КІБЕРНЕТИКА та КОМП'ЮТЕРНІ ТЕХНОЛОГІЇ

УДК 519.688 DOI:10.34229/2707-451X.24.4.9

П.П. МАСЛЯНКО, К.І. ПАВЛОВСЬКА

КОНЦЕПТУАЛЬНА МОДЕЛЬ ТА NLP-СИСТЕМА "TEXT TO IMAGE"

Запропоновано концептуальну модель та NLP-систему "Text to image" на основі методології системної інженерії систем Data Science, архітектуру та програмне забезпечення системи генерації зображень на основі латентної дифузійної моделі. Запропоновано покращення базової архітектури латентної дифузійної моделі шляхом використання дифузійного трансформера. Побудова динамічного дерева розгалужень та нелінійних оцінок дозволяє прискорити процес пошуку оптимального розв'язку, але суттєво залежить від початкової задачі, що ускладнює розробку загального алгоритму. Розробка концептуальної моделі, та NLP-системи "Text to image" дозволяє реалізувати ефективне перетворення текстових даних у зображення, що є актуальним питанням у сфері візуалізації даних.

Ключові слова: *системна інженерія, Data Science, NLP-системи "Text to image.*

Вступ. Коли люди сприймають текстову і аудіо інформацію, або самі є джерелом такої інформації, то підсвідомо у них виникають уявні образи, що дає можливість краще зрозуміти й візуалізувати зміст отриманої інформації. Такий процес візуального уявлення відіграє вирішальну роль у різних когнітивних функціях, таких як запам'ятовування, міркування та критичне мислення. Оскільки технології продовжують розвиватися, зростає інтерес до розробки науково обґрунтованих підходів, систем та прикладних застосунків, які моделюють зв'язок між текстовою і аудіо інформацію та візуальним сприйняттям і розумінням такої інформації. Розробка таких систем та прикладних застосунків стала можливим завдяки швидкому прогресу в техніці обробки зображень і застосуванні комп'ютерного зору, якому сприяв успіх штучного інтелекту та глибокого навчання. Сферою, яка виникла в результаті цих досягнень і є, власне, генерація тексту в зображення. У випадку генерації зображення-картинки з текстових даних модель приймає написаний людиною опис як вхідні дані та створює повнокольорове зображення, яке відповідає заданому тексту. Для забезпечення ступеня відповідності між текстом і зображенням система/застосунок має враховувати не тільки синтаксичні та семантичні аспекти тексту, а й контекстуальні й прагматичні нюанси. Наприклад, у текстовому описі можуть міститися як явні, так і приховані значення, що потребують моделювання багаторівневої семантики.

Концептуальна модель та система "Text to image" базується на методології системної інженерії систем Data Science на основі бізнес-профіля Еріксона – Пенкера.

1. Постановка задачі

Об'єкт дослідження – теоретичні інструменти та інструментальні засоби трансформації текстової інформації у зображення.

Предмет дослідження – концептуальна модель, інструменти Data Science та застосунок трансформації текстової інформації у зображення.

Мета роботи – це розробка концептуальної моделі та системи для перетворення текстових описів у зображення, що базується на методології системної інженерії, сучасних методах глибинного навчання та бізнес профілі Еріксона – Пенкера.

Кінцевий результат – наукове обґрунтування концептуальної моделі, модель системи для перетворення текстових описів у зображення, імплементація системи, верифікація і валідація системи.

2. Огляд існуючих рішень

Розглянемо та проаналізуємо, насамперед, ті наявні наукові та інженерні рішення, що вже частково або повністю вирішують поставлену задачу.

Мета роботи – це розробка системи для перетворення текстових описів у зображення, що базується на сучасних методах глибинного навчання.

До найпоширеніших методів машинного навчання для перетворення текстових даних в зображення можна віднести генеративно-змагальні мережі та дифузійні моделі [1]. Генеративна змагальна мережа (GAN) – це система, яка використовує два ключові компоненти: генератор і дискримінатор для виконання свого завдання зі створення нових екземплярів даних. Генератор відповідає за створення синтетичних даних, які дуже схожі на навчальні приклади, тоді як дискримінатор визначає, чи є надісланий приклад реальним, чи згенерований генератором [2]. Важливо зазначити, що GAN використовують дискримінаційні ознаки, але їхній підхід відрізняється від типових класифікаторів. На відміну від дискримінаційних моделей, які лише ідентифікують клас, до якого належить об’єкт, GAN використовують вхідні об’єкти для створення нових екземплярів даних [3, 4]. Вони поєднують завдання генерування нових даних з класифікацією, що забезпечується дискримінаційними моделями. Генеративна змагальна мережа, що показано на рис. 1, має структуру, що складається з двох ключових компонентів: генератора G і дискримінатора D. Як правило, ці компоненти реалізуються за допомогою нейронних мереж, наприклад, згорткових нейронних. Метод опорних векторів також може служити альтернативним дискримінатором [5].

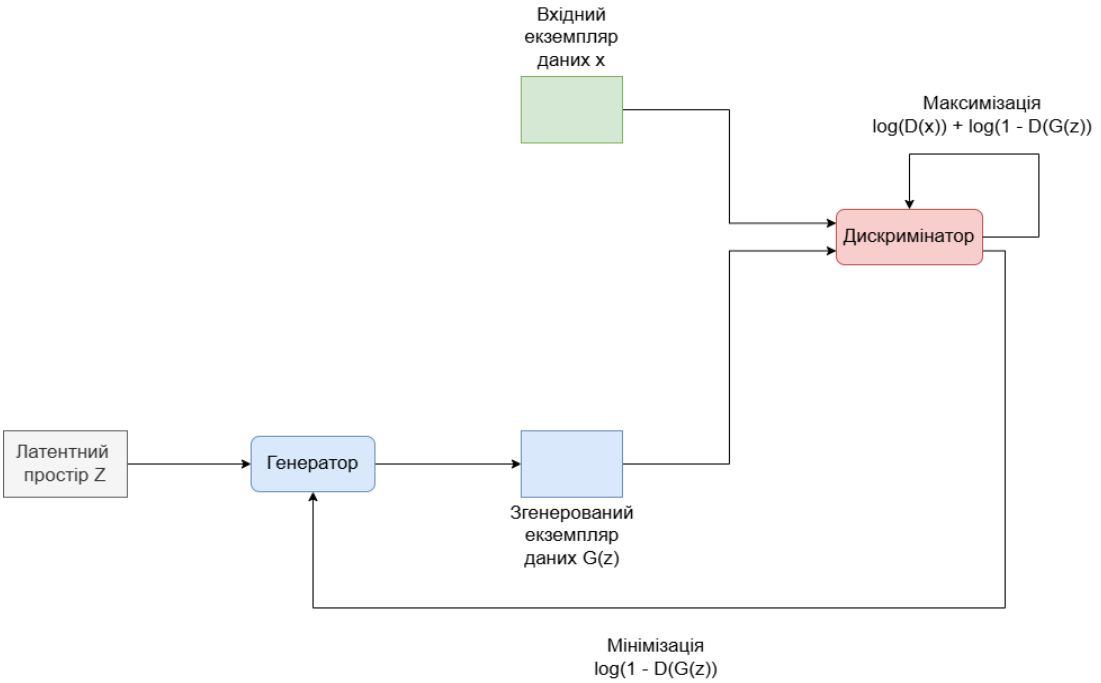


РИС. 1. Типова архітектура генеративної змагальної мережі

Щоб надати більш формальне пояснення проблеми, можна розглядати генератор як відображення, яке приймає вхідні дані з латентного простору Z і перетворює їх на елементи реального простору $\mathbb{R}^n$. Зокрема, представимо генератор як G: $Z \to \mathbb{R}^n$, тоді як дискримінатор відображає елементи з реального простору $\mathbb{R}^n$ на інтервал (0,1), позначений як D: $\mathbb{R}^n \to (0,1)$. Основне завдання дискримінатора – це присвоєння значення кожному зразку на основі його сприйняття того, чи є вхідний екземпляр реальним чи згенерованим [6]. Значення 1 призначається для позначення реального екземпляра, тоді як значення 0 призначається для позначення згенерованого екземпляра. Ці два виходи, оцінки дискримінатора реальних або згенерованих екземплярів, потім використовуються для оцінки продуктивності відповідних моделей [7]. Генератор навчений мінімізувати функцію $\log(1-D(G(z)))$ яка стимулює його створювати зображення, які дискримінатор не може легко ідентифікувати як згенеровані, прагнучи до виходу дискримінатора, близького до 1 (тобто $D(G(z)) \approx 1$). З іншого боку, дискримінатор навчений максимізувати функцію $\log(D(x)) + \log(1-D(G(z)))$ з метою точної ідентифікації екземплярів $\mathrm{D(x)}$ і розпізнавання екземплярів $\mathrm{D(G(z))}$.

Дифузійні моделі можна класифікувати як моделі латентних змінних, вказуючи на те, що вони характеризуються базовим і прихованим безперервним простором ознак. Ця якість нагадує про схожість, яку вони мають із варіаційними автокодувальниками. На практиці ці моделі формулюються з використанням ланцюга Маркова, що складається з T кроків [8]. Враховуючи точку даних x_0, відібрану з розподілу даних $q(x)$, можна встановити процес прямої дифузії шляхом введення шуму.

