CSK-SNIFFER: Commonsense Knowledge for Sniffing Object Detection Errors

Anurag Garg[1], Niket Tandon[2] and Aparna S. Varde[3]

[1]PQRS Research, Dehradun, India
[2]Allen Institute for AI, Seattle, USA
[3]Montclair State University, Montclair, USA

Abstract

This paper showcases the demonstration of a system called CSK-SNIFFER to automatically predict failures of an object detection model on images in big data sets from a target domain by identifying errors based on commonsense knowledge. CSK-SNIFFER can be an assistant to a human (as sniffer dogs are assistants to police searching for problems at airports). To cut through the clutter after deployment, this "sniffer" identifies where a given model is probably wrong. Alerted thus, users can visually explore within our demo, the model's explanation based on spatial correlations that make no sense. In other words, it is impossible for a human without the help of a sniffer to flag false positives in such large data sets without knowing ground truth (unknown earlier since it is found after deployment). CSK-SNIFFER spans human-AI collaboration. The AI role is harnessed via embedding commonsense knowledge in the system; while an important human part is played by domain experts providing labeled data for training (besides human commonsense deployed by AI). Another highly significant aspect is that the human-in-the-loop can improve the AI system by the feedback it receives from visualizing object detection errors, while the AI provides actual assistance to the human in object detection. CSK-SNIFFER exemplifies visualization in big data analytics through spatial commonsense and a visually rich demo with numerous complex images from target domains. This paper provides excerpts of the CSK-SNIFFER system demo with a synopsis of its approach and experiments.

1. Introduction

Human-AI collaboration, the realm of humans and AI systems working together, typically achieves better performance than either one working alone [1]. Big data visualization and analytics can be used to foster interaction [2]. Such areas receive attention, e.g. NEIL (Never Ending Image Learner) [3], active learning approaches [4], human-in-the-loop learning [5] etc. To that end, we demonstrate a system "CSK-SNIFFER" exemplifying human-AI collaboration via enhancing object detection by visualizing potential errors in large complex data sets, harnessing spatial commonsense. This system "sniffs" errors in object detection using spatial collocation anomalies, assisting humans analogous to sniffer dogs aiding police at airports. The process, (Figure 1), is as follows, with the CSK-SNIFFER system (C), human-in-the-loop (H), and inference model (M) for object detection.

System C interacts with human H and provides object detection output visualizing potential errors, over model M, by deploying commonsense knowledge through a spatial knowledge base (KB). The KB is derived by capturing spatial commonsense, especially as collocation anomalies. Then H sees the visualized errors (output by C) and can thereby enhance C by increasing its precision and recall, based on the spatial KB. Hence, the two directions in this learning loop are as follows.

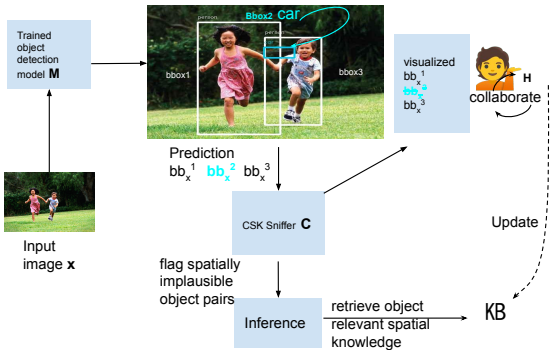

Figure 1: CSK-SNIFFER and the human-in-the-loop: A car was detected in the image which was flagged bad by CSK-SNIFFER based on its spatial knowledge w.r.t KB which the human-in-the-loop can update after visualizing errors

- H to C: Feedback-based interactive learning
- C to H: Assistance in object detection

Thus, the human and the AI work together with the goal of enhancing object detection in big data. Further, the inference model M can potentially improve, as an added benefit of this adversarial learning via human-AI collaboration. The obtained information can be used to supply more examples to M on the misclassified categories to make it more robust. If certain labels are inappropriate

Published in the Workshop Proceedings of the EDBT/ICDT 2022 Joint Conference (March 29-April 1, 2022), Edinburgh, UK
anuraggarg1209@gmail.com (A. Garg); nikett@allenai.org (N. Tandon); vardea@montclair.edu (A. S. Varde)
© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3135/bigvis_short2.pdf

consistently across examples, it is a valuable insight. As data sets get bigger in volume and variety, such automation is even more significant in assisting object detection errors.

