SemOI: A Semantical Augmentation for the Open Images Detection

Achim Reiz, Kurt Sandkuhl, Birger Lantow
Rostock University
Rostock, Germany
achim.reiz, kurt.sandkuhl, birger.lantow@uni-rostock.de

***Abstract*—In this paper, we present a semantic extension to the popular Open Image dataset. Using a python script, we transferred the JSON-representation into RDF(S) and then extended the 576 detectable classes by 392 scenes connected by 57 intermediate objects called occasions. The extension allows the inferring of possible locations and suitable semantic objects for the detected items. Using a GUI, a user can upload a picture and calculate a numerical value for the possibility of a given semantic augmentation.**

I. Introduction

Open Images is a popular open-source dataset comprising of about 9 million annotated images [1]. It can be used to train and set up domain-independent object recognition systems. With TensorFlow Hub, there is already a pre-trained neural network (NN) publicly available that can easily be integrated into existing applications, bypassing the need to train a new NN using the 550 Gb large Open Images repository.

Fig. 1. Sample input for OpenImage - detected (among others) tent, car, wheel, and person

While the trained object detection delivers promising results, the detected items are isolated, and a shared context is missing. In Fig. 1, TensorFlow discovered (among others) *Tent, Car,* and *Man.* While the identification of these objects is correct, we still cannot grasp the corresponding situation. A high-level semantic is missing, allowing the interpretation of the results beyond the elements that are visible. Such can be achieved using ontology methods [2]. All object detectors have an ontological twin in our semantic augmentation component, with the same name and the same unique classifier-ID. These twins are connected to additional contextual knowledge called occasions, which are further connected to scenes, allowing us to infer, for this case, the return values *public-place*, *campsite*, or *youth_hostel*.

The paper is structured as followed. The next section presents relevant related work, followed by an overview of the ontology. The service architecture and the ontology's request handling are presented in section four, followed by a preliminary evaluation. The paper concludes with a discussion and an outlook on future research directions.

II. Related Work

This work mainly falls into the category of the connection of semantic technologies with deep learning, especially image recognition. Recently, Ding et al. [2] and Bhandari and Kulikajevas [3] published literature reviews in this subarea. Both show application scenarios for the interdisciplinary combination of the two technologies. Ding. et al. presented possibilities for improving detection accuracy using WordNet and the categorization of sport-videos through enhanced tagging. Further, he gave examples for the inferring of user behavior and support for robot vision. Bhandari and Kulikajevas presented methodologies for automated image annotation and improvement possibilities through the capturing of `Part-Of` relationships. Additionally, relevant are the works by Zambrano et al. [4] with a semantic-based analysis of CCTV and the work by Abburu and Anandhi, who created an ontology-based soccer-video tagging [5].

The main idea and the services' architecture are often similar: The image analysis detects the low-level objects within a picture, the semantic augments this detection by inferring high-level relationships. However, in detail, the various approaches differ widely in their requirements and goals. The CCTV analysis builds upon extensive image data while the possible events are limited. The improving of segmentation analysis through `Part-Of` relationships described by [2, 3] induces a significantly different ontology structure. There is currently no similar approach for the augmentation of an extensive library like Open Images towards a contextual classification to the best of our knowledge.

III. Semantic Augmentation

The ontology for the semantic augmentation of the image recognition technology builds on the detector-IDs and labels. Open Images' Webpage provides a JSON-hierarchy of the detector IDs and a CSV to translate the IDs into human-readable labels. Using a python script, we merged these two inputs and

ISSN 2305-7254

built an RDF(S)-hierarchy with the label-name as the class name and an annotation that contains the detector ID. The translated hierarchy was then converted into a valid OWL-representation using the ontology editor "Protégé" [6]. Reusing an ontology out of the fashion domain, we imported `Scenes` with a total of 379 subclasses that indicate locations like `bar`, `cafeteria`, or `pier`. As the mapping of the 576 detectable objects to 379 possible scenes creates significant overhead, these two classes are connected using intermediate objects called occasions (57 in total). [7] further describes the originating ontology.

