УДК 004.8

Задачи и методы автоматического описания изображений

Коршунова К. П.

Постановка проблемы: *задачи автоматического описания изображений – сложные мультимодальные задачи искусственного интеллекта (предполагающие совместную обработку разнородной информации: графической и текстовой). Данный класс задач является сравнительно новым, исследования по теме содержат определенные противоречия: во-первых, среди исследователей отсутствует единство в употреблении различных понятий, терминов и формулировок для обозначения решаемых задач, во-вторых, отсутствует качественная классификация задач и методов их решения, а предлагаемые классификации являются неполными.* ***Цель работы:*** *обзор современных исследований по решению задач автоматического описания изображений, уточнение понятийно-терминологического аппарата, создание классификации задач и методов автоматического описания изображений.* ***Результат:*** *в результате анализа многочисленных исследований выделены следующие классы рассматриваемых задач: аннотирование изображений (описание набором ключевых слов), поиск описаний изображений (поиск наиболее удачного описания в некотором имеющемся множестве) и генерирование описаний изображений (составление нового описания в виде предложения на естественном языке). Предложена следующая классификация методов решения: поисковые (в том числе поиск по изображениям и поиск по описаниям), генеративные и гибридные методы. Проведен обзор наиболее удачных методов, способов и моделей поиска и генерирования описаний, а также сформулированы основные тенденции их развития. Проведен обзор публично доступных баз данных и показателей оценки качества для систем автоматического описания изображений.* ***Практическая значимость:*** *данная работа является первым подробным обзором задач, методов, а также способов, алгоритмов, моделей автоматического описания изображений на русском языке. Результаты исследования могут быть использованы для изучения имеющихся и разработки новых методов решения рассматриваемых задач.*

Ключевые слова: *автоматическое описание изображений, генерирование описаний изображений, поиск описаний изображений, автоматическое аннотирование изображений, глубокие нейронные сети, мультимодальные задачи, машинное обучение, искусственный интеллект.*

Введение

Большинство окружающей нас информации представлено в одном из двух видов: *графическом* (визуальная информация: фото- и видеоизображения) и *текстовом* (письменный или устный текст на естественном языке).

Количество информации обоих видов, с которой сталкиваются пользователи информационных технологий каждый день, растет с постоянно увеличивающейся скоростью. Например, на сегодняшний день ресурс *You Tube* насчитывает более 1 миллиарда пользователей, на его серверах хранятся тысячи видеороликов общей длительностью более 70 млн часов [1].

Библиографическая ссылка на статью:
Коршунова К. П. Задачи и методы автоматического описания изображений // Системы управления, связи и безопасности. 2018. № 1. С. 30–77. URL: http://sccs.intelgr.com/archive/2018-01/02-Korshunova.pdf

Reference for citation:
Korshunova K. P. Automatic Image Captioning: Tasks and Methods. *Systems of Control, Communication and Security*, 2018, no. 1, pp. 30–77. Available at: http://sccs.intelgr.com/archive/2018-01/02-Korshunova.pdf (in Russian).

Однако, несмотря на бо́льшую «информационную емкость» визуальной информации («*a picture is worth a thousand words*»), информация в виде текстов на естественном языке является более распространенной, привычной и удобной для коммуникации между людьми.

В связи с этим в последние годы возрастает интерес к комплексным задачам искусственного интеллекта, которые предполагают совместную обработку разнородной информации: как графической, так и текстовой. Подобные задачи и методы их решения называют разномодальными или мультимодальными (англ. *multimodal*). Под «модальностью» в данном случае понимается тип обрабатываемых данных: текст, графика. Решение мультимодальных задач требует комбинации подходов и принципов интеллектуального анализа информации разной природы, что обуславливает чрезвычайную сложность данного класса задач: в добавление к существующим проблемам анализа графических и текстовых данных возникают новые, связанные с комбинированием разнородной информации.

Интерес к исследованиям в данной области связан с многочисленными практическими потребностями, а также огромным ее потенциалом. Подобные задачи возникают во многих отраслях. Вот далеко неполный список приложений мультимодальных методов интеллектуального анализа фото- и видеоизображений и текста:

– системы управления технологическими процессами (в том числе промышленные роботы, автономные транспортные средства);
– эффективный поиск изображений или видео в больших коллекциях (*content-based image retrieval* (*CBIR*) – поиск графической информации на основе ее содержания);
– помощь слабовидящим людям;
– взаимодействие компьютер-человек;
– виртуальная (дополненная) реальность.

Данная статья посвящена задачам автоматического описания изображений [2] (англ. *Automatic Image Description*, *Image Captioning*), первые попытки решения которых предпринимались еще в 1990-х годах [3, 4]. Система автоматического описания изображений получает на входе фото- или видеоизображение, анализирует его содержание и на выходе выдает текстовое описание данного изображения в виде одного или нескольких предложений на естественном языке (ЕЯ).

Задача автоматического описания изображений является мультимодальной и находится на стыке двух областей анализа данных: теории распознавания образов (*pattern recognition*) и обработки естественного языка (*natural language processing*), – так как предполагает не только распознавание графических образов (идентификацию объектов на изображении, определение их свойств, отношений между ними), но и последующее описание результатов распознавания на ЕЯ.

Сложность решения данной задачи заключается в следующем.

1. *С точки зрения распознавания графических образов* от компьютера требуется понимание содержания изображенных сцен: очевидно, что распознавания отдельных несложных визуальных признаков (цвета, формы, классов изображенных объектов) в данном случае недостаточно. От методов решения подобных задач требуются интеллектуальные способности устанавливать пространственно-временные, причинно-следственные и др. связи и отношения между изображенными объектами.
2. *С точки зрения обработки естественного языка* от компьютера требуется генерирование предложений на естественном языке, с одной стороны, точно описывающих заданное изображение, с другой стороны, грамматически правильных:
   - одно изображение может быть описано на естественном языке огромным количеством способов. В зависимости от конкретной предметной области, важными могут быть разные аспекты, а также может потребоваться разная степень детализации: компьютер должен самостоятельно решать, о каких из изображенных объектов, их свойств и отношений должна идти речь в описании;
   - кроме того, от методов решения подобных задач требуется правильное обращение с многочисленными грамматическими и лексическими особенностями языка.
3. *С точки зрения комбинации теории распознавания образов и обработки естественного языка*:
   - в общем случае в описаниях изображений могут упоминаться объекты, явления, их свойства или отношения, которые непосредственно не изображены (например, на картине с описанием «сильный шторм» показаны только высокие волны, дождь и тучи). На организацию визуальной информации в процессе человеческого зрения оказывают существенное влияние такие факторы, как опыт субъекта, его установки [5], которые были сформированы в результате анализа не только предыдущей визуальной информации, но и всей информации, поступившей из окружающей среды с помощью различных органов чувств. В связи с этим, для составления точного текстового описания изображения только информации об объектах, их свойствах и отношениях недостаточно: требуются некоторые априорные знания о предметной области;
   - в общем случае объекты, их свойства и отношения на изображениях, а также их описания на естественном языке очень разнообразны, поэтому для построения методов решения рассматриваемых задач требуются обширные репрезентативные массивы данных, содержащие изображения и соответствующие им текстовые описания;
   - как графическая, так и текстовая информация на естественном языке очень специфична, разнообразна и далеко не всегда может быть опи-

сана формально или обработана с применением известных математических операций (=, <, > и других). До сих пор не решена фундаментальная проблема автоматического оценивания качества методов генерирования естественного языка [6, 7]. Еще бо́льшую сложность представляет разработка показателя качества работы мультимодальных методов. В случае комплексных интеллектуальных задач, которые с легкостью решаются человеком, но для компьютера представляют большую сложность, лучший способ оценивания – это привлечение экспертов (людей). Однако в случае большого количества данных этот способ может стать слишком дорогостоящим. Поэтому для успешного решения задач автоматического описания изображений требуется наличие качественного показателя оценивания точности работы систем решения рассматриваемых задач.

В виду большого объема статьи материал был декомпозирован на следующие подразделы.

1. Задачи автоматического описания изображений.
2. Методы автоматического описания изображений.
   2.1. Классификации методов.
   2.2. Поисковые методы.
      2.2.1. Поиск по изображениям.
      2.2.2. Поиск по описаниям.
   2.3. Генеративные методы.
   2.4. Гибридные методы.
   2.5. Основные тенденции развития методов автоматического описания изображений.
3. Наборы данных.
4. Оценка качества работы методов автоматического описания изображений.

1. Задачи автоматического описания изображений

В настоящее время среди исследователей, занимающихся рассматриваемыми задачами, отсутствует единство в употреблении различных понятий, терминов и формулировок для обозначения решаемых задач. В различных публикациях по теме применяется множество понятий: *Generating Image Descriptions* (англ. «генерирование описаний изображений»), *Image Captioning* («подпись изображений»), *Image Annotation* («аннотирование изображений»), *Sentence-based Image Annotation* («аннотирование изображений на основе предложений»), *Sentence-based Image Description* («описание изображений на основе предложений»), *Composing Image Descriptions* («составление описаний изображений»), *Framing Image Description* («выработка описаний изображений»), *Retrieval* («поиск») *and Generation* («генерация») *of Image Descriptions* («описаний изображений»).

Несмотря на то, что задача автоматического описания изображений привлекает исследователей уже несколько лет, накопленные знания в данной

предметной области до сих пор не систематизированы, исследования по теме содержат следующие противоречия:

1. Во многих публикациях под одними и теми же терминами понимаются принципиально разные задачи и подходы к их решению. Так, например, термином «*generating*» («генерирование») обозначают как задачу *поиска* наилучшего описания изображения в заданном наборе, так и задачи *составления нового описания* в виде предложения на естественном языке, а также в виде множества ключевых слов.

2. Отсутствует классификация рассматриваемых мультимодальных задач. Лишь в работах [8, 9] представлены два направления в рассматриваемой области анализа данных: *generating image descriptions* (англ. *description* – описание) и *image captioning* (англ. *caption* – заголовок, титр, подпись). Первое (*generating image descriptions*) предполагает формирование предложения на естественном языке, описывающее те и только те объекты, их свойства и действия, которые изображены на входной картинке. Второе направление (*image captioning*) допускает, что выходное описание может содержать информацию о том, что не изображено на картинке непосредственно. То, что может быть описано на естественном языке, гораздо шире, чем то, что может быть «увидено» методами компьютерного зрения [9]. В данном случае описания могут содержать личный, культурный, исторический контекст, связанный с изображением (например, «на картине изображена Мона Лиза»).

3. Предлагаемые классификации методов и моделей решения рассматриваемой задачи ([8, 10, 11]) предполагают принципиально разный взгляд на саму решаемую ими задачу и поэтому относятся в первую очередь к задачам, а не к методам. Так, в [8] методы делятся на *поисковые* и *генеративные*. В первом случае решается поисковая задача (поиск наиболее подходящего описания из рассматриваемой базы изображений и их описаний), во втором – задача генерирования новых предложений на естественном языке на основе информации о заданном изображении.

Подавляющее большинство публикаций, посвященных решению данного класса задач, англоязычные, за исключением единичных исследований на русском языке [12, 13, 14], в которых встречаются термины: аннотация/аннотирование изображений, синтез естественного языка, генерирование текста.

Исходя из всего вышесказанного, предложим классификацию задач автоматического описаний изображений и соответствующих им методов и моделей (рис. 1) и введем собственные термины для их обозначения.

1. *Задача аннотирования изображений* (*annotating*) – задача описания входного изображения набором ключевых слов (из заданного словаря). Поскольку данную задачу можно рассматривать как задачу классификации с пересекающимися классами (когда объекты могут принадлежать одновременно многим классам), то в данной статье она рассмотрена не будет, а основное внимание будет уделено более сложным задачам двух других классов.

2. *Задача поиска описания изображения* (*retrieval*) – выбор наилучшего описания для заданного изображения из описаний, входящих в заданное ограниченное множество. Методы решения данного класса задач – *поисковые*.

3. *Задача генерирования описания изображения* (*generation*) – формирование новых предложений на ЕЯ, описывающих заданное изображение. Методы решения данного класса задач – *генеративные* и *гибридные* (см. п. 2.1).

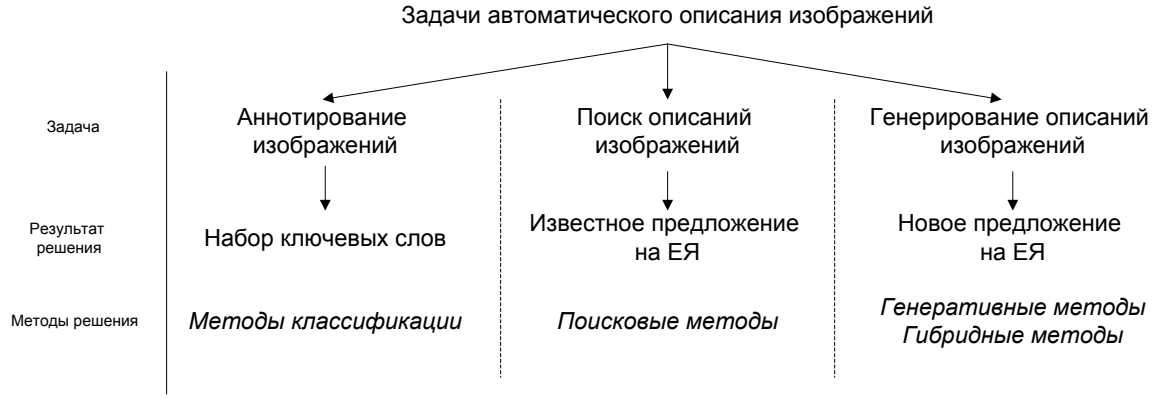


Рис. 1. Классификация задач автоматического описания изображений

Предложенная классификация разделяет задачи аннотирования изображений, поиска и генерирования описаний изображений по двум критериям.

1. Тип информации, получаемой в результате решения задачи: результатом решения первой является одна или несколько меток классов (ключевых слов), а вторых – предложение (-ия) на ЕЯ.

2. Методы, применяемые для решения задач (методы классификации изображений для случая пересекающихся классов, методы поиска описаний изображений, генеративные и гибридные методы соответственно).

Хотя задача аннотирования гораздо проще задач двух других классов, не является мультимодальной и методы ее решения в статье не рассмотрены, в предложенную классификацию она включена, так как во многих публикациях под терминами *generating* (*composing, framing*) *image description* («получение описание изображений») понимается именно она или, наоборот, понятием *annotation* («аннотирование») обозначается поиск или генерация текстового описания.

2. Методы автоматического описания изображений

2.1. Классификации методов

В разных публикациях встречаются различные классификации методов автоматического описания изображений.

В работах [10, 11] предложена следующая классификация:

1) методы на основе шаблонов: результаты работы алгоритмов компьютерного зрения (текстовые метки распознанных объектов и их отношений) «помещаются» в некоторую заранее определенную структуру – шаблон выходного предложения;

2) композиционные (поисковые) методы (*transfer-based approaches*): в качестве выходного описания просто копируется описание наиболее похожего изображения из рассматриваемой базы изображений и их описаний, либо результат работы метода генерируется на основе описаний нескольких похожих изображений;

3) языковые (в т. ч. нейросетевые) методы: работа методов основана на предварительном изучении вероятностного распределения в общем «визуально-языковом» пространстве, в том числе с помощью мультимодальных глубоких нейронных сетей (и их комбинаций).

В работе [8] предложена следующая классификация методов, применяемых для решения рассматриваемых задач:

1) собственно генеративные методы: сначала с помощью алгоритмов компьютерного зрения обнаруживаются определенные визуальные признаки (объекты, свойства, отношения), затем на основе данной информации система генерирования предложений на естественном языке составляет выходное описание;

2) поисковые методы: рассматривают задачу как поисковую (см. предыдущую классификацию).

Данные методы, в свою очередь, делятся на две подгруппы в зависимости от того, каким образом представляется входное изображение и измеряется «похожесть» изображений:

– в первом случае для поиска похожих изображений используется визуальное пространство;
– во втором случае используется общее мультимодальное пространство (включающее представление как изображений, так и текста).

В данной работе предлагается классификация методов автоматического описания изображений, соответствующая классификации задач, представленной в п. 1.

1. *Поисковые методы* – методы решения задач поиска описания изображения. В данном случае перед исследователями стоит задача не сгенерировать описание какого-либо изображения, а найти наиболее подходящее или ранжировать заданные описания. На выходе системы – не новое описание изображения, а одно или несколько наиболее релевантных из заданного множества.

2. *Генеративные методы* – методы решения задач генерирования описания изображения. В этом случае речь идет о собственно генеративных методах, которых предполагают непосредственное генерирование (то есть создание) выходного описания на основе заданного изображения (визуальной информации от него).

3. *Гибридные методы* предполагают комбинацию процедур, используемых в поисковых и генеративных методах. С помощью гибридных методов решается задача генерирования описания изображения, но используются некоторые приемы поисковых методов.

В данной статье приведен обзор существующих поисковых, генеративных и гибридных методов, а также реализующих их способов, алгоритмов, моделей и других решений задач автоматического поиска и генерирования описаний изображений.

2.2. Поисковые методы

В состав поисковых методов входит два следующих метода:

– метод поиска по изображениям (визуальным признакам): для входного изображения осуществляется поиск наиболее похожего изображения (изображений) и в качестве выхода системы используется описание найденного изображения (изображений) [15, 16, 17, 18, 19];
– метод поиска по описаниям (поиск в едином визуально-текстовом пространстве): осуществляется поиск наиболее подходящих описаний из некоторого набора [20, 21, 22, 23, 25, 26, 27].

