УДК 004.8

ПРИМЕНЕНИЕ VLM ДЛЯ АНАЛИЗА ВИЗУАЛЬНЫХ СЦЕН

В.В. Венгеренко
Младший научный сотрудник государственного научного учреждения «Объединенный институт проблем информатики Национальной академии наук Беларуси»
vengerenko@lsi.bas-net.by

В.В. Венгеренко
Является аспирантом ОИПИ НАН Беларуси. Область научных интересов связана с разработкой методов и алгоритмов машинного обучения для решения задач анализа данных и компьютерного зрения.

Аннотация. Визуально-языковые модели (Vision-Language Models, VLM) представляют собой сложное достижение в области искусственного интеллекта (ИИ, Artificial Intelligence, AI), объединяющее возможности как компьютерного зрения (Computer Vision, CV), так и обработки естественного языка (Natural Language Processing, NLP) для обеспечения более целостного понимания данных. В отличие от традиционных моделей, которые фокусируются на одном типе входных данных – визуальном или текстовом, – VLM предназначены для обработки и понимания мультимодальных данных, объединяя визуальную и текстовую информацию для формирования более содержательных выводов.

Цель работы – определить особенности использования VLM для анализа визуальных сцен.

Рассмотрены ключевые компоненты VLM. Исследованы существующие подходы к анализу визуальных сцен на основе VLM. Выполнен обзор методов оценки качества таких моделей.

Ключевые слова: искусственный интеллект, глубокое обучение, языковая модель, модальность, токен.

Введение. За последнее десятилетие в области компьютерного зрения произошел переход от приложений, основанных на восприятии, таких как классификация изображений и обнаружение объектов, к более сложным задачам, таким как понимание сцены, реляционное моделирование и ответы на визуальные вопросы (Visual Question Answering, VQA) [1]. Но вопрос о том, как обеспечить системам визуального контроля истинную способность к рассуждению, до сих пор остается нерешенным.

Визуальное рассуждение (visual reasoning) выходит далеко за рамки распознавания объектов и предполагает выводы, основанные на атрибутах объектов, их отношениях и причинно-следственных зависимостях в пространстве и времени. Оно служит важным связующим звеном между низкоуровневым восприятием и высокоуровневым познанием. Несмотря на значительные успехи, сохраняются серьезные пробелы в объяснении, обобщении между разными предметными областями и способности решать сложные задачи.

Последние достижения в области VLM продемонстрировали все более сложные возможности в интерпретации и анализе визуального контента.

Эти модели могут выполнять сложные задачи, такие как ответы на открытые визуальные вопросы, понимание документов и мультимодальный диалог.

Задача исследования состоит в обзоре существующих подходов к анализу визуальных сцен на основе VLM.

Компоненты VLM. VLM – это мультимодальные, генеративные модели ИИ, способные понимать и обрабатывать видео, изображения и текст [2].

Большинство VLM соответствуют архитектуре из трех частей (рисунок 1):

– визуальный энкодер (vision encoder);

– проектор (projector);

– большая языковая модель (Large Language Model, LLM).

