Извлечение предпочтений пользователя на основе методов автоматического порождения текстовых описаний изображений фотоальбома

А.С. Харчевникова[1], А.В. Савченко[1]
[1]Национальный исследовательский университет «Высшая школа экономики», 603155, Россия, г. Нижний Новгород, ул. Большая Печерская, д.25/12

Аннотация

В работе рассматривается задача извлечения предпочтений пользователя по его фотоальбому. Предложен новый подход на основе автоматического порождения текстовых описаний фотографий и последующей классификации таких описаний. Проведен анализ известных методов создания аннотаций по изображению на основе свёрточных и рекуррентных (Long short-term memory) нейронных сетей. С использованием набора данных Google's Conceptual Captions обучены новые модели, в которых объединяются характерные признаки фотографии и выходы блока рекуррентной нейронной сети. Исследовано применение алгоритмов обработки текстов для преобразования полученных аннотаций в пользовательские предпочтения. Проведены экспериментальные исследования с помощью наборов данных Microsoft COCO Captions, Flickr8k и специально собранного набора данных, отражающего интересы пользователя. Показано, что наилучшее качество определения предпочтений достигается с помощью методов поиска ключевых слов и суммаризации текстов из Watson API, которые оказываются на 8 % точнее по сравнению с традиционным латентным размещением Дирихле. При этом описания, порожденные с помощью обученных моделей, классифицируются на 1 – 7 % точнее известных аналогов.

Ключевые слова: анализ предпочтений пользователя, обработка изображений, текстовое описание изображений, сверточные нейронные сети.

Цитирование: **Харчевникова, А.С.** Извлечение предпочтений пользователя на основе методов автоматического порождения текстовых описаний изображений фотоальбома / А.С. Харчевникова, А.В. Савченко // Компьютерная оптика. – 2020. – Т. 44, № 4. – С. 618-626. – DOI: 10.18287/2412-6179-CO-678.

Citation: Kharchevnikova AS, Savchenko AV. Visual preferences prediction for a photo gallery based on image captioning methods. Computer Optics 2020; 44(4): 618-626. DOI: 10.18287/2412-6179-CO-678.

Введение

Задачи, связанные с построением рекомендательных систем или сервисов контекстной рекламы, в настоящий момент являются одними из наиболее актуальных приложений технологий искусственного интеллекта [1 – 3]. При этом чаще всего в качестве исходных данных для таких систем используется структурированная или текстовая информация, содержащая, например, отзывы пользователя о некоторых продуктах [2, 4, 5]. К сожалению, во многих случаях в таких системах возникает проблема «холодного старта»: имеющихся данных об отзывах или покупках, сделанных конкретным пользователем, может быть недостаточно для выдачи надежных рекомендаций. Поэтому в настоящей работе предлагается для извлечения первоначальной информации о предпочтениях пользователя воспользоваться галереей его фотографий. Действительно, в современном мире одними из самых достоверных источников данных о пользователе являются его мобильные устройства. Гигабайты информации, хранящиеся в фотоальбомах пользователей, могут послужить основой для построения профиля интересов. Например, если в галерее пользователя преобладают фотографии, связанные со спортом или здоровым образом жизни, то в путешествии ему будут предложены спортивные товары и места для посещения. Рассматриваемая задача определения предпочтений пользователя по его фотоальбому представляет несомненную практическую значимость, но все ещё недостаточно исследована.

Стоит отметить, что изображения в фотоальбоме зачастую содержат персональную информацию, поэтому накладываются существенные ограничения на прямую передачу данных на сервер и последующую обработку с помощью современных высокоточных методов глубокого обучения [6]. К сожалению, вычислительная эффективность и сложность по затратам памяти таких методов является недостаточной для их реализации даже на современных мобильных устройствах. В связи с этим в настоящее время наблюдается заметная тенденция к разработке эффективных архитектур свёрточных нейронных сетей (англ. convolutional neural network, CNN), которые могут быть использованы непосредственно на мобильном устройстве [7]. Однако их точность в ряде приложений может быть значительно ниже точности CNN с большим числом скрытых слоев.

В настоящей работе предлагается воспользоваться альтернативным подходом, в котором вначале автоматически получается текстовое описание фотографии, а далее полученные описания преобразуются в пользовательские предпочтения. Словарь для получаемых текстовых описаний может быть выбран так, что получаемые аннотации изображений не содержат персональной информации. Поэтому при таком подходе описания фотографий могут обрабатываться и на удаленном вычислительном сервере для повышения надежности решения. Исследованию в этом актуальном направлении и посвящена предлагаемая статья.

1. Постановка задачи

Задача извлечения предпочтений пользователя по его фотоальбому состоит в том, чтобы поступившему на вход фотоальбому (галереи) – множеству фотографий – выделить наиболее интересные для пользователя из некоторого множества номеров категорий $\{1, 2, \ldots, L\}$, где L – общее число категорий. Такими категориями могут являться искусство, спорт, высокие технологии, еда, музыка и другие возможные увлечения и предпочтения пользователя. В общем случае множество категорий, интересных пользователю, является подмножеством множества всех категорий $\{1, 2, \ldots, L\}$.

Предполагается, что для обучения системы для каждой категории задано множество соответствующих ей изображений сцен и событий. Например, для категории «спорт» собирается множество фотографий разнообразных спортивных сцен. Задача анализа предпочтений тогда сводится к известной задаче мультиклассовой (multi-label) классификации набора изображений [2, 8].

Для решения задачи каждое изображение из фотоальбома можно обрабатывать изолированно с использованием известных методов распознавания сцен [9]. При этом для настройки классификатора может применяться собранное обучающее множество. Здесь может использоваться как традиционное дообучение CNN, так и предварительное детектирование объектов (примитивов), из которых состоят сцены [10]. В последнем случае необходимо сформировать алфавит признаков, на основе которых изображения следует относить к той или иной категории [11]. В обоих вариантах результатом анализа предпочтений можно считать частоты встречаемости распознанных сцен, соответствующих каждой категории на всех фотографиях из галереи.

