ОТ СЛАБОГО ИИ К ОБЩЕМУ УНИВЕРСАЛЬНОМУ ИНТЕЛЛЕКТУ (ОБЗОР ТЕНДЕНЦИЙ 2020-2023)

Визильтер Юрий Валентинович, д.ф.-м.н., профессор РАН,
директор по направлению – руководитель научного комплекса
«Искусственный интеллект и техническое зрение» ФАУ «ГосНИИАС»
viz@gosniias.ru

Начало нынешнего этапа развития технологий искусственного интеллекта (ИИ) принято отсчитывать от 2011 г. с появлением в области компьютерного зрения глубоких сверточных нейронных сетей (ГНС, CNN), которые позволили решать задачи распознавания визуальных образов на уровне человека [1]. Внедрение CNN позволило уже к 2015-16 гг. решить целый спектр практически значимых задач в области компьютерного зрения и анализа больших данных.

В 2016-2017 гг. получил широкое распространение ряд новых подходов и научных результатов в области машинного обучения, таких как генеративно-состязательные сети (GAN) для синтеза реалистичных сигналов и изображений [2], архитектуры типа Transformer [3], GPT [4], BERT [5], на основе т.н. «модулей внимания» (Attention [3]) для анализа текстов на естественном языке (NLP), обучение ГНС методом подкрепления (Reinforcement Learning [6]-[8]), глубокое обучение с использованием структурных моделей (Graph CNN [9]), автоматическое конструирование и обучение глубоких сетей (Auto-ML, NAS, HPO [10]).

На современном этапе (2020-2023 гг.) можно указать следующие основные тенденции развития технологий машинного обучения.

В области компьютерного зрения следует, в первую очередь, отметить тренд перехода лидерства от CNN к визуальным трансформерам [11]-[13]. При этом перспективы CNN и гибридных архитектур также сохраняются (см., например, ConvNeXt [14], ConvNeXt v2 [15]).

В области анализа текстов на естественном языке (NLP) главным трендом стало создание и использование больших языковых (LLM) и фундаментальных (многомодальных) моделей (Foundation Models [16]), таких как GPT-3, ChatGPT [17], GPT-4 [18].

В области обучения с подкреплением (RL) новые тенденции были связаны с появлением таких методов и результатов как обучение LLM с человеком в обратной связи (RLHF [17]), совместное обучение универсальных агентов анализу данных и игровым задачам (GATO [19]), обучение с открытым списком виртуальных сред и целевых задач для приобретения когнитивного поведения (Open-Ended Learning [20]), обучение с подкреплением путем программирования LLM для поиска управляющих решений (GITM [21]).

Представляется, что последняя работа [21] вместе с рядом работ из других областей (см., например, [22]) указывает на появление новой парадигмы программирования в ИИ, которую можно назвать «программированием на LLM для извлечения и применения знаний». Эта новая дисциплина разработки и оптимизации запросов для эффективного использования языковых моделей (LM) получила название «промпт-инжиниринг». Она используется для повышения прозрачности и безопасности LLM, извлечения и добавления знаний, организации использования внешних инструментов и взаимодействия LLM.

В области реалистичной генерации данных мы наблюдаем переход лидерства от GAN [23] к диффузным моделям (Diffusion Models [24], [25]).

В области универсальных моделей для анализа данных и управления наиболее значимыми и перспективными представляются такие подходы как универсальные агенты (GATO [19]), ГНС для одновременного анализа данных разных типов (Perceiver IO [26], Unified-IO [27]), а также кооперативные модели ГНС, общающихся между собой на естественном языке с целью совместного решения задач (Socratic Models [28]). Работа [28] – пример реализации нового подхода к созданию интеллектуальных агентов, которые функционируют и обучаются одновременно в двух средах – физической и языковой (символьной), вследствие чего выучивают 2 типа поведения (физическое и языковое), которые вместе помогают им быть эффективными в обеих средах.

