8-2023

Multi-Object Tracking: A Computer Vision Paradigm

Natalie Friede

Multi-Object Tracking: A Computer Vision Paradigm

Multi-Object Tracking: A Computer Vision Paradigm

A thesis submitted in partial fulfillment
of the requirements for the degree of
Batchelor of Science in Computer Science

By

Natalie Friede
Bachelor of Science in Computer Science, 2023

July 2023
University of Arkansas

Abstract

This paper delves into advancements and hurdles encountered in multi-object tracking, a critical aspect of computer vision, with a special emphasis on 'referring understanding.' This technique integrates natural language queries into multi-object tracking tasks, thus broadening the scope for practical applications. The innovative referring multi-object tracking (RMOT) approach emerges as a promising solution in this regard. The effectiveness of RMOT was tested using the Refer-KITTI dataset, a dataset specializing in traffic scenes. The evaluation revealed RMOT's ability to handle a diverse range of referent objects, its robust temporal dynamics, and a high level of adaptability. While the paper acknowledges the significant strides made with this approach, it also illuminates a few inherent limitations and new challenges such as multi-object prediction and cross-frame association. In addressing these issues, the paper attempts to retrain an end-to-end differentiable framework for RMOT, building on the latest DETR framework, suggesting promising prospects for future advancements in this domain. The ultimate goal of this paper is to refine the RMOT model further, promote a more profound understanding of the computer vision landscape, and underscore the technology's potential for future research and applications.

ACKNOWLEDGEMENTS

Thank you to my mentor Pha Nguyen as well as my faculty sponsor Dr. Khoa Luu. They were both exceptionally helpful when I was going through a tough time in life and I would not have been able to get this far without them.

TABLE OF CONTENTS

LIST OF FIGURES

1 Introduction

In the field of computer vision, the emergence of referring understanding —an innovative concept that marries natural language processing with scene perception—signifies a notable stride forward [1, 2]. This approach aims to pinpoint regions of interest within visual media, such as images or videos, using natural language queries. This development holds considerable promise for various applications, ranging from autonomous driving, to video editing, to tracking objects throughout video. Recognized benchmarks in the field have played pivotal roles in advancing image and video based referring tasks along with the development of new data sets to train these models on a wider variety of video feed [2].

Despite these significant advancements, observations show that two principal limitations are present in the existing benchmarks. The natural language expressions and targets can incorrectly end up operating on a one-to-one basis, while real-world scenarios often use one expression to correspond to multiple objects, meaning there is an under representation of multi-object scenarios in current data sets. The second principal limitation can arise when a single phrase can refer to a state that can be entered or left by an object. For instance, consider the expression 'the car which is turning.' Current systems would need to predict the entire trajectory, which could end up including after the car's completion of its turn. Such an instance shows the inability of a single expression to cover all the short-term states of a target. Consequently, existing data sets struggle to encapsulate scenarios involving multiple referent targets and temporal status variations [1].

Newer solutions, such as the referring multi-object tracking (RMOT) approach,

are starting to address these challenges. RMOT specifically is able to target multiple objects in a single video by utilizing a natural language expression as a reference to identify all corresponding objects within a video [1]. This unique approach of RMOT brings the technology closer to real-world applications, as it allows a single expression to involve multiple objects to be tracked throughout a video.

The developers of the RMOT approach have tested it using the Refer-KITTI data set, which specializes in traffic scenes. This data set has unique features like high flexibility with objects, high temporal dynamics, and low labeling cost, therefore advancing existing benchmarks while creating its own benchmarks.

The RMOT model is built using transformer architecture, capitalizing on its inherent ability to handle sequential data, coupled with RoBERTa, a model shown to have exceptional natural language processing capabilities built off the BERT framework. The transformer's ability to capture long-range dependencies within the data makes it an ideal choice for tasks like object tracking in videos, where contextual understanding across multiple frames is critical. On the language understanding side, RMOT leverages RoBERTa, a robustly optimized version of BERT, for processing and understanding the natural language queries. By combining the strengths of both transformer and RoBERTa, RMOT provides a powerful framework for tracking multiple objects in videos using natural language references, merging computer vision and natural language understanding in a highly effective manner [2, 7, 8].

