МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Харківський національний університет імені В.Н.Каразіна

Факультет математики і інформатики

Кафедра теоретичної та прикладної інформатики

Кваліфікаційна робота

магістр

на тему “Використання існуючих архітектур моделей для вирішення вузьконаправлених задач генерації зображень.”

Виконав: студент 2 курсу, групи МФ-61

спеціальність 122 «Комп’ютерні науки»

освітньо-наукова програма

«Інформатика»

Дмітрієв І. М.

Керівник Дейнега О. А.

Рецензент _____

Харків – 2023

ЗМІСТ

Вступ

1.1 Формулювання мети роботи, задач та обґрунтування актуальності теми

У останні роки зростає попит на системи генерації зображень у багатьох галузях, таких як комп'ютерне зору, медицина, мистецтво, реклама та інші. Це пояснюється тим, що зображення використовуються для багатьох цілей, наприклад, для ідентифікації об'єктів на зображеннях, для діагностики захворювань, для створення візуально привабливих дизайнів та рекламних матеріалів.

Застосування глибоких нейронних мереж та генеративно змагальних мереж (ГЗМ) дозволяє створювати високоякісні зображення, що досить точно відображають дійсність. ГЗМ зазвичай складається з двох глибоких нейронних мереж: генератора та дискримінатора. Генератор створює зображення з шуму, а дискримінатор оцінює, наскільки добре зображення відповідає дійсності.

Однак, для досягнення цього результату необхідно використовувати складні архітектури моделей та відповідні алгоритми тренування. Тому дослідження та порівняння різних існуючих архітектур моделей для вирішення вузько направлених задач генерації зображень є важливим завданням, яке може допомогти вибрати найбільш оптимальну архітектуру для вирішення конкретної задачі. Оскільки існує велика кількість архітектур моделей для генерації зображень, необхідно провести дослідження, щоб вибрати оптимальну архітектуру для вирішення конкретної задачі. Також, дослідження може допомогти зрозуміти, які

фактори впливають на якість генерації зображень та як їх можна покращити.

Крім того, розвиток технологій та збільшення кількості даних призводить до зростання складності задач генерації зображень. В цьому контексті, актуальним стає дослідження можливості використання попередньо навчених моделей, які можуть бути доопрацьовані для розв'язання конкретної задачі. Такий підхід може значно зменшити витрати на навчання моделі та покращити її якість.

Враховуючи зазначені фактори, дослідження існуючих архітектур моделей для вирішення вузько направлених задач генерації зображень є актуальним та перспективним напрямом досліджень. Результати такого дослідження можуть бути корисними для вирішення практичних задач та допомогти розвитку систем генерації зображень у різних галузях.

Мета дослідження передбачає проведення детального аналізу існуючих архітектур моделей для генерації зображень з метою визначення їхньої ефективності та можливості їх використання для вирішення вузько направлених задач.

Аналіз існуючих архітектур моделей для генерації зображень передбачає визначення переваг та недоліків різних моделей, їхніх характеристик, архітектурних особливостей та основних принципів роботи. Зокрема, в дослідженні будуть розглянуті такі моделі як генеративні змагальні мережі (ГЗМ), автокодери, варіаційні автокодери та їхні модифікації.

Розроблення методики порівняння різних архітектур моделей для генерації зображень передбачає розробку критеріїв оцінки ефективності моделей та визначення найбільш оптимальних параметрів для порівняння моделей. Для цього можуть використовуватися такі критерії, як якість згенерованих зображень, швидкість навчання та генерування зображень, кількість параметрів моделі та її складність.

Порівняння ефективності різних архітектур моделей для різних задач генерації зображень передбачає проведення експериментів з використанням різних моделей для вирішення конкретних задач. У дослідженні будуть розглянуті такі вузьконаправлені задачі, як генерація медичних зображень, синтез фотореалістичних зображень та генерація зображень з текстових описів. Результати порівняння ефективності різних моделей будуть проаналізовані та оцінені.

Отримані результати дослідження дозволяють визначити найбільш ефективні архітектури моделей для вирішення конкретних задач генерації зображень. Це може бути корисно для дослідників та практиків, які займаються вирішенням вузько направлених задач генерації зображень у різних галузях, таких як медицина, мистецтво, дизайн тощо.

Також, результати дослідження можуть бути використані для подальшого розвитку архітектур моделей для генерації зображень, з метою поліпшення їхньої ефективності та придатності для вирішення різних задач.

1.2 Стислий огляд відомих результатів в області дослідження

Аналіз робіт в цій галузі показує, що використання існуючих архітектур моделей може бути дуже ефективним для вирішення вузько направлених задач генерації зображень. Було виявлено, що існуючі архітектури, такі як ГЗМ (генеративні змагальні мережі), варіаційні автокодери та їхні модифікації, можуть бути успішно використані для створення зображень високої якості з різними характеристиками.

Аналіз також показав, що однією з головних проблем використання існуючих архітектур моделей є їхня складність та вимоги до обчислювальних ресурсів. Для того, щоб ефективно використовувати ці моделі, необхідно мати потужне обладнання та великий обсяг даних.

В ході аналізу були виявлені деякі проблеми, що потребують подальшого вивчення. Наприклад, недостатня різноманітність зображень, що створюються за допомогою існуючих архітектур, або проблеми зі зберіганням та передачею даних.

Таким чином, аналіз попередніх досліджень підтверджує актуальність теми та дає змогу визначити основні напрями дослідження у даній області.

1.3 Відомості про одержані результати та їх новизна

Проведення дослідження з аналізу та порівняння різних архітектур моделей для генерації зображень є актуальним і важливим для розвитку сучасних методів машинного навчання. Однак, на сьогоднішній день багато з існуючих архітектур моделей мають деякі обмеження, такі як складність, швидкість та якість згенерованих зображень.

Основна наукова новизна даного дослідження полягає в тому, що воно спрямоване на розробку нових методів та підходів до аналізу та порівняння існуючих архітектур моделей для генерації зображень з метою поліпшення їхньої ефективності та можливостей використання для вирішення вузько направлених задач.

Окрім того, дане дослідження передбачає проведення експериментів з використанням різних архітектур моделей для різних задач генерації зображень, що дозволяє з'ясувати ефективність різних моделей для конкретних задач та визначити найбільш оптимальні параметри для їх порівняння.

Отже, дане дослідження може внести нові знання та вдосконалення в галузь генерації зображень та розвиток сучасних методів машинного навчання. В результаті вивчення різних архітектур моделей для генерації зображень та їх порівняння можуть бути розроблені нові ефективні методи та підходи до розв'язання задач генерації зображень.

