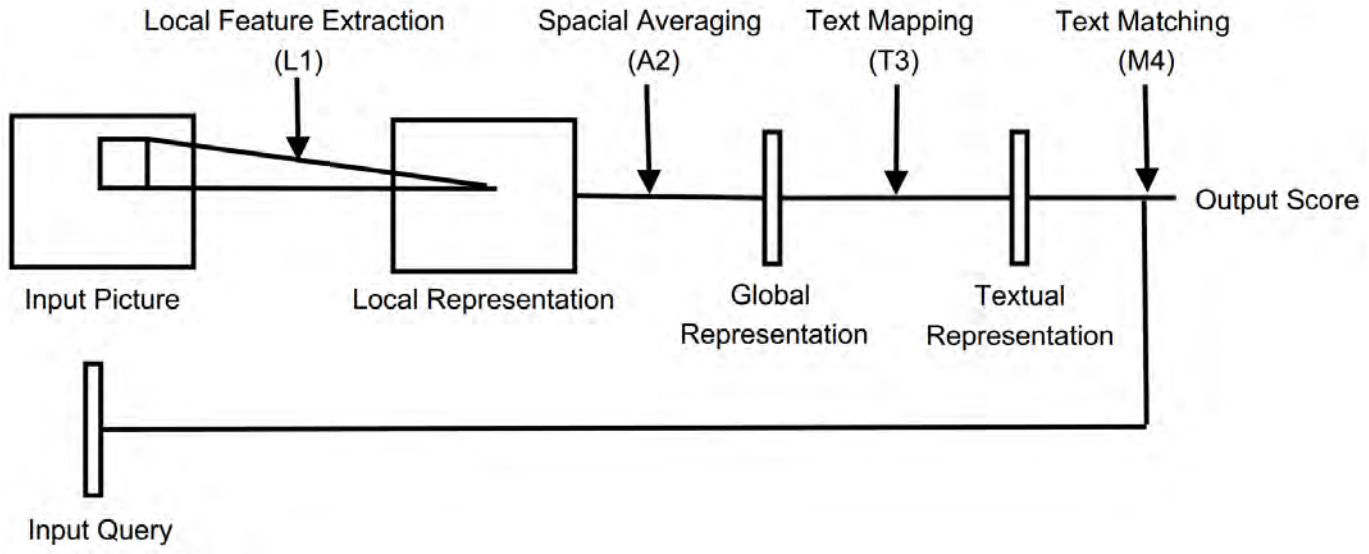


Рис. 7: Блочная архитектура нейронной сети, представленная в [12]

скрипторы изображения[18], близость которых к текстовому запросу определяется как скалярное произведение их векторных представлений (M4). Основными отличиями предлагаемой блочной нейронной сети от модели [12] являются:

- построение дескрипторов изображения: последние исследования и соревнования в области компьютерного зрения доказывают превосходство сверточных нейронных сетей над инженерными дескрипторами в задачах распознавания образов;
- обработка текстового запроса: в [12] в качестве представления запроса используется мешок слов, однако размер словаря составляет 179 слов, что позволяет использовать скалярное произведение как меру сходства. В предлагаемой модели размер векторного представления запроса заранее не ограничен, что делает вычислительно сложным преобразование на шаге T3 в силу наличия матрицы, вторая размерность которой может достигать сотен тысяч.

Другой подход к решению задачи построения гетерогенной меры схожести описан в [17, 16] и заключается в построении общего векторного пространства для

[18]На шаге L1 вычисляются локальные дескрипторы для каждого блока в разбиении изображения: конкатенируются гистограмма распределения цветов пикселей и гистограмма распределения направлений градиента, полученная с помощью uniform Local Binary Pattern [14, 43]. Затем на стадии A2 локальные представления усредняются по всем блокам для получения глобального описания изображения, которое преобразуется к размеру словаря, используемого для построения мешка слов, с помощью полносвязного слоя на шаге T3.

объектов из множеств разной природы на основе их признаковых описаний. Рассмотрим два множества: $T = \{t_i\}_{i=1}^k$, $t_i \in \mathbb{R}^n$ и $R = \{r_i\}_{i=1}^k$, $r_i \in \mathbb{R}^m$, где изображение r_i релевантно текстовому запросу t_i. Проекции $U \in \mathbb{R}^{n \times d}$ и $V \in \mathbb{R}^{m \times d}$ в общее векторное пространство $\mathbb{R}^d$ обучаются таким образом, чтобы корреляция между представлениями объектов r_i и t_i была максимальна, то есть рассматривается следующая задача оптимизации: $\text{corr}(TU, RV) \to \max_{U,V}$. Один из методов этого семейства, canonical correlation analysis, используется в рамках данного исследования для сравнения с блочной нейронной сетью. В качестве дескрипторов текста используется Word2vec, в качестве представлений изображений — выход первого полносвязного слоя сверточной нейронной сети AlexNet.

4.2 Оптимизация ошибки сети

В данном исследовании в качестве представлений запросов рассматриваются мешок слов и Word2vec, в качестве функционалов ошибки — среднеквадратическая ошибка и кросс-энтропия. Ниже представлены различные варианты сочетаний дескрипторов запросов и функционалов ошибки с целью выбора наилучшей модели.

