System koncepcyjnego etykietowania segmentów w nagraniach naukowych

Jedną z wielu form przekazywania wiedzy są materiały filmowe podczas których omawiane są koncepcje naukowe, często z wykorzystaniem prezentacji. Zazwyczaj jednak problemem jest zlokalizowanie najistotniejszych momentów w takim nagraniu. Na platformie YouTube możliwe jest zarówno manualne jak i automatyczne oznaczenie wybranych segmentów wraz z ich krótkim opisem [1]. Niestety nagrania, zwłaszcza w przypadku wykładów, mają nawet kilka godzin długości, stąd manualne ich oznaczenie wymaga sporego wkładu czasowego od twórców. Wersja automatyczna natomiast często nie pozwala na osiągnięcie zadowalającej jakości i dokładności proponowanych segmentów.

Pojawia się w związku z tym szansa na opracowania udoskonalonej metody etykietowania nagrań z wykorzystaniem uczenia maszynowego. Zadanie opisane w poprzednim akapicie znane jest w literaturze jako Dense Video Captioning [2] i dostępne są metody generujące predykcje na podstawie zarówno wizualnym jak i audio [3]. Większość metod skupia się jednak na oznaczaniu przedstawionego zachowania w materiale wideo, na przykład “mężczyzna gra na pianinie” lub “para osób tańczy na scenie”. Pożądana metoda charakteryzowałaby się wyższym poziomem ekstrakcji kontekstu zarówno na podstawie przedstawionych materiałów naukowych w wideo jak i w głosowym opisie zagadnień.

Celem projektu jest przygotowanie rozwiązania, które pozwoli na identyfikację i oznaczenie opisem słownym nienakładających się segmentów w naukowych materiałach wideo. Etykiety powinny w jasny i ogólny sposób przedstawiać omawianą koncepcję, na przykład “Generative Adversarial Network” czy “Equation describing loss function of YOLO architecture”. System powinien również umożliwiać wyszukiwanie semantyczne w wygenerowanych opisach – na podstawie badania podobieństwa pomiędzy zapytaniem a znalezionymi konceptami. Dodatkową funkcjonalnością systemu powinno być również wychwytywanie najważniejszych semantycznie klatek obrazu (wykresy, schematy, wzory itp.) i ich zapisywanie jako metadane dla plików wideo.

Literatura:

1. Lecture 1 | Introduction to Convolutional Neural Networks for Visual Recognition. Stanford University School of Engineering https://www.youtube.com/watch?v=vT1JzLTH4G4 [Dostęp 21 marca 2024].
2. Krishna R, Hata K, Ren F, Fei-Fei L, Carlos Niebles J. Dense-captioning events in videos. InProceedings of the IEEE international conference on computer vision 2017 (pp. 706-715)
3. Lashin V, Rahtu E. Multi-modal dense video captioning. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops 2020 (pp. 958-959).