Fig. 2 contains an example of the relationships between the classes. The left side shows the detectable `Drinks` in Open Images. These elements connect with `party` and `leisure-activity` from the intermediate class `occasion`. These `occasions` relate to `scenes`, in this case, `beer_hall`, `bar`, `cafeteria`, and `porch`, differentiated into `inside` and `outside` (for enhanced readability, Fig. 2 contains just a fraction of the related scenes). In the specific example, the semantic augmentation can infer that a detected object `Beer` fits into the location `beer_hall`, or bar but not a `cafeteria` or `porch`.

The ontological relationships themselves do not contain a confidence-measurement for a scene's probability based on a detected property. To better grade the semantic service results, we implemented a measurement based on the confidence of the image recognition service and the frequency of the detected objects. Taking the vector $\overrightarrow{oi}$ for the confidence of the image detection regarding the objects i, and the matrix S for the scenes k inferred by the semantic, the multiplication of $\overrightarrow{oi} * S$, followed by normalization, returns the semantic confidence SC of the system.

$$SC = \overrightarrow{oi} * S = \begin{pmatrix} oi_1 \\ \vdots \\ oi_i \end{pmatrix} \begin{pmatrix} s_{11} & \cdots & s_{1k} \\ \vdots & \ddots & \vdots \\ s_{i1} & \cdots & s_{ik} \end{pmatrix}$$

Taking the example from above, we could imagine the following calculation for the semantic confidence values:

$$SC = \begin{pmatrix} 0{,}77\ beer \\ 0{,}2\ wine \\ 0{,}2\ cocktail \end{pmatrix} \begin{pmatrix} beerhall & 0 & beerhall \\ bar & 0 & bar \\ 0 & cafeteria & 0 \\ 0 & porch & 0 \end{pmatrix}$$

$$SC = 0{,}97\ beerhall + 0{,}97\ bar + 0{,}2\ cafeteria + 0{,}2\ porch$$

$$SC_{normalized} = 100\%\ beerhall + 100\%\ bar + 20{,}6\%\ cafeteria + 20{,}6\%\ porch$$

IV. The Architecture of Augmentation Service

This section gives a manual for using the newly developed augmentation service and presents the architecture and orchestration of the different components.

A. Using the Newly Developed Service

The service is publicly available on the servers of Rostock University on semoi.informatik.uni-rostock.de. Here, a user can upload a JPG file and trigger the analysis. Two possible object detection algorithms are available: SSD/Mobilenet v2 and Inception Resnet. While the former delivers coarse results in about 30-40 seconds for a 4,5 MB picture with 4032x3024 pixels, the latter offers more precision, but is computational more expensive and takes about 230 seconds for the same image. [8] presents a rigorous comparison between the different neural nets. After a successful analysis, the preview window shows the annotated picture with the bounding boxes of the object detection algorithm. Fig. 3 presents a screenshot of the functioning tool. The image recognition's confidence values are shown on the bottom left, while the window on the bottom right presents the inferred scenes from the semantic augmentation, including the semantic confidence value.