Цей процес передбачає додавання шуму Гаусса з дисперсією β_t до попередньої латентної змінної x_{t-1} на кожному кроці ланцюга Маркова. Отже, генерується нова прихована змінна x_t, що відповідає розподілу $q(x_t \mid x_{t-1})$ [9]:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \mu_t = \sqrt{1-\beta_t}\, x_{t-1}, \Sigma_t = \beta_t I\right) x_0,$$

де β_t – вказує на кожному кроці компроміс між інформацією, яку потрібно зберегти з попереднього кроку, та шумом, який необхідно додати. У контексті багатовимірного сценарію важливо зазначити, що матриця ідентичності I представляє кожен вимір, що має однакове стандартне відхилення, позначене β_t. Розподіл $q(x_t \mid x_{t-1})$ все ще є нормальним розподілом, який характеризується середнім μ та дисперсією, де $\mu_t = \sqrt{1-\beta_t}\, x_{t-1}$, а дисперсія = $\mathrm{B}_t I$ [10]. Це дозволяє ефективно обчислювати перехід від початкових даних до кінцевих даних x_T у закритій формі.

Якщо визначимо α_t та $\bar{\alpha}_t$, що відповідають нормальному розподілу $\epsilon_0, \ldots, \epsilon_{t-2}, \epsilon_{t-1} \sim \mathcal{N}(0, I)$, можна використати підхід перепараметризації рекурсивним способом для доведення наступного [11]:

$$x_t = \sqrt{1-\beta_t}\, x_{t-1} + \sqrt{\beta_t}\, \epsilon_{t-1} = \sqrt{\alpha_{t-2}}\, x_{t-2} + \sqrt{1-\alpha_t}\, \epsilon_{t-2} = \ldots = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon_0 .$$

Розглянемо існуючі моделі перетворення текстових даних в зображення, а саме DALL-E, Stable Diffusion та Midjourney (табл. 1).

Таблиця 1. Порівняльний аналіз систем T2IMG

Характеристика	DALL-E	Stable Diffusion	Midjourney
Основний принцип	Використовує трансформерні моделі для генерації зображень з текстових описів	Diffusion model, що генерує зображення з текстових описів	Генерування зображень з текстових запитів, використовуючи власні алгоритми штучного інтелекту. Специфічні деталі технології не повністю розкриті
Якість зображення	Висока, з високою деталізацією та вірністю до	Висока, з гнучкістю у стилях та деталізації	Висока, з унікальним стилем та артистичним виразом
Контроль зображення	Дозволяє точне управління зображенням через деталізовані текстові запити	Модифікований, дозволяє користувачу впливати на процес генерації	Дозволяючи детально налаштовувати запити, хоча іноді результати можуть бути менш передбачувані
Ліміти використання	Обмежений доступ через API, з обмеженнями на комерційне використання	Відкритий доступ, що дозволяє широке комерційне та особисте використання	На момент проведення дослідження – доступ надається лише за підпискою (через Discord-бот)
Можливість кастомізації (параметризації)	Обмежена можливість кастомізації	Розширена можливість кастомізації через варіативність параметрів моделі	Обмежена можливість кастомізації порівняно з іншими платформами

На основі порівняльного аналізу систем DALL-E, Stable Diffusion та Midjourney можна сформувати такі висновки:

– DALL-E використовує модель Transformer для створення зображень з текстових описів. Це забезпечує високу якість та деталізацію зображень і дозволяє точно управляти зображенням через деталізовані текстові запити. Однак, доступ до системи обмежений через API, і є певні обмеження на комерційне використання, а також обмежені можливості кастомізації [12].

– Stable Diffusion від Stability AI, заснована на дифузійній моделі, також генерує зображення з текстових описів, але з більшою гнучкістю у стилях і деталізації. Ця система забезпечує модифікований контроль зображення, дозволяючи користувачу більше впливати на процес генерації. Завдяки відкритому доступу вона дозволяє широке комерційне та особисте використання, а також пропонує розширену можливість кастомізації [13].

– Midjourney вирізняється своїм підходом до генерації зображень з текстових запитів, використовуючи власні алгоритми штучного інтелекту. Характеризується високою якістю зображень з унікальним стилем і артистичним виразом. Ця система дозволяє детальне налаштування запитів, хоча результати можуть бути менш передбачувані.

Доступ до Midjourney надається лише за підпискою через Discord-бот, і можливості кастомізації є обмеженими порівняно з іншими, але достатніми для створення зображень.

3. Концептуальна модель системи “Текст – картинка” на основі бізнес-профілю Еріксона – Пенкера

Розробка концептуальної моделі системи “Text to image” потребує визначення класів головних сутностей і множин областей знань у термінах яких може бути формалізована така система. Ключова ідея розробки системи “Text to image” спирається на методологію системної інженерії систем Data Science на основі бізнес профіля Еріксона – Пенкера опубліковану в роботах [14, 15].

Наведемо визначення класів сутностей представлення предметної області Data Science у вигляді діаграми Венна на рис. 2 [14].

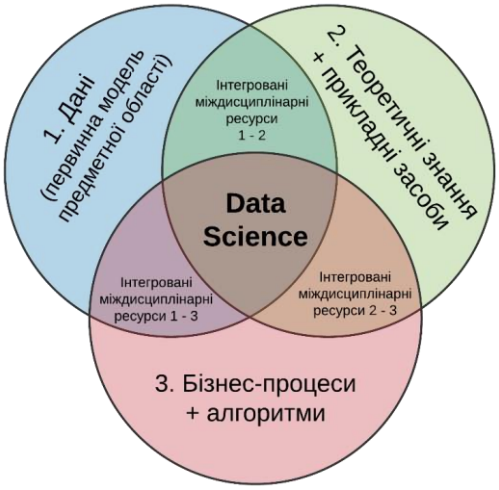


РИС. 2. Представлення класів сутностей Data Science у вигляді діаграми Венна [14]

Перед застосуванням системної інженерії до розробки систем Data Science необхідно чітко визначити поняття множини даних d. Оскільки у світі постійно генерується величезна кількість різнорідних даних, важливо структурувати цей інформаційний хаос, виділивши базовий елемент множини даних – формалізований ресурс даних d. Формалізований ресурс даних d представляє собою зібрану та структуровану інформацію, яка підпорядковується певним принципам класифікації.

В зоні 1, показаній на рис. 2, зібрані дані можуть бути класифіковані за такими ознаками:

– тип носія: можна виділити цифрові носії (жорсткі диски, SSD, USB-накопичувачі тощо) та класичні носії (книги, журнали, наскельні або настінні написи). Сюди ж можна віднести нематеріалізовані дані, такі як мова, думки чи рухи об’єктів;

– структура даних: дані можуть бути структурованими (наприклад, числові масиви, табличні дані або стандартні типи мов програмування) та неструктурованими (аудіо, відео, зображення, текстові документи);

– доступність: дані можуть бути конфіденційними або загальнодоступними.

Якщо формалізований ресурс даних d може бути оброблений відповідними технологіями, то він потрапляє до множини інтегрованих міждисциплінарних ресурсів 1 – 2.

Для прикладу, наведемо декілька формалізованих ресурсів:

– *числові масиви* – це один із найпростіших типів даних для обробки комп'ютером. Масиви можуть мати різні форми, розміри та кількість елементів, а самі числа також можуть виступати як формалізовані ресурси;

– *цифрові зображення* – це структурно впорядковані числові масиви, однак з точки зору сприйняття людиною відіграють особливу роль і, відповідно, можуть класифікуватися як окремий тип формалізованих даних;

– *відео-файли* – більш складний тип даних, що об'єднує зображення й аудіо, включаючи механізм їхньої синхронізації для правильного відтворення.

У контексті Data Science усі зазначені типи даних, а також багато інших, узагальнюються у формалізовані ресурси даних – набори даних (datasets), які можуть містити будь-яку кількість записів (екземплярів), кожен з яких має однакові характеристики.

В термінах Data Science, такі концепції як таблиці баз даних, електронні таблиці (.xlsx файли) та набори даних можна розглядати як синоніми.

Набори даних можуть бути декомпозовані таким чином:

– екземпляри – це кожен набір даних складається з екземплярів (рядків таблиці), які представляють окремі записи;

– характеристики – це кожен екземпляр має певну кількість характеристик (стовпців таблиці), що ідентифікують його. Серед характеристик можуть бути цільові змінні, які використовуються для порівняння передбачених значень у процесі контрольованого навчання з реальними значеннями. Іноді цільові змінні можуть бути відокремлені від основних характеристик на етапі збору даних, особливо в задачах прогнозування. Однак зазвичай їхня декомпозиція виконується вже в процесі аналізу та підготовки даних до моделювання.

Теоретичні знання та прикладні засоби включають будь-які формалізовані або матеріалізовані знання та навички, що дозволяють виконувати операції або забезпечують глибше розуміння даних.

Для формалізації цього поняття визначимо елементарну одиницю цієї множини як інструмент i. Ці інструменти можуть мати різний рівень абстракції та масштабу, утворюючи окремі кластери знань, які об'єднуються в цілі галузі.

Серед найбільш значущих для Data Science можна виділити наступні області:

– статистика – це фундаментальна дисципліна, яка включає безліч підрозділів (k), але в контексті Data Science використовуються лише певні її компоненти, такі як елементи описової статистики на етапах розвідувального аналізу: мода, медіана, середнє арифметичне, стандартне відхилення. Цей процес є частиною розвідувального аналізу даних, що є ключовим етапом у Data Science, забезпечуючи базовий огляд та первинну структуру інформації для подальшого аналізу;

– Machine Learning – витоки цієї галузі датуються роботою Артура Самуеля "Some Studies in Machine Learning Using the Game of Checkers", де було запропоновано концепцію "навчання" машини. Сутність машинного навчання полягає в оптимізації поведінки комп'ютера шляхом адаптації на основі минулого досвіду. Це дозволяє системам самостійно покращувати свої результати без явного програмування алгоритмів для кожної конкретної задачі.

