Можем ли мы убедить модели видеть мир по-другому?

Дата: 2025-03-05 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2403.09193

Рейтинг: 65
Адаптивность: 70

Ключевые выводы:

Исследование направлено на изучение того, как языковые подсказки (промпты) могут влиять на визуальное восприятие мультимодальных моделей (VLM). Основной вывод: VLM наследуют определенные визуальные предпочтения (bias) от своих энкодеров зрения, но эти предпочтения можно частично изменять с помощью языковых промптов.

Объяснение метода:

Исследование предлагает практические методы управления визуальным восприятием VLM через промпты, что полезно для целенаправленного взаимодействия с моделями. Пользователи могут направлять внимание модели на форму или текстуру объектов без технических знаний. Однако степень влияния ограничена (~20-25%), а полное понимание требует технической подготовки.

Ключевые аспекты исследования: 1. **Изучение текстурно-формовой предвзятости в VLM**: Исследование анализирует, как мультимодальные модели зрения-языка (VLM) воспринимают визуальные признаки, особенно выбор между текстурой и формой при распознавании объектов.

Влияние языка на визуальное восприятие: Авторы обнаружили, что VLM по умолчанию больше ориентированы на форму объектов (около 60-70%), чем чисто зрительные модели (22%), хотя и не достигают человеческого уровня (96%).

Управление визуальными предвзятостями через промпты: Исследование демонстрирует, что через языковые промпты можно влиять на то, какие визуальные признаки (форма или текстура) модель будет использовать для принятия решений.

Мультимодальное слияние и обработка информации: Языковая часть VLM активно влияет на визуальное восприятие, а не просто наследует предвзятости от зрительного энкодера.

Расширение на другие визуальные предвзятости: Авторы показали, что подобное управление возможно не только для текстуры/формы, но и для

высоко/низкочастотных визуальных признаков.

Дополнение:

Исследование действительно не требует дообучения или специального API для применения основных методов и концепций. Ключевая ценность работы в том, что она демонстрирует возможность влиять на визуальное восприятие моделей через обычные текстовые промпты в стандартном чате с VLM.

Концепции и подходы, которые можно применить в стандартном чате:

Направленные промпты для управления вниманием - пользователи могут включать в запросы фразы типа "Определи объект по его форме" или "Опиши текстуру/поверхность объекта", чтобы направить внимание модели на конкретные визуальные аспекты.

Использование синонимов для усиления эффекта - исследование показало, что синонимы слов "форма" и "текстура" также эффективны. Можно использовать термины как "контур", "силуэт", "очертание" для формы или "поверхность", "узор", "материал" для текстуры.

Постепенное уточнение запросов - если модель фокусируется не на тех аспектах изображения, пользователь может уточнять запрос, указывая на конкретные визуальные характеристики.

Понимание базовых предпочтений VLM - зная, что VLM по умолчанию больше ориентированы на форму (~60-70%), чем чисто зрительные модели, пользователи могут соответствующим образом формулировать запросы.

Ожидаемые результаты: - Более целенаправленные описания изображений, фокусирующиеся на нужных пользователю аспектах - Возможность получить альтернативные интерпретации одного и того же изображения - Более контролируемые ответы модели при работе со сложными или неоднозначными изображениями - Улучшенное взаимодействие в сценариях, где важно различать форму и текстуру (например, при анализе произведений искусства, дизайне, медицинской визуализации)

Важно отметить, что степень влияния ограничена (~20-25% смещения в сторону формы или текстуры), но даже такое частичное управление может быть полезным во многих практических сценариях.

Анализ практической применимости: **Изучение текстурно-формовой предвзятости в VLM** - Прямая применимость: Средняя. Пользователи могут лучше понимать, как VLM "видят" объекты, но непосредственного применения для обычных задач мало. - Концептуальная ценность: Высокая. Понимание того, что VLM больше ориентируются на форму, чем чистые зрительные модели, помогает осознать их преимущества и особенности. - Потенциал для адаптации: Средний. Знание о предвзятостях может помочь пользователям формулировать более эффективные

запросы.

Влияние языка на визуальное восприятие - Прямая применимость: Низкая. Скорее академический результат о внутренней работе моделей. - Концептуальная ценность: Высокая. Понимание, что языковые запросы влияют на то, как модель "видит" изображение, важно для эффективного взаимодействия. - Потенциал для адаптации: Средний. Можно использовать как концептуальную основу для работы с VLM.

Управление визуальными предвзятостями через промпты - Прямая применимость: Высокая. Пользователи могут напрямую использовать специальные промпты для управления тем, как модель будет интерпретировать изображения. - Концептуальная ценность: Высокая. Демонстрирует гибкость VLM и возможность настройки их поведения без переобучения. - Потенциал для адаптации: Высокий. Принципы управления предвзятостями можно применить к разным задачам и контекстам.

Мультимодальное слияние и обработка информации - Прямая применимость: Низкая. Технические детали работы моделей мало применимы для обычного пользователя. - Концептуальная ценность: Средняя. Помогает понять, почему VLM работают иначе, чем отдельные зрительные или языковые модели. - Потенциал для адаптации: Низкий. Знание внутренних механизмов сложно адаптировать для практического применения.

Расширение на другие визуальные предвзятости - Прямая применимость: Средняя. Показывает, что принципы управления применимы к разным типам визуального восприятия. - Концептуальная ценность: Высокая. Демонстрирует универсальность подхода управления через промпты. - Потенциал для адаптации: Высокий. Методы можно адаптировать для управления различными аспектами визуального восприятия.

Prompt:

Использование знаний о визуальных предпочтениях VLM в промптах ## Ключевое понимание исследования

Исследование показывает, что мультимодальные модели (VLM) имеют определенные визуальные предпочтения, но эти предпочтения можно корректировать с помощью языковых промптов. Важно, что VLM обычно больше ориентируются на форму объектов (shape bias ~60-70%), чем на текстуру.

Пример промпта для усиления фокуса на текстуре

[======] Проанализируй это изображение, обращая особое внимание на ТЕКСТУРУ и поверхностные характеристики объектов. Сначала опиши детально текстурные элементы (материал, узор, поверхностные качества), а затем уже форму и другие характеристики. Какие материалы и текстуры ты видишь на изображении в первую

очередь? [======]

Объяснение работы промпта

Этот промпт использует знания из исследования следующим образом:

Преодоление естественного предпочтения формы: Поскольку VLM имеют встроенное предпочтение формы (~60-70%), промпт явно направляет внимание модели на текстурные характеристики, которым она уделяет меньше внимания по умолчанию.

Использование прямых инструкций: Исследование показало, что языковые инструкции могут значительно изменить визуальные предпочтения (с 49% до 72%), поэтому промпт напрямую указывает модели, на что обращать внимание.

Структурирование ответа: Запрашивая сначала описание текстуры, а затем формы, промпт использует знание о том, что VLM могут принимать высоко уверенные решения, игнорируя один из визуальных сигналов, поэтому мы явно просим учесть оба.

Приоритизация: Финальный вопрос закрепляет приоритет текстурной информации, противодействуя естественной склонности модели к форме.

Аналогичным образом можно создавать промпты для усиления фокуса на форме или для поиска баланса между различными визуальными характеристиками изображения.