ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»

Кафедра моделирования и системного анализа

**Дисциплина: «Технологии математического моделирования и анализа данных»**

*Направление подготовки: «Прикладная математика и информатика»*

*Профиль: «Прикладное машинное обучение»*

*Факультета информационных технологий и анализа больших данных*

*Форма обучения очная*

*Учебный 2024/2025 год, 7 семестр*

**Тема**:

«Визуальные трансформеры»

Выполнил:

студент группы ПМ21-5 Факультета информационных технологий и анализа больших данных

\_\_Клевицкий Д.Б.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Оглавление

[Введение 3](#_Toc186549597)

[Теоретический обзор 3](#_Toc186549598)

[Понятие трансформеров в машинном обучении 3](#_Toc186549599)

[Развитие технологии трансформеров 4](#_Toc186549600)

[Принципы работы визуальных трансформеров 5](#_Toc186549601)

[Архитектура визуальных трансформеров 5](#_Toc186549602)

[Механизм внимания и его роль в обработке изображений 7](#_Toc186549603)

[Сравнение с традиционными методами 10](#_Toc186549604)

[Преимущества и недостатки по сравнению с конволюционными нейронными сетями (CNN) 10](#_Toc186549605)

[Примеры улучшений и ограничений 11](#_Toc186549606)

[Применение визуальных трансформеров 13](#_Toc186549607)

[Обнаружение и распознавание объектов. Улучшение точности и скорости идентификации объектов на изображениях 13](#_Toc186549608)

[Сегментация изображений. Разделение изображения на локализованные участки и классификация этих участков 14](#_Toc186549609)

[Анализ медицинских изображений. Диагностика и автоматический анализ рентгеновских снимков, МРТ и других медицинских изображений 15](#_Toc186549610)

[Обработка видео. Извлечение информации и отслеживание объектов в видеопотоке 16](#_Toc186549611)

[Стабилизация и улучшение изображений. Применение для повышения качества изображений и устранения шумов 17](#_Toc186549612)

[Заключение 18](#_Toc186549613)

Введение

В последние годы область компьютерного зрения переживает революцию благодаря развитию глубоких нейронных сетей и увеличению вычислительных возможностей. Трансформеры, изначально разработанные для задач обработки естественного языка, показали выдающиеся результаты в моделировании длинных зависимостей в последовательностях. Их успешная адаптация в области компьютерного зрения привела к появлению визуальных трансформеров, которые начинают конкурировать с традиционными свёрточными нейронными сетями (CNN).

Актуальность исследований в области визуальных трансформеров обусловлена несколькими факторами. Во-первых, они обладают способностью эффективно захватывать глобальные взаимосвязи в изображениях благодаря механизмам самовнимания. Во-вторых, визуальные трансформеры предоставляют более унифицированный подход к моделированию данных различных модальностей, что открывает возможности для мультимодальных приложений. Наконец, учитывая быстрый рост данных и необходимость в более гибких моделях, исследования в этой области становятся всё более значимыми для развития технологий искусственного интеллекта.

Теоретический обзор

Понятие трансформеров в машинном обучении

Трансформеры — это класс моделей глубокого обучения, которые стали основой для современных достижений в области обработки последовательных данных. Изначально представленные в 2017 году в работе "Attention is All You Need" исследователями из Google Brain, трансформеры были разработаны для решения задач обработки естественного языка (NLP), таких как машинный перевод, отвечая на необходимость эффективного моделирования длинных зависимостей в текстовых последовательностях.

Ключевым компонентом трансформеров является механизм самовнимания (self-attention), который позволяет модели обращать внимание на различные части входной последовательности при генерации каждого выходного элемента. В отличие от рекуррентных нейронных сетей (RNN) и длинной краткосрочной памяти (LSTM), трансформеры не зависят от последовательной обработки данных, что обеспечивает лучшую параллелизацию и скорость обучения.

Архитектура трансформера состоит из энкодера и декодера, каждый из которых включает многоголовый механизм внимания и позиционные кодировки для сохранения информации о порядке элементов в последовательности. Эта архитектура позволила достичь значительных улучшений в качестве моделей и установила новые стандарты в различных задачах NLP.

Развитие технологии трансформеров

После успешного применения в обработке естественного языка идея трансформеров начала распространяться на другие области машинного обучения. Исследователи обратили внимание на потенциал механизмов внимания в обработке данных различных модальностей, включая изображения, аудио и видео.