2. Предлагаемый подход

Рассмотрим функциональную схему (рис. 1) предлагаемой системы извлечения предпочтений пользователя по его фотоальбому на основе методов автоматического порождений описаний изображений. Здесь решение поставленной задачи выявления предпочтений пользователя состоит из двух основных этапов. На первом этапе получается текстовое представление каждой из анализируемых фотографий: автоматически формируется некоторое сочетание слов или целое предложение в зависимости от качества снимка, количества изображенных объектов, их взаимодействия, а также наличия всевозможных деталей фона, к примеру зрители матча на стадионе. Полученные текстовые описания характеризуют представленные на изображении объекты и связи между ними, например, взаимное расположение этих объектов. Для этого каждая фотография из фотоальбома поочередно подается на вход блока обработки изображения, где производится изменение масштаба и нормировка. После чего преобразованные фотографии попадают на вход CNN. Выходом блока являются векторы характерных признаков, полученные с предпоследнего слоя. Далее на основе извлеченного вектора признаков выполняется порождение текста. В зависимости от выбранной архитектуры сети выполняется предсказание слова по текущей текстовой последовательности – аннотации каждого изображения. Полученные текстовые описания объединяются в общий текст, описывающий весь фотоальбом.

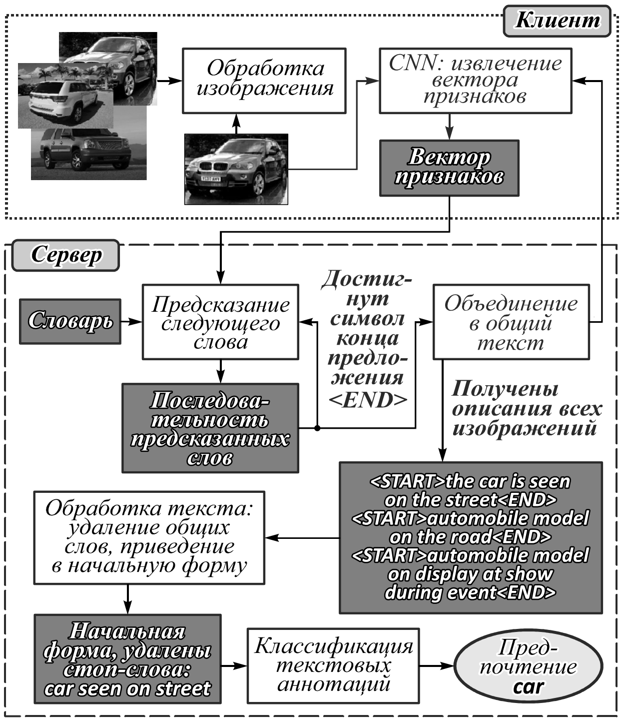

Рис. 1. Функциональная схема предлагаемой системы извлечения интересов пользователя

На втором этапе в заключительном блоке, в котором полученные текстовые аннотации классифицируются и принимается итоговое решение о потенциальной сфере предпочтений пользователя. Рассмотрим три альтернативных способа преобразования полученных аннотаций в пользовательские предпочтения. Наиболее интуитивным способом является подсчет ключевых слов в сформированных текстах. Будем полагать, что слова, встречающиеся в тексте максимальное количество раз, и отвечают за целевое предпочтение для текущего пользователя:

$$l^* = \underset{l=1,\ldots,L}{\arg\max} \sum_{k=1}^{K} I[i_k = l], \tag{1}$$

где k – номер слова в последовательности, l – индекс слова-предпочтения в заданном словаре, I – функция-индикатор. Такой подход является реализацией рассмотренного в предыдущем параграфе варианта решения задачи, в котором детектирование объектов заменяется на порождение текстовых описаний. Алфавит признаков, использующийся для описания каждой категории предпочтений, состоит из ключевых слов, предварительно выбранных для каждой темы экспертом вручную с учётом, например, распределения терминов в теме. В результате этот подход не использует семантическое значение входных слов, а также не выявляет значимые для результата устойчивые словосочетания, что накладывает существенные ограничения на формирование общей темы для группы фотографий.

Как известно, в задачах обработки естественного языка особое место занимают алгоритмы тематического моделирования, которые для заданной коллекции текстов автоматически формируют степень принадлежности всех документов к каждой из тем [12, 13]. Для обучения подобных моделей требуется большая выборка текстов с тематическими метками. В данной работе будет использоваться латентное размещение Дирихле (англ. Latent Dirichlet Allocation, LDA) [14], которое автоматически выделяет различные темы в текстах и слова, наиболее характерные для каждой темы.

Наконец, ещё один способ формирования предпочтений пользователя состоит в автоматической суммаризации текстов с использованием API коммерческой когнитивной системы Watson компании IBM для выделения ключевых фраз в тексте [15, 16, 17]. В частности, использовались сервисы «Natural Language Classifier» и «Natural Language Understanding», которые позволяют выявлять тематическую принадлежность для неструктурированных текстов.

Заметим, что предложенный подход может как полностью выполняться на мобильном устройстве, так и реализовываться с помощью технологии клиент-сервер, в которой характерные признаки каждой фотографии извлекаются непосредственно на устройстве, а порождение текстовых описаний и их преобразование в предпочтения пользователя осуществляется с помощью высокоточных нейросетевых архитектур на вычислительном сервере. Рассмотрим подробнее основной компонент предлагаемого подхода – нейросетевые модели автоматического порождения описаний изображений.

3. Порождение текстовых описаний изображения

Задача автоматического порождения описаний изображений состоит в том, чтобы вновь поступающее на вход изображение X представить как некоторую последовательность слов определенной длины $W = \{\mathbf{w}_{i_1}, \ldots, \mathbf{w}_{i_k}\}$, $i_k \in \{1, \ldots, N\}$. Здесь $\mathbf{w}_{i_k}$ – векторное представление слова с уникальным идентификатором i_k в заранее выбранном словаре, N – общее количество слов в словаре, k – номер слова в последовательности.

Для решения задачи на предварительном этапе для каждого доступного изображения осуществляется извлечение характерных признаков [18]. В настоящий момент наиболее часто для настройки классификатора применяется не доступное обучающее множество, а сверхбольшая коллекция дополнительно собранных изображений. Такая коллекция используется для обучения глубокой CNN [19], состоящей из нескольких чередующихся слоев свертки и подвыборки, выход которых поступает на вход последовательно соединенных полносвязных слоев [6]. Подавая на вход сети изображение X, извлекают его D-мерный вектор характерных признаков $\mathbf{x}$ на выходе из $D >> 1$ значений предпоследнего слоя.