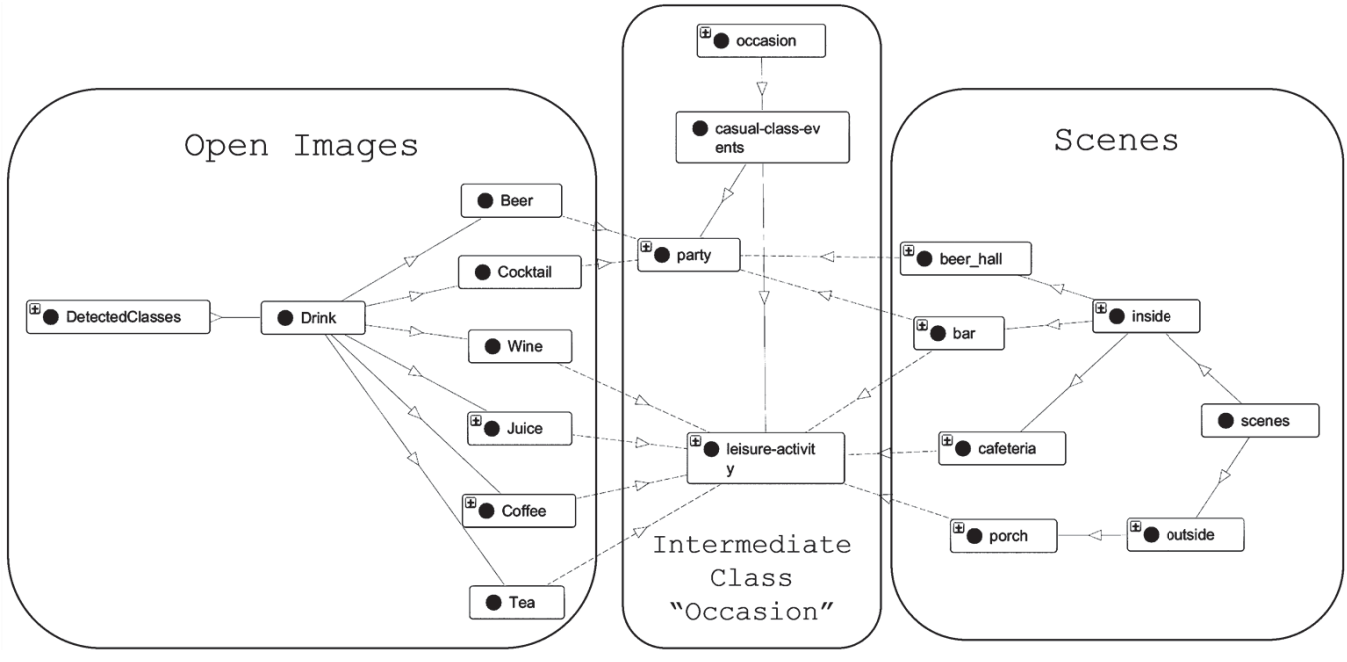

Fig. 2. An Excerpt out of the augmentation ontology

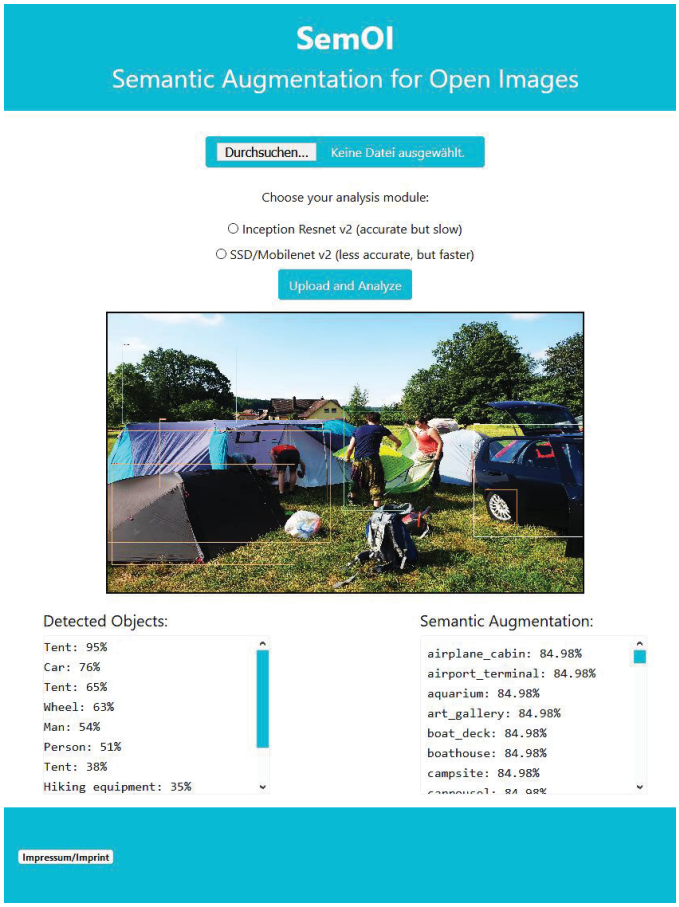

Fig. 3.Screenshot of *the* SemOI webpage (semoi.informatik.uni-rostock.de)

