Object Tracking

I. Định nghĩa bài toán

Theo dõi đối tượng là nhiệm vụ lấy một tập hợp các phát hiện đối tượng ban đầu, tạo một ID duy nhất cho mỗi phát hiện ban đầu, sau đó theo dõi từng đối tượng khi chúng di chuyển xung quanh các khung hình trong video, duy trì việc gán ID.

- Input: video có các đối tượng được quan tâm
- Output: các id được gán cho các đối tượng trong toàn bộ video qua từng frame

II. Thách thức:

- Tốc độ đào tạo và theo dõi: Các thuật toán để theo dõi các đối tượng được cho là không chỉ thực hiện chính xác việc phát hiện và khoanh vùng các đối tượng quan tâm mà còn làm như vậy trong khoảng thời gian ít nhất có thể. Tăng cường tốc độ theo dõi đặc biệt cấp thiết đối với các mô hình theo dõi đối tượng thời gian thực. Để quản lý thời gian thực hiện một mô hình, thuật toán được sử dụng để tạo mô hình theo dõi đối tượng cần được tùy chỉnh hoặc lựa chọn cẩn thận. Fast R-CNN và Faster R-CNN có thể được sử dụng để tăng tốc độ của cách tiếp cận R-CNN phổ biến nhất. Vì CNN (Mạng nơ ron chuyển đổi) thường được sử dụng để phát hiện đối tượng, các sửa đổi của CNN có thể là yếu tố phân biệt giữa mô hình theo dõi đối tượng nhanh hơn và mô hình chậm hơn. Các lựa chọn thiết kế bên cạnh khung phát hiện cũng ảnh hưởng đến sự cân bằng giữa tốc độ và độ chính xác của mô hình phát hiện đối tượng.
- Sự phân tâm trong nền (Background Distractions**)**: Nền của hình ảnh được nhập vào hoặc hình ảnh được sử dụng để đào tạo mô hình theo dõi đối tượng cũng ảnh hưởng đến độ chính xác của mô hình. Nền bận rộn của các đối tượng cần theo dõi có thể khiến các đối tượng nhỏ khó bị phát hiện hơn. Với nền mờ hoặc một màu, hệ thống AI sẽ dễ dàng phát hiện và theo dõi các đối tượng hơn. Nền quá bận, có cùng màu với đối tượng hoặc quá lộn xộn có thể khiến khó theo dõi kết quả cho một đối tượng nhỏ hoặc một đối tượng có màu nhạt.
- Nhiều thang đo không gian (Multiple Spatial Scales**):** Các đối tượng được theo dõi có thể có nhiều kích thước và tỷ lệ khung hình. Các tỷ lệ này có thể khiến các thuật toán theo dõi đối tượng nhầm lẫn thành các đối tượng tin rằng được thu nhỏ lớn hơn hoặc nhỏ hơn kích thước thực của chúng. Nhận thức sai về kích thước có thể tác động tiêu

cực đến việc phát hiện hoặc tốc độ phát hiện. Để chống lại các vấn đề về tỷ lệ không gian khác nhau, các lập trình viên có thể thực hiện các kỹ thuật như bản đồ đặc trưng, hộp neo, kim tự tháp hình ảnh và kim tự tháp đặc trưng.

- Hộp neo (anchor box) : Hộp neo là tập hợp các hộp giới hạn có chiều cao và chiều rộng xác định. Các hộp có nghĩa là để thu được quy mô và tỷ lệ co của các đối tượng quan tâm. Chúng được chọn dựa trên kích thước đối tượng trung bình của các đối tượng trong một tập dữ liệu nhất định. Hộp neo cho phép phát hiện nhiều loại đối tượng khác nhau mà không cần các tọa độ hộp giới hạn xen kẽ trong quá trình bản địa hóa.
- Bản đồ đặc trưng (feature map): Bản đồ đặc trưng là hình ảnh đầu ra của một lớp khi Mạng thần kinh kết hợp (CNN) được sử dụng để ghi lại kết quả của việc áp dụng các bộ lọc cho hình ảnh đầu vào đó. Bản đồ đặc trưng cho phép hiểu sâu hơn về các đối tượng địa lý được CNN phát hiện. Máy dò một lần phải tính đến vấn đề nhiều thang đo vì chúng phát hiện các đối tượng chỉ bằng một lần đi qua khung CNN. Điều này sẽ xảy ra trong việc giảm khả năng phát hiện đối với các hình ảnh nhỏ. Các hình ảnh nhỏ có thể mất tín hiệu trong quá trình lấy mẫu xuống trong các lớp tổng hợp, đó là khi CNN được đào tạo về một tập hợp con thấp của các hình ảnh nhỏ hơn đó. Ngay cả khi số lượng đối tượng giống nhau, việc lấy mẫu giảm có thể xảy ra vì CNN không thể phát hiện các hình ảnh nhỏ và đếm chúng vào kích thước mẫu. Để ngăn chặn điều này, nhiều bản đồ đặc trưng có thể được sử dụng để cho phép máy dò ảnh chụp một lần tìm kiếm các đối tượng trong các lớp CNN - bao gồm các lớp trước đó với hình ảnh có độ phân giải cao hơn. Máy dò một lần vẫn không phải là một lựa chọn lý tưởng để theo dõi vật thể nhỏ vì chúng gặp khó khăn khi phát hiện vật thể nhỏ. Việc phân nhóm chặt chẽ có thể đặc biệt khó khăn. Ví dụ, các bức ảnh chụp bằng máy bay không người lái từ trên cao về một nhóm động vật bầy đàn sẽ khó theo dõi bằng cách sử dụng máy dò một lần.
- Hình ảnh và đại diện tính năng kim tự tháp (Image and Feature Pyramid Representations) : Kim tự tháp nổi bật, còn được gọi là bản đồ đối tượng nhiều cấp vì cấu trúc hình chóp của chúng, là một giải pháp sơ bộ cho sự thay đổi tỷ lệ đối tượng khi sử dụng bộ dữ liệu theo dõi đối tượng. Do đó, các kim