В целом появление подобных универсальных и кооперативных агентов на основе языковых и мультимодальных моделей представляется важным свидетельством в пользу начинающегося «большого объединения» всех методов, подходов и задач машинного обучения, а также их объединения с технологиями ИИ на основе формализации знаний и логического вывода. Результатом такого «большого объединения» может стать создание «прозрачного» и «объяснимого»

универсального ИИ для решения, в том числе, задач автономного управления сложными техническими объектами.

СПИСОК ЛИТЕРАТУРЫ

1 Krizhevsky A., Sutskever I., Hinton G.E. ImageNet classification with deep convolutional neural networks // Communications of the ACM. 2017. V. 60. № 6. P. 84–90.

2 Goodfellow I.J., et al. Generative Adversarial Networks // arXiv:1406.2661.

3 Vaswani A., et al. Attention Is All You Need // arXiv:1706.03762.

4 Radford A., Karthik N. Improving Language Understanding by Generative Pre-Training. 2018.

5 Devlin J., et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018.

6 Mnih V., et al. Human-level control through deep reinforcement learning // Nature. 2015.

7 Schulman J., Wolski F., Dhariwal P., Radford A., Klimov O. Proximal policy optimization algorithms // arXiv preprint arXiv:1707.06347. 2017.

8 Badia A.P., et al. Agent57: Outperforming the atari human benchmark // arXiv preprint arXiv:2003.13350. 2020.

9 Zhang S., Tong H., Xu J., Maciejewski R. Graph convolutional networks: a comprehensive review // Comput Soc Netw 6. 2019. №11.

10 Xin H., Kaiyong Z., Xiaowen C. AutoML: A Survey of the State-of-the-Art // arXiv:1908.00709v6. 2021.

11 Dosovitskiy A., et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale // arXiv:2010.11929. 2021.

12 Liu Z., et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows // arXiv:2103.14030. 2021.

13 Radford A., et al. Learning transferable visual models from natural language supervision // arXiv preprint arXiv:2103.00020. 2021.

14 Zhuang L., et al. A ConvNet for the 2020s // arXiv:2201.03545. 2022.

15 Sanghyun W., et al. ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders // arXiv:2301.00808. 2023.

16 Rishi B., et al. On the Opportunities and Risks of Foundation Models // arXiv:2108.07258. 2021.

17 Ouyang L., et al. Training language models to follow instructions with human feedback //Advances in Neural Information Processing Systems. 2022. T. 35. C. 27730-27744.

18 OpenAI, GPT-4 technical report // arXiv preprint arXiv: 2303.08774. 2023.

19 Scott R., et al. A Generalist Agent // arXiv:2205.06175. 2022.

20 Adam S., et al. Open-Ended Learning Leads to Generally Capable. Agents // arXiv:2107.12808. 2021.

21 Xizhou Z., et al. Ghost in the Minecraft: Generally Capable Agents for Open-World Environments via Large Language Models with Text-based Knowledge and Memory // arXiv:2305.17144. 2023.

22 Fantechi A., Gnesi S., Semini L. Rule-based NLP vs ChatGPT in Ambiguity Detection, a Preliminary Study // Joint Proceedings of REFSQ-2023 Workshops. 2023.

23 Jun-Yan Z. et al. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks // ICCV. 2017.

24 Ling Y., et al. Diffusion Models: A Comprehensive Survey of Methods and Applications // arXiv:2209.00796. 2022.

25 Aditya R., et al. Hierarchical Text-Conditional Image Generation with CLIP Latents // arXiv:2204.06125. 2022.

26 Andrew J., et al. Perceiver IO: A General Architecture for Structured Inputs & Outputs // arXiv:2107.14795. 2022.

27 Lu et al., Unified-IO: A Unified Model for Vision, Language, and Multi-Modal Tasks // arXiv:2206.08916. 2022.

28 Andy Z., et al. Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language // arXiv:2204.00598. 2022.