4.2.1 Мешок слов + среднеквадратическая ошибка

График обучения нейронной сети представлен на рисунке 8, где по оси абсцисс отложено время обучения блочной нейронной сети, а по оси ординат — значение среднеквадратической ошибки. Стратегия понижения константы скорости обучения: 0.01 до 19 эпохи, 0.005 до конца обучения. Уменьшение среднеквадратической ошибки на валидационной выборке говорит о сходимости метода.

4.2.2 Мешок слов + кросс-энтропия

Графики обучения нейронной сети представлены на рисунках 9a и 9b, где по оси абсцисс отложено время обучения блочной нейронной сети, а по оси ординат — значение кросс-энтропии и доли неправильно классифицированных объектов соответственно. Стратегия понижения константы скорости обучения: 0.005 до

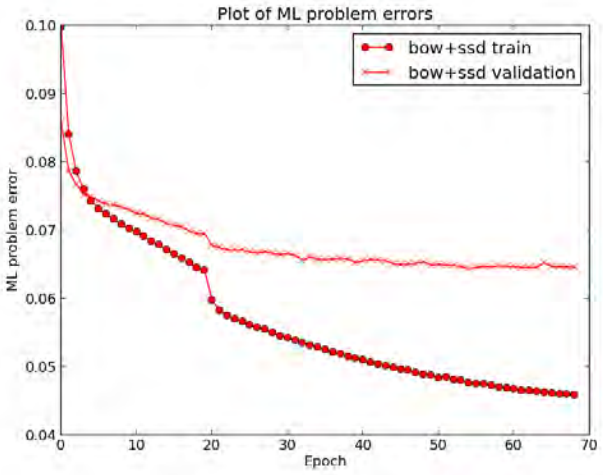
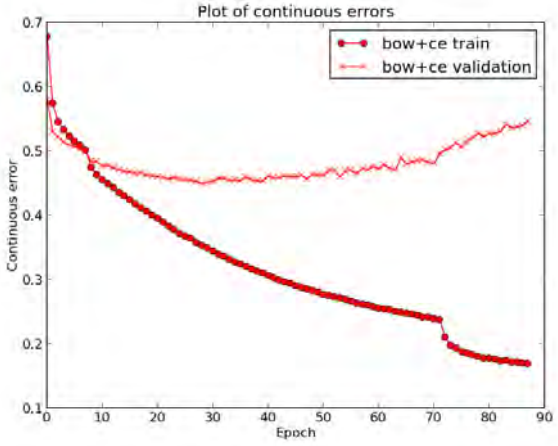
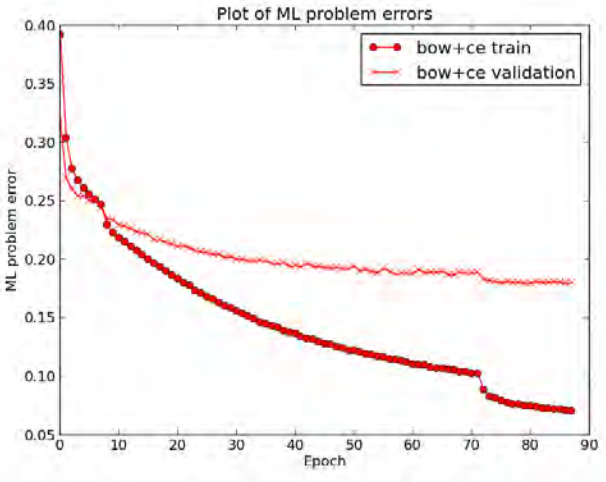


Рис. 8: Мешок слов + среднеквадратическая ошибка



(a) Кросс-энтропия



(b) Доля неправильно классифицированных объектов

Рис. 9: Мешок слов + кросс-энтропия

7 эпохи[19], 0.002 до конца обучения. Уменьшение доли неправильно классифицированных объектов на валидационной выборке говорит о сходимости метода.

4.2.3 Word2vec + среднеквадратическая ошибка

График обучения нейронной сети представлен на рисунке 10, где по оси абсцисс отложено время обучения блочной нейронной сети, а по оси ординат — значение среднеквадратической ошибки. Стратегия понижения константы скорости обучения: 0.01 до 19 эпохи, 0.002 до 36 эпохи и 0.0005 до конца обучения. Уменьшение

[19]Первый раз константа скорости обучения была понижена рано в следствии ошибки, результат эксперимента мог быть лучше.

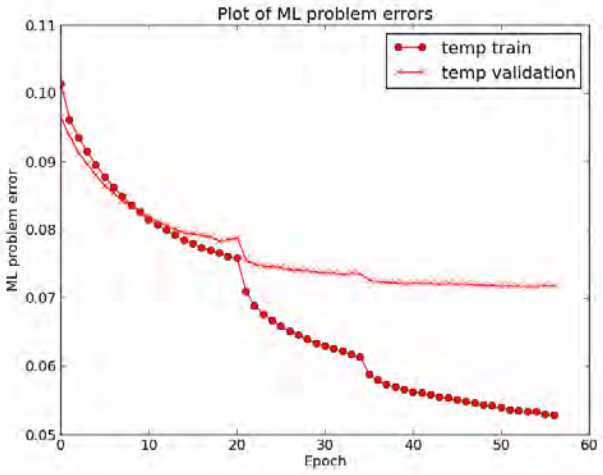


Рис. 10: Word2vec + среднеквадратическая ошибка

среднеквадратической ошибки на валидационной выборке говорит о сходимости метода.

