Генеративний фреймворк для побудови візуально-текстових датасетів на основі онтологій

Чоловський С.О., Здирко В.В.. / Cholovskyi S. Zdyrko V.
Національний університет "Києво-Могилянська Академія" / National University of Kyiv-Mohyla Academy
04655, Київ, вул. Григорія Сковороди, 2, факультет інформатики, кафедра інформатики
E-mail: s.cholovskyi@ukma.edu.ua, v.zdyrko@ukma.edu.ua

This work presents a general framework for generating VQA (Visual Question Answering) datasets across arbitrary knowledge domains. Logically complex questions are derived from OWL-ready formatted ontologies, and correct answers are obtained using SPARQL queries. The diversity of generated questions is enhanced through paraphrasing with a large language model. Relevant scenes are generated using Stable Diffusion with CLIP-score–based post-filtering. We believe this hybrid approach enables efficient creation of high-quality, semantically rich datasets.

Одним з напрямків комп'ютерного зору є візуально-лінгвістична обробка зображень, він включає в себе, зокрема, такі задачі як опис(captioning), відповіді на питання за зображенням(далі VQA - visual question answering), та загальне розуміння(visual common sense reasoning). Задача VQA полягає в тому, щоб на основі зображення надати коротку відповідь на розгорнуте питання (в common sense задачах відповідь може бути довільною). Постановка задачі VQA є більш простою ніж у задачі опису зображень яка полягає в тому, щоб з заданим зображенням будувати його текстовий опис. Збір даних для задачі Visual Question Answering (VQA) є складним і ресурсоємним процесом, оскільки вимагає ретельно узгоджених наборів зображень, запитань і відповідей. Одним із перспективних напрямів часткової автоматизації цього процесу є використання онтологічних моделей та методів логічного виведення, що забезпечує можливість автоматичного формування запитань із наперед визначеними правильними відповідями. Для підвищення варіативності як змісту запитань, так і візуальних сцен, доцільним є застосування генеративних нейронних мереж, здатних синтезувати нові текстові формулювання запитань та створювати синтетичні зображення, які містять релевантну інформацію для відповіді.

Ми пропонуємо підхід що дозволяє автоматично будувати VQA -датасети на основі онтологій. За основу взято підхід описаний в[1]. Для побудови питань онтологія має відповідати формату owlready2. Першим кроком є побудова табличного представлення онтології, для чого обробляється ієрархія класів, кортежі відношень та пари об'єктів. Формування списку словників сутність відношення з ієрархії, для кожного рівня ієрархії. Перетворення словника в таблицю з пяти стовпців класи об'єктів, класи атрибутів, сутності, атрибути та відношення. Для генерації запитань із таблиці випадковим чином обираються концепти, ролі та сутності, щоб сформувати блок речення. питання формуються шляхом поєднання блоків логічними операторами. Правильна відповідь формується на основі SPARQL запиту. Результатом є пара первинне питання, та набір атомарних сутностей, що є відповідями на це питання. Нехай Q — множина усіх можливих первинних питань, сформованих алгоритмом, а O — множина усіх об'єктів (сутностей) у розгорнутій онтології (базі знань). Результат роботи блоку R формалізується як множина пар:

$$R=\{(q,L)\vee q\in Q, L\subseteq O\}$$

Де:

- q — це **первинне питання** (string/текст), що сформоване за шаблоном(рис.1).
- L — це **множина об'єктів-відповідей** (set), де кожен об'єкт $o\in L$ задовольняє **усі логічні умови**, виражені у питанні q.

