Создание многозадачной унифицированной нейронной сети типа ERNIE 3 для анализа и генерации русскоязычных текстов

Екатерина Тотмина

НГУ

Современные большие языковые модели (LLM) такие, как GPT и T5, сталкиваются с трудностями в понимании контекста, интеграции знаний и обобщении на нестандартных данных. Они малоэффективны в многозадачных сценариях и при обработке текстов разной длины, а seq2seq-архитектуры часто испытывают проблемы с передачей градиента, что может приводить к расходимостям в обучении. Эти ограничения требуют решений, обеспечивающих устойчивость, универсальность и производительность в реальных задачах.

Для решения поставленных проблем в рамках проекта разрабатывается универсальная и робастная модель типа seq2seq на основе архитектуры трансформера (encoder-decoder), предназначенная для решения задач анализа (Natural Language Understanding, NLU) и генерации текста (Natural Language Generation, NLG) на русском языке. Ключевой особенностью модели является возможность её эффективного дообучения в условиях ограниченных данных (few-shot tuning), что позволяет адаптировать её под широкий спектр аналитических и генеративных задач без необходимости в больших объемах обучающей выборки.

Модель реализует стратегию иерархической многозадачности: энкодеры специализируются на задачах NLU, а единый декодер обслуживает задачи NLG. Это позволяет эффективно распределять ресурсы и оптимизировать инференс, обеспечивая высокую производительность даже при решении сложных задач, требующих глубокого анализа и генерации текстов. Применение инновационных методов минимизации инвариантного риска делает модель устойчивой к изменениям в данных, включая вариативность длины текста и его синтаксическую сложность.

Проектом предусмотрено создание открытой библиотеки, предоставляющей удобные инструменты для интеграции модели в существующие проекты. Библиотека будет включать адаптивные алгоритмы для эффективного few-shot тюнинга, а также поддерживает задачи, такие как распознавание именованных сущностей (NER), генерация парафраз, суммаризация текста и другие.

В результате предполагается значительное улучшение по сравнению с существующими подходами, включая генерализацию на нестандартных данных, ускорение инференса и снижение требований к обучающим данным. Ожидается, что проект окажет существенное влияние на развитие систем обработки русскоязычных текстов, включая аналитику социальных медиа, системы рекомендаций, автоматизированный перевод и генерацию контента.