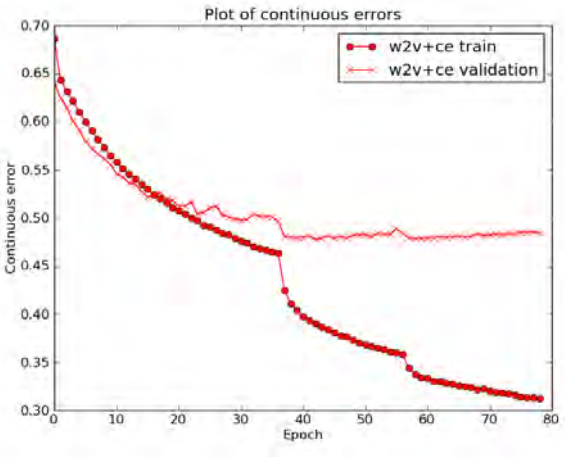
4.2.4 Word2vec + кросс-энтропия

Графики обучения нейронной сети представлены на рисунках 11a и 11b, где по оси абсцисс отложено время обучения блочной нейронной сети, а по оси ординат — значение кросс-энтропии и доли неправильно классифицированных объектов соответственно. Стратегия понижения константы скорости обучения: 0.005 до 37 эпохи, 0.002 до 57 эпохи и 0.0005 до конца обучения. Уменьшение доли неправильно классифицированных объектов на валидационной выборке говорит о сходимости метода.

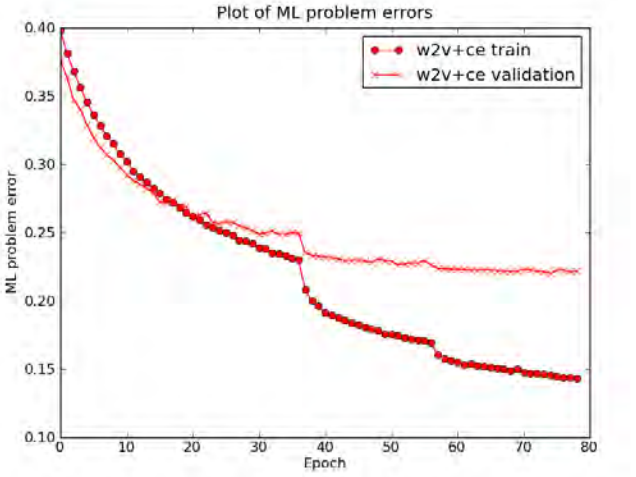
4.3 Использование предсказанной релевантности для ранжирования

Для оценки качества предсказания релевантности был проведено два эксперимента с использованием запросов из заранее отложенной тестовой выборки, репрезентативной по отношению к потоку запросов поисковой системы:

1. Ранжирование всех размеченных изображений для 50 запросов;
2. Ранжирование первых ≈ 100 изображений поисковой выдачи для 773 запросов.



(a) Кросс-энтропия



(b) Доля неправильно классифицированных объектов

Рис. 11: Word2vec + кросс-энтропия

В обоих случаях результаты ранжирования согласно предсказанной релевантности сравниваются с canonical correlation analysis.

Целевой метрикой является следующая величина: $\sum_{i=1}^{30} \frac{\text{rel}(q,d_i)}{30}$, где $\text{rel}(q, d_i)$ — функция, определяющая релевантность i-ого изображения запросу q:

$$\text{rel}(q,d) = \begin{cases} 1.0 & : \text{RELEVANT\_PLUS} \\ 0.5 & : \text{RELEVANT\_MINUS} \\ 0.0 & : \text{IRRELEVANT} \end{cases},$$

что, по сути, соответствует взвешенной точности среди первых тридцати изображений после ранжирования согласно предсказанной релевантности.

Помимо абсолютной величины целевого функционала ранжирования рассматривается достигаемый уровень значимости гипотезы об эквивалентности ранжирований согласно критерию знаковых рангов Уилкоксона для связных выборок [46] с двусторонней альтернативой.

4.3.1 Ранжирование случайных изображений

В качестве запросов были взяты 50 случайных запросов из заранее отложенной валидационной корзины. Для каждого запроса были взяты все размеченные по нему изображения, таким образом для эксперимента было построено декартово

Метка класса	Количество объектов	Доля от общего числа объектов
RELEVANT_PLUS	2738	0.010442
RELEVANT_MINUS	683	0.002605
IRRELEVANT	258779	0.986953

Таблица 3: Распределение размеченных триплетов по классам при ранжировании случайных изображений

произведение множества запросов и изображений: для каждой пары выбиралась асессорская оценка из разметки, если таковой не было, изображение считалось нерелевантным. Итоговое распределение объектов по классам представлено в таблице 3.

