УДК 004.89:004.91

# СТВОРЕННЯ ТА КОРЕКЦІЯ ЗОБРАЖЕННЯ ЗА ДОПОМОГОЮ ТЕКСТОВИХ ПІДКАЗОК У ДІАЛОЗІ

Левицький К.Ю., Терзіян В.Я.
e-mail: kyrylo.levytskyi@nure.ua, vagan.terziyan@nure.ua
Харківський національний університет радіоелектроніки, каф. ШІ
м. Харків, Україна

Generative AI models for image creation use text prompts by encoding them into numerical representations with transformers and applying diffusion models to refine noisy images. The process involves multiple convolution operations to gradually enhance image quality. These models understand not only keywords but also contextual relationships, allowing for complex scene generation. To enhance generated images, high-resolution upscaling methods reduce blurriness and ensure stylistic consistency. Further research is needed to improve precision, reduce variability, and optimize user control over AI-generated visuals.

З кожним роком системи штучного інтелекту роблять величезні кроки у своєму розвитку. Якщо, донедавна, були системи, які могли аналізувати простий текст і на його основі робити висновки, то сучасні системи дозволяють створювати зображення, або відео з нічого на основі текстових інструкцій та підказок через діалог з користувачем [1].

Створення зображень за допомогою штучного інтелекту є одним із найперспективніших напрямів сучасних генеративних технологій. Використовуючи потужні нейронні мережі, такі як дифузійні моделі, GAN або трансформери, ШІ здатен перетворювати текстові описи на реалістичні чи стилізовані візуальні образи.

Створення зображень за текстовими підказками ґрунтується на генеративних моделях, які поєднують семантичне розуміння тексту та можливість візуалізації відповідних об'єктів, сцен і стилів. Модель спочатку кодує текст у числове представлення за допомогою трансформерів, після чого це представлення використовується для генерації або зміни зображення. Основним підходом є дифузійні моделі, які поступово очищують випадковий шум, створюючи зображення, що відповідає опису.

В основі генерації зображень лежить модель, яка починається з зашумленого зображення – набору матриць випадкових чисел. Ці матриці розбиваються на менші підматриці, до яких застосовується послідовність згорток (математичних операцій), що дає очищений, менш зашумлений результат. Кожна згортка передбачає операцію множення та накопичення. Цей процес згладжування повторюється кілька разів, доки не буде отримано нове, покращене кінцеве зображення. На рисунку 1 зображено процес створення зображення [2].

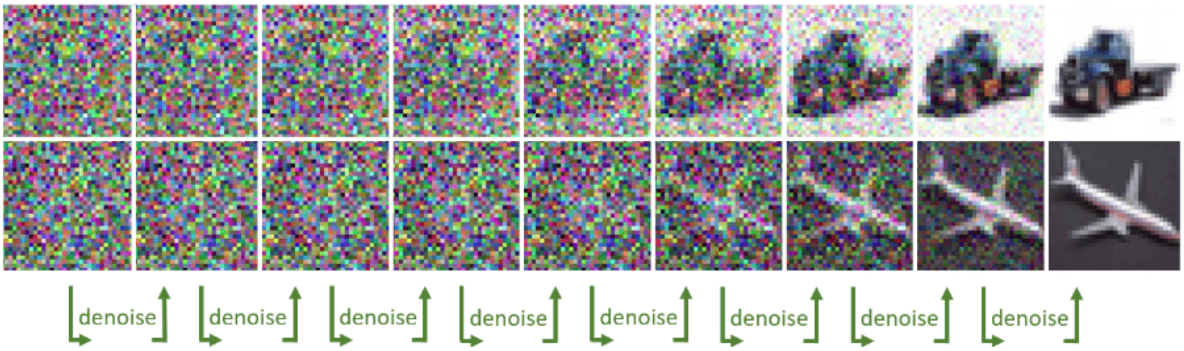

Рисунок 1 – Процес створення зображення

Щоб модель зрозуміла текстовий запит, вона спочатку перетворює слова у числові вектори за допомогою мовних моделей. Ці вектори потім спрямовуються в генератор, який відповідає за побудову зображення. Завдяки цьому штучний інтелект може враховувати не тільки окремі ключові слова, а й взаємозв'язки між ними, розуміючи контекст і деталі сцени. Наприклад, запит «схід сонця над горами в стилі імпресіонізму» спонукатиме модель до створення пейзажу з характерною розмитістю мазків та ніжними кольорами [1].