However, while RMOT offers potential advancements, it also presents new challenges in terms of multi-object prediction and cross-frame association. To tackle these challenges, the model for RMOT has been based off of the Modulated Detection for End-to-End Multi-Modal Understanding (MDTER) framework [3]. It showcases the power of cross-modal reasoning and cross-frame conjunction within an encoder-decoder archi-

tecture, offering a promising avenue for future progress in the field. It is evident that this technology holds tremendous potential for growth and refinement, making it a subject of significant interest for future research in the realm of computer vision.

1.1 Goals

In this paper, the goal is to replicate the success of RMOT while re-training the model as well as to understand the growing field of computer vision.

2 Related works

In the realm of multi-object tracking, other innovative models have had similar approaches. A model that simulated human attention through the use of transformer architecture, TrackFormer, is able to track as a frame-to-frame set prediction challenge. The model uses attention for data association between frames and evolves a set of track predictions throughout a video sequence. This new tracking-by-attention paradigm, while simple in design, showed remarkable performance on tasks and segmentation benchmarks, thereby setting a new state-of-the-art standard, which RMOT was able to build off of [3].

Similarly, the approach End-to-End Referring Video Object Segmentation with Multi-Modal Transformers (MTTR) acts as somewhat of a precursor to RMOT. MTTR has a more simple, one stage pipeline that focuses more on predicting which object sequences of text refer to. This model is a more specialized model built to segment objects to be associated with a natural language query. It also utilizes a similar transformer architecture to assist with natural language processing to disseminate the information in the query to help assign the segments to their language counterparts. RMOT has expanded upon this idea by being able to not just identify objects in video, but to be able to identify and track multiple objects pertaining to the same query [6].

3 Methods

3.1 Data Preparation

For this study, the Refer-KITTI data set provided by the RMOT codebase was used, which is specifically designed for interactions while driving. The data set contains labeled data where each expression in a video is annotated with an average of 10.7 objects. This data set provides the model with a robust representation of real-world scenarios.

3.2 Model Configuration and Dependency Resolution

The base of the model is the DETR framework, which combines transformers with set prediction tasks. The RMOT model being trained here utilizes multi-scale deformable attention (MDTER) as the backbone of the model. To build and configure the model, hundreds of dependencies worked in tandem, with the most prominent (and problematic) being pytorch, torchtext, and opencv-python.

3.3 Training Procedure

Training the RMOT model was done using standard back propagation methods. The model is trained end-to-end, and all parameters are updated via Adam, a popular optimization algorithm. The encoder-decoder architecture facilitated the attention based learning.

During training, regular validation checks using a held-out portion of the data were set to monitor the model's performance and ensure that it is not over-fitting the

training data.

3.4 Testing

Testing the model was done using the testing data provided by the KITTI dataset. A training command was utilized and the parameters which were used are outlined in the code snippet bellow.

The following code was generated to test the model

```
python3 inference.py \
--meta_arch rmot \
--dataset_file e2e_rmot \
--with_box_refine \
--epoch 200 \
--lr_drop 100 \
--lr 2e-4 \
--lr_backbone 2e-5 \
--batch_size 1 \
--sample_mode random_interval \
--sample_interval 1 \
--sampler_steps 50 90 150 \
--sampler_lengths 2 3 4 5 \
--update_query_pos \
--random_drop 0.2 \
--fp_ratio 0.3 \
--query_interaction_layer QIM \
--extra_track_attn \
--resume exps/default/checkpoint0099.pth \
--output_dir exps/default2 \
--num_workers 4 \
--visualization \
--sgd \
--seed 42 \
```

These parameters were chosen mostly based off what is standard, such as the amount of encoder/decoder layers. A small batch size and shorter epochs were chosen

so as to produce results in a time efficient manner and not overwhelm the machine. The model was tested with higher epochs and a larger batch size, but the differences did not warrant the extra time and resources. The interface command and the model are from the RMOT repository and modified from Deformable DETR [2, 3].

3.5 Evaluation Metrics

After training, evaluate the model using standard metrics for multi-object tracking tasks. These may include Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), and other relevant metrics such as the number of missed detections, false positives, and identity switches. Unfortunately, due to something going wrong during the testing phase, the data was annotated correctly on the output images, but each separate object was not given a proper id, was predicted more than once in the same spot, or there was an issue in the evaluating script which caused the ”Tracker predicts the same ID more than once in a single time step” to occur. Manual evaluation was done by looking through the predictions and manually identifying if the model was performing correctly. This is not an ideal method, but since many of the issues were ocurring in the same areas (namely queries with state based identifiers (like ”moving”), identifying people, and identifying between ”cars” and other vehicles), it was somewhat effective to manually check these problem areas for improvement.

