УДК 004.8

Двухуровневая диффузия для генерации текста в дискретном латентном пространстве

Гаврилов А. О. (ИТМО)
Газзаев А.-Б. В. (ИТМО)
Научный руководитель – к.т.н. Муравьев Сергей Борисович (ИТМО)

Введение. Современные генеративные модели текста чаще всего авторегрессионны, что ограничивает параллелизм и усложняет контроль над глобальной структурой. Диффузионные подходы предлагают альтернативу через итеративное «очищение» шумных представлений, однако на токен-уровне они сталкиваются с длинными последовательностями и высокой вычислительной стоимостью. Мы предлагаем пайплайн двухуровневой диффузии: сначала генерация в сжатом дискретном латентном пространстве, затем уточнение на более детальном уровне, что позволяет сочетать глобальную согласованность и локальную точность без тяжёлой аналитики.

Основная часть. Пайплайн строится вокруг диффузионной модели и иерархического автоэнкодера, работающего на разных масштабах последовательности. Первый (верхний) уровень отвечает за грубую структуру текста в короткой последовательности латентных кодов, второй (нижний) уровень — за детальную реконструкцию и согласование с условием (промптом/контекстом) в более длинной последовательности [1],[2],[3].

Предлагаемый пайплайн включает следующие этапы:

- Дискретизация: токенизация текста и обучение (или использование) дискретного автоэнкодера (VQ-VAE), который сжимает последовательность в K раз и отображает её в последовательность кодов из кодбука(ов) [4].
- Верхний уровень (диффузия): обучение дискретной / маскированной диффузии, которая генерирует короткую последовательность латентных кодов, моделируя глобальную семантику и крупномасштабную структуру.
- Нижний уровень (первый уровень автоэнкодера): обучение автоэнкодера на более детальном представлении (например, на кодах нижнего кодбука), который восстанавливает детали, используя сэмпл верхнего уровня как условие.
- Декодирование: полученная и уточнённая латентная последовательность декодируется вторым уровнем автоэнкодера обратно в текст; при необходимости добавляется пост-обработка (детокенизация, фильтрация, ограничение по длине).

Выводы. Двухуровневая диффузия позволяет разделить задачу генерации на «глобальный план» и «локальное уточнение». Это даёт возможность ускорить моделирование длинных текстов за счёт работы в сжатом пространстве, а затем повысить качество за счёт отдельного шага детализации, сохранив модульность пайплайна (второй уровень автоэнкодера + верхний уровень автоэнкодера + diffusion).

Список использованных источников.

[1] Sohl-Dickstein J. et al. Deep Unsupervised Learning using Nonequilibrium Thermodynamics // International Conference on Machine Learning (ICML). – 2015.

[2] Ho J., Jain A., Abbeel P. Denoising Diffusion Probabilistic Models // Advances in Neural Information Processing Systems (NeurIPS). – 2020.

[3] Austin J. et al. Structured Denoising Diffusion Models in Discrete State-Spaces // Advances in Neural Information Processing Systems (NeurIPS). – 2021.

[4] van den Oord A., Vinyals O., Kavukcuoglu K. Neural Discrete Representation Learning // Advances in Neural Information Processing Systems (NeurIPS). – 2017.

Гаврилов А. О. (автор) _____

Газзаев А.-Б. В. (автор) _____

Муравьев С. Б. (научный руководитель) _____