A Proposal to Estimate Trajectory of Objects Without Using Images or Video

Bharat Bhargava

Intern, Samsung Shishya Program 2020

Problem Statement

Given only a set of predictions of **bounding box coordinates** from object detection models, how will you "associate" these predictions together. Current implementations are using the Hungarian algorithms and graph based primitives for this. But, we are more interested in treating the association step as a **separate deep learning problem**.
So, if we give you only a .txt file containing **object detections** and **classes**, how will you link those detections across a sequence of frames for trajectory estimation?

Data Association- An Overview

Data association can be divided into three broad categories:

1. **Feature extraction on a per-frame basis:** This would reduce our problem to feature comparison, rather than comparison of entire images. Co-ordinates of the bounding boxes can serve our purpose here.
2. **Feature comparison across detections:** Association would certainly involve comparisons of features (actually, objects, but with very high computational efficiency), treating the entire problem as a supervised learning problem.
3. **Intra-video feature aggregation:** Video datasets carry along with them rich temporal information. With our model predicting per-frame associations, linking such (single-frame) associations to obtain the trajectory would complete our task of object tracking. A "naive" thinking would be- why not perform a time-series data analysis here? Note that we also have the classes of objects provided; afterwards, we can apply a soft-max classification for class prediction and through the use of an appropriate loss function, we can train the model in an end-to-end fashion.

Purported Methodology

1. The detections will first pass through a network extracting out the spatial features. These features can be a representative of multiple levels of abstraction, and hence a single output feature vector for a particular object can be a conglomeration of vectors obtained at several intermediate stages.
2. Thereafter, an exhaustive permutation of such feature vectors from two different frames can be the input to the network which would estimate similarity between objects in a supervised manner by predicting the correlation between two feature vectors per estimation.
3. Eventually, data association would simply involve performing a comprehensive time-series data analysis of the feature matrices obtained earlier, and we would store a feature matrix per frame for future reference for trajectory estimation.

Feature Extraction

- Run two separate deep networks for feature extraction, which may share the parameters.
- Conglomerate only the outputs of a few intermediate CNN layers so as to obtain a spatial feature vector per object per frame.
- See the output as a **spatial feature matrix**, with a feature vector for each object.
- Finally, combine these feature vectors in every possible way so as to form a **tensor** to be processed by the subsequent sub-network.

Feature Comparison

- Reduce the tensor obtained into a matrix. Can be achieved by a **compression network**.
- Concatenate an extra row and column to the output of the compression network (an N_m X N_m matrix) to cater to the objects present in one frame but absent in the other frame.
- Compare with the ground truth after approriate **thresholding** of the indices.
- Back-propagate the error to train the network for detecting similarity between detections across different frames.

Temporal Feature Aggregation

- Pass the detections through the feature extractor to obtain comprehensive spatial feature matrices.
- Perform batch-wise time-series data analysis on these matrices along with affinity estimation (performed separately by feature comparator) through the use of RNN/GRU/LSTM to possibly get a **video-level feature descriptor**.
- Apply soft-max to predict the class for an object, compare with ground truth class.
- This sub-network enhances the **temporal continuity** of the data provided, by linking the associations by performing a (probable) time-wise ordering.

Conclusion

The poster summarizes a theoretical attempt to propose data association using a deep network, when no image/video data is provided and the only data available is the bounding boxes of detected objects and their classes.

- New features formed from the given data can be fed to a spatial feature extraction sub-network.
- The output of this sub-network can be used to perform a highly exhaustive feature comparison.
- Simultaneously, a comprehensive, supervised time-series analysis can be used to not only associate features across frames, but also to impart a temporal sequence to the associations.

References

- Arulkumar Subramaniam, Athira Nambiar, and Anurag Mittal. Co-segmentation inspired attention networks for video-based person re-identification. The IEEE International Conference on Computer Vision (ICCV), 2019.
- ShiJie Sun, Naveed Akhtar, HuanSheng Song, Ajmal Mian, and Mubarak Shah. Deep affinity network for multiple object tracking. Journal of LaTeX class files, 13(9), 2017.
- Zhongdao Wang, Liang Zheng, Yixuan Liu, and Shengjin Wang. Towards real-time multi-object tracking. arXiv preprint arXiv:1909.12605, 2019.