В первом случае определяется некоторая процедура измерения степени визуального сходства изображений, во втором – процедура измерения степени ассоциации (соответствия) описания и изображения.

2.2.1. Поиск по изображениям

Для осуществления поиска похожих изображений на первом шаге необходимо получить некоторое представление входного изображения в виде набора визуальных признаков, которые затем используются для сравнения с другими изображениями заданного множества. В качестве выходного описания просто копируется описание (описания) наиболее похожего изображения (изображений) из рассматриваемой базы изображений и их описаний.

Метод поиска по изображениям состоит из следующих этапов.

1. Представление изображений (извлечение визуальных признаков).
2. Сравнение изображений.
3. Выбор наиболее похожих изображений и соответствующих им описаний.

Схематичное представление этапов работы метода поиска по изображениям показано на рис. 2 (англ. *The lady in a hat with feathers* – дама в шляпе с перьями, *The girl in a hat* – девушка в шляпе).

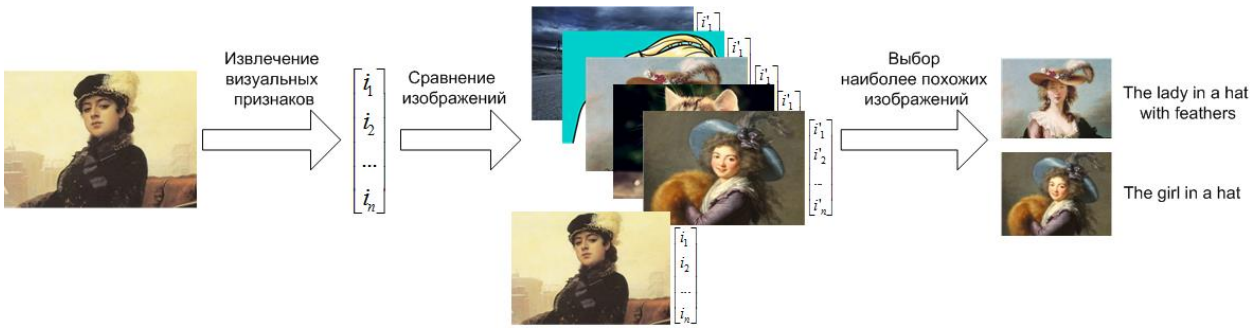


Рис. 2. Этапы работы метода поиска по изображениям

В [15, 18] сходство изображений оценивается с использованием суммы значений глобальных дескрипторов: GIST [28] и «сырых» пикселей сжатого исходного изображения (32×32), – а также с использованием различных оценок сходства содержания изображений по следующим признакам: объекты, матери-

алы, люди (действия), фон, TFIDF-веса (англ. *TF – term frequency, IDF – inverse document frequency*) [29].

В [16] из изображения извлекаются 4 атрибута: GIST, гистограмма направленных градиентов [30, 31], самоподобие, гистограммы геометрического контекста; для измерения сходства изображений применяется евклидово расстояние.

В [17] для представления изображения используется вектор признаков, полученный от сверточной нейронной сети (СНС), для сравнения изображений также применяется евклидово расстояние.

В [19] похожие изображения ищутся с помощью алгоритма *k* ближайших соседей: для заданного изображения выбирается *k* ближайших соседей с использованием GIST-дескрипторов, а также векторов признаков от СНС.

2.2.2. Поиск по описаниям

В случае поиска по описаниям формируется единое визуально-текстовое пространство, в котором представления изображения и описания находятся тем «ближе» друг к другу, чем больше подходит данное описание к данному изображению (и наоборот). Поиск описаний осуществляется в этом едином пространстве с помощью некоторой процедуры измерения релевантности, «близости» изображения и описания (часто расстояния в векторном пространстве).

Метод поиска по описаниям состоит из следующих этапов.

1. Представление изображений, представление описаний.
2. Построение единого мультимодального пространства.
3. Выбор наиболее «близкого» описания для заданного изображения в построенном пространстве.

Схематичное представление этапов работы метода поиска по описаниям показано на рис. 3 (англ. *The cat playing with a ball* – кот, играющий с мячиком, *The girl in front of a mirror* – девушка перед зеркалом).

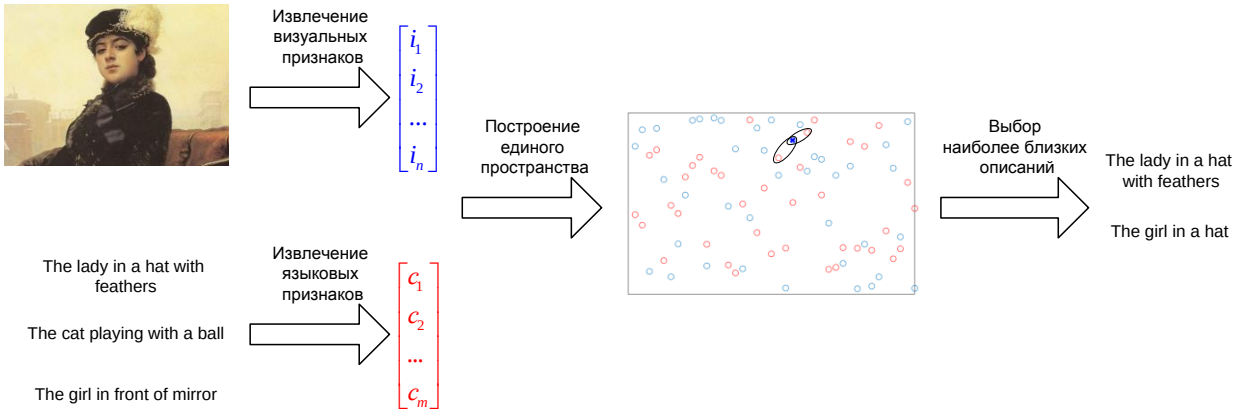


Рис. 3. Этапы работы метода поиска по описаниям

В [20] для представления изображения используются: GIST, детекторы на основе гистограммы направленных градиентов; для получения представления описания применяется система синтаксического разбора *Curran & Clark* [32].

Общее пространство, представляющее собой набор троек:

<объект, действие, сцена>,

строится в виде графовой модели – марковского случайного поля.

В [21] для представления изображения используется вектор признаков, полученный нейронной сетью, а описание представляется в виде вектора, полученного разновидностью рекурсивной нейронной сети. Для выбора наиболее «близкого» описания обучается целевая функция от двух векторов (представлений изображения и описания).

В [22] для описаний используется представление «мешок слов», для изображений строятся гистограммы распределения значений низкоуровневых признаков, учитывающих цвет (координаты цветового пространства CIELAB), текстуру [33] и форму (дескрипторы SIFT). Близость между изображением и описанием определяется средствами канонического корреляционного анализа (*Kernel Canonical Correlation Analysis, KCCA*).

В [23] как изображения, так и текст представляются в виде вероятностного распределения частоты ключевых слов (для изображений – «визуальных ключевых слов», которые определены алгоритмом *k* средних). Для построения общего пространства используется модификация метода опорных векторов *Structural SVM* [34]. При этом поиск осуществляется в неразмеченном множестве предложений на естественном языке (то есть наборе описаний без соответствующих им изображений).

В [24] предлагается способ решения задачи на основе единой нейросетевой модели, которая на выходе выдает оценку степени близости изображения и описания, при этом входом сети служат фрагменты изображения и словосочетания, извлеченные из описания методом синтаксического разбора [35]. Для разбиения заданного изображения на фрагменты используется разновидность сверточной нейронной сети *Region Convolutional Neural Network (R-CNN)* [36].

В [25] в качестве представления изображений используется вектор признаков с выхода последнего слоя СНС, в качестве представления описаний – стандартное представление «мешок слов». Для построения единого векторного пространства используется канонический корреляционный анализ (с нормализацией).

В [26] предложен способ на основе единой нейросетевой модели, включающей сверточные слои для анализа изображения и *LSTM*-слои (*Long-Short Term Memory* [53]) для анализа текста, на выходе которой формируется оценка степени релевантности заданного текста заданному изображению. На вход сети подается также метка класса, к которому относятся изображенные объекты. В процессе обучения сети максимизируется степень релевантности описания и соответствующего ему изображения и минимизируется степень релевантности описания изображениям других классов.

В [27] вводится новая задача локализации выражений (*phrase localization*) как предваряющая задачу поиска описаний. По заданным изображению и соответствующему описанию необходимо определить конкретные фрагменты изображения (ограничивающие прямоугольники) для каждой сущности (объекта) из описания. Степень релевантности описания изображению затем оценивается

как сумма степеней релевантности всех пар фрагмент-выражение. Для решения задачи локализации выражений предложен набор данных, содержащий изображения, их описания, а также соответствия между фрагментами изображений и словосочетаниями из описаний.

2.3. Генеративные методы

Задача генерирования описания на ЕЯ для заданного изображения представляет собой перевод из одного представления (пространства визуальных признаков) в другое (текстовое представление). В этом отношении она сходна с задачей машинного перевода: требуется перевести представление данных в одном языке (модальности) I в представление в другом языке (модальности) C, максимизируя вероятность $p(C|I)$ [37]. Системы решения рассматриваемой задачи, как правило, имеют архитектуру в виде двух подсистем: кодер и декодер. Первая кодирует входное изображение в вектор визуальных признаков, вторая декодирует данный вектор в текстовое представление (описание на ЕЯ).

Генеративные методы состоят из двух этапов.

1. Представление входного изображения (кодер).
2. Генерирование описания на ЕЯ (декодер).

Этапы работы генеративных методов схематично представлены на рис. 4.

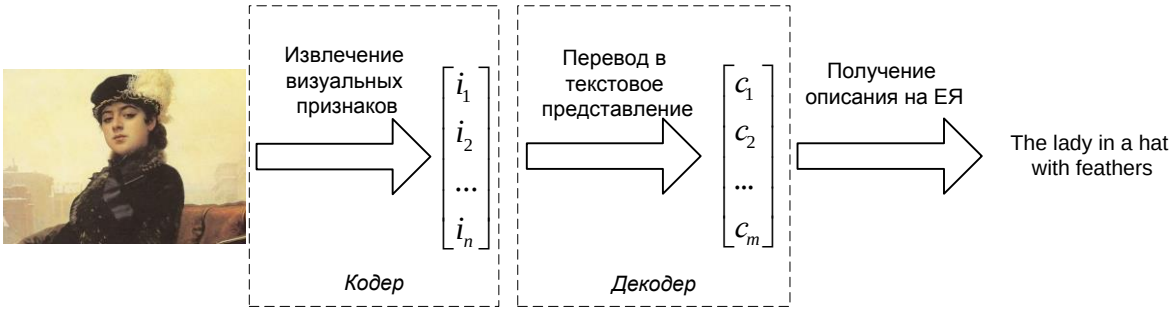


Рис. 4. Этапы работы генеративных методов

В [38] для комплексного представления изображений используется сложная обученная графовая модель *AoG* [39], которая включает как визуальные элементы (примитивы, части, объекты и сцены), так и синтаксические (композиционные) и семантические отношения (категориальные, пространственные, временные и функциональные) между ними. Входное изображение подвергается подробному *полуавтоматическому* разбору, результаты которого с помощью *AoG* конвертируются в смысловое представление в форме специальной онтологии *Web Ontology Language* (*OWL*) [40]. Генерация текста разбивается на две части: планирование (составляет пары *«признак-значение»*, определяя концепты, о которых пойдет речь: объект, его действия, место и время действия и др.) и реализация (генерирование грамматически правильных предложений).

В [41] предлагается способ решения задачи на основе генеративной модели, для построения которой не требуется набор данных, содержащий пары «изображение-описание», а достаточно двух независимых наборов (набор изображений и корпус текстов). По входному изображению определяются объекты и сцены (с помощью известных мощных алгоритмов компьютерного зрения) –

существительные. Затем при помощи обученной на корпусе текстов статистической модели определяются наиболее вероятные действия (глагол) и препозиция (предлог) для данных объектов и сцены. С помощью скрытой марковской сети получается наиболее вероятная четверка:

«объект – действие – сцена – препозиция»,

которая переводится в предложение на ЕЯ (с учетом основных грамматических правил языка).

В [42] для генерирования описания используются тройки вида

((*свойство*1, *объект*1), *препозиция,* (*свойство*2, *объект*2)),

составленные для каждой пары объектов, найденных на изображении известными методами компьютерного зрения. С помощью *n*-граммной модели (модели, которая рассчитывает вероятность последнего слова *n*-граммы – последовательности из *n* слов, – если известны все предыдущие) *Google Web 1T data* определяется несколько предложений-кандидатов, из которых затем составляется наилучшее.

В [43] основой способа автоматического описания изображений *The Midge* служит многоэтапный процесс построения синтаксического дерева на основе информации об объектах, их атрибутах, действиях и пространственных отношениях, полученной алгоритмами компьютерного зрения, а также на основе статистической информации о совместном использовании слов в предложениях на ЕЯ.

В [44] введено понятие представления визуальных зависимостей (*visual dependency representation*, *VDR*). *VDR* для входного изображения представляет собой ациклический граф, который описывает структуру изображения, главным образом именно геометрические зависимости между фрагментами (выше, ниже, за, перед, напротив и др.). На основе данного графа строится описание изображения по заданному шаблону. В работе [45] авторы снова обращаются к представлению визуальных зависимостей. Предлагается автоматический способ построения *VDR*-графа для изображения: с помощью модели распознавания объектов (R-CNN [36]) из входного изображения извлекаются объекты, а затем устанавливаются пространственные отношения между ними (выше, ниже, перед, вокруг и др.). Данный граф используется для генерирования текста на ЕЯ по заданному шаблону.

В [46] из входного изображения с помощью набора различных классификаторов извлекаются объекты, их атрибуты и пространственные отношения между ними. Затем используется модель условного случайного поля (*conditional random field*) для вывода нескольких троек вида

<(*свойство*1, *объект*1), *препозиция,* (*свойство*2, *объект*2)>.

Далее применяется *n*-граммная модель для генерации описания (или описание составляется на основе заданного шаблона).

В [47] регионам изображения, представленным векторами признаков от сверточной нейронной сети, ставятся в соответствия метки классов – отдельные слова. С помощью статистической языковой модели на основе критерия максимизации энтропии из данного набора слов генерируются предложения на ЕЯ. Затем предложения ранжируются с использованием нейросетевой модели, обу-

ченной для определения степени релевантности описания изображению. В качестве выходного описания выбирается наилучшее.

В [48] для генерирования текста на ЕЯ используется билинейная модель (*Log-Bilinear Model* [49]), представляющая собой нейронную сеть прямого распространения с одним скрытым слоем, которая предсказывает следующее слово в контексте предыдущих слоев. В модель добавлены дополнительные входы – признаки изображения (полученные от сверточной сети) – для генерирования описания заданного изображения. Модель обучается стандартным алгоритмом «обратного распространения ошибки».

В [50] генерируются описания интерьеров, поэтому главным образом учитываются пространственные отношения между объектами. Изображение анализируется известными методами распознавания объектов, после чего для входного изображения строится граф, отражающий объекты, их атрибуты и пространственные отношения между ними. По данному графу генерируются выходные предложения с учетом грамматических особенностей ЕЯ.

Начиная с 2014 г. началась активная разработка генеративных методов и способов решения задачи автоматического описания изображений на основе глубоких нейронных сетей: сверточных (СНС) для представления изображений (кодер) и рекуррентных (РНС) для генерирования предложений на ЕЯ (декодер). Основой метода/способа решения задачи автоматического описания изображений является сложная нейросетевая модель; новизна предлагаемого решения главным образом определяется новизной архитектуры предлагаемой модели, а также алгоритмов ее обучения и функционирования.

В [51] предложена мультимодальная генеративная модель на основе простой РНС: изображение кодируется с помощью СНС, затем вектор представления изображения (выходы предпоследнего слоя СНС) подается на вход РНС. РНС обучается для предсказания очередного слова описания в зависимости от предыдущих сгенерированных слов. Информация об изображении при этом учитывается только на первой итерации работы РНС. Модель представлена на рис. 5 (англ. *straw hat* – соломенная шляпа).

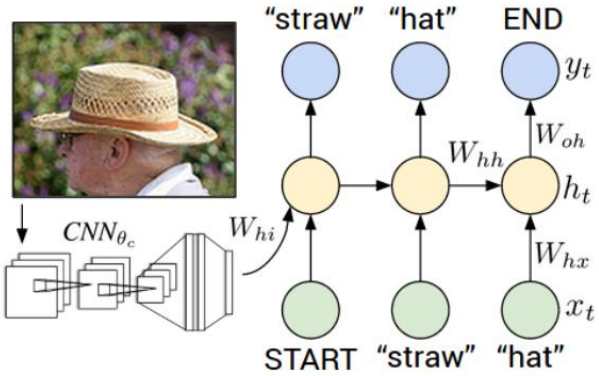


Рис. 5. Генеративная модель на основе простой РНС

В [37, 52] предлагается модель аналогичной архитектуры (рис. 6), но для генерирования текста используется разновидность РНС *LSTM*.