Бізнес-процеси та алгоритми (множина P) – це ключові компоненти для синхронізації взаємодії між даними (множина D) та інструментами (множина I) у процесах Data Science. Для забезпечення систематичної взаємодії теоретичних знань та прикладних засобів з даними, використовується набір методологій системної інженерії в Data Science. Мета цих процесів – оптимізація витягу знань та інформації з формалізованих ресурсів даних d за допомогою відповідних інструментів i. Всі дії, які потрібно здійснити для досягнення цього, організовані у вигляді процесів p. Взаємозв'язок між $d,\ i,\ p$ можна проілюструвати за допомогою концепції інтеграції міждисциплінарних ресурсів.

Розглянувши загальне визначення систем Data Science, переходимо до визначення підгалузі цієї технології.

Перш ніж поглиблюватися в систематизований підхід до інженерії програмних продуктів, слід спочатку розібратися з поняттям системи перетворення текстових даних у зображення. Щоб краще

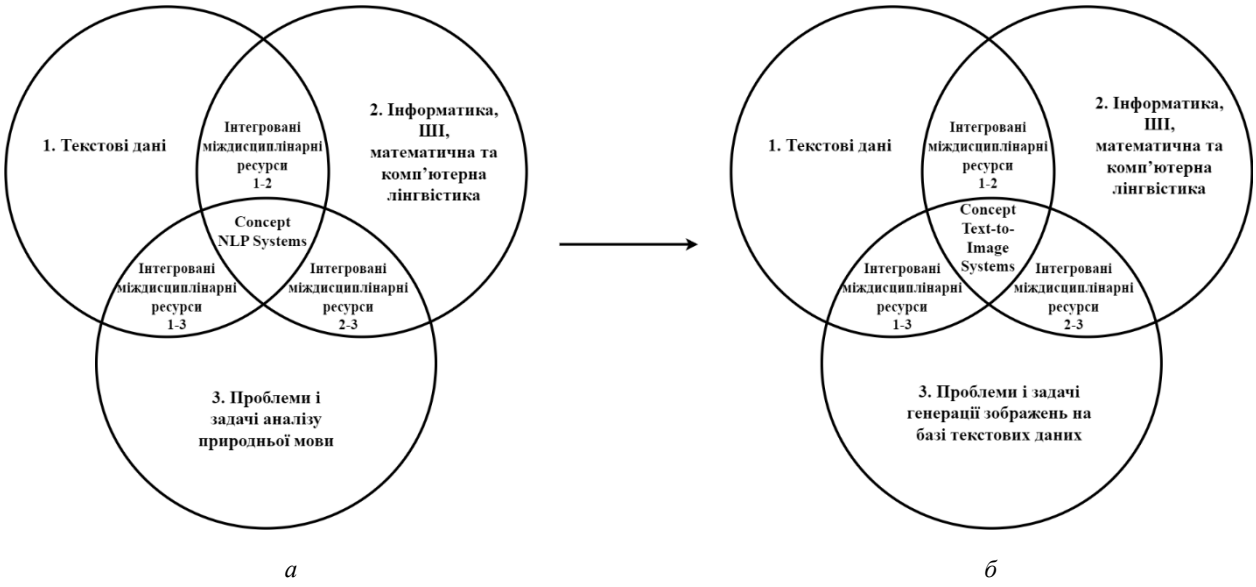
зрозуміти систематичний підхід до розробки програмного забезпечення в контексті таких систем, потрібно розглянути основні компоненти. Діаграма Венна, яку можна застосувати до цієї системи, відображає складові процесу перетворення, об'єднуючи комп'ютерні науки, машинне навчання та генеративні моделі. Аналіз такої діаграми дозволяє визначити взаємодію між різними областями знань, необхідними для створення системи. У результаті аналізу предметної області Text-to-Image, концепція цієї системи може бути розкладена на три класи сутностей:

– множина сутностей Текстові дані та їхнє розуміння;

– множина сутностей Комп'ютерні науки, штучний інтелект, машинне навчання, генеративні моделі;

– множина сутностей Завдання генерації зображень на основі тексту.

Далі можна подати власне визначення класів сутностей предметної області Text-to-Image у вигляді діаграми Венна, що ілюструє взаємозв'язки між текстовими даними, алгоритмами та генеративними методами (рис. 3).



а *б*

РИС. 3. *а* – концепт NLP систем; *б* – концепт систем генерації зображень [14]

1. Сутності Природної Мови, Множина N – сукупність сутностей природної мови в рамках концепції NLP визначається як множина N, де $N \in$ охоплює всі наявні елементи та аспекти, що стосуються розпізнавання, аналізу та розуміння природної мови. У цій множині включаються лінгвістичні одиниці, такі як слова, фрази та речення, а також мовні структури, наприклад, синтаксичні та семантичні взаємозв'язки. До складу N входять також сутності, пов'язані із семантичним аналізом, такі як іменовані сутності, події, атрибути та інші лінгвістичні конструкції. Крім того, вона враховує контекстуальні аспекти, такі як емоційний тон, стилістика та інші мовні нюанси, що допомагають досягти більш точного розуміння тексту та взаємодії з природною мовою в широкому спектрі застосувань, включаючи розробку систем автоматичного перекладу та створення інтелектуальних асистентів.

2. Сутності Інформатика, Штучний Інтелект (ШІ), математична та комп'ютерна лінгвістика, Множина A – ця множина охоплює сукупність концепцій і методів, пов'язаних із використанням

інформатики, штучного інтелекту і математичної лінгвістики для обробки тексту та його перетворення в зображення. До цієї множини входять алгоритми обробки природної мови, методи машинного та глибокого навчання, статистичні моделі, а також комп'ютерно-лінгвістичні техніки.

3. Сутності, що стосуються проблем та завдань аналізу природної мови, формують Множину *P*, яка в рамках концепції NLP охоплює різні виклики та задачі, пов'язані з ефективністю систем обробки природної мови при аналізі та генерації мовлення. До цієї множини входять ключові елементи, які спрямовані на вирішення практичних питань, пов'язаних з обробкою текстової та мовної інформації. Множина *P* включає завдання розпізнавання мовленнєвих шаблонів, класифікації текстів, визначення семантики та ідентифікації сутностей, а також аналізу тональності та емоційного забарвлення мовлення. Сюди також входять такі завдання, як автоматичний переклад, генерація текстових резюме, аналіз сентименту і синтез природної мови. Крім того, до цієї множини відносяться більш складні задачі, а саме вирішення амбівалентності, обробка невизначеностей у текстах та інші аспекти, які забезпечують більш глибокий та точний аналіз і синтез природної мови. На діаграмі Венна, яка показана на рис. 2, перетин множин 1-2, 1-3 і 2-3, що відповідають сутностям N, A і P, формує інтегровані міждисциплінарні ресурси, які виділяються через наявність функціональних зв'язків між сутностями різних типів, таких як множини Set N, Set A і Set P.

Інтегровані міждисциплінарні ресурси 1-2 (T-A) об'єднують знання та технології текстової обробки й штучного інтелекту для аналізу тексту і його перетворення в зображення. Інтегроваі міждисциплінарні ресурси 1-3 (T-P) зосереджується на проблемах візуалізації тексту, враховуючи емоційний і контекстуальний зміст, і забезпечує відповідність зображення текстовим характеристикам. Інтегроваі міждисциплінарні ресурси 2-3 (A-P) поєднує інформатику, ШІ та лінгвістику для вирішення завдань візуальної відповідності, точності та емоційної передачі тексту у зображеннях.

Для розробки концептуальної моделі застосовується метод системної інженерії інформаційно комунікаційних систем на основі бізнес профіля Еріксона – Пенкера [1, 14, 15]. В першу чергу, структура системи моделюється у вигляді бізнес-профіля Еріксона – Пенкера [1, 2, 14, 15] (рис. 4): визначаються її проблема, мета, ресурси, процеси, цілі, правила, а також проєктуються детальні структурне та динамічне представлення.

Розберемо детально кожен з класів діаграми (рис. 4):

– клас *Проблема* – актуальне питання, що потребує відповідних рішень; основна мотивація розробки системи, яка спонукає до формулювання конкретної мети. Проблема даної роботи: *ефективне перетворення текстових даних у зображення, що є актуальним питанням в сфері візуалізації даних*;

– клас *Мета* – виражає глобальну ціль роботи, покликана вирішити поставлену проблему. Мета даної роботи: *розробка та валідація моделі перетворення текстових описів на зображення, що здатна ефективно інтерпретувати текстові дані у формі художньо виражених зображень*;

– клас *Процес* – складові загальної діяльності системи, в результаті якої досягається мета; чітко визначена послідовність дій/підпроцесів, що призводить до виконання певного завдання. Процеси даної системи це: *Завантаження зображень, Попередня обробка зображень, Навчання моделі нейронної мережі для генерації зображень, Генерація зображень з шумового вектору, Збільшення розмірів зображень, Перетворення згенерованих даних у зображення, Функціонування інтерфейсу користувача*;

– клас *Зміна стану* – можливі зміни певних ресурсів у результаті роботи процесів. Система налічує 5 різних змін станів.