tự tháp đặc trưng mô hình hóa thông tin hữu ích nhất liên quan đến các đối tượng có kích thước khác nhau trong một biểu diễn từ trên xuống và do đó giúp phát hiện các đối tượng có kích thước khác nhau dễ dàng hơn. Các chiến lược như kim tự tháp hình ảnh và kim tự tháp đặc trưng rất hữu ích để ngăn ngừa các vấn đề về tỷ lệ. Kim tự tháp đối tượng dựa trên bản đồ đối tượng nhiều tỷ lệ, sử dụng ít năng lượng tính toán hơn so với kim tự tháp hình ảnh. Điều này là do các kim tự tháp hình ảnh bao gồm một tập hợp các phiên bản đã thay đổi kích thước của một hình ảnh đầu vào, sau đó được gửi đến máy dò khi thử nghiệm.

III. Các phương pháp tiếp cận Object Tracking:

1. *Video tracking (Multi-Object Tracking):*

   Theo dõi video là một ứng dụng theo dõi đối tượng, nơi các đối tượng chuyển động nằm trong thông tin video. Do đó, các hệ thống theo dõi video có thể xử lý cảnh quay trực tiếp, thời gian thực và cả các tệp video đã ghi. Các quy trình được sử dụng để thực hiện các tác vụ theo dõi video khác nhau dựa trên loại đầu vào video nào được nhắm mục tiêu. Các ứng dụng theo dõi video khác nhau đóng một vai trò quan trọng trong phân tích video, hiểu cảnh cho an ninh, quân sự, giao thông vận tải và các ngành công nghiệp khác. Ngày nay, một loạt các ứng dụng học tập sâu và thị giác máy tính thời gian thực sử dụng các phương pháp theo dõi video.

2. *Visual tracking:*

   Theo dõi trực quan hoặc theo dõi mục tiêu trực quan là một chủ đề nghiên cứu về thị giác máy tính được áp dụng trong một loạt các tình huống hàng ngày. Mục tiêu của theo dõi trực quan là ước tính vị trí trong tương lai của mục tiêu trực quan đã được khởi tạo mà không có phần còn lại của video.

3. *Object tracking camera (Online Multi-Object Tracking):*

   Các phương pháp theo dõi đối tượng hiện đại có thể được áp dụng cho các luồng video thời gian thực về cơ bản của bất kỳ máy ảnh nào. Do đó, nguồn cấp dữ liệu video của camera USB hoặc camera IP có thể được sử dụng để thực hiện theo dõi đối tượng, bằng cách cung cấp các khung hình riêng lẻ cho một thuật toán theo

dõi. Bỏ qua khung hình hoặc xử lý song song là các phương pháp phổ biến để cải thiện hiệu suất theo dõi đối tượng với nguồn cấp dữ liệu video thời gian thực của một hoặc nhiều máy ảnh.

IV. 3 SOTA của các phương pháp tiếp cận object tracking:

1. *Video tracking:*
   - Dataset: MOT16
   - Paper:
     - TransMOT: Spatial-Temporal Graph Transformer for Multiple Object Tracking
     - FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking
     - Do Different Tracking Tasks Require Different Appearance Models?
2. *Visual tracking:*
   - Dataset: TrackingNet
   - Paper:
     - SwinTrack: A Simple and Strong Baseline for Transformer Tracking
     - Learning Spatio-Temporal Transformer for Visual Tracking
     - Siam R-CNN: Visual Tracking by Re-Detection
3. *Object tracking camera:*
   - Dataset: MOT16
   - Paper:
     - Track to Detect and Segment: An Online Multi-Object Tracker
     - Tracking without bells and whistles
     - Real-time Multiple People Tracking with Deeply Learned Candidate Selection and Person Re-Identification