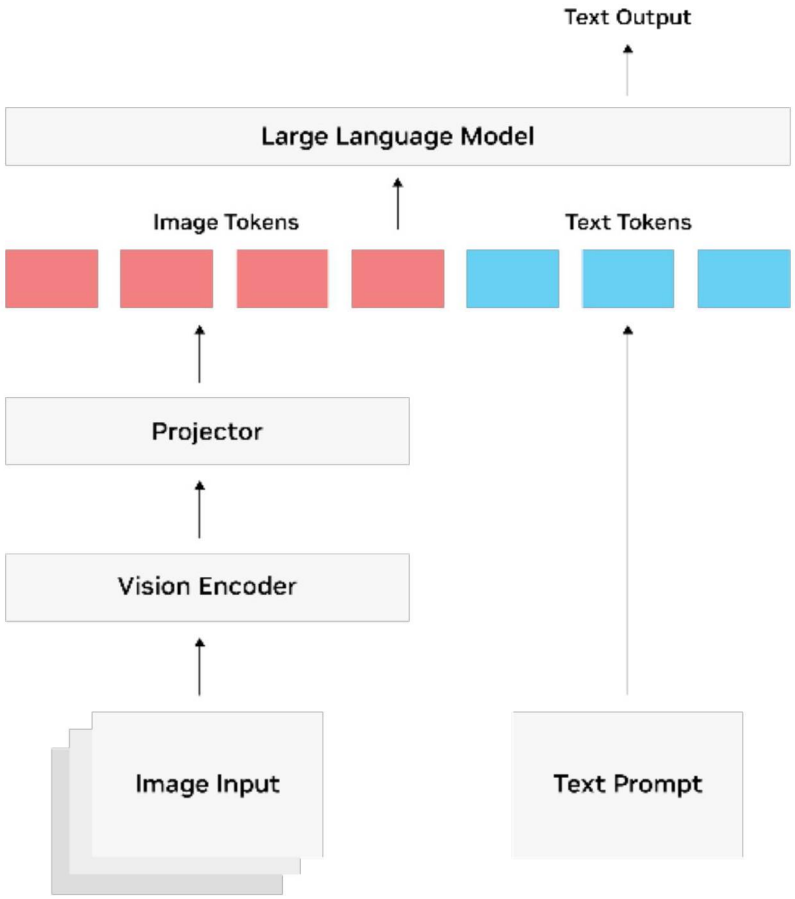

Рисунок 1. Общая архитектура VLM

LLM – это категория моделей глубокого обучения (Deep Learning, DL), обученных на огромных объемах данных, что делает их способными понимать и генерировать естественный язык и другие типы контента для выполнения широкого спектра задач [3].

Наиболее широко используемыми архитектурами LLM являются архитектуры только с энкодером (encoder-only), только с декодером (decoder-only) и архитектуры типа энкодер-декодер. Большинство из них основаны на архитектуре трансформер (Transformer) [4] (рисунок 2).

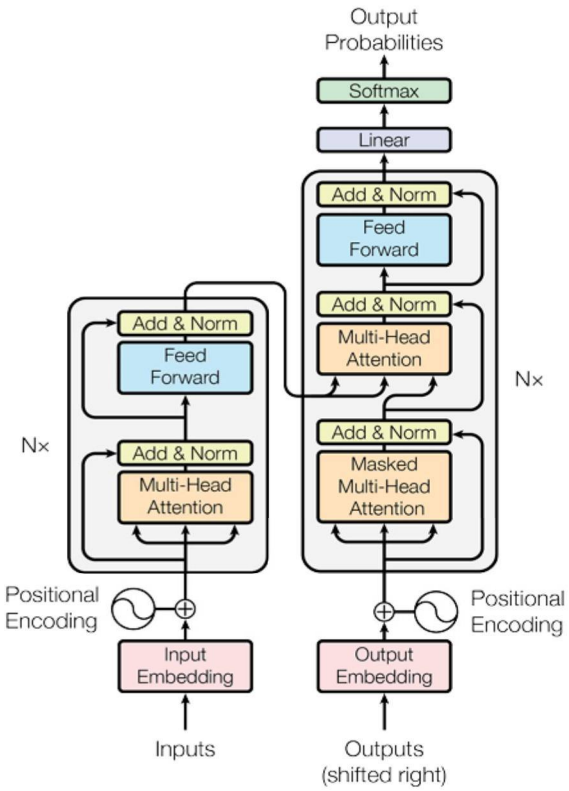

Рисунок 2. Архитектура трансформер [5]