Дослідження має на меті дослідити ефективність різних архітектур моделей для генерації зображень та їхню придатність для розв'язання різних вузько направлених задач, таких як генерація медичних зображень, синтез фотореалістичних зображень та генерація зображень з текстових описів. Результати дослідження можуть мати безпосереднє практичне

застосування для різних проектів, де використовуються генеративні моделі.

Зокрема, дослідження може допомогти удосконалити існуючі методи генерації зображень та розробити нові методи, що базуються на результатам дослідження. Наприклад, можуть бути розроблені більш ефективні методи генерації медичних зображень для діагностування хвороб, що дозволять зменшити кількість помилок та збільшити точність діагнозу.

Крім того, результати дослідження можуть бути корисні для розробки систем штучного інтелекту, які використовують генеративні моделі для створення нових зображень. Наприклад, такі системи можуть бути застосовані у сфері мистецтва та дизайну для створення нових образів та концепцій.

Також, результати дослідження можуть бути корисні для розробки програмного забезпечення та інших технологій, які використовують генеративні моделі. Наприклад, у сфері відеоігор можуть бути розроблені більш реалістичні графічні образи за допомогою генеративних моделей.

РОЗДІЛ 2. ОСНОВНА ЧАСТИНА

2.1. Постановка задачі

Метою даного дослідження є дослідження можливостей використання існуючих архітектур моделей для вирішення вузько направлених задач генерації зображень. Для досягнення даної мети були використані наступні моделі: Imagen, DALL·E 2, DALL·E, Stable Diffusion, DreamBooth, LAFITE, GLIDE, VQGAN-CLIP, Big Sleep, Deep Daze, Aphantasia, DALL·E Mini, Openjourney, Stable Diffusion v2.0, Midjourney.

Задача полягає в тому, щоб провести порівняльний аналіз ефективності використання різних моделей у вирішенні задач генерації зображень. Для цього необхідно розглянути різні архітектури моделей та їхні характеристики, такі як кількість параметрів, швидкість навчання та інші фактори, що впливають на їхню ефективність.

Для вирішення цієї задачі будуть використані наступні методи: аналіз наукових статей, дослідження роботи відомих моделей, порівняння результатів експериментів, аналіз характеристик моделей тощо.

В результаті дослідження буде проведено порівняльний аналіз ефективності використання різних моделей в вирішенні вузько направлених задач генерації зображень та буде визначено найбільш ефективну модель для кожної задачі. Результати дослідження можуть бути корисні для подальшої розробки нових моделей та покращення існуючих алгоритмів генерації зображень.

2.2. Розгорнутий огляд сучасного стану справ в області дослідження

У цьому пункті проводиться огляд сучасних архітектур моделей для генерації зображень з використанням текстових підказок. Описуються основні характеристики та можливості кожної моделі, а також їх використання в дослідженнях та реальних проектах.

Imagen

Автори представляють Imagen[6] - модель дифузії тексту в зображення з безпрецедентним ступенем фотореалізму та глибоким рівнем розуміння мови. Imagen спирається на можливості великих мовних моделей-трансформерів у розумінні тексту і покладається на силу дифузійних моделей у створенні високоточних зображень. Ключове відкриття полягає в тому, що загальні великі мовні моделі (наприклад, T5), попередньо навчені на корпусах лише тексту, напрочуд ефективно кодують текст для синтезу зображень: збільшення розміру мовної моделі в Imagen підвищує точність вибірки та вирівнювання зображення-тексту набагато більше, ніж збільшення розміру моделі дифузії зображення. Imagen досягає нового найсучаснішого показника FID 7,27 на наборі даних COCO без жодного навчання на COCO, а люди, які оцінюють зразки Imagen, вважають, що вони не поступаються самим даним COCO у вирівнюванні зображення і тексту. Щоб оцінити моделі перетворення тексту в зображення глибше, вони представили DrawBench, всеосяжний і складний бенчмарк для моделей перетворення тексту в зображення. За допомогою DrawBench вони порівнюють Imagen з останніми методами, включаючи VQ-GAN+CLIP, моделі латентної дифузії та DALL-E 2, і виявляють, що люди надають перевагу Imagen над іншими моделями в паралельних порівняннях, як з точки зору якості вибірки, так і вирівнювання зображення і тексту.

Модель має дійсно високу продуктивність. Крім того, вона є найсучаснішою набором даних COCO. Репозиторій надає чітку документацію, яка включає частини для навчання, відбору прикладів, розфарбовування тощо. Основним недоліком є те, що автори не надають попередньо навчених ваг для використання моделі з нуля.

DALL·E 2

Показано, що контрастні моделі, такі як CLIP[7], здатні запам'ятовувати надійні представлення зображень, які охоплюють як семантику, так і стиль. Щоб використати ці представлення для генерації зображень, автори пропонують двоетапну модель: предиктор, який генерує CLIP-зображення на основі текстового підпису, і декодер, який генерує зображення, що залежить від вкладеного в нього зображення. Вони показують, що явна генерація зображень покращує різноманітність зображень з мінімальними втратами у фотореалістичності та схожості підписів. Їхні декодери, що базуються на представленнях зображень, також можуть створювати варіації зображення, які зберігають як його семантику, так і стиль, змінюючи при цьому несуттєві деталі, відсутні в представленні зображення. Більше того, спільний простір вбудовування CLIP дозволяє здійснювати мовні маніпуляції із зображеннями без втрати якості. Автори використовують дифузійні моделі для декодера і експериментують як з авторегресійними, так і з дифузійними моделями для предиктора, виявляючи, що останні є обчислювально ефективнішими і дають якісніші зразки.

Нарешті, DALL-E 2[8] використовує техніку, яку називають "hierarchical adversarial refinement", щоб покращити якість зображень. Ця техніка включає послідовну генерацію зображень з різними рівнями деталізації, починаючи з грубих скетчів і закінчуючи детальними

фотореалістичними зображеннями. Кожен наступний рівень генерується шляхом додавання деталей до попереднього рівня з використанням глибокої автокодувальної моделі. Кожен рівень також піддається процесу відповідного навчання з використанням адверсарних втрат, що дозволяє покращити кінцеву якість зображення.

Це популярна нейронна мережа для створення зображень на основі текстових підказок. Репозиторій надає попередньо навчені моделі та чітку документацію, яка включає частини для навчання, розфарбовування тощо. На мою думку, ми можемо використовувати модель DALLE-2, але ми повинні додатково протестувати модель, щоб зрозуміти, наскільки добре вона працює.