В компьютерном зрении традиционно доминировали свёрточные нейронные сети (CNN), которые эффективно обрабатывали пространственные зависимости в изображениях. Однако CNN имеют ограничения в моделировании глобальных взаимосвязей из-за локальности свёрток. В попытке объединить преимущества трансформеров и CNN, были предложены гибридные модели, но полностью трансформерные подходы в компьютерном зрении оставались редкостью из-за вычислительной сложности.

Прорыв произошёл с появлением Vision Transformer (ViT) в 2020 году. ViT адаптировал архитектуру трансформера для обработки изображений, разбивая их на последовательности патчей (небольших участков изображения) и применяя к ним механизм самовнимания. Эта модель показала, что при наличии большого объёма данных для обучения трансформеры могут превосходить CNN в задачах классификации изображений.

Дальнейшее развитие технологий привело к созданию более эффективных и адаптированных моделей:

* **DeiT (Data-efficient Image Transformers)**: предложил методы обучения ViT на меньших датасетах без потери точности, используя стратегии обучения, такие как твердые учителя (strong teachers) и дистилляция знаний.
* **Swin Transformer**: ввёл иерархическую архитектуру и концепцию скользящих окон (sliding windows) для внимания, что улучшило эффективность обработки изображений разного разрешения и масштабов.
* **Perceiver и Perceiver IO**: расширили идею трансформеров, предложив модели, способные эффективно обрабатывать мультимодальные данные, объединяя информацию из различных источников.

Кроме того, трансформеры нашли применение в задачах генерации изображений, сегментации и детекции объектов, что свидетельствует о их универсальности и мощи. Развитие технологий трансформеров продолжается, с акцентом на улучшение эффективности, снижение вычислительных затрат и расширение областей применения в различных сферах искусственного интеллекта.

Принципы работы визуальных трансформеров

Архитектура визуальных трансформеров

Визуальные трансформеры представляют собой адаптацию архитектуры трансформеров, изначально разработанных для задач обработки естественного языка, к области компьютерного зрения. Основная идея заключается в том, чтобы представить изображение как последовательность элементов и применить к ним механизмы внимания для моделирования зависимостей между различными частями изображения.

**Разбиение изображения на патчи**

Визуальные трансформеры, такие как Vision Transformer (ViT), начинают обработку с разбиения изображения на небольшие фрагменты — патчи фиксированного размера, например, (16x16) пикселей. Каждый патч разворачивается в одномерный вектор путем линейного преобразования, создавая последовательность эмбеддингов, аналогичных токенам в задачах NLP.

**Позиционные кодировки**

Поскольку трансформеры не учитывают порядок входных данных по умолчанию, к эмбеддингам патчей добавляются позиционные кодировки. Это позволяет модели сохранять информацию о пространственном расположении каждого патча в исходном изображении, что критически важно для задач компьютерного зрения.

**Энкодер на основе трансформеров**

Полученные эмбеддинги поступают в энкодер, состоящий из нескольких слоев трансформера. Каждый слой включает в себя:

* **Многоголовый механизм самовнимания**: позволяет модели обращать внимание на различные части изображения, выявляя взаимосвязи между патчами.
* **Нормализацию**: стабилизирует процесс обучения и улучшает сходимость модели.
* **Полносвязные слои**: применяются после механизма внимания для обработки информации и создания новых представлений данных.

**Классификационный токен**

Для задач классификации вводится специальный токен ([CLS]), который инициализируется определенным образом и проходит через все слои энкодера вместе с эмбеддингами патчей. После энкодера этот токен содержит агрегированную информацию обо всем изображении и используется для предсказания класса.

**Выходной слой**

Наконец, эмбеддинг классификационного токена передается в полносвязный слой с функцией активации softmax (или другой подходящей для задачи), который выдает вероятности принадлежности изображения к различным классам.

**Преимущества архитектуры визуальных трансформеров**

* **Глобальный контекст**: механизм внимания позволяет учитывать взаимосвязи между отдаленными частями изображения.
* **Унифицированная обработка**: отсутствует необходимость в специальных операциях, таких как свертки или пулинг.
* **Гибкость**: архитектура легко адаптируется к различным размерам изображений и задачам.

Механизм внимания и его роль в обработке изображений

Механизм внимания является ключевым компонентом трансформеров и обеспечивает их преимущество в обработке последовательностей и изображений.

