КРОСС-МОДАЛЬНЫЙ ПЕРЕНОС ЗНАНИЙ КАК ФАКТОР РАЗВИТИЯ ЛОГИЧЕСКОГО РАССУЖДЕНИЯ В ЯЗЫКОВЫХ МОДЕЛЯХ

Оруджев М.М.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Масейчик Е.А. – ассистент

В данной работе рассматривается феномен кросс-модального переноса знаний (cross-modal transfer learning) в контексте больших языковых и мультимодальных моделей. Утверждается, что включение видеоданных в процесс предобучения систематически улучшает способность моделей к логическому рассуждению в текстовых задачах – эффект, не объяснимый исключительно увеличением объёма текстового корпуса.

Принятая парадигма в масштабировании языковых моделей предполагает, что повышение когнитивных способностей ИИ достигается посредством увеличения объёма обучающих текстов [1]. Однако, накопленные эмпирические данные свидетельствуют как раз об обратном: разнообразие модальностей, но отнюдь не масштаб однородных данных, является в большей степени фундаментальным фактором для развития общих способностей к рассуждению.

Основным эмпирическим подтверждением данного тезиса служит работа [2], представившая модель MERLOT, предобученную на 180 миллионах видеофрагментов с транскриптами. Авторы зафиксировали значимое улучшение показателей модели на текстовых бенчмарках в области здравого смысла (HellaSwag, VCR) в сравнении с текстовыми базовыми линиями аналогичного масштаба. Авторы именно таким образом интерпретируют этот результат, а именно: видеопоследовательности обеспечивают модель темпоральными причинно-следственными структурами, то есть «законами мира». В явном виде вывести их из текста крайне сложно. В системе VideoBERT получены подобные результаты [3]. Совместное обучение по видео вместе с текстом повысило точность заполнения масок в языковых задачах, тем самым косвенно указывая на обогащение семантических репрезентаций.

В то же самое время модель CLIP показала, что контрастивное обучение на парах изображение-текст формирует семантически намного более плотные представления в сравнении с обучением на одном из каналов, а в работе о системе Gato показали, что единая мультимодальная политика, обученная на разнородных задачах (включая игровые видеоданные), демонстрирует способность к zero-shot генерализации на новые текстовые задачи, а это свидетельствует о переносе абстрактных структур рассуждения.

С когнитивно-научной точки зрения, этот эффект в целом согласуется с теорией embodied cognition: концептуальные репрезентации укоренены в перцептивном опыте и также в моторном опыте. Во время наблюдения моделью за физикой движения объектов, столкновений, гравитации, она формирует неявные каузальные схемы, которые затем активируются при решении текстовых задач на рассуждение. Йошуа Бенжио вместе с другими в теории глубоких представлений предположили, что признаки должны отражать основные факторы вариативности наблюдаемых данных, при этом видеоданные делают эти факторы (движение, время, причинность) явными для модели .

Полученные результаты ставят под сомнение саму идею насчет «текстоцентричной» гипотезы масштабирования. Логическое мышление отчасти является эмерджентным продуктом мультимодального восприятия. Значит, достичь artificial general intelligence можно не увеличением текстовых данных, а созданием систем, которые объединяют разные типы данных в общие причинно-следственные модели мира.

Cross-modal transfer learning являет собой не технический артефакт, а скорее основной механизм для формирования общего интеллекта. Обучение с помощью видео улучшает текстовую логику, потому что предоставляет моделям доступ к структурам реального мира. В чисто лингвистическом пространстве подобные структуры оказываются недоступными для модели.

Список использованных источников:

1. Scaling laws for neural language models / J. Kaplan [et al.] // arXiv preprint. – 2020. – arXiv:2001.08361.

2. MERLOT: Multimodal neural script knowledge models / R. Zellers [et al.] // Advances in Neural Information Processing Systems (NeurIPS). – 2021. – Vol. 34. – P. 23634–23651.

3. VideoBERT: A joint model for video and language representation learning / C. Sun [et al.] // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). – 2019. – P. 7464–7473.