Methods for multi-object representation learning in images and videos

mgr inż. Piotr Zieliński

Abstract

Visual representation learning facilitates the extraction of useful information from scenes presented in images and videos, which is essential for numerous computer vision applications. Global representation learning methods typically merge all scene elements into a single embedding, limiting the ability to distinguish individual objects clearly. Multi-object representation learning explicitly models scenes as collections of distinct entities, enabling structured, human-like understanding crucial for tasks such as visual reasoning, object tracking, and robotics.

Despite significant advancements, several challenges persist in this area. Early approaches relied on sequential inference, limiting scalability to visually complex scenes. Convolutional grid-based methods improved object discovery efficiency but still required sequential processing of object glimpses and were constrained by fixed spatial resolution, making it difficult to represent objects of varying sizes. Fully unsupervised models often fail to disentangle objects reliably in realistic settings, frequently encoding multiple objects as a single entity. Temporal extensions for videos further amplified these challenges, additionally exhibiting computational complexity, which restricts their practical applicability.

This thesis addresses these gaps by integrating advances from modern one-stage object detection architectures into multi-object representation learning frameworks, which was motivated by an investigation into using visual features from an object detection network to guide deep reinforcement learning in robotic navigation. It proposes SSDIR, a novel method leveraging multi-scale feature maps within a parallel spatial grid-based encoding strategy, using pre-trained object detectors as a robust foundation for unsupervised representation learning and precise object localisation in complex, real-world settings. Additionally, RDIR introduces an implicit temporal extension through a recurrent architecture, ensuring consistent object representations across video frames. Furthermore, a diffusion-based model, DetDiff, conditioned on detection-guided representations enhances generative quality, enabling controllable image synthesis. Extensive experiments confirm improvements in representation quality and their performance in downstream tasks, demonstrating the efficacy and versatility of the proposed approaches across diverse visual domains.

Piotr Zieliński