1. Необроблені зображення → Попередньо оброблені зображення (процес *Попередня обробка зображень*);

2. Ініціалізована модель → Навчена модель (процес *Навчання моделі нейронної мережі для генерації зображень*);

3. Дані у вигляді шумового вектору → Згенероване зображення;

4. Згенеровані зображення низької якості → Зображення вищої якості;

5. Набір згенерованих даних → Результуючі зображення.

– клас *Ресурс* – будь-які сутності (матеріальні чи ні), що споживаються та продукуються розроблюваною системою.

1. Клас *Вихід бізнес-процесу* – ресурси, що продукуються системою, кінцевий результат її функціонування. До цього можна віднести згенеровані зображення та навчену модель нейронної мережі.

2. Клас *Ресурс забезпечення виконання бізнес-процесу* – ресурси, що забезпечують виконання процесів, але не є кінцевим результатом її роботи: модель обробки зображень, модель покращення якості зображень, генератор.

3. Клас *Вхід бізнес-процесу* – первинні ресурси входу початкових процесів, які ініціалізують цикл роботи системи: необроблені зображення, параметри генерації зображень, ініціалізована модель нейронної мережі.

– клас *Подія* – виникає за рахунок певних зовнішніх факторів або ж як результат взаємодії між процесами.

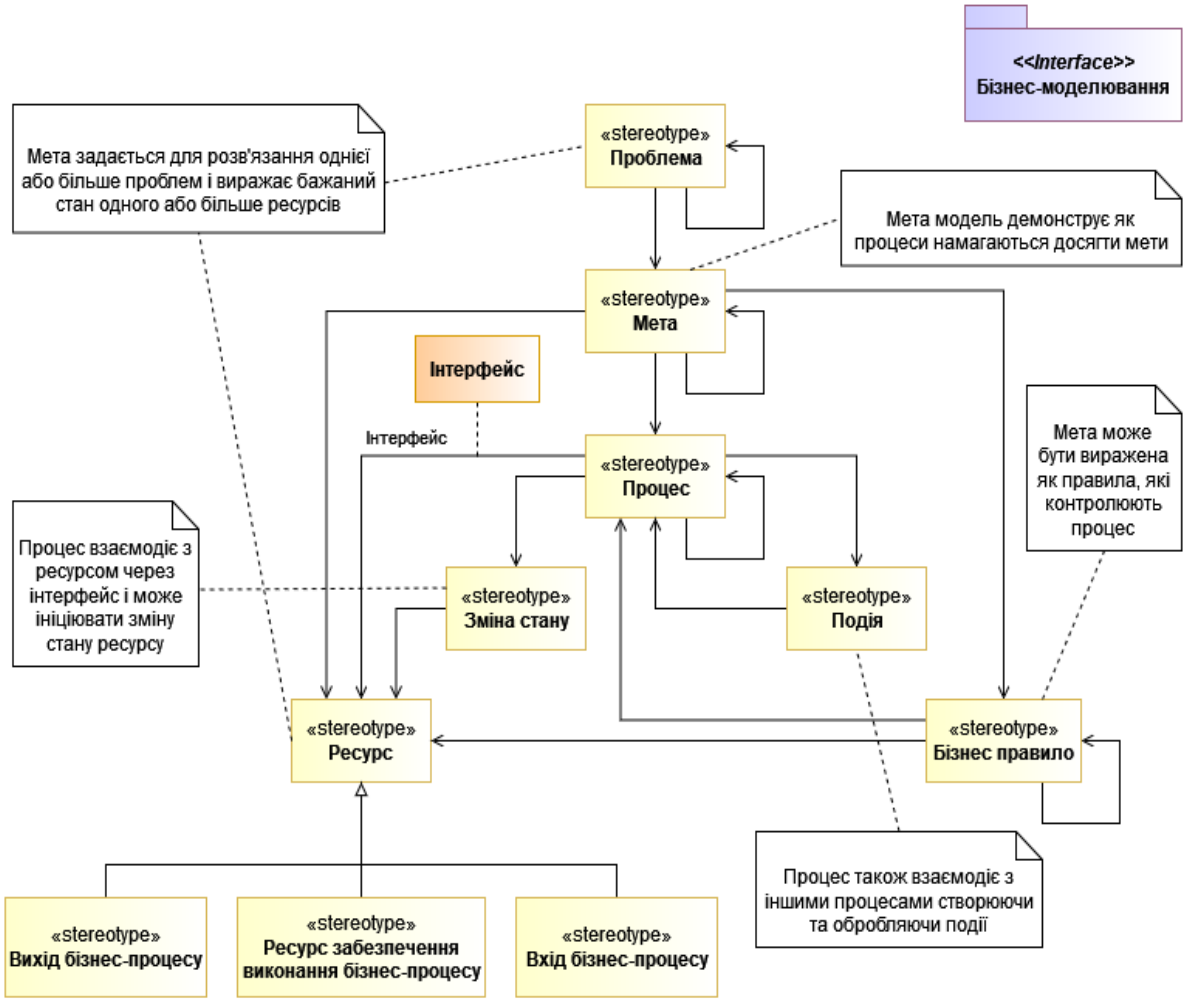


РИС. 4. Концептуальна модель системи “Text to image” на основі бізнес-профіля Еріксона – Пенкера. Діаграма класів у нотації UML [15]

Модель системи “Text to image” у вигляді структурного зображення показано на рис. 5.

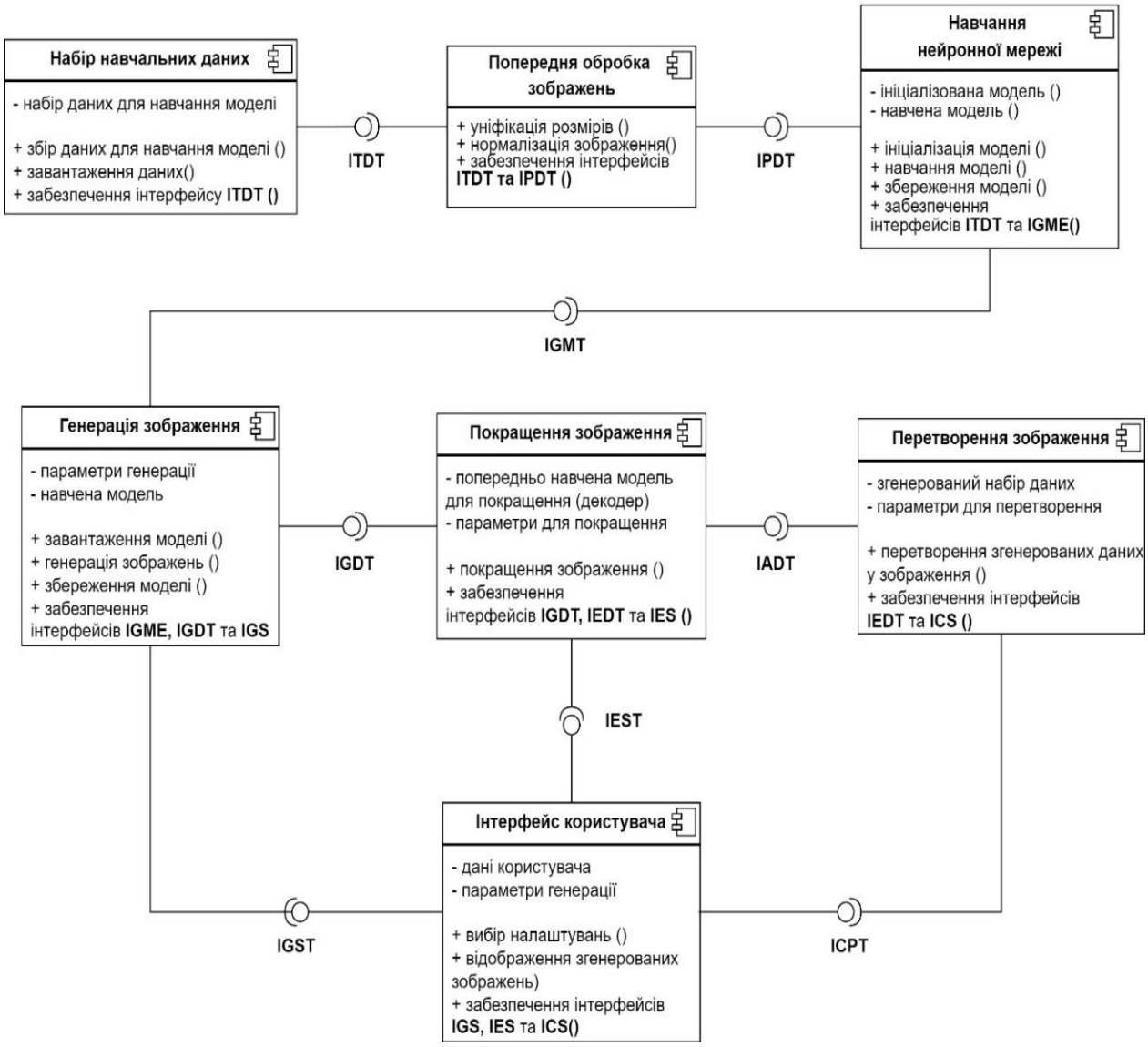


РИС. 5. Модель системи “Text to image”. Діаграма компонентів у нотації UML

Функціональність і призначення компонентів системи:

– набір навчальних даних: локальне сховище даних, де зберігаються набір зображень, які будуть використані для навчання нейронної мережі;

– попередня обробка зображень: набір інструментів, який використовується для попередньої обробки та підготовки зображень;

– навчання нейронної мережі: набір інструментів, який ініціалізує архітектуру нейронної мережі та проводить навчання на наборі даних;

– генерація зображення: набір інструментів, які генерують зображення за заданими користувачем параметрами з використанням навченої мережі;

– покращення зображення: попередньо навчена модель (декодер), яка використовується для збільшення розмірів зображень (upscaling);

– перетворення зображення: набір інструментів, які використовуються для перетворення отриманих даних у формат зображення;

– інтерфейс користувача: компонент, який уособлює інтерфейс взаємодії з користувачем. Відповідний інтерфейс відповідає за збір даних необхідних для генерації зображення (текстовий опис, параметри для генерації, розмір зображення), а також відображення процесу генерації.