Для оценки качества ранжирования значения целевого функционала усреднялись по запросам. Для получения значения функционала при случайном ранжировании изображения были перемешаны 100 раз, а результаты усреднены по перемешиваниям и запросам. Значение целевого функционала для случайной перестановки составило 0.011907. Значение целевого функционала для canonical correlation analysis составило 0.045000, что подтверждает состоятельность этого метода. В таблице 4 представлены значения функционала для четырех рассмотренных выше блочных моделей вместе с соответствующими достигаемыми уровнями значимости по сравнению с базовым методом. Очевидно, гипотеза о эквивалентности ранжирований принимается на уровне значимости $\alpha = 0.05$ для моделей на основе Word2vec представления текстовых запросов, при этом ранжирование моделями на основе мешка слов значимо проигрывает базовому методу[20]. Также отметим, что значимых отличий между различными функциями потерь для фиксированного текстового представления не замечено.

4.3.2 Ранжирование поисковой выдачи

В качестве запросов были взяты 773 случайных запроса из заранее отложенной валидационной корзины. По каждому запросу было размечено ≈ 100 первых изображений поисковой выдачи. Итоговое распределение объектов по классам представлено в таблице 5.

[20]Заметим, что выборка, по которой оценивается статистическая эквивалентность ранжирований состоит всего из 50 запросов, что может привести к понижению чувствительности критерия знаковых рангов Уилкоксона.

Модель	Значения функционала ранжирования	Достигаемый уровень значимости
Мешок слов + среднеквадратическая ошибка	0.017000	0.005787
Мешок слов + кросс-энтропия	0.010667	0.002023
Word2vec + среднеквадратическая ошибка	0.054667	0.573599
Word2vec + кросс-энтропия	0.053000	0.250376

Таблица 4: Значения целевого функционала ранжирования при ранжировании случайных изображений

Метка класса	Количество объектов	Доля от общего числа объектов
RELEVANT_PLUS	47311	0.541049
RELEVANT_MINUS	9635	0.110186
IRRELEVANT	30497	0.348765

Таблица 5: Распределение размеченных триплетов по классам при ранжировании поисковой выдачи

Для оценки качества ранжирования значения целевого функционала усреднялись по запросам. Для получения значения функционала при случайном ранжировании изображения были перемешаны 100 раз, а результаты усреднены по перемешиваниям и запросам. Значение целевого функционала для случайной перестановки составило 0.527235. Значение целевого функционала для canonical correlation analysis составило 0.535489, что подтверждает состоятельность этого метода. В таблице 6 представлены значения функционала для четырех рассмотренных выше блочных моделей вместе с соответствующими достигаемыми уровнями значимости по сравнению с базовым методом. Очевидно, гипотеза о эквивалентности ранжирований отвергается на уровне значимости $\alpha = 0.05$ для моделей на основе Word2vec представления текстовых запросов в силу их превосходства, при этом ранжирование моделями на основе мешка слов проигрывает базовому методу, но этот проигрыш нельзя уверенно назвать значимым. Также отметим, что значимых отличий между различными функциями потерь для фиксированного текстового представления не замечено.

Модель	Значения функционала ранжирования	Достигаемый уровень значимости
Мешок слов + среднеквадратическая ошибка	0.526973	0.014677
Мешок слов + кросс-энтропия	0.528655	0.060269
Word2vec + среднеквадратическая ошибка	0.548879	0.000103
Word2vec + кросс-энтропия	0.547865	0.000314

Таблица 6: Значения целевого функционала ранжирования при ранжировании поисковой выдачи

4.3.3 Визуализация ранжирования случайных изображений

В данном разделе представлены визуальные результаты эксперимента по ранжированию случайных изображений для модели Word2vec + среднеквадратическая ошибка. Рассмотрим три характерных примера запросов:

1. На рисунке 12 представлен конкретный запрос [волосы укладка], в данном случае все изображения релевантны запросу;
2. На рисунке 13 представлен более сложный запрос [шляпы на конкурс в детский сад], в данном случае нейронная сеть улавливает тематику запроса, однако изображения не релевантны;
3. На рисунке 14 представлен тяжелый запрос [доминатор новый], с которым нейронная сеть не справилась, хотя сложно сказать, что пользователь ожидает увидеть по такому запросу.

Отметим, что изображения приведены к размеру 256×256, используемому на входе сверточного блока представленной нейронной сети.

