УДК 004.853

СНИЖЕНИЕ ЭФФЕКТА КАТАСТРОФИЧЕСКОГО ЗАБЫВАНИЯ ПРИ ДООБУЧЕНИИ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ НА ДАННЫХ МЕДИЦИНСКОГО ДОМЕНА С ИСПОЛЬЗОВАНИЕМ МЕТОДА ЭКСПОНЕНЦИАЛЬНОГО СГЛАЖИВАНИЯ

Свириденко Д.К. (НИЯУ МИФИ)
Научный руководитель – доктор технических наук, профессор Зайцев К.С.
(НИЯУ МИФИ)

Введение. Внедрение больших языковых моделей (LLM) в высокотехнологичные и социально значимые отрасли, такие как медицина, требует их дообучения на специализированных корпусах данных для обеспечения точности терминологии и фактологии. Однако данный процесс неизбежно сопряжен с эффектом «катастрофического забывания» — фундаментальной проблемой искусственных нейронных сетей, при которой модель, оптимизируясь под новое распределение данных, необратимо искажает веса, отвечающие за ранее приобретенные знания. Это приводит к деградации общих когнитивных способностей модели: снижается качество генерируемого текста на общие темы и утрачивается способность решать задачи, не связанные с целевым доменом [1]. В современной практике для минимизации эффекта забывания применяются различные подходы: методы регуляризации весов (например, EWC), ограничивающие изменение значимых параметров; архитектурные методы, выделяющие отдельные подсети под новые задачи; а также методы воспроизведения, требующие смешивания новых данных с историческими выборками [2, 3]. Однако последние часто неприменимы из-за политик конфиденциальности данных, а существующие методы регуляризации ограничивают адаптивность модели. Актуальной задачей является разработка вычислительно эффективного метода, обеспечивающего баланс между усвоением новых специализированных знаний и сохранением генеративных способностей базовой модели.

Основная часть. Для решения поставленной задачи предлагается использование метода экспоненциального скользящего среднего в рамках параметрически-эффективной настройки (PEFT). Суть решения заключается в замене статичного регуляризатора на динамическую модель-учителя, веса которой формируются и обновляются как скользящее среднее параметров обучаемой модели непосредственно в процессе адаптации [4]. Это позволяет модели-учителю плавно обучаться на новых данных вместе с моделью-учеником, обеспечивая стабильность обучения без необходимости хранения старых данных.

В экспериментах использовалась модель Qwen2-1.5B с 4-битным квантованием весов. Она была обучена 3 способами: дообучением с учителем без регуляризации, дообучением с учителем с регуляризацией в виде модели-учителя с неизменяемыми весами и предлагаемым методом с использованием экспоненциального скользящего среднего.

Выводы. Разработан метод параметрически-эффективной тонкой настройки для адаптации больших языковых моделей к данным нового домена на языке программирования Python. Проведен сравнительный анализ качества адаптации моделей, обученных различными способами, на данных медицинского домена, а также сохранения общих знаний на бенчмарке MMLU.

Список использованных источников:

1. Y. Luo, Z. Yang, F. Meng, Y. Li, J. Zhou, and Y. Zhang, “An empirical study of catastrophic forgetting in large language models during continual fine-tuning,” arXiv preprint arXiv:2308.08747, 2023.
2. Kirkpatrick J. et al. Overcoming catastrophic forgetting in neural networks // Proceedings of the national academy of sciences. – 2017. – T. 114. – №. 13. – C. 3521-3526.
3. Zhang, Y., Jiang, S., Zhao, M., Li, Y., Fan, Y., Wu, X., & Chen, Q. (2025). Gere: Towards efficient anti-forgetting in continual learning of llm via general samples replay. arXiv preprint arXiv:2508.04676.
4. Tarvainen A., Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results // Advances in neural information processing systems. – 2017. – T. 30. – C. 1195-1204.