Список інтерфейсів:

– ITDT (Interface of Training Data Transferring) – інтерфейс для передавання даних із модуля сховища навчальних даних на вхід відповідного модуля Попередня обробка зображень;

– IPDT (Interface of Processed Data Transferring) – інтерфейс для передавання попередньо оброблених даних із модуя Попередня обробка зображеня на вхід до модуля Навчання нейронної мережі;

– IGMT (Interface of Generation Model Transferring) – інтерфейс для передавання навченої моделі нейронної мережі із модуля Навчання нейронної мережі на вхід до модуля Генерація зображень;

– IGDT (Interface of Generated Data Transferring) – інтерфейс для передавання згенерованого зображення із модуля Генерація зображень на вхід до модуля Покращення зображення;

– IGST (Interface of Generation Setting Transferring) – інтерфейс для передавання параметрів генерації зображень із модуля Інтерфейсу користувача на вхід модуля Генерація зображення;

– ICPT (Interface of Conversion Parameters Transferring) – інтерфейс для передавання параметрів перетворення зображення із модуля Інтерфейсу користувача на вхід модуля Перетворення зображення;

– IADT (Interface of Upscaled Data Transferring) – інтерфейс для передавання зображення високої роздільної здатності із модуля Покращення зображення на вхід до модуля Перетворення зображення.

На рис. 6 показана Модель системи “Text to image” у формі діаграми діяльності першого рівня у нотації UML.

На рис. 7 показано модель системи “Text to image” формалізовану у вигляді деталізованої діаграми діяльності з водними доріжками другого рівня, де показані внутрішня структура компонентів системи.

Для реалізації генерації зображень було обрано використовувати підхід на основі латентної дифузійної моделі.

Модель латентної дифузії відрізняється від звичайної моделі дифузії тим, що вона ініціює процес дифузії в латентному просторі, а не в просторі пікселів, що призводить до зниження витрат на навчання [16].

Основна концепція включає як процес прямої дифузії, так і зворотний процес (реконструкція), подібний до традиційної моделі дифузії.

Модель латентної дифузії ефективно розділяє перцептивне та семантичне стиснення за допомогою методів генеративного моделювання.

Спочатку вона усуває надлишковість на рівні пікселів за допомогою автокодувальника, а потім маніпулює та генерує семантичні концепції за допомогою процесу дифузії, застосованого до отриманих латентних даних.

Для того, щоб оптимізувати обробку вхідних даних і зменшити обчислювальні вимоги до навчання моделі дифузії, запропоновано перетворити вхідні дані в латентне представлення через мережу кодувальників. Цей підхід спрямований на обробку вхідних даних у просторі меншої розмірності. Згодом загальна модель U-Net, використовується для створення нових даних, які далі аналізуються мережею декодування [17].

Отже, при забезпеченні кодером ε і латентним представленням z, функція втрат для моделі латентної дифузії визначається відповідно до рівняння

$$L_{\text{LDM}} = \mathbb{E}_{E(x),\epsilon\sim\mathcal{N}(0,1),t}\left[\left\|\epsilon - \epsilon_0(z_t, t)\right\|_2^2\right],$$

де $E(x)$ – кодувальник, який перетворює зображення x в латентний простір, $\epsilon \sim \mathcal{N}(0,1)$ – гауссівський шум, z_t – латентне представлення на етапі t, ϵ_0 – передбачення шуму на основі параметрів моделі, t – ітеративний час дифузії. На кожному часовому кроці нейронна мережа UNet передбачає шум у репрезентації зображення на поточному кроці часу [18].

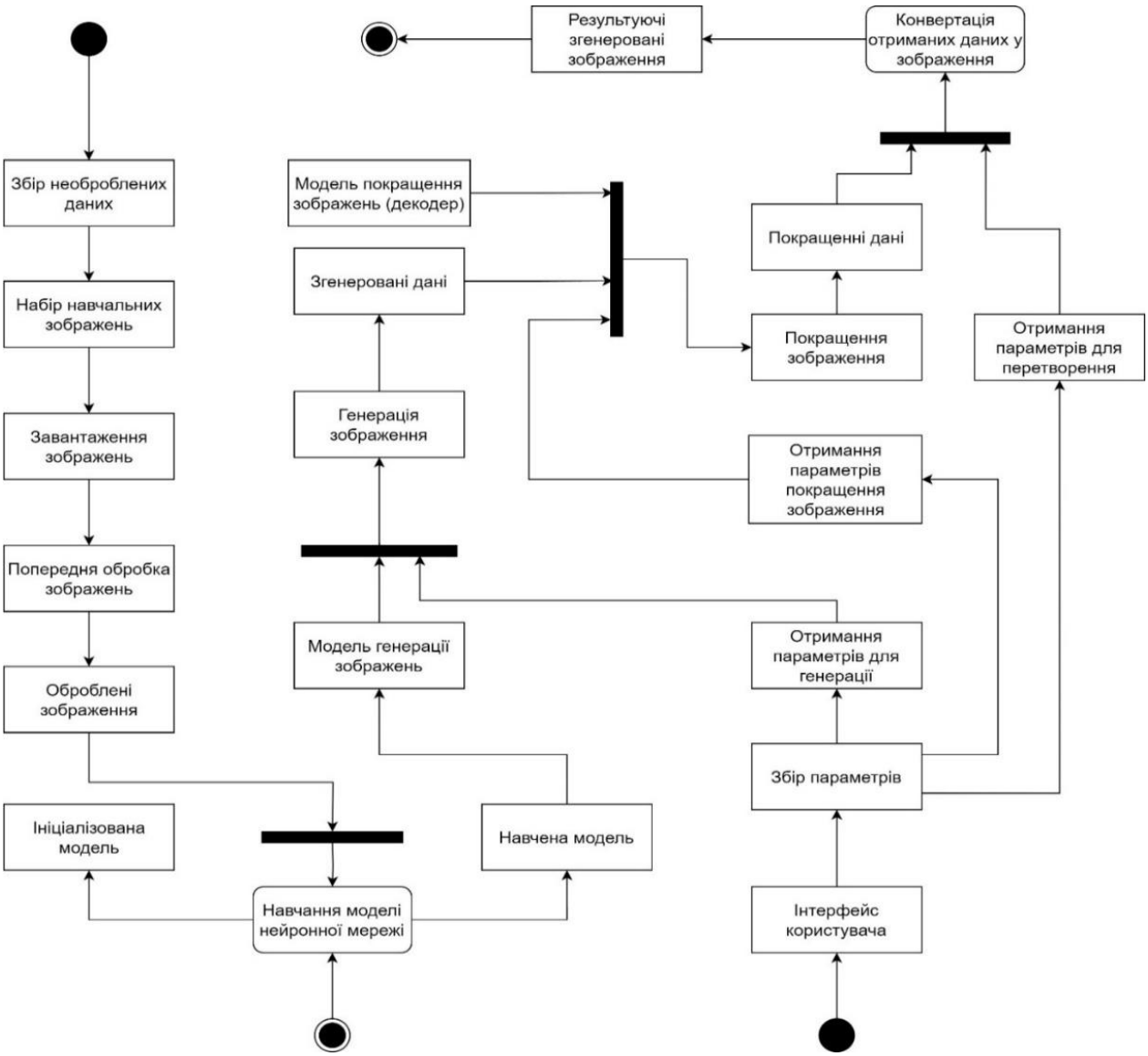


РИС. 6. Модель системи “Text to image”. Діаграма діяльності першого рівня у нотації UML