B. *The architecture of Augmentation Service*

The application builds upon a microservice architecture. In total, the service consists of three parts: The frontend and image-recognition container, the semantic augmentation web-service, and a triplestore. The former serves the webpage, stores the picture temporarily, and performs the object detection using TensorFlow-hub with a pre-trained model for Open Images. In the next step, the detected objects are transferred to the semantic advisory component. This component performs a SPARQL-request for every detected object on an apache fuseki server. Based on the inferred objects, the SC-value is calculated for each detected element and returned to the fronted to display the results.

The code is written in python and utilizes the Django- and Django-Rest-Framework and Docker for containerization. For ontology development, we used the tool Protégé. Further, the code is published under an open-source license (LGPL) on Github (https://github.com/Uni-Rostock-Win/SemOI), including the dockerfile for the service orchestration and the ontology-file.

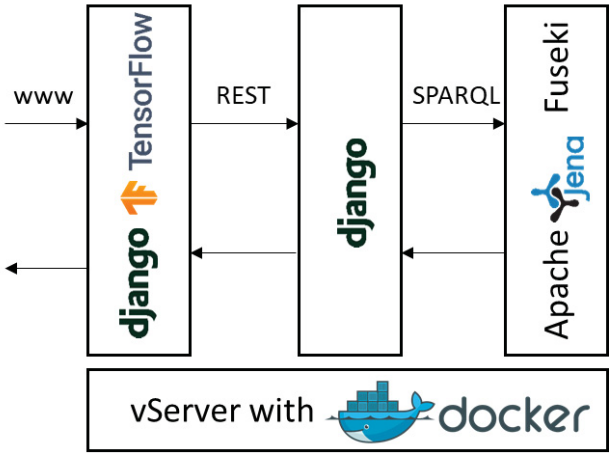

Fig. 4. Architecture of semantic augmentation service

V. PRELIMINARY EVALUATION

The following section shows the results of the object detection and semantic augmentation for Fig. 1. Please note that the newly developed software is currently not yet ready for a thorough evaluation but is mainly a proof of concept for the feasibility of integrating semantic into the Open Images recognition framework. Therefore, the section's goal is primarily the motivation for the further planned developments described in the following section.

TABLE I. FIRST TEN RESULTS FOR THE CALCULATION OF FIG. 1

Open Image Recognition	**Semantic Augmentation (Scenes)**
Tent: 95%	castle: 100.0%
Car: 76%	fountain: 100.0%
Tent: 65%	market_outdoor: 100.0%
Wheel: 63%	moat_water: 100.0%
Man: 54%	promenade: 100.0%
Person: 51%	public-place: 100.0%
Tent: 38%	ruin: 100.0%
Hiking equipment: 35%	viaduct: 100.0%
Tree: 33%	airplane_cabin: 84.98%
Person: 32%	airport_terminal: 84.98%

Table 1 shows the items detected by the image recognition and the semantic augmentation service. The low quality of the mappings is explained through the intermediate object `Occasion`. As the mapping originates from a fashion-domain ontology, these mappings are not fully transferable to a general-purpose domain. Fig. 5 presents the relationships of the class `Tent` towards the items that do not fit, like `airport_terminal` or `ruin`. While it is arguably possible to wear items fitting into a camping situation also during a visit to a castle or a viaduct, it does not verify these items' general fit. While the technology stack is now ready, the refinement for the underlying data is the necessary next step.

