УДК 004.93

ЭВОЛЮЦИЯ TEXT-TO-VIDEO ГЕНЕРАЦИИ И РЕДАКТИРОВАНИЯ

Детков Н.С. (Университет ИТМО)

Научный руководитель – кандидат физико-математических наук, Фильченков А.А. (Университет ИТМО)

Введение. Text-to-Video генерация — это сложная задача, которая заключается в создании высококачественных и реалистичных видео с произвольными объектами, стилями и сценариями на основе текстовых описаний. Смежной к ней является задача Text-to-Video редактивания, то есть редактирования входного видео по произвольному текстовому запросу. Последние достижения в области глубокого обучения и генеративных моделей, в особенности, диффузионных нейронных сетей, привели к значительному прогрессу в этой области. Цель данного доклада - изучить современные методы и нейронные сети, используемые для Text-to-Video генерации и редактирования, а также обсудить их ограничения и будущие направления. Мы рассмотрим недавние работы и методы, чтобы изучить достижения и проблемы в генерации видео из текста, а также проанализируем их и предложим направления для улучшения.

Основная часть. Задача Text-to-Video генерации заключается в создании высококачественных и реалистичных видео с произвольными объектами, стилями и сценариями на основе текстовых описаний. Эта задача особенно сложна, поскольку требует от модели понимания семантики входного текста и синтеза реалистичных видео, соответствующих описанию. Столь же сложна, и имеет дополнительные особенности, задача редактирования видео по текстовому описанию.

На данный момент, Text-to-Video генеративные модели, могут использовать текст лишь как часть для условной генерации [1], также уславливаясь на первый кадр для задачи предсказания видео, или уславливаясь на другое видео для более строгого следования генерации. Более того, можно генерировать долгие видео, используя последовательность текстовых запросов [2]. С другой стороны, редактирование также имеет следующие под-задачи, такие как стилизация [3], изменение объекта интереса, редактирование по маске и редактирование с текстом-нструкцией.

Множество проблем, относящиеся к более высокому качеству, более высокому разрешению, улучшению разнообразия, лучшему сочетанию соотношению текста к результату, остаются на повестке, и мы скрупулёзно показываем их, и предлагаем решения.

Выводы. Менее чем за 2 года, начиная с генерации и редактирования видео из текста по кадрам, исследователи добились значительного прогресса. До сих пор эта тема является очень активной областью исследований, в которой есть нерешённые проблемы, подсвеченные нами с предложениями по их решению.

Список использованных источников:

1. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., and Shi, H. Text2video-zero: Text-to-image diffusion models are zero-shot video generators

2. Villegas, R., Babaeizadeh, M., Kindermans, P.-J., Moraldo, H., Zhang, H., Saffar, M. T., Castro, S., Kunze, J., and Erhan, D. Phenaki: Variable length video generation from open domain textual description, 2022

3. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., and Germanidis, A. Structure and content-guided video synthesis with diffusion models, 2023

Детков Н.С. (автор) Подпись: _____

Фильченков А.А. (научный руководитель) Подпись: _____