The use case in our work focuses on the smart mobility domain [6]. It entails autonomous vehicles, self-operating traffic signals, energy-saving street lights dimming / brightening as per pedestrian usage etc. In such AI systems, it is crucial to detect objects accurately, especially due to issues such as safety. CSK-SNIFFER plays an important role here, generating large adversarial training data sets by sniffing object detection errors.

2. The CSK-SNIFFER Approach

We summarize the CSK-SNIFFER approach as per its design and execution [7]. In this approach. we represent and construct a KB, along with the function $f(bbox)$ that generates a triple, i.e. $< o_i, \hat{v}_{ij}, o_j >$ from the predicted bounding boxes of objects o_i and o_j such that v_{ij} is a binary vector over relations $rel(KB)$.

Gather X_T using `action-vocab`(T): Table 1 presents some examples from `action-vocab`(T) where T =*smart mobility* domain. These entries represent mostly typical and some unique scenes in this domain. This list was manually compiled by a domain expert. Images X_T are compiled using Web queries $\in$ `action-vocab`(T) (on an image search engine), and an object detector predicts bounding boxes over $x_T \in X_T$.

People crossing city streets on pedestrian crossings
Vehicles coming to a full hault at red signals
Vehicles stopping or slowing down at stop signs
Street lights dimming when occupants are few
Street lights brightening when occupants are many
Buses running on traffic-optimal routes
Service dogs helping blind people
People charging phones at WiFi stations
People reading useful information at roadside kiosks
People parking bikes at share-ride spots
Vehicles flashing turning lights for L/R turns
Bikes riding on bike routes only
Traffic cops making hand signals in regular operations
Vehicles driving beneath an overpass
Dogs on a leash walking with their owners
People jogging on sidewalks
People entering and leaving trains when doors open
People using prams for kids in buses
Trees existing on sidewalks
Ropeways carrying passengers to tourist spots
Bikers wearing smart watches
Maglev trains running between airports and cities
Grass existing on freeway sides and city streets
Solar panels existing on roofs of buildings
People using smartphones for talking anytime anywhere
Canal lights dimming when occupants are few
Canal lights brightening with many occupants
People wheeling shopping carts in grocery stores

Table 1
$\sim$10% examples from `action-vocab`(T) where T =*smart mobility* domain. These are used as queries to compile the input to the object detector, and then CSK-SNIFFER can flag images in T where the detector failed to predict the correct bounding boxes.

KB construction: While in principle, we can directly use existing KBs, these have errors as elaborated in some works [8]. CSK-SNIFFER isolates the effect of these errors by instead manually creating a KB at a very low cost. The KB is defined over a set of objects O and relations $rel(KB)$. The relation set $rel(KB)$ comprises 5 relations (`isAbove`, `isBelow`, `isInside`, `isNear`, `overlapsWith`). We are inspired by other works in the literature such as [9] in picking these relations, and because our initial analysis proposed their suitability for bounding box relative relations. An entry in the KB comprises of a pair of objects $o_i, o_j \in O$ and a binary vector v_{ij} denoting o_i, o_j's and their spatial relations over $rel(KB)$. These spatial relations are manually annotated by a domain expert, according to general likelihood e.g., it is more likely that a dog is observed near a human, and much less likely that it is observed near a whale. The k most popular objects on MSCOCO training data make up O; in our experiments $k = 10$, and this leads to k^2 entries in KB that need to be annotated with v_{ij}. It is remarkable that our experiments demonstrate that even with $k = 10$, the KB allows CSK-SNIFFER to achieve good performance. We can infer that selecting a popular subset helps, even if it is small. An entry in the KB is denoted as $< o_i, v_{ij}, o_j >$ The KB is publicly available at https://tinyurl.com/kb-for-csksniffer

Function f(bbox): Similar to the triples in the KB, we define a function $f(bbox)$ to construct triples $< o_i, \hat{v}_{ij}, o_j >$ using the predicted bounding boxes of image x_T. The $f(bbox)$ input consists of predicted bounding boxes on an image, and the output is a list of triples in the format: $< o_i, \hat{v}_{ij}, o_j >$, for every pair of objects $o_i, o_j \in$ the objects detected in the image. For every such pair, $f(bbox)$ compares the coordinates of the bounding boxes of o_i and o_j (this is a known process e.g. [9]). We illustrate this for the `isInside` relation. Let coordinates of a bounding box be x_1, y_1, x_2, y_2, then $o_i.y_1$ denotes y_1 coordinate of o_i. If $o_i.y_2 \leq o_j.y_1$ and $o_j.x_2 > o_i.x_1$ and $o_j.x_1 < o_i.x_2$ then o_i `is inside` o_j. Similarly, other relations in $rel(KB)$ are built, compiling which provides $\hat{v}_{ij}$. For anomaly detection, we compare vectors $\hat{v}_{ij}$ and v_{ij} for overlapping object-pairs o_i, o_j, detected in the image and present in the KB.

