УДК 004.89

ОГЛЯД ЛАТЕНТНИХ ДИФУЗІЙНИХ МОДЕЛЕЙ

Овсієнко В.С.
Науковий керівник – доц., к.т.н., доц.каф. ШІ Дейнеко А.О.
Харківський національний університет радіоелектроніки, каф. ШІ
м. Харків, Україна
тел.: +38(066) 192-60-59, e-mail: vladyslava.ovsiienko@nure.ua

In recent years, algorithms for solving a wide variety of generative modeling tasks, such as image generation, super-resolution, inpainting, editing, have evolved rapidly. Diffusion models raising the bar to a new level in these fields. Latent diffusion model is an extension of the classical diffusion model and takes into account the underlying structure and dependencies in an image to generate more realistic and efficient image samples with less training and less time to inference. In this work, the concept and architecture of these models were reviewed and described.

Дифузійні моделі належать до категорії глибоких генеративних моделей, які останнім часом досягли вражаючих генеративних можливостей, починаючи від високого рівня деталізації до різноманітності згенерованих прикладів. Їх можна умовно класифікувати залежно від того, де проводиться попередня дифузія – в піксельному просторі або в латентному просторі. Перший клас методів генерує зображення безпосередньо з високо розмірного піксельного рівня, наприклад, GLIDE та Imagen. Інший клас є латентними дифузійними моделями (LDM) до яких належать Stable Diffusion, VQ-дифузія та DALL-E 2 [1].

В моделі латентної дифузії як і в звичайній дифузійній моделі є процес прямої дифузії та зворотній процес (реконструкція), але процес дифузії запускається в латентному просторі, а не в піксельному, що робить вартість навчання нижчою, а швидкість виведення швидшою. Це мотивовано спостереженням, що більшість бітів зображення допомагають сприйняттю деталей і семантична, концептуальна композиція все ще залишається після сильного стиснення. LDM вільно декомпозує перцептивне і семантичне стиснення за допомогою генеративного моделювання, спочатку відсікаючи надлишковість на рівні пікселів за допомогою автокодувальника, а потім маніпулюючи/генеруючи семантичні концепції за допомогою процесу дифузії на вивчених латентних даних [2].

Процес перцептивної компресії ґрунтується на моделі автокодувальника. Кодувальник ε використовується для стиснення вхідного зображення $x \in R^{H \times W \times 3}$ до меншого двовимірного латентного вектора $z = \varepsilon(x) \in R^{h \times w \times c}$, де частота дискретизації

$f = \frac{H}{h} = \frac{W}{w} = 2^m, m \in \mathbb{N}$. Потім декодувальник D відновлює зображення з латентного вектора $\tilde{x} = D(z)$ (рисунок 1).

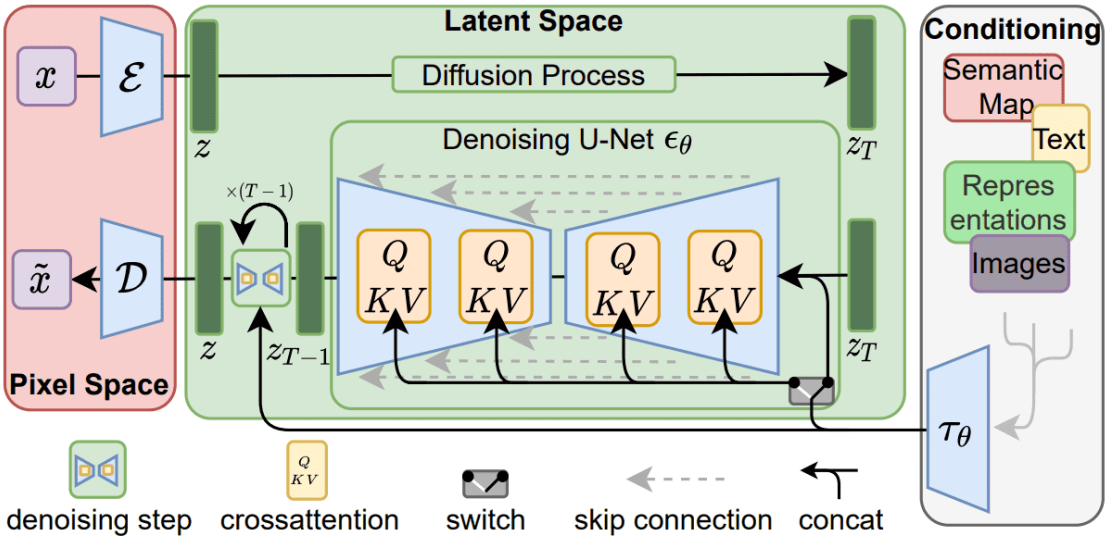

Рисунок 1 – Архітектура латентної дифузійної моделі

Процеси дифузії та знешумлення відбуваються на латентному векторі z. Модель знешумлення – це обумовлена в часі U-Net, доповнена механізмом перехресної уваги для обробки гнучкої умовної інформації для генерації зображень (наприклад, мітки класів, семантичні карти, розмиті варіанти зображень). Структура еквівалентна об'єднанню представлення різних модальностей у модель з механізмом перехресної уваги. Кожен тип умовної інформації пов’язаний з специфічним для домену кодувальником τ_θ для проектування вхідної інформації y на проміжну репрезентацію, яку можна зіставити з компонентом перехресної уваги, $\tau_\theta(y) \in R^{M \times d_\tau}$:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V,$$

де $Q = W_Q^{(i)} \cdot \varphi(z_i)$, $K = W_K^{(i)} \cdot \tau_\theta(y)$, $V = W_V^{(i)} \cdot \tau_\theta(y)$, $W_Q^{(i)} \in R^{d \times d_\varepsilon^i}$, $W_K^{(i)}, W_V^{(i)} \in R^{d \times d_\tau}$, $\varphi(z_i) \in R^{N \times d_\varepsilon^i}$, $\tau_\theta(y) \in R^{M \times d_\tau}$ [3].

Список використаних джерел:

1. Croitoru, F.-A., Hondru, V., Ionescu, R. T., & Shah, M. (2023). Diffusion Models in Vision: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–20.

2. Zhang, C., Zhang, C., Zhang, M., & Kweon, I. S. (2015). Text-to-image Diffusion Model in Generative AI: A Survey. JOURNAL OF LATEX CLASS FILES, 14(8).

3. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. У 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).