3.6 Fine-tuning and Iterative Improvement

Finally, additional fine-tuning was done, and ideally the model would have iterated based on the performance metrics. This iterative process of training, evaluation, and fine-tuning would help to enhance the performance of the RMOT model over time.

4 Results

The results of the training were as follows. The model was able to correctly identify, track, and label the objects that pertained to the natural language query. The model had issues where it was double-bounding certain objects, which does not show visually in the produced annotated images, but it made it impossible to run the evaluation on the model. From what is shown in the images, the model did decently in a variety of scenarios outlined in the introduction as being problematic, namely multi-object association, identification of objects based off temporary states, and identification of objects that become obscured, but it did run into a few issues, mainly in identifying objects in certain states. Below are some examples of the model's predictions. Figures are generated from individual frames in the annotated videos.

Figure 4.1, which shows the results of identifying "the cars with a light color," easily shows how the model is able to correctly identify objects based off of color and shape descriptions (with the shape being "car", which would not include semis). While "lighter color" is a somewhat subjective term, the model is able to identify all cars meeting this parameter and is able to decide which colors it considers "lighter".

Figure 4.2, which focuses on identifying pedestrians who are moving, shows how the model is able to identify smaller objects in a state-based action, but how it can lose track of them when obscured behind a pole in low light. Luckily the model was able to pick back up on the walking pedestrian, but was not able to identify it as the same pedestrian from earlier.

Figure 4.3, which tries to identify ”moving cars,” shows some of the difficulties in state based identification. The model is able to correctly identify the car in front of our car as ”moving”, which shows that it is able to identify a physical frame of reference and identify things that are ”moving”, but not moving in terms of their placement on the screen. The model does tend to over-index, however, and incorrectly identifies cars that are parked as ”moving”, when they are only moving on camera. This issue occurs with a few other queries, like ”cars that are parking”, or ”vehicles turning” where the perspective of the camera and movements of the filming car may make objects appear to be taking actions they are not.

Figure 4.4, which identifies ”standing women,” struggles from similar issues as figure 4.3. The model not only has to identify between men, women, and children, but also differentiate between standing, walking, and sitting people. The model is able to do so in low light, but does perform slightly more confidently when figures are up close and well lit. In a few instances, the model mistakes sitting men for standing women. The differentiation of men and women appears to be strongly influenced by stature, hair length, and the size of the person’s chest, so some taller women with shorter hair were not classified when they should have been.

Figure 4.5 shows another tricky situation where the model must identify both ”women” and ”bag carriers” and marry the two. The way the query was separated placed more weight on identifying individuals carrying ”bag” than just women. The model does correctly identify multiple women with bags, but also tends to over-guess and confidently identifies men and women or men without bags as well. From the queries tested, this seemed to be one of the most difficult for the model. There were no instances found where a woman with a bag was in frame but not identified, however there were

many cases of misidentifying pedestrians who did not match both descriptions. To fix this, the model could be re-trained with higher penalties for mis-identification.

Figure 4.1: Query: The cars with a light color

(a) The initial identification includes cars on the side of the road

(b) The model is able to identify moving cars with light color and keep track of previously identified cars

(c) Another example with darker cars not being identified

Figure 4.2: Query: The pedestrians who are moving

(a) The model is able to identify the pedestrian

(b) The pedestrian is lost behind a pole in a dark setting

(c) The pedestrian is re-identified, but as a different object

Figure 4.3: Query: Moving cars

(a) The model is able to identify all the moving cars in the image and ignores the parked car