Based on this discussion, the following algorithm summarizes the execution of CSK-SNIFFER.

Algorithm 1: CSK-SNIFFER Approach

Input: Object detector M trained on source domain S

Manually compiled `action-vocab`(T) in target domain T

Images X_T compiled using Web queries $\in$ `action-vocab`(T)

1. Define $rel(KB)$ comprising 5 relations: `isAbove`, `isBelow`, `isInside`, `isNear`, `overlapsWith`
2. Define commonsense KB, each entry $< o_i, v_{ij}, o_j >$ where v_{ij} is a binary vector over $rel(KB)$
3. Generate triples $< o_i, \hat{v}_{ij}, o_j >$ from predicted bounding boxes of $x_T \in X_T$ using function $f(bbox)$.
4. For each image $x_T \in X_T$, compare $\hat{v}_{ij}$ and v_{ij}, from bounding box triples $< o_i, \hat{v}_{ij}, o_j >$ and KB triples $< o_i, v_{ij}, o_j >$
5. For each x_T, if $\hat{v}_{ij} \neq v_{ij}$ then flag $x_T : wrong$, add x_T to X'_T

Output: Subset X'_T where M failed

3. Excerpts from System Demo

We have built a live demo to depict the working of CSK-SNIFFER. This demo illustrates the functioning of CSK-SNIFFER to enhance its actual comprehension and augment its usage. In addition, this demo paper presents the principles behind the human-in-the-loop functioning of CSK-SNIFFER for sniffing object detection errors in large, complex data sets, thereby being added contributions over our earlier work [7]. While this human-in-the-loop functioning is explained in the introduction with an illustration and theoretical justification, its detailed empirical validations with respect to interactive KB updates constitute ongoing work, based on CSK-SNIFFER being actively deployed in real-world settings. In fact, this demo paves the way for such interactive KB updates via augmenting the usage of CSK-SNIFFER in suitable applications to provide the human-in-the-loop feedback for the addition of such interactive KB updates.

We present some screenshots illustrating the demo. Many more can be provided in a live setting. The user enters any search query related to smart mobility [6]. Images are downloaded from Google Images based on this query. Object detection is then performed on the images using YOLO [10] to start predicting triples in the image using the $f(bbox)$ function. Once the triples are predicted, the demo moves to the home page. This contains details on the output files generated. The “Images” option displays downloaded images as shown in Figure 2 herewith.

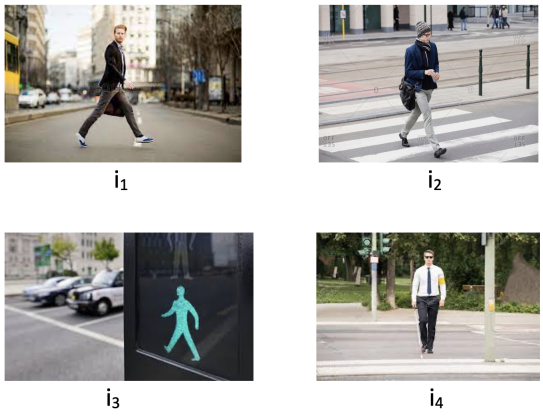

Figure 2: A sample from the gathered images (X_T) by searching the Web for "People crossing city streets on pedestrian crossings" in `action-vocab`(T)

Inferred spatial relation on predicted bounding boxes	Frequency
person, overlapsWith, person	18
car, is_near, car	10
person, overlapsWith, car	8
car, overlapsWith, person	8
car, overlapsWith, car	8
person, is_near, backpack	4
backpack, is_near, person	4
car, is_near, backpack	4
backpack, is_near, car	4
traffic light, is_near, traffic light	4

Table 2
Distribution of spatial relations in X_T. Spatial relations are of the form: $< o_i, \hat{v}_{ij}, o_j >$

Output files generated by CSK-SNIFFER are illustrated as follows. Table 2 shows the first output file “Collocations Map” with triples predicted by CSK-SNIFFER in images with their respective counts. The final output file “Error Set” Table 3 contains names of images with some odd visual collocations. It also indicates the triple that actually got predicted versus the expectation from the model. These files help fathom the functioning of CSK-SNIFFER.

