УДК 004.93

ВИКОРИСТАННЯ МЕТОДУ LORA ДЛЯ АДАПТАЦІЇ МОДЕЛЕЙ STABLE DIFFUSION

Ішу А.О.
e-mail: anastasiia.ishu@nure.ua
Науковий керівник – ст. викл. Бобнєв Р.О.
Харківський національний університет радіоелектроніки, каф. МІРЕС
м. Харків, Україна

The article is devoted to the LoRA (Low-Rank Adaptation) method for adapting Stable Diffusion models. The basic principles of this work are reviewed, looking at the mechanism for making low-level changes in the parameters of the model to change the payment costs. The key advantages of LoRA are analyzed, such as efficiency, flexibility and ease of integration, as well as possible shortcomings associated with it.

Актуальність проблеми. Stable Diffusion – потужна генеративна модель, яка створює високоякісні зображення на основі текстових описів.

Адаптація моделі LORA до конкретних завдань, таких як створення зображень з певним стилем або специфічними особливостями, може вимагати додаткового навчання. Ці моделі глибокого навчання достатньо великі за обсягом. Розмір файлу може становити кілька гігабайт. І перенавчання моделі, тобто оновлення багатьох вагових коефіцієнтів, є складним завданням. Наприклад, іноді необхідно модифікувати модель стабільної дифузії, щоб визначити нові інтерпретації підказок або змусити модель генерувати різні стилі малювання за замовчуванням [1].

Сучасні генеративні моделі часто використовують великі, попередньо навчені моделі, які дають високоякісні результати. Однак адаптація таких моделей до конкретних завдань і даних вимагає значних обчислювальних ресурсів і часу на навчання.

Мета роботи. Одним з найбільш перспективних методів вирішення цих проблем є низькорангова адаптація (Low-Rank Adaptation, LoRA) – технологія, яка дозволяє ефективно адаптувати великомасштабні моделі з мінімальними додатковими витратами на навчання.

Виклад основного матеріалу. Використання даного методу для адаптації зі стабільною дифузією може значно скоротити час навчання та обчислювальні ресурси, роблячи процес персоналізації моделі швидшим та доступнішим. LoRA вносить невеликі зміни до шару перехресної уваги, найважливішої частини моделі стабільної дифузії. Дослідники виявили, що для ефективного навчання вистачить налаштувати саме цю частину, де зустрічаються образи та підказки. Шар перехресної уваги – це жовта частина архітектури моделі стабільної дифузії (частини QKV предиктора шуму U-Net) [2], що зображена нижче на рисунку 1. Він містить у собі вагові коефіцієнти, що розміщуються в матрицях.

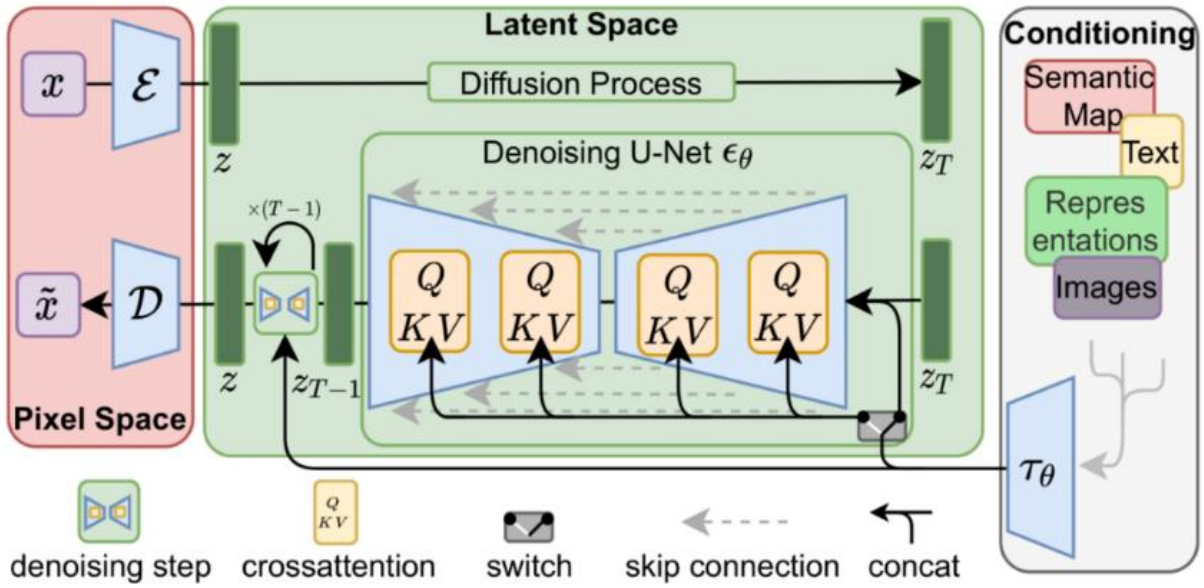

Рисунок 1 – Етапи роботи стабільної дифузії [3]

Метод зменшує матрицю до необхідного розміру шляхом перестановки ваг шарів перехресної уваги в матриці і таким чином згортає матрицю в дві зменшені сітки низького рівня. Це значно зменшує кількість файлів моделі, які потрібно зберігати, і зменшує розмір моделі LoRA.

Дана технологія має декілька переваг:

–ефективне використання ресурсів шляхом адаптації моделі без повного перенавчання забезпечує зменшення часу на навчання і зниження використання обчислювальних ресурсів;

–мінімізація перенавчання отримується завдяки низькорівневих адаптацій, що зберігає високу якість генерації при обмежених чи рідкісних наборах даних;

–LoRA може бути застосована до різних типів моделей, що робить її універсальним інструментом в багатьох сферах: від обробки тексту до генерації зображень;

–технологія легко інтегрується з іншими методами та архітектурами нейромереж, зокрема з популярними генеративними моделями.

Попри численні переваги, LoRA має й певні обмеження.

По-перше, ефективність методу може знижуватися при роботі з надто великими або складними даними, де низькорівневі адаптації можуть не забезпечити достатнього покращення результатів.

По-друге, метод може бути менш ефективним у випадках, коли необхідно змінити значну частину моделі, а не лише окремі її частини.

Крім того, процес налаштування LoRA вимагає глибоких знань про структуру моделі та точну настройку її параметрів.

Висновки. Метод LoRA є потужним інструментом завдяки своїй здатності знижувати обчислювальні витрати та зберігати високу ефективність. Хоча він має певні обмеження, його застосування в генеративних моделях відкриває нові можливості для персоналізації та оптимізації процесів генерації зображень в багатьох сферах нашого життя. Тому даний підхід буде розвиватися і використовуватися в майбутньому.

Список використаних джерел:

1. Mehreen K. Using LoRA in Stable Diffusion. Machine learning mastery. URL: https://machinelearningmastery.com/using-lora-in-stable-diffusion/ (дата звернення: 01.03.2025).

2. Andrew. What are LoRA models and how to use them in AUTOMATIC1111. Stable Diffusion Art. URL: https://stable-diffusion-art.com/lora/ (дата звернення: 01.03.2025).

3. High-resolution image synthesis with latent diffusion models / R. Rombach та ін. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022. С. 10684–10695. DOI: https://doi.org/10.48550/arXiv.2112.10752 (дата звернення: 01.03.2025).