SEQ-NMS FOR VIDEO OBJECT DETECTION

Wei Han[1]*, **Pooya Khorrami**[1]*, **Tom Le Paine**[1]*, **Prajit Ramachandran**[1],
Mohammad Babaeizadeh[1], **Honghui Shi**[1], **Jiana Li**[2]
Shuicheng Yan[2], **Thomas S. Huang**[1]

[1]University of Illinois at Urbana-Champaign
{weihan3, pkhorra2, paine1, prmchnd2
mb2, hshi10, t-huang1}@illinois.edu
[2]National University of Singapore
{elev373, eleyans}@nus.edu.sg

ABSTRACT

Video object detection is challenging because objects that are easily detected in one frame may be difficult to detect in another frame within the same clip. Recently, there have been major advances for doing object detection in a single image. These methods typically contain three phases: (i) object proposal generation (ii) object classification and (iii) post-processing. We propose a modification of the post-processing phase that uses high-scoring object detections from nearby frames to boost scores of weaker detections within the same clip. We show that our method obtains superior results to state-of-the-art single image object detection techniques. Our method placed 3^{rd} in the video object detection (VID) task of the ImageNet Large Scale Visual Recognition Challenge 2015 (ILSVRC2015).

1 INTRODUCTION

Single image object detection has experienced large performance gains in the last few years Girshick et al. (2014), Girshick (2015), Ren et al. (2015), Russakovsky et al. (2015), He et al. (2015). Video object detection, on the other hand, still remains an open problem. This is mainly because objects that are easily detected in one frame may be difficult to detect in another frame within the same video clip. Some reasons for this difficulty include: (i) drastic scale changes (ii) occlusion and (iii) motion blur. In this work we propose a simple extension of single image object detection to help overcome these difficulties.

The main contributions of our work are as follows:

1. We present Seq-NMS, a method to improve object detection pipelines for video data. Specifically, we modify the post-processing phase to use high-scoring object detections from nearby frames in order to boost scores of weaker detections within the same clip.
2. We evaluate Seq-NMS on the ImageNet VID dataset and show that it outperforms state-of-the-art single image-based methods. We show that our method is helpful in cases where single frames contain objects that are at extreme scales, occluded, or blurred. We present specific instances where our Seq-NMS improves performance.
3. Our method placed 3^{rd} in the video object detection (VID) task of the ImageNet Large Scale Visual Recognition Challenge 2015 (ILSVRC2015).

2 OUR APPROACH

We use ideas from the "tracking by detection" literature Wolf et al. (1989), Bercla et al. (2006), Perera et al. (2006) to combine individual detections into sequences, and then we use the sequences to re-score individual bounding boxes. Consider the example video clip given in Figure 1. For each

*Authors contributed equally to this work

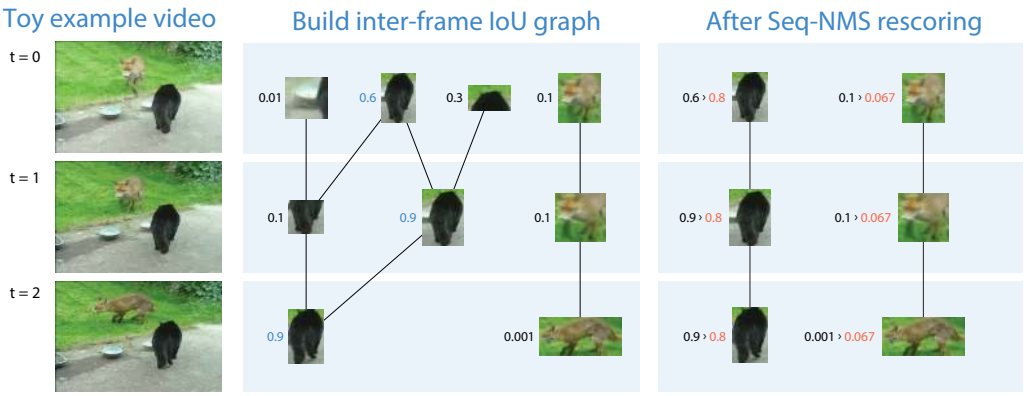

Figure 1: Simple Example of Seq-NMS. Given a video sequence of region proposals and their corresponding class scores, Seq-NMS links bounding boxes in adjacent frames iff they have an IoU greater than 0.5. It then selects boxes to maximize a sequence's sum of scores. The selected boxes are then used to suppress overlapping boxes in their respective frames and are subsequently rescored using either the maximum score in the sequence or the average (above).