4.3.4 Анализ полученных результатов

Проведенные эксперименты соответствуют двум стадиям ранжирования, присутствующим в большинстве поисковых систем, работающих с большими коллекциями ресурсов [27]: быстрое извлечение небольшого подмножества ресурсов

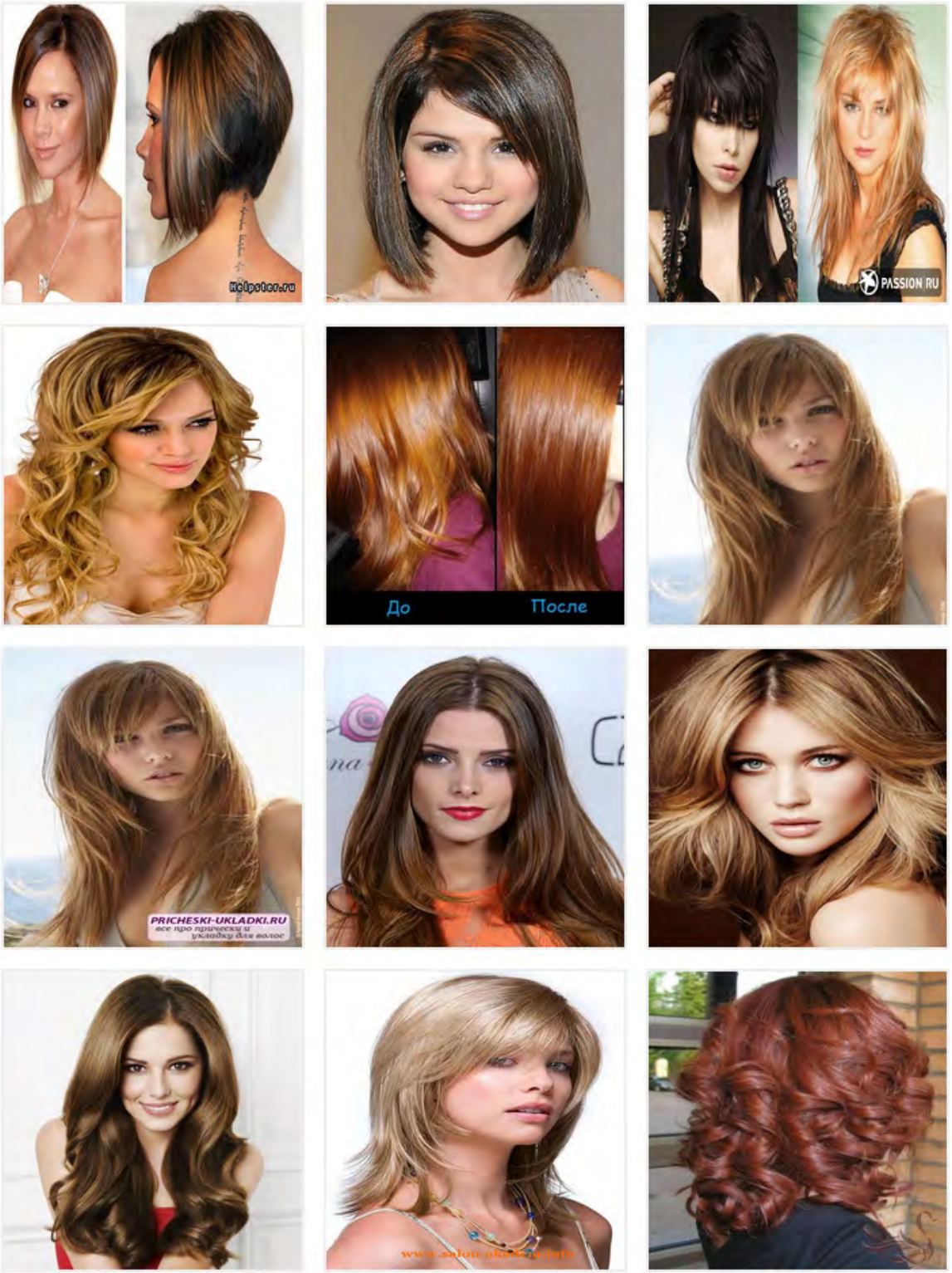


Рис. 12: [волосы укладка]

Рис. 13: [шляпы на конкурс в детский сад]

