**Дистилляция на уровне текстов: как создать маленький но эффективный энкодер предложений.**

**Малашенко Б.Т.**

(ИТМО, Санкт-Петербург)

Кодирование текстов остаётся одной из ключевых задач в области обработки естественного языка. Векторные представления текстов — эмбеддинги — широко используются для классификации, кластеризации, поиска, суммаризации и других задач. Недавние исследования показали, что удаление каузальной маски в больших языковых моделях (БЯМ) и их дообучение под задачу кодирования позволяет создавать энкодеры с миллиардами параметров. Однако для задач векторного поиска требуется высокая пропускная способность модели и низкая размерность латентного пространства, особенно при работе с большими базами данных, содержащими сотни миллиардов скрытых представлений.

Одним из решений является дистилляция на уровне текстовых представлений, при которой модель-ученик заимствует знания от более крупной модели-учителя. Это позволяет сохранить высокое качество при уменьшении размера модели. Возможно использование нескольких учителей для дистилляции в одну модель-ученика. Важно, чтобы в процессе дистилляции не нарушались представления модели-ученика, если она уже прошла фазу предобучения. Для этого перед дистилляцией необходимо спроецировать латентное пространство учителя на пространство ученика, минимизируя различия в начальной инициализации векторов.

Результаты экспериментов показали, что дистилляция на уровне текстовых представлений позволяет создать качественную и быструю модель. В качестве модели ученика использована архитектура, аналогичная bert\_uncased\_L-2\_H-128\_A-2, а модель rubert-tiny использована как бейзлайн. После дистилляции удалось превзойти бейзлайн по агрегации метрик на бенчмарке RuMTEB.