Одним из главных вопросов, возникающих при создании модели для порождения текстовых аннотаций, является выбор нейролингвистической архитектуры, в которой вектор признаков $\mathbf{x}$ преобразуется в текстовое описание W. В современной литературе определяют два основных вида таких сетей, которые по своей структуре отличаются только тем, как CNN соединяется с рекуррентной сетью типа Long-Short Term Memory (LSTM) [20]. С одной стороны, вектор признаков изображения может быть подан на вход рекуррентной нейронной сети совместно с последовательностью слов в процессе обучения модели. Такой тип архитектур называется «injecting» (рис. 2) [20, 21]. В качестве примера такой архитектуры можно привести модель **im2txt**, которая победила на первом соревновании по порождению аннотаций изображений «MS COCO Captions Image Captioning Challenge» в 2015 году [22]. Модель представляет собой пример процесса кодирования-декодирования. Она сначала преобразует входное изображение в векторное представление, а далее переводит результат в текстовое представление.

Другим типом нейролингвистических архитектур являются модели, выполняющие объединение («merge», рис. 3) вектора признаков изображения с выходом рекуррентной сети (LSTM) для генерации текстовых описаний. В отличие от первой группы здесь CNN не является частью рекуррентной нейронной сети, так что последняя обрабатывает исключительно лингвистическую информацию [20, 23]. При этом предполагается, что объёма обучающей выборки текстовых описания достаточно для автоматического извлечения рекуррентной сетью грамматических и семантических связей, поэтому для работы системы (рис. 1) необходимо указать только словарь.

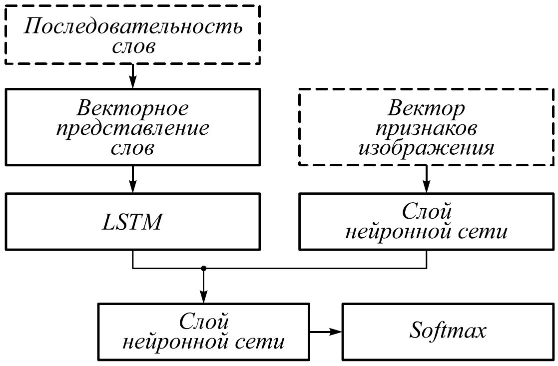

Рис. 2. Архитектура порождения описаний изображений: «injecting»

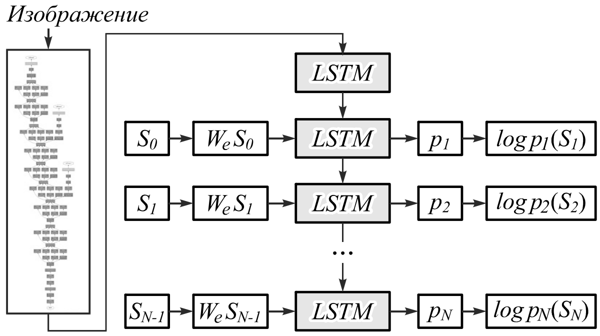

Рис. 3. Архитектура порождения описаний изображений: «merge»

В любой такой архитектуре для порождения текстового описания на вход модели подается вектор признаков **x**, а также последовательность уже предсказанных ранее слов $\{\mathbf{w}_{i_1},\ldots,\mathbf{w}_{i_k}\}$. В самом начале эта последовательность состоит из одного элемента – векторного представления специального символа начала последовательности <BEGIN>. Выход модели в слое Softmax выдает оценку апостериорной вероятности $P(\mathbf{w}_{i_k}|X;\mathbf{w}_{i_1},\ldots,\mathbf{w}_{i_{k-1}})$ появления слова i_k в качестве следующего слова в текстовом описании [6]:

$$P\left(\mathbf{w}_n \middle| X;\mathbf{w}_{\mathbf{i}_1},\ldots,\mathbf{w}_{\mathbf{i}_{k-1}}\right) = \operatorname{softmax} z_n(X) = \\ = \exp(z_n(X)) / \sum_{j=1}^{N} \exp(z_j(X)). \tag{2}$$

Здесь $z_n(X)$ – выход предпоследнего слоя нейронной сети. Обычно решение принимается с помощью «жадного алгоритма» в пользу индекса слова с максимальной апостериорной вероятностью (2):

$$i_k = \underset{n=1,2,\ldots,N}{\operatorname{argmax}} P\left(\mathbf{w}_n \middle| X;\mathbf{w}_{\mathbf{i}_1},\ldots,\mathbf{w}_{\mathbf{i}_{k-1}}\right). \tag{3}$$

Так как на каждом шаге могут быть порождены сразу несколько подходящих по смыслу слов, «жадный» выбор только одного из них не всегда приводит к выбору наиболее вероятной последовательности слов. В таком случае к лучшим результатам обычно приводит применение алгоритма **Beam Search,** определяющего стратегию поиска для выбора нескольких лучших результатов из всех возможных кандидатов [24]. На каждом шаге выделяется фиксированное число B (параметр алгоритма) наиболее вероятных следующих слов для B выбранных последовательностей длины $k-1$, после чего среди получаемых новых B^2 последовательностей выбирается B лучших, соответствующих наибольшим значениям оценок апостериорной вероятности (2).

Процесс рекуррентного порождения текстовых описаний повторяется до тех пор, пока в порожденной последовательности не появится специальный символ окончания описания <END>. Получаемая в результате последовательность слов $\{i_2,\ldots,i_{k-1}\}$, из которой удалены специальные символы начала и конца последовательности, и возвращается в качестве описания фотографии X.

4. Результаты экспериментальных исследований моделей порождения текстовых описаний

Для тестирования качества порождения текстовых описаний изображений использовались следующие наборы данных, специально разработанные для решения подобных задач:

1. **Flickr8k** предоставляет коллекцию 8000 эталонных изображений, для каждого из которых доступно пять различных текстовых описаний [25].
2. **MS COCO Captions** – набор данных от компании Microsoft [26, 27], состоящий из более 500000 размеченных изображений.

В экспериментальном исследовании использовались несколько нейросетевых архитектур порождения текстовых описаний изображений. В качестве базовой применялась предварительно обученная модель **im2txt** (рис. 2) [22], доступная в официальном репозитории TensorFlow.

Кроме того, было обучено несколько архитектур вида (рис. 3). Модель **Flickr8k**, состоящая из CNN VGG16 [19] и LSTM, была обучена на наборе данных **Flickr8k**. Кроме того, для обучения применялся набор данных **Google's Conceptual Captions (CC)** [28], состоящий из трех миллионов ссылок на картинки совместно с их текстовыми аннотациями, автоматически собранных из разных источников в Интернете, а не размеченных вручную. В настоящей работе с использованием этого набора данных было обучено две модели (**CC MobileNet** и **CC Inception**), отличающиеся видом CNN, использовавшейся для извлечения характерных признаков изображения: MobileNet v1 [7] и InceptionV3 [6] соответственно.