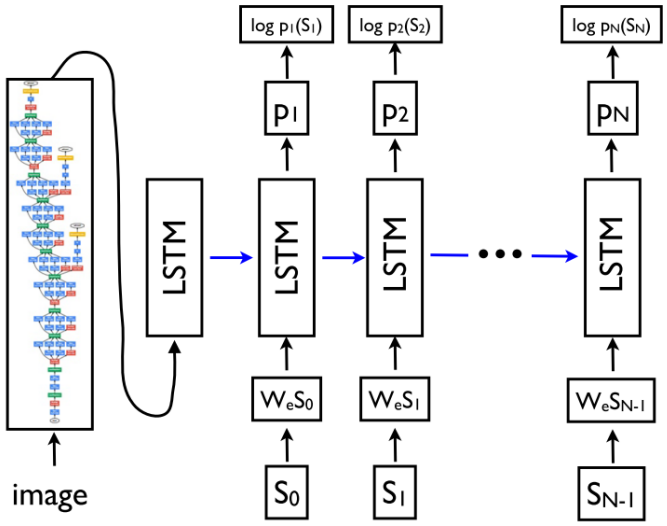


Рис. 6. Генеративная модель на основе *LSTM*

В [54, 55] предложены аналогичные модели (СНС + *LSTM*), но информация об изображении при этом учитывается на каждой итерации работы РНС. В структуру РНС вводятся скрытые переменные состояния (для возможности решения обратной задачи – генерирования признаков изображения по заданному описанию на ЕЯ). Для обучения используется алгоритм «обратного распространения ошибки сквозь время» (*Backpropagation Through Time* (*BPTT*) [56]).

В [57, 58] предлагается единая модель, состоящая из СНС для извлечения признаков изображения, РНС для генерирования очередного слова предложения на ЕЯ в зависимости от предыдущих слов, а также промежуточного мультимодального слоя, который, используя выходы предыдущих частей, предсказывает наиболее вероятное слово в описании входного изображения. Модель обучается алгоритмом «обратного распространения ошибки». Схематичное изображение модели представлено на рис. 7 (англ. *Embedding* – представление, *Fully Connected* – полносвязанный слой, *Deep Image Feature Extraction* – извлечение признаков изображения на основе глубокой сети, *Image Feature* – вектор признаков изображения).

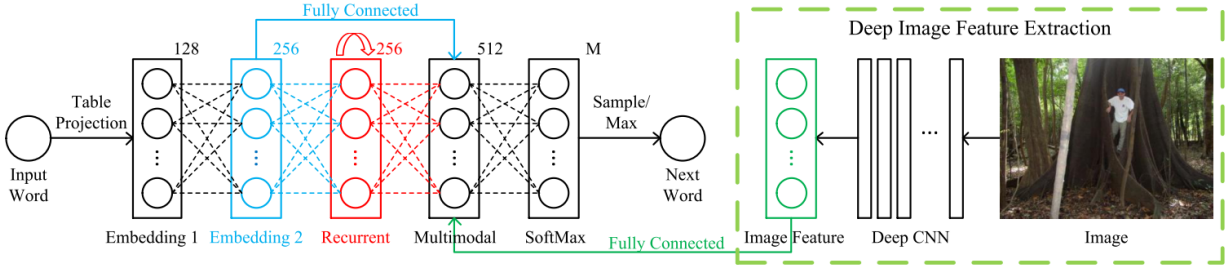


Рис. 7. Генеративная модель СНС+РНС
с промежуточным мультимодальным слоем

В [10] СНС применяется для представления изображений, РНС (*LSTM*) – для представления предложений, на основе чего строится общее мультимодальное векторное пространство. Для генерирования описаний применяется нейросетевая модель (*Structure-content neural language model*, *SC-NLM*), исполь-

зующая как вектор представления изображения в данном пространстве, так и дополнительную информацию о структуре предложений на ЕЯ. Данный способ решения задачи проиллюстрирован на рис. 8 (англ. *Steam ship at the dock* – пароход на пристани, *Multimodal space* – мультимодальное пространство, *content* – содержание, *structure* – структура, *context* – контекст).

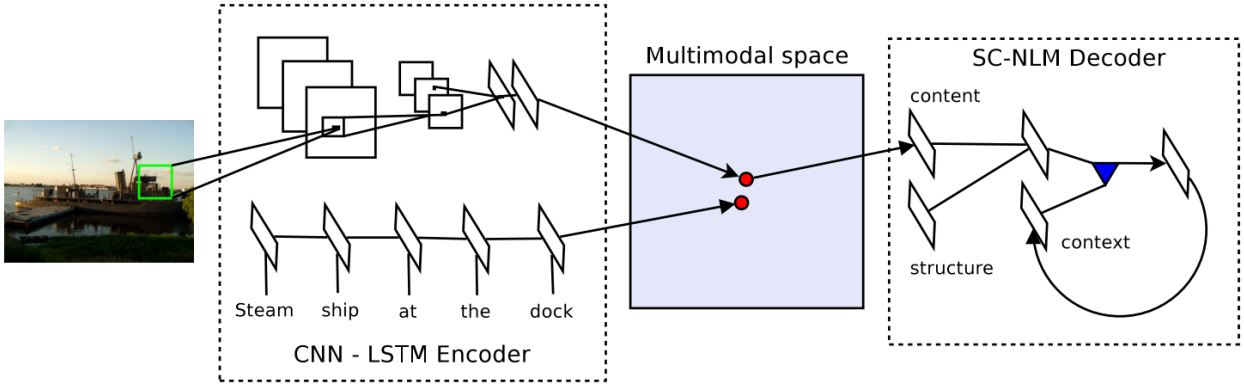


Рис. 8. Способ решения задачи на основе общего мультимодального пространства и нейросетевой модели *SC-NLM*

В [59] предлагается единая модель, названная сверточно-рекуррентной (*Long-term recurrent convolutional network* (*LRCN*) *model*): СНС обрабатывает изображение и вектор его представления подает на вход РНС (*LSTM*). Обучение обеих составных частей модели осуществляется в один этап (*end-to-end*). Данная модель решает также связанные задачи распознавания и описания видео (последовательности изображений). Единая сверточно-рекуррентная модель представлена на рис. 9 (англ. *Visual Features* – визуальные признаки, *Sequence Leqrning* – генерирование последовательностей).

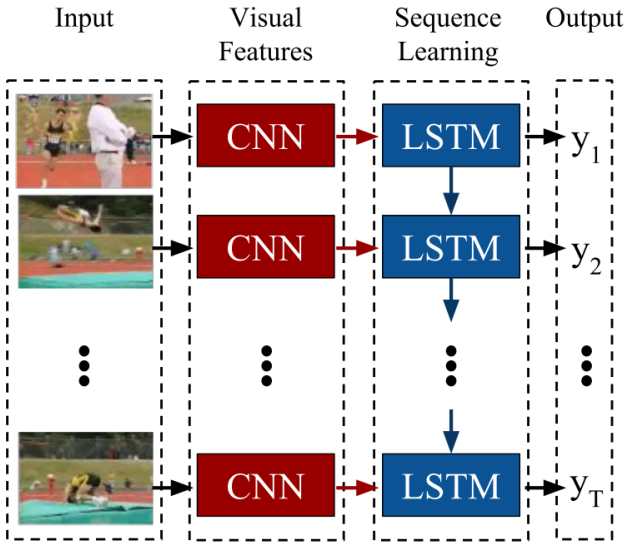


Рис. 9. Единая сверточно-рекуррентная модель

В [60] решается задача, объединяющая задачу описания изображения и локализации, *Dense Captioning* (англ. *dense* – плотный, густой) – генерирование описаний смысловых фрагментов изображений: на входном изображении необходимо выделить некоторые фрагменты и описать каждый из них текстом на

ЕЯ. Предложенная в [60] классификация мультимодальных задач представлена на рис. 10 (англ. *Whole image* – целое изображение, *Image regions* – фрагменты изображения, *Single label* – одиночная метка, *Sequence* – последовательность, *Density* – плотность, *Complexity* – сложность, *Classification* – задача классификации, *Detection* – задача идентификации, *Skateboard* – скейтборд, *A cat riding a skateboard* – кот едет на скейтборде, *Orange spotted cat* – оранжевый пятнистый кот, *Skateboard with red wheels* – скейтборд с красными колесами, *Brown hardwood flooring* – коричневый деревянный пол).

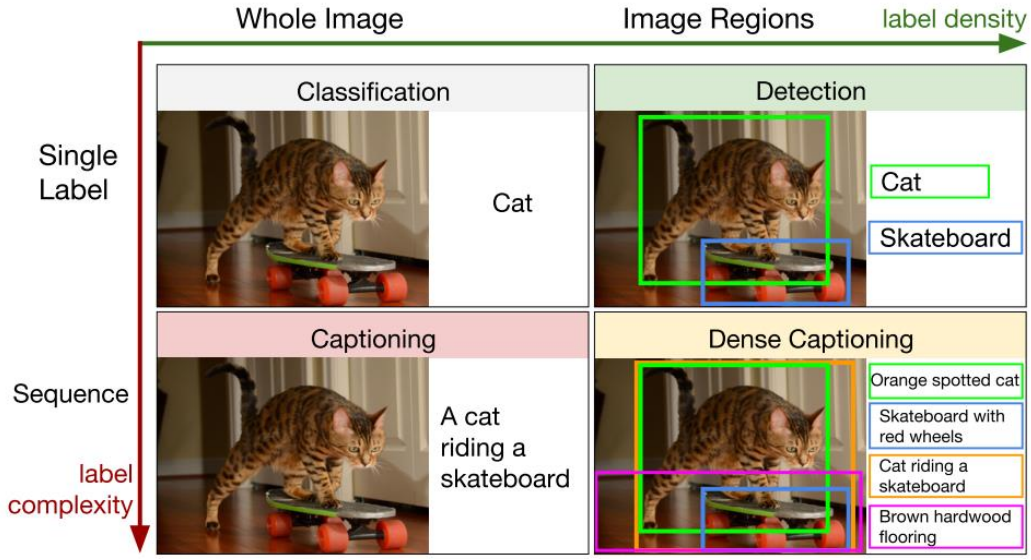


Рис. 10. Классификация мультимодальных задач анализа изображений

Для решения введенной задачи используется единая модель, состоящая из трех частей: СНС (архитектуры VGG-16 [61]), слоя локализации (который предсказывает фрагменты изображения, содержащие ключевые объекты) и РНС (*LSTM*) для генерации текста. Данная модель представлена на рис. 11 (англ. *Localization Layer* – слой локализации, *Striped gray cat* – полосатый серый кот, *Cats watching TV* – коты смотрят телевизор).

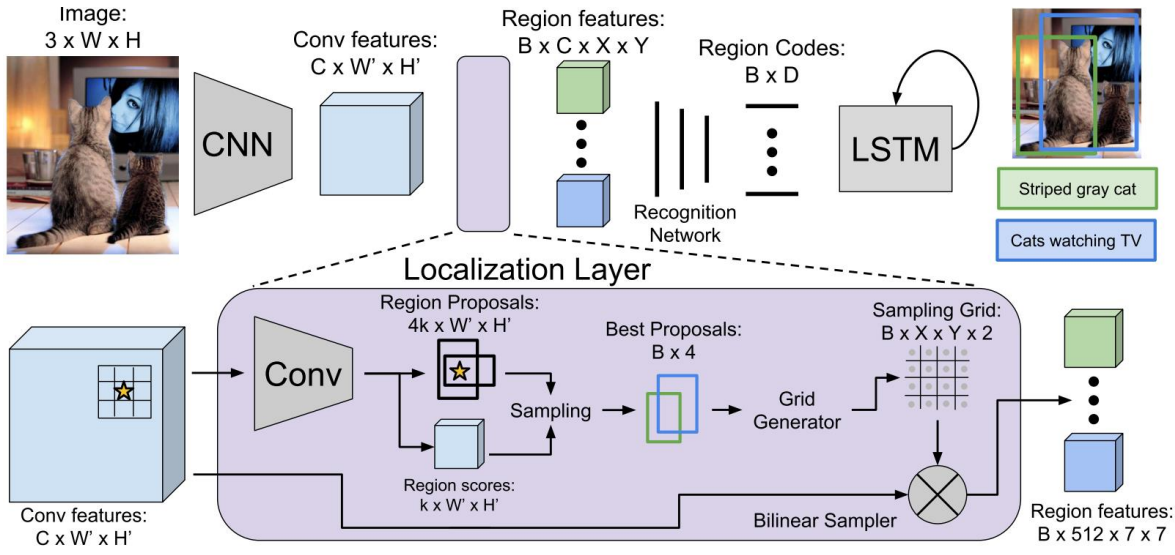


Рис. 11. Модель решения задачи генерирования описаний смысловых фрагментов изображений

В [11] предложена разновидность РНС *LSTM-A* (*Long Short-Term Memory with Attributes*), которая генерирует текст, учитывая не только вектор представления изображения, но и вектор его атрибутов. В качестве атрибутов используются текстовые метки (1000 наиболее часто встречающихся слов из обучающей выборки). Для обучения детекторов атрибутов применяются алгоритмы «с незначительным привлечением учителя» (*weakly-supervised learning*) [62].

В [63] для генерирования текста предложена двунаправленная разновидность *LSTM*-сети (*Bidirectional LSTM, Bi-LSTM*). Модель представлена на рис. 12 (англ. *three footballer in a tackle* – схватка трех футболистов).

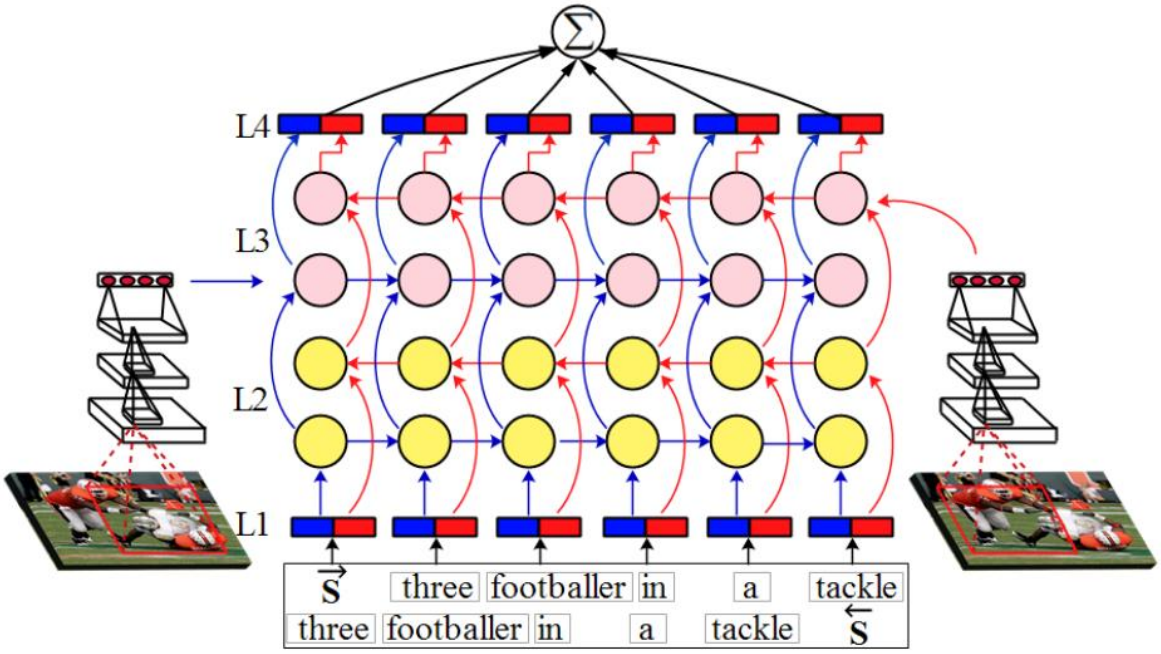


Рис. 12. Решение задачи на основе двунаправленной разновидности *LSTM*-сети

В отличие от однонаправленной *LSTM*-сети предложенная разновидность учитывает не только предыдущий, но и будущий контекст. Кроме того, предлагается увеличить глубину сети добавлением промежуточных полносвязанных слоев с целью упрощения процедуры обучения.

В [64] предложено включить в декодер, представленный *LSTM*-сетью, небольшую нейронную подсеть, называемую *механизмом внимания* (*attention mechanism* [65]), которая принимает на входе предыдущее состояние РНС, а также вектор представления изображения. Выход данной подсети обобщает информацию обо всем исходном изображении, но с разным акцентом на разных фрагментах (объектах). Таким образом, на каждой итерации работы *LSTM* исходное изображение маскируется для того, чтобы при генерации очередного слова учитывались разные фрагменты изображения. Механизм внимания обучается с помощью стохастического градиентного спуска. Способ проиллюстрирован на рис. 13 (англ. *A bird flying over a body of water* – птица, пролетающая над водным пространством).

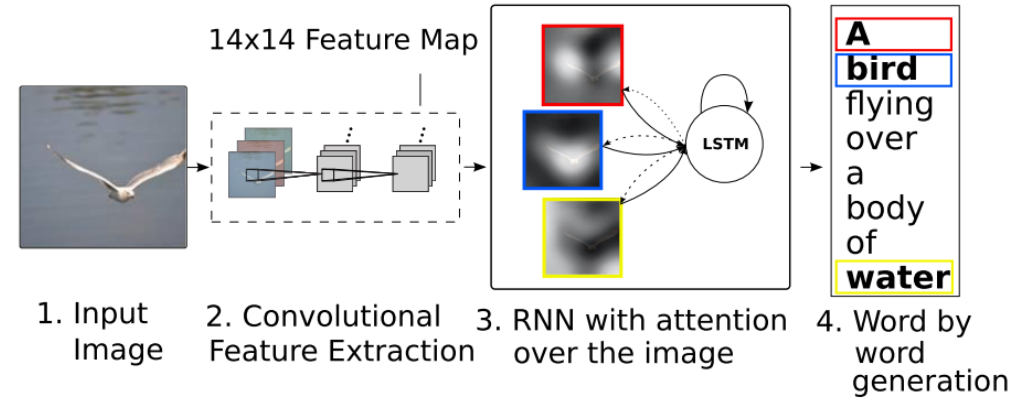


Рис. 13. Способ решения задачи
с использованием механизма внимания

В [66] для обучения подобной модели применяется разновидность алгоритма обучения с подкреплением [67].

