Review for Learning a Neural Solver for Multiple Object Tracking

Irem ARPAG

May 15, 2021

As a long-standing research problem of computer vision, Multiple Object Tracking (MOT) has been dominated, in recent years, by tracking -by-detection models that is formed of two steps; the first is to get frame-by-frame object detections through learning based detectors and the second is to link detections to create trajectories via data association which is constructed as graph partitioning task aiming to find the detections belonging to the same trajectory by a cost function within a graph optimization framework. Constituting efficient graph formulation and learning better costs were the primary tasks of the previous works on graph-based MOT but focusing on improving the graph optimization framework or the feature extraction cause a dilemma. In order to prevent it, the authors of the article suggest to concentrate these two tasks together with a joint learning-based solver that performs learning directly in the graph domain with a message passing network (MPN) that learns to join deep features into high-order information across the graph and predict final partitions of the graph into trajectories. The pipeline is constituted of four parts; the first one is graph construction where nodes correspond to detection and edges correspond to connections between nodes; the second part is feature encoding, initializing the node appearance feature embeddings from a CNN; the third one is neural message passing, yielding updated embeddings for nodes and edges and the final part is training by using the cross-entropy loss.

Although the model depends on a simple graph formulation, it achieves global interactions among detections. Their fully differentiable framework does not require heavily engineered features but still it is very fast comparing with the traditional graph partitioning methods. Their ingenious, time-aware, neural message passing network (MPN) perform-feature learning and final solution prediction together efficiently. Plain architecture design results in a substantial performance improvement with respect to state-of-art in three public benchmarks, with both MOTA and IDF1 metrics.

As being an off-line recording that is a standard way for video analysis tasks, the model does not provide a real-time-performance.

For all experiments, MOT Challenge Pedestrian dataset is used including challenging tracking 2DMOT2015, MOT16, MOT17 benchmarks. To comprehend the functions of the model, researchers present an ablation study, compar-

ing the three fundamental components of the model. The first one is to evaluate the performance of time-aware node update that demonstrate 98.8 in percentage constraint satisfaction where the baseline has only 82.1 in percentage success which confirm its remarkable achievement of linking detections. The second one is the effect of number of steps in network training, for both IDF1 and MOTA, a clear upward tendency is observed with the increasing of the steps, but twelve messages are the end of development. Finally, the impact of the features for edges that are time difference, relative position and the Euclidian distance in CNN embeddings between the two bounding boxes are evaluated. Results show that relative position information has a significant importance on overall performance. Comparing of the model with previous state-of-art methods prove the accuracy of learned solver and its fastness especially with IDF1 measure with a rising performance of 11, 6.4, and 6.6 percentage.

With its fully differentiable pipeline, the model carries on the task of MOT from tracking-by-detection approach to entirely learning problem in which feature extraction and data association can be learned together that open a new way for future work.