Все модели были реализованы средствами языка Python и библиотек Keras и TensorFlow. Их обучение проходило в течение 30 эпох на сервере 64-bit CPU AMD Ryzen Threadripper 1920X 12-Core, 64GB RAM, оснащенном тремя GPU GeForce GTX 1080Ti.

В табл. 1 указаны среднее время преобразования одного изображения в текст на ноутбуке Lenovo T480 Core i5-8350 CPU, 64-разрядной операционной системе Windows 10, 16GB RAM. Кроме того, в этой же таблице приведен размер моделей CNN и LSTM. Здесь и далее лучшие результаты выделены полужирным шрифтом.

Табл. 1. Среднее время создания аннотаций

Модель	**Время, с**	**Размер CNN, Мбайт**	**Размер остальной части модели, Мбайт**
Flickr8k	6,03	528	166
Im2txt	**1,19**	90	**142**
CC MobileNet	**1,67**	**53**	257
CC Inception	3,09	90	259

Наиболее быстрыми архитектурами оказались CC MobileNet и Im2txt. В основу CC MobileNet заложена легковесная и вычислительно эффективная сверточная сеть MobileNet. Размер модели и скорость выполнения операций накладывает существенное ограничение на внедрение архитектуры Flickr8k из-за существенных затрат, связанных с использованием сети VGG16.

Табл. 2. Оценки качества «жадного» алгоритма порождения текстовых описаний для тестового набора Flickr8k

Метрика	**Flickr 8k**	**im2txt**	**CC MobileNet**	**CC Inception**
BLEU-1	**0,548**	0,433	0,334	0,348
BLEU-2	**0,294**	0,187	0,110	0,118
BLEU-3	**0,197**	0,118	0,041	0,039
METEOR	**0,233**	0,201	0,197	0,199
ROUGE-L	**0,358**	0,327	0,302	0,305
CIDEr	**0,721**	0,682	0,653	0,662

Оценка качества работы моделей выполнялась с помощью метрик Corpus BLEU-n [29], METEOR [30], ROUNGE-L [30] и CIDEr [31]. В табл. 2 представлены результаты оценки точности порождения описаний на основе «жадного» алгоритма (3) для тестовой выборки набора данных Flickr8k. Результаты метода Beam Search приведены в табл. 3.

Здесь модель Flickr8k продемонстрировала наилучшие результаты по каждой из метрик. Эта сеть была обучена на тренировочной выборке Flickr8k, поэтому словарь модели содержит слова, близкие к размеченным описаниям в тестовом наборе данных. Обученные на других наборах данных модели также показали достаточно высокую точность. Например, разница между CC MobileNet и CC Inception по метрике BLEU-1 составила всего 1 %, но первая архитектура показала лучшие результаты по скорости генерирования текста (табл. 1).

В табл. 4 и 5 приведены результаты генерирования аннотаций для тестового набора данных MS COCO Captions с использованием «жадного» алгоритма (3) и алгоритмом Beam Search соответственно.

Табл. 3. Оценки качества алгоритма Beam Search порождения текстовых описаний для тестового набора Flickr8k

Метрика	**Flickr 8k**	**im2txt**	**CC MobileNet**	**CC Inception**
		$\boldsymbol{B=3}$		
BLEU-1	0,542	0,442	0,338	0,33
BLEU-2	0,268	0,189	0,116	0,122
BLEU-3	0,183	0,120	0,139	0,038
METEOR	0,145	0,192	0,132	0,139
ROUGE-L	0,354	0,321	0,317	0,303
CIDEr	0,695	0,591	0,584	0,592
		$\boldsymbol{B=5}$		
BLEU-1	0,531	0,343	0,296	0,315
BLEU-2	0,252	0,147	0,102	0,119
BLEU-3	0,171	0,058	0,036	0,04
METEOR	0,137	0,132	0,129	0,135
ROUGE-L	0,308	0,107	0,199	0,201
CIDEr	0,644	0,479	0,481	0,484

Здесь метод Beam Search продемонстрировал небольшое увеличение точности для моделей CC MobileNet и CC Inception. Например, поиск трех наиболее вероятных предложений на 0,2 % точнее, чем поиск по максимальной вероятности.

Табл. 4. Оценки качества «жадного» алгоритма порождения текстовых описаний для тестового набора MS COCO Captions

Метрика	**Flickr 8k**	**im2txt**	**CC MobileNet**	**CC Inception**
BLEU-1	0,39	**0,724**	0,37	0,383
BLEU-2	0,117	**0,557**	0,214	0,229
BLEU-3	0,01	**0,413**	0,148	0,159
METEOR	0,165	**0,252**	0,185	0,19
ROUGE-L	0,258	**0,535**	0,312	0,32
CIDEr	0,314	**0,513**	0,347	0,377

Табл. 5. Оценка качества алгоритма Beam Search порождения текстовых описаний для тестового набора MS COCO Captions

Метрика	**Flickr 8k**	**im2txt**	**CC MobileNet**	**CC Inception**
		$\boldsymbol{B=3}$		
BLEU-1	0,315	**0,708**	0,37	**0,388**
BLEU-2	0,215	**0,541**	0,214	**0,229**
BLEU-3	0,116	**0,405**	**0,293**	0,257
METEOR	0,309	**0,253**	0,384	**0,391**
ROUGE-L	0,25	**0,527**	0,3	**0,313**
CIDEr	0,214	**0,523**	0,244	**0,271**
		$\boldsymbol{B=5}$		
BLEU-1	0,287	**0,701**	0,375	**0,392**
BLEU-2	0,116	**0,536**	0,218	**0,23**
BLEU-3	0,01	**0,39**	0,152	**0,158**
METEOR	0,15	**0,251**	0,186	**0,192**
ROUGE-L	0,236	**0,514**	0,301	**0,316**
CIDEr	0,312	**0,517**	0,345	**0,371**

Худшее качество порождения текстовых описаний соответствует модели Flickr8k. Также заметно увеличение точности по каждой из метрик для модели im2txt, которая была обучена на подмножестве MS

COCO Captions. Разница в 3 % по метрике BLEU-1 с результатами из табл. 3 говорит о том, что эта модель хорошо обучена на словаре из своего набора данных, но значительно хуже работает с изображениями из других предметных областей. Наконец, модели CC MobileNet и CC Inception показали стабильные результаты, поэтому их применение на практике может считаться наиболее приемлемым. Более глубокая сеть Inception ожидаемо оказалась лучше на 0,5 – 1 % большинства моделей по каждой из метрик качества.