pair of adjacent frames, a box in the first frame is linked with a box in the second frame if their IoU is above a threshold. We find such linkages across the entire clip. Then, we attempt to find the maximum score chain across the entire clip, by finding the sequence of linkages such that the sum of object scores of all the boxes in the sequence is maximized. This can be found efficiently using a very simple dynamic programming algorithm that maintains the maximum score chain so far at each box. The algorithm returns a set of chained boxes $(b_i, b_{i+1}, \ldots, b_j)$. The chained boxes are then removed from the set of boxes we link over. We also do suppression within frames such that if a bounding box in frame $t, t \in [i, j]$, has an IoU with b_t over some threshold, it is also removed from the set of candidate boxes. This algorithm is then repeated until a length 1 chain is returned.

After chains are extracted, the constituent bounding boxes can be reweighted. The simplest reweighting scheme is to set each bounding box's object score to the average of object scores across the chain. Another approach is to set the score to the maximum score across the chain. We experimented with both approaches.

3 Experiments

We validate our approach by conducting experiments on both the validation and test sets of the ImageNet VID dataset. Our system uses the region proposal network (RPN) and the classifier of the Faster R-CNN framework Ren et al. (2015). The RPN is based is a Zeiler Fergus style network Zeiler & Fergus (2014) while the classifier is a VGG16 network Simonyan & Zisserman (2014) pre-trained on the ImageNet DET challenge. During the post-processing phase, we considered three different techniques: (i) single image NMS (ii) Seq-NMS (avg) (iii) Seq-NMS (max). Seq-NMS (avg) and Seq-NMS (max) rescored the sequences selected by Seq-NMS using the average or max detection scores respectively. Table 1 shows our performance on the ImageNet VID validation and test sets. Our Seq-NMS (avg) model achieved 3^{rd} place in VID task of the ImageNet 2015 competition [1]. Figure 2 shows which classes experienced the largest gains in performance when switching from single image NMS to Seq-NMS (avg). In Figure 3, we present clips from the ImageNet VID dataset where Seq-NMS improved performance.

Acknowledgments

The six Tesla K40 GPUs used for this research were donated by the NVIDIA Corporation.

[1] http://image-net.org/challenges/LSVRC/2015/results

Table 1: Method comparison on ImageNet VID validation and test set.

Method	mAP(%) - (Validation)	mAP(%) - (Test)
VGG net + NMS	44.9	43.4
VGG net + Seq-NMS (max)	50.5	47.5
VGG net + Seq-NMS (avg)	**51.4**	**48.7**

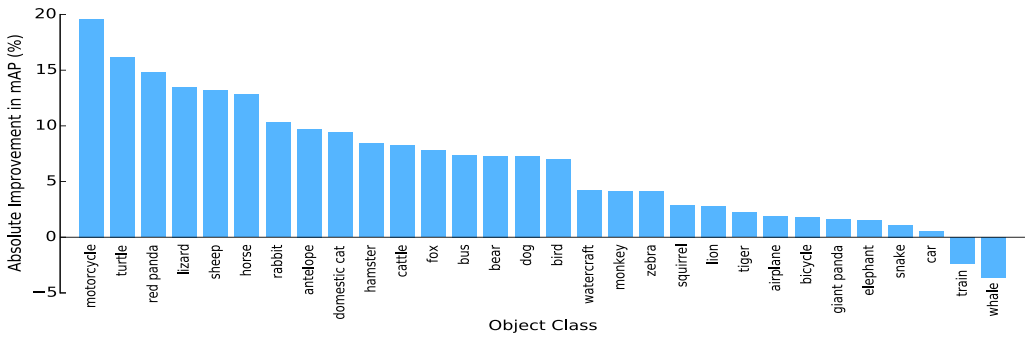

Figure 2: Absolute improvement in mAP (%) using Seq-NMS. The improvement is relative to single image NMS. Note that 7 classes have higher than 10% improvement, and only two classes show decreased performance (train and whale).