В [68] используется похожая идея, предполагающая учет разных фрагментов изображения на разных итерациях РНС. Введена подсеть локализации, которая на каждой итерации выделяет в исходном изображении несколько фрагментов, ранжирует пары фрагмент-слово по их соответствию друг другу и использует полученную информацию для генерирования очередного слова из описания. Вместо *LSTM* используется другой тип РНС – *GRU* (*gated recurrent units* [69]).

В [70] в состав генеративной модели включены две подсети: одна из них решает задачу предсказывания «заметного положения» (фрагмента, на котором будет сфокусировано внимание человека при взгляде на изображение, *saliency attention*), вторая предсказывает контекстную информацию из изображения (*context attention*). Выходы обеих подсетей учитывается на каждой итерации работы РНС. Способ проиллюстрирован на рис. 14 (англ. *baseball player* – бейсболист, *ball* – мяч).

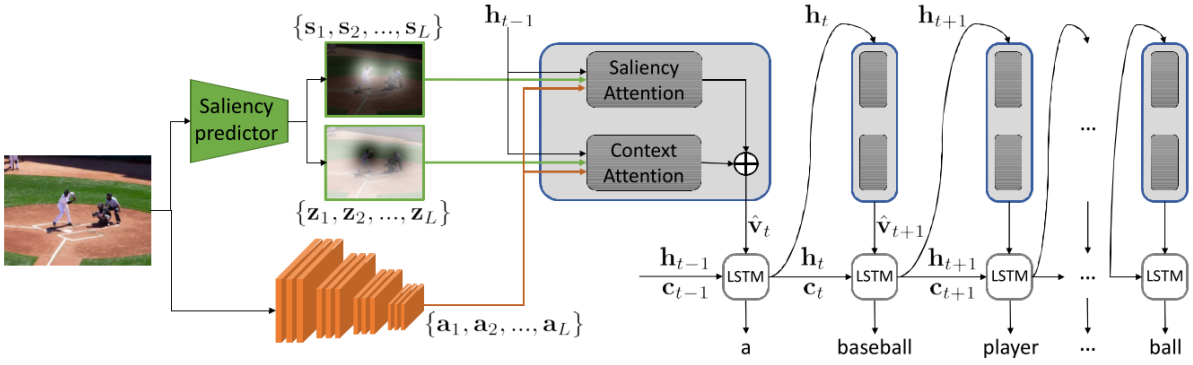


Рис. 14. Способ решения задачи с использованием вспомогательных
подсетей для предсказывания «заметного положения»
и контекстной информации

В [71] подсеть механизма внимания на каждой итерации работы *LSTM* определяет не только фрагмент изображения, который необходимо учитывать

на данной итерации, но и вообще необходимость учета визуальной информации об изображении (например, при генерации артиклей, предлогов и т.п. достаточно опираться только на языковую модель). Модель обучается алгоритмом *Adam* [72].

В [73] предложена иерархическая модель, состоящая из кодера (CHC) и нескольких последовательно соединенных декодеров (*LSTM*), которые генерируют описание изображения, переходя от более общего к более подробному описанию, учитывая при этом визуальные признаки от CHC, а также описание, полученное РНС на предыдущем уровне. На рис. 15 представлен пример модели, состоящей из трех уровней (англ. *a cat in front of a mirror* – кот перед зеркалом, *a cat standing in front of a mirror* – кот стоит перед зеркалом, *a black and while cat looking at itself in a mirror* – черно-белый кот смотрит на себя в зеркало).

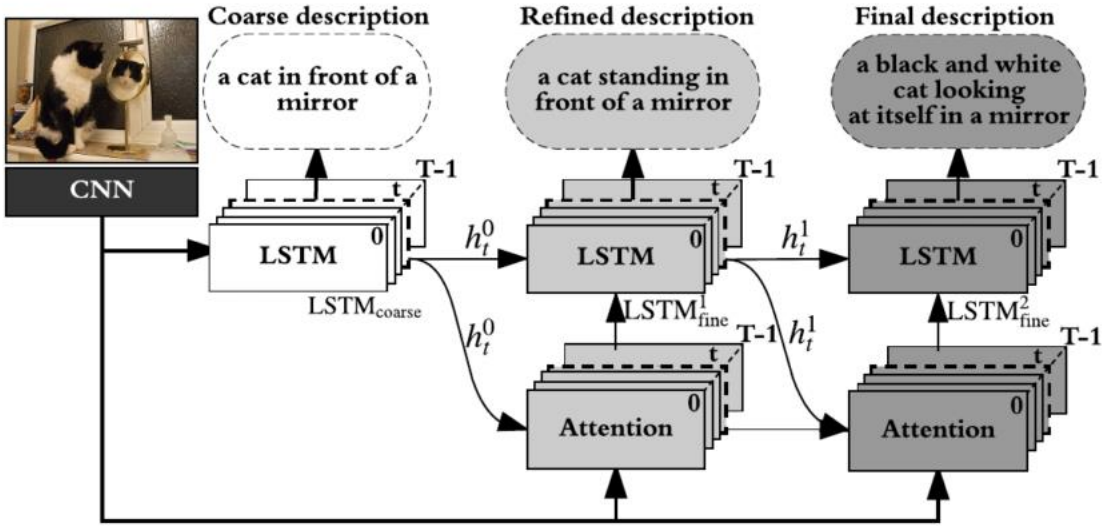


Рис. 15. Иерархическая модель на основе CHC, LSTM и механизма внимания

Для получения более подробных описаний в генеративных сетях второго и старших уровней используется механизм внимания. Модель обучается с помощью алгоритма обучения с подкреплением.

В [74] в модель вводится дополнительная подсеть (РНС), которая работает синхронно с декодирующей РНС, на каждой итерации оценивая, насколько точно декодером сгенерировано очередное слово. Для обучения применяется разновидность обучения с подкреплением – способ «Актера-Критика» (*actor-critic approach* [75]). Предложенная модель представлена на рис. 16.

В [76] предложен «сверточный» способ решения задачи описания изображений: вместо РНС для генерирования текста используется CHC, реализующая маскированные свертки (*masked convolutions*), позволяющие при обработке очередного слова учитывать предыдущий контекст, как и в случае использования РНС. Замена РНС на CHC вызвана тем, что рекуррентные вычисления довольно сложно распараллелить из-за неизбежной последовательности вычисли-

тельного процесса по времени. В предложенном решении обработка всех слов описания производится одновременно и поэтому может быть реализована более эффективно.

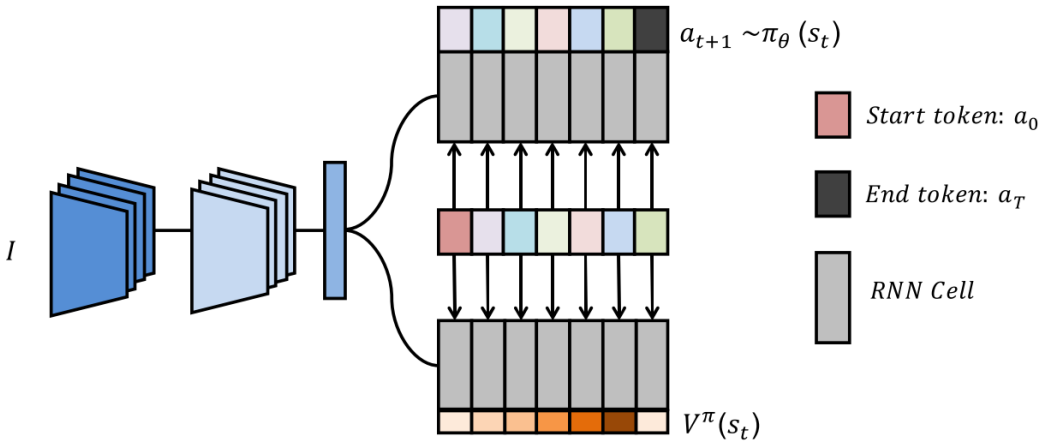


Рис. 16. Модель решения задачи с дополнительной оценочной подсетью

В [77] не только предложен новый способ автоматического описания изображений, но и введен показатель оценки качества работы способов решения рассматриваемой задачи *SPIDEr* (см. п. 4). Для обучения модели (CHC+PHC) предлагается использовать алгоритм обучения с подкреплением – *policy gradient* [78], работа которого заключается в подборе параметров модели, максимизирующих введенный показатель *SPIDEr*.

2.4. Гибридные методы

Гибридные методы автоматического описания изображений предполагают комбинацию процедур, используемых в поисковых и генеративных методах. Для входного изображения осуществляется поиск наиболее релевантных описаний, затем из них извлекаются отдельные словосочетания или выражения (англ. *phases*) – «строительные блоки», – из которых на последнем шаге составляется («синтезируется») выходное описание.

Гибридные методы состоят из следующих этапов.

1. Поиск наиболее релевантных описаний.
2. Извлечение словосочетаний.
3. Синтез выходного описания из словосочетаний.

Схематичное представление этапов работы гибридных методов изображено на рис. 17 (англ. *The lady wearing a headdress with feathers* – дама в головном уборе с перьями, *The pretty girl having a braid* – симпатичная девушка с косой).

При этом поиск может осуществляться двумя способами, описанными в п. 2.2: по изображениям [79, 81, 9] и по описаниям (в едином мультимодальном пространстве) [82, 83, 84].

Рис. 17. Этапы работы гибридных методов

В [79] содержимое изображения представляется в виде набора объектов, материалов, сцен, полученных методами компьютерного зрения. Поиск похожих изображений осуществляется отдельно для каждого элемента данного представления. Затем из описаний найденных изображений составляется выходное предложение путем решения задачи математической оптимизации – целочисленного линейного программирования [80].

В [81] описания из рассматриваемого набора данных предварительно обрабатываются алгоритмом NLP и представляются в виде набора выражений вида (*субъект, действие*), (*объект, действие*), (*действие, предлог, объект*) и других. Изображения представляются в виде числового вектора, содержащего значения нескольких дескрипторов (полученных в результате работы алгоритмов Gabor, Haar, GIST, SIFT и других), и сравниваются с помощью вычисления расстояния в векторном пространстве. Наборы выражений, относящиеся к наиболее похожим изображениям, ранжируются в порядке близости ко входному изображению и используются для построения выходного описания с помощью известных алгоритмов обработки ЕЯ [85, 86].

В [9] из имеющихся описаний извлекаются выражения 4 типов: объекты, действия, материалы, сцены. Для их поиска используются детекторы цвета, текстуры, формы, а также информация, полученная в результате синтаксического разбора текстовых описаний. Из данных выражений-кандидатов составляется выходное предложение (предложения), состоящее из четырех частей, соответствующих четырем типам выражений, для чего решается задача целочисленного линейного программирования.

В [82] для извлечения словосочетаний предлагается использовать общее визуально-текстовое векторное пространство, учитывающее отношения между изображениями и словосочетаниями, которое, по сути, представляет собой классификатор, обучающийся методом градиентного спуска. С использованием этого пространства входному изображению ставится в соответствие несколько словосочетаний. Из данных словосочетаний составляется выходное предложение с использованием методов комбинаторной оптимизации. Для представления изображений используются векторы Фишера и СНС.

В [83, 84] также вводится общее визуально-текстовое пространство, которое входному изображению (представленному в виде вектора признаков от СНС) ставит в соответствие несколько словосочетаний. Данные словосочетания подаются на вход простой языковой модели, которая генерирует текст на ЕЯ исходя из принятой в языке структуры предложения и статистической информации о сочетании слов.

2.5. Основные тенденции развития методов автоматического описания изображений

В результате анализа существующих методов, способов и моделей решения рассматриваемой задачи можно выделить следующие основные тенденции их развития.

1. В связи с увеличением общего объема данных в мировом информационном пространстве и усложнением практических задач, усложняются и изображения, обрабатываемые алгоритмами автоматического описания: увеличивается число и разнообразие классов изображаемых объектов, свойств и отношений. Это влечет за собой рост размера словаря и экспоненциальное увеличение количества возможных предложений для их описания. А вероятность того, что предопределенное предложение будет соответствовать новому изображению, резко уменьшается, если число таких предложений также не растет экспоненциально [52]. В связи с этим наибольший интерес для исследователей в настоящее время представляет задача *генерирования* текстовых описаний, хотя ранее рассматриваемую задачу часто сводили к поисковой.

2. Наиболее интенсивно разрабатываются различные методы и способы на основе глубоких нейронных сетей (CHC для анализа изображений и РНС для обработки текста).

3. Для обучения нейросетевых моделей активно применяются различные алгоритмы обучения с подкреплением, что подразумевает рассмотрение обучения как процесса поиска стратегии поведения системы (в отличие от классических алгоритмов обучения с учителем или без учителя).

4. Многие успешно используемые модели включают различные подсети, позволяющие учитывать дополнительные аспекты в процессе анализа изображения и/или генерирования текста (например, механизм внимания).

3. Наборы данных

Для построения систем решения задачи автоматического описания изображений, способных конкурировать с человеком в данной области, необходимы обширные репрезентативные массивы данных, содержащие изображения и соответствующие им текстовые описания, удовлетворяющие, кроме того, следующим требованиям.

1. Учитывая специфику и большую вариативность выходных данных (каждое изображение может быть описано огромным количеством разных синонимичных предложений на ЕЯ), желательно, чтобы в наборе данных одному изображению соответствовало несколько различных описаний.

2. Для построения интеллектуальных с точки зрения пользователя моделей, способных заменить человека в некоторых практических задачах, требуется, чтобы описания в наборе данных были составлены людьми.

К настоящему времени существует множество наборов данных, предназначенных для решения рассматриваемой задачи [15, 22, 27, 44, 87–92], однако наиболее популярными являются следующие.

1. *PASCAL1K* [87]. Один из наиболее ранних наборов данных, составленный в 2010 году по базе изображений PASCALVOC (2008) [93], включающий

1000 изображений, по 5 описаний на каждое изображение. Описания составлены людьми и собирались с помощью краудсорсинговой площадки *Amazon's Mechanical Turk (MTurk)* [94]. Пример изображения и 5 соответствующих описаний представлены на рис. 18 (англ. *Two men playing cards at a table* – двое мужчин играют в карты за столом).

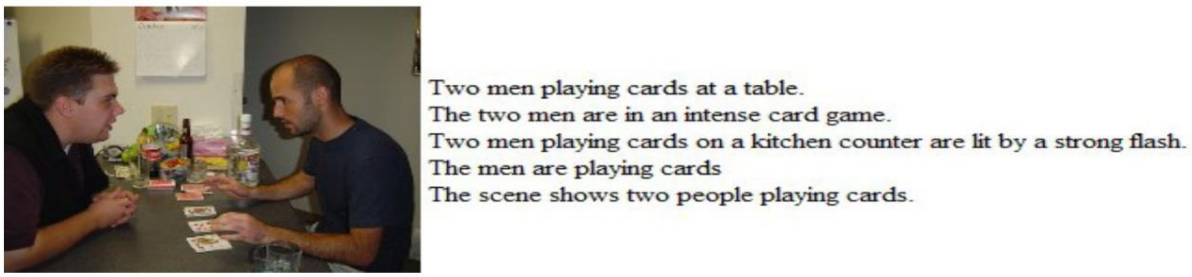


Рис. 18. Пример данных из набора PASCAL1K

Сами изображения являются фотографиями пользователей ресурса Flickr [95] и представлены в 20 категориях (по 50 изображений на категорию): самолет, велосипед, птица, лодка, бутылка, автобус, машина, кошка, кресло, корова, стол, собака, лошадь, мотоцикл, человек, растение в горшке, овца, диван, самолет, ТВ/монитор. На одной фотографии может быть изображено более 1 объекта, в том числе объекты разных категорий. На рис. 19 представлена гистограмма распределения количества изображений (*images*) и объектов (*objects*) на них по категориям.

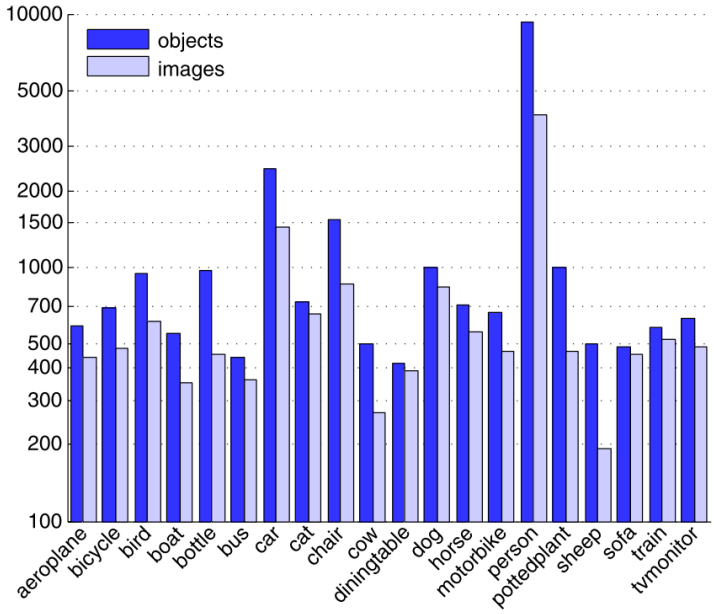


Рис. 19. Распределение количества изображений и объектов из набора PASCAL1K по категориям

2. *Flickr8K* [22]. Набор данных собран из фотографий пользователей ресурса Flickr и содержит 8092 изображения и их описания (по 5 описаний на изображение). Набор был собран в 2013 г. с помощью MTurk той же командой исследователей, что и PASCAL1K, однако он более обширен и разнообразен по

сравнению с предыдущим набором и содержит в основном изображения людей и животных (собак). Примеры приведены на рис. 20 (англ. *A man is doing tricks on a bicycle on ramps in front of a crowd* – мужчина делает трюки на велосипеде перед толпой; *A group of people sit at a table in front of a large building* – группа людей сидит за столом перед большим зданием).

