УДК 004.8

РАЗРАБОТКА АЛГОРИТМА СЕМАНТИКО-ГРАФОВОЙ ОЦЕНКИ И АДАПТИВНОГО РАСПРЕДЕЛЕНИЯ ТЕКСТОВЫХ ФРАГМЕНТОВ ДЛЯ РАСШИРЕНИЯ КОНТЕКСТНОГО ОКНА МАЛЫХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Воробьев Л. Г.[1], Бравичев К. А.[1]
Научный руководитель – Бравичев К. А.[1]
[1]Университет ИТМО
extrasparrow@gmail.com

Введение

В настоящее время большие языковые модели активно внедряются в разные области разработки программ, в том числе и в индустрии компьютерных игр. Они применяются как в технических аспектах (написание программного кода, процедурная генерация окружения), так и в творческих (создание сюжетов, диалогов и нарративных элементов) [1]. Отдельным направлением стало развитие малых языковых моделей, которые получили широкое распространение благодаря возможности локального развертывания на устройствах с ограниченным объемом видеопамяти.

Однако фундаментальная архитектура языковых моделей обладает серьезным ограничением, заключающимся в квадратичной сложности механизма самовнимания при линейном росте Key-Value кэша (KV) [2]. В ситуациях, когда диалог становится многооборотным и применяется Retrieval Augmented Generation (RAG), модель может терять контекстную связь с ранним диалогом или генерировать галлюцинации.

Существует несколько подходов к решению данной проблемы: применение стратегий вытеснения данных для управления памятью KV-кэша [3], проведение графового анализа перед непосредственным инференсом модели [4]. В данной работе предлагается комбинированный алгоритм, который, в отличие от других решений, одновременно учитывает семантическую близость и графовую связность между запросом пользователя и текстовыми фрагментами из базы знаний. Наиболее важные фрагменты остаются в видеопамяти, остальные – архивируются в оперативной памяти (ОЗУ) для возможного повторного использования, что снижает требования к видеопамяти и сохраняет контекстную связь между частями диалога.

Основная часть

Предлагаемый в данной работе алгоритм использует гибридную оценку текстовых фрагментов на основе двух взаимодополняющих критериев: семантическое сходство и графовый анализ. Каждому критерию присваивается весовой коэффициент, определяющий его значимость для конкретного сценария. Значения весов определяются эмпирически на основе тестирования системы.

При семантическом анализе на первичном этапе текст запроса пользователя и кандидаты текстовых фрагментов из базы знаний преобразуются в векторные представления (эмбеддинг) с помощью специализированных моделей, которые кодируют семантические данные текста и позволяют сравнивать их смысловую близость. Сравнение производится через косинусную метрику расстояния. Полученный результат нормализуется в необходимый диапазон и используется как первая оценка контекстного соответствия.

Графовый анализ начинается с извлечения сущностей из текста запроса и каждого проверяемого текстового фрагмента – люди, места, события, термины. Из полученных данных строится направленный граф, в структуру которого входят узлы – сущности, а ребра показывают степень их взаимосвязанности. Чем больше общих сущностей, тем выше вероятность, что фрагмент наиболее ценен для текущего контекста. Степень связи

рассчитывается через длину кратчайшего пути между узлами графа с помощью алгоритма Дейкстры. Полученное расстояние нормализуется и используется как вторая метрическая оценка в системе.

На каждой итерации система определяет набор кандидатов для оценивания. Для каждого из кандидатов рассчитывается гибридная оценка из двух критериев. На основе полученных оценок выполняется ранжирование фрагментов. Фрагменты, чья оценка превышает целевое пороговое значение, загружаются в видеопамять для текущей итерации работы системы, тогда как все остальные архивируются в ОЗУ.

Данный подход обеспечивает два ключевых преимущества. Во-первых, сохраняется весь найденный контекст для его повторного использования в последующих итерациях диалога. Во-вторых, алгоритм адаптивен к меняющимся условиям работы диалога. В отличие от традиционных стратегий вытеснения (FIFO, LIFO) предложенное решение учитывает актуальный для текущего диалога контекст, что особенно важно в длительных сценариях ведения диалога.

Прототип системы был разработан и установлен на устройстве с объемом видеопамяти 16 GB. Для оценки эффективности алгоритма планируется использовать метрики Recall@K, Multi-turn Accuracy, Coverage Score, Latency с применением инструмента RAGAS.

Выводы

На основе анализа существующих решений разработан алгоритм, применяющий гибридную оценку семантического и графового анализа [5]. Предложенное решение предотвращает потерю информации и обеспечивает эффективное использование видеопамяти, что обеспечивает перспективы для интеграции в системы обработки естественного языка в студиях по разработке компьютерных игр и компаний, которым требуется локальное развертывание при ограниченных аппаратных ресурсах.

Литература

1. Maleki M. F., Zhao R. Procedural Content Generation in Games: A Survey with Insights on Emerging LLM Integration // AIIDE '24: Proceedings of the Twentieth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment. 2024. Vol. 20, no. 1. P. 167–178. https://doi.org/10.1609/aiide.v20i1.31877.
2. Li B. [и др.] LLM Inference Serving: Survey of Recent Advances and Opportunities // 2024 IEEE High Performance Extreme Computing Conference (HPEC). 2024. P. 1–8.
3. Kwon W. [и др.] Efficient Memory Management for Large Language Model Serving with PagedAttention // SOSP '23: Proceedings of the 29th Symposium on Operating Systems Principles. 2023. P. 611–626. https://doi.org/10.1145/3600006.3613165.
4. Wang X. [и др.] Knowledge Graph-Based Semantic Ranking for Efficient Semantic Query // 2022 IEEE 10th International Conference on Computer Science and Network Technology (ICCSNT). Dalian, China, 2022. P. 75–79.
5. Воробьев Л. Г. Репозиторий S-GAS Manager [Электронный ресурс]. – Режим доступа: https://github.com/TownSparrow/S-GAS_Manager (дата обращения: 10.01.2026).