**Самовнимание (Self-Attention)**

Самовнимание позволяет модели оценивать, какие части входных данных наиболее важны для каждого элемента последовательности. В контексте изображения это означает, что каждый патч может обращать внимание на другие патчи и взвешивать их вклад при обновлении своего представления.

**Вычисление внимания**

Процесс самовнимания включает в себя следующие шаги:

1. **Генерация запросов, ключей и значений**: для каждого эмбеддинга патча с помощью линейных преобразований вычисляются три вектора — запрос (Q), ключ (K) и значение (V).
2. **Вычисление скоринговой функции**: для каждого патча вычисляются скалярные произведения его запроса (Q\_i) с ключами всех других патчей (K\_j), что отражает степень их взаимодействия.
3. **Нормализация**: полученные значения проходят через функцию softmax, превращаясь в весовые коэффициенты, суммирующиеся до 1.
4. **Агрегация значений**: взвешенные значения (V\_j) суммируются, формируя обновленное представление патча (i).

**Многоголовое внимание**

Механизм многоголового внимания позволяет модели учитывать различные аспекты информации путем параллельного вычисления нескольких самовниманий с разными параметрами. Это расширяет способность модели захватывать сложные зависимости и особенности изображения.

**Роль внимания в обработке изображений**

* **Учет контекста**: внимание способствует интеграции информации из разных частей изображения, что особенно важно для понимания глобальной структуры и объектов в контексте сцены.
* **Изменение фокуса**: модель может адаптивно выделять более важные патчи, игнорируя менее существенные детали, что улучшает эффективность обработки.
* **Снижение зависимости от локальных признаков**: в отличие от свёрточных сетей, которые изначально сфокусированы на локальных паттернах, внимание позволяет моделировать долгосрочные зависимости без увеличения сложности.

**Применение в задачах компьютерного зрения**

* **Классификация**: улучшение качества предсказаний за счет интеграции глобальной информации.
* **Детекция объектов**: более точное определение границ и местоположения объектов посредством учета контекста.
* **Сегментация**: повышение точности сегментации объектов путем объединения информации из различных областей изображения.

**Преимущества механизма внимания**

* **Гибкость**: легко масштабируется и адаптируется к различным размерам и типам данных.
* **Параллельность**: позволяет эффективно использовать современные вычислительные ресурсы благодаря возможности параллельного вычисления.

Сравнение с традиционными методами

Преимущества и недостатки по сравнению с конволюционными нейронными сетями (CNN)

**Преимущества визуальных трансформеров по сравнению с CNN:**

* **Моделирование глобальных взаимосвязей**: Визуальные трансформеры используют механизм самовнимания, который позволяет им учитывать взаимосвязи между отдалёнными частями изображения. Это особенно полезно для задач, где контекст играет важную роль, позволяя модели захватывать долгосрочные зависимости.
* **Гибкость архитектуры**: В отличие от CNN, которые построены на основе свёрток с фиксированным размером ядра, трансформеры могут обрабатывать входные данные различного размера и формы без необходимости изменения архитектуры.
* **Параллельная обработка**: Трансформеры не зависят от последовательных операций и позволяют эффективно использовать параллелизм на уровне обработки патчей, что может ускорить обучение на современных вычислительных платформах.
* **Унифицированный подход ко всем типам данных**: Трансформеры успешно применяются в различных областях, включая NLP и компьютерное зрение, что упрощает интеграцию мультимодальных данных и разработку комплексных моделей.

**Недостатки визуальных трансформеров по сравнению с CNN:**

* **Требование больших объёмов данных**: Визуальные трансформеры, как правило, требуют значительно больше данных для эффективного обучения по сравнению с CNN. Отсутствие встроенных индуктивных предубеждений, таких как локальность и трансляционная инвариантность, означает, что модели должны самостоятельно изучать эти свойства из данных.
* **Высокая вычислительная сложность**: Механизм самовнимания имеет квадратичную сложность по отношению к размеру входа, что приводит к высоким требованиям к вычислительным ресурсам и памяти, особенно при обработке изображений высокого разрешения.
* **Отсутствие локальных индуктивных предубеждений**: CNN эффективно распознают локальные паттерны благодаря свёрткам и пулингу. Визуальные трансформеры могут не справляться с этим так же хорошо, если не использовать дополнительные архитектурные улучшения.
* **Сложность оптимизации**: Большое количество параметров и отсутствие структурных ограничений могут затруднять процесс обучения, делая модели более чувствительными к гиперпараметрам и требующими тщательной настройки.

