ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 508, с. 146–148

ПЕРЕДОВЫЕ ИССЛЕДОВАНИЯ В ОБЛАСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И МАШИННОГО ОБУЧЕНИЯ

УДК 004.8

FusionBrain: ИССЛЕДОВАТЕЛЬСКИЙ ПРОЕКТ ПО МУЛЬТИМОДАЛЬНОМУ И МУЛЬТИЗАДАЧНОМУ ОБУЧЕНИЮ

© 2022 г. Д. В. Димитров[1,2,*], А. В. Кузнецов[1,2], А. А. Мальцева[1], Е. Ф. Гончарова[2]

Представлено академиком РАН С.С. Гончаровым
Поступило 28.10.2022 г.
После доработки 28.10.2022 г.
Принято к публикации 01.11.2022 г.

FusionBrain – это исследовательский проект, основными задачами которого являются разработка эффективных мультизадачных и мультимодальных моделей и применение их для решения широкого круга практических задач. Общая цель и идея проекта – научиться создавать модели, которые смогут как можно более эффективно извлекать дополнительные важные знания из большого количества модальностей и задач при обучении, и за счет этого лучше решать разные другие задачи. Исследования проводятся во многих модальностях: тексты, изображения, аудио, видео, языки программирования, графы (например, молекулярные структуры), временные ряды и так далее. Список решаемых задач очень большой: от классических задач CV и NLP до задач, вовлекающих разные модальности: VideoQA, Visual Commonsense Reasoning, IQ tests (эти задачи сложны даже для человека). Изучается также способность моделей решать задачи, сформулированные на естественном или визуальном языках, и даже справляться со скрытыми задачами (для которых в обучающей выборке отсутствовали примеры). Исследования сосредоточены в том числе на сокращении данных, человеческих и вычислительных ресурсов, необходимых для обучения и инференса различных моделей. В рамках данного материала мы поделимся полученными результатами в рамках исследования и разработки некоторых мультимодальных и мультизадачных архитектур.

Ключевые слова: мультимодальность, мультизадачность, компьютерное зрение, обработка естественного языка, нейронные сети, трансформеры, фундаментальные модели, FusionBrain

DOI: 10.31857/S2686954322070244

Информация, обрабатываемая мозгом человека в каждый момент времени и необходимая для принятия даже самых простых повседневных решений, имеет разную природу и представлена в самом разном виде (по-другому – представлена в разных модальностях). Для восприятия такой разнородной информации человек использует свои органы чувств, а для ее анализа – специальные зоны головного мозга (и, как следствие, специализированные знания, полученные в течение жизни): так, визуальная информация требует зрительного восприятия, слуховая информация предполагает восприятие и анализ звука, обработка текстов на естественном языке предполагает знание языка, и так далее. Почти всегда при этом для успешного решения возникающих задач и проблем в реальном мире необходимо использовать одновременно информацию, поступающую из разных модальностей, так как сами задачи по своей природе вовлекают несколько таких модальностей (например, вождение автомобиля, просмотр фильма, ответы на разные вопросы, и так далее).

Тем не менее в науке о данных и машинном обучении исторически сложилось так, что изучению и способам обработки каждой из основных модальностей посвящены отдельные области, часто не сильно пересекающиеся: например, в рамках CV (computer vision или компьютерного зрения) разрабатываются модели, которые решают задачи, связанные с анализом изображений, 3D-объектов или видео, в рамках NLP (natural language processing или обработки естественного языка) изучаются архитектуры, которые умеют работать с текстовыми данными на разных языках, в рамках PLP (program language processing) – с кодом на разных языках программирования, от-

[1] *ПАО Сбербанк, Москва, Россия*
[2] *AIRI, Москва, Россия*
*****E-mail: Dimitrov.D.V@sberbank.ru*

дельно стоят модели, работающие с временными рядами разной природы и табличными данными. Из-за этого разрабатываемые модели (особенно те, которые работают и используются в реальных бизнес-процессах) в большинстве своем умеют работать строго с одним типом данных и решать ровно одну узкоспециализированную задачу, на которую и были обучены.

Но последние несколько лет все больше исследований ведется в области разработки мультимодальных и мультизадачных архитектур. Помимо того, что это современный тренд, это еще и большой научный и инженерный вызов, еще один шаг на пути к созданию сильного искусственного интеллекта. В настоящее время ведется большее количество исследований в области мультимодальных и мультизадачных моделей. Все разработки ведутся в нескольких направлениях, исследуются либо трансформеры с архитектурами энкодер-декодер [1, 2], либо только декодерные трансформеры [3]. В своих исследованиях авторы экспериментально подбирают задачи, которые используются на претрейне, чтобы затем обученная мультимодальная модель могла решать множество задач в режиме zero-shot.

