**СОДЕРЖАНИЕ**

[**ВВЕДЕНИЕ** 3](#_Toc184504993)

[**ГЛАВА 1** 4](#_Toc184504994)

[**1.1 Предмет анализа** 4](#_Toc184504995)

[**1.2 Машинное зрение** 4](#_Toc184504996)

[**1.3 Генерация текста** 5](#_Toc184504997)

[**1.4 Классические подходы решения проблемы базируются на 3-х шагах:** 5](#_Toc184504998)

[**1.5 Где может использоваться машинное зрение и генерация текста?** 7](#_Toc184504999)

# **ВВЕДЕНИЕ**

**Актуальность** обоснована ростом визуального контента, чей спад не будет наблюдаться в ближайшее время. Анализ большого количества данных стал рутинным и сложным делом. Для такого необходим инструментарий в виде приложения для обработки визуального контента и его точного описания. Данные технологии необходимы во множестве отраслях, науке и сфере услуг и развлечений.

Технология image captioning широко распространена в современном мире. Без неё не уже сложно представить жизнь любого человека даже в моменты, когда напрямую не пользуется ей, так как технология является полезным и важным инструментом.

**Цель:** разработка приложения для автоматической генерации текстового описания изображения.

**Объект исследования:** нейросети, машинное зрение, методология работы с генерацией естественного текста.

**Предмет исследования:** технология автоматической генерации текста, технология машинного зрения, технология автоматической генерации текста по изображению.

**Задача:**

- изучение предметной области в сфере автоматической генерации текстового описания изображения

- анализ уже существующих решений

- создание приложения для автоматической генерации текста по изображению.

# **ГЛАВА 1**

## **1.1 Предмет анализа**

Автоматическая генерация текстового описания изображения (Image Captioning) – технология описания изображения посредством методов машинного зрения.

Дальнейшая теория основана на ресурсе Papers with Code и его широкой базе данных различных исследований и статей в области машинного обучения. В рамках курсовой работы будут взяты статьи по теме Image Captioning и «наборов с перекрёстными описаниями COCO (Crisscrossed Captions, CxC).

## **1.2 Машинное зрение**

Одно из первых упоминаний для широких масс людей технология получила 14 апреля 2015 года вместе с выходом статьи «From Captions to Visual Concepts and Back», спонсированная Microsoft Research (MSR), что специализируется на исследованиях в области информатики. Текст статьи подробно раскрывает концепцию и логику работы машинного зрения. Это было важным витком развития, так как технология стала активно развиваться и интересовать не только научных деятелей, но и общественность.

Основная цель для создания и изучения машинного зрения – получения ключевых слов, которые характеризуют изображение. А именно процесс проходит следующим образом:

- **Извлечение признаков:** Машинное зрение позволяет извлекать важные признаки из изображений, такие как цвета, текстуры, формы и объекты. Эти признаки могут быть использованы для дальнейшего анализа и кластеризации.

- **Обнаружение объектов:** С помощью алгоритмов обнаружения объектов (например, YOLO, Faster R-CNN) можно идентифицировать и классифицировать объекты на изображении. Каждое обнаруженное событие может быть связано с ключевым словом.

- **Классификация изображений:** Модели глубокого обучения, такие как сверточные нейронные сети (CNN), могут быть обучены для классификации изображений по категориям. Каждая категория может быть представлена набором ключевых слов.

## **1.3 Генерация текста**

Данная статья ссылается на исследования UMIACS в другой статье «Corpus-Guided Sentence Generation of Natural Images», в которой подробно описан процесс автоматической генерации текста (Speech/Text Generation).

Серьёзной проблемой для генерации текста всегда является Natural language generation (NLG) или же естественное понимание языка. При генерации текста важно не только определить ключевые слова, но и составить в логическую правильную последовательность с соблюдением норм любого человеческого языка.

## **1.4 Классические подходы решения проблемы базируются на 3-х шагах:**

- выбор;

- планирование;

- реализация;

Точнее всего первый шаг изображён в самой статье   
в виде таблицы (рис. 1)

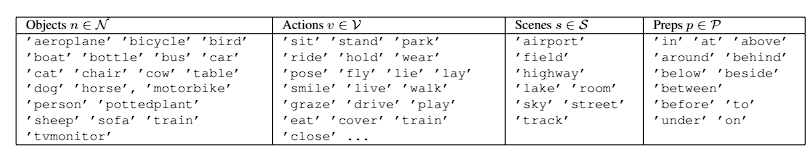


Рис. 1 Выбор подходящих слов под генерацию

Для описания множества образцов изображений необходимо понимать основные переменные, из которых и будет складываться строка описания:

- Объекты на изображении;

- Действия, которые происходят на изображении;

- Окружение сцены и самих объектов;

- Грамматически необходимые слова. В данном случае статья и пример на английском языке;

Все они содержатся в единой структуре T∗ = {n ∗ , v∗ , s∗ , p∗} (визуально проще воспринимать в виде таблицы, как на рис. 1).

На этапе планирования создается структура предложения, основанная на выбранных словах. Это включает в себя определение грамматических ролей слов и их последовательности в предложении, что критично для естественного звучания текста.