Примеры улучшений и ограничений

**Примеры улучшений:**

* **Классификация изображений**: Модель Vision Transformer (ViT) показала, что трансформеры способны достигать и превосходить уровень точности современных CNN в задачах классификации, особенно при наличии больших объемов данных для обучения. ViT обучалась на масштабных датасетах и продемонстрировала, что механизм внимания эффективно захватывает как локальные, так и глобальные особенности изображений.
* **Улучшенная обобщающая способность**: Визуальные трансформеры могут лучше обобщать на новых данных благодаря способности моделировать сложные взаимосвязи. Это особенно заметно в задачах, где контекст и взаимодействие между объектами играют ключевую роль.
* **Мультимодальные приложения**: Благодаря унифицированной архитектуре, трансформеры легко интегрируются в модели, которые обрабатывают разные типы данных, например, сочетание текста и изображений. Это открывает новые возможности для приложений, таких как визуальный вопрос-ответ или генерация описаний к изображениям.

**Примеры ограничений:**

* **Эффективность на ограниченных данных**: На небольших или специализированных датасетах CNN часто превосходят визуальные трансформеры, так как свёрточные сети лучше используют встроенные индуктивные предубеждения, позволяющие им эффективно обучаться на ограниченном количестве данных.
* **Вычислительные ресурсы**: Высокие требования к аппаратным ресурсам затрудняют использование визуальных трансформеров в условиях ограниченных вычислительных возможностей, например, на мобильных устройствах или в реального времени.
* **Сложность интерпретации**: Несмотря на то, что механизм внимания даёт некоторую интерпретируемость, общее число параметров и сложность модели могут затруднять понимание и объяснение того, как именно принимаются решения, что важно в критически важных приложениях.
* **Не всегда превосходят CNN**: В задачах, где локальные особенности являются ключевыми, и глобальный контекст менее важен, CNN могут показывать более высокую эффективность и точность по сравнению с визуальными трансформерами.

Сравнение визуальных трансформеров с традиционными сверточными нейронными сетями показывает, что, хотя трансформеры обладают значительным потенциалом и преимуществами в определённых областях, они не заменяют полностью CNN. Выбор между этими архитектурами зависит от конкретной задачи, объёма доступных данных и вычислительных ресурсов. Визуальные трансформеры представляют собой мощный инструмент для задач, требующих учёта глобального контекста и сложных взаимодействий, в то время как CNN остаются эффективными и надёжными для широкого спектра приложений компьютерного зрения.

Применение визуальных трансформеров

Обнаружение и распознавание объектов. Улучшение точности и скорости идентификации объектов на изображениях

Визуальные трансформеры значительно продвинули вперед область обнаружения и распознавания объектов на изображениях. Традиционные модели на основе свёрточных нейронных сетей (CNN) достигли впечатляющих результатов в этом направлении, однако внедрение трансформеров привнесло новые возможности.

**Обнаружение объектов с помощью трансформеров**

Модель DETR (DEtection TRansformer) стала одним из первых примеров успешного применения трансформеров в задаче обнаружения объектов. DETR объединяет CNN для извлечения начальных признаков и трансформер для моделирования глобальных зависимостей между объектами и фоном изображения. Это позволяет модели напрямую предсказывать набор объектов с их классами и координатами, обходясь без сложных процедур якорей и не максимального подавления (NMS), используемых в традиционных методах.

**Преимущества в точности и скорости**

* **Точность**: Механизм внимания в трансформерах помогает лучше различать объекты, особенно в сложных сценах с множеством перекрывающихся или мелких деталей. Это повышает качество обнаружения и снижает число пропущенных объектов или ложных срабатываний.
* **Скорость**: несмотря на вычислительную сложность трансформеров, оптимизация архитектуры и использование мощных графических процессоров позволяют достичь конкурентоспособной скорости обработки. Кроме того, упрощение этапа постобработки, как в DETR, сокращает общее время предсказания.

**Применение в реальных задачах**

* **Видеонаблюдение**: Обнаружение и трекинг объектов в режиме реального времени для обеспечения безопасности и мониторинга.
* **Автономные системы**: Распознавание дорожных знаков, пешеходов и других транспортных средств для навигации и предотвращения столкновений.
* **Робототехника**: Идентификация объектов и препятствий для взаимодействия с окружающей средой.

