УДК 004.896

Ускорение инференса больших языковых моделей с помощью метода спекулятивного декодинга

Гарипов Р.И. (ИТМО)
Научный руководитель – кандидат технических наук, доцент Аксенов В.Е.
(ИТМО)

Введение. Современные методы спекулятивного декодинга [1] позволяют существенно ускорить авторегрессионную генерацию текста в больших языковых моделях за счет предсказания и валидации последовательностей с использованием вспомогательных моделей. Однако классические подходы ограничены строгими проверками корректности предсказанных токенов, что снижает потенциальный выигрыш в скорости. Спекулятивный декодинг с потерями ослабляет эти ограничения, позволяя добиться еще большего ускорения без значительной деградации качества генерации. В данной работе исследуется влияние этого подхода на задачи, требующие рассуждений, например [2] GSM8K, предлагается метод контроля ошибок при ослабленной проверке токенов, а также рассматриваются различные схемы ослабления на основе базового метода speculative decoding.

Основная часть. В рамках исследования была реализована имплементация speculative decoding и разработан пайплайн для проверки гипотез и экспериментов. Спекулятивный декодинг ускоряет генерацию, используя вспомогательную модель (draft model) для предсказания нескольких токенов, которые затем проверяются основной моделью (target model) в одном батче. Это позволяет за один вызов target model и несколько вызовов draft model получить несколько токенов, сокращая вычислительные затраты и эффективно утилизирую видео-карту.

Классический speculative decoding требует строгой проверки токенов, ограничивая выигрыш в скорости. Ослабление гарантий о сохранении распределения текста, позволяет достичь еще большего ускорения без значительной деградации качества. Эксперименты на таких задачах GSM8K подтверждают эффективность подхода, обеспечивая баланс между скоростью и точностью генерации.

Выводы. Speculative decoding с ослабленными гарантиями позволяет ускорить генерацию текста, минимизируя потери в качестве.

Список использованных источников:

1. Leviathan Y., Seidman A., Shazeer N., Parmar N., Piatetsky J. Fast Inference from Transformers via Speculative Decoding // arXiv preprint. – 2022. – URL: https://arxiv.org/abs/2211.17192 (дата обращения: 12.02.2025).
2. Cobbe K., Kosaraju V., Bavarian M., Chen M., Jun H., Kaiser L., Plappert M., Tworek J., Hilton J., Nakano R., Hesse C., Schulman J. Training Verifiers to Solve Math Word Problems // arXiv preprint. – 2021. – URL: https://arxiv.org/abs/2110.14168 (дата обращения: 12.02.2025).

Автор _____ Гарипов Р.И.

Научный руководитель _____ Аксенов В.Е.