Video2Agent: ІНТЕРАКТИВНИЙ АГЕНТ ДЛЯ ОБРОБКИ ВІДЕОКОНТЕНТУ НА ОСНОВІ СЕМАНТИЧНОГО ПОШУКУ

Хаханов І.В.
Харківський національний університет радіоелектроніки, Харків, Україна

Розвиток технологій обробки природної мови (Natural Language Processing, NLP), векторних баз даних та мультимодальних моделей створив передумови для формування нових підходів до аналізу та пошуку інформації у відеоконтенті.

Традиційні методи базуються на пошуку за ключовими словами або ручному перегляді транскриптів, що є неефективним для обробки великих відеокорпусів і не враховує візуальний контекст сцени.

У роботі досліджується проблема семантичного та візуально-контекстуального аналізу відео, спрямована на створення інтелектуальної системи, здатної здійснювати інтерактивну взаємодію з мультимедійним контентом на основі аналізу текстових і візуальних ознак.

Метою дослідження є розроблення AI-керованого чатбота, що забезпечує діалогову взаємодію користувача з YouTube-відео, використовуючи семантичний пошук у векторній базі даних та аналіз ключових кадрів відеоряду.

Запропонована система реалізує повний функціональний цикл обробки мультимедійного контенту, який включає:

1) автоматичне отримання транскрипту відео з підтримкою кількох мов (англійська, українська тощо);
2) сегментацію транскрипту на фіксовані семантичні чанки, що слугують основними одиницями пошуку;
3) виділення ключових кадрів для кожного чанка, що репрезентують візуальний контекст відповідного фрагмента;
4) векторизацію текстових і візуальних даних із використанням моделей OpenAI та CLIP-подібних підходів до побудови спільного векторного простору;
5) збереження мультимодальних вкладень у векторних базах даних (Pinecone або Milvus);
6) семантично-візуальний пошук, який дозволяє системі враховувати як зміст транскрипту, так і візуальну складову сцени;
7) генерацію контекстно обґрунтованих відповідей з використанням великої мовної моделі (LLM) OpenAI.

Система реалізована у вигляді двостороннього веб-додатку на основі Streamlit, який забезпечує динамічне відображення метаданих, візуалізацію кадрів і збереження історії діалогу для підтримки контекстної узгодженості відповідей.

Експериментальна апробація проведена на відео англійською та українською мовами.

Результати дослідження показали, що поєднання текстових і візуальних ознак:

8) підвищує точність і релевантність результатів семантичного пошуку;

9) забезпечує глибше розуміння контенту сцени, що відображає контекст транскрипту;

10) скорочує час отримання потрібної інформації у великому відеоконтенті;

11) підвищує когерентність і змістовність діалогової взаємодії користувача з системою.

Наукова новизна роботи полягає у комплексній інтеграції семантичного аналізу тексту, векторного представлення зображень і генеративних мовних моделей для створення мультимодальної системи семантичного пошуку у відео.

Отримані результати підтверджують ефективність мультимодального підходу до побудови інтелектуальних інтерфейсів взаємодії з відеоконтентом.

Подальші дослідження планується спрямувати на покращення якості векторних репрезентацій зображень, розширення підтримки мов і дослідження адаптації системи до потокового відео.

Список літератури

1. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT 2019. arXiv preprint arXiv:1810.04805. DOI: https://doi.org/10.48550/arXiv.1810.04805

2. Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense Passage Retrieval for Open-Domain Question Answering. In Proceedings of EMNLP 2020. arXiv preprint arXiv:2004.04906. DOI: https://doi.org/10.48550/arXiv.2004.04906

3. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., & Schwenk, H. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Advances in Neural Information Processing Systems (NeurIPS 2020). arXiv preprint arXiv:2005.11401. DOI: https://doi.org/10.48550/arXiv.2005.11401

4. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing (NeurIPS 2019). arXiv preprint arXiv:1910.01108. DOI: https://doi.org/10.48550/arXiv.1910.01108

5. Katharopoulos, A., Vyas, A., Pappas, N., & Fleuret, F. (2020). Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. In Proceedings of ICML 2020. arXiv preprint arXiv:2006.16236. DOI: https://doi.org/10.48550/arXiv.2006.16236

6. Pinecone, Inc. (2024). Pinecone Vector Database. https://www.pinecone.io/

7. OpenAI. (2024). OpenAI API Documentation. https://platform.openai.com/docs/api-reference