A man is doing tricks on a bicycle on ramps in front of a crowd.
A man on a bike executes a jump as part of a competition while the crowd watches.
A man rides a yellow bike over a ramp while others watch.
Bike rider jumping obstacles.
Bmx biker jumps off of ramp.

A group of people sit at a table in front of a large building.
People are drinking and walking in front of a brick building.
People are enjoying drinks at a table outside a large brick building.
Two people are seated at a table with drinks.
Two people are sitting at an outdoor cafe in front of an old building.

Рис. 20. Примеры данных из набора Flickr8K

3. *Flickr30K* [88] – расширенная версия набора Flickr8K. Содержит 31 783 фотографии и 158 915 описаний. Примеры изображений и соответствующих им описаний представлены на рис. 21 (англ. *Gray haired man in black suit and yellow tie working in a financial environment* – седой мужчина в черном костюме с желтым галстуком, работающий в финансовой среде; *A butcher cutting an animal to sell* – мясник разрезает животное для продажи).

Gray haired man in black suit and yellow tie working in a financial environment.
A graying man in a suit is perplexed at a business meeting.
A businessman in a yellow tie gives a frustrated look.
A man in a yellow tie is rubbing the back of his neck.
A man with a yellow tie looks concerned.

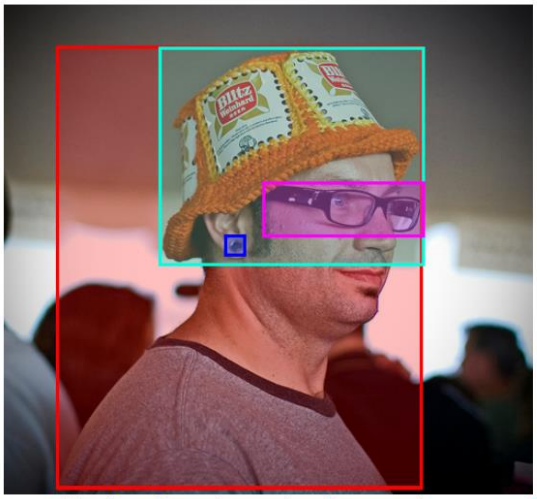
A butcher cutting an animal to sell.
A green-shirted man with a butcher's apron uses a knife to carve out the hanging carcass of a cow.
A man at work, butchering a cow.
A man in a green t-shirt and long tan apron hacks apart the carcass of a cow while another man hoses away the blood.
Two men work in a butcher shop; one cuts the meat from a butchered cow, while the other hoses the floor.

Рис. 21. Примеры данных из набора Flickr30K

4. *Flickr30k Entities* [27]. В 2015 г. на основе Flickr30k был создан набор Flickr30k Entities, включающий, помимо описаний, относящихся к целым изображениям, детальную информацию о фрагментах изображений и ассоциированных с ними сущностях (объектах, свойствах, действиях). Дополнительная информация из набора данных используется для решения задачи локализации выражений (*phrase localization*).

Примеры изображений и ассоциированной с ними информации представлены на рис. 22 (англ. *A man with pierced ears is wearing glasses and an orange hat* – мужчина с проколотыми ушами в очках и оранжевой шляпе; *During a gay pride parade in an Asian city, some people hold up rainbow flags to show their support* – во время гей-парада в азиатском городе несколько человек держат ра-

дужные флаги для выражения поддержки). Фрагменты изображений и соответствующие им словосочетания из описаний выделены одинаковыми цветами.



A man with pierced ears is wearing glasses and an orange hat.
A man with glasses is wearing a beer can crotched hat.
A man with gauges and glasses is wearing a Blitz hat.
A man in an orange hat starring at something.
A man wears an orange hat and glasses.

During a gay pride parade in an Asian city, some people hold up rainbow flags to show their support.
A group of youths march down a street waving flags showing a color spectrum.
Oriental people with rainbow flags walking down a city street.
A group of people walk down a street waving rainbow flags.
People are outside waving flags .

Рис. 22. Примеры данных из набора Flickr30k Entities

На рис. 23 представлены категории объектов и распределение по ним количества объектов на изображениях набора данных (англ. *people* – люди, *clothing* – одежда, *body parts* – части тела, *animals* – животные, *vehicles* – транспорт, *instruments* – инструменты, *scene* – фон, окружение, место действия, *other* – другое).

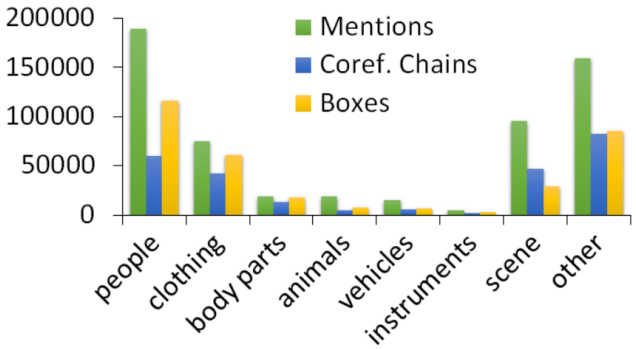


Рис. 23. Распределение количества объектов на изображениях набора Flickr30k Entities по категориям

5. *Microsoft COCO Caption* [89] содержит изображения из набора *Microsoft Common Objects in COntext* (*COCO*) [96]. Описания (по 5 на каждое изображения) также собирались с помощью Amazon's Mechanical Turk. Набор данных содержит более 330 000 изображений, относящихся к 80 категориям и 91 типу сцен. Примеры приведены на рис. 24 (англ. *The man at bat readies to swing at the pitch while the umpire looks on* – мужчина с битой готовится отбить подачу, пока судья смотрит на него; *A large bus sitting next to a very tall building* – большой

автобус стоит рядом с очень высоким зданием; *A horse carrying a large load of hay and two people sitting on it* – лошадь несет большой тюк сена и двух человек, сидящих на нем; *Bunk bed with a narrow shelf sitting underneath it* – двухъярусная кровать с узкой полкой под ней).

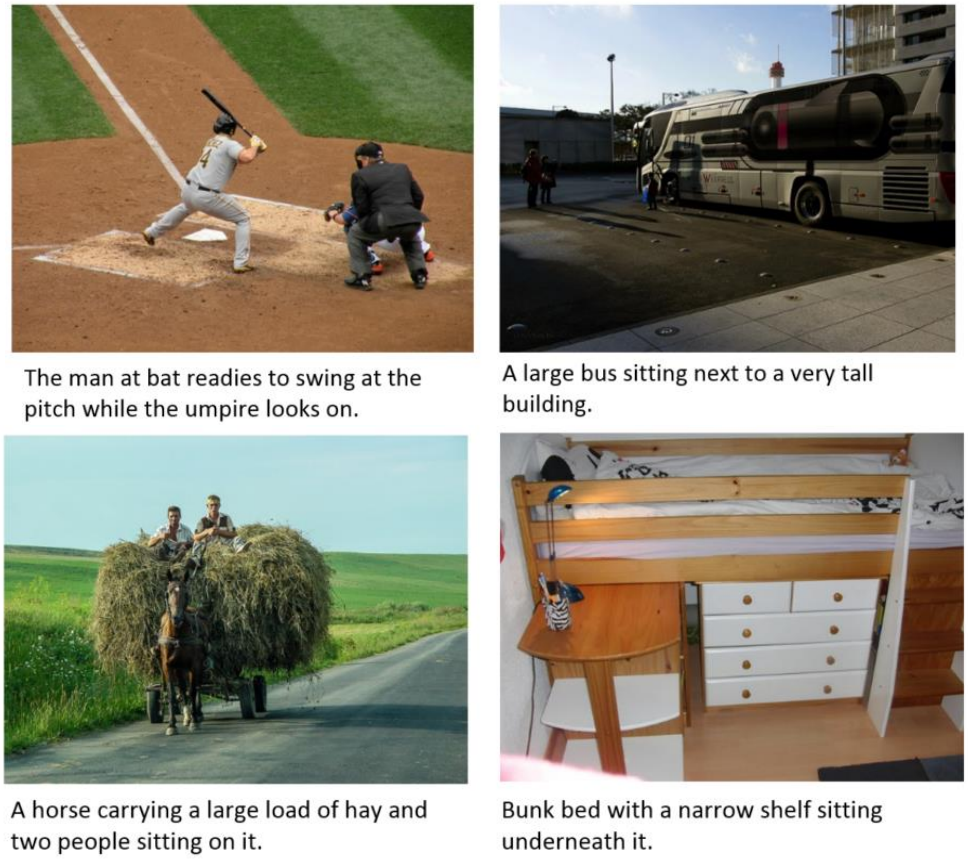


Рис. 24. Примеры данных из набора Microsoft COCO Caption

На рис. 25 представлена гистограмма распределения количества объектов на изображениях наборов PASCAL1K и COCO по категориям (англ. *Instances per category*).

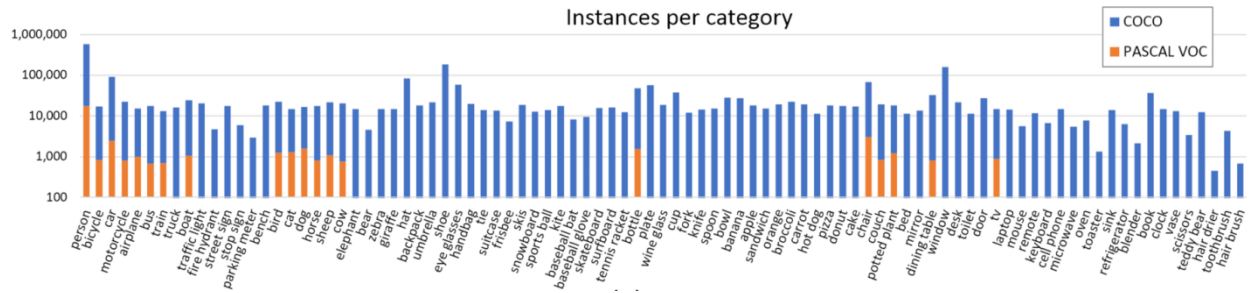


Рис. 25. Распределение количества объектов на изображениях наборов PASCAL1K и MS COCO по категориям

В настоящее время большинство исследователей для обучения и тестирования методов автоматического описания изображений используют наборы Flickr30k и Microsoft COCO Caption как наиболее обширные и разнообразные.

4. Оценка качества работы методов автоматического описания изображений

В случае комплексных интеллектуальных задач, которые с легкостью решаются человеком, но для компьютера представляют большую сложность, лучший способ оценить качество решения (качество работы того или иного метода/способа решения) – это привлечь экспертов (людей). Во многих исследованиях авторами организуется сбор экспертных оценок результатов решения задачи с помощью какой-либо Интернет-площадки. Пользователю предлагается оценить, насколько точно заданное предложение описывает заданную картинку, например, по шкале от 1 до 5 [79] или от 1 до 4 [22].

При оценке качества сгенерированного (найденного) описания заданного изображения применяются следующие критерии [8].

1. Описание точно отражает содержимое изображения.
2. Описание правильно с точки зрения грамматики.
3. Описание не содержит неверной информации.
4. Описание напоминает предложение, составленное человеком.

Собранные данные обрабатываются, затем по статистическим характеристикам (средняя оценка эксперта, плотность распределения и т. д.) можно говорить о качестве работы метода решения задачи.

Однако в случае большого количества данных способы, связанные с вовлечением экспертов, могут стать слишком дорогостоящими. Поэтому чаще применяются различные автоматические способы оценивания.

Выбор метода оценки качества работы системы описания изображений зависит от решаемой задачи и от того, как организован тестовый набор данных. Возможны два варианта проведения автоматической оценки качества решения рассматриваемой задачи.

1. Если речь идет о поисковой задаче, а тестовый набор данных состоит из пар вида «изображение x – идентификатор/идентификаторы описания из заданного множества описаний $y^* \in Y$», то выходное значение (результат работы метода) y должно точно совпадать с эталонным значением (одним из эталонных значений) из тестового набора: $y = y^*$. Тогда оценивается точность (*Precision*) и/или полнота (*Recall*) работы метода.

Чаще других применяется показатель *Recall@K* – полнота работы для первых *K* выходов – доля случаев, для которых по крайне мере одно из эталонных значений y^* встретилось среди первых *K* выходных результатов. В ряде работ применяется показатель *AP@K* (*Average Precision*) – точность работы для первых *K* выходов – доля совпадений выходного и эталонного значений среди первых *K* выходных результатов.

2. Если тестовый набор содержит изображения x и соответствующие им описания y^*, не входящие в заданное множество Y, по которому производился поиск описания, либо речь идет о задаче генерирования описания, то требование о точном совпадении результата работы метода y с эталонным значением y^* становится невозможным в силу специфики получаемого результата – тек-

стов на ЕЯ. В этом случае оценка качества работы метода основывается на некоторой функции, оценивающей сходство двух предложений на естественном языке (y и $y*$).

К настоящему времени предложено несколько показателей оценки качества решения задач автоматического описания изображений, наиболее популярными из которых являются: BLEU [97], ROUGE [98], METEOR [99], CIDEr [101], SPICE [102].

Пусть имеется тестовый набор данных, состоящий из множества изображений I и множества соответствующих описаний S. В результате работы того или иного метода получено множество C сгенерированных (найденных) описаний. При описании показателей будут использованы следующие обозначения: $I_i \in I$ – входное изображение, $c_i \in C$ – выходное описание, $S_i = \{s_{i1}, ..., s_{im}\} \in S$ – множество верных описаний изображения I_i, $m \geq 1$. Описания представляются в виде набора *n*-грамм $\{\omega_1, ..., \omega_k, ...\} \in \Omega$ (*n*-грамма есть последовательность из *n* слов, *n>0*), $k \geq 1$. Количество раз, которое *n*-грамма ω_k встречается в описании, обозначается $h_k(s_{ij})$ или $h_k(c_i)$.

1. *BLEU* [97]. Показатель был предложен ранее других и заимствован из области машинного перевода. Сходство двух предложений оценивается по вхождению в них одинаковых *n*-грамм. Для заданного *n* рассчитывается точность работы метода по формуле (1):

$$CP_n(C,S) = \frac{\sum_i \sum_k \min(h_k(c_i), \max\limits_{j \in \{1..m\}} h_k(s_{ij}))}{\sum_i \sum_k h_k(c_i)}. \quad (1)$$

Чаще применяются значения *n* не более 4. Общее значение показателя BLEU вычисляется как взвешенное среднее геометрическое точности работы метода по *n*:

$$BLEU(C, S) = b(C, S)\exp(\sum_{j=1}^{n} \omega_j \log(CP_j(C, S))), \quad (2)$$

где: $\omega_n = const$, $n = 1, 2, 3, 4$, $b(C, S)$ – «штраф» за длинные описания, вычисляемый по формуле (3):

$$b(C, S) = \begin{cases} 1, & если\ l_c > l_s \\ e^{1-l_s/l_c}, & если\ l_c \leq l_s \end{cases}, \quad (3)$$

где: l_c – общая длина предложений на выходе метода, l_s – общая длина предложений из тестового набора данных.

2. *ROUGE* [98]. Показатель представляет собой набор из трех показателей, заимствованных из задачи автоматического реферирования (*Automatic Text Summarization* [103]).

2.1. $ROUGE_N$ – показатель аналогичен показателю $BLEU(C, S)$, но представляет собой полноту работы метода:

$$ROUGE_N(c_i, S_i) = \frac{\sum_j \sum_k \min(h_k(c_i), h_k(s_{ij}))}{\sum_j \sum_k h_k(s_{ij})}. \quad (4)$$

2.2. $ROUGE_L$ вычисляется на основе самой длинной общей подпоследовательности (*LCS, Longest Common Subsequence*) – набора слов, которые встречаются в обоих сравниваемых предложениях в одинаковой последовательности (но, в отличие от *n*-грамм, между словами *LCS* могут встречаться другие слова, не входящие в *LCS*). Пусть $l(c_i, s_{ij})$ – длина *LCS*, тогда $ROUGE_L$ вычисляется как *F*-мера по формулам:

$$R_l(c_i, S_i) = \max_j \frac{l(c_i, s_{ij})}{|s_{ij}|}, \quad (5)$$

$$P_l(c_i, S_i) = \max_j \frac{l(c_i, s_{ij})}{|c_i|}, \quad (6)$$

$$ROUGE_L(c_i, S_i) = \frac{(1+\beta^2)R_l P_l}{R_l + \beta^2 P_l}, \quad (7)$$

где: $\beta = 1,2$ – наиболее предпочтительное значение отношения P_l / R_l.

3. $ROUGE_S$ – показатель основан на понятии *skip bigrams* (скачущие биграммы) – последовательность из 2 слов. Как и в *LCS*, слова в биграмме не обязательно расположены друг за другом, а могут разделяться другими словами. Пусть $f_k(s_{ij})$ – количество биграмм в предложении s_{ij}, тогда $ROUGE_S$ вычисляется как F-мера по формулам:

$$R_S(c_i, S_i) = \max_j \frac{\sum_k \min(f_k(c_i), f_k(s_{ij}))}{\sum_k f_k(s_{ij})}, \quad (8)$$

$$P_S(c_i, S_i) = \max_j \frac{\sum_k \min(f_k(c_i), f_k(s_{ij}))}{\sum_k f_k(c_i)}, \quad (9)$$

$$ROUGE_S(c_i, S_i) = \frac{(1+\beta^2)R_S P_S}{R_S + \beta^2 P_S}. \quad (10)$$

4. *METEOR* [99]. Показатель также заимствован из области машинного перевода. Для оценки схожести 2 предложений сначала нужно провести выравнивание – установление соответствия между *n*-граммами. При этом последовательно применяются следующие виды соответствия: точное совпадение слов в *n*-грамме, соответствие основ (после процедуры стемминга – выделения основы), соответствие синонимов (с использованием информации из *WordNet* [104]). Примеры выравниваний приведены на рис. 26 (англ. *The cat sat on the mat* – кот сидит на коврике).