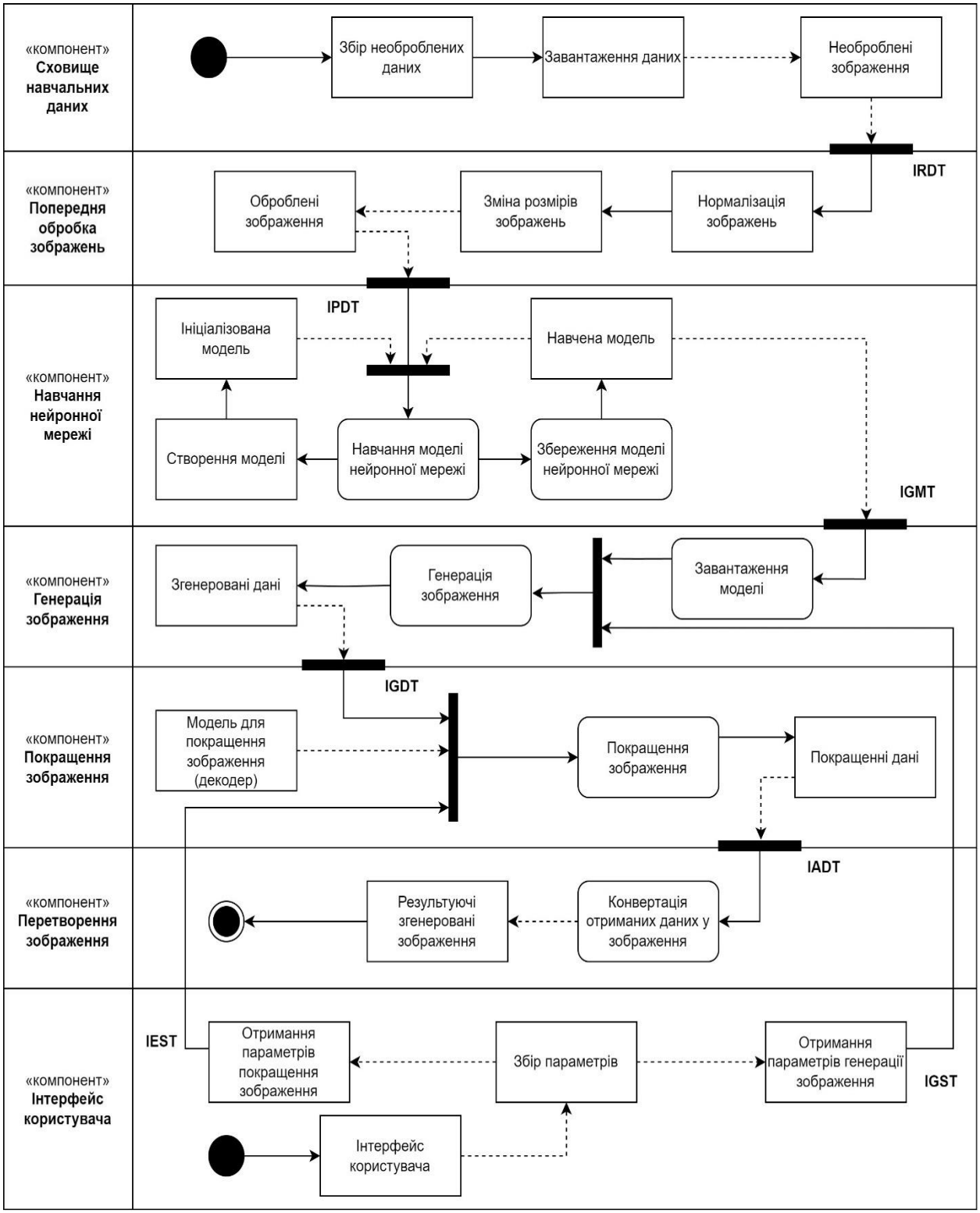


РИС. 7. Модель системи “Text to image”. Діаграма діяльності з водними доріжками другого рівня у нотації UML

На рис. 8 показана блок-схема роботи алгоритму генерації зображень з текстових даних з використанням латентної дифузійної моделі.

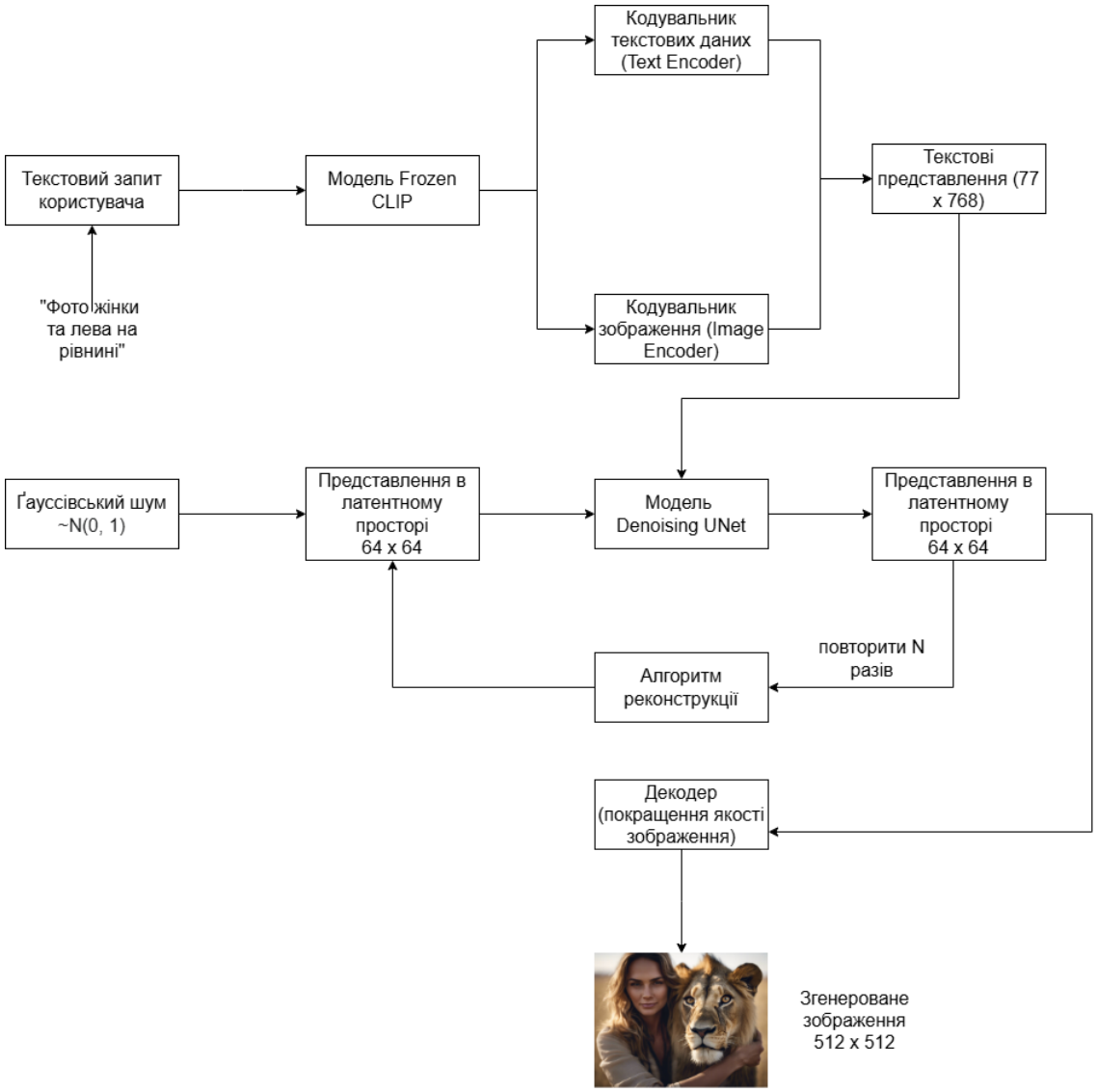


РИС. 8. Алгоритм перетворення текстових даних у зображення з використанням моделі латентної дифузії

На початковій стадії VQGAN використовується для обробки прихованого представлення, після чого використовується модель прихованої дифузії відповідно до архітектури DALL-E, моделі, яка покладається на VQVAE для сприйняття візуального коду. VQGAN покращує структуру VQVAE, інтегруючи конкурентну мету для підвищення автентичності створених зображень. Крім того, попередньо навчена модель VAE вносить модифікації у процес прямої дифузії, вносячи шум у прихований простір. Метод включає перехресну увагу як вимогу для різноманітних контекстних сигналів (текст).

Спочатку, використовуючи кодувальник ε, зображення піддається стисненню, що здійснюється у прихованому просторі кодувальник, а не в просторі пікселів. Згодом, під час прогресування дифузії, гауссівський шум інтегрується у представлення. Після цього представлення подається через

UNet, яка відповідає за передбачення. Ця ітераційна процедура повторюється T разів послідовно, доки не буде досягнуто z, яке потім зміщується з латентного простору в простір пікселів декодером D. Кінцева фаза включає постійне обумовлення шляхом узгодження з іншою модальністю, наприклад, текст. Модальність введення y зазнає перетворення через спеціалізований кодувальник і згодом пов'язується з проміжними рівнями U-Net за допомогою того самого рівня перехресної уваги, який представлений у трансформерах.