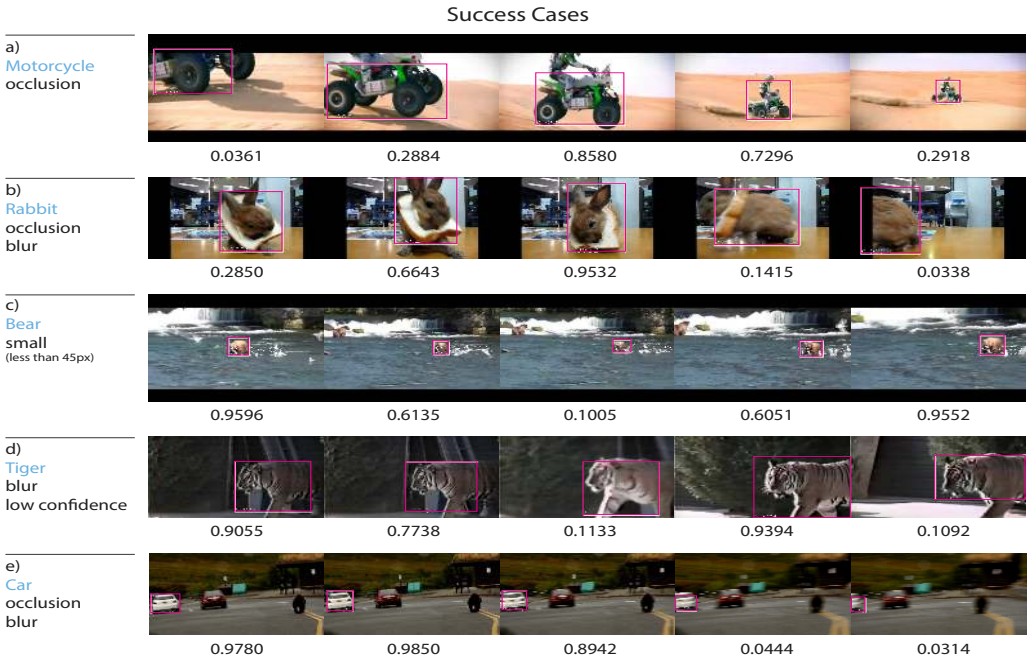

Figure 3: Example video clips where Seq-NMS improves performance. The boxes represent a sequence selected by Seq-NMS. Clips are subsampled to provide examples of high and low scoring boxes. In clips **a**, **b**, and **e**, the object becomes more and more occluded as it exits the frame, leading to lower scores. Meanwhile, in clips **c** and **d**, the object of interest has a low classifier score because it is either very small or blurred, respectively. In all of these cases, Seq-NMS' rescoring significantly boosts the weaker detections by using the strong detections from adjacent frames.

REFERENCES

Jérôme Bercla, Francois Fleuret, and Pascal Fua. Robust people tracking with global trajectory optimization. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pp. 744–750. IEEE, 2006.

Ross Girshick. Fast R-CNN. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2015.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

AG Amitha Perera, Chukka Srinivas, Anthony Hoogs, Glen Brooksby, and Wensheng Hu. Multi-object tracking through simultaneous long occlusions and split-merge conditions. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pp. 666–673. IEEE, 2006.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Neural Information Processing Systems (NIPS)*, 2015.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Jack K Wolf, Audrey M Viterbi, and Glenn S Dixon. Finding the best set of k paths through a trellis with application to multitarget tracking. *Aerospace and Electronic Systems, IEEE Transactions on*, 25(2):287–296, 1989.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, pp. 818–833, 2014.