(b) The model incorrectly assumes parked car is moving

(c) The model correctly identifies all moving cars and ignores the semi

Figure 4.4: Query: Standing women

(a) The model is able to pick up 2 women standing in the distance

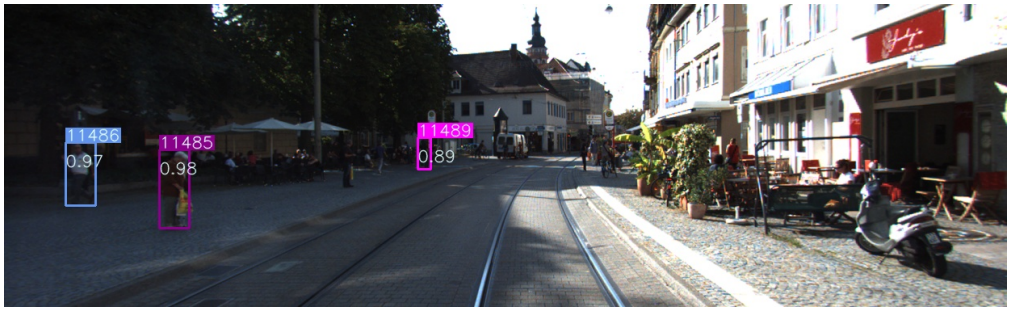
(b) As the women get closer, the model is slightly more confident

(c) The model is able to identify 3 women and ignore 1 man

(d) The model identifies a standing women

(e) The model un-marks the woman once she is no longer standing and starts to walk

Figure 4.5: Query: Women with bags



(a) The model identifies one woman with a bag (right), a man with a bag (middle), and a man without a bag

(b) The model correctly identifies 3 women with bags

(c) The model identifies one woman with a bag, but incorrectly identifies a sitting man

5 Conclusion

In conclusion, the model was able to correctly identify multiple objects given a single natural language phrase in most cases. The main situations where there were tricky cases where the frame of reference was not clear, such as ”cars breaking,” ”cars moving,” and ”cars turning,”. The model did appear to perform well on simple identification tasks using color or descriptor, and was able to identify some of the state based actions, although not perfectly. Since the evaluation script was not able to be run due to the issue with how the model reported the predictions, it is impossible to know exactly how good it is without manually looking through the thousands of frames of video.

While there were hangups in the model, it was able to analyze video relatively quickly, with a 340 frame video taking 39 seconds to annotate, or .11 seconds per frame. In real word scenarios, faster performance would be required for self driving cars that have to make split-second decisions, but for the purpose of identifying and tracking the frames per second of 8.7 is slightly higher performing than similar models [4].

6 Bibliography

[1] L. Ye, M. Rochan, Z. Liu, et al, ”Cross-Modal Self-Attention Network for Referring Image Segmentation,” in CVPR, University of Manitoba, Canada, and Shanghai University, China, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.1904.04745

[2] D. Wu, W. Han, T. Wang, et al, ”Referring Multi-Object Tracking,” in CVPR, 2023. [Online]. Available: https://arxiv.org/abs/2303.03366

[3] A. Kamath, M. Singh, Y. LeCun, et al, ”MDETR - Modulated Detection for End-to-End Multi-Modal Understanding,” arXiv:2104.12763v2 [cs.CV], 12 Oct 2021. [Online]. Available: https://arxiv.org/abs/2104.12763v2

[4] T. Meinhardt, A. Kirillov, L. Leal-Taixe´, et al, ”TrackFormer: Multi-Object Tracking with Transformers,” Technical University of Munich and Facebook AI Research (FAIR). [Online]. Available: https://arxiv.org/pdf/2101.02702.pdf

[5] X. Zhu, W. Su, L. Lu, et al, ”DEFORMABLE DETR: DEFORMABLE TRANSFORMERS FOR END-TO-END OBJECT DETECTION” arXiv:2010.04159v4 [cs.CV], 18 Mar 2021. [Online]. Available: https://arxiv.org/abs/2010.04159v4.

[6] A. Botach, E. Zheltonozhskii, C. Baskin, ”End-to-End Referring Video Object Segmentation with Multimodal Transformers,” arXiv:2111.14821v2 [cs.CV], 3 Apr 2022. [Online]. Available: https://arxiv.org/abs/2111.14821v2.

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, ”BERT: Pre-training of Deep Bidi-

rectional Transformers for Language Understanding," arXiv:1810.04805v2 [cs.CL], 24 May 2019. [Online]. Available: https://arxiv.org/abs/1810.04805v2.

[8] Y. Liu, M. Ott, N. Goyal, et al, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv:1907.11692v1, 26 Jul 2019. [Online]. Available: https://arxiv.org/abs/1907.11692v1.

A Identification of All Software Used in Research and Thesis Generation

Computer #1:

Model Name: AMD Ryzen 7 5800 8-Core Processor

Location: CVIU Lab

Owner: CVIU Lab

Software Name
RMOT source code found at https://github.com/wudongming97/RMOT

Computer #2:

The laptop used to remotely access Computer #1. Owner: Natalie Friede