Из множества возможных выравниваний выбирается выравнивание с минимальным количеством пересечений. Итоговое значение показателя METEOR вычисляется как гармоническое среднее между точностью и полнотой для наилучшей пары предложений:

$$R_m = \frac{|m|}{\sum_k h_k(s_{ij})}, \quad (11)$$

$$P_m = \frac{|m|}{\sum_k h_k(c_i)}, \quad (12)$$

где: m – количество общих n-грамм.

$$F_{mean} = \frac{R_m P_m}{\alpha P_m + (1-\alpha) R_m}, \quad (13)$$

$$Pen = \gamma(\frac{ch}{m})^{\theta}, \quad (14)$$

где: ch – количество смежных n-грамм с одинаковым порядком слов в обоих предложениях, Pen – штраф для учета словосочетаний.

Итоговое значение показателя METEOR:

$$METEOR = (1-Pen)F_{mean}. \quad (15)$$

Значения параметров γ, α, θ выбираются в ходе экспериментов таким образом, чтобы обеспечить наибольшую корреляцию с экспертной оценкой.

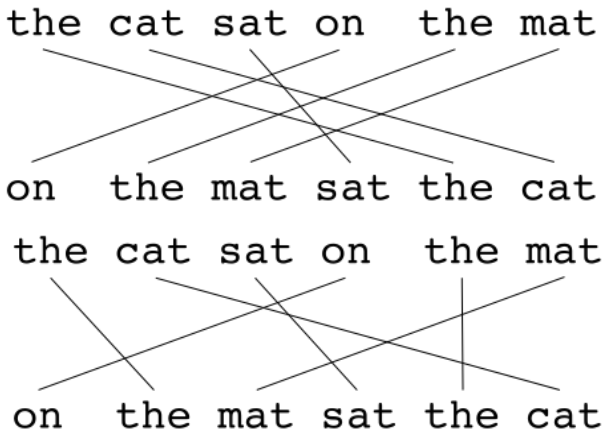


Рис. 26. Примеры выравниваний для вычисления показателя METEOR [100]

5. *CIDEr* [101]. В отличие от предыдущих показателей CIDEr был разработан специально для оценки качества задачи автоматического описания изображений. С помощью показателя измеряется сходство сгенерированного предложения с набором предложений из тестового набора данных. Предварительно проводится стемминг. Для каждой n-граммы рассчитывается значение TF-IDF (статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса [105]) $g_k(s_{ij})$:

$$g_k(s_{ij}) = \frac{h_k(s_{ij})}{\sum_{\omega_l \in \Omega} h_l(s_{ij})} \log(\frac{|I|}{\sum_{I_p \in I} \min(1, \sum_q h_k(s_{pq}))}), \quad (16)$$

где первый множитель вычисляет частоту (TF) n-граммы ω_k, а второй – вычисляет «редкость» n-граммы с помощью IDF, таким образом, второй множитель уменьшает вес n-грамм, которые часто встречаются во всех описаниях тестового набора.

Для заданного *n* показатель CIDEr вычисляется следующим образом:

$$CIDEr_N(c_i, S_i) = \frac{1}{m}\sum_j \frac{g(c_i)^n g(s_{ij})^n}{\left\|g(c_i)^n\right\|\left\|g(s_{ij})^n\right\|}, \quad (17)$$

где: $g(c_i)^n$, $g(s_{ij})^n$ – вектора, составленные из $g_k(c_i)$ и $g_k(s_{ij})$ для всех *n*-грамм длины *n*, $\left\|g(c_i)^n\right\|$, $\left\|g(s_{ij})^n\right\|$ – модули векторов $g(c_i)^n$ и $g(s_{ij})^n$ соответственно.

Общее значение показателя:

$$CIDEr(c_i, S_i) = \sum_{n=1}^{N} \frac{1}{N} CIDEr_N(c_i, S_i). \quad (18)$$

В [101] предложена также модификация показателя CIDEr – CIDEr-D, – отличающаяся введением штрафа, основанного на длине сравниваемых предложений:

$$CIDErD_N(c_i, S_i) = \frac{10}{m}\sum_j e^{\frac{-(l(c_i)-l(s_{ij}))^2}{2\sigma^2}} \frac{\min(g(c_i)^n, g(s_{ij})^n)\cdot g(s_{ij})^n}{\left\|g(c_i)^n\right\|\left\|g(s_{ij})^n\right\|}, \quad (19)$$

где: $l(c_i)$, $l(s_{ij})$ – длины предложений c_i и s_{ij} соответственно. Значение параметра $\sigma = 6$.

Общее значение показателя:

$$CIDErD(c_i, S_i) = \sum_{n=1}^{N} \frac{1}{N} CIDErD_N(c_i, S_i). \quad (20)$$

6. *SPICE* [102]. Для вычисления показателя описания представляются не в виде набора *n*-грамм, а в виде набора логических кортежей, полученных из графового представления описания – «графа сцены» (*scene graph*). Вершинами данного графа являются объекты, их атрибуты и отношения между ними. Для представления текстового описания в виде графа используется специальный синтаксический разбор на основе *Stanford Scene Graph Parser* [106]. Примеры графа и соответствующего набора логических кортежей представлены на рис. 27 (англ. *A young girl standing on top of a tennis court* – юная девушка стоит на кромке теннисного корта).

Показатель SPICE рассчитывается как *F*-мера с использованием следующих формул:

$$P(c_i, S_i) = \frac{\left|T(G(c_i)) \otimes T(G(S_i))\right|}{\left|T(G(c_i))\right|}, \quad (21)$$

$$R(c_i, S_i) = \frac{\left|T(G(c_i)) \otimes T(G(S_i))\right|}{\left|T(G(S_i))\right|}, \quad (22)$$

$$SPICE(c_i, S_i) = \frac{2P(c_i, S_i)R(c_i, S_i)}{P(c_i, S_i) + R(c_i, S_i)}, \quad (23)$$

где: $G(c_i)$, $G(S_i)$ – графовые представления описаний c_i и S_i соответственно, $T(G)$ – набор логических кортежей графа G, $\otimes$ – бинарная операция, возвращающая общие кортежи двух графов. При установлении соответствия между кортежами учитывается синонимичность слов в ЕЯ по информации из *WordNet*.

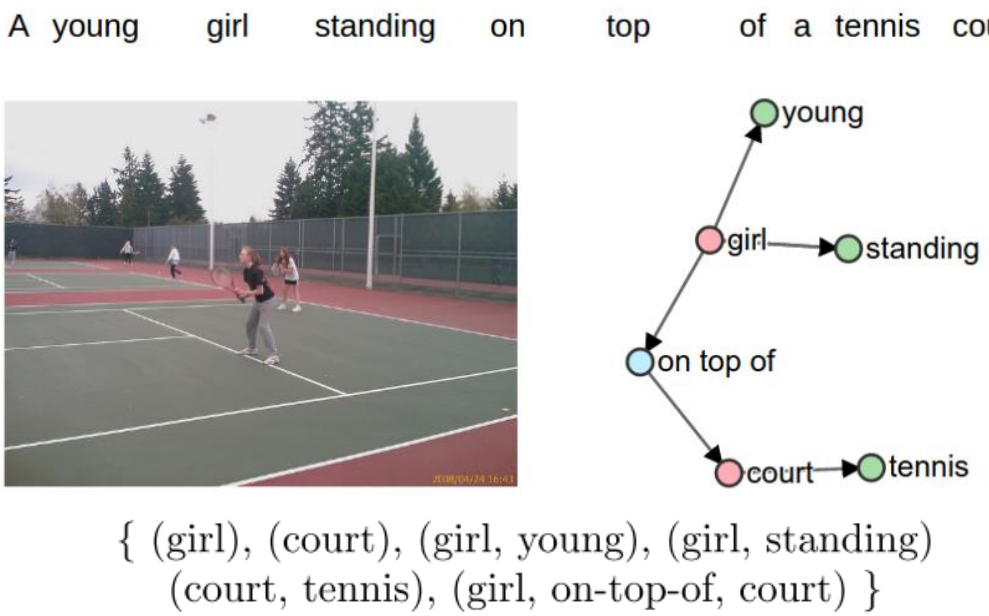


Рис. 27. Пример графового представления описания и набора логических кортежей для вычисления показателя SPICE (красные вершины – объекты, зеленые – атрибуты, свойства, голубые – отношения)

7. *SPIDEr* [77]. Новый показатель, представляющий собой линейную комбинацию показателей CIDEr и SPICE и учитывающий таким образом как синтаксическую (CIDEr), так и семантическую (SPICE) точность описаний.

Создание качественного показателя автоматической оценки точности генерирования текстовых описаний изображений, способного заменить экспертное оценивание по критериям, приведенным выше, – возможно, самая важная проблема в данной области [107]. Широко применяемые в настоящее время показатели оценки качества (1–4) часто обсуждаются и подвергаются критике. Так, в [22, 46, 44] показано, что данные показатели плохо коррелируют с экспертными оценками (оценками качества, выполненными людьми). При оценке качества описания изображений с помощью данных показателей многие разработанные модели «превзошли» человека [8]. Однако машинные описания по-прежнему намного хуже с точки зрения смысла и грамотности, чем описания, составленные людьми [47].

Показатели SPICE и SPIDEr учитывают не только совпадение *n*-грамм, но и семантическую точность описаний, поэтому неплохо коррелируют с экспертными оценками [102].

В [77] введено 2 критерия, которым должен соответствовать показатель оценки качества методов описания изображений:

1) описания, которые по мнению экспертов (людей) точно описывают заданные изображения, должны иметь высокие значения показателя;

2) описания, имеющие высокие значения показателя, должны быть высоко оценены экспертами.

Существующие показатели удовлетворяют критерию 1, но не 2. По-прежнему актуальной является задача разработки метода оценивания качества

автоматического описания изображений, удовлетворяющего сформулированным критериям и способного заменить экспертное оценивание.

Заключение

В данной статье рассмотрены задачи автоматического описания изображений – актуальные мультимодальные задачи искусственного интеллекта, находящиеся на стыке разных областей анализа данных: теории распознавания образов и обработки естественного языка. Проведен анализ и предложена классификация рассматриваемых задач, а также методов их решения. Приведены уточнения терминов и формулировок, используемых зарубежными учеными при обозначении решаемых задач, для применения в русскоязычных исследованиях. Описаны наиболее эффективные методы автоматического поиска и генерирования описаний изображений, а также сформулированы основные тенденции их развития. Приведено описание наиболее популярных наборов данных для решения рассматриваемых задач и показателей оценки качества работы методов и реализующих их способов и алгоритмов.

Данная работа является первым подробным обзором задач и методов автоматического описания изображений на русском языке. Результаты исследования могут быть использованы для изучения существующих и разработки новых методов, а также способов, алгоритмов и моделей решения рассматриваемых задач.

Работа выполнена при поддержке гранта РФФИ, проект № 18-07-00928.

Литература

1. YouTube for Press // Youtube [Электронный ресурс]. – URL: https://www.youtube.com/yt/about/press (дата обращения: 01.02.2018).

2. Борисов В. В., Коршунова К. П. Постановка прямой и обратной задачи поиска и генерирования текстовых описаний по изображениям // Энергетика, информатика, инновации - 2017 (электроэнергетика, электротехника и теплоэнергетика, математическое моделирование и информационные технологии в производстве). Сборник трудов VII Международной научно-технической конференции. Т 1. - Смоленск, 2017. - С. 228-230.

3. Abella A., Kender J. R., Starren J. Description Generation of Abnormal Densities found in Radiographs // Proc. Symp. Computer Applications in Medical Care, Journal of the American Medical Informatics Association. 1995. P. 542-546.

4. Gerber R., Nagel N. H. Knowledge representation for the generation of quantified natural language descriptions of vehicle traffic in image sequences // Proceedings of the International Conference on Image Processing. 1996. P. 805-808.

5. Szeliski R. Computer Vision: Algorithms and Applications // Springer Science & Business Media. 2010. P. 10-16.

6. Dale R., White M. Shared Tasks and Comparative Evaluation in Natural Language Generation: Position Papers, Arlington. VA, USA. 2007.

7. Reiter E., Belz A. An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems // Computational Linguistics. 2009. № 35(4). P. 529-558.

8. Bernardi R., Cakici R., Elliott D., Erdem A., Erdem E., Ikizler-Cinbis N., Keller F., Muscat A., Plank B. Automatic description generation from images: A survey of models, datasets, and evaluation measures // IJCAI International Joint Conference on Artificial Intelligence. 2017. P. 4970-4974.

9. Kuznetsova P., Ordonez V., Berg T., Choi Y. TREETALK: Composition and Compression of Trees for Image Descriptions // Transactions of the Association for Computational Linguistics. 2014. Vol. 2. P. 351-362.

10. Kiros R., Salakhutdinov R., Zemel R. S. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models // Association for Computational Linguistics. 2014.

11. Yao T., Pan Y., Li Y., Qiu Z., Mei T. Boosting Image Captioning with Attributes // International Conference on Computer Vision. 2017. P. 1-11.

12. Мельниченко А. С. Автоматическая аннотация изображений на основе глобальных признаков // Известия ЮФУ. Технические науки. 2009. № 8. С. 189-200.

13. Курбатов С. С., Найденова К. А., Хахалин Г. К. О схеме взаимодействия в комплексе «анализ и синтез естественного языка и изображений» // КИИ-2010. Труды конференции. – Тверь, 2010. – С. 234-242.

14. Проскурин А. В. Автоматическое аннотирование ландшафтных изображений // Сибирский журнал науки и технологий. 2014. № 3 (55). С. 120-125.

15. Ordonez V., Kulkarni G., Berg T. L. Im2Text: Describing Images Using 1 Million Captioned Photographs // Neural Information Processing Systems. 2011. P. 1143-1151.

16. Mason R., Charniak E. Nonparametric Method for Data-driven Image Captioning // Association for Computational Linguistics. 2014. P. 592-598.

17. Yagcioglu S., Erdem E., Erdem A. A Distributed Representation Based Query Expansion Approach for Image Captioning // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015. Vol. 1. P. 106-111.

18. Ordonez V., Han X., Kuznetsova P., Kulkarni G., Mitchell M., Yamaguchi K., Stratos K., Goyal A., Dodge J., Mensch A., Daume H., Berg A. C., Choi Y., Berg T. L. Large Scale Retrieval and Generation of Image Descriptions // International Journal of Computer Vision. 2016.

19. Devlin J., Gupta S., Girshick R., Mitchell M., Zitnick C. L. Exploring Nearest Neighbor Approaches for Image Captioning // arXiv.org [Электронный ресурс]. 2015. – URL: http://arxiv.org/abs/1505.04467 (дата обращения: 01.02.2018).

20. Farhadi A., Hejrati M., Sadeghi M. A., Young P., Rashtchian C., Hockenmaier J., Forsyth D. Every picture tells a story: Generating sentences from images // Lecture Notes in Computer Science (Including Subseries Lecture Notes in

Artificial Intelligence and Lecture Notes in Bioinformatics). 2010. 6314 LNCS (PART 4). P. 15-29.

21. Socher R., Karpathy A., Le Q. V., Manning C. D., Ng A. Y. Grounded Compositional Semantics for Finding and Describing Images with Sentences // Transactions of the Association for Computational Linguistics. 2014. P. 207-218.

22. Hodosh M., Young P., Hockenmaier J. Framing image description as a ranking task: Data, models and evaluation metrics // IJCAI International Joint Conference on Artificial Intelligence. 2013. P. 4188-4192.

23. Verma Y., Jawahar C. V. Im2Text and Text2Im : Associating Images and Texts for Cross-Modal Retrieval // British Machine Vision Conference. 2014.

24. Karpathy A., Joulin A., Fei-Fei L. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping // Neural Information Processing Systems. 2014.

25. Gong Y., Wang L., Hodosh M., Hockenmaier J., Lazebnik S. Improving image-sentence embeddings using large weakly annotated photo collections // Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2014. 8692 LNCS (PART 4). P. 529-545.

26. Reed S., Akata Z., Schiele B., Lee H. Learning Deep Representations of Fine-grained Visual Descriptions // Conference on Computer Vision and Pattern Recognition. 2016.

27. Plummer B. A., Wang L., Cervantes C. M., Caicedo J. C., Hockenmaier J., Lazebnik S. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models // International Journal of Computer Vision. 2017. № 123 (1). P. 74-93.

28. Oliva A., Hospital W., Ave L. Modeling the Shape of the Scene : A Holistic Representation of the Spatial Envelope // International Journal of Computer Vision. 2001. № 42 (3). P. 145-175.

29. TF-IDF // Википедия: свободная энциклопедия [Электронный ресурс]. 23.01.2018. – URL: http://ru.wikipedia.org/?oldid=90463885 (дата обращения: 23.01.2018).

30. Felzenszwalb P., McAllester D., Ramanan D. A Discriminatively Trained, Multiscaled, Deformable Part Model // Conference on Computer Vision and Pattern Recognition. 2008. P. 1-8.

31. Hoiem D., Divvala S., Hays J. Pascal VOC 2009 Challenge // PASCAL challenge workshop in European Conference on Computer Vision. 2009.

