Video Object Detection and Tracking using the Attention Mechanism: A Comparative Study

Jeremy Reist
University of Toronto
jeremy.reist@mail.utoronto.ca

Travis Shao
University of Toronto
travis.shao@mail.utoronto.ca

Abstract

Video object detection and tracking are critical tasks in computer vision, with applications spanning across various domains, including autonomous vehicles, robotics, and surveillance. While recent advances in deep learning have significantly improved object detection performance, challenges remain in accurately detecting and tracking objects in video sequences. This study investigates the effectiveness of the DETR (DEtection TRansformer) architecture and its variants, Deformable DETR and TrackFormer, for video object detection and tracking. We evaluate the performance of these models on the car subset of the YouTube-Objects dataset and the MOT20-07 video, using mean Intersection over Union (MIoU) and mean Average Precision (mAP) metrics. Our results demonstrate that TrackFormer outperforms both DETR and Deformable DETR on the MOT20 dataset, while DETR yields the highest MIoU on the YouTube-Objects car subset. This comparative study provides valuable insights into the strengths and limitations of each method, highlighting their applicability for various video object detection and tracking scenarios, and paving the way for future research in this field. A repository with our setup and findings can be found at https://github.com/jeremyreist/DETR-Comparative-Study

1 Introduction

1.1 Background and Motivation

Object detection and tracking are essential tasks in computer vision, with a wide range of practical applications such as surveillance, robotics, and autonomous vehicles. Accurately detecting and tracking objects in videos remains a challenging problem due to variations in object appearance, lighting conditions, occlusions, and complex motion patterns. Traditional methods, which rely on hand-crafted features and heuristics, often struggle to handle these challenges effectively.

Recent advances in deep learning have led to the development of end-to-end object detection and tracking systems that can automatically learn relevant features from data. One such promising approach is the DETR (DEtection TRansformer) architecture, which employs a self-attention mechanism to perform object detection and association in a single feedforward pass. However, the performance of DETR in video object detection and tracking has not been extensively explored, and it remains unclear how well it can address the challenges posed by real-world video sequences, such as motion blur, camera jitter, and frame-to-frame variations.

1.2 State-of-the-art

The field of object detection and tracking has made significant advancements in recent years, particularly with the advent of deep learning-based approaches. Current state-of-the-art object detection

Preprint. Under review.

models such as Faster R-CNN, YOLO, and SSD have achieved impressive accuracy and speed while being capable of detecting multiple object classes in an image. However, these models still face limitations when tracking objects in videos, especially in situations where objects are occluded, partially visible, or rapidly moving.

Several recent extensions of the DETR architecture, such as Deformable DETR and TrackFormer, aim to address these challenges in video object detection and tracking. These extensions leverage attention mechanisms to improve performance in various deep learning models, particularly in natural language processing and computer vision. Therefore, it is crucial to investigate the potential benefits of attention-based approaches for video object detection and tracking, particularly in challenging scenarios where traditional methods may struggle.

1.3 Objectives and Scope of the Study

The primary objective of this study is to conduct a comparative analysis of the DETR architecture and its selected variants, Deformable DETR and TrackFormer, to assess their effectiveness in video object detection and tracking. Due to time restrictions and compatibility issues, other initially considered models were excluded from the study. We aim to evaluate the performance of these models on the car subset of the YouTube-Objects dataset and the MOT20-07 video, using mean Intersection over Union (MIoU) and mean Average Precision (mAP) metrics.

This study focuses on identifying the strengths and limitations of each method in various video object detection and tracking scenarios. Our findings will provide valuable insights for researchers and practitioners working on video analysis, contributing to the development of more robust and accurate object detection and tracking systems in the future.

2 Related Work

In this section, we review the related work on the DETR (DEtection TRansformer) architecture and its selected variants, Deformable DETR and TrackFormer, which are the focus of our study. These models leverage attention mechanisms to improve the performance of video object detection and tracking.

DETR (DEtection TRansformer) DETR is a novel approach to object detection that combines a transformer-based architecture with self-attention mechanisms. Unlike traditional two-stage object detection pipelines, which consist of region proposal generation followed by object classification and bounding box regression, DETR performs object detection in a single stage. The model encodes an input image using a convolutional neural network (CNN) and then processes the resulting feature maps with a transformer to predict object classes and bounding boxes simultaneously. This end-to-end framework simplifies the object detection pipeline and has shown competitive performance compared to state-of-the-art methods.