5. Результаты тестирования алгоритмов. Предпочтения пользователя

Для тестирования качества оценивания пользовательских предпочтений был собран собственный набор данных **User Preferences Dataset (User Dataset)**, состоящий из десяти категорий интересов некоторых пользователей, а именно: природа, спорт, музыка, театр, мода, собаки, кошки, путешествия, еда. Для каждой категории интереса более 1000 изображений были загружены вручную или взяты из известных наборов данных, принадлежащих целевой тематике. С помощью этого набора данных проводилось сравнение описанных ранее алгоритмов поиска ключевых слов (1), LDA и API Watson. На вход алгоритмов поиска ключевых слов (1) и LDA подаются векторные представления всех слов в текстовых описаниях, для которых удалены стоп-слова и выполнен стемминг [33]. На вход Watson API подавался обычный текст, использовался метод NaturalLanguageUnderstandingV1, который возвращает ответ на запрос и сохраняет полученные оценки в формате JSON. Алгоритм LDA возвращает вероятности близких слов, принадлежащих потенциальной теме. Он был обучен на 80 % сформированных описаний для набора данных User Dataset по каждой из моделей, поэтому для LDA далее представлены результаты обработки аннотаций из оставшейся выборки.

Использовались описания, сформированные с помощью «жадного» алгоритма, максимизирующего оценки апостериорных вероятностей слова в последовательности (3), который, как показано в предыдущем параграфе, практически не уступает по точности методу Beam Search, но при этом получает текстовое описание намного быстрее. В табл. 6 приведены примеры текстовых описаний, порожденных использовавшимися в эксперименте моделями, на вход которых поступило изображение на рис. 4.

Табл. 7 и 8 отражают предпочтения по нескольким категориям, выявленные моделями im2txt и CC MobileNet соответственно.

Согласно приведенным табл. 7 и 8, лучшим из рассматриваемых методов определения предпочтений оказался сторонний функционал, предоставленный Watson API. Алгоритм поиска ключевых слов также достаточно точно определил интересы пользователей для текстов, сформированных сетями обеими моделями. Наиболее достоверные результаты получаются с использованием модели, обученной на большом наборе данных Google’s Conceptional Captions (CC MobileNet). В то же время качество извлекаемых с помощью LDA тем-предпочтений оказалось не удовлетворительным.

Рис. 4. Примеры изображений с пользовательскими предпочтениями

Табл. 6. Аннотации, полученные для изображения (рис. 4)

Модель	Порожденный текст
Flickr8k	a dog is on the street
Im2txt	a dog on a leash standing next to a bicycle
CC MobileNet	dog on the street in the city
CC Inception	person and his dog on the street

Табл. 7. Примеры предпочтений, извлеченных для модели im2txt

Предпочтение	Ключевые слова	LDA	Watson API
nature	• Mountain • Back-ground	• Mountain • Back-ground • Forest • Cloud • person	/sports/boat racing (score: 0.99)
sport	• play • tennis	• player • game • team • snowbord • train	/sports/tennis (score: 0.99) /sports/skateboarding (score: 0.96)
food	• plate • food	• food • cake • cook • tabl • view	• /food and drink/food (score: 0.99) • /food and drink/food/grains and pasta (score: 0.96) • /food and drink/gastronomy /slow food (score: 0.95)

Для того, чтобы количественно оценить точность работы предлагаемых алгоритмов, было решено обучить классификатор ближайшего соседа. Для простоты предполагалось, что рассматриваемый пользователь интересуется только одной конкретной областью, поэтому в качестве решения выбирался только один класс. Общая выборка набора данных User Dataset была разбита на обучающее и тестовое множества в соотношении 90 / 10 % для каждого класса предпочтений. Далее согласно предлагаемому подходу (рис. 1) каждая фотография преобразовывалась в текст. Все слова в получившихся аннотациях кодировались с помощью *Word2vec* [32] и подавались на вход методу ближайшего соседа. Оценки точности классификации предпочтений представлены в табл. 9.

Табл. 8. Примеры предпочтений, извлеченных для модели CC MobileNet

Предпочтение	**Ключевые слова**	**LDA**	**Watson API**
travel	• View • Beach	• Mountain • Beach • Person • Tree • Solar	• /travel/transports/air travel/airplanes (score: 0.73) • /society/social institution (score: 0.72)
food	• Food • Recipe	• Person • Food • Best • Cake • View	• /food and drink/food (score: 0.98) • /food and drink/desserts and baking (score: 0.98) • /food and drink/cuisines (score: 0.94)
car	• Automobile • Model	• Automobil • Person • Vehicle • Driver • Generat	• /automotive and vehicles/cars (score: 0.99) • /automotive and vehicles/cars/car culture (score: 0.99)

Табл. 9. Точность (%) классификации предпочтений

Модель	**Ключевые слова**	**LDA**	**Watson API**
Flickr8k	81,032	77,963	**83,902**
Im2txt	87,941	78,177	**89,001**
CC MobileNet	88,435	79,362	**89,723**
CC Inception	88,501	79,668	**89,754**

Здесь наилучшее качество определения предпочтений достигается с помощью суммаризации текстов из Watson API, который оказался на 1–2 % точнее остальных алгоритмов. Худшие результаты показал алгоритм LDA, который в среднем оказался на 8 % менее точным, чем другие методы. Описания, сгенерированные с помощью CC MobileNet и CC Inception, классифицируются на 1–7 % точнее по сравнению с остальными нейролингвистическими моделями.

Заключение

В настоящей работе предложен новый подход (рис. 1) к решению задачи определения предпочтений пользователя по его фотоальбому, основанный на применении методов автоматического порождения текстовых описаний изображений. Проведено экспериментальное исследование нескольких нейросетевых методов автоматического порождения текстовых описаний, результаты которого показали преимущества модели (рис. 3), обученной на основе набора данных Google's Conceptional Captions. По результатам исследования методов классификации предпочтений на основе полученных текстовых описаний, наилучшие методы поиска ключевых слов (1) и сервис IBM Watson оказались намного точнее традиционного для подобных задач метода LDA (табл. 9).