32. Curran J., Clark S., Bos J. Linguistically motivated large-scale NLP with C&C and boxer // Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics on Interactive Poster and Demonstration Sessions. 2007. P. 33-36

33. Varma M., Zisserman A. A statistical approach to texture classification from single images // International Journal of Computer Vision. 2005. № 62. P. 61-81.

34. Tsochantaridis I., Hofmann T., Joachims T., Altun Y. Support vector machine learning for interdependent and structured output spaces // Proc. Intl. Conf. Machine Learning. 2004.

35. De Marneffe M.-C., Maccartney B., Manning C. D. Generating Typed Dependency Parses from Phrase Structure Parses // International Conference on Language Resources and Evaluation. 2006. Vol. 6. 449-454.

36. Girshick R., Donahue J., Darrell T., Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation // Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2014.

37. Vinyals O., Toshev A., Bengio S., Erhan D. Show and Tell: A Neural Image Caption Generator // Conference on Computer Vision and Pattern Recognition. 2015. P. 1-10.

38. Yao B. Z., Yang X., Lin L., Lee M. W., Zhu S. C. I2T: Image parsing to text description // Proceedings of the IEEE. 2010. № 98(8). P. 1485-1508.

39. Chen H., Xu Z. J., Liu Z. Q., Zhu S. C. Composite templates for cloth modeling and sketching // Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2006. Vol. 1. P. 943-950.

40. Berners-Lee T., Hendler J., Lassila O. The Semantic Web // Scientific American. 2001. № 284. P. 29-37.

41. Yang Y., Teo C. L., Daume H., Aloimonos Y. Corpus-Guided Sentence Generation of Natural Images // Proceedings of The Conference on Empirical Methods on Natural Language Processing. 2011. P. 444-454.

42. Li S., Kulkarni G., Berg T., Berg A., Choi Y. Composing simple image descriptions using web-scale n-grams // Conference on Computational Natural Language Learning. Association for Computational Linguistics. 2011. P. 220-228.

43. Mitchell M., Dodge J., Goyal A., Yamaguchi K., Stratos K., Mensch A., Berg A., Berg T., Daume H. Midge: Generating Image Descriptions From Computer Vision Detections // The European Chapter of the Association for Computational Linguistics. 2012. P. 747-756.

44. Elliott D., Keller F. Image Description using Visual Dependency Representations // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013. P. 1292-1302.

45. Elliott D., Vries A. P. Describing Images using Inferred Visual Dependency Representations // Association for Computational Linguistics. 2015. P. 42-52.

46. Kulkarni G., Premraj V., Ordonez V., Dhar S., Li S., Choi Y., Berg A., Berg T. L. Baby talk: Understanding and generating simple image descriptions // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2013. Vol. 35 (12). P. 2891-2903.

47. Fang H., Gupta S., Iandola F. From Captions to Visual Concepts and Back // Conference on Computer Vision and Pattern Recognition. 2015.

48. Kiros R., Salakhutdinov R., Zemel R. Multimodal Neural Language Models // International Conference on Machine Learning. 2014. P. 595-603.

49. Mnih A., Hinton G. Three new graphical models for statistical language modelling // Proceedings of the 24th International Conference on Machine Learning. 2007. Vol. 62. P. 641-648.

50. Lin D., Kong C., Fidler S., Urtasun R. Generating Multi-Sentence Lingual Descriptions of Indoor Scenes // British Machine Vision Conference (BMVC). 2015. P. 1-13.

51. Karpathy A., Fei-Fei L. Deep visual-semantic alignments for generating image descriptions // Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2015.

52. Vinyals O., Toshev A., Bengio S., Erhan D. Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2016. Vol. 39 (4). P. 652-663.

53. Hochreiter S., Urgen Schmidhuber J. Long Short-Term Memory // Neural Computation. 1997. Vol. 9 (8). P. 1735-1780.

54. Chen X., Zitnick C. L. Learning a Recurrent Visual Representation for Image Caption // arXiv.org [Электронный ресурс]. 2014. – URL: https://arxiv.org/abs/1411.5654 (дата обращения: 01.02.2018).

55. Chen X., Zitnick C. L. Mind's eye: A recurrent visual representation for image caption generation // Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2015. P. 2422-2431.

56. Williams R. J., Zipser D. Experimental Analysis of the Real-Time Recurrent Learning Algorithm // Connection Science. 1989. Vol. 1 (1). P. 87-111.

57. Mao J., Xu W., Yang Y., Wang J., Yuille A. L. Explain Images with Multimodal Recurrent Neural Networks // arXiv.org [Электронный ресурс]. 2014. – URL: https://arxiv.org/abs/1410.1090 (дата обращения: 01.02.2018).

58. Mao J., Xu W., Yang Y., Wang J., Huang Z., Yuille A. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN) // International Conference on Learning Representations. 2015. Vol. 1090. P. 1-17.

59. Donahue J., Hendricks L. A., Rohrbach M., Venugopalan S., Guadarrama S., Saenko K., Darrell T. Long-term Recurrent Convolutional Networks for Visual Recognition and Description // Conference on Computer Vision and Pattern Recognition. 2015.

60. Johnson J., Karpathy A., Fei-Fei L. DenseCap: Fully Convolutional Localization Networks for Dense Captioning // Conference on Computer Vision and Pattern Recognition. 2015.

61. Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition // arXiv.org [Электронный ресурс]. 2014. – URL: https://arxiv.org/abs/1409.1556 (дата обращения: 01.02.2018).

62. Maron O., Lozano-Perez T. A Framework for Multiple-Instance Learning // Conference on Neural Information Processing Systems. 1998.

63. Wang C., Yang H., Bartz C., Meinel C. Image Captioning with Deep Bidirectional LSTMs // Proceedings of the 2016 Association for Computing Machinery on Multimedia Conference. 2016. P. 988-997.

64. Xu K., Ba J., Kiros R., Cho K., Courville A., Salakhutdinov R., Zemel R., Bengio Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention // International Conference on Machine Learning. 2015.

65. Bahdanau D., Cho K., Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate // International Conference on Learning Representations. 2014. P. 1-15.

66. Rennie S. J., Marcheret E., Mroueh Y., Ross J., Goel V. Self-critical Sequence Training for Image Captioning // Conference on Computer Vision and Pattern Recognition. 2016.

67. Sutton R. S., Barto G. Reinforcement learning: an introduction // University College London, Computer Science Department, Reinforcement Learning Lectures. 2017.

68. Pedersoli M., Lucas T., Schmid C., Verbeek J. Areas of Attention for Image Captioning // International Conference on Computer Vision. 2017.

69. Cho K., Van Merrienboer B., Bahdanau D., Bengio Y. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches // Syntax, Semantics and Structure in Statistical Translation. 2014.

70. Cornia M., Baraldi L., Serra G., Cucchiara R. Paying More Attention to Saliency: Image Captioning with Saliency and Context Attention // arXiv.org [Электронный ресурс]. 2017. – URL: https://arxiv.org/abs/1706.08474 (дата обращения: 01.02.2018).

71. Lu J., Xiong C., Parikh D., Socher R. Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning // Conference on Computer Vision and Pattern Recognition. 2017.

72. Kingma D. P., Ba J. Adam: A Method for Stochastic Optimization // International Conference on Learning Representations. 2015. P. 1-15.

73. Gu J., Cai J., Wang G., Chen T. Stack-Captioning: Coarse-to-Fine Learning for Image Captioning // Association for the Advancement of Artificial Intelligence. 2018.

74. Zhang L., Sung F., Liu F., Xiang T., Gong S., Yang Y., Hospedales T. M. Actor-Critic Sequence Training for Image Captioning // arXiv.org [Электронный ресурс]. 2017. – URL: https://arxiv.org/abs/1706.09601 (дата обращения: 01.02.2018).

75. Barto A. G., Sutton R. S., Anderson C. W. Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems // IEEE Transactions on Systems, Man and Cybernetics, SMC-13(5). 1983. P. 834-846.

76. Aneja J., Deshpande A., Schwing A. Convolutional Image Captioning // arXiv.org [Электронный ресурс]. 2017. – URL: https://arxiv.org/abs/1711.09151 (дата обращения: 01.02.2018).

77. Liu S., Zhu Z., Ye N., Guadarrama S., Murphy K. Improved Image Captioning via Policy Gradient optimization of SPIDEr // International Conference on Computer Vision. 2017.

78. Sutton R. S., Mcallester D., Singh S., Mansour Y. Policy Gradient Methods for Reinforcement Learning with Function Approximation // Advances in Neural Information Processing Systems. 1999. Vol. 12. P.1057-1063.

79. Kuznetsova P., Ordonez V., Berg A. C., Berg T. L., Choi Y., Brook S. Collective Generation of Natural Image Descriptions // Association for Computational Linguistics. 2012. P. 359-368.

80. Целочисленное программирование // Википедия: свободная энциклопедия [Электронный ресурс]. 23.01.2018. – URL:

https://ru.wikipedia.org/wiki/Целочисленное_программирование (дата обращения: 23.01.2018).

81. Gupta A., Verma Y., Jawahar C. V. Choosing Linguistics over Vision to Describe Images // Association for the Advancement of Artificial Intelligence. 2012. P. 606-612.

82. Ushiku Y., Yamaguchi M., Mukuta Y., Harada T. Common subspace for model and similarity: Phrase learning for caption generation from images // Proceedings of the IEEE International Conference on Computer Vision. 2015. P. 2668-2676.

83. Lebret R., Pinheiro P. O., Collobert R. Simple Image Description Generator via a Linear Phrase-Based Approach // International Conference on Learning Representations Workshop. 2015.

84. Lebret R., Pinheiro P. O., Collobert R. Phrase-based Image Captioning // International Conference on Machine Learning. 2015.

85. Reape M., Mellish C. Just what is aggregation anyway? // Proceedings of the 7th European Workshop on Natural Language Generation. 1999. P. 20-29.

86. Gatt A., Reiter E. SimpleNLG: A realisation engine for practical applications // Proceedings of the 12th European Workshop on Natural Language Generation, ENLG. 2009. P. 91-93.

87. Rashtchian C., Young P., Hodosh M., Hockenmaier J. Collecting Image Annotations Using Amazon's Mechanical Turk // Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. 2010. P. 139-147.

88. Young P., Lai A., Hodosh M., Hockenmaier J. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions // Transactions of the Association for Computational Linguistics (TACL). 2014. P. 67-78.

89. Chen X., Fang H., Lin T. Y., Vedantam R., Gupta S., Dollár P., Zitnick C. L. Microsoft COCO Captions: Data Collection and Evaluation Server // arXiv.org [Электронный ресурс]. 2015. - URL:. https://arxiv.org/abs/1504.00325 (дата обращения: 01.02.2018).

90. Zitnick C. L., Parikh D. Bringing semantics into focus using visual abstraction // Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2013. P. 3009-3016.

91. Feng Y., Lapata M. Automatic caption generation for news images // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2013. Vol. 35 (4). P. 797-812.

92. Chen J., Kuznetsova P., Warren D., Choi Y. Déjà Image-Captions: A Corpus of Expressive Descriptions in Repetition // Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. 2015. P. 504-514.

93. Everingham M., Van Gool L., Williams C. K. I., Winn J., Zisserman A. The pascal visual object classes (VOC) challenge // International Journal of Computer Vision. 2010.

94. Amazon Mechanical Turk [Электронный ресурс]. – URL: https://www.mturk.com/ (дата обращения: 01.02.2018).

95. Find your inspiration | Flickr [Электронный ресурс]. – URL: https://www.flickr.com/ (дата обращения: 01.02.2018).

96. Lin T. Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Dollár P., Zitnick C. L. Microsoft COCO: Common objects in context // European conference on computer vision. 2014. P.740-755.

97. Papineni K., Roukos S., Ward T., Zhu W. BLEU: a method for automatic evaluation of machine translation // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 2002. P. 311-318.

98. Lin C. Y. Rouge: A package for automatic evaluation of summaries // Proceedings of the Workshop on Text Summarization Branches out (WAS 2004). 2004. Vol. 1. P. 25-26.

99. Denkowski M., Lavie A. Meteor Universal: Language Specific Translation Evaluation for Any Target Language // Proceedings of the Ninth Workshop on Statistical Machine Translation. 2014. P. 376-380.

100. METEOR (Metric for Evaluation of Translation with Explicit ORdering) // Википедия: свободная энциклопедия [Электронный ресурс]. 23.01.2018. – URL: https://ru.wikipedia.org/wiki/METEOR (дата обращения: 23.01.2018).

101. Vedantam R., Zitnick C. L., Parikh D. CIDEr: Consensus-based image description evaluation // Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2015. P. 4566-4575.

102. Anderson P., Fernando B., Johnson M., Gould S. SPICE: Semantic propositional image caption evaluation // Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9909 LNCS. 2016. P. 382-398.

103. Automatic summarization // Википедия: свободная энциклопедия [Электронный ресурс]. 23.01.2018. – URL: https://en.wikipedia.org/w/index.php?title=Automatic_summarization (дата обращения: 23.01.2018).

104. George A. Miller. WordNet: A Lexical Database for English // Communications of the Association for Computing Machinery. 1995. Vol. 38. № 11. P. 39-41.

105. Robertson S. Understanding inverse document frequency: on theoretical arguments for IDF // Journal of Documentation. 2004. № 60 (5). P. 503-520.

106. Schuster S., Krishna R., Chang A., Fei-Fei L., Manning C. D. Generating Semantically Precise Scene Graphs from Textual Descriptions for Improved Image Retrieval // Conference on Empirical Methods in Natural Language Processing. 2015. P. 70-80.

107. Elliott D., Keller F. Comparing Automatic Evaluation Measures for Image Description // Association for Computational Linguistics. 2014. P. 452-457.

References

1. YouTube for Press. *Youtube*, 2018. Available at: https://www.youtube.com/yt/about/press (accessed: 01 February 2018).

2. Borisov V. V., Korshunova K. P. Direct and Reverse Image Captioning problem definition. *Postanovka priamoi i obratnoi zadachi poiska i generirovaniia*

tekstovykh opisanii po izobrazheniiam. Energetika, informatika, innovatsii - 2017 (elektroenergetika, elektrotekhnika i teploenergetika, matematicheskoe modelirovanie i informatsionnye tekhnologii v proizvodstve). [Power engineering, computer science, innovations - 2017. Proceedings of the VII international scientific conference]. Smolensk, 2017, pp. 228-230 (in Russian).

3. Abella A., Kender J. R., Starren J. Description Generation of Abnormal Densities found in Radiographs. *Proc. Symp. Computer Applications in Medical Care, Journal of the American Medical Informatics Association*, 1995, pp. 542-546.

4. Gerber R., Nagel N. H. Knowledge representation for the generation of quantified natural language descriptions of vehicle traffic in image sequences. *Proceedings of the International Conference on Image Processing*, 1996, pp. 805-808.

5. Szeliski R. Computer Vision: Algorithms and Applications. *Springer Science & Business Media*, 2010, pp. 10-16.

6. Dale R., White M. Shared Tasks and Comparative Evaluation in Natural Language Generation: Position Papers, Arlington, VA, USA, 2007.

7. Reiter E., Belz A. An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems. *Computational Linguistics*, 2009, no. 35 (4), pp. 529-558.

8. Bernardi R., Cakici R., Elliott D., Erdem A., Erdem E., Ikizler-Cinbis N., Keller F., Muscat A., Plank B. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *IJCAI International Joint Conference on Artificial Intelligence*, 2017, pp. 4970-4974.

9. Kuznetsova P., Ordonez V., Berg T., Choi Y. TREETALK: Composition and Compression of Trees for Image Descriptions. *Transactions of the Association for Computational Linguistics*, 2014, Vol. 2, pp. 351-362.

10. Kiros R., Salakhutdinov R., Zemel R. S. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *Association for Computational Linguistics*, 2014.

11. Yao T., Pan Y., Li Y., Qiu Z., Mei T. Boosting Image Captioning with Attributes. *International Conference on Computer Vision*, 2017, pp. 1-11.

12. Mel'nichenko A. S. Automatic image annotationing based on global features. *Avtomaticheskaia annotatsiia izobrazhenii na osnove global'nykh priznakov. Izvestiia IuFU. Tekhnicheskie nauki.* [Bulletin of The Southern Federal University. Technical science], 2009, no. 9, pp.189-200 (in Russian).

13. Kurbatov S. S., Naidenova K. A., Khakhalin G. K. On the scheme of interaction in the complex «analysis and synthesis of natural language and images». *O skheme vzaimodeistviia v komplekse «analiz i sintez estestvennogo iazyka i izobrazhenii» // KII-2010. Trudy konferentsii.* [Proceedings of the CAI - Conference on Artificial intelligence]. Tver, 2010, pp. 234-242 (in Russian).

14. Proskurin A. V. Automatic landscape image annotation. *Avtomaticheskoe annotirovanie landshaftnykh izobrazhenii. Sibirskii zhurnal nauki i tekhnologii.* [Siberian Journal of Science and Technology]. 2014, no. 3 (55), pp.120-125 (in Russian).

15. Ordonez V., Kulkarni G., Berg T. L. Im2Text: Describing Images Using 1 Million Captioned Photographs. *Neural Information Processing Systems*, 2011, pp. 1143-1151.

16. Mason R., Charniak E. Nonparametric Method for Data-driven Image Captioning. *Association for Computational Linguistics*, 2014, pp. 592-598.

17. Yagcioglu S., Erdem E., Erdem A. A Distributed Representation Based Query Expansion Approach for Image Captioning. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 2015, Vol. 1, pp. 106-111.