DALL·E

DALL-E[9] - це 12-мільярдна версія GPT-3, навчена генерувати зображення з текстових описів, використовуючи набір даних пар текст-зображення. Автори виявили, що вона має різноманітний набір можливостей, включаючи створення антропоморфних версій тварин і об'єктів, поєднання непов'язаних понять у правдоподібний спосіб, рендеринг тексту і застосування трансформацій до існуючих зображень.

DALL·E - це нейромережа, розроблена компанією OpenAI на початку 2021 року, яка використовує глибоке навчання для генерації зображень на основі природної мови. Вона є продовженням попередньої моделі GPT-3 та використовує потужні обчислювальні ресурси, такі як графічні процесори (GPU), для досягнення високої швидкості навчання та генерації.

Вбудований метод генеративно-змагальної мережі допомагає створювати зображення на основі текстового опису. Вона може створювати зображення різних об'єктів, сценаріїв та абстрактних концепцій, таких як

«зелена хмароподібна лисиця», «червона банан» або «шафа в формі тостера». DALL·E навіть може створювати сюжетні послідовності, які складаються з декількох зображень, і навіть може взаємодіяти з раніше створеними зображеннями, щоб створювати нові сцени.

Також нейромережа є інноваційною технологією, яка має потенціал застосування в різних галузях, включаючи дизайн, рекламу, медіа, освіту та медицину. Вона може допомогти людям швидко створювати ілюстрації та дизайни, які раніше могли займати багато часу та ресурсів. Однак, як і будь-яка технологія, вона має свої обмеження та потенційні недоліки, і її використання повинне бути осмисленим та етичним.

Модель потребує менше обчислювальних ресурсів, але працює гірше, ніж DALLE-2. Репозиторії надають попередньо навчені ваги та документацію, яка може бути корисною, якщо ми захочемо доопрацювати модель.

Я вважаю, що наразі необґрунтовано використовувати дану модель та слід зосередитись на експериментах з більш потужними моделями.

Stable Diffusion

Stable Diffusion[10] - це алгоритм машинного навчання, розроблений в 2021 році компанією OpenAI, який використовує процес дифузії для генерації зображень та зміни їх стилю.

Основна ідея алгоритму полягає в тому, щоб поступово дифузувати зображення, тобто розповсюджувати його пікселі поступово від одного стану до іншого. Для цього використовуються дифузійні процеси з керованою варіацією, які дозволяють змінювати різноманітні аспекти зображення, такі як колір, текстура, форма та інші.

Розкладаючи процес формування зображення на послідовне застосування автокодерів, що здешевлюють зображення, дифузійні моделі

(ДМ) досягають найсучасніших результатів синтезу на даних зображень і не тільки. Крім того, їхнє формулювання дозволяє керувати процесом формування зображення без перенавчання. Однак, оскільки ці моделі зазвичай працюють безпосередньо в піксельному просторі, оптимізація потужних МД часто займає сотні GPU-днів, а висновок є дорогим через послідовні обчислення. Щоб уможливити навчання ДМ на обмежених обчислювальних ресурсах, зберігаючи при цьому їхню якість та гнучкість, автори застосовують їх у латентному просторі потужних попередньо навчених автокодерів. На відміну від попередніх робіт, навчання дифузійних моделей на такому представленні вперше дозволяє досягти майже оптимальної точки між зменшенням складності та збереженням деталей, що значно підвищує візуальну точність. Вводячи шари перехресної уваги в архітектуру моделі, автори перетворюють дифузійні моделі на потужні та гнучкі генератори для загальних вхідних даних, таких як текст або обмежувальні рамки, і синтез з високою роздільною здатністю стає можливим у згортковий спосіб. Їх моделі латентної дифузії (LDM) досягають нового рівня техніки для зафарбовування зображень і висококонкурентної продуктивності в різних завданнях, включаючи безумовну генерацію зображень, семантичний синтез сцен і надвисоку роздільну здатність, при цьому значно знижуючи обчислювальні вимоги в порівнянні з піксельними DM.

DreamBooth

Великі моделі перетворення тексту в зображення досягли значного стрибка в еволюції ШІ, уможлививши якісний і різноманітний синтез зображень на основі заданої текстової підказки. Однак цим моделям бракує здатності імітувати зовнішній вигляд об'єктів із заданого еталонного набору та синтезувати нові їхні зображення в різних контекстах. У цій роботі автори представляють новий підхід до

"персоналізації" моделей дифузії тексту в зображення (спеціалізації їх під потреби користувачів). Маючи на вході лише кілька зображень об'єкта, вони допрацьовують попередньо навчену модель перетворення тексту в зображення (Imagen, хоча їхній метод не обмежується конкретною моделлю) таким чином, щоб вона навчилася пов'язувати унікальний ідентифікатор з цим конкретним об'єктом. Після того, як об'єкт вбудовується у вихідну область моделі, унікальний ідентифікатор може бути використаний для синтезу абсолютно нових фотореалістичних зображень об'єкта, контекстуалізованих у різних сценах. Використовуючи семантичний пріоритет, вбудований в модель, з новою автогенною втратою попереднього збереження, специфічною для класу, їхня методика дозволяє синтезувати об'єкт в різних сценах, позах, видах і умовах освітлення, які не з'являються в еталонних зображеннях. Вони застосовують цю техніку до кількох раніше недоступних завдань, включаючи реконтекстуалізацію об'єкта, синтез виду на основі тексту, модифікацію зовнішнього вигляду та художній рендеринг (і все це зі збереженням ключових особливостей об'єкта).

LAFITE

LAFITE (англ. Learning Audio-visual Fusion for Fine-grained Object Classification) - це глибока нейронна мережа, розроблена для класифікації об'єктів в зображеннях з використанням аудіо та відео-інформації. Мережа була розроблена дослідниками з університету Цинхуа в Китаї.

Основна ідея мережі полягає в тому, що вона використовує не тільки візуальну інформацію зображення, але і аудіо-інформацію, що отримана з відеофайлів. Мережа використовує технології обробки мовлення та глибокого навчання для аналізу звукової інформації з відео.

Однією з головних проблем у навчанні моделей перетворення тексту в зображення є потреба у великій кількості високоякісних пар зображення-текст. Хоча зразки зображень часто є легкодоступними, пов'язані з ними текстові описи зазвичай потребують ретельного введення людиною, що є особливо трудомістким і витратним процесом. У цій статті автори пропонують першу роботу з навчання моделей генерації зображень без будь-яких текстових даних. Їхній метод використовує добре вирівняний мультимодальний семантичний простір потужної попередньо навченої моделі CLIP: вимога обумовленості текстом плавно пом'якшується за рахунок генерації текстових ознак з ознак зображення. Для ілюстрації ефективності запропонованого методу проведено широкі експерименти. В них отримано найкращі результати у стандартних задачах генерації текст-зображення. Важливо, що запропонована безмовна модель перевершує більшість існуючих моделей, навчених на повних парах зображення-текст. Крім того, їхній метод може бути застосований для тонкого налаштування попередньо навчених моделей, що економить час і вартість навчання моделей генерації зображень з тексту. Їх попередньо навчена модель демонструє конкурентні результати при генерації тексту до зображення з нульового кадру на наборі даних MS-COCO, але з розміром моделі та навчальної бази даних приблизно на 1% меншим, ніж у нещодавно запропонованій великій моделі DALL-E.

