Metody uczenia reprezentacji wielu obiektów na obrazach i materiałach wideo

mgr inż. Piotr Zieliński

Streszczenie

Uczenie reprezentacji wizualnej umożliwia wydobycie istotnych informacji ze scen zawartych na obrazach i materiałach wideo, co stanowi fundament wielu zastosowań w dziedzinie wizji komputerowej. Globalne metody uczenia reprezentacji zazwyczaj łączą wszystkie elementy sceny w jedno osadzenie, co utrudnia precyzyjne rozróżnianie poszczególnych obiektów. Uczenie reprezentacji wielu obiektów modeluje sceny jako zbiory odrębnych encji, co pozwala na ustrukturyzowaną analizę scen, podobną do ludzkiego sposobu ich postrzegania. Jest to kluczowe dla zadań takich jak rozumowanie wizualne, śledzenie obiektów czy robotyka.

Pomimo znaczących postępów, w tym obszarze nadal istnieje kilka wyzwań. Wczesne rozwiązania polegały na wnioskowaniu sekwencyjnym, co ograniczało skalowalność do złożonych wizualnie scen. Metody oparte na konwolucyjnych siatkach cech poprawiły efektywność wykrywania obiektów, jednak wciąż wymagały sekwencyjnego przetwarzania wycinków obiektów oraz nie potrafiły tworzyć reprezentacji przy zmiennych rozmiarach obiektów z powodu sztywnej rozdzielczości map cech. W pełni nienadzorowane podejścia często nie mogły niezawodnie rozróżniać poszczególnych obiektów w realistycznych scenach, łącząc wiele z nich jako jedno osadzenie. Rozszerzenia temporalne dla filmów wideo pogłębiły te wyzwania, dodatkowo zwiększając złożoność obliczeniową, co ogranicza ich praktyczne zastosowanie.

Niniejsza rozprawa podejmuje te wyzwania, integrując postępy z jednoetapowych architektur wykrywania obiektów z metodami uczenia reprezentacji wielu obiektów, co było zainspirowane badaniami nad zastosowaniem cech z modeli detekcji w nawigacji robotycznej z użyciem głębokich modeli uczenia ze wzmocnieniem. Zaproponowano metodę SSDIR, wykorzystującą wieloskalowe mapy cech w metodzie kodowania opartej na przestrzennej siatce cech, wykorzystując wstępnie wytrenowaną sieć do detekcji obiektów jako fundament do nienadzorowanego uczenia reprezentacji i precyzyjnego ustalania położenia obiektów w złożonych, rzeczywistych warunkach wizualnych. Metoda RDIR rozszerza model na wideo, wprowadzając architekturę rekurencyjną dla spójnych reprezentacji obiektów w kolejnych klatkach. Model dyfuzyjny DetDiff, warunkowany reprezentacjami wzbogaconymi przez detekcje, poprawia zdolności generatywne, umożliwiając kontrolowane tworzenie obrazów. Szerokie eksperymenty potwierdzają polepszenie jakość reprezentacji i ich efektywne zastosowanie w zdaniach docelowych, wykazując skuteczność i wszechstronność proponowanych podejść w różnych dziedzinach wizualnych.

Piotr Zieliński