Review on “Detect to Track and Track to Detect”

Authors: C. Feichtenhofer, A. Pinz, A. Zisserman

# Short Summary

This paper tackles the task of detection and tracking of objects in video. Specifically, the authors propose a new ConvNet architecture which is intended to improve upon previous methods by jointly performing both detection and tracking. Doing away with exhausting post-processing steps on top of frame-based detected, this paper demonstrates that it is possible to infer a tracklet over multiple frames by simultaneously performing detection and tracking. The approach is termed Detection and Tracking (D&T).

Detection is performed by extending a R-FCN with a tracker that utilizes correlation and regression. The model can be trained end-to-end accepting multiple frames as input, passing them through a ConvTrunk (ResNet-101) to produce features, computing convolutional cross-correlation between adjacent frames to estimate displacement, employing RoI-Pooling to classify and regress box proposals, and RoI-tracking to regress box transformations across frames.

The model achieves state-of-the-art on ImageNet VID (79.8% mAP) at the time of publishing and also improves static object detection.

# Main Contributions

1. Proposed a ConvNet architecture for simultaneous learning of frame-based object detection and across-frame track regression.
2. Introduce correlation maps that represent object co-occurrence through time to aid the network with tracking
3. Link detections and tracklets to produce high accuracy detections in video
4. Show that temporal sliding improves tracker speed
5. Achieves state-of-the-art on ImageNet VID at time of publishing

# High-Level Evaluation of Paper

This paper sets out to simplify existing methods and in doing so finds that the proposed model performs better on ImageNet VID. Although a lot of the theory in the paper is difficult to follow, Figures 2 and 3 in the paper do an excellent job of providing an overview of the model. I didn’t find that the paper did a good job in articulating, with words, what they were proposing or why it yields better performance.

# Discussion on Evaluation Methodology

The model is evaluated on the ImageNet VID dataset using per-class % AP and mean % AP. Additionally, it is compared to historical baselines and state-of-the-art in previous years. They also show how the inclusion of tracking loss improves upon R-FCN in a similar manner. Overall, the evaluation is fairly clear.

# Future Directions

Two possible directions for future work are described in the paper: removing detections with low scores along a tube and investigating how additional frames influence detector performance.