Визуальный энкодер играет решающую роль в проецировании визуальных компонентов в признаки векторного представления (embedding), которые соответствуют векторным представлениям LLM для таких задач, как генерация текста или изображений [6]. Он обучается извлекать содержательные визуальные признаки из изображений или видеоданных, обеспечивая интеграцию с языковыми представлениями. В частности, визуальные энкодеры, используемые во многих VLM, предварительно обучаются на больших объемах мультимодальных или графических данных: эти энкодеры совместно обучаются на парах изображение-текст, что позволяет им эффективно улавливать визуальные и языковые связи. Известными примерами являются CLIP (Contrastive Language-Image Pretraining), который сопоставляет изображения и текстовые векторные представления с помощью контрастивного обучения, и BLIP (Bootstrapping Language-Image Pretraining), который использует предобучение с начальной загрузкой для сопоставления языка и изображений. Существуют энкодеры, предварительно обученные на ImageNet или аналогичных крупномасштабных наборах данных. Они обучаются на огромных объемах размеченных визуальных данных или с помощью самоконтролируемого (self-supervised) обучения, что позволяет им фиксировать визуальные признаки, характерные для конкретной предметной области. Будучи изначально одномодальными, эти энкодеры, такие как ResNet или Vision Transformers (ViTs), могут быть адаптированы для решения мультимодальных задач. Многие современные VLM обычно включают в себя предобученные визуальные энкодеры, которые не только обеспечивают надежное и осмысленное визуальное представление, но и являются высокоэффективными для трансферного (transfer) обучения. Они превосходят случайно инициализированные энкодеры, используя полученные визуальные знания из своих областей обучения.

Текстовый энкодер (text encoder) проецирует токенизированные текстовые последовательности в пространство векторного представления, аналогично тому, как визуальные энкодеры обрабатывают изображения [6]. Такие модели, как CLIP, BLIP и

ALIGN, используют как графический, так и текстовый энкодер. В этих моделях используется контрастивное обучение для сопоставления графических и текстовых векторных представлений в общем скрытом пространстве, что позволяет эффективно улавливать межмодальные связи. Однако в более новых моделях часто отсутствует специальный текстовый энкодер. Вместо этого они полагаются на LLM для понимания текста, интегрируя визуальные данные с помощью проекционных слоев или механизмов перекрестного внимания.

Текстовый декодер (text decoder) использует LLM в качестве основного генератора текста, используя визуальные энкодеры для проецирования признаков изображения [6]. В этих моделях обычно используется минимальный механизм визуальной проекции, что позволяет мощному языковому декодеру генерировать контекстуально насыщенные выходные данные. Для обучения VLM с нуля обычно требуется отдельный текстовый декодер, в то время как при использовании LLM в качестве основы часто используются исходные декодеры из LLM.

Механизмы перекрестного внимания (cross-attention mechanisms) обеспечивают визуально-текстовое взаимодействие, позволяя токенам из одной модальности (визуальной) влиять на токены из другой модальности (текстовой) [6]. Эти слои вычисляют показатели внимания в разных модальностях, но не все модели их используют.

Проектор отображает визуальные признаки, извлеченные с помощью визуального энкодера, в общее пространство векторного представления, согласованное с текстовыми векторными представлениями LLM [6]. Обычно он состоит из многослойного персептрона (Multilayer Perceptron, MLP), который преобразует трехмерные визуальные представления в компактные токены векторного представления, совместимые с текстовой модальностью. Проектор можно обучать совместно с остальной частью модели для оптимизации межмодальных задач или фиксируя определенные части модели, такие как LLM, для сохранения предобученных знаний.

Анализ сцен. Анализ сцен – фундаментальная задача компьютерного зрения, направленная на извлечение семантической информации из изображений или видео для идентификации объектов, сцен и их взаимосвязей [7]. Областями его применения выступают автономное вождение, интеллектуальное видеонаблюдение и медицинская диагностика. Последние достижения в области VLM продвинули анализ сцен на новый уровень развития.

Визуальное рассуждение можно систематически разделить на несколько основных типов, каждый из которых соответствует определенному способу вывода. В литературе широко описываются пять основных парадигм: символическое (symbolic), реляционное (relational), причинно-следственное (causal), временное (temporal) и рассуждение на основе здравого смысла или намерений (commonsense or intent-driven) [1].

Символическое рассуждение предполагает манипулирование дискретными структурированными абстракциями, полученными на основе визуального ввода. Обычно оно основано на преобразовании необработанных изображений в промежуточные символьные представления, такие как графы сцен, логические деревья или программы, с последующим выводом на основе правил. Это позволяет создавать легко интерпретируемые конвейеры рассуждений. Хотя такая явная прослеживаемость выгодна для верификации и объяснимости, этот метод требует аннотированной семантической информации, такой как маски сегментации или атрибуты объектов, которые зачастую трудно извлечь непосредственно из пикселей.