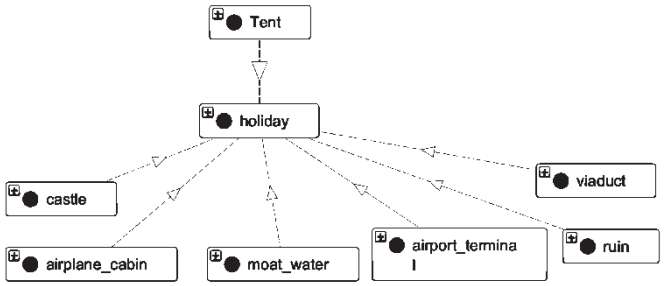

Fig. 5. The mapping through the intermediate object occasion is currently too coarse

Besides the values shown above, the semantic augmentation also inferred 70 additional items associated with SC-values ranging from 84,98% to 15,02 %. While some of them fit the given situation like `campsite`, `camera`, `hills`, `orchard`, `field road`, `hiking equipment`, `hayfield`, `shorts`, and `reception`, others like `mausoleum`, `airplane_cabin`, or `airport_terminal` are not matching the situation.

VI. Discussion and Further Outlook

This paper presented a novel connection of deep learning technologies and ontological reasoning, creating a semantic augmentation for the Open Images database. This augmentation also comprises a newly developed calculation for semantic confidence.

The proof of concept is the first step towards generalized, domain-independent ontology augmentation for public training-data. The new approach was revealed to be technically feasible. The semantic can infer additional knowledge without annotating new detectors and retraining the neural net expensively. However, the current approach also has its limitations. The inferred objects are, at times, insufficient regarding the accuracy of the predictions. While this sometimes originates in a false prediction of the neural net, it mostly happens due to the ontology's coarse relation mapping. The semantic confidence (SC) roots merely on the occurrence-count of similar reasoned objects, which does not always reflect a concept's importance. The missing fine granularity further amplifies this problem.

Future efforts should consider the further extension of the ontology to create more meaningful relationships to tackle the shortcomings. The mapping between `scenes` and the detected elements needs to have a higher resolution. Besides the concept `scenes` that mainly represent locations, the further extension with additional items can increase possible applications. Furthermore, the current simple n:m mapping in the ontology does not unleash semantic technologies' full potential. Utilizing owl's full expressivity to create complex relationships between detected objects could drastically improve the analysis's significance. An extension of the semantic confidence (SC) value to include the relative size of an object in the image analysis can further increase its' significance.

Acknowledgment

Parts of the code resulted from a student project. We thank Roman Kempke, Maximilian Niese, Henry Hilse, and Tim Schröder for their support in realizing the prototype.

References

[1] I. Krasin *et al.,* "OpenImages: A public dataset for large-scale multi-label and multi-class image classification," *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017.

[2] Z. Ding, L. Yao, B. Liu, and J. Wu, "Review of the Application of Ontology in the Field of Image Object Recognition," in *Proceedings of the 11th International Conference on Computer Modeling and Simulation - ICCMS 2019*, North Rockhampton, QLD, Australia, 2019, pp. 142–146.

[3] S. Bhandari and A. Kulikajevas, "Ontology based image recognition: A review," *CEUR Workshop Proceedings*, vol. 2145, 2018. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85050967415&partnerID=40&md5=4d2e4f8d3188f101d97047cfcbfdd193

[4] Alejandro Zambrano, Carlos Toro, Cesar Sanín, Edward Szczerbicki, Marcos Nieto, and Ricardo Sotaquira, "Video Semantic Analysis Framework based on Run-time Production Rules - Towards Cognitive Vision," doi: 10.3217/jucs-021-06-0856.

[5] S. Abburu and R. J. Anandhi, "Concept ontology construction for sports video," in *Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India - A2CWiC '10*, Coimbatore, India, 2010, pp. 1–6.

[6] M. A. Musen, "The Protégé Project: A Look Back and a Look Forward," *AI matters*, vol. 1, no. 4, pp. 4–12, 2015, doi: 10.1145/2757001.2757003.

[7] A. Reiz and K. Sandkuhl, "Design Decisions and Their Implications: An Ontology Quality Perspective," in *Lecture Notes in Business Information Processing, Perspectives in Business Informatics