УДК 004.4-004.9

ОПТИМИЗАЦИЯ КВАНТОВАНИЯ ЭМБЕДДИНГОВ В GOOGLE GEMINI

Сушкевич Е.П., Прихожий А.А.
Белорусский национальный технический университет
Минск, Беларусь

Современные робототехнические системы, системы искусственного интеллекта, системы анализа научных публикаций и многие другие требуют эффективных способов представления семантической информации. Векторные представления (эмбеддинги) стали стандартом для моделирования семантики в числовом пространстве, где близость векторов соответствует смысловой схожести текстов [1]. Высокая размерность эмбеддингов создает нагрузку на системы хранения и обработки информации. Методы сжатия, квантования данных помогают решить эту проблему. Семантические эмбеддинги формируются в процессе обучения нейросетевых моделей на больших массивах текстовых данных. Современные архитектуры, включая трансформеры, кодируют смысловые отношения в виде векторов фиксированной размерности. Передовые модели, подобные Google Gemini, используют технологию Matryoshka Representation Learning, позволяющую адаптировать размерность вектора в зависимости от решаемой задачи. Это обеспечивает баланс между точностью представления и вычислительной эффективностью. Предлагаемая методика квантования заключается в преобразовании 32-битных чисел с плавающей точкой в 8-битные целые числа: 1) исходные значения нормализуются к диапазону [-1, 1]; 2) полученные значения масштабируются и округляются до целочисленных значений в интервале [-127, 127]. Несмотря на потерю точности, семантическая информация сохраняется благодаря распределенному характеру представления данных в многомерном пространстве. Для подтверждения эффективности методики проведено преобразование эмбеддинга научной статьи из базы данных arXiv (идентификатор arXiv:0704.0001). Квантование сократило объем занимаемой памяти с 3072 байт до 768 байт. Качественная оценка результатов основывалась на предположении о сохранении относительных расстояний между векторами в семантическом пространстве. Теоретические выкладки подтверждают, что ошибка квантования не влияет существенно на результаты операций сравнения эмбеддингов, таких как вычисление косинусного сходства. Достигнутый результат — четырехкратное сокращение объема памяти для хранения эмбеддинга — имеет практическое значение для построения масштабируемых систем обработки больших данных.

1. Хобсон, Л., Ханнес Х., Коул Х. Обработка естественного языка в действии. - СПб.:Питер, 2021. -576 с.