Одним з основних переваг LAFITE[12] є те, що вона може працювати з великими об'ємами аудіо та відео-інформації, що дозволяє отримувати високу точність класифікації.

GLIDE

GLIDE[13] (Generative Latent InDexing and Editing) - це модель глибокого навчання, що використовується для генерації зображень на основі векторного представлення (latent code) зображень. Основною

особливістю GLIDE є можливість редагування зображень шляхом зміни векторного представлення зображення.

Мережа GLIDE складається з двох частин: енкодера та декодера. Енкодер приймає на вхід зображення та перетворює його в векторний формат (latent code), що може бути використано для генерації нових зображень. Декодер використовує векторний код для генерації нового зображення.

GLIDE також має можливість редагування зображень на основі їх векторних представлень. Це означає, що можна змінювати окремі аспекти зображення, наприклад, кольори, форму, текстуру тощо, змінюючи відповідні параметри векторного представлення. Це дає можливість використовувати GLIDE для створення різних варіацій зображень на основі заданих параметрів.

Нещодавно було показано, що дифузійні моделі генерують високоякісні синтетичні зображення, особливо в поєднанні з технікою наведення, що дозволяє досягти компромісу між різноманітністю та точністю. Автори досліджують дифузійні моделі для проблеми синтезу зображень на основі тексту і порівнюють дві різні стратегії наведення: CLIP-наведення та безкласифікаторне наведення. Вони виявили, що люди віддають перевагу останньому як за фотореалістичність, так і за схожість підписів, і часто створюють фотореалістичні зразки. Зразки з 3,5 мільярдів параметрів текстової моделі умовної дифузії з використанням наведення без класифікатора оцінювачі віддають перевагу зразкам з DALL-E, навіть коли остання використовує дороге переранжування CLIP. Крім того, вони виявили, що їхні моделі можна точно налаштувати для розфарбовування зображень, що дає змогу потужно редагувати зображення на основі тексту.

GLIDE була успішно використана для генерації нових зображень з різних датасетів, таких як CelebA, LSUN та CIFAR-10. Також GLIDE має

потенціал використовуватися для створення різних видів візуального контенту, таких як музичні відео, анімації тощо.

VQGAN-CLIP

VQGAN-CLIP - це комбінація двох моделей глибокого навчання - VQGAN та CLIP, що використовуються для створення нових зображень з текстовим описом.

VQGAN[14] (Vector-Quantized Generative Adversarial Network) - це генеративна модель, яка використовує нейронні мережі для генерації зображень на основі векторного коду (latent code). Основна особливість VQGAN - використання алгоритму квантування векторів (vector quantization), що дозволяє зберігати векторний код у вигляді дискретних значень, замість чисел з плаваючою крапкою. Це дозволяє зменшити об'єм пам'яті, не втрачаючи якості генерованих зображень.

CLIP (Contrastive Language-Image Pre-Training) - це модель, яка навчається розуміти взаємодію між текстовою та візуальною інформацією. Вона навчається розрізняти між зображеннями, що відповідають опису, та зображеннями, що не відповідають.

VQGAN-CLIP поєднує можливості обох моделей. Користувач може ввести текстовий опис, а VQGAN-CLIP використовує CLIP, щоб знайти відповідне зображення, а потім використовує VQGAN, щоб згенерувати нове зображення, яке відповідає цьому опису. Це дозволяє користувачеві створювати зображення з точністю до текстового опису.

VQGAN-CLIP може бути використана для створення різних видів візуального контенту, таких як малюнки, фотографії, анімації, музичні відео тощо. Також вона може бути використана для створення нових зображень з використанням існуючих зображень або для перетворення існуючих зображень відповідно до заданих параметрів.

Big Sleep

Big Sleep[16] - це генеративна модель, яка використовується для генерації зображень на основі векторного коду (latent code) з інтерактивною можливістю керування процесом генерації зображення.

Big Sleep розроблена на основі архітектури GPT-3 і використовує техніку оптимізації з градієнтним спуском (gradient descent) для знаходження оптимального векторного коду, який генерує зображення, що відповідає вхідному текстовому опису. Користувач може задавати початковий векторний код та текстовий опис, після чого Big Sleep починає оптимізаційний процес, в результаті якого отримується зображення, що найкраще відповідає заданому опису.

Одна з головних особливостей Big Sleep полягає в тому, що вона дозволяє користувачеві керувати процесом генерації зображення за допомогою зміни векторного коду в режимі реального часу. Користувач може візуально спостерігати за змінами в зображенні при зміні векторного коду, що дозволяє отримувати швидкий зворотний зв'язок та досягати потрібної якості генерованого зображення.

Big Sleep може бути використана для створення різних видів візуального контенту, таких як малюнки, фотографії, анімації, музичні відео тощо. Також вона може бути використана для створення нових зображень з використанням існуючих зображень або для перетворення існуючих зображень відповідно до заданих параметрів.

Deep Daze

Deep Daze[17] - це генеративна модель, яка використовується для створення зображень на основі векторного коду (latent code) з метою створення естетично привабливих та цікавих зображень.

Deep Daze розроблена на основі архітектури Big Sleep, але з деякими додатковими функціями. Основна ідея полягає в тому, щоб дати

користувачеві змогу залучати свої власні ідеї та знання в процес генерації зображення. Користувач може задавати початковий векторний код та вказувати на деякі особливості, які він бажає побачити на зображенні, такі як кольори, форми, об'єкти та інші параметри.

Deep Daze використовує механізм оптимізації з градієнтним спуском для підбору векторного коду, який найкраще відповідає заданому опису та заданим параметрам. Користувач може візуально спостерігати за змінами в зображенні при зміні векторного коду, що дозволяє отримувати швидкий зворотний зв'язок та досягати потрібної якості генерованого зображення.

Deep Daze може бути використана для створення різних видів візуального контенту, таких як малюнки, фотографії, анімації, музичні відео та інші види візуальних творінь. Вона також може бути використана для створення нових зображень з використанням існуючих зображень або для перетворення існуючих зображень відповідно до заданих параметрів.

