УДК 004.855.5

Интеграция количественных и пространственных признаков в латентное пространство предобученных визуально-лексических нейронных сетей

Дьячков Д.А. (Университет ИТМО)

Устинова В.Е. (Университет ИТМО)

Научный руководитель – к.т.н. Ефимова В.А. (Университет ИТМО)

Введение. Современный мир уже сложно представить без генеративного ИИ и диффузионных моделей, но помимо достоинств у диффузионных моделей также присутствуют недостатки. Так, проблема генерации заданного большого количества объектов на изображении по-прежнему не решена. Исследования показывают [1], что значительная часть этой проблемы обусловлена кодировщиком, преобразующим текст в визуально-лингвистическое латентное пространство. Это вызвано недостаточным пониманием кодировщиком чисел и их взаимосвязи с количеством объектов в визуальном представлении латентного пространства. Причиной этого является недостаток данных о количестве объектов в наборе, на котором проводилось обучение. В данной статье предлагается метод решения этой проблемы, основанный на дообучении визуально-лингвистических кодировщиков на синтетическом наборе данных, содержащем геометрические примитивы.

Основная часть. Синтетический датасет, состоящий из различных изображений с геометрическими примитивами и аннотаций к этим изображениям, сильно отличается от данных, которые визуально-лингвистические модели видели при обучении. Это является проблемой не только само по себе, но и в сценариях, когда визуально-лингвистические модели используются в качестве одного из элементов многостадийного процесса, как это обычно делается для задачи условной генерации в моделях диффузии. Всё это ведёт к существенному дрейфу данных в результате дообучения на таком синтетическом датасете. Классические методы, такие как изменение стиля синтетических данных [2] не подойдут, ведь изначальная стилистика набора данных была выбрана специально и несёт в себе цель упростить задачу дообучения. Более того, необходимо, чтобы дрейф данных после дообучения визуально-лингвистической модели стремился к нулю, ведь целью дообучения на синтетическом датасете является не запоминании моделью новых типов объектов, а выявление пространственной и количественной информации об уже известных визуально-лингвистической модели объектах, таких как связь количества с нужным типом объекта и пространственного положения конкретного объекта.

Чтобы добиться такого эффекта, было принято решение добавить к визуально-лингвистической сети адаптеры LORA [3], показавшие качественные результаты при дообучении больших языковых моделей. Это приведёт не только к уменьшению вычислительных затрат на дообучение визуально-лингвистических моделей по сравнению с полным дообучением (за счёт снижения количества обучаемых параметров), но и позволит избежать дрейфа данных, ведь основные веса визуально-лингвистической модели будут заморожены в процессе дообучения.

Чтобы обогатить визуально-лингвистическое латентное пространство новыми признаками, отвечающими исключительно за количество и пространственное расположение объектов, было решено применить подход, аналогичный применённому в SpatialRGPT фьюжн-модулю [4], заключающемуся в конкатенации предыдущего латентного пространства, замороженного при дообучении, с новыми признаками, добавленными для дообучения модели и предназначенными для хранения количественных и пространственных свойств объектов (признаки самих объектов хранятся в ванильном визуально-лингвистическом латентном пространстве, не подвергающемуся процессу дообучения).

Для того, чтобы добавленные признаки не содержали свойства, которые описывают непосредственно объект, а не количество или пространственное положение, в функцию потерь было решено добавить компонент, отвечающий за разницу между двумя текстовыми аннотациями, в которых различаются типы объектов, но пространственные и количественные свойства сохраняются (пример таких аннотаций: «в правом углу красный квадрат» и «в правом углу синий круг»). Для каждой такой аннотации вычисляется визуально-лингвистическое латентное представление, после чего для добавленных в процессе дообучения частей латентного пространства вычисляется среднеквадратичное отклонение и с коэффициентом (который является гиперпараметром при дообучении) прибавляется к общей функции потерь.

Выводы. Предложенный метод позволяет эффективно дообучать визуально-лингвистические модели нейронных сетей на синтетических данных, которые существенно отличаются от реальных и имеют большой сдвиг распределений. Дообученная предложенным образом модель, осуществляющая преобразование текста в визуально-лингвистическое латентное пространство, имеет не только более качественное понимание композиционных свойств текстового запроса, но и способна улучшить качество условной генерации диффузионных моделей с точки зрения количественного и пространственного соответствия объектов на сгенерированном изображении текстовому запросу. Это улучшение позволит расширить область применения диффузионных моделей в сфере промышленного графического дизайна и позволит вывести качество условной генерации на уровень профессионального дизайнера не только в области визуальной составляющей, но и соответствия результатов генерации реальным запросам клиентов по наполнению изображения.

Список использованных источников:

[1] Paiss R. et al. Teaching clip to count to ten //Proceedings of the IEEE/CVF International Conference on Computer Vision. – 2023. – С. 3170-3180.

[2] Mullick K. et al. Domain adaptation of synthetic driving datasets for real-world autonomous driving //arXiv preprint arXiv:2302.04149. – 2023. "],

[3] Hu E. J. et al. Lora: Low-rank adaptation of large language models //ICLR. – 2022. – Т. 1. – №. 2. – С. 3.

[4] Cheng A. C. et al. SpatialRGPT: Grounded Spatial Reasoning in Vision Language Models //arXiv preprint arXiv:2406.01584. – 2024.