УДК 004.92

ДОСЛІДЖЕННЯ ГЕНЕРАТИВНО-ЗМАГАЛЬНИХ МОДЕЛЕЙ ДЛЯ СИНТЕЗУ ТЕКСТУ В ЗОБРАЖЕННЯ

Воронюк К. Л.
Науковий керівник – проф. каф. ШІ Рябова Н. В.
Харківський національний університет радіоелектроніки, каф. ШІ
м. Харків, Україна

This work is devoted to the study of the architecture and work of GAN networks for text-to-image synthesis. The paper highlights some common problems that a developer may encounter when using vanilla GAN. Ways to solve these problems in the form of GAN models for generating high quality images have been presented. We critically examine current strategies for evaluating text-to-image synthesis models, identify the drawbacks and merits of each model. As a result, the best solution in the form of using a suitable model for high-quality image generation from text descriptions is selected.

Завдяки останнім досягненням у науці, машини здатні малювати оригінальні картини в певному стилі, писати абзаци зв'язного тексту, розробляти виграшні стратегії для складних ігор та інше. І це лише початок генеративної революції та здібності генеративно-змагальних мереж.

Генеративно-змагальні мережі (GAN) – це тип алгоритму глибокого навчання, який набув величезної популярності в останні роки завдяки своїй здатності генерувати високореалістичні та різноманітні синтетичні дані [1]. GAN складаються з двох нейронних мереж – генератора та дискримінатора. Мережа-генератора генерує синтетичні вибірки даних, а мережа-дискримінатора оцінює справжність як реальних, і синтетичних вибірок даних. Завдяки цьому змагальному процесу мережа-генератор навчається створювати все більш реалістичні синтетичні дані, тоді як мережа-дискримінатор краще розрізняє справжні та підроблені дані.

GANs вирішують ряд завдань, пов'язаних із генерацією нових даних. Основні задачі, які вирішує GAN, включають – генерація відео та тексту, підвищення якості зображень, поєднання зображень, пошук прихованих зв'язків, а також синтез зображень.

Розповсюдженою і актуальною із цих задач є саме область синтезу зображень, яка є процесом створення нових зображень на основі текстових описів. Синтез тексту у зображення може мати безліч застосувань, включаючи створення зображень для ігор, автоматичного створення ілюстрацій і багато іншого. Однак, як і будь-яка інша технологія, вона також може мати свої обмеження, проблеми та потенційні ризики, які необхідно враховувати.

Проблеми GAN при синтезі тексту у зображення пов'язані з тим, що завдання генерації зображень з текстового опису є досить складним і

вимагає врахування багатьох факторів. Однією з основних проблем є необхідність навчання моделей на великій кількості даних, щоб вони могли коректно відтворювати елементи зображення. Іншою проблемою є складність визначення метрик для оцінки якості зображень, вони можуть бути неефективними для вимірювання якості зображень. Також виникають проблеми з урахуванням контексту, особливо коли текстові описи містять складні зв'язки та відносини між об'єктами [2].

Для вирішення основних проблем (врахування складних зв'язків між об'єктами) у моделях GAN пропонується використання більш складних моделей. Існує кілька моделей GAN для синтезу тексту зображення.

Таким чином, головною метою роботи було проведення порівняльного аналізу таких моделей як AttnGAN, StackGAN та DALL-E. Кожна модель мала свої переваги та недоліки, і вибір конкретної моделі залежав від конкретного завдання та доступних ресурсів.

AttnGAN – це модель GAN, яка використовує механізм уваги для синтезу зображень на основі текстового опису. В результаті дослідження виявилось, що AttnGAN може створювати зображення з високою якістю та деталізацією. Однак у AttnGAN є недолік у тому, що вона може мати проблеми з описом зображень, які не відповідають наданим текстовим описам.

StackGAN – це модель, яка використовує два ступені (stages) для генерації зображень. У першому ступені StackGAN згенерувала зображення низької роздільної здатності, а потім у другому ступені воно покращилось до більш високої роздільної здатності. Одним із недоліків StackGAN виявилось те, що її навчання може зайняти велику кількість часу.

Модель DALL-E також використовується для синтезу зображень на основі текстового опису, але на відміну від інших моделей GAN вона показала високі результати, працюючи зі складними текстовими описами, а також може створювати зображення на основі декількох предметів одночасно. Але недоліком DALL-E є те, що вона потребує великої кількості обчислювальних ресурсів та часу для навчання.

Доцільним слід вважати, що застосування цих моделей потребує великих обчислювальних ресурсів та вибір моделі залежить від специфічних потреб і можливостей.

Список використаних джерел:

1. A. Kailash, Generative Adversarial Networks. Packt Publishing, 2019.
2. Adversarial text-to-image synthesis: A review. URL: https://www.sciencedirect.com/science/article/pii/S0893608021002823