На этапе реализации создаётся структурированный естественный для восприятия человеком текст. Для этого используются различные способы, например:

- **Шаблонные методы:** Используют заранее заданные структуры предложений, такие как "Объект + Действие + Место", что позволяет генерировать текст на основании шаблонов;

- **Грамматические модели:** Алгоритмы, которые применяют правила грамматики для построения синтаксически верных предложений;

**- Статистические модели:** Развиваются на основе вероятностных методов;

- **N-Граммовые модели:** Определяют вероятность появления слова на основе N предыдущих слов. Эти модели широко использовались для обработки текстов до появления глубокого обучения;

- **Hidden Markov Models:** Модели для обработки последовательных данных, используются для генерации текстов из скрытых элементов, таких как части речи;

- **Модели на основе машинного обучения**: Эти модели начали доминировать в последние годы. Они требуют больших затрат на создание и поддержание, но показывают достаточно высокую эффективность относительно многих моделей;

- **LSTM (Long Short-Term Memory):** Рекуррентные нейронные сети, которые обеспечивают сохранение контекста в длительных последовательностях при генерации текста;

- **GRU (Gated Recurrent Unit):** Подобные LSTM модели, но с меньшим количеством параметров, которые также эффективны для генерации текста;

- **Глубокие нейронные сети и трансформеры:** Современные и наиболее продвинутые подходы;

- **Трансформеры:** Используют механизм внимания для обработки последовательностей информации, что позволяет моделям эффективно учитывать контекст. Примеры включают модели, такие как BERT и GPT (Generative Pre-trained Transformer), которые демонстрируют выдающиеся результаты в генерации естественного языка;

- **Генеративные модели:** Модели, которые фокусируются на создании нового контента;

- **GANs (Generative Adversarial Networks):** Характеризуются двумя компонентами — генератор и дискриминатор; применяются для создания правдоподобных текстов или изображений;

## **1.5 Где может использоваться машинное зрение и генерация текста?**

Вместе или раздельно данные технологии могут использоваться в различных системах. Такие есть как зарубежные, так и российские, например:

- **Google Cloud Vision API:** компания Google предоставляет возможность пользоваться технологией image captioning через свои серверы с помощью API. Данная услуга широко распространена и ей пользуется большое количество компаний в своих продуктах для автоматической разметки изображений, повышения эффективности функций поиска и других задач;

- **Социальные сети:** VK, а также запрещённые на территории Российской Федерации Facebook, Instagram используют технологии image captioning для генерации описаний к фото, определения пользователей на фотографиях. Pinterest использует технологию для создания описаний к изображению при сохранении их на свои «доски»;

- **Медицина:** аппараты МРТ или устройства анализа рентгеновских снимков используют технологию image captioning для генерации описания потенциальных заболеваний;

- **AI в роботах:** такие компании, как Tesla или Yandex разрабатывают системы анализа окружения в реальном времени. Для этого используются и совершенствуются методы и алгоритмы технологии image captioning;

Данная технология может использоваться не только в узконаправленных сферах науки, но и в сфере развлечения. Подобные технологии предоставляет в качестве услуг компания OpenAI. Чат-бот Chat GPT умеет распознавать изображения и давать текстовое описание. Реализован бот на сайте. Данный функционал возможно реализовать и в приложении;

# **ГЛАВА 2**

## **2.1 В ходе анализа данных по теме image captioning выяснилось:**

Для создания приложения с функционалом этой технологии необходимо извлекать ключевые данные из изображения (как на рис. 1), затем преобразовывать в текст.

Для такого наиболее удачным вариантом будет язык программирования Python. Внутри обширное количество библиотек, где уже реализованы данные возможности.

## **2.2 Для этого будут использованы ключевые библиотеки:**

- **Pillow** для обработки изображений;

Pillow и его предшественник PIL — это оригинальные библиотеки Python для работы с изображениями. Несмотря на то, что существуют другие библиотеки Python для обработки изображений, Pillow остается важным инструментом для понимания и работы в целом. Для оперирования и обработки изображений Pillow предоставляет инструменты, аналогичные тем, которые можно найти в программном обеспечении, таком как Photoshop. Некоторые из более современных библиотек обработки изображений Python построены на основе Pillow и часто предоставляют более продвинутую функциональность.

- **ImageAI** для компьютерного зрения;

ImageAI — это простая, но очень мощная и продвинутая библиотека Python с открытым исходным кодом, которая дает разработчикам программного обеспечения возможность разрабатывать приложения и программные утилиты с автономными возможностями глубокого обучения и компьютерного зрения.

- **Transformers** для получения текста на естественном языке;

Transformers — это мощная библиотека Python, созданная Hugging Face, которая позволяет вам загружать, управлять и запускать тысячи предварительно обученных моделей ИИ с открытым исходным кодом.

- **Tkinter** для создания графического интерфейса приложения;

Tkinter – это пакет для Python, предназначенный для работы с библиотекой Tk. Библиотека Tk содержит компоненты графического интерфейса пользователя (graphical user interface – GUI). Эта библиотека написана на языке программирования Tcl.

## **2.3 Для реализации работы функционального GUI необходимо:**

- Поле или кнопка, куда будет заносится изображение;

- Поле для вывода строки текста;

- Возможность использовать «Обзор» для упрощения внесения изображения в приложение;

## **2.4 Общий функционал должен включать:**

- Загрузка и вывод изображения;

- Обработка изображения для последующего получения объекта с необходимыми ключевыми словами;

- Генерация текста на естественном языке на основе полученных данных с объекта;

- Вывод текста и в GUI пользователю;

- Внесения сразу нескольких изображений для упрощения процесса внесения данных изображений в приложение;

**ГЛАВА 3**

*1 глава анализ теории,решений, требования к своей работе*

*2 глава: концепция*

*Выбор инструментов для создания*

*Дизайн интерфейса*

*3 глава:*

*Пояснение основных кусков кода*

*Инструкция как пользоваться продуктом*

*В заключении цель достигнута или нет и с каким результатом решены задачи*

*Список литературы ссылка, название, автор/коллектив авторов, дата на день просмотра*