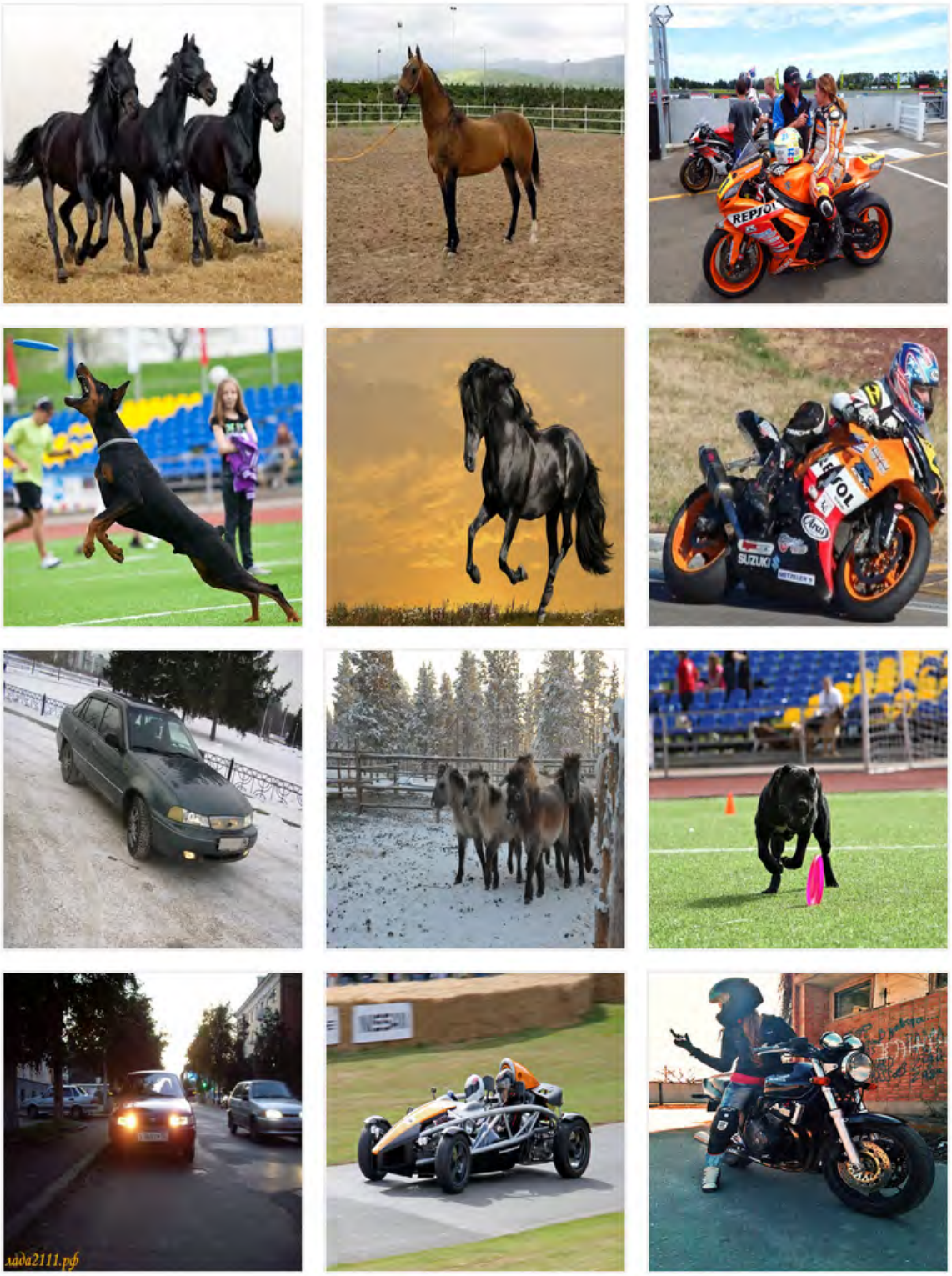


Рис. 14: [доминатор новый]

из коллекции и переранжирование этого подмножества более точным алгоритмом. Результаты раздела 4.3.1 иллюстрируют потенциальное улучшение целевой метрики в случае использования полученной оценки релевантности в качестве одного из факторов ранжирования на первой стадии поиска изображений в больших коллекциях. Результаты раздела 4.3.2 демонстрируют значимое улучшение целевой метрики в случае использования полученной оценки релевантности на второй стадии поиска. Таким образом, представленный в данной работе алгоритм достаточно универсален с точки зрения области применимости, однако предсказание релевантности с помощью глубокой нейронной сети требует значительных вычислительных затрат, что затрудняет применение подобных моделей на первой стадии поиска в связи с ограничениями на время работы алгоритма.

5 Заключение

В данной работе представлен подход к построению гетерогенной меры схожести между текстовым запросом и контентом мультимедийного ресурса с помощью блочной нейронной сети. Описанная дискриминативная модель может обучаться как на асессорской разметке, так и на логах поисковых систем, что позволяет использовать ее в широком спектре гетерогенных задач информационного поиска. Проведенные эксперименты для задачи поиска изображений в больших коллекциях показывают, что полученная таким образом оценка близости контента изображения текстовому запросу может быть использована для улучшения систем ранжирования больших коллекций изображений.

Список литературы

[1] Amato, G. Large scale image retrieval using vector of locally aggregated descriptors / G. Amato, P. Bolettieri, F. Falchi et al. // Similarity Search and Applications. — 2013. — P. 245–256. — (Lecture Notes in Computer Science; Vol. 8199).

[2] Arandjelovic, R. All About VLAD / R. Arandjelovic, A. Zisserman // IEEE Conference on Computer Vision and Pattern Recognition (CVPR'13). — 2013. — P. 1578–1585.

[3] Babenko, A. Neural Codes for Image Retrieval / A. Babenko, A. Slesarev, A. Chigorin, V. Lempitsky // 13th European Conference on Computer Vision (ECCV'14) : proceedings. — 2014. — P. 584–599. — (Lecture Notes in Computer Science, Vol. 8689, Prt. 1).

[4] Barnard, K. Matching words and pictures / K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei, M. I. Jordan // The Journal of Machine Learning Research. — 2003. — Vol. 3. — P. 1107–1135.

[5] Bay, H. SURF: Speeded Up Robust Features / H. Bay, A. Ess, T. Tuytelaars, L. van Gool // Computer Vision and Image Understanding. — 2006. — Vol. 110, No. 3. — P. 346–359.

[6] Bishop C. M. Pattern Recognition and Machine Learning. — New York : Springer, 2006. — xx, 738 p., ill.

[7] Delhumeau, J. Revisiting the VLAD Image Representation / J. Delhumeau, Ph.-H. Gosselin, H. Jegou, P. Perez // 21st ACM international conference on Multimedia : proceedings. — 2013. — P. 653–656.