В то же время следует отметить необходимость проведения ряда дополнительных исследований. В частности, получаемые текстовые аннотации (табл. 6) существенно обедняют представления изображений по сравнению с традиционными характерными признаками, извлекаемыми с помощью глубокой CNN. Поэтому в будущем целесообразно исследовать применение ансамблей моделей, состоящих из традиционных классификаторов и предлагаемого подхода (рис. 1). Кроме того, необходимо исследовать более сложные методы обработки текстов для достижения точности классификации предпочтений, сравнимой с точностью коммерческого сервиса IBM Watson API. Наконец, необходимо рассмотреть альтернативы реализованному в статье простейшему способу агрегации всех описаний, объединяемых в единый текст для последующего извлечения предпочтений.

Благодарности

Статья подготовлена в результате проведения исследования (№ 19-04-004) в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ)» в 2019 г. и в рамках государственной поддержки ведущих университетов Российской Федерации "5-100".

Библиография

1. **Singhal, A.** Use of deep learning in modern recommendation system: A summary of recent works [Electronical Resource] / A. Singhal, P. Sinha, R. Pant // arXiv preprint arXiv:1712.07525. – 2017. – URL: https://arxiv.org/abs/1712.07525 (request date 4.12.2019).
2. **Demochkin, K.V.** Visual product recommendation using neural aggregation network and context gating / K.V. Demochkin, A.V. Savchenko // Journal of Physics: Conference Series. – 2019. – Vol. 1368, Issue 3. – 032016.
3. **Kharchevnikova, A.S.** Neural networks in video-based age and gender recognition on mobile platforms / A.S. Kharchevnikova, A.V. Savchenko // Optical Memory and Neural Networks (Information Optics). – 2018. – Vol. 27, Issue 4. – P. 246-259.
4. **Grechikhin, I.** User modeling on mobile device based on facial clustering and object detection in photos and videos / I. Grechikhin, A.V. Savchenko. – In: Proceedings of the iberian conference on pattern recognition and image analysis (IbPRIA) / ed. by A. Morales, J. Fierrez, J. Sánchez, B. Ribeiro. – Cham: Springer, 2019. – P. 429-440.

5. **Rassadin, A.G.** Scene recognition in user preference prediction based on classification of deep embeddings and object detection / A.G. Rassadin, A.V. Savchenko. – In: Proceedings of international symposium on neural networks (ISNN) / ed. by H. Lu, [et al.]. – Springer Nature Switzerland AG, 2019. – P. 422-430.
6. **Szegedy, C.** Going deeper with convolutions / C. Szegedy // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2015. – P. 1-9.
7. **Howard, A.G.** MobileNets: Efficient convolutional neural networks for mobile vision applications [Electronical Resource] / A.G. Howard, M. Zhu, B Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam. – arXiv preprint arXiv:1704.04861. – 2017. – URL: https://arxiv.org/abs/1704.04861 (request date 4.12.2019).
8. **Wang, R.** Covariance discriminative learning: A natural and efficient approach to image set classification / R. Wang, H. Guo, L.S. Davis, Q. Dai // IEEE Conference on Computer Vision and Pattern Recognition. – 2012. – P. 2496-2503.
9. **Wang, L.** Transferring deep object and scene representations for event recognition in still images / L. Wang, Z. Wang, Y. Qiao, L. Van Gool // International Journal of Computer Vision. – 2018. – Vol. 126, Issues 2-4. – P. 390-409.
10. **Xiong, Y.** Recognize complex events from static images by fusing deep channels / Y. Xiong, K. Zhu, D. Lin, X. Tang // Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). – 2015. – P. 1600-1609.
11. **Фурман, Я.А.** Точечные поля и групповые объекты / Я.А. Фурман, А.А. Роженцов, Р.Г. Хафизов, Д.Г. Хафизов, А.В. Кревецкий, Р.В. Еруcланов; под ред. Я.А. Фурмана. – М: Физматлит, 2014. – 440 с. – ISBN: 978-5-9221-1604-6.
12. **Vorontsov, K.** Additive regularization of topic models / K. Vorontsov, A. Potapenko // Machine Learning. – 2015. – Vol. 101. – P. 303-323.
13. **Rosen-Zvi, M.** The author-topic model for authors and documents / M. Rosen-Zvi // Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. – 2004. – P. 487-494.
14. **Blei, D.M.** Latent Dirichlet allocation / D.M. Blei, A.Y. Ng, M.I. Jordan // Journal of Machine Learning Research. – 2003. – Vol. 3. – P. 993-1022.
15. **Ferrucci, D.A.** Introduction to “this is Watson” / D.A. Ferrucci // IBM Journal of Research and Development. – 2012. – Vol. 56, Issue 3.4. – P. 1:1-1:15.
16. **Lally, A.** Question analysis: How Watson reads a clue / A. Lally, J. Prager, M. McCord, B. Boguraev, S. Patwardhan, J. Chu-Carroll // IBM Journal of Research and Development. – 2012. – Vol. 56, Issue 3.4. – P. 2:1-2:14.
17. **Fan, J.** Automatic knowledge extraction from documents / J. Fan, A. Kalyanpur, D. Gondek, D. Ferrucci // IBM Journal of Research and Development. – 2012. – Vol. 56, Issue 3.4. – P. 5:1-5:10.
18. **Савченко, А.В.** Тригонометрическая система функций в проекционных оценках плотности вероятности нейросетевых признаков изображений / А.В. Савченко // Компьютерная оптика. – 2018. – Т. 42, № 1. – С. 149-158. – DOI: 10.18287/2412-6179-2018-42-1-149-158.
19. **Simonyan, K.** Very deep convolutional networks for large-scale image recognition [Electronical Resource] / K. Simonyan, A. Zisserman. – arXiv preprint arXiv:1409.1556. – 2014. – URL: https://arxiv.org/abs/1409.1556 (request date 4.12.2019).
20. **Tanti, M.** Where to put the image in an image caption generator / M. Tanti, A. Gatt, K.P. Camilleri // Natural Language Engineering. – 2018. – Vol. 24, Issue 3. – P. 467-489.
21. **Wang, M.** A parallel-fusion RNN-LSTM architecture for image caption generation / M. Wang, L. Song, X. Yang, C. Luo // Proceedings of the IEEE International Conference on Image Processing (ICIP). – 2016. – P. 4448-4452.
22. **Vinyals, O.** Show and tell: A neural image caption generator / O. Vinyals, A. Toshev, S. Bengio, D. Erhan // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2015. – P. 3156-3164.
23. **Kiros, R.** Multimodal neural language models / R. Kiros, R. Salakhutdinov, R. Zemel // Proceedings of the International Conference on Machine Learning (ICML). – 2014. – P. 595-603.
24. **Vijayakumar, A.K.** Diverse beam search: Decoding diverse solutions from neural sequence models [Electronical Resource] / A.K. Vijayakumar, M. Cogswell, R. Selvaraju, Q. Sun, S. Lee, D. Crandall, D. Batra. – arXiv preprint arXiv:1610.02424. – 2016. – URL: https://arxiv.org/abs/1610.02424 (request date 4.12.2019).
25. **Bernardi, R.** Automatic description generation from images: A survey of models, datasets, and evaluation measures / R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, B. Plank // Journal of Artificial Intelligence Research. – 2016. – Vol. 55. – P. 409-442.
26. **Lin, T.Y.** Microsoft COCO: Common objects in context / T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, C. Zitnick // Proceedings of the European conference on computer vision (ECCV). – 2014. – P. 740-755.
27. **Chen, X.** Microsoft COCO captions: Data collection and evaluation server [Electronical Resource] / X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollar. – arXiv preprint arXiv:1504.00325. – 2015. – URL: https://arxiv.org/abs/1504.00325 (request date 4.12.2019).
28. **Sharma, P.** Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning / P. Sharma, N. Ding, S. Goodman, R. Soricut // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). – 2018. – Vol. 1. – P. 2556-2565.
29. **Papineni, K.** BLEU: a method for automatic evaluation of machine translation / K. Papineni, S. Roukos, T. Ward, W.J. Zhu // Proceedings of the 40th annual meeting on association for computational linguistics (ACL). – 2002. – P. 311-318.
30. **Denkowski, M.** Meteor universal: Language specific translation evaluation for any target language / M. Denkowski, A. Lavie // Proceedings of the Ninth Workshop on Statistical Machine Translation. – 2014. – P. 376-380.
31. **Vedantam, R.** CIDEr: Consensus-based image description evaluation / R. Vedantam, C.L. Zitnick, D. Parikh // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2015. – P. 4566-4575.
32. **Goldberg, Y.** Word2Vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method [Electronical Resource] / Y. Goldberg, O. Levy. – arXiv preprint arXiv:1402.3722. – 2014. – URL: https://arxiv.org/abs/1402.3722 (request date 4.12.2019).
33. **Manning, C.D.** Foundations of statistical natural language processing / C.D. Manning, H. Schütze. – MIT Press, 1999.
34. **Харчевникова, А.С.** Свёрточные нейронные сети в задаче распознавания пола и возраста по видеоизображению / А.С. Харчевникова, А.В. Савченко. – В кн.: Сборник трудов IV Международной конференции и молодёжной школы "Информационные технологии и нанотехнологии" (ИТНТ 2018). – Самара: Предприятие "Новая техника", 2018. – С. 916-924.

