Context Matters: Refining Object Detection in Video with Recurrent Neural Networks

Subarna Tripathi[1]
http://acsweb.ucsd.edu/~stripath/
Zachary C. Lipton[1]
zlipton@cs.ucsd.edu
Serge Belongie[2,3]
sjb344@cornell.edu
Truong Nguyen[1]
tqn001@eng.ucsd.edu

[1] University of California San Diego
La Jolla, CA, USA
[2] Cornell University
Ithaca, NY, USA
[3] Cornell Tech
New York, NY, USA

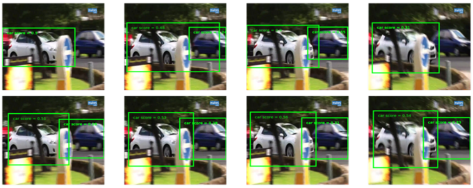

Figure 1: RNN (bottom) recognizes multiple objects more accurately than a state of the art frame-level model (top).

In this paper, we introduce a framework for improving object detection in videos by capturing temporal context and encouraging temporally consistent predictions. First, we train a *pseudo-labeler*, that is, a domain-adapted convolutional neural network for object detection. The pseudo-labeler is first trained individually on the subset of labeled frames, and then subsequently applied to all frames. Then we train a recurrent neural network (RNN) that takes as input sequences of pseudo-labeled frames and optimizes an objective that encourages both accuracy on the target frame and consistency across consecutive frames.

The approach incorporates strong supervision of target frames, weak-supervision on context frames, and regularization via a smoothness penalty. Building on YOLO, a domain-adapted frame-level object detection model [3], we demonstrate that for the sparsely annotated *YouTube Objects* dataset [2], our method achieves mean Average Precision (mAP) of 68.73 on test data, as compared to a best published result of 37.41 [4] and 61.66 for YOLO alone.

As with YOLO [3], our fine-tuned $pseudo-labeler$ takes 448×448 frames as input and regresses on category types and locations of possible objects at each one of 7×7 non-overlapping grid cells. For each grid cell, the model outputs class conditional probabilities as well as 2 bounding boxes and their associated confidence scores.

Then, to incorporate temporal context, we train an RNN with gated recurrent units (GRUs) [1] to refine the provisional predictions. This net takes as input sequences of *pseudo-labels*. For this recurrent model, we demonstrate an efficient and effective training strategy. The objective encourages predictions to be close to true labels (for labeled frames), not to deviate too far from the pseudo-labels, and to be similar across adjacent frames. As demonstrated experimentally, our framework proves effective, achieving state-of-the art mAP and producing compelling visual examples.

[1] KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *Proc. Workshop on Syntax, Semantics and Structure in Statistical Translation*, 2014.

[2] Alessandro Prest, Vicky Kalogeiton, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari. Youtube-objects dataset v2.0, 2014. URL `calvin.inf.ed.ac.uk/datasets/youtube-objects-dataset`. University of Edinburgh (CALVIN), INRIA Grenoble (LEAR), ETH Zurich (CALVIN).

[3] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016.

[4] Subarna Tripathi, Serge J. Belongie, Youngbae Hwang, and Truong Q. Nguyen. Detecting temporally consistent objects in videos through object class label propagation. *WACV*, 2016.