18. Ordonez V., Han X., Kuznetsova P., Kulkarni G., Mitchell M., Yamaguchi K., Stratos K., Goyal A., Dodge J., Mensch A., Daume H., Berg A. C., Choi Y., Berg T. L. Large Scale Retrieval and Generation of Image Descriptions. *International Journal of Computer Vision*, 2016.

19. Devlin J., Gupta S., Girshick R., Mitchell M., Zitnick C. L. Exploring Nearest Neighbor Approaches for Image Captioning. *arXiv.org*, 2015. Available at: http://arxiv.org/abs/1505.04467 (accessed: 01 February 2018).

20. Farhadi A., Hejrati M., Sadeghi M. A., Young P., Rashtchian C., Hockenmaier J., Forsyth D. Every picture tells a story: Generating sentences from images. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2010, 6314 LNCS (PART 4), pp. 15-29.

21. Socher R., Karpathy A., Le Q. V, Manning C. D., Ng A. Y. Grounded Compositional Semantics for Finding and Describing Images with Sentences. *Transactions of the Association for Computational Linguistics*, 2014, pp. 207-218.

22. Hodosh M., Young P., Hockenmaier J. Framing image description as a ranking task: Data, models and evaluation metrics. *IJCAI International Joint Conference on Artificial Intelligence*, 2013, pp. 4188-4192.

23. Verma Y., Jawahar C. V. Im2Text and Text2Im : Associating Images and Texts for Cross-Modal Retrieval. *British Machine Vision Conference*, 2014.

24. Karpathy A., Joulin A., Fei-Fei L. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. *Neural Information Processing Systems*, 2014.

25. Gong Y., Wang L., Hodosh M., Hockenmaier J., Lazebnik S. Improving image-sentence embeddings using large weakly annotated photo collections. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014, 8692 LNCS (PART 4), pp. 529-545.

26. Reed S., Akata Z., Schiele B., Lee H. Learning Deep Representations of Fine-grained Visual Descriptions. *Conference on Computer Vision and Pattern Recognition*, 2016.

27. Plummer B. A., Wang L., Cervantes C. M., Caicedo J. C., Hockenmaier J., Lazebnik S. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. *International Journal of Computer Vision*, no. 123 (1), 2017, pp. 74-93.

28. Oliva A., Hospital W., Ave L. Modeling the Shape of the Scene : A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, no. 42 (3), 2001, pp. 145-175.

29. TF-IDF. *Wikipedia*, 2018. Available at: http://ru.wikipedia.org/?oldid=90463885 (accessed: 01 February 2018) (in Russian).

30. Felzenszwalb P., McAllester D., Ramanan D. A Discriminatively Trained, Multiscaled, Deformable Part Model. *Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1-8.

31. Hoiem D., Divvala S., Hays J. Pascal VOC 2009 Challenge. *PASCAL challenge workshop in European Conference on Computer Vision*, 2009.

32. Curran J., Clark S., Bos J. Linguistically motivated large-scale NLP with C&C and boxer. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics on Interactive Poster and Demonstration Sessions*, 2007, pp. 33-36.

33. Varma M., Zisserman A. A statistical approach to texture classification from single images. *International Journal of Computer Vision*, 2005, no. 62, pp. 61-81.

34. Tsochantaridis I., Hofmann T., Joachims T., Altun Y. Support vector machine learning for interdependent and structured output spaces. *Proc. Intl. Conf. Machine Learning*, 2004.

35. De Marneffe M.-C., Maccartney B., Manning C. D. Generating Typed Dependency Parses from Phrase Structure Parses. *International Conference on Language Resources and Evaluation*, 2006, vol. 6, pp. 449-454.

36. Girshick R., Donahue J., Darrell T., Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014.

37. Vinyals O., Toshev A., Bengio S., Erhan D. Show and Tell: A Neural Image Caption Generator. *Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1-10.

38. Yao B. Z., Yang X., Lin L., Lee M. W., Zhu S. C. I2T: Image parsing to text description. *Proceedings of the IEEE*, 2010, no. 98 (8), pp. 1485-1508.

39. Chen H., Xu Z. J., Liu Z. Q., Zhu S. C. Composite templates for cloth modeling and sketching. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006, vol. 1, pp. 943-950.

40. Berners-Lee T., Hendler J., Lassila O. The Semantic Web. *Scientific American*, 2001, no. 284, pp. 29-37.

41. Yang Y., Teo C. L., Daume H., Aloimonos Y. Corpus-Guided Sentence Generation of Natural Images. *Proceedings of The Conference on Empirical Methods on Natural Language Processing*, 2011, pp. 444-454.

42. Li S., Kulkarni G., Berg T., Berg A., Choi Y. Composing simple image descriptions using web-scale n-grams. *Conference on Computational Natural Language Learning. Association for Computational Linguistics*, 2011, pp. 220-228.

43. Mitchell M., Dodge J., Goyal A., Yamaguchi K., Stratos K., Mensch A., Berg A., Berg T., Daume H. Midge: Generating Image Descriptions From Computer Vision Detections. *The European Chapter of the Association for Computational Linguistics*, 2012, pp. 747-756.

44. Elliott D., Keller F. Image Description using Visual Dependency Representations. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1292-1302.

45. Elliott D., Vries A. P. Describing Images using Inferred Visual Dependency Representations. *Association for Computational Linguistics*, 2015, pp. 42-52.

46. Kulkarni G., Premraj V., Ordonez V., Dhar S., Li S., Choi Y., Berg A., Berg T. L. Baby talk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, Vol. 35 (12), pp. 2891-2903.

47. Fang H., Gupta S., Iandola F. From Captions to Visual Concepts and Back. *Conference on Computer Vision and Pattern Recognition*, 2015.

48. Kiros R., Salakhutdinov R., Zemel R. Multimodal Neural Language Models. *International Conference on Machine Learning*, 2014, pp. 595-603.

49. Mnih A., Hinton G. Three new graphical models for statistical language modelling. *Proceedings of the 24th International Conference on Machine Learning*, 2007, Vol. 62, pp. 641-648.

50. Lin D., Kong C., Fidler S., Urtasun R. Generating Multi-Sentence Lingual Descriptions of Indoor Scenes. *British Machine Vision Conference (BMVC)*, 2015, pp. 1-13.

51. Karpathy A., Fei-Fei L. Deep visual-semantic alignments for generating image descriptions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015.

52. Vinyals O., Toshev A., Bengio S., Erhan D. Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, Vol. 39 (4), pp. 652-663.

53. Hochreiter S., Urgen Schmidhuber J. Long Short-Term Memory. *Neural Computation*, 1997, Vol. 9 (8), pp. 1735-1780.

54. Chen X., Zitnick C. L. Learning a Recurrent Visual Representation for Image Caption. *arXiv.org*, 2015. Available at: https://arxiv.org/abs/1411.5654 (accessed: 01 February 2018).

55. Chen X., Zitnick C. L. Mind's eye: A recurrent visual representation for image caption generation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2422-2431.

56. Williams R. J., Zipser D. Experimental Analysis of the Real-Time Recurrent Learning Algorithm. *Connection Science*, 1989, no. 1 (1), pp. 87-111.

57. Mao J., Xu W., Yang Y., Wang J., Yuille A. L. Explain Images with Multimodal Recurrent Neural Networks. *arXiv.org*, 2014. Available at: https://arxiv.org/abs/1410.1090 (accessed: 01 February 2018).

58. Mao J., Xu W., Yang Y., Wang J., Huang Z., Yuille A. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). *International Conference on Learning Representations*, 2015, no. 1090, pp. 1-17.

59. Donahue J., Hendricks L. A., Rohrbach M., Venugopalan S., Guadarrama S., Saenko K., Darrell T. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. *Conference on Computer Vision and Pattern Recognition*, 2015.

60. Johnson J., Karpathy A., Fei-Fei L. DenseCap: Fully Convolutional Localization Networks for Dense Captioning. *Conference on Computer Vision and Pattern Recognition*, 2015.

61. Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv.org*, 2014. Available at: https://arxiv.org/abs/1409.1556 (accessed: 01 February 2018).

62. Maron O., Lozano-Perez T. A Framework for Multiple-Instance Learning. *Conference on Neural Information Processing Systems*, 1998.

63. Wang C., Yang H., Bartz C., Meinel C. Image Captioning with Deep Bidirectional LSTMs. *Proceedings of the 2016 Association for Computing Machinery on Multimedia Conference*, 2016, pp. 988-997.

64. Xu K., Ba J., Kiros R., Cho K., Courville A., Salakhutdinov R., Zemel R., Bengio Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *International Conference on Machine Learning*, 2015.

65. Bahdanau D., Cho K., Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. *International Conference on Learning Representations*, 2014, pp. 1-15.

66. Rennie S. J., Marcheret E., Mroueh Y., Ross J., Goel V. Self-critical Sequence Training for Image Captioning. *Conference on Computer Vision and Pattern Recognition*, 2016.

67. Sutton R. S., Barto G. Reinforcement learning: an introduction. *University College London, Computer Science Department, Reinforcement Learning Lectures*, 2017.

68. Pedersoli M., Lucas T., Schmid C., Verbeek J. Areas of Attention for Image Captioning. *International Conference on Computer Vision*, 2017.

69. Cho K., Van Merrienboer B., Bahdanau D., Bengio Y. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *Syntax, Semantics and Structure in Statistical Translation*, 2014.

70. Cornia M., Baraldi L., Serra G., Cucchiara R. Paying More Attention to Saliency: Image Captioning with Saliency and Context Attention. *arXiv.org*, 2015. Available at: https://arxiv.org/abs/1706.08474 (accessed: 01 February 2018).

71. Lu J., Xiong C., Parikh D., Socher R. Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning. *Conference on Computer Vision and Pattern Recognition*, 2017.

72. Kingma D.P., Ba J. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*, 2015, pp. 1-15.

73. Gu J., Cai J., Wang G., Chen T. Stack-Captioning: Coarse-to-Fine Learning for Image Captioning. *Association for the Advancement of Artificial Intelligence*, 2018.

74. Zhang L., Sung F., Liu F., Xiang T., Gong S., Yang Y., Hospedales T. M. Actor-Critic Sequence Training for Image Captioning. *arXiv.org*, 2015. Available at: https://arxiv.org/abs/1706.09601 (accessed: 01 February 2018).

75. Barto A. G., Sutton R. S., Anderson C. W. Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems. *IEEE Transactions on Systems, Man and Cybernetics*, 1983, SMC-13(5), pp. 834-846.

76. Aneja J., Deshpande A., Schwing A. Convolutional Image Captioning. *arXiv.org*, 2015. Available at: https://arxiv.org/abs/1711.09151 (accessed: 01 February 2018).

77. Liu S., Zhu Z., Ye N., Guadarrama S., Murphy K. Improved Image Captioning via Policy Gradient optimization of SPIDEr. *International Conference on Computer Vision*, 2017.

78. Sutton R. S., Mcallester D., Singh S., Mansour Y. Policy Gradient Methods for Reinforcement Learning with Function Approximation. *Advances in Neural Information Processing Systems*, 1999, vol. 12, pp.1057-1063.

79. Kuznetsova P., Ordonez V., Berg A. C., Berg T. L., Choi Y., Brook S. Collective Generation of Natural Image Descriptions. *Association for Computational Linguistics*, 2012, pp. 359-368.

80. Integer programming *Wikipedia*, 2018. Available at: https://en.wikipedia.org/wiki/Integer_programming (accessed: 01 February 2018) (in Russian).

81. Gupta A., Verma Y., Jawahar C. V. Choosing Linguistics over Vision to Describe Images. *Association for the Advancement of Artificial Intelligence*, 2012, pp. 606-612.

82. Ushiku Y., Yamaguchi M., Mukuta Y., Harada T. Common subspace for model and similarity: Phrase learning for caption generation from images. *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2668-2676.

83. Lebret R., Pinheiro P. O., Collobert R. Simple Image Description Generator via a Linear Phrase-Based Approach. *International Conference on Learning Representations Workshop*, 2015.

84. Lebret R., Pinheiro P. O., Collobert R. Phrase-based Image Captioning. *International Conference on Machine Learning*, 2015.

85. Reape M., Mellish C. Just what is aggregation anyway? *Proceedings of the 7th European Workshop on Natural Language Generation*, 1999, pp. 20-29.

86. Gatt A., Reiter E. SimpleNLG: A realisation engine for practical applications. *Proceedings of the 12th European Workshop on Natural Language Generation*, *ENLG*, 2009, pp. 91-93.

87. Rashtchian C., Young P., Hodosh M., Hockenmaier J. Collecting Image Annotations Using Amazon's Mechanical Turk. *Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk,* 2010, pp. 139-147.

88. Young P., Lai A., Hodosh M., Hockenmaier J. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions. *Transactions of the Association for Computational Linguistics (TACL)*, 2014, pp. 67-78.

89. Chen X., Fang H., Lin T. Y., Vedantam R., Gupta S., Dollár P., Zitnick C. L. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv.org*, 2015. Available at: https://arxiv.org/abs/1504.00325 (accessed: 01 February 2018).

90. Zitnick C.L., Parikh D. Bringing semantics into focus using visual abstraction. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3009-3016.

91. Feng Y., Lapata M. Automatic caption generation for news images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, no. 35 (4), pp. 797-812.

92. Chen J., Kuznetsova P., Warren D., Choi Y. Déjà Image-Captions: A Corpus of Expressive Descriptions in Repetition. *Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2015, pp. 504-514.

93. Everingham M., Van Gool L., Williams C. K. I., Winn J., Zisserman A. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 2010.

94. Amazon Mechanical Turk, 2018. Available at: https://www.mturk.com/ (accessed: 01 February 2018).

95. Find your inspiration. *Flickr*, 2018. Available at: https://www. flickr.com (accessed: 01 February 2018).

96. Lin T. Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Dollár P., Zitnick C. L. Microsoft COCO: Common objects in context. *European conference on computer vision*, 2014, pp.740-755.

97. Papineni K., Roukos S., Ward T., Zhu W. BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311-318.

98. Lin C. Y. Rouge: A package for automatic evaluation of summaries. *Proceedings of the Workshop on Text Summarization Branches out (WAS 2004)*, 2004, vol. 1, pp. 25-26.

99. Denkowski M., Lavie A. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 2014, pp. 376-380.

100. METEOR (Metric for Evaluation of Translation with Explicit ORdering). *Wikipedia*, 2018. Available at: https://ru.wikipedia.org/wiki/METEOR (accessed: 01 February 2018).

101. Vedantam R., Zitnick C. L., Parikh D. CIDEr: Consensus-based image description evaluation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4566-4575.

102. Anderson P., Fernando B., Johnson M., Gould S. SPICE: Semantic propositional image caption evaluation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9909 LNCS, 2016, pp. 382-398.

103. Automatic summarization. *Wikipedia*, 2018. Available at: https://en.wikipedia.org/w/index.php?title=Automatic_summarization&oldid=822496672 (accessed: 01 February 2018).

104. George A. Miller. WordNet: A Lexical Database for English. *Communications of the Association for Computing Machinery*, 1995, vol. 38, no. 11, pp. 39-41.

105. Robertson S. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, 2004, no.60(5), pp. 503-520.

106. Schuster S., Krishna R., Chang A., Fei-Fei L., Manning C. D. Generating Semantically Precise Scene Graphs from Textual Descriptions for Improved Image

Retrieval. Conference on Empirical Methods in Natural Language Processing, 2015, pp. 70-80.

107. Elliott D., Keller F. Comparing Automatic Evaluation Measures for Image Description. *Association for Computational Linguistics*, 2014, pp. 452-457.

Статья поступила 13 февраля 2018 г.

Информация об авторах

Коршунова Ксения Петровна – соискатель ученой степени кандидата технических наук. Аспирант кафедры вычислительной техники. Филиал ФГБОУ ВО «Национальный исследовательский университет «МЭИ» в г. Смоленске. Область научных интересов: искусственный интеллект; машинное обучение; интеллектуальная поддержка принятия решений. E-mail: ksenya-kor@mail.ru

Адрес: Россия, 14013, г. Смоленск, Энергетический проезд, дом 1.

Automatic Image Captioning: Tasks and Methods

K. P. Korshunova

Problem definition. *Automatic Image Captioning is a challenging multimadal problem of artificial intelligence that require processing of combination of visual and linguistic information. These tasks are quite new and the studies on the subject have certain contradictions: first of all, there are no sustainable system of terms and definitions, secondly, there are no good classification of the tasks and approaches for image captioning.* ***Purpose.*** *The purpose of the present paper is to analyze the current state-of-the-art knowledge on automatic image captioning, to propose a system of terms and definitions for Russian-language researches and developments and to propose tasks and methods classification.* ***Results.*** *We analyzed many various papers and classify image captioning tasks: automatic image annotation (providing a set of key words), image description retrieval (searching the best description in a database) and image description generation (framing a novel description for the certain image). We provide a methods classification: retrieval (that includes image retrieval and description retrieval), generative and hybrid methods. We provide a detailed review of existing retrieval, generative and hybrid methods, highlighting main trends of their development. We give an overview of the public available datasets and evaluation measures.* ***Practical relevance.*** *The presented paper is the first detailed survey of the tasks, methods and approaches of Automatic Image Captioning in Russian. The results can be used for investigating current state-of-the-art methods and development novel Automatic Image Captioning approaches.*

Key words: *automatic image captioning, image description generation, image description retrieval, automatic image annotation, multimodal tasks, deep neural networks, machine learning, artificial intelligence.*

Information about Author

Kseniya Petrovna Korshunova – The postgraduate student of the Dept of Computer Engineering. The Branch of National Research University «Moscow Power Engineering Institute» in Smolensk. Field of research: artificial intelligence, machine learning, intellectual decision-making support. E–mail: ksenya-kor@mail.ru

Address: Russia, 214013, Smolensk, Energeticheskiy proezd, 1.