Щоб розробити автокодер для стиснення зображень, реалізовано методологію, подібну до тієї, що використовується у VQGAN (рис. 8 і 9) [19]:

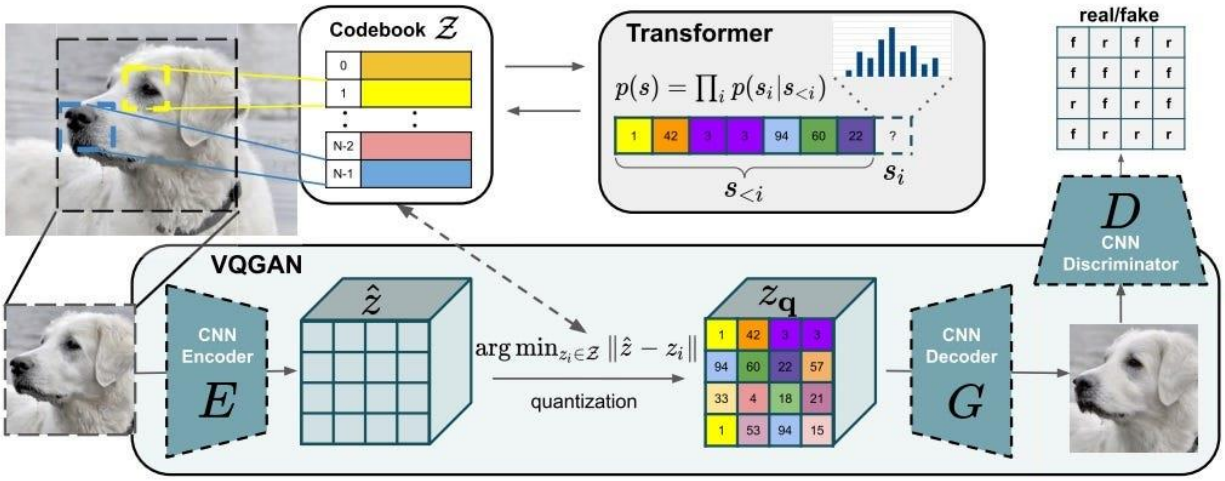


РИС. 9. Архітектура VQGAN, яка використовується для вивчення візуальних частин, композиція яких згодом моделюється за допомогою архітектури авторегресійного трансформер

Цільова функція для навчання має вигляд

$$L_{\text{Autoencoder}} = \min_{E,D} \max_{\psi} \left(L_{\text{rec}}(x, D(E,x)) - L_{\text{adv}}(D(E(x))) + \log\left(D_{\psi}(x)\right)\right),$$

де L_{adv} – adversial loss (змагальна втрата), L_{rec} – квадратична помилка між x та рекунструйованим $\hat{x}$ (втрата реконструкції), D_{ψ} – дискримінатор.

У процесі регуляризації можна вибрати або регулярізацію Кульбака – Лейблера, або використовувати рівень векторного квантування. Перший метод передбачає апроксимацію розподілу латентної змінної для узгодження зі стандартним нормальним розподілом [20]. Це коригування має на меті структурувати та консолідувати латентний простір так, щоб коли дві приховані змінні, z_1 та z_2, знаходяться поруч, їхні представлення, $D(z_1)$ і $D(z_2)$, демонстрували певний ступінь подібності. Отже, декодер отримує вхідні дані, що містять лише комбінації векторів кодової книги – по суті, дискретизовану або квантовану версію прихованого простору. Цей підхід передбачає вбудовування дифузійної моделі в текст шляхом її інтеграції в проміжні шари моделі U-Net за допомогою перехресної уваги, подібно до архітектури трансформатора. Точніше, за допомогою доменно-спеціального кодувальник вхідні дані (наприклад, текстовий вміст) піддаються перетворенню в проміжне представлення. Згодом це представлення включається до рівнів U-Net після обходу через модуль перехресної уваги:

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,\ldots,\text{head}_h)W^O,$$

де $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$,

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

де $Q = W_Q^i \cdot \varphi(z_i)$, $W_Q^i \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $K = W_K^i \cdot \tau_0(y)$, $W_K^i \in \mathbb{R}^{d_{\text{model}} \times d_\tau}$, $\varphi(z_i)$ – проміжне представлення UNet, τ_0 – кодувальник, який виконує про'кцію y на представлення $\tau_0(y) \in \mathbb{R}^{d_{\text{model}} \times d_\tau}$.

4. Верифікація та валідація системи “Text to image”

В табл. 2 наведено порівняльний аналіз імплементації системи “Text to image” з існуючими системами перетворення текстових даних у зображення.

Для наочності, розроблену систему буде названо TransformerLD, де LD – Latent Diffusion.

Таблиця 2. Порівняльний аналіз розробленої системи з існуючими

Модель	FID Score (менше – краще)	Inception Score (більше – краще)
TransformerLD	35.65	2.28
Imagen	27.27	9.85
DALL-E 2	32.14	4.37
Stable Diffusion v2.0	31.59	4.89
GLIDE	32.24	8.95
XMC-GAN	31.33	7.45

Примітка: сформовано автором на основі власних розрахунків та [21, 22].

Модель Imagen демонструє найкращий результат за обома показниками. Її FID Score (27.27) свідчить про найменшу різницю між розподілом реальних та згенерованих зображень, що означає високу якість і реалістичність вихідних зображень. Крім того, Inception Score (9.85) вказує на те, що модель генерує різноманітні зображення, які легко розпізнаються нейронними мережами. XMC-GAN також досягає хороших результатів з FID Score 31.33 і Inception Score 7.45, що робить її другою за якістю після Imagen, хоча її Inception Score трохи нижчий, ніж у лідерів, що може свідчити про меншу різноманітність або складність зображень. Stable Diffusion v2.0 демонструє сильні результати з FID Score 31.59 та Inception Score 4.89. Ця модель генерує зображення з відносно високою реалістичністю та різноманітністю, хоча і поступається Imagen у загальній якості. DALL-E 2 також показує достойні результати з FID Score 32.14 та Inception Score 4.37, демонструючи хорошу якість зображень, хоча різниця між реальними та згенерованими зображеннями трохи більша, ніж у Imagen або XMC-GAN. Модель GLIDE із FID Score 32.24 та Inception Score 8.95 показує нижчу реалістичність порівняно з іншими моделями, але все ще підтримує хорошу візуальну якість та різноманітність зображень. TransformerLD демонструє FID Score (35.65) та Inception Score (2.28) серед представлених моделей, що свідчить про більшу різницю між реальними та згенерованими зображеннями. Однак, попри ці показники, TransformerLD має кілька значних переваг. Однією з головних є семантична

узгодженість між текстовим запитом та згенерованим зображенням завдяки використанню трансформерної архітектури. Це дозволяє моделі краще обробляти складні текстові запити та створювати зображення, які детально відповідають опису.

Розроблена система є безкоштовною, що відрізняє її від більшості аналогічних систем на ринку, які вимагають платних підписок або платежів для доступу до повного функціоналу. Відсутність плати за користування робить систему доступною для ширшого кола користувачів, у тому числі для дослідників, студентів та початківців у галузі генеративних моделей. Такий підхід сприяє популяризації технології серед різних аудиторій, стимулюючи розвиток нових застосувань у різних сферах, де створення зображень на основі текстових запитів може принести користь.

Розробка концептуальної моделі є значним досягненням у процесі створення системи, оскільки вона забезпечує структуровану основу для майбутнього вдосконалення та адаптації. Модель побудована таким чином, що її елементи можуть бути легко замінені або оновлені в залежності від потреб користувачів та вимог ринку. Наприклад, можливість заміни компонентів моделі дозволяє інтегрувати більш передові методи уваги або генеративні алгоритми, підвищуючи продуктивність і точність генерації зображень. Це забезпечує гнучкість і масштабованість системи, що є важливими факторами для її подальшого розвитку та адаптації до нових вимог та технологічних тенденцій.

Висновки

1. В роботі запропоновано NLP систему “Text to image” на основі методології системної інженерії систем Data Science, архітектуру та програмне забезпечення системи генерації зображень на основі латентної дифузійної моделі. Запропоновано покращення базової архітектури латентної дифузійної моделі шляхом використання дифузійного трансформера. Дифузійний трансформер застосовується для покращення якості синтезованих зображень, особливо в тих випадках, коли генерація базується на текстових інструкціях або інших мультимодальних вхідних даних.

2. На відміну від підходів, що базуються на архітектурі U-Net, DiTs працюють з латентними патчами, забезпечуючи кращу масштабованість та підвищену продуктивність.

3. Проведено верифікацію та валідацію розробленої системи для перетворення текстових даних в зображення. Результати генерації демонструють точне відтворення ключових елементів, що свідчить про високу якість відповідності зображення текстовому опису. В результаті проведення порівняльного аналізу продуктивності моделей визначено, що система TransformerLD, хоч і поступається моделям Stable Diffusion і DALL-E 2 за показниками FID та IS, все ж залишається конкурентоспроможною.

4. Можливі напрямки покращення системи генерації зображень на основі латентної дифузійної моделі з використанням дифузійного трансформера можуть включати оптимізацію архітектури трансформера для підвищення ефективності та точності генерації, зокрема шляхом покращення механізмів уваги. Це дозволить краще враховувати глобальні та локальні залежності між елементами зображення, що сприятиме підвищенню якості синтезованих зображень, особливо у випадках з великими розмірами зображень. Крім цього, навчання на більш різноманітних і якісних наборах даних також може підвищити здатність системи до генерації зображень, у відповідності до більш складних інструкцій.

Авторські внески.

Маслянко П.П.: методологія системної інженерії систем Data Science на основі бізнес-профіля Еріксона-Пенкера, метод системної інженерії NLP-систем, концептуалізація та модель NLP-системи “Text to image”, узагальнення, написання і редагування, висновки.

Павловська К.І.: огляд існуючих рішень, імплементація NLP-системи “Text to image”, програмування, верифікація і валідація, написання і редагування, висновки.

