**Реферат на тему “Современное состояние систем распознавания образов и компьютерного зрения"**

Дисциплина: Распознавание образов

Подготовил студент группы

932001, Соколов В.С.

Томск 2023

**Введение**

Компьютерное зрение является одним из самых популярных и многообещающих направлений в компьютерном зрении, так как из визуальных данных можно извлечь больше всего информации.

Новинками в этой области можно считать нейронные сети на основе трансформеров со слоями внимания. Такие сети показали преимущество относительно традиционных для данной области сверточных нейронных сетей, также имели некоторые недостатки.

Генеративные модели на основе диффузии дали толчок к развитию качественной генерации изображений.

Мультимодальные модели научились работать с информацией разного типа и решать сложные задачи, решения которых раньше не существовало.

Также осуществляются попытки работать с 3D образами, такие технологии могут решить задачи создания компьютерного 3D образа окружающего пространства для более точного позиционирования автономных аппаратов, а также помогут с решением задач в AR. В эту категорию можно включить не только работу с изображениями, но и с облаками точек на примере лидарных данных, которые используются также для позиционирования и ориентирования автономных аппаратов.

Также новыми методами можно считать применение методов обучения с подкреплением к задачам компьютерного зрения.

Далее рассмотрим некоторые описанные пункты подробнее.

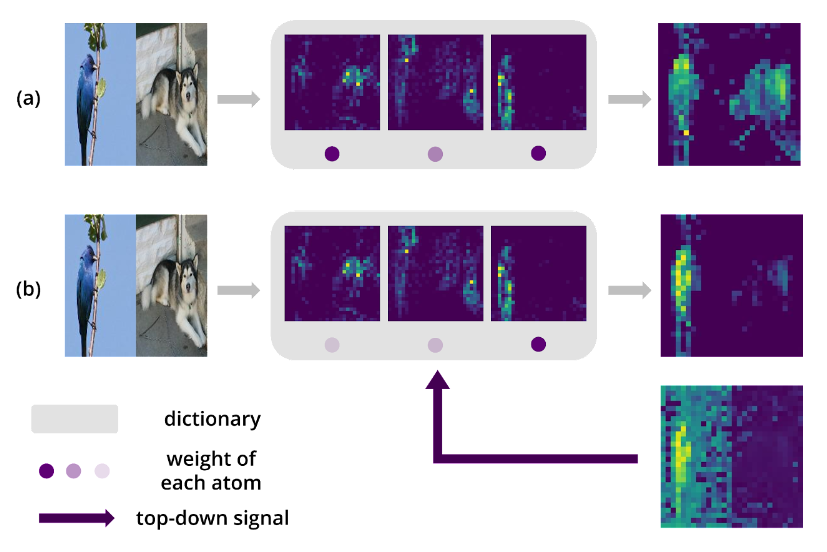
1. **Трансформеры (ViTs)**

Трансформерные архитектуры нейронных сетей были настолько точны для задач обработки естественного языка, что специалисты из Google Research использовали трансформер для обработки изображений в качестве альтернативы сверточным архитектурам. С помощью трансформеров удалось достигнуть наибольшей точности на многих задачах и наборах данных компьютерного зрения. Однако остаётся проблема производительности из-за большого числа параметров, следовательно, эффективности и надёжности, а также необходимости большого количества данных для обучения.

Классическая архитектура ViT разделяет изображение на последовательность фрагментов (патчей), затем к последовательности применяет несколько блоков трансформера: слой внимания (multi-head attention layer) и линейный слой (feed-forward layer).

* 1. AbSViT (обобщение)

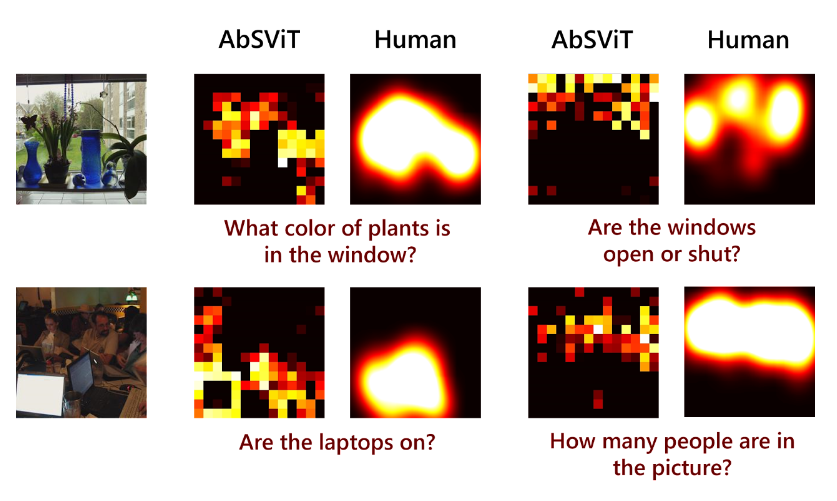
Зрительное внимание человека часто ориентировано на выполнение конкретной задачи. Мы склонны фокусироваться на разных объектах при выполнении задач. Когда мы отвечаем на вопросы об изображении, мы обращаем внимание только на образы, имеющие отношение к вопросам.