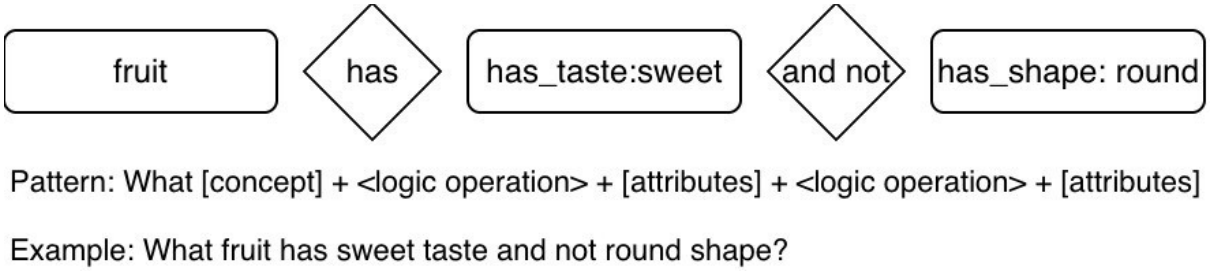

рисунок 1. Схема типового шаблона, із вибором випадкогово параметру з таблиці, з довільним логічним оператором. Приклад сформованого початкового питання.

Попри те, що автоматично згенеровані питання абсолютно коректні, вони не варіативні. Ми використовуємо велику мовну модель для перефразування первинного питання q і отримання q_{para}. Цей прийом вже успішно використовується з image captioning[3].Наведемо приклад згенерованих варіацій (для онтології з оригінальної статті). Первинне питання “Which food on photo has cylindtical shape and has nutricious fats” та його варіанти “Which food in the photo is cylindrical in shape and contains nutritious fats?” та “Which cylindrical food in the photo is rich in healthy fats?”.

Для генерації зображень ми використовуємо мережу StableDiffusion. Для вибору об’єктів, що мають бути на зображенні, з L випадковим чином обирається набір I**(included)** від одного до чотирьох атомарних об’єктів, решту об’єктів позначимо як E**(excluded)**. До набору I також включаємо зашумлюючі об`єкти $N \subseteq L^C$. Шаблон запиту генерації “{general scene description} with {objects from I}, {objects from N}.” Для фільтрації результатів ми відсікаємо CLIP-score згенерованого зображення з кожною з ключових частин нашого запиту та для усіх об’єктів з E. Таким чином забезпечується відповідність зображення домену та наявність відповіді при збереженні певної варіат ивності.

В результаті виконання отримуємо триплет: $(q_{para}, image, I)$ та проміжні результати (L, P, q, N, E). Для подальшого оцінювання описаним методом побудовано набір з 1000 зображень,питань та відповідей до них. Дані будуть оцінені експертами, а для зображень також будуть обчислені такі кількісні показники як FID(відстань сприйняття за Фреше), IS(inception score) та MUSIQ.

Таким чином, запропонований фреймворк включає в себе інструменти для генерації та розширення VQA-датасетів. Розділення етапів генерації дозволяє при потребі гнучко замінити їх необхідними для конкретної задачі. Наприклад, використання LLM дозволяє отримати запитання новою мовою, без зміни логіки побудови питань та базової онтології. Варто відмітити, що сучасний стан розвитку генеративних моделей накладає певні обмеження на варіативність отриманих даних. Якість перефразування деградує зі зростанням складності та довжини питань. Дифузійні моделі демонструють обмежену здатність до генерації складних сцен, тому було додано механізм, що забезпечує узгодженість згенерованих зображень із відповідними питаннями та відповідями.

Джерела

1. LoRA: a logical reasoning augmented dataset for visual question answering / Jingying Gao [et al.] // Proceedings of the 37th International Conference on Neural Information Processing Systems. – Red Hook, {NY}, {USA}, 2023. – P. 30579–30591.
2. VQA: Visual Question Answering [Electronic resource] / Aishwarya Agrawal [et al.]. – Mode of access: https://doi.org/10.48550/arXiv.1505.00468 (date of access: 02.11.2025). – Title from screen.
3. Xiao C. Multimodal Data Augmentation for Image Captioning using Diffusion Models [Electronic resource] / Changrong Xiao, Sean Xin Xu, Kunpeng Zhang // Proceedings of the 1st Workshop on Large Generative Models Meet Multimodal Applications. – [S. l.], 2023. – P. 23–33. – Mode of access: https://doi.org/10.1145/3607827.3616839 (date of access: 02.02.2024). – Title from screen.