[8] Deng, L. Deep Learning: Methods and Applications / L. Deng, D. Yu // Foundations and Trends in Signal Processing. — 2014. — Vol. 7 (3–4), P. 197–387.

[9] Fei-Fei, L. A Bayesian Hierarchical Model for Learning Natural Scene Categories / L. Fei-Fei, P. Perona // IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). — 2005. — P. 524–531.

[10] Gao, J. Modeling Interestingness with Deep Neural Networks / J. Gao, P. Pantel, M. Gamon, X. He, L. Deng, Y. Shen // Conference on Empirical Methods in Natural Language Processing : proceedings. — 2014. — P. 2–14.

[11] Grangier, D. A Discriminative Approach for the Retrieval of Images from Text Queries / D. Grangier, S. Bengio // European Conference on Machine Learning (ECML'06) : proceedings. — 2006. — P. 162–173.

[12] Grangier, D. A Neural Network to Retrieve Images from Text Queries / D. Grangier, S. Bengio // 6th International Conference on Artificial Neural Networks: Biological Inspirations (ICANN'06). — P. 24—34. — (Lecture Notes in Computer Science; Vol. 4132).

[13] Gutmann, M. U. Noise-Contrastive Estimation of Unnormalized Statistical Models, with Applications to Natural Image Statistics / M. U. Gutmann, A. Hyvarinen // The Journal of Machine Learning Research. — 2012. — Vol. 13, Iss. 1. — P. 307–361.

[14] He, D.-C. Texture Unit, Texture Spectrum, And Texture Analysis / D.-C. He, L. Wang // IEEE Transactions on Geoscience and Remote Sensing. — 1990. — Vol. 28, № 4. — P. 509–512.

[15] Hinton, G. E. Improving Neural Networks by Preventing Co-adaptation of Feature Detectors [Electronic resourse] / G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R. R. Salakhutdinov // arXiv. — 2012. — URL: `http://arxiv.org/abs/1207.0580` (date of access: 11.05.2015).

[16] Hardoon, D. R. Canonical Correlation Analysis: An Overview with Application to Learning Methods / D. R. Hardoon, S. Szedmak, J. Shawe-Taylor // Neural computation. — 2004. — Vol. 16, № 12. — P. 2639–2664.

[17] Hotelling, H. Relations Between Two Sets of Variates // Biometrika. — 1936. — Vol. 28, № 3–4. — P. 321–377.

[18] Hua, X.-H. Clickage: towards bridging semantic and intent gaps via mining click logs of search engines / X.-S. Hua, L. Yang, J. Wang, J. Wang, M. Ye, K. Wang, Y. Rui, J. Li // ACM Multimedia. — 2013. — P. 243–252.

[19] Huang, P.-S. Learning Deep Structured Semantic Models for Web Search using Clickthrough Data / P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, L. Heck // 22nd ACM international conference on Conference on information & knowledge management : proceedings. — 2013. — P. 2333–2338.

[20] Jarvelin, K. Cumulated Gain-Based Evaluation of IR Techniques / K. Jarvelin , J. Kekalainen // ACM Transactions on Information Systems. — 2002. — Vol. 20. — P. 422–446.

[21] Jegou, H. Aggregating Local Descriptors into a Compact Image Representation // IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10). — 2010. — P. 3304–3311.

[22] Joachims, T. Optimizing Search Engines using Clickthrough Data / T. Joachims // ACM Conference on Knowledge Discovery and Data Mining : proceedings. — 2002. — P. 133-–142.

[23] Jones, K. S. A Statistical Interpretation of Term Specificity and its Application in Retrieval / K. S. Jones // Journal of Documentation. — 2004. — Vol. 60, № 5. — P. 493–502.

[24] Kiros, R. Deep Representations and Codes for Image Auto-Annotation / R. Kiros, C. Szepesvari // Advances in Neural Information Processing Systems, 25. — 2012. — P. 917–925.

[25] Krizhevsky, A. ImageNet Classification with Deep Convolutional Neural Networks / A. Krizhevsky, I. Sutskever, G. E. Hinton // Advances in Neural Information Processing Systems 25. — 2012. — P. 1106–1114.

[26] LeCun, Y. Gradient-based Learning Applied to Document Recognition / Y. LeCun, L. Bottou, Y. Bengio, P. Haffner // Proceedings of the IEEE. — 1998. — Vol. 86, Iss. 11. — P. 2278–2324.

[27] Liu, T.-Y. Learning to Rank for Information Retrieval / T.-Y. Liu. Berlin : Springer Berlin, 2010. — 300 p.

[28] Lowe, D. G. Distinctive Image Features from Scale-Invariant / D. G. Lowe // International Journal of Computer Vision. — 2004. — Vol. 60, Iss. 2. — P. 91–110.

[29] Mikolov, T. Efficient Estimation of Word Representations in Vector Space [Electronic resourse] / T. Mikolov, K. Chen, G. Corrado, J. Dean // arXiv.org. — 2013. — URL: http://arxiv.org/pdf/1301.3781v3.pdf (date of access: 11.05.2015).