Мы предлагаем модель RUDOLPH – мультизадачную модель-декодер, способную решать ряд задач на стыке двух модальностей: текст и изображение. На претрейне RUDOLPH обучался на 3 типах задач – text2image, image2text и text2text. На вход в модель подается следующая последовательность токенов: текстовые, картиночные, текстовые. Такая комбинация позволяет обучать текстово-визуальные и визуально-текстовые задачи. Наряду с текстовыми и визуальными токенами, используемыми в трансформере, вводятся спецтокены, отражающие конкретную задачу на обучение. Эти спецтокены явно подсказывают модели, какая конкретно задача пришла на обучение в текущий момент. Благодаря такому обучению, на инференсе модель способна сама определить задачу, при этом качество генерации становится выше. Существует три версии модели RUDOLPH: 350M, 1.3B, 2.7B.

Мы стремимся способствовать развитию такой перспективной и сложной области, как мультимодальные исследования, и проводим соревнование Fusion Brain Challenge 2.0. В рамках данной задачи предлагается построить единую multitask-модель, которая бы успешно решала подзадачи в двух модальностях (визуальной и текстовой), принимая на вход описания подзадач, выраженные на естественном русском языке, например: “сгенерируй изображение”, “опиши изображение”, “ответь на вопрос” и т.д. В состав входит 12 подзадач, из которых 6 известны участникам с момента начала Конкурса (открытые подзадачи), а 6 неизвестны (скрытые подзадачи) и представляют собой частные случаи открытых задач (имеют некоторые отличительные особенности в постановке). Основная задача участников заключается в построении и обучении единой мультимодальной мультизадачной архитектуры, которая позволила бы получить максимальные значения метрик для каждой отдельной подзадачи и, как следствие, достичь максимального значения интегральной метрики на 12 подзадачах.

К открытым подзадачам относятся Text QA, Mathematical QA, Image Generation, Image Captioning, Visual QA, Text Recognition in the Wild. Подзадача Text QA – задание на понимание прочитанного текста. Для успешного решения модель должна уметь устанавливать причинно-следственные связи, разрешать кореференции, а также определять правильную последовательность действий, учитывая временную информацию. Подзадача Mathematical QA проверяет способность модели выполнять простейшие арифметические действия, необходимые для решения линейных уравнений или систем линейных уравнений, а также производить операции сравнения. Подзадача Image Generation подразумевает генерацию изображений на основе текстовых описаний на русском языке. Ответом на подзадачу является изображение, чье содержание соответствует входному текстовому описанию. Подзадача Image Captioning подразумевает генерацию текстовых описаний на русском языке к изображениям. Ответом на подзадачу является текстовая строка, содержащая текстовое описание входного изображения. Подзадача Visual QA предполагает, что обученная модель способна формировать ответ на вопрос по изображению. В этой подзадаче на вход модели подается пара вида “текстовый вопрос – картинка”, а выходом является соответствующий текстовый ответ. Подзадача Text Recognition in the Wild – задание на распознавание текста в городской или иной подобной местности (вывески, дорожные знаки, рекламные объявления и т.п.). Данные представляют собой фотографии объектов с изображенным на них текстом. Ответом на подзадачу является текстовая строка. Скрытые задачи относятся к этим же модальностям и позволяют оценить обобщающую способность модели. То есть модель, обученная на открытых задачах, должна использовать свои знания в решение скрытых задач.

Baseline решение для соревнования FusionBrain основано на модели RUDOLPH. В качестве базовой модели мы использовали модель RUDOLPH 2.7B, дообученную для решения шести открытых задач FBC2.

В заключение отметим, что наш вклад заключался в следующем – были подготовлены данные

как тестовые, так и для обучения, была определена постановка задачи и подготовлена платформа для соревнования Fusion Brain Challenge 2.0. Также был разработан baseline, обученный на 6 открытых задачах и сочетающий мультимодальный и мультизадачный подход. Помимо этого, были разработаны специализированные метрики под каждую задачу и общая метрика для оценки моделей.

СПИСОК ЛИТЕРАТУРЫ

1. *Wang P. et al.* Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework//CoRR. 2022.
2. *Wang W. et al.* Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks//arXiv preprint arXiv:2208.10442. 2022.
3. *Reed S. et al.* A Generalist Agent // arXiv preprint arXiv:2205.06175. 2022.