Deformable DETR Deformable DETR is an extension of the original DETR architecture that aims to enhance the detection and tracking of deformed objects. This model addresses the slow convergence and limited feature spatial resolution observed in the original DETR by incorporating a data-dependent sparse attention module inspired by deformable convolution. The deformable attention module allows the model to learn more flexible and expressive attention patterns, leading to improved performance on object detection tasks. Deformable DETR has demonstrated better performance than the original DETR with significantly fewer training epochs.

TrackFormer TrackFormer is a multi-object tracking (MOT) approach based on a transformer architecture that formulates the MOT task as a frame-to-frame set prediction problem. The model leverages attention mechanisms to perform data association between frames, evolving a set of track predictions throughout a video sequence. TrackFormer introduces a new tracking-by-attention paradigm that jointly performs tracking and detection without the need for additional graph optimization or modeling of motion and appearance.

TrackFormer's architecture combines a detection transformer, which predicts object classes and bounding boxes, with a tracking transformer that associates object predictions across frames. This ap-

proach enables the model to achieve state-of-the-art performance on multi-object tracking benchmarks while maintaining real-time processing capabilities.

We also considered two other models, but could not test them due to reasons elaborated on below.

TransVOD/TransVOD++ TransVOD uses transformers to encode temporal information and spatial features. It makes use of a novel temporal transformer to link object queries and memory encoding outputs simultaneously. TransVOD needs several frames of input for one input, increasing the total latency between frame recording and prediction. The paper also proposes a TransVOD Lite model, which uses sequential hard query mining to speed up image detection by up to 6x, and a TransVOD++ model, which is more accurate than the base TransVOD but with less of a speed boost than TransVOD Lite.

All models required inputs to be formatted in the same way as the COCO dataset. Conversion from MOT20 and YT-Objects proved not very seamless, and we could not get it working.

Efficient DETR Efficient DETR addresses the issue of slow convergence and low performance in small object detection in previous end-to-end detectors such as DETR and Deformable DETR. It achieves this by leveraging both dense and sparse set detection to initialize object containers more efficiently, which leads to improved performance and faster convergence. Efficient DETR, with only 3 encoder layers and 1 decoder layer, achieves competitive performance with state-of-the-art object detection methods and is robust in crowded scenes. Unfortunately, there was no open source implementation of their model, so we skipped this model due to time constraints.

3 Methodology

In this section, we describe the datasets, evaluation metrics, and implementation details used in our study to evaluate the performance of DETR, Deformable DETR, and TrackFormer for video object detection and tracking.

3.1 Datasets

3.1.1 MOT20

The MOT20 dataset is a widely used benchmark for multi-object tracking, consisting of high-quality video sequences with diverse scenarios, such as crowded scenes and varying object sizes.

For this study, we focus on the MOT20-07 video, which provides a challenging test case for object detection and tracking due to its complex motion patterns, occlusions, and varying object appearances.

3.1.2 YouTube-Objects

The YouTube-Objects dataset consists of video sequences collected from YouTube, containing multiple object classes. The dataset is suitable for evaluating the performance of object detection and tracking algorithms in real-world video scenarios.

In our study, we specifically focus on the car subset of the YouTube-Objects dataset. This subset provides a diverse range of car appearances, motion patterns, and challenging scenarios for object detection and tracking.

3.2 Evaluation Metrics

3.2.1 MIoU (Mean Intersection over Union)

Mean Intersection over Union (MIoU) is a popular metric for evaluating object tracking performance. It computes the average of the intersection over union (IoU) values between ground truth and predicted bounding boxes across all frames. Higher MIoU values indicate better tracking performance.

3.3 mAP (Mean Average Precision)

Mean Average Precision (mAP) is a widely used metric for object detection tasks, which measures the precision of the detection results across different object categories. However, due to the limitations in the annotations of the YouTube-Objects dataset, which only had one weakly annotated ground truth bounding box even in frames with multiple cars, we opted not to use mAP as it would lead to inaccurate evaluations of false positives.

