УДК 004.85 DOI: https://doi.org/10.30837/IYF.IIS.2024.408

ТЕНДЕНЦІЇ РОЗВИТКУ ГЕНЕРАЦІЇ ТЕКСТОВОГО ОПИСУ ЗОБРАЖЕНЬ

Голобородько Б. Ю.
Науковий керівник – д.т.н., проф. Смеляков К. С.
Харківський національний університет радіоелектроніки, каф. ПІ
м. Харків, Україна
e-mail: bohdan.holoborodko@nure.ua

This paper covers advanced approaches and challenges in generating textual descriptions for images. Modern models, such as BLIP, CLIP, and others that use self-learning, contrastive learning, and noise robustness to improve the accuracy and reliability of descriptions are considered. The main trends and problems of the development of textual descriptions for images are viewed.

У сфері штучного інтелекту генерація текстового опису до зображень стали важливою технологією, яка зменшує розрив між візуальним контентом і його лінгвістичною інтерпретацією. Це технологія, яка використовує синергію комп'ютерного зору та обробки природної мови і спрямована на автоматичне створення текстових описів зображень. Подібне завдання вимагає не лише розпізнавання об’єктів на зображенні, але й розуміння їхнього контексту та взаємодії. На практиці, дана проблема виходить за межі базового розуміння зображень, пропонуючи значні переваги в різних сферах де відбувається взаємодія з візуальною інформацією, наприклад: цифрова доступність веб сайтів для людей з вадами зору, більш ефективне управління графічним контентом, E-Commerce, соціальні мережі, ШІ-асистенти і т.д.

Застосуванні генерації текстового опису до зображень викликає питання щодо точності, надійності та достовірності даних. Варіативність якості зображень, різноманітність сцен і тонкощі людської мови додають до цього завдання ще більше складнощів. Ці складнощі пов'язані з необхідністю точно розуміти і описувати вміст зображення у контекстно-релевантний і лінгвістично зв'язний спосіб. Такі моделі, як BLIP (Bootstrapped Language Image Pretraining) і CLIP (Contrastive Language-Image Pretraining), а також їхні наступники, такі як NLIP (Noise-robust Language-Image Pretraining), вирішують ці проблеми за допомогою інноваційних підходів. Розглянемо дані моделі.

BLIP використовує метод самонавчання, який дозволяє моделі покращувати своє розуміння зображень та асоційованого з ними тексту шляхом взаємного навчання між модулями обробки зображень та тексту [1]. Це дозволяє створювати більш точні та контекстуально релевантні описи.

CLIP впроваджує контрастивне навчання, зв'язуючи візуальний та текстовий контент через велику кількість прикладів. Ідея такого підходу

полягає в тому, щоб модель вчилася розрізняти "позитивні" (де текст правильно описує зображення) та "негативні" (де текст не відповідає зображенню) пари. Це дозволяє моделі встановлювати глибші зв'язки між зображеннями та їх текстовими описами, покращуючи здатність до узагальнення та розуміння контексту.

NLIP ще більше розширює підходи, впроваджуючи механізми стійкості до шуму в даних [2]. Це дозволяє моделі бути більш стійкою до помилок у вхідних даних та підвищує точність генерації описів у складних умовах.

Окрім загального розуміння контексту, сучасним моделям генерації тексту до зображень все ще важко узагальнювати різні домени або типи зображень. Моделі на основі BLIP та CLIP попередньо навчаються на різноманітних зображеннях та текстах з Інтернету, що допомагає їм краще працювати в різних доменах, не потребуючи специфічних навчальних даних. Попереднє навчання моделі на загальних даних відображає ще один тренд у розвитку Image Captioning. Широке попереднє навчання дозволяє краще справлятися з неоднозначністю, використовуючи величезну кількість вивчених візуальних і текстових даних, щоб робити більш обґрунтовані припущення про неоднозначний вміст зображення.

Значною перевагою CLIP та його наступників є їхня здатність до навчання з нуля, коли модель може точно підписувати зображення без попереднього навчання на подібних прикладах. Це особливо корисно для створення підписів до нових або рідкісних зображень.

Ще одна тенденція, яку варто відзначити – це розвиток багатомодальних моделей, які можуть одночасно обробляти та аналізувати інформацію з різних джерел, забезпечуючи більш глибоке розуміння контенту.

Також, варто зазначити що продовжується активна інтеграція глибокого навчання і нейронних мереж: стрімкий розвиток глибокого навчання і конволюційних нейронних мереж (CNN) для обробки зображень, разом з послідовними нейронними мережами (RNN) або трансформерами для обробки тексту, значно підвищує якість генерації текстових описів.

Основні проблеми включають упередження в навчальних даних, високі вимоги до обчислювальних ресурсів, та складнощі інтерпретації генерованих описів [3]. Упередження в даних може призвести до некоректних описів, що викликає етичні та соціальні проблеми. Великі обсяги даних та складні моделі вимагають значних обчислювальних потужностей, обмежуючи доступність технології. Також існує ризик того, що без належного контролю та аналізу використання методів штучного інтелекту [4], можуть виникнути непередбачувані результати.

Ще однією проблемою є динамічність контенту. Багато зображень, особливо в соціальних мережах, є динамічними або містять елементи, що

швидко застарівають. Системи генерації опису повинні бути здатними швидко адаптуватися до змін у візуальному контенті та вміти розрізняти нові контексти.

У підсумку, хоча сучасні моделі генерації текстових описів зображень демонструють значно краще розуміння візуального контенту, вони одночасно стикаються з потребою в значно більших обсягах даних для навчання та збільшених обчислювальних ресурсах. Це ставить ефективність моделей та їх здатність до генерації найбільш оптимального опису зображень на передній план. Таким чином, подальші дослідження та розробки повинні зосередитися на оптимізації цих аспектів для забезпечення ширшого впровадження та ефективності в реальних сценаріях застосування.

Список використаних джерел

1. Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. arXiv:2201.12086.

2. Huang, R., Long, Y., Han, J., Xu, H., Liang, X., Xu, C., & Liang, X. (2023, June). Nlip: Noise-robust language-image pre-training. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 1, pp. 926-934).

3. Xu, L., Tang, Q., Lv, J., Zheng, B., Zeng, X., & Li, W. (2023). Deep Image Captioning: A Review of Methods, Trends and Future Challenges. Neurocomputing, 126287.

4. Шарун Д. А. Методи штучного інтелекту в системах прийняття рішень і управління / Д. А. Шарун // Радіоелектроніка та молодь у XXI столітті: зб. матеріалів 27-го Міжнар. молодіжн. форуму, 10–12 травня 2023 р. – Харків: ХНУРЕ, 2023. – Т. 6, Ч. ІІ. (конф. «Інформаційні інтелектуальні системи»). – С. 192–193.