Реляционное рассуждение, напротив, делает акцент на обнаружении и моделировании парных или более сложных отношений между визуальными объектами. Эти отношения могут охватывать пространственные, функциональные или семантические измерения. Вместо использования предопределенных правил реляционные модели, часто реализуемые с помощью графовых нейронных сетей (Graph Neural Networks, GNN), трансформеров или

реляционных сетей, учатся неявно фиксировать зависимости посредством архитектурного проектирования. Такие возможности имеют решающее значение для таких задач, как создание графов сцен, понимание визуальных эталонных выражений и анализ изображений на основе аналогий.

Причинно-следственное рассуждение выходит за рамки корреляции, моделируя механизмы, посредством которых одно событие влияет на другое. В визуальных областях это часто включает выявление направленных зависимостей, а не просто совпадений. Такие методы, как структурные причинно-следственные модели (Structural Causal Models, SCM), контрфактические оценщики и do-исчисление, обеспечивают основу для получения интервенционных и гипотетических выводов на основе визуальных данных. Эти подходы особенно актуальны в рассуждениях, основанных на видео, где хронология событий кодирует множество причинно-следственных сигналов.

Временное рассуждение направлено на динамический визуальный контент, который развивается с течением времени. Такие задачи, как прогнозирование движения, сегментация действий или упорядочение событий, требуют понимания как временной эволюции отдельных объектов, так и взаимодействий между объектами. В зависимости от сложности задачи, архитектуры варьируются от рекуррентных нейронных сетей (Recurrent Neural Networks, RNN), таких как LSTM (Long Short-Term Memory) и временные сверточные сети (Temporal Convolutional Networks, TCN), до пространственно-временных трансформеров, основанных на внимании, которые фиксируют долгосрочные зависимости. Временное рассуждение особенно важно в видеонаблюдении, обеспечении качества (Quality Assurance, QA) видео и автономном восприятии. Рассуждения на основе здравого смысла и намерений связаны с неявным пониманием действий, целей и контекстуальных сигналов человека, что позволяет системам визуализации делать правдоподобные выводы даже в неоднозначных или частично наблюдаемых условиях. Этот тип рассуждений играет центральную роль в реальных приложениях, включающих вспомогательную робототехнику, прогнозирование намерений водителя и анализ социальной обстановки. Для поддержки такой формы рассуждения модели обычно включают в себя внешние базы знаний, сети с расширенной памятью или предобученные языковые модели.

Задача описания изображений имеет большое значение во многих практических приложениях, включая взаимодействие человека и компьютера и мультимодальные рекомендательные системы. Современные модели обычно обучаются на наборах данных Microsoft COCO (MS-COCO) и Flickr30k, размеченных людьми. Как у MS-COCO, так и у Flickr30k есть определенные ограничения: они в основном состоят из изображений обычных объектов и сцен, но в них отсутствуют изображения редких или сложных событий, таких как бедствия, спорт или искусство; описания и запросы, связанные с изображениями, часто являются упрощенными и повторяющимися, или неточными, что не отражает использование естественного языка и ожидания пользователя. Таким образом, существует потребность в более реалистичных и крупномасштабных наборах данных, которые могут отражать разнообразие и насыщенность визуальной и лингвистической информации в реальном мире. Стоит отметить, что сбор высококачественных размеченных наборов данных, содержащих пары изображение-текст, требует значительных временных и финансовых затрат [8].

Наборы данных с описаниями изображений обычно делятся на две основные категории: размеченные человеком, такие как MS-COCO и Flickr30k, и собранные из сети Internet, такие как CC, CC12M и SBU Captions. Эти наборы данных в совокупности формируют основу для VLP (Vision Language Pretraining) и впоследствии настраиваются для последующих задач. Как правило, наборы данных, размеченные человеком, меньше по размеру, но содержат значительно меньше шума по сравнению с наборами данных, собранными из сети Internet. Однако обе категории характеризуются относительно лаконичными описаниями.