Список літератури

1. Yin L. A Review of Text-to-Image Synthesis Methods. *2024 5th International Conference on Computer Vision*. 2024. P. 858–861. https://ieeexplore.ieee.org/document/10603609
2. Li H. et al. On the Scalability of Diffusion-based Text-to-Image Generation. *2024 Conference on Computer Vision and Pattern*. 2024. P. 9400–9409. https://ieeexplore.ieee.org/document/10655871
3. Patel M., Kim C., Cheng S., Baral C., Yang Y. ECLIPSE: A Resource-Efficient Text-to-Image Prior for Image Generations. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2024. P. 9069–9078. https://ieeexplore.ieee.org/document/10656952
4. Zhang Y., Song Y., Yu J., Pan H., Jing Z. Fast Personalized Text to Image Synthesis with Attention Injection. *ICASSP 2024 IEEE International Conference on Acoustics, Speech and Signal Processing.* 2024. P. 6195–6199. https://ieeexplore.ieee.org/document/1044704
5. Rauniyar A., Raj A., Kumar A., Kandu A.K., Singh A., Gupta A. Text to Image Generator with Latent Diffusion Models. *International Conference on Computational Intelligence and Networking.* 2023. P. 144–148. https://ieeexplore.ieee.org/document/10140348
6. Prerak S. Addressing Bias in Text-to-Image Generation: A Review of Mitigation Methods. *2024 Third International Conference on Smart Technologies and Systems for Next Generation Computing*. 2024. P. 1–6. https://ieeexplore.ieee.org/document/10671230
7. Shi J., Xiong W., Lin Z., Jung H.J. InstantBooth: Personalized Text-to-Image Generation without Test-Time Finetuning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024. P. 8543–8552. https://ieeexplore.ieee.org/document/10657619
8. Yamac A., Genc D., Zaman E., Gerschner F., Klaiber M., Theissler A. Open-Source Text-to-Image Models: Evaluation using Metrics and Human Perception. *Annual Computers and Applications Conference*. 2024. P. 1659–1664. https://ieeexplore.ieee.org/document/1063362
9. Text-to-image: latent diffusion models. Nicd: офіційний веб-сайт. https://nicd.org.uk/knowledge-hub/image-to-text-latent-diffusion-models (звернення: 21.11.2024)
10. TokenCompose: Text-to-Image Diffusion with Token-level Supervision. https://mlpc-ucsd.github.io/TokenCompose/ (звернення: 21.11.2024)
11. Zhang S. e't al. Learning Multi-Dimensional Human Preference for Text-to-Image Generation. *2024 Conference on Computer Vision and Pattern Recognition (CVPR),* Seattle, WA, USA, 2024. P. 8018–8027. https://ieeexplore.ieee.org/document/10655849/
12. Maung A., Nguyen H.H., Kiya H., Echizen I. Fine-Tuning Text-To-Image Diffusion Models for Class-Wise Spurious Feature Generation. *2024 IEEE International Conference on Image Processing (ICIP).* 2024. P. 3910–3916. https://ieeexplore.ieee.org/document/1064762
13. Peebles W., Xie S. Scalable Diffusion Models with Transformers. arXiv. 2022. https://arxiv.org/abs/2212.09748 (звернення: 21.11.2024)
14. Maslianko P., Sielskyi Y. Data Science — Definition and Structural Representation. *System Research & Information Technologies*. 2021. No. 1. P. 61–78. https://doi.org/10.20535/SRIT.2308-8893.2021.1.05
15. Маслянко П.П., Сєльський Є.П. Метод системної інженерії систем нейронного машинного перекладу. *Наукові вісті КПІ*. 2021. № 2. С. 46–55. https://doi.org/10.20535/kpisn.2021.2.236939
16. Kandwal S., Nehra V. A Survey of Text-to-Image Diffusion Models in Generative AI. *International Conference on Cloud Computing.* 2024. P. 73–78. https://ieeexplore.ieee.org/document/1046337
17. Ahamed S., Al Amin A., Ahsan S.M.M. Synthesizing Realistic Images from Textual Descriptions: A Transformer-Based GAN Approach. *2023 International Conference on Next-Generation Computing*. 2023. P. 1–6. https://ieeexplore.ieee.org/document/10212565
18. Rethinking FID: Towards a Better Evaluation Metric for Image Generation. https://arxiv.org/html/2401.09603 (звернення: 21.11.2024)
19. Zhou D., Li Y., Ma F., Zhang X., Yang Y. MIGC: Multi-Instance Generation Controller for Text-to-Image Synthesis. *IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2024. P. 6818–6828. https://ieeexplore.ieee.org/document/10658514
20. He F. et al. CartoonDiff: Training-free Cartoon Image Generation with Diffusion Transformer Models. *IEEE International Conference on Acoustics, Speech and Signal Processing.* 2024. P. 3825–3829. https://ieeexplore.ieee.org/document/10447821
21. Quantitative Comparison of Stable Diffusion, Midjourney and DALL-E 2. http://bit.ly/3BVLwwV (звернення: 21.11.2024)

22. Akar C.A., Luckow A., Obeid A., Beddawi C., Kamradt M., Makhoul A. Enhancing Complex Image Synthesis with Conditional Generative Models and Rule Extraction. 2023. P. 136–143. https://ieeexplore.ieee.org/document/10459883

Одержано 12.11.2024

Маслянко Павло Павлович,
кандидат технічних наук, старший науковий співробітник
Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського",
Україна, Київ,
https://orcid.org/0000-0003-4001-7811
masliankop@gmail.com

Павловська Катерина Ігорівна,
магістрант
Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського",
Україна, Київ.

УДК 519.688

П.П. Маслянко [*], К.І. Павловська

Концептуальна модель та NLP-система "Text to image"

Національний технічний університет України "КПІ імені Ігоря Сікорського", Україна, Київ
[*] Листування: masliankop@gmail.com

Вступ. Розробка теоретичних інструментів та інструментальних засобів трансформації текстової інформації у зображення є актуальною проблемою для різних галузей діяльності людини та організаційних систем різноманітного призначення.

В статті запропоновано концептуальну модель та NLP-систему "Text to image" на основі методології системної інженерії систем Data Science, архітектуру та програмне забезпечення системи генерації зображень на основі латентної дифузійної моделі. Запропоновано покращення базової архітектури латентної дифузійної моделі шляхом використання дифузійного трансформера.Встановлено, що на відміну від підходів, що базуються на архітектурі U-Net, DiTs працюють з латентними патчами, забезпечуючи кращу масштабованість та підвищену продуктивність.

Мета роботи – розробка науково обгрунтованої концептуальної моделі та системи для перетворення текстових описів у зображення, що базується на методології системної інженерії, сучасних методах глибинного навчання та бізнес профілі Еріксона – Пенкера.

Результати. Побудовано оціночні задачі, властивості яких регулюються параметром, для задачі розміщення об'єктів в евклідовому просторі. Досліджені властивості оціночної задачі в залежності від значення параметра та показані межі значення параметра, дотримання яких дозволяє отримувати оцінки, адекватні початковій задачі.

Проведено верифікацію та валідацію розробленої NLP-систему "Text to image" для перетворення текстових даних у зображення. Результати генерації демонструють точне відтворення ключових елементів, що свідчить про високу якість відповідності зображення текстовому опису. В результаті проведення порівняльного аналізу продуктивності моделей визначено, що система TransformerLD, хоч і поступається моделям Stable Diffusion і DALL-E 2 за показниками FID та IS, все ж залишається конкурентоспроможною.

Висновки. Побудова динамічного дерева розгалужень та нелінійних оцінок дозволяє прискорити процес пошуку оптимального розв'язку, але суттєво залежить від початкової задачі, що ускладнює розробку загального алгоритму.

Розробка концептуальної моделі, та NLP-системи "Text to image" дозволяє реалізувати ефективне перетворення текстових даних у зображення, що є актуальним питанням в сфері візуалізації даних.

Ключові слова: системна інженерія, Data Science, NLP-системи "Text to image.

UDC 519.688

Pavlo Maslianko [*], Kate Pavlovska

Conceptual Model and NLP-System "Text to Image"

Igor Sikorsky Kyiv Polytechnic Institute,Ukraine
[*] Correspondence: masliankop@gmail.com

Introduction. The development of theoretical tools and instrumental means of transforming text information into images is an urgent problem for various fields of human activity and organizational systems of various purposes. The article proposes a conceptual model and NLP system "Text to image" based on the methodology of system engineering of Data Science systems, architecture, and software of the image generation system based on the latent diffusion model. It is proposed to improve the basic architecture of the latent diffusion model by using a diffusion transformer. It is found that unlike approaches based on U-Net architecture, DiTs work with latent patches, providing better scalability and increased performance.

The purpose of the work is to develop a scientifically based conceptual model and system for transforming text descriptions into images, based on the methodology of system engineering, modern methods of deep learning and business profile of Erikson – Penker.

Results. Estimation problems, the properties of which are regulated by a parameter, have been constructed for the problem of placing objects in Euclidean space. The properties of the evaluation problem depending on the value of the parameter are studied and the limits of the value of the parameter are shown, the observance of which allows obtaining estimates adequate to the initial problem. Verification and validation of the developed NLP system "Text to image" for converting text data into images was carried out. The generation results demonstrate the exact reproduction of key elements, which indicates the high quality of the correspondence between the image and the text description. As a result of a comparative analysis of the performance of the models, it was determined that the TransformerLD system, although inferior to the Stable Diffusion and DALL-E 2 models in terms of FID and IS, still remains competitive.

Conclusions. The construction of a dynamic branching tree and nonlinear estimations allows speeding up the process of finding the optimal solution, but it depends significantly on the initial problem, which complicates the development of a general algorithm. The development of the conceptual model and the NLP system "Text to image" allows implementing the effective transformation of text data into images, which is a topical issue in the field of data visualization.

Keywords: system engineering, Data Science, NLP-systems “Text to image.