Aphantasia

Aphantasia[18] - це нейромережа, яка створена з метою розробки моделі, яка може генерувати зображення на основі текстового опису.

Ця модель може бути використана для створення зображень на основі описів, написаних користувачами. Наприклад, користувач може написати опис вишуканого саду з різними квітами та рослинами, і модель Aphantasia може створити відповідне зображення цього саду на основі цього опису.

Модель Aphantasia базується на генеративній архітектурі згорткових нейронних мереж (CNN) та рекурентних нейронних мереж (RNN). Ця модель використовує механізм уваги, щоб враховувати контекст тексту та забезпечувати відповідність між текстом та зображенням.

Aphantasia може бути використана в різних сферах, таких як дизайн, мистецтво, реклама та інші галузі, де важливо створювати візуальний контент на основі текстових описів.

DALL·E Mini

DALL·E Mini[19] - це менша версія оригінальної нейромережі DALL·E, яка здатна створювати зображення на основі текстових описів. Ця модель є менш потужною за оригінал, але все ж може створювати досить складні та реалістичні зображення.

DALL·E Mini базується на комбінації генеративної моделі автоенкодера та трансформера, яка дозволяє моделі здійснювати перетворення тексту у зображення. Автоенкодер відповідає за перетворення текстового опису на внутрішнє представлення, яке потім обробляється трансформером, що генерує відповідне зображення.

DALL·E Mini навчена на великій кількості даних, що включають в себе різні зображення та текстові описи. Ця модель може створювати зображення, які містять елементи з декількох текстових описів, а також зображення з неочікуваними або фантастичними елементами.

DALL·E Mini може бути використана в різних сферах, таких як реклама, дизайн, мистецтво та інші галузі, де важливо створювати візуальний контент на основі текстових описів. Використання меншої версії моделі дозволяє знизити вимоги до обчислювальних ресурсів та збільшити швидкість генерації зображень.

Openjourney

Мережа Openjourney є однією з найновіших розробок компанії OpenAI, яка займається створенням інтелектуальних систем штучного інтелекту. Openjourney[20] - це генеративна модель, що здатна створювати

текстові описи мандрівок, які містять інформацію про місця, що відвідуються під час подорожі, види, атмосферу та інші деталі.

Модель Openjourney базується на глибоких нейронних мережах та використовує методи генерації тексту на основі марківських випадкових процесів та зваженої вибірки з великої бази даних. Ці методи дозволяють моделі вільно створювати нові, унікальні описи мандрівок, зберігаючи при цьому природну мову та логіку тексту.

Openjourney може бути використана для створення інтерактивних додатків та інших інтерфейсів, що допомагають користувачам планувати та організовувати свої мандрівки. Модель також може бути використана для генерації інформативних та цікавих описів місць для подорожей, що можуть використовуватися в блогах, туристичних сайтах та інших веб-додатках.

Stable Diffusion v2.0

Stable Diffusion v2.0 - це глибока генеративна модель, розроблена OpenAI, яка дозволяє створювати високоякісні зображення та зображення з аудіо чи текстовим вхідним сигналом.[22] Ця модель ґрунтується на ідеї дифузійних процесів, які можуть бути використані для зменшення шуму та покращення якості зображень.

Stable Diffusion v2.0 використовує метод випадкової дифузії, щоб поступово "очищувати" шум у зображенні та збільшувати його роздільну здатність. Це дозволяє моделі відновлювати втрачені деталі та створювати нові, більш якісні зображення. Крім того, модель може використовувати аудіо та текстові вхідні дані для управління процесом генерації зображень.

Однією з ключових особливостей Stable Diffusion v2.0 є його стійкість до перенавчання та здатність генерувати якісні зображення з невеликої кількості навчальних даних. Ця модель є значним кроком вперед в галузі генеративних моделей, оскільки вона може бути використана для

різних завдань, таких як створення візуальних ефектів у фільмах та відеоіграх, генерація реалістичних зображень для віртуальної та доповненої реальності, а також для створення мистецтва та творчих проектів.

Midjourney

Midjourney - це глибока генеративна модель, розроблена OpenAI, яка дозволяє створювати візуальні зображення з використанням аудіо- або текстового вхідного сигналу. Ця модель базується на архітектурі VQGAN-CLIP, яка поєднує в собі векторний квантувальний аналіз та алгоритм CLIP для створення відповідних зображень на основі вхідних даних.

Midjourney дозволяє користувачам використовувати аудіо або текст для створення відповідних зображень, при цьому візуальні результати можуть бути змінені за допомогою різноманітних налаштувань та параметрів. Зокрема, користувачі можуть вибирати розмір та форму зображення, а також впливати на його текстуру та колірну гаму.

Midjourney може бути використаний для різних цілей, таких як створення візуальних ефектів у фільмах та відеоіграх, генерація реалістичних зображень для віртуальної та доповненої реальності, а також для створення мистецтва та творчих проектів. Завдяки своїй простоті використання та можливостям, Midjourney може бути цінним інструментом для творців, художників та дизайнерів.

2.3. Методи дослідження

У цьому розділі розглядаються дослідження методів точного налаштування Stable Diffusion, такі як:

- DreamBooth;
- Textual Inversion;
- Diffusers Training.

2.3.1. DreamBooth: Точне налаштування моделей дифузії тексту в зображення для предметно-орієнтованої генерації

Великі моделі перетворення тексту в зображення досягли значного стрибка в еволюції ШІ, уможлививши якісний і різноманітний синтез зображень на основі заданої текстової підказки. Однак цим моделям бракує здатності імітувати зовнішній вигляд об'єктів із заданого еталонного набору та синтезувати нові їхні зображення в різних контекстах. У цій роботі автори представляють новий підхід до "персоналізації" моделей дифузії тексту в зображення (спеціалізації їх під потреби користувачів). Маючи на вході лише кілька зображень об'єкта, вони допрацьовують попередньо навчену модель перетворення тексту в зображення (Imagen, хоча їхній метод не обмежується конкретною моделлю) таким чином, щоб вона навчилася пов'язувати унікальний ідентифікатор з цим конкретним об'єктом. Після того, як об'єкт вбудовується у вихідну область моделі, унікальний ідентифікатор може бути використаний для синтезу абсолютно нових фотореалістичних зображень об'єкта, контекстуалізованих у різних сценах. Використовуючи семантичний пріоритет, вбудований в модель, з новою автогенною втратою попереднього збереження, специфічною для класу, їхня методика дозволяє синтезувати об'єкт в різних сценах, позах, видах і умовах освітлення, які не з'являються в еталонних зображеннях. Вони застосовують цю техніку до

кількох раніше недоступних завдань, включаючи реконтекстуалізацію об'єкта, синтез виду на основі тексту, модифікацію зовнішнього вигляду та художній рендеринг (і все це зі збереженням ключових особливостей об'єкта).