Сведения об авторах

Харчевникова Ангелина Сергеевна, 1996 года рождения, в 2018 году окончила Национальный исследовательский университет Высшая школа экономики – Нижний Новгород поспециальности «Бизнес-Информатика». В 2020 окончила магистерскую программу «Интеллектуальный анализ данных» в НИУ Высшая школа экономики – Нижний Новгород. С 2018 года работает в компании Intel в должности Machine Learning Engineer. Область научных интересов: распознавание образов, глубокое обучение.
E-mail: *angelina.kharchevnikova@gmail.com* .

Савченко Андрей Владимирович, 1985 года рождения, в 2008 году окончил Нижегородский государственный технический университет им. Р.Е. Алексеева по специальности «Прикладная математика и информатика». В 2010 году защитил диссертацию на соискание ученой степени кандидата технических наук по специальности 05.13.18 «Математическое моделирование, численные методы и комплексы программ». В 2015 г. присвоено ученое звание доцента по специальности 05.13.18. В 2016 году присуждена ученая степень доктора технических наук по специальности 05.13.01 «Системный анализ, управление и обработка информации». В настоящее время работает профессором кафедры информационных систем и технологий и старшим научным сотрудником лаборатории алгоритмов и технологий анализа сетевых структур в Национальном исследовательском университете Высшая школа экономики – Нижний Новгород. Автор более 100 научных работ. Область научных интересов: обработка мультимедийной информации, распознавание образов.
E-mail: *avsavchenko@hse.ru* .

ГРНТИ: 28.23.15
Поступила в редакцию 13 декабря 2019 г. Окончательный вариант – 06 марта 2020 г.

Visual preferences prediction for a photo gallery based on image captioning methods

A.S. Kharchevnikova[1], A.V. Savchenko[1]
[1]National Research University Higher School of Economics, Nizhny Novgorod, Russia

Abstract

The paper considers a problem of extracting user preferences based on their photo gallery. We propose a novel approach based on image captioning, i.e., automatic generation of textual descriptions of photos, and their classification. Known image captioning methods based on convolutional and recurrent (Long short-term memory) neural networks are analyzed. We train several models that combine the visual features of a photograph and the outputs of an Long short-term memory block by using Google's Conceptual Captions dataset. We examine application of natural language processing algorithms to transform obtained textual annotations into user preferences. Experimental studies are carried out using Microsoft COCO Captions, Flickr8k and a specially collected dataset reflecting the user's interests. It is demonstrated that the best quality of preference prediction is achieved using keyword search methods and text summarization from Watson API, which are 8 % more accurate compared to traditional latent Dirichlet allocation. Moreover, descriptions generated by trained neural models are classified 1 – 7 % more accurately when compared to known image captioning models.

Keywords: user modeling, image processing, image captioning, convolutional neural networks.
Citation: Kharchevnikova AS, Savchenko AV. Visual preferences prediction for a photo gallery based on image captioning methods. Computer Optics 2020; 44(4): 618-626. DOI: 10.18287/2412-6179-CO-678.
Acknowledgements: The work was partly funded within the Academic Fund Program at the National Research University Higher School of Economics (HSE University) in 2019 (grant No 19-04-004) and by the Russian Academic Excellence Project «5-100».