С развитием глубокого обучения в создании описаний к изображениям произошел значительный прогресс. В последнее время особое внимание уделяется архитектурам на основе трансформеров. ExpansionNet v2 использует Swin-трансформер в качестве энкодера и декодер на основе расширения для улучшенной генерации описаний, включая механизмы динамического расширения и множественного внимания (multi-head attention). ViT-GPT2 сочетает в себе ViT-энкодер и GPT2-декодер для простого и эффективного мультимодального подхода, используя преимущества предварительно обученных унимодальных моделей. Во многих исследованиях используется стратегия предварительной подготовки больших VLM, а затем их адаптации к конкретным задачам, таким как создание описаний. BLIP-2 объединяет предобученные VLM и NLP модели, используя трансформер запросов (Q-Former) для объединения энкодера изображений и LLM. Данная модель использует двухэтапную стратегию предварительной подготовки для выравнивания представления и генеративного обучения. Современные исследования в основном направлены на дальнейшее масштабирование этих методов, основанных на предварительном обучении, и интеграцию различных визуально-языковых задач на этапе предварительного обучения. OFA – это независимая от задач модель, объединяющая представления изображений и текста с помощью трансформеров, оптимизированная с помощью перекрестной энтропии и CIDEr для описания изображений. GIT использует энкодер изображений на основе ViT и трансформер в качестве декодера текста, предварительно обученные на больших наборах данных типа изображение-текст для преобразования визуальных входных данных в текстовые описания. I-Tuning – это облегченная модель для описания изображений, которая представляет новый модуль перекрестного внимания, объединяющий фиксированный языковой декодер GPT2 с визуальным энкодером CLIP-ViT. CA-Captioner улучшает иерархию предложений, используя механизм концентрации внимания, состоящий из трех компонентов: абсолютное позиционное кодирование блока внимания (Head Absolute Positional Encoding, HAPE), фиксирующее пространственные отношения; обучаемый механизм разреживания (Learnable Sparse Mechanism, LSM), фильтрующий шум и выделяющий ключевые объекты; локальное улучшение признаков (Local Feature Enhancement, LFE), объединяющее локальные признаки. E

VCap – это дополненная поиском модель описания изображений, которая расширяет знания фиксированной LLM названиями объектов из внешней визуально-текстовой памяти. Эта модель использует легковесный модуль объединения внимания, комбинирующий извлеченные имена и визуальные признаки, обеспечивая понимание открытого мира с помощью 3,97 млн обучаемых параметров без тонкой настройки на данных, выходящих за пределы предметной области. MSRM генерирует детализированное описание сцены с точными взаимосвязями благодаря трем ключевым инновациям: семантическому анализатору, извлекающему динамические семантические сигналы; семантическому средству отображения, устанавливающему детализированные реляционные взаимодействия; адаптивному связующему декодеру, динамически объединяющему признаки различной степени детализации.

VIPCap – это визуальный запрос на основе поиска для упрощенного описания изображений, преобразующий полученный для изображения текст в семантические признаки, используя изучаемое распределение Гаусса, и сопоставляющий их с визуальными признаками для создания визуальных запросов. Вышеописанные методы создают описания с нуля, оптимизируя контролируемую функцию потерь на подобранных наборах данных [8].

Современные подходы направлены на информативное описание изображений путем объединения нескольких источников или включения дополнительной семантической информации, полученной из визуального контента. LaCLIP использует LLM для перезаписи необработанных описаний, но часто сталкивается с галлюцинациями из-за

ограниченной визуальной привязки (grounding) и низкого качества входных описаний. FuseCap улучшает качество описаний с помощью детекторов объектов. VeCLIP использует LLM для объединения необработанных и синтетических описаний, но напрямую зависит от уже существующей LLM для вывода и не имеет четких указаний по использованию знаний из необработанных описаний или синтаксических подсказок из синтетических. CapsFusion решает эту проблему, настраивая LLM с открытым исходным кодом на данных, сгенерированных ChatGPT, и включая подробные инструкции, которые помогают модели принимать более обоснованные решения во время объединения результатов. Инструмент визуальной проверки фактов направлен на уменьшение галлюцинаций в расширенных описаниях путем интеграции результатов двух мультимодальных описателей. Для проверки фактов используется обнаружение объектов, а для проверки и обобщения результатов в виде согласованного окончательного описания используется LLM. QAC улучшает детализацию описания, объединяя Questioner (запрашивает объекты, пространственные отношения и мировые знания), ChatGPT (для получения исчерпывающих ответов) и Answerer (обосновывает ответы визуальными данными), а окончательное описание синтезируется специальным Captioner [8].