3.4 Implementation Details

3.4.1 Preprocessing

For both the MOT20-07 and YouTube-Objects car subset datasets, we preprocess the video sequences by extracting individual frames. The frames are then resized and normalized to match the input size and format required by the respective models.

3.4.2 Pre-trained models

We used the models (DETR, Deformable DETR, and TrackFormer) using their respective pre-trained weights, on the car subset of the YouTube-Objects dataset and the MOT20-07 video. We followed the original authors' recommended procedures for running said models.

3.5 Inference

For the inference phase, we applied the trained models to the test sets of the MOT20-07 and YouTube-Objects datasets. We then calculated the MIoU scores for each model, providing a quantitative assessment of their performance in video object detection and tracking.

4 Experimental Results and Analysis

In this section, we present the experimental results and analysis of the DETR, Deformable DETR, and TrackFormer models on the MOT20-07 and YouTube-Objects Car Subset datasets.

4.1 DETR Performance

4.1.1 MOT20

On the MOT20-07 dataset, DETR achieved an MIoU score of 0.329127. This performance indicates that the model can effectively detect and track some, but not all objects in challenging conditions, such as crowded scenes and complex motion patterns.

4.1.2 YouTube-Objects

For the YouTube-Objects Car Subset, DETR obtained an MIoU score of 0.916971. Despite the weak annotations in the dataset, DETR demonstrated a strong performance in detecting and tracking car objects.

4.2 Deformable DETR Performance

4.2.1 MOT20

On the MOT20-07 dataset, Deformable DETR achieved an MIoU score of 0.112819. This result shows that the model underperformed in comparison to DETR in detecting and tracking objects in complex scenes.

4.2.2 YouTube-Objects

In the YouTube-Objects Car Subset, Deformable DETR scored an MIoU of 0.856639. The model's performance was lower than DETR, indicating that its deformable attention mechanism did not significantly improve the detection and tracking of cars in this dataset.

4.3 Trackformer Performance

4.3.1 MOT20

TrackFormer achieved an impressive MIoU score of 0.813151 on the MOT20-07 dataset. This result demonstrates the model's superior performance in detecting and tracking multiple objects in challenging conditions.

4.3.2 Limitations on YouTube-Objects

Unfortunately, we could not evaluate TrackFormer on the YouTube-Objects Car Subset, as the model is specifically designed for detecting and tracking people. This limitation highlights the need for future research to adapt the model for different or potentially multiple object categories.

4.4 Comparative Analysis

4.4.1 Object Detection Performance

In terms of object detection, DETR outperformed Deformable DETR on both datasets. This finding suggests that the deformable attention mechanism may not provide significant benefits for general object detection tasks.

4.4.2 Object Tracking Performance

TrackFormer demonstrated superior object tracking performance on the MOT20-07 dataset compared to DETR and Deformable DETR. Its tracking-by-attention paradigm and set prediction formulation allowed it to effectively associate object predictions across frames.

4.4.3 Processing Time

Our benchmark YT-obj video is 17 seconds long, and our benchmark MOT20 video will be 400 frames long. On a system with a Ryzen 5 1600, RTX2070 Super, and 16GB of RAM, DETR finished processing the YT-obj video in 1 hour 14 minutes and finished processing the MOT20 video in 1 hour and 2 minutes. Deformable DETR finished processing the YT-obj video in 3 hours and 29 minutes and finished processing the MOT20 video in 59 minutes. Trackformer finished processing the MOT20 video in 2 minutes.

4.4.4 Performance in Crowded Scenes

TrackFormer exhibited the best performance in crowded scenes, while Deformable DETR showed the weakest results. This outcome highlights the potential of attention-based methods, like TrackFormer, in handling complex scenarios where traditional object detection and tracking approaches may struggle.

4.5 Discussion

In this section, we discuss the strengths and limitations of the evaluated methods, the best approaches for video object detection and tracking, and the challenges and future directions for research in this area.

4.5.1 Strengths and Limitations of Each Method

DETR demonstrated strong object detection performance in both datasets, indicating its general applicability for object detection tasks. However, its object tracking performance was inferior to that of TrackFormer. Deformable DETR showed some improvement in handling deformed objects, but it did not outperform DETR in our experiments. TrackFormer exhibited superior performance in object tracking, robustness to occlusion and motion blur, and processing time, making it a promising approach for video analysis. Its limitation, however, is that it is currently designed to detect and track people, which restricts its applicability for other object categories.