Image id	Inferred spatial relation on predicted bounding boxes	Expected spatial relation between these objects present in KB
i_1, i_3	person, overlapsWith, car	person, is_near, is_inside, car
i_1, i_3	car, overlapsWith, person	car, is_near, person
i_1, i_4	backpack, is_inside, person	backpack, is_near, overlapsWith, person
i_1	backpack, is_near, car	backpack, is_inside, car
i_2	traffic light, is_inside, traffic light	traffic light, is_near, is_above, traffic light
i_2	traffic light, overlapsWith, traffic light	traffic light,is_near, is_above, traffic light
i_3	truck, overlapsWith, car	truck, is_near, car
i_4	person, is_above, backpack	person,is_near, overlapsWith, backpack
i_4	backpack, is_below, person	backpack, is_near, overlapsWith, person
i_4	backpack, is_inside, backpack	backpack, is_near, backpack
i_4	backpack, overlapsWith, backpack	backpack, is_near, backpack

Table 3
Canonical examples of errors flagged by CSK-SNIFFER If inferred spatial relations over model-generated bounding boxes are not consistent with expected spatial relations between objects, then predicted bounding boxes are flagged as erroneous.

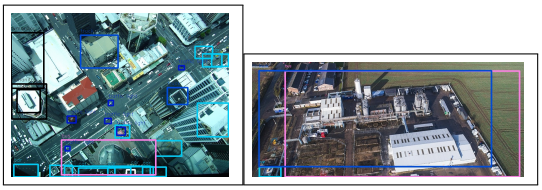

Figure 3: Images actually bad, flagged bad (In the 1st image "buildings" are detected as "truck" and "TV monitor"; in the 2nd image "buildings" are detected as "bus" and "truck").

Figure 4: Images actually good, flagged good. CSK-SNIFFER has a high success rate in not flagging images with meaningful bounding box collocations.

4. Experimental Evaluation

We present examples from our experiments, showing the correct and wrong predictions made by CSK-SNIFFER, along with the error analysis. Here, "bad" refers to images containing object detection errors while "good" refers to correctly identified images with no such errors.

4.1. Appropriate Identifications

Actually bad, flagged bad: Figure 3 illustrates examples in this category. Experimental evaluation shows that our model is good at identifying odd bounding boxes.

Actually good, flagged good: Figure 4 portrays examples of this type. Experimental evaluation shows that CSK-SNIFFER is able to distinguish good predictions.

Other benefits: Interestingly, while analyzing mistakes of CSK-SNIFFER , we find that ∼10% of the reference data on which M is trained (MSCOCO, expected to be a high quality), contains wrong bounding boxes. This provides insights into potentially improving MSCOCO, constituting an added benefit of this work.

On the whole, the human and the AI collaborate with each other, such that the AI (CSK-SNIFFER) provides a visual demo of the object detection errors sniffed by spatial CSK, thus generating large adversarial data sets to assist object detection, while the human can use this feedback to enhance the performance of CSK-SNIFFER, thereby playing its role in the learning loop.

4.2. Error Analysis

We now present the precision and recall shortcomings.

Recall issues: Actually bad, flagged good: Figure 5 depicts examples of these types of images. The reason for CSK-SNIFFER predicting these images as good instead of bad is that the objects wrongly detected in the image are not present in our KB, hence it does not check for their locations. Thus, they are not found in any of the triples predicted, they are skipped so that they do not make their way to the error set.

Precision issues: Actually good, flagged bad: Our investigation of the source of these errors (see Figure 6), concluded that the KB relations are authored with a **3D space** in perspective, while the images only contain $2D$ information. Therefore, relations such as `above` and `below` may be confused with `farther` and `nearer`. For example, if a `car` is detected in the background, $f(bbox)$ function make an incorrect interpretation as `car is above person`. The KB will flag this as unlikely and hence an erroneous detection, leading to a possibly good prediction flagged as an error.

Addressing 2D vs. 3D errors: We calculate the area covered by a bounding box, such that if the area is less

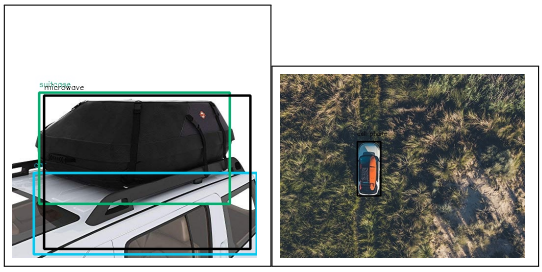

Figure 5: Images with bounding boxes actually bad, flagged good (In the 1st image "suitcase" is also detected as "microwave"; in the 2nd image, "car" is detected as "cell phone").