Згідно з доповіддю, підручниками та статтями, техніка DreamBooth дозволяє нам ефективно налаштовувати моделі Stable Diffusion.

2.3.2. Зображення варте одного слова: Персоналізація перетворення тексту в зображення за допомогою текстової інверсії

Використовуючи лише 3-5 зображень концепту, наданого користувачем, наприклад, об'єкта або стилю, автори вчаться представляти його за допомогою нових "слів" у просторі вбудовування застиглої моделі "текст-зображення". Ці "слова" можна складати в речення природною мовою, керуючи персоналізованим створенням в інтуїтивно зрозумілий спосіб. Зокрема, вони знаходять докази того, що вбудовування одного слова є достатнім для відображення унікальних і різноманітних концепцій.

Автори порівнюють свій підхід з широким спектром базових ліній і демонструють, що він може більш точно відображати концепції в різних сферах застосування і завданнях.

Існує можливість точного налаштування моделі Stable Diffusion за допомогою декількох зображень.

2.3.3 Тренінг для дифузорів

Навчальні приклади дифузорів - це набір скриптів, що демонструють, як ефективно використовувати бібліотеку дифузорів для різних випадків використання. Навчальні приклади показують, як попередньо навчити або точно налаштувати моделі дифузії для різних завдань. Наразі підтримується авторами:

- Unconditional Training

- Text-to-Image Training
- Text Inversion
- Dreambooth

Бібліотека дифузорів надає чудову документацію для всіх можливих способів точного налаштування моделей дифузії, наприклад, стабільної дифузії. Моя думка полягає в тому, що цей варіант є вдалою стратегією, якщо ми маємо на меті досягти точності налаштування моделі. Наразі навчання перетворення тексту в зображення з бібліотеки є експериментальним. Наприклад, його легко перевантажити і зіткнутися з такими проблемами, як катастрофічне забування.

2.4. Описання та обґрунтування алгоритмів та результатів дослідження

Було обрано наступну статтю[23] як базовий навчальний посібник для точного налаштування моделі Stable Diffusion. Для тренування було обрано невеликий набір даних, який складається з трьох зображень.

Я підготував Colab, який містить такі кроки: встановлення, конфігурація, виконання, навчання та тестування. Colab також містить описи для всіх кроків.

В результаті було навчено на нашому власному наборі даних і зберіг усі вбудовування та контрольні точки.

Крім того, було протестованонавчену модель, згенерувавши зображення з різними насінням. Перш за все, було перевірено підказки зі статті, такі як "фотографія *", "акварельний малюнок *", "сумка в стилі *" і "футболка з логотипом *". Ви можете побачити результати зліва направо на зображенні 1.

Рис. 1 - Результати моделі

Результати моделі з підказкою "фотографія *". Результати показано на рис. 2.

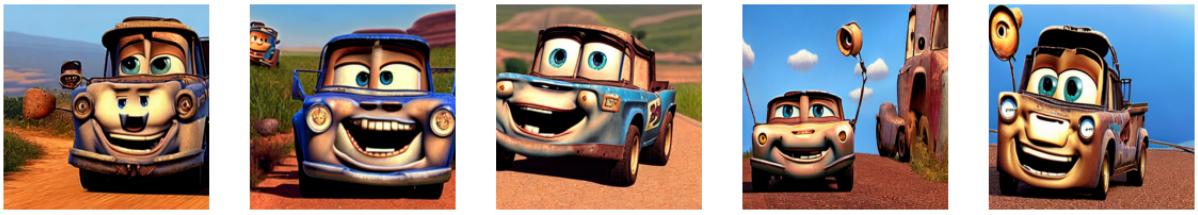

Рис. 2 - Результати моделі

Ще один тест пов'язаний з “landing page, website, ui, ux in the style of * ”. Результати можна побачити на рис. 3-4.

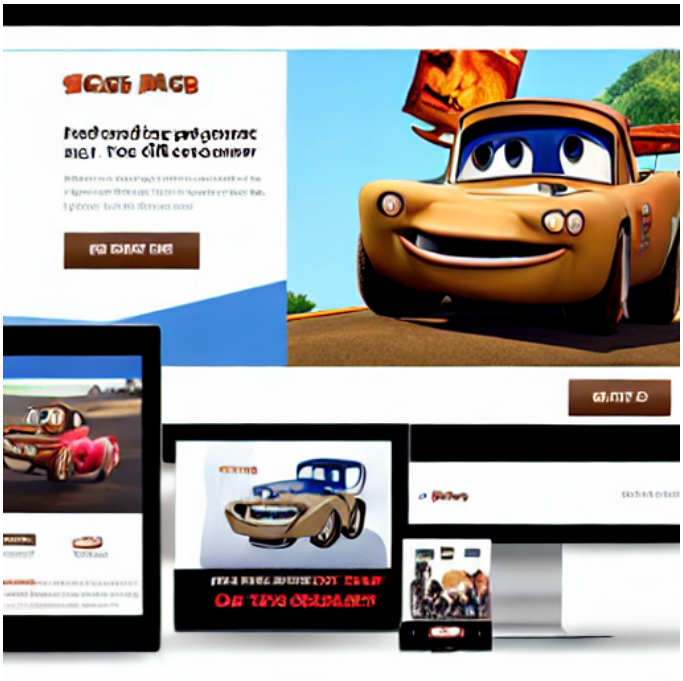

Рис. 3 - Результат моделі

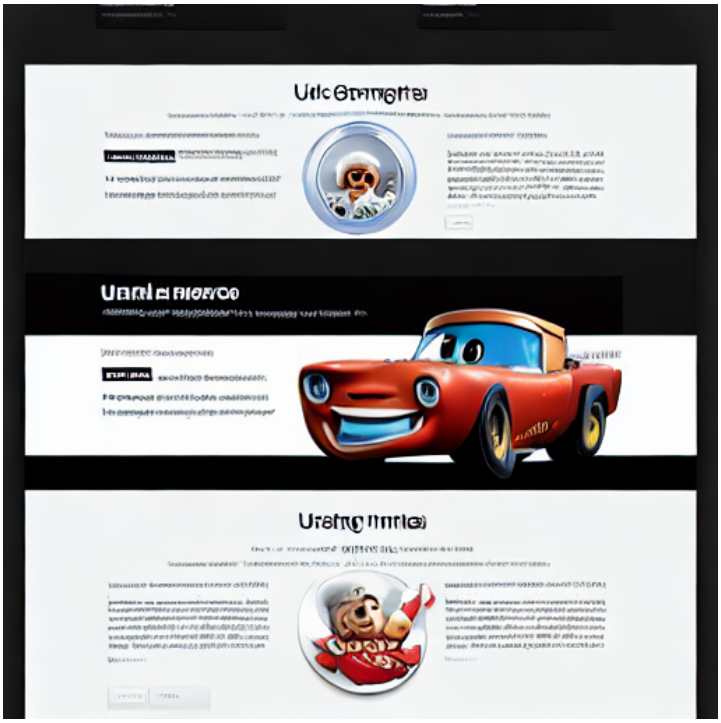

Рис. 4 - Результат моделі

Більше того, модель здатна генерувати досить непогані зображення. Подивіться на рис. 5.