Сегментация изображений. Разделение изображения на локализованные участки и классификация этих участков

Сегментация изображений является ключевой задачей в компьютерном зрении, где требуется определить точные границы и классы объектов на пиксельном уровне. Визуальные трансформеры способствуют улучшению результатов сегментации благодаря глобальному вниманию к контексту сцены.

**Трансформеры в задачах сегментации**

Модель SETR (SEgmentation TRansformer) заменила классические свёрточные блоки на трансформерный энкодер, позволяя обрабатывать глобальные связи между пикселями. Это особенно полезно для объектов со сложной структурой или в случаях, когда объекты состоят из неоднородных частей.

**Преимущества визуальных трансформеров**

* **Глобальное восприятие**: Способность учитывать информацию из отдаленных областей изображения помогает точно определять границы и классифицировать сложные объекты.
* **Устойчивость к масштабам**: Трансформеры эффективно обрабатывают объекты различных размеров без необходимости специальных архитектурных решений.

**Практические применения**

* **Медицинская сегментация**: Выделение опухолей, органов и других анатомических структур на медицинских изображениях для диагностики и планирования лечения.
* **Автономное управление**: Сегментация дорожных покрытий, препятствий и окружающей среды для безопасного передвижения транспортных средств.
* **Распознавание сцен**: Классификация различных элементов в сцене, например, зданий, зелени, дорог на аэрофотоснимках.

Анализ медицинских изображений. Диагностика и автоматический анализ рентгеновских снимков, МРТ и других медицинских изображений

Медицинская визуализация является одной из наиболее критически важных областей применения компьютерного зрения, где точность и надежность моделей являются жизненно важными. Визуальные трансформеры предлагают новые возможности для улучшения качества диагностики и автоматизации анализа изображений.

**Диагностика заболеваний**

* **Обнаружение аномалий**: Трансформеры используются для выявления патологий на рентгеновских снимках, КТ и МРТ, включая опухоли, поражения тканей и другие отклонения.
* **Классификация заболеваний**: Модели могут различать типы заболеваний и стадии их развития, что помогает врачу в постановке диагноза.

**Автоматическая сегментация**

* **Выделение органов и структур**: Точное определение границ органов, сосудов и других анатомических структур облегчает планирование операций и других медицинских процедур.
* **Анализ изображений высокого разрешения**: Способность трансформеров обрабатывать большое количество данных помогает в работе с детальными медицинскими изображениями.

**Преимущества и вызовы**

* **Преимущества**: Повышение точности диагностики, ускорение процесса анализа, уменьшение нагрузки на медицинский персонал.
* **Вызовы**: Необходимость большого объема качественных данных для обучения, обеспечение объяснимости и прозрачности моделей для доверия со стороны врачей и пациентов.

Обработка видео. Извлечение информации и отслеживание объектов в видеопотоке

Видеоданные содержат дополнительный временной аспект, который необходимо учитывать при анализе. Визуальные трансформеры расширяются для обработки видео, эффективно моделируя как пространственные, так и временные зависимости.

**Спатио-временные трансформеры**

* **Моделирование движений**: Трансформеры способны учитывать последовательность кадров, выявляя изменения и перемещения объектов во времени.
* **Отслеживание объектов**: Механизм внимания позволяет следить за объектами на протяжении видеопотока, даже если они временно скрываются или изменяют внешний вид.

**Применение в различных областях**

* **Безопасность и наблюдение**: Анализ видеопотоков для обнаружения подозрительных действий, контроля доступа и мониторинга общественных мест.
* **Развлечения и спорт**: Технологии захвата движений для создания спецэффектов, анализа спортивных матчей и тренировок.
* **Автономные системы**: Восприятие окружающей среды в реальном времени для навигации и взаимодействия с динамическими объектами.

**Технические аспекты**

* **Обработка больших объемов данных**: Видео генерирует огромные объемы информации, требующие эффективных алгоритмов и оптимизации вычислений.
* **Синхронизация пространственных и временных признаков**: Важно правильно балансировать внимание модели между текущей информацией и предыдущим контекстом.

Стабилизация и улучшение изображений. Применение для повышения качества изображений и устранения шумов

Улучшение качества изображений и видео является важной задачей в компьютерном зрении, особенно в условиях низкого освещения, шума или движений камеры. Визуальные трансформеры предлагают инновационные подходы для решения этих проблем.

**Стабилизация изображений**

* **Коррекция смещения**: Трансформеры могут анализировать последовательность кадров для выявления и корректировки нежелательных движений камеры.
* **Предсказание движения**: Модель обучается предсказывать траектории движения, что позволяет сглаживать видео и устранять дрожание.