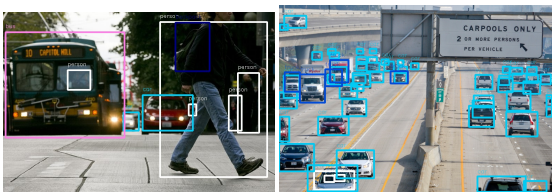

Figure 6: Images actually good, flagged bad (In the 1st image CSK-SNIFFER predicts "person inside person"; in the 2nd image it predicts "car above car", hence flagged as bad).

than an empirically estimated threshold that object is considered to be detected in the background and therefore does not predict the triple, e.g. `car above person` in that image. This helps to increase accuracy to ∼80%.

5. Conclusions and Roadmap

This paper synopsizes the demo (with approach and experiments) of a system "CSK-SNIFFER" that "sniffs" object detection errors in a big data on an unseen target domain using spatial commonsense, with high accuracy at no additional annotation cost. Based on human-AI collaboration, the AI angle entails spatial CSK imbibed in the system deployed via visual analytics to assist humans, while an important human role comes from the domain expert perspective in image tagging and task identification for training the system (in addition to the obvious human contribution of commonsense knowledge in the system). More significantly, the human and the AI make contributions to the learning loop by feedback-based interactive learning, and assistance in object detection respectively. It is promising to note that our approach based on simplicity can automatically discover errors in data of significant volume and variety, and be potentially useful in this learning setting. We demonstrate that with high quality, we can generate large complex adversarial datasets on target domains such as smart mobility.

Future work includes harnessing existing, potentially noisy and incomplete commonsense KBs in CSK-SNIFFER Another direction is to study whether automatic adversarial datasets compiled with assistance from CSK-SNIFFER help train better models on novel target domains. Our work presents interesting facets from big data visualization and analytics along with human-AI collaboration.

6. Acknowledgments

A. Varde has NSF grants 2018575 (MRI: Acquisition of a High-Performance GPU Cluster for Research & Education); 2117308 (MRI: Acquisition of a Multimodal Collaborative Robot System (MCROS) to Support Cross-Disciplinary Human-Centered Research & Education). She is a visiting researcher at Max Planck Institute for Informatics, Germany.

References

[1] D. Wang, E. Churchill, P. Maes, X. Fan, B. Shneiderman, Y. Shi, Q. Wang, From human-human collaboration to human-ai collaboration: Designing ai systems that can work together with people, in: CHI, 2020, pp. 1–6.

[2] F. Zuo, J. Wang, J. Gao, K. Ozbay, X. J. Ban, Y. Shen, H. Yang, S. Iyer, An interactive data visualization and analytics tool to evaluate mobility and sociability trends during covid-19, arXiv:2006.14882 (2020).

[3] X. Chen, A. Shrivastava, A. Gupta, Neil: Extracting visual knowledge from web data, in: ICCV, 2013, pp. 1409 – 1416.

[4] K. Konyushkova, R. Sznitman, P. Fua, Learning active learning from data, Advances in Neural Information Processing Systems 30 (2017).

[5] D. Xin, L. Ma, J. Liu, S. Macke, S. Song, A. Parameswaran, Accelerating human-in-the-loop machine learning: Challenges and opportunities, in: ACM SIGMOD (DEEM workshop), 2018, pp. 1–4.

[6] A. Orlowski, P. Romanowska, Smart cities concept: Smart mobility indicator, Cybernetics and Systems (Taylor & Francis) 50 (2019) 118–131.

[7] A. Garg, N. Tandon, A. S. Varde, I am guessing you can't recognize this: Generating adversarial images for object detection using spatial commonsense, in: AAAI, 2020, pp. 13789–13790.

[8] N. Tandon, A. S. Varde, G. de Melo, Commonsense knowledge in machine intelligence, ACM SIGMOD Record 46 (2017) 49–52.

[9] M. Yatskar, V. Ordonez, A. Farhadi, Stating the obvious: Extracting visual common sense, NAACL (2016).

[10] J. Redmon, A. Farhadi, Yolo9000: Better, faster, stronger, CVPR (2016).