[30] Mikolov, T. Distributed Representations of Words and Phrases and their Compositionality / T. Mikolov, I. Sutskever, K. Chen, G. S Corrado, J. Dean // Advances in Neural Information Processing Systems. — 2013. — P. 3111–3119.

[31] Nair, V. Rectified Linear Units Improve Restricted Boltzmann Machines / V. Nair, G. E. Hinton // 27th International Conference on Machine Learning (ICML'10) : proceedings. — 2010. — P. 807–814.

[32] Nesterov, Y. Introductory Lectures on Convex Optimization. A Basic Course / Y. Nesterov. — Boston : Kluwer Academic Publishers, 2004. — xviii, 236 p. — (Applied optimization; Vol. 87).

[33] Ng, A. Y. On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes / A. Y. Ng, I. J. Michael // Advances in Neural Information Processing Systems 14 (NIPS'01). — 2001. — P. 841–848.

[34] Ojala, T. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns / T. Ojala, M. Pietikainen, T. Maenpaa // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 2002. — Vol. 24, Iss. 7. — P. 971–987.

[35] Pearson, K. On Lines and Planes of Closest Fit to Systems of Points in Space / K. Pearson // Philosophical Magazine. — 1901. — Vol. 2, No. 6. — P. 559–572.

[36] Rosenblatt, F. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms / F. Rosenblatt. — Buffalo, N.Y. : Cornell Aeronautical Laboratory, 1961. — xviii, 622 p. — (Cornell Aeronautical Laboratory; Report no. VG-1196-G-8).

[37] Rosten, E. Machine Learning for High-Speed Corner Detection / E. Rosten, T. Drummond // European Conference on Computer Vision. — 2006. — P. 430–443. — (Lecture Notes in Computer Science; Vol. 3951).

[38] Rumelhart, D. E. Learning Representations by Back-Propagating Errors / D. E. Rumelhart, G. E. Hinton, R. J. Williams // Nature. — 1986. — № 323. — P. 533–536.

[39] Russakovsky, O. ImageNet Large Scale Visual Recognition Challenge [Electronic resourse] / Olga Russakovsky et al. // arXiv.org. — 2014. — URL: `http://arxiv.org/pdf/1409.0575v3.pdf` (date of access: 11.05.2015).

[40] Schmidhuber, J. Deep Learning in Neural Networks: An Overview / J. Schmidhuber // Neural Networks. — 2015. — Vol. 61. — P. 85–117.

[41] Shen, Y. A Latent Semantic Model with Convolutional-Pooling Structure for Information Retrieval / Y. Shen, X. He, J. Gao, L. Deng, G. Mesnil // 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM'14) : proceedings. — 2014. — P. 101–110.

[42] Sutskever, I. On the importance of initialization and momentum in deep learning / I. Sutskever, J. Martens, G. Dahl, G. Hinton // 30th International Conference on Machine Learning : proceedings. — 2013. — Vol. 28. — P. 1139–1147.

[43] Wang, L. Texture Classification Using Texture Spectrum / L. Wang, D.-C. He // Pattern Recognition. — 1990. — Vol. 23, Iss. 8. — P. 905–910.

[44] Wang, X.-J. AnnoSearch: Image Auto-Annotation by Search / X.-J. Wang, L. Zhang, F. Jing, W.-Y. Ma // IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). — 2006. — Vol. 2. — P. 1483–1490.

[45] White, R. W. Investigating the Querying and Browsing Behavior of Advanced Search Engine Users / R. W. White, D. Morris // 30th annual international ACM SIGIR conference on Research and development in information retrieval : proceedings. — 2007. — P. 255–262.

[46] Wilcoxon, F. Individual comparisons by ranking methods / F. Wilcoxon // Biometrics Bulletin. — 1945. — Vol. 1, № 6. — P. 80–83.

[47] Wolf, L. Joint word2vec Networks for Bilingual Semantic Representations / L. Wolf, Y. Hanani, K. Bar, N. Dershowitz // International Journal of Computational Linguistics and Applications. — 2014. — Vol. 5, No. 1. — P. 27-–44.

[48] Xu, Z. Cross-Media Relevance Mining for Evaluating Text-Based Image Search Engine / Z. Xu, Y. Yang, A. Kassim, S. Yan // IEEE International Conference on Multimedia and Expo Workshops (ICMEW'14). — 2014. — P. 1–4.

[49] Yosinski, J. How Transferable are Features in Deep Neural Networks? / J. Yosinski, J. Clune, Y. Bengio, H. Lipson // Advances in Neural Information Processing Systems 27 (NIPS'14). — 2014. — P. 3320–3328.

[50] Zhang, Y. Image Search Reranking with Query-Dependent Click-Based Relevance Feedback / Y. D. Zhang, X. P. Yang, T. Mei // IEEE Transactions on Image Processing. — 2014. — Vol.23, No 10. — P. 4448-4459.

[51] Zhang, Y. Text Understanding from Scratch [Electronic resourse] / X. Zhang, Y. LeCun // arXiv. — 2015. — URL: `http://arxiv.org/pdf/1502.01710.pdf` (date of access: 11.05.2015).