4.5.2 Best Approaches for Video Object Detection and Tracking

Based on our experimental results, TrackFormer emerges as the best approach for video object detection and tracking, particularly in challenging conditions involving people such as crowded scenes and complex motion patterns. Its attention-based tracking mechanism and set prediction formulation allow it to effectively associate objects across frames. For applications requiring detection and tracking of other object categories, DETR serves as a strong baseline, and researchers could explore the adaptation of the TrackFormer model to handle different object types.

4.5.3 Challenges and Future Directions

Our study identified several challenges and future directions for research in video object detection and tracking:

Strict software requirements: While not a problem unique to our study, every model having very strict requirements on python packages and CUDA versions ended up taking considerable time to solve. Worth highlighting was our inability to use certain versions of pytorch that depended on specific CUDA versions. Some of those versions were unavailable on our stronger GPU due to being too new. This problem will likely exacerbate as the later papers become more dated and rely on newer versions of CUDA.

Adapting models like TrackFormer to handle different object categories: TrackFormer's current limitation in detecting and tracking only people highlights the need for future research to extend the model to other object types.

Investigating the impact of different attention mechanisms: Our results indicate that attention mechanisms can play a crucial role in improving object detection and tracking performance. Future research should explore other attention mechanisms and their potential benefits for video analysis tasks.

Exploring methods to handle weak annotations: Our experiments with the YouTube-Objects Car Subset revealed the challenges posed by weak annotations in the dataset. Developing models that can handle such annotations effectively could lead to better performance in real-world scenarios.

Real-time processing and scalability: As video analysis applications often require real-time processing, it is essential to develop models that are computationally efficient and can scale to large video streams.

Robustness to various challenges: Future research should focus on improving model robustness to various challenges, such as occlusion, motion blur, and changing lighting conditions, to enhance the applicability of these models in real-world scenarios.

5 Conclusion

In this study, we conducted a comparative analysis of three attention-based models for video object detection and tracking: DETR, Deformable DETR, and TrackFormer. Our experiments were performed on two benchmark datasets: the MOT20-07 video and the car subset of the YouTube-Objects dataset. The results demonstrated that TrackFormer outperformed the other models in object tracking, robustness to occlusion and motion blur, and processing time, making it a promising approach for video analysis tasks. However, its current limitation lies in detecting and tracking only people, which restricts its applicability for other object categories.

DETR and Deformable DETR exhibited strong object detection performance, with DETR showing better overall results in our experiments. This suggests that DETR could serve as a strong baseline for video object detection tasks, while researchers could explore adapting the TrackFormer model to handle different object types.

Our study also identified several challenges and future research directions, including the adaptation of models like TrackFormer to handle various object categories, exploring different attention mechanisms, handling weak annotations, real-time processing and scalability, and improving robustness to various challenges.

Overall, our findings provide valuable insights for researchers and practitioners working on video analysis and pave the way towards more robust and accurate object detection and tracking systems in the future.

References

[1] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S. (2020, May 28). End-to-end object detection with Transformers. arXiv.org. Retrieved March 15, 2023, from https://arxiv.org/abs/2005.12872

[2] Meinhardt, T., Kirillov, A., Leal-Taixe, L., Feichtenhofer, C. (2022, April 29). Trackformer: Multi-object tracking with Transformers. arXiv.org. Retrieved March 15, 2023, from https://arxiv.org/abs/2101.02702

[3] Yao, Z., Ai, J., Li, B., Zhang, C. (2021, April 3). Efficient detr: Improving end-to-end object detector with dense prior. arXiv.org. Retrieved March 15, 2023, from https://arxiv.org/abs/2104.01318

[4] Zhou, Q., Li, X., He, L., Yang, Y., Cheng, G., Tong, Y., Ma, L., amp; Tao, D. (2022, November 22). TransVOD: End-to-end video object detection with spatial-temporal transformers. arXiv.org. Retrieved March 15, 2023, from https://arxiv.org/abs/2201.05047

[5] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J. (2021, March 18). Deformable detr: Deformable Transformers for end-to-end object detection. arXiv.org. Retrieved March 15, 2023, from https://arxiv.org/abs/2010.04159