Авторы AbSViT предлагают новый подход «Анализ через синтез». Основу метода составляет утверждение, что человеческое зрение представляет собой систему байесовского вывода: наше понимание зависит от представления о мире.

Программно метод работает примерно также, промежуточное представление модулируется высокоуровневой задачей. Мы можем заставить модель концентрировать внимание на различных образах одном изображении.

Архитектурно предлагается прямой канал (обычный ViT) и канал с обратной связью, который содержит декодирующие блоки. Изображение проходит через прямой канал, далее изменяются веса выходящих токенов в соотвествии с текстовым описанием или задачей, токены отправляются обратно через канал обратной связи. Затем изображение проходит через прямой канал снова, но также промежуточно принимает токены с канала обратной связи.



Такая модель показывает результаты выше, чем ViT.

* 1. EfficientViT (эффективность)

Исследователи из Массачусетского технологического института, Лаборатории искусственного интеллекта MIT-IBM Watson и других организаций разработали более эффективную модель компьютерного зрения, которая существенно снижает вычислительную сложность этой задачи.

Исследователями были устранены проблемы с избыточным использованием памяти, избыточными вычислениями и использовании параметров. Были выявлены и устранены узкие места архитектуры ViT. Нелинейная функция преобразования на первом этапе преобразования изображений в токены была заменены на линейную, что ухудшило точность, но ускорило вычисления в 10 раз. Для улучшения точности были добавлены дополнительные блоки.

Такая модель может обрабатывать изображения лучшего качества и с большей скоростью, относительно её предшественников, что позволяет развёртывать её на меньших вычислительных системах, к примеру автомобили или мобильные утройства.

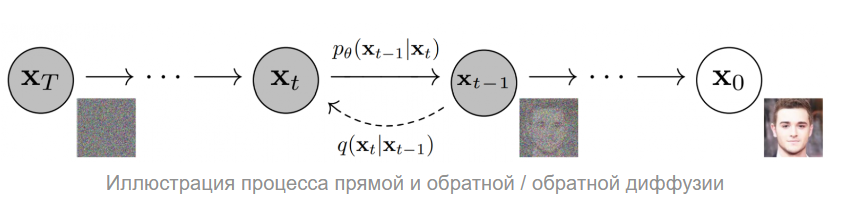
1. **Генеративные модели (GAI)**

Задачи генерации и редактирования изображений набирали популярность с момента создания генеративных состязательных сетей. В 2020 была разработана концепция диффузионных моделей, которая открыла возможности для генерации более качественных и разнообразных изображений при сохранении их семантической структуры. До сих пор имеется потенциал для развития и изучения диффузионных моделей.

Если обратиться к физике, то можно дать такое определение диффузии – перемещение частиц или молекул из области высокой концентрации в область низкой концентрации, обусловленное градиентом концентрации. Процесс диффузии обусловлен случайным движением молекул или частиц.

Диффузионные модели генерируют изображение из случайного шума поэтапно, что напоминает диффузию.

Идея заключается в том, что можно постепенно преобразовать одно распределение в другое. Структура исходного изображения разрушается путём добавления шума, затем модель восстанавливает изначальное изображения, таким образом выучивая исходное распределение. Этапность генерации данных позволяет точнее смоделировать изначальную выборку и получить изображение, лучшее по качеству и смыслу, в отличие от других моделей генерации. Популярной архитектурой считается UNet, который обучается описанным выше образом. Одной из распространённых моделей является Stable Diffusion.



Также модели диффузии позволили генерировать видеофрагменты.

1. **Мультимодальные модели**

Изобретение ViT позволило объединить работу с различными видами информации в одной модели.

Мультимодальные модели обычно состоят из нескольких унимодальных нейронных сетей, которые обрабатывают свой вид данных отдельно. Каждая такая сеть кодирует данные, далее информация должна быть объединена и интегрирована с помощью другой модели. Затем информация может быть использована для различных задач, таких как генерирование, классификация, регрессия.

С помощью мультимодальных моделей решают такие задачи как ответ на вопрос, который относится к изображению, генерировать изображение из текста (DALL-E), создавать видео, имея на входе только звук и т.д.

**Заключение**

Можно уследить тенденцию к усложнению архитектур нейронных сетей, увеличению возможностей взаимодействия с внешним миром, упрощению решению многих задач для человека. Стоит отметить, что человечество стоит только на пороге эры машинного обучения и больших данных. С увеличением вычислительных мощностей будут создаваться более сложные и качественные модели, которые в некоторых рутинных задачах превзойдут человека.