Оценка VLM. По мере нарастающей интеграции языковых моделей в повседневные приложения, включая цифровых ассистентов, ботов для обслуживания клиентов, образовательные инструменты, медицинскую диагностику и даже юридический анализ, их оценка приобретает критически важное значение. Оценка охватывает важнейшие аспекты устойчивости (robustness) и достоверности (trustworthiness). Эти факторы приобретают все большее значение при всесторонней оценке производительности языковых моделей. Устойчивость изучает стабильность системы при столкновении с неожиданными входными данными [9]. В частности, устойчивость к входным данным, выходящим за пределы обучающей выборки (Out-of-Distribution, OOD) и состязательная (adversarial) устойчивость к противодействию являются популярными темами исследований в области устойчивости.

Языковые модели способны генерировать связный и, казалось бы, достоверный текст. Однако полученная информация может содержать фактические неточности или утверждения, не имеющие отношения к реальности. Это явление известно как галлюцинация (hallucination) [9].

Среди автоматических метрик оценки стоит отметить [10]:

– BLEU вычисляет профиль перекрытия n-грамм сгенерированного описания с эталонными описаниями. Данная метрика широко используется и легко вычисляется, но не учитывает семантическое сходство или беглость, а также может быть чувствительна к точному совпадению слов.

– ROUGE измеряет перекрытие n-грамм (в частности, униграмм, биграмм и самых длинных общих подпоследовательностей). Метрика подходит для оценки качества обобщения, ориентирована на полноту (Recall). Ограничения аналогичны BLEU и в основном связаны с точным совпадением слов.

– METEOR вычисляется на основе среднего гармонического значения точности (precision) и полноты униграммы, где полнота имеет большее значение. Лучшая корреляция с человеческими суждениями, чем у BLEU, включающая сопоставление корней слов и синонимии. Тем не менее, может пропускать семантические связи более высокого уровня.

– CIDEr вычисляет взвешенное по TF-IDF (Term Frequency-Inverse Document Frequency) сходство n-грамм с несколькими эталонными описаниями. Это метрика, специально разработанная для задачи описания изображений, которая придает больший вес более редким n-граммам. Иногда она может отдавать предпочтение чрезмерно избыточным описаниям.

– SPICE оценивает описания на основе семантического предлагаемого контента, включая объекты, атрибуты и отношения. Метрика улавливает семантическое значение за пределами перекрытия n-грамм и хорошо коррелирует с человеческими суждениями.

Однако ее вычисление может вызывать сложность, и она не всегда эффективно обрабатывает синтаксическое разнообразие.

– CLAIR использует zero-shot LLM для оценки описаний. У данной метрики лучшее соответствие человеческим суждениям по сравнению с текущими показателями, и она обеспечивает интерпретируемость при наличии помех. У нее высокая вычислительная стоимость, а воспроизводимость может зависеть от стабильности базовой LLM.

– FLEUR – метрика, использующая большую мультимодальную модель для сопоставления описания с изображением и не нуждающаяся в эталонных описаниях. Она объяснима и демонстрирует лучшие показатели при оценке без использования эталона. Однако производительность зависит от качества и возможностей базовой большой мультимодальной модели.

В ответ на растущую сложность моделей был разработан ряд эталонных тестов для стандартизации оценки языковых моделей при решении широкого круга задач:

– MS COCO – наиболее широко используемый набор данных, который содержит большую коллекцию изображений с подробными аннотациями, такими как сегментация объектов и несколько созданных человеком описаний к изображению [10]. Наличие нескольких описаний для одного изображения позволяет оценить разнообразие генерируемых описаний.

