I Am Guessing You Can't Recognize This: Generating Adversarial Images for Object Detection Using Spatial Commonsense (Student Abstract)

Anurag Garg
PQRS Research and DIT
Dehradun, India
anuraggarg1209@gmail.com

Niket Tandon
Allen Institute for AI
Seattle, WA, USA
nikett@allenai.org

Aparna S. Varde
Montclair State University
Montclair, NJ, USA
vardea@montclair.edu

Abstract

Can we automatically predict failures of an object detection model on images from a target domain? We characterize errors of a state-of-the-art object detection model on the currently popular *smart mobility* domain, and find that a large number of errors can be identified using spatial commonsense. We propose CSK-SNIFFER, a system that automatically identifies a large number of such errors based on commonsense knowledge. Our system does not require any new annotations and can still find object detection errors with high accuracy (more than 80% when measured by humans). This work lays the foundation to answer exciting research questions on domain adaptation including the ability to automatically create adversarial datasets for target domain.

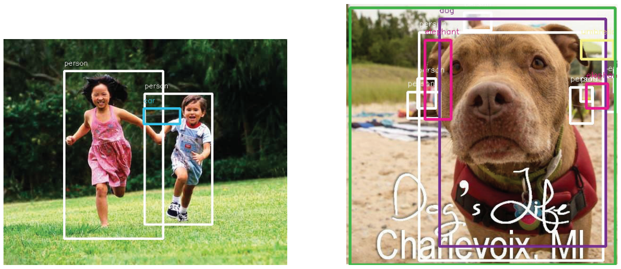

Figure 1: (left) YOLO model incorrectly predicts a car between 2 kids (in blue bounding box), and (right) a person is predicted in the large white bounding box. This paper automatically compiles such errors.

.

Introduction

Motivation and Problem Definition: Object detection has seen tremendous progress on various benchmark datasets, yet, on real-world images these models encounter large variance in view points, object appearance, backgrounds, illumination and image quality (Chen et al. 2018). This domain shift between the training and test data causes models to perform poorly (Gopalan, Li, and Chellappa 2011). Domain adaptation is an active research field to address this problem, with the goal of maintaining accuracy on new domains while using minimal additional annotation. To advance domain adaptation, *can we* ***predict*** *failures of an object detection model on images from any domain?* This is the question our paper addresses, answering which would enable:

- Automatic error prediction to allow performance estimates or creation of an adversarial dataset for any domain.
- Such adversarial datasets across domains to measure object detection performance holistically (an approach trending in NLP as well (Talmor and Berant 2019)).
- Studying an open issue on whether the adversarial dataset produces better trained models than large, randomly selected training data on a new target domain.

We observed that a large fraction (≈37.5%) of the errors made by a state-of-the-art object detection model, YOLO (Redmon and Farhadi 2016), on images from *smart mobility* (Vienna 2015), were odd according to our commonsense knowledge on relative locations of objects (see Fig. 1). Finding such mistakes automatically is the goal of this paper.

Related Work: Commonsense knowledge (CSK) in general (Chowdhury et al. 2018), and spatial commonsense in particular has helped to improve object detection (Shiang et al. 2017), especially when there is limited data or when moving to a new domain or unseen concepts (Kumar S. et al. 2018). Spatial knowledge has been extracted either directly from images (Yatskar, Ordonez, and Farhadi 2016) or mined from text (Xu, Lin, and Zhu 2018). Our work does not aim to acquire spatial commonsense or to inject it in object detection models, instead we use it to *automatically generate* adversarial sets for new target domains, opening interesting avenues for further research in domain adaptation.

Contributions: The contributions of this work are:

1. We propose CSK-SNIFFER, a system to automatically predict failures of any object detection model on *any* domain. CSK-SNIFFER verifies the predicted bounding boxes of objects, finding whether their relative locations are in accordance with commonsense knowledge.
2. We automatically create an adversarial dataset for *smart mobility*, and establish its high quality (indicating high quality of CSK-SNIFFER). Humans verify that 80% of the flagged images indeed had object detection errors.

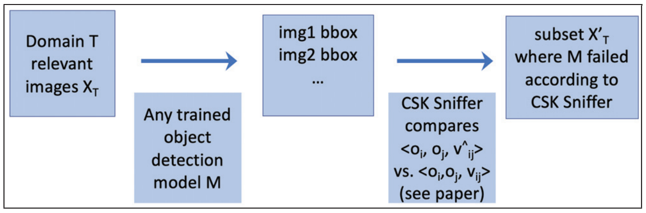
Domain T relevant images X_T
Any trained object detection model M
img1 bbox
img2 bbox
…
CSK Sniffer compares <o_i, o_j, $v^{\wedge}_{ij}$> vs. <o_i,o_j, v_{ij}> (see paper)
subset X'_T where M failed according to CSK Sniffer

Figure 2: CSK-SNIFFER flow (details in approach section)

Proposed System: CSK-SNIFFER

Task: The inputs to the system are:

- Object detection model M trained on a source domain S
- Typical objects, actions in a target domain T: $\texttt{vocab}(T)$
- Image search engine (e.g., Bing/ Google) to query entries in $\texttt{vocab}(T)$, leading to a large collection of images X_T.