Одна з головних переваг створення зображень ШІ – можливість отримати нові, унікальні композиції, які можуть бути складними або навіть неможливими для традиційного малювання. Однак є й труднощі, зокрема обмежена передбачуваність результату. Наприклад, одна й та ж текстова підказка може щоразу давати трохи різні варіанти зображень. Це пов'язано з тим, що генерація містить елемент випадковості, який робить результат варіативним.

Ще одним важливим аспектом є контроль над деталями. Якщо потрібно створити точний образ, наприклад, персонажа з конкретними рисами обличчя, штучний інтелект може давати неточні результати, оскільки йому важко запам'ятати та відтворювати специфічні характеристики без додаткових вказівок. Для цього можна використовувати інструменти на кшталт керованої генерації або комбінувати різні підходи, наприклад, поєднуючи автоматичну генерацію з ручним редагуванням.

Корекція через текстові інструкції є складнішою задачею, ніж первинна генерація, оскільки вимагає збереження вихідної структури зображення та внесення цільових змін. Тут важливою є можливість роботи з масками, що дозволяє редагувати окремі частини, а також механізм негативних підказок, який допомагає уникати небажаних елементів. Наприклад, фраза «без різких тіней» підштовхує модель до згладжування контрастних переходів [3].

Основні труднощі полягають у тому, що текстові інструкції можуть бути розмитими або неоднозначними, а очікування користувача – надто суб'єктивними. Щоб покращити точність, можуть застосовуватися

проміжні етапи перегляду та уточнення результату, де користувач бачить частковий результат і може вносити коригування в режимі реального часу. Також можна використовувати додаткові контролюючі механізми, наприклад, управління кольорами, композицією або навіть деталізацією текстур. У майбутньому розвиток таких моделей може привести до ще більш точного та інтуїтивного керування процесом генерації через текст.

Для вирішення задачі генерації зображень необхідно, щоб користувач спочатку описує бажане зображення, а система перетворює цей опис на вхідний запит для моделі генерації. Після першого рендеру зображення користувач може вносити коригування за допомогою додаткових текстових команд, які система інтерпретує та застосовує через редагування існуючого зображення або регенерацію окремих його частин [4].

Для виконання задачі внесення змін до створеного зображення можна використовувати методи, такі як inpainting (заповнення вибраної області) або стилізація (адаптація кольорів, текстур, освітлення). Система аналізує текстові вказівки та адаптує параметри моделі, змінюючи деталі зображення без необхідності повної регенерації.

Після проведення процесу генерації і редагування зображення необхідно мати високоякісне зображення. Для того, щоб отримати чіткіше та більш деталізоване зображення, можна застосувати технологію покращення роздільної здатності. Ця технологія дозволяє отримати зображення якомога більш якісним, зменшуючи розмитість та дотримання єдиного стилю зображення.

Список використаних джерел:

1. Terziyan V. Deep Learning for Cognitive Computing. University of Jyväskylä. 1089 p. URL: https://ai.it.jyu.fi/vagan (дата звернення: 04.03.2025);

2. Terziyan, V., & Vitko, O. (2023). Causality-Aware Convolutional Neural Networks for Advanced Image Classification and Generation. Procedia Computer Science, 217, 495-506. Elsevier. https://doi.org/10.1016/j.procs.2022.12.245.

3. Generate Stunning Images with Stable Diffusion XL on the NVIDIA AI Inference Platform | NVIDIA Technical Blog. NVIDIA Technical Blog. URL: https://developer.nvidia.com/blog/generate-stunning-images-with-stable-diffusion-xl-on-the-nvidia-ai-inference-platform/ (дата звернення: 04.03.2025).

4. Gupta M. How does DALL-E, the text-to-image generator work?. Medium. URL: https://medium.com/data-science-in-your-pocket/how-does-dall-e-the-text-to-image-generator-work-c2d9f4a0f26c (дата звернення: 04.03.2025).

5. The Comprehensive Guide to Text-to-Image Models. Everypixel Journal – Your Guide to the Entangled World of AI. URL: https://journal.everypixel.com/guide-to-text-to-image-models (дата звернення: 04.03.2025).