– Flickr8K – набор данных, предоставляющий пять детальных описаний для каждого изображения [10].

– Flickr30K – расширенная версия Flickr8K [10].

– Conceptual Captions – большой набор данных, созданный путем получения изображений и описаний программным путем с веб-сайтов, которые предоставляют большое количество зашумленных, но разнообразных пар изображение-текст [10]. Одному изображению соответствует одно описание.

– IU X-Ray – набор данных, который часто используется для описаний медицинских изображений [10]. Он содержит рентгенологические снимки и диагностические отчеты. Для подобных наборов необходимо, чтобы модели изучали специализированные словари и понимали визуальные особенности конкретной предметной области. Для каждого изображения используется только одно описание, что ограничивает лингвистическую избыточность, но обеспечивает более целенаправленное текстовое описание.

– UnIVAL – унифицированный эталонный тест, охватывающий несколько модальностей (изображение, видео, аудио, текст), позволяющий проводить оценку в классических визуально-языковых задачах, а также обобщать результаты на видео и аудио-текст [11]. Данный тест охватывает аудио-текстовые задачи, VQA, создание описаний к изображениям, ответы на вопросы в видео в рамках единого фреймворка и выпущен в качестве ресурса для оценки унифицированных мультимодальных LLM (MLLM).

Некоторые подходы для оценки выполнения задач обнаружения, сегментации, классификации изображений и обработки видео с помощью VLM используют такие классические наборы данных, как ImageNet-1K, CIFAR-10/100, ADE20K, Cityscapes, Pascal VOC, YouTube-VOS [11].

Несмотря на достижения в области автоматической оценки, оценка человеком остается решающей для всесторонней оценки качества описания изображений, поскольку люди, выполняющие оценку, могут дать тонкие суждения о связности, соответствии и общей естественности сгенерированных описаний [10]. В качестве компромисса некоторые системы используют гибридные методы оценки – автоматизированная оценка для задач большого объема в сочетании с периодическими проверками человеком областей с высоким уровнем риска или значительным влиянием [12].

Стремительное развитие LLM значительно расширило их возможности в восприятии и рассуждении в долгосрочном контексте, что все больше популяризирует их использование в качестве оценщиков в различных задачах NLP. На основе концепции LLM-

as-a-Judge возникла идея MLLM-as-a-Judge [13], означающая возможность использования MLLM для оценки выполнения задач.

Заключение. В последнее время LLM достигли значительного прогресса, сравнявшись или даже превзойдя возможности человека в решении целого ряда сложных задач. Несмотря на значительные достижения, VLM по-прежнему отстают от LLM в основных возможностях рассуждения. В то время как LLM демонстрируют сильные навыки символического и абстрактного рассуждения, VLM испытывают трудности с существенными аспектами визуального понимания. Они часто неправильно подсчитывают объекты в загроможденных или перекрывающихся сценах, плохо справляются с пространственным рассуждением и не могут точно определить местоположение или распознать важные объекты во время визуального поиска. Более того, композиционное и реляционное рассуждение, где понимание критически зависит от структурированных отношений между объектами, остается серьезной проблемой. Эти ограничения указывают на то, что VLM, несмотря на их мощь, не имеют ключевых механизмов для структурированного визуального рассуждения. Следовательно, их эффективность в этих областях еще не соответствует эффективности LLM в сопоставимых задачах рассуждения на основе текста. Проведенное в работе исследование выполнено в рамках совместного белорусско-турецкого проекта № Ф25ТУРГ-001.

Список литературы

[1] Sarkar A., Idris M. Y. I., Yu Z. Reasoning in Computer Vision: Taxonomy, Models, Tasks, and Methodologies. arXiv preprint arXiv:2508.10523v1 [cs.CV]. 2025. DOI: 10.48550/arXiv.2508.10523.

[2] What are Vision-Language Models? | NVIDIA Glossary [Electronic resource]. Mode of access: https://www.nvidia.com/en-us/glossary/vision-language-models/. Date of access: 16.03.2026.