The output of the system is a subset of images X'_T on which the model is likely to make wrong bounding box predictions.

Approach: A naive approach could be to manually identify incorrect bounding boxes in each image in X_T to construct X'_T, but has a highly prohibitive manual labeling cost.

Instead, CSK-SNIFFER finds anomalies in the relative locations of an image's bounding boxes. These anomalies are defined w.r.t. a commonsense KB containing likely relative locations of a pair of objects o_i, o_j. An entry in the KB is a triple: $< o_i, o_j, v_{ij} >$ where v_{ij} is a binary vector over relations $rel(KB)$. The KB used in our work has $rel(KB)$ defined on 5 relations (`isAbove`, `isBelow`, `isInside`, `isNear`, `overlapsWith`). We can use any other external KB that is available, e.g. (Yatskar, Ordonez, and Farhadi 2016)) or construct a KB manually as we currently do. To detect anomalies, we first generate triples $< o_i, o_j, \hat{v}_{ij} >$ from the predicted bounding boxes of image x_T using a function $f(bbox)$. Then, CSK-SNIFFER compares them against the KB triples: e.g., $< o_i, o_j, v_{ij} >$. If $\hat{v}_{ij} \neq v_{ij}$, then the image is flagged as potentially wrong. Figure 2 illustrates the flow of CSK-SNIFFER, that "sniffs" for subtle intuitive aspects in object detection. See **Appendix** for additional details on the KB and $f(bbox)$.

Experimental Evaluation

Experimental setup: In our experiments, S=$\texttt{MSCOCO}$, T=$\texttt{smart mobility}$, $\texttt{vocab}(T)$ comprises 200+ entries on *smart mobility* e.g. "canal lights dimming when occupants are few", "people parking bikes at share-ride spots", "kids running in playgrounds near roads" etc. X_T comprises of 14977 images, obtained by collecting top-75 images retrieved from Google image search for entries in $\texttt{vocab}(T)$. The output dataset X'_T comprises 4872 images.

Results: Human evaluation performed by 3 annotators based on 100 random images from X'_T found that 80% (of 4900) indeed had errors. This shows that if our method predicts an error, then it very likely is an error. To perform a large-scale, automated evaluation we use the gold annotated images from development set of MSCOCO. Ideally, if CSK-SNIFFER is 100% accurate then no error should be predicted in gold images. We found that 86.2% of the images (690 out of 5000) were flagged as containing erroneous bounding boxes. This demonstrates that CSK-SNIFFER can effectively discriminate good bounding 86.2% of the times. So, our model has both high precision (it can tell apart bad and good bounding boxes, see Table 1) and has high recall (it found nearly 5000 bad predictions from 15000 images and manual sampling finds that only 28% bad predictions went undetected). See **Appendix** for details on the evaluation, error analysis and examples of $\texttt{vocab}(T)$, X_T and X'_T and screenshots from a demo of the system.

Setting	Accuracy
good predictions, flagged as good	86.2%
bad predictions, flagged as bad	80.0%

Table 1: Accuracy numbers of CSK-SNIFFER

Conclusions and Future Work

This paper proposes a commonsense-based system that "sniffs" object detection errors on a previously unseen domain, with high accuracy, at no additional annotation cost. The surprising finding is that our simple method can automatically discover errors and might be useful in an active learning setting. We demonstrate that with high quality, we can create adversarial datasets on target domains such as *smart mobility*. Our work opens up several research issues, some of which we are investigating in our ongoing work:

- Leverage existing, potentially noisy and incomplete commonsense KBs in CSK-SNIFFER.
- Improve model M with signals from CSK-SNIFFER.
- Determine whether automatic adversarial datasets X'_T help train better models on a new target domain.

References

Chen, Y.; Li, W.; S., C.; Dai, D.; and V.G., L. 2018. Domain adaptive faster r-cnn for object detection in the wild. *CVPR*.

Chowdhury, S. N.; Tandon, N.; F., H.; and Weikum, G. 2018. Visir: Visual and semantic image label refinement. *WSDM*.

Gopalan, R.; Li, R.; and Chellappa, R. 2011. Domain adaptation for object recognition: unsupervised approach. *ICCV*.

Kumar S., K.; D., S.; Farhadi, A.; and Jae, Y. 2018. Dock: Detecting objects by transferring commonsense. *ECCV*.

Redmon, J., and Farhadi, A. 2016. Yolo9000: Better, faster, stronger. *CVPR*.

Shiang, S.-R.; Rosenthal, S.; G., A.; Carbonell, J.; and Oh, J. 2017. Vision-language fusion for object recognition. *AAAI*.

Talmor, A., and Berant. 2019. Multiqa: Generalization and transfer in reading comprehension. *ACL*.

Vienna, T. 2015. European smart cities. *Technical report*.

Xu, F. F.; Lin, B. Y.; and Zhu, K. 2018. Automatic extraction of commonsense locatednear knowledge. *ACL*.

Yatskar, M.; Ordonez, V.; and Farhadi, A. 2016. Stating the obvious: Extracting visual common sense. *NAACL*.