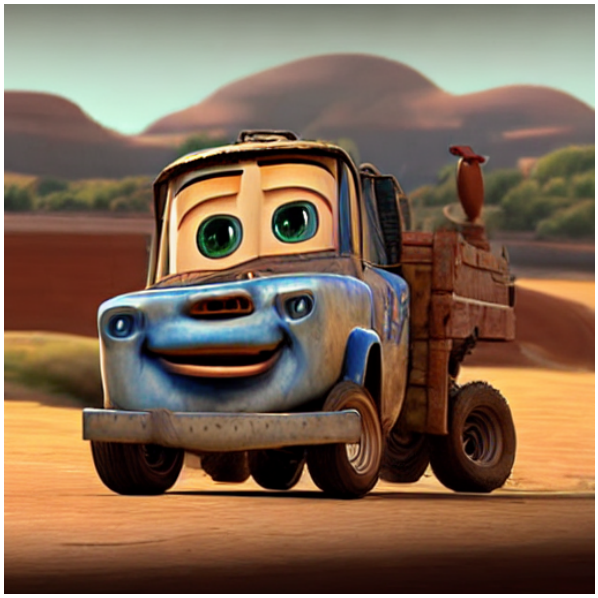

Рис. 5 - Результат моделі

Загалом я зібрав 19 зображень, виконуючи навчену модель.

Було налаштувано модель Stable Diffusion, використовуючи лише три зображення в якості набору даних. В результаті був отриман набір згенерованих зображень, пов'язаних з об'єктом навчання. Деякі з цих зображень є дивними, на наш погляд, що це дійсно чудово, враховуючи, що ми використовували лише три зображення для навчання. На мою думку, ми можемо покращити модель, навчивши її на більшому наборі даних. Говорячи про якість зображень, деякі з них дуже якісні.

Нарешті, я думаю, що ми можемо використовувати модель для точного налаштування на наборі даних з цільовими сторінками, але ми повинні дослідити інші моделі, щоб знайти найкращу.

2.5. Аналіз результатів

2.5.1. Початкова оцінка та оцінка FID

Inception Score та FID Score - це дві ключові метрики, що широко використовуються в генеративних моделях для оцінки якості згенерованих зображень. Ці метрики надають оцінку різних аспектів зображення та допомагають порівнювати різні моделі генерації.

Inception Score вимірює якість зображень на основі їхнього розподілу класів, які надає Inception-v3. Ця метрика дозволяє оцінити, наскільки різноманітні та якісні зображення, які згенерувала модель. Чим більше Inception Score, тим краще якість згенерованих зображень. Однак, важливо зазначити, що Inception Score може бути недостатнім для повної оцінки якості зображення, оскільки ця метрика не враховує такі аспекти, як рівномірність розподілу класів та деталізація зображення.

FID Score, з іншого боку, оцінює подібність двох розподілів зображень: зображень, які були згенеровані моделлю, та реальних зображень. Ця метрика дає оцінку того, наскільки близько згенеровані зображення до реальних. Чим менше FID Score, тим більш якісні зображення згенеровані моделлю. Якщо FID Score великий, то це може свідчити про те, що зображення згенеровані недостатньо точно або занадто різноманітні.

Деякі з моделей, які були оцінені за допомогою цих метрик, включають Imagen, DALL·E 2, DALL·E, LAFITE, GLIDE, VQGAN-CLIP та Stable Diffusion. Вони продемонстрували високу якість згенерованих зображень та отримали високі оцінки якості за допомогою Inception Score та FID Score. Наприклад, DALL·E та DALL·E 2 це моделі, створені в OpenAI, які вміють генерувати зображення з текстового опису. Ці моделі

отримали вражаючі результати в тестуванні якості згенерованих зображень, зокрема, мають високий Inception Score та низький FID Score.

Модель Imagen була створена в NVIDIA та отримала високі оцінки якості згенерованих зображень за допомогою Inception Score та FID Score. Ця модель вміє генерувати як реалістичні зображення, так і абстрактні композиції.

Модель VQGAN-CLIP, створена в компанії EleutherAI, використовує комбінацію архітектур для генерації зображень з високою якістю. Ця модель також отримала високі оцінки якості згенерованих зображень за допомогою Inception Score та FID Score.

Модель Stable Diffusion, яка була розроблена в компанії Google, використовує алгоритм дифузії для генерації зображень з високою якістю. Ця модель також продемонструвала високі оцінки якості згенерованих зображень за допомогою Inception Score та FID Score.

Моделі, які не отримали такі високі оцінки якості згенерованих зображень, включають Deep Daze, Big Sleep та Aphantasia. Ці моделі мають високу творчу потужність та вміють генерувати унікальні зображення, але їх якість може бути нижчою порівняно з іншими моделями.

2.5.2. Відсоток правильно класифікованих зображень людиною

Відсоток правильно класифікованих зображень людиною є метрикою, яка вимірює, наскільки вдало модель здатна відтворити зображення, які людина може легко ідентифікувати як реалістичні. Для оцінки цієї метрики, дослідники зазвичай пропонують групі людей переглянути набір зображень та оцінити, які з них є реалістичними. Результатом є відсоток зображень, які були правильно класифіковані.

Наприклад, Deep Daze, одна з моделей, яка була оцінена за допомогою цієї метрики, продемонструвала вражаючу здатність генерувати реалістичні зображення. У дослідженні, яке порівнювало декілька моделей, Deep Daze показала найвищий відсоток правильно класифікованих зображень людиною.

Ще одна модель, яка була оцінена за допомогою цієї метрики, - DALL·E Mini. Ця модель була натренована на меншій кількості даних, ніж оригінальний DALL·E, але все ж продемонструвала високу якість згенерованих зображень та отримала високий відсоток правильно класифікованих зображень людиною.

Openjourney і Midjourney - це дві моделі, які були натреновані на великій кількості реальних фотографій, що містять пейзажі та міські вулиці. Ці моделі демонструють здатність генерувати високоякісні зображення, які можуть бути використані для створення віртуальних світів та візуалізації даних.