[3] What Are Large Language Models (LLMs)? | IBM [Electronic resource]. Mode of access: https://www.ibm.com/think/topics/large-language-models. Date of access: 17.03.2026.

[4] Minaee S., Mikolov T., Nikzad N., Chenaghlu M., Socher R., Amatriain X., Gao J. Large Language Models: A Survey. arXiv preprint arXiv:2402.06196v3 [cs.CL]. 2025. DOI: 10.48550/arXiv.2402.06196.

[5] Vaswani A., Jones L., Shazeer N., Parmar N., Gomez A. N., Uszkoreit J., Kaiser Ł., Polosukhin I. Attention Is All You Need. Proc. of the 31st International Conference on Neural Information Processing Systems (NIPS'17). 2017;6000-6010.

[6] Li Z., Wu X., Du H., Liu F., Nghiem H., Shi G. A Survey of State of the Art Large Vision Language Models: Alignment, Benchmark, Evaluations and Challenges. arXiv preprint arXiv:2501.02189v6 [cs.CV]. 2025. DOI: 10.48550/arXiv.2501.02189.

[7] Wang J., Chen Y., Si L., Zheng C. Advancing Complex Wide-Area Scene Understanding with Hierarchical Coresets Selection. Proc. of the 33rd ACM International Conference on Multimedia (MM '25). 2025;2663-2672. DOI: 10.1145/3746027.3754707.

[8] Celona L., Bianco S., Donzella M., Napoletano P. Improving image captioning descriptiveness by ranking and LLM-based fusion. Neural Computing and Applications. 2025;37:27279-27299. DOI: 10.1007/s00521-025-11672-x.

[9] Chang Y., Wang X., Wang J., Wu Y., Yang L., Zhu K., Chen H., Yi X., Wang C., Wang Y., Ye W., Zhang Y., Chang Y., Yu P. S., Yang Q., Xie X. A Survey on Evaluation of Large Language Models. ACM Transactions on Intelligent Systems and Technology. 2024;15(3):1-45. DOI: 10.1145/3641289.

[10] Panchal P., Polara V., U S., Baz A., Patel S. K. Deep learning–driven image captioning: Progress through transformers and large language models. PLoS One. 2026;21(3):e0345012. DOI: 10.1371/journal.pone.0345012.

[11] Deng Z., Wang Y., Liang Y., Du J., Yang Y., Fang L., He L., Han Y., Zhu Y., Miao C., Zhang W., Chen J., Li Y., Zhao W., Yu P. S. A Survey of Multimodal Models on Language and Vision: A Unified Modeling Perspective. Data Mining and Machine Learning. 2025;1(1).

[12] Pandhare H. V. Evaluating Large Language Models: Frameworks and Methodologies for AI/ML System Testing. International Journal of Scientific Research and Management (IJSRM). 2024;12:1467-1486. DOI: 10.18535/ijsrm/v12i09.ec08.

[13] Chen D., Chen R., Zhang S., Wang Y., Liu Y., Zhou H., Zhang Q., Wan Y., Zhou P., Sun L. MLLM-as-a-Judge: assessing multimodal LLM-as-a-Judge with vision-language benchmark. Proc. of the 41st International Conference on Machine Learning (ICML'24). 2024;6562-6595.

APPLICATION OF VLM FOR VISUAL SCENE ANALYSIS

V.V. Vengerenko
Junior Researcher at The State Scientific Institution
«The United Institute of Informatics Problems of the
National Academy of Sciences of Belarus»

Abstract. Vision-Language Models (VLMs) represent a sophisticated advancement in artificial intelligence, integrating the capabilities of both computer vision and natural language processing (NLP) to provide a more holistic understanding of data. Unlike traditional models that focus on a single type of input – either visual or textual – VLMs are designed to process and understand multimodal data, combining visual and textual information to generate richer insights.

The purpose of the research is to determine the specifics of using VLMs for visual scene analysis.

The key components of VLMs are examined. The existing approaches to the visual scene analysis based on VLMs are investigated. A review of methods for evaluating the quality of such models is performed.

Keywords: artificial intelligence, deep learning, language model, modality, token.