**Устранение шумов и улучшение резкости**

* **Денойзинг**: Удаление различных типов шума с сохранением деталей и текстур изображения благодаря глобальному вниманию к контексту.
* **Суперразрешение**: Повышение разрешения изображения путем восстановления недостающей информации на основе соседних пикселей и глобальных зависимостей.

**Применения**

* **Фотография и видеосъемка**: Улучшение качества снимков, сделанных в сложных условиях, стабилизация видео, снятого без штатива.
* **Медицинская визуализация**: Улучшение диагностической ценности медицинских изображений за счет повышения четкости и контрастности.
* **Спутниковые и аэрофотосъемки**: Обработка изображений Земли для улучшения видимости деталей и облегчения дальнейшего анализа.

**Преимущества визуальных трансформеров**

* **Глобальное взаимодействие пикселей**: Моделирование зависимостей между отдаленными частями изображения позволяет более эффективно устранять дефекты.
* **Адаптивность**: Модели могут быть обучены на различных типах шумов и искажений, что делает их универсальными инструментами для улучшения изображений.

Заключение

В ходе данного реферата были подробно рассмотрены визуальные трансформеры — современный подход в области компьютерного зрения, основанный на использовании механизмов внимания. Мы исследовали теоретические основы трансформеров, их адаптацию для обработки изображений и преимущества, которые они предоставляют по сравнению с традиционными методами, такими как сверточные нейронные сети (CNN).

Визуальные трансформеры показали высокую эффективность в моделировании глобальных зависимостей в изображениях, что значительно расширяет возможности анализа и интерпретации визуальных данных. Их архитектура, основанная на механизмах самовнимания, позволяет учитывать взаимосвязи между отдаленными частями изображения, что особенно важно для сложных задач распознавания и понимания сцены.

Сравнение с CNN выявило, что, несмотря на более высокую вычислительную сложность и потребность в больших объемах данных для обучения, визуальные трансформеры обладают рядом существенных преимуществ. Они способны превосходить CNN в точности и универсальности при решении сложных задач, таких как обнаружение объектов, сегментация и анализ медицинских изображений.

Практические приложения визуальных трансформеров охватывают широкий спектр областей, включая компьютерное зрение, медицинскую диагностику, обработку видео и улучшение качества изображений. В каждой из этих сфер трансформеры продемонстрировали потенциал для повышения эффективности, точности и скорости обработки данных.

Будущее визуальных трансформеров обещает быть динамичным и перспективным. Одним из ключевых направлений развития является оптимизация моделей для работы с ограниченными вычислительными ресурсами. Разработка более легковесных архитектур, таких как компактные или разреженные трансформеры, позволит расширить их применение на мобильных устройствах и в реальном времени.

Другим важным направлением является улучшение обучаемости трансформеров на меньших объемах данных. Это может быть достигнуто за счет внедрения индуктивных предубеждений, характерных для CNN, таких как локальность и трансляционная инвариантность, в архитектуру трансформеров. Комбинация сильных сторон обеих архитектур может привести к созданию гибридных моделей, объединяющих преимущества трансформеров и свёрточных сетей.

Развитие технологий мультимодального обучения откроет новые горизонты в интеграции визуальных трансформеров с другими типами данных, такими как текст, аудио и сенсорная информация. Это позволит создавать более интеллектуальные системы, способные понимать и обрабатывать информацию так же комплексно, как это делает человек.

В области медицинской визуализации перспективы включают в себя разработку специализированных трансформерных моделей, способных обрабатывать трёхмерные данные и выполнять сложный анализ с высокой степенью точности и объяснимости. Это повысит качество диагностики и персонализированного подхода к лечению пациентов.

Кроме того, важным направлением является повышение интерпретируемости и прозрачности работы визуальных трансформеров. Разработка методов визуализации внимания и объяснения решений моделей поможет повысить доверие пользователей и облегчить внедрение трансформеров в критически важные области.

Наконец, с развитием аппаратного обеспечения и появлением новых технологий, таких как квантовые вычисления, возможно дальнейшее ускорение работы трансформеров и преодоление текущих ограничений по вычислительной сложности.

В целом, визуальные трансформеры имеют огромный потенциал для преобразования способов обработки и анализа визуальной информации. Их дальнейшее исследование и развитие будут способствовать продвижению искусственного интеллекта и открытию новых возможностей во многих сферах человеческой деятельности.