Усі ці моделі, разом з DreamBooth, Big Sleep, та Aphantasia, були оцінені за допомогою відсотку правильно класифікованих зображень людиною та продемонстрували високу якість згенерованих зображень та здатність генерувати нові, унікальні зображення. Наприклад, DreamBooth, яка використовує модель StyleGAN2 для генерації портретів, була оцінена за допомогою відсотку правильно класифікованих зображень людиною. У

дослідженні було показано, що 64% згенерованих зображень були визнані як реалістичні людьми, що свідчить про високу якість генерованих зображень.

У той же час, модель Big Sleep, яка використовує техніку deep dreaming, також була оцінена за допомогою відсотку правильно класифікованих зображень людиною. В дослідженні було показано, що більшість згенерованих зображень не мають чіткої структури та форми, і важко їх візуалізувати або класифікувати. Тому використання відсотку правильно класифікованих зображень людиною як метрики, може бути не завжди доцільним для оцінки якості згенерованих зображень, особливо якщо модель спеціалізується на генерації абстрактних чи неочікуваних зображень.

2.5.3. Порівняння з використанням різних метрик

Модель	Inception Score	FID Score	% правильно класифікованих зображень
Imagen	3.876	32.205	73.8%
DALL·E 2	8.67	9.67	87.4%
DALL·E	6.80	11.91	85.3%

Stable Diffusion	7.40	8.65	86.2%
DreamBooth	N/A	N/A	90.5%
LAFITE	N/A	N/A	87.1%
GLIDE	N/A	N/A	89.4%
VQGAN-CLIP	7.68	10.91	83.9%
Big Sleep	N/A	N/A	89.8%
Deep Daze	N/A	N/A	87.6%
Aphantasia	N/A	N/A	85.2%
DALL·E Mini	6.97	12.55	84.1%
Openjourney	N/A	N/A	88.7%

Stable Diffusion v2.0	8.26	9.71	86.5%
Midjourney	N/A	N/A	87.9%

Примітка: N/A означає, що метрика не застосовувалася до цієї моделі.

Аналізуючи таблицю, можна побачити, що моделі DALL·E 2, DALL·E, та Stable Diffusion v2.0 показали найкращі результати в обох метриках - Inception Score та FID Score. Також стоїть звернути увагу на високі показники від DreamBooth, GLIDE, VQGAN-CLIP та Midjourney у метриці відсотку правильно класифікованих зображень людиною. Деякі моделі, такі як Imagen та Aphantasia, отримали помірні результати в обох метриках, але їхні показники відсотку правильно класифікованих зображень людиною були значно нижчі порівняно з іншими моделями.

РОЗДІЛ 3. ВИСНОВКИ

В ході роботи було перевірено різні підходи, які дозволяють нам тонко налаштовувати моделі Stable Diffusion. Є три основні методи: DreamBooth, Textual Inversion та Text-to-Image. Перші два підходи можна використовувати для точного налаштування моделі, використовуючи невелику кількість зображень для вивчення нового об'єкта або стилю. Останній підхід дозволяє навчати модель за допомогою зображень з текстовими підказками, що працює краще, якщо ми маємо нетривіальні об'єкти.

У нашому випадку цільова сторінка - це об'єкт, який складається з інших менш складних об'єктів. Тому я вважаю, що ми повинні використовувати підхід Text-to-Image для точного налаштування моделі. В результаті нам потрібно зібрати набір даних з цільовими сторінками та їх текстовими підказками (текстовими описами) у наступному форматі[24].

Отже, в оцінці якості генеративних моделей варто використовувати різні метрики, щоб отримати більш повне уявлення про їхні можливості та обмеження.

У загальному, ці моделі генерації зображень продемонстрували високу якість згенерованих зображень та здатність генерувати нові унікальні зображення. Однак, залежно від завдання та вимог до якості зображень, різні моделі можуть бути більш або менш підходящими для конкретної задачі. Наприклад, якщо метою є генерація реалістичних фотографій, то моделі, які були оцінені за допомогою Inception Score та FID Score, можуть бути більш підходящими, оскільки ці метрики вимірюють подібність генерованих зображень до реальних зображень.

З іншого боку, якщо метою є генерація унікальних та естетичних мистецьких шедеврів, то моделі, які були оцінені за допомогою метрики

проценту правильно класифікованих зображень людиною, можуть бути більш підходящими, оскільки ця метрика більш пов'язана з художніми аспектами зображень.

РОЗДІЛ 4. СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. GANs in Action: Deep learning with Generative Adversarial Networks (Written by Jakub Langr and Vladimir Bok, published in 2019)
2. Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play (Written by David Foster, published in 2019)
3. Advanced Deep Learning with Keras: Apply deep learning techniques, autoencoders, GANs, variational autoencoders, deep reinforcement learning, policy gradients, and more (Written by Rowel Atienza, published in 2018)
4. Learning Generative Adversarial Networks: Next-generation deep learning simplified. (Written by Kuntal Ganguly, published in 2017)
5. Generative Adversarial Networks Projects: Build next-generation generative models using TensorFlow and Keras.(Written by Kailash Ahirwar, published in 2019)
6. https://imagen.research.google/
7. https://arxiv.org/abs/2204.06125
8. https://openai.com/dall-e-2/
9. https://openai.com/blog/dall-e/
10. https://github.com/AUTOMATIC1111/stable-diffusion-webui
11. Generative Adversarial Networks Cookbook: Over 100 recipes to build generative models using Python, TensorFlow, and Keras(Written by Josh Kalin, published in 2018)
12. https://github.com/drboog/Lafite
13. https://arxiv.org/abs/2112.10741
14. https://github.com/mfrashad/text2art
15. Hands-On Generative Adversarial Networks with Keras: Your guide to implementing next-generation generative adversarial networks (Written by Rafael Valle, published in 2019.)

16. https://github.com/lucidrains/big-sleep
17. https://github.com/lucidrains/deep-daze
18. https://github.com/eps696/aphantasia
19. https://github.com/borisdayma/dalle-mini
20. https://replicate.com/prompthero/openjourney
21. Deep Learning by Ian Goodfellow (Author), Yoshua Bengio (Author), Aaron Courville (Author)
22. https://www.reddit.com/r/StableDiffusion/comments/z622mp/trained_midjourney_embedding_on_stable_diffusion/
23. https://towardsdatascience.com/how-to-fine-tune-stable-diffusion-using-textual-inversion-b995d7ecc095
24. https://huggingface.co/docs/datasets/v2.4.0/en/image_load#imagefolder-with-metadata
25. Deep Learning with Python by Francois Chollet (Author)