References

[1] Singhal A, Sinha P, Pant R. Use of deep learning in modern recommendation system: A summary of recent. Source: ⟨https://arxiv.org/abs/1712.07525⟩.
[2] Demochkin KV, Savchenko AV. Visual product recommendation using neural aggregation network and context gating, J Phys Conf Ser 2019; 1368(3): 032016.
[3] Kharchevnikova AS, Savchenko AV. Neural networks in video-based age and gender recognition on mobile platforms. Opt Mem Neural Network 2018; 27(4): 246-259.
[4] Grechikhin I, Savchenko AV. User modeling on mobile device based on facial clustering and object detection in photos and videos. In: Book: Morales A, Fierrez J, Sánchez J, Ribeiro B, eds. Proceedings of the iberian conference on pattern recognition and image analysis (IbPRIA). Cham: Springer; 2019: 429-440.
[5] Rassadin AG, Savchenko AV. Scene recognition in user preference prediction based on classification of deep embeddings and object detection. In Book: Lu H, et al, eds. Proceedings of international symposium on neural networks (ISNN). Springer Nature Switzerland AG; 2019: 422-430.
[6] Szegedy C. Going deeper with convolutions. Proc CVPR 2015: 1-9.
[7] Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H, MobileNets: Efficient convolutional neural networks for mobile vision applications. Source: ⟨https://arxiv.org/abs/1704.04861⟩.
[8] Wang R. Covariance discriminative learning: A natural and efficient approach to image set classification. IEEE CVPR 2012: 2496-2503.
[9] Wang L, Wang Z, Qiao Y, Van Gool L. Transferring deep object and scene representations for event recognition in still images. Int J Comput Vis 2018; 126(2-4): 390-409.
[10] Xiong Y, Zhu K, Lin D, Tang X. Recognize complex events from static images by fusing deep channels. Proc CVPR 2015: 1600-1609.
[11] Furman YaA, ed. Point fields and group objects [In Russian]. Moscow: "Fizmatlit" Publisher; 2014. ISBN: 978-5-9221-1604-6.
[12] Vorontsov K, Potapenko A. Additive regularization of topic models. Mach Learn 2015; 101: 303-323.
[13] Rosen-Zvi M. The author-topic model for authors and documents. Proc 20th CUAI 2004: 487-494.
[14] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res 2003; 3: 993-1022.
[15] Ferrucci DA. Introduction to "this is Watson". IBM J Res Dev 2012; 56(3.4): 1.
[16] Lally A, Prager J, McCord M, Boguraev B, Patwardhan S, Chu-Carroll J, Question analysis: How Watson reads a clue, IBM J Res Dev 2012; 56(3.4): 2.
[17] Fan J, Kalyanpur A, Gondek D, Ferrucci D. Automatic knowledge extraction from documents. IBM J Res Dev 2012; 56(3.4): 5.
[18] Savchenko AV. Trigonometric series in orthogonal expansions for density estimates of deep image features. Computer Optics 2018; 42(1): 149-158. DOI: 10.18287/2412-6179-2018-42-1-149-158.
[19] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Source: ⟨https://arxiv.org/abs/1409.1556⟩.
[20] Tanti M, Gatt A, Camilleri KP. Where to put the image in an image caption generator. Nat Lang Eng 2018; 24(3): 467-489.
[21] Wang M, Song L, Yang X, Luo C. A parallel-fusion RNN-LSTM architecture for image caption generation. Proc IEEE ICIP2016: 4448-4452.

[22] Vinyals O, Toshev A, Bengio, Erhan D. Show and tell: A neural image caption generator. Proc IEEE CVPR 2015: 3156-3164.

[23] Kiros R, Salakhutdinov R, Zemel R. Multimodal neural language models. Proc ICML 2014: 595-603.

[24] Vijayakumar AK, Cogswell M, Selvaraju R, Sun Q, Lee S, Crandall D, Batra D. Diverse beam search: Decoding diverse solutions from neural sequence models. Source: ⟨https://arxiv.org/abs/1610.02424⟩.

[25] Bernardi R, Cakici R, Elliott D, Erdem A, Erdem E, Ikizler-Cinbis N, Plank B. Automatic description generation from images: A survey of models, datasets, and evaluation measures. J Artif Intell Res 2016; 55: 409-442.

[26] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Zitnick C. Microsoft COCO: Common objects in context. Proc ECCV 2014: 740-755.

[27] Chen X, Fang H, Lin T, Vedantam R, Gupta S, Dollar P, Microsoft COCO captions: Data collection and evaluation server. Source: ⟨https://arxiv.org/abs/1504.00325⟩.

[28] Sharma P, Ding N, Goodman S, Soricut R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics 2018; 1: 2556-2565.

[29] Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics 2002: 311-318.

[30] Denkowski M, Lavie A. Meteor universal: Language specific translation evaluation for any target language. Proc 9th Workshop on Statistical Machine Translation 2014: 376-380.

[31] Vedantam R, Zitnick CL, Parikh D. CIDEr: Consensus-based image description evaluation. Proc IEEE CVPR 2015: 4566-4575.

[32] Goldberg Y, Levy O. Word2Vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. Source: ⟨https://arxiv.org/abs/1402.3722⟩.

[33] Manning CD, Schütze H. Foundations of statistical natural language processing. MIT Press; 1999.

[34] Kharchevnikova AS, Savchenko AV. Convolutional Neural Networks in age/gender video-based recognition. Proceedings of the IV International Conference "Information Technologies and Nanotechnologies" (ITNT 2018). Samara: "Novaja Tehnika" Publisher; 2018: 916-924.

Authors' information

Angelina Sergeevna Kharchevnikova (b. 1996) graduated from National Research University Higher School of Economics, Nizhny Novgorod in 2018, majoring in Business Informatics. Shedefended her Master's Thesis in Data Mining in National Research University Higher School of Economics, Nizhny Novgorod. Shealso has been working in the Intel company as a Machine Learning Engineer since 2018. Research interests include pattern recognition and deep learning. E-mail: *angelina.kharchevnikova@gmail.com* .

Andrey Vladimirovich Savchenko (b. 1985) graduated from N. Novgorod State Technical University in 2002, majoring in Applied Mathematics and Informatics. He defended his PhD in Mathematical Modeling, Numeric Methods and Software Complexes in 2010. He received the Doctor of Science degree in System Analysis, Control and Information Processing in 2016. Currently he works as the professor of Information Systems and Technologies department and senior researcher of the laboratory of Algorithms and Technologies in Network Analysis in National Research University Higher School of Economics, Nizhny Novgorod. He is the co-author of more than 100 scientific papers. Research interests include multimedia processing and pattern recognition. E-mail: *avsavchenko@hse.ru* .

Received December 13, 2019. The final version – March 06, 2020.