Название статьи (авторы статьи): **Flamingo: a Visual Language Model for Few-Shot Learning DeepMind 2022** (Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech and others)

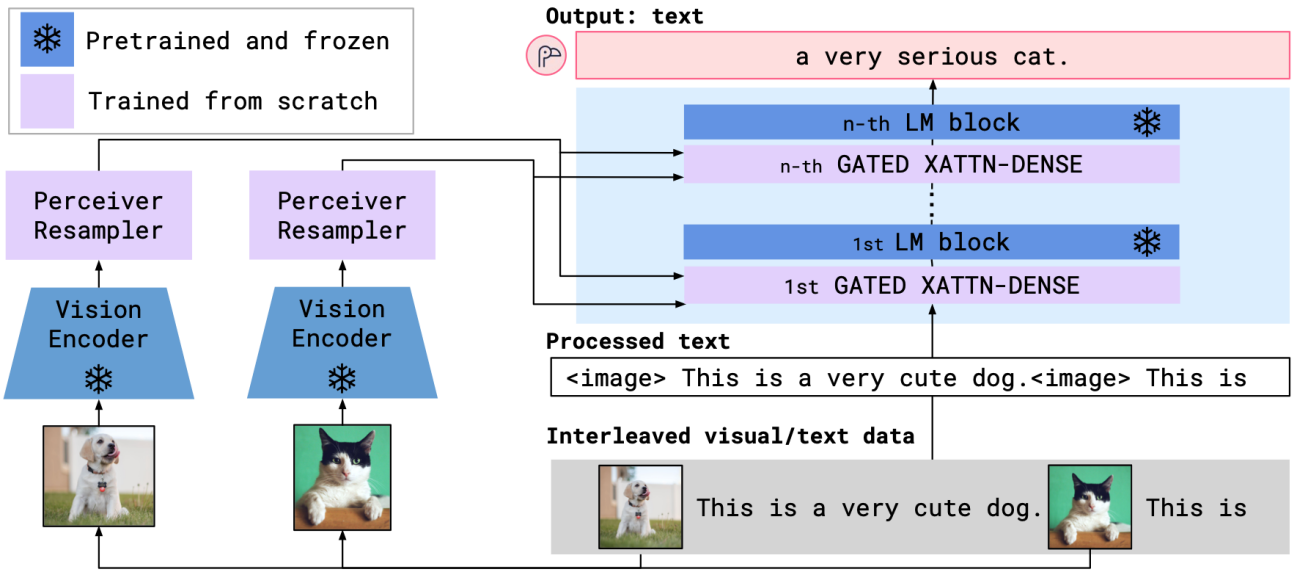
Автор обзора-рецензии: Маша Тимонина

В работе предлагается новая VLM (visual-language) модель, Flamingo. Она показала высокое качество ответа (6 SOTA результатов на paperworks весной 2022) в целом ряде визуально-языковых задач закрытого (например, множественный вопрос ответа из списка вариантов) и открытого типа, таких как ответ на вопросы по фотографии, описание действия на видео, решение арифметических примеров - достаточно лишь дать несколько примеров работы в новой постановке задачи (дообучить). Авторы подчеркивают, что особенный вклад работы состоит в адаптации нейросети к работе с видео и достижения высокого результата в режиме few-shot постановки задачи.

Flamingo опирается на две работы, тоже выпущенные в DeepMind: Chinchilla (70B параметров (other large autoregressive language models GPT-3, PaLM, Gopher)) - это огромная языковая модель для генерации, Perceiver IO - мультимодальная модель классификации. Модель строится на основе ранее известных больших и эффективных предварительно обученных чисто языковой модели и модели, исключительно работающих с изображениями. Параметры используемых языковой (Chinchilla) и визуальной моделей (трансформер) фиксируются (замораживаются), а промежуточные блок perceiver resampler и gated cross-attention слои внутри языковой части в предлагаемой архитектуре обучаются на смешанном наборе данных. Через обучаемые токены query внутри perceiver resampler можно получать выходы нужной структуры для генерации сложных выходов произвольного размера. Модель может принимать на вход последовательность чередующихся текстов и изображений/видеороликов и генерировать текстовый вывод.

В сентябре 2022 года публикация принята на международную конференцию NeurIPS 2022. Основные авторы статьи ведут исследовательскую деятельность в DeepMind. Трое из них - Jean-Baptiste Alayrac, Pauline Luc, Antoine Miech - ранее проходили обучение в Ecole Normale Supérieure (Франция) и состояли в группе научных разработок INRIA WILLOW под руководством Ивана Лаптева и Джозефа Цивика. Группа занимается исследованиями в задачах представления графических объектов (эмбеддинги), распознавания предметов и описания сцен через применение методов глубинного обучения в сочетании с теоретическими работами по статистическим методам

машинного обучения для генерации. Jean-Baptiste Alayrac имеет обширный опыт в NLP тематике. В 2021 году вместе с Antoine Miech участвовал в статье “Thinking Fast and Slow: Efficient text-to-visual retrieval with transformers”, а также опубликовался на ICLR 2022 со статьей “Perceiver IO: A General Architecture for Structured Inputs & Outputs”, базовой для одного из ключевых блоков Flamingo. Jeff Donahue в 2014 году совместно с Ross Girshick (FAIR - мы знаем о нем по RCNN, Masked Autoencoders, Detectron2, ViTDet) занимался вопросами извлечения признаков в object detection, а в последние годы исследовал генеративные adversarial сети, text-to-speech. Pauline Luc имеет серию публикаций по семантической сегментации изображений, тоже интересуется получением универсальных представлений для аудио данных. Antoine Miech специализируется на задачах распознавания и описания видеопотока, вопросно-ответных системах. Таким образом, мы можем пронаблюдать, как сотрудничество исследователей из смежных областей дает сегодня результаты в теме создания мультизадачных моделей (VLM).



В задаче создания визуально-языковых моделей параллельно развивается несколько направлений. В разделе Zero-Shot Cross-Modal Retrieval on COCO 2014 рейтинга PapersWithCode лидирующие позиции занимают модели, использующие в качестве главной функции потерь вариации contrastive loss [TCL (Vision-Language Pre-Training - Triple Contrastive Learning, 2022)]. При этом такие модели не обладают способностью принимать мультимодальный вход - либо только картинка, либо текст.

Рассматриваемая статья совсем новая, поэтому рано говорить о значительных продолжениях. Тем не менее, в ноябре 2022 года она имеет уже 103 цитирования. Из них интерес может представлять обзорная статья “Emergent Abilities of Large Language Models” от авторов из Google Research и Stanford University, исследующая свойство эмерджентности в очень больших языковых моделях (появление у системы свойств, не присущих её элементам в отдельности). Эта тема

возникала во время доклада на семинаре как вопрос о существовании нижнего предела на размер модели, при котором она уже не справляется выучивать осмысленные черты.

Среди сильных сторон работы можно выделить структурированность статьи, подробное описание архитектуры и общую легкость подачи материала, наглядные схемы. Использован ряд интересных приемов (vision feature resampling, cross-modal attention), позволяющих соединить готовые языковые и текстовые модели. Наконец, авторы из DeepMind (видимо не только основные, но и с помощью соавторов из длинного списка фамилий) смогли подготовить качественную коллекцию мультимодальных датасетов: слабо связанные текст или картинка в виде текста веб страниц и layout'a, и сильно связанные датасеты с парами картинка-текст и видео-текст. Авторы отмечают, что хороший подбор и чистка обучающих данных сыграли решающую роль в получении такого высокого качества работы модели в режиме few-shot.

Среди ревьюеров NeurIPS перед утверждением работы на конференцию имело место оживленное обсуждение и даже спор с авторами в комментариях. Центральная линия критики - отказ команды DeepMind от публикации весов обученной модели и собранного мультимодального датасета. В сложившейся ситуации никто не сможет повторить эксперименты статьи и проверить их результаты, поскольку обучение такой большой модели требует объема ресурсов, имеющегося только у единичных корпораций. Более того, авторы подтвердили, что не намерены в ближайшем будущем предоставлять доступ в виде API внешним пользователям, поскольку не могут гарантировать безопасность технологии, а значит использование алгоритма для прикладных задач будет полностью ограничено. В защиту своих результатов авторы напомнили, что большинство экспериментов из статьи проводилось на бенчмарках со скрытыми тестовыми данными (VQAv2, VizWiz, STAR, VisDial, TextVQA, etc.). Приходится отмечать, что разработка нейросетевых алгоритмов порождает все больше этических дискуссий.

В ревью статьи к конференции разрослось детальное обсуждение работы CM3: A Causal Masked Multimodal Model of the Internet (Aghajanyan et al., 2022, Facebook AI Research) - из-за оговорки авторов Flamingo о наследовании в CM3 идей Flamingo. Впоследствии команде DeepMind даже пришлось внести корректировки в раздел Related Works своей оригинальной статьи и заменить формулировку на “concurrent”. В действительности работы имеют куда больше различий, чем общих черт. CM3 обрабатывает вход-картинку сразу как текстовый токен, на всех слоях использует all-to-all self-attention между токенами. В Flamingo визуальная информация обрабатывается Perceiver Resampler и затем уже в LM с помощью gated cross-attention слоев смещает внимание к последнему предшествующему изображению. Есть и другие существенные отличия. Чтобы построить совместную вероятность слов и картинок, в CM3 ко входу применяют дискретизацию VQ-VAE. Этот подход позволяет на выход генерировать не только текст, но и изображения, а именно html-разметку страницы сайта, что не входит в круг возможных задач Flamingo. Авторы CM3 в то же время отмечают, что дискретизация приводит к упущению паттернов и мелкого текста на входных картинках, с чем Flamingo хорошо справляется через совершенно другую архитектуру (комбинация causal и masked подходов в

CM3). Третий уровень различий - подходы к выбору задач и оценке качества моделей. Flamingo позиционируется как модель для решения целого спектра задач multiple input - to - text, которая делает особые успехи в режиме few-shot и может обрабатывать видео. CM3 - это модель в первую очередь для суммаризации и устранения неоднозначности (упрощения) в hypertext входе (страничках сайтов), при файн-тюнинге показывает в этих задачах SoTA результаты. В то время как команда Flamingo ставит эксперименты на множестве популярных бенчмарков, CM3 описывает именно zero-shot результаты на CC-NEWS+EnglishWikipedia (Common Crawl News) и COCO. Таким образом, эти модели трудно считать сравнимыми, что может навести нас на мысли о излишней концентрации DeepMind на своих успехах и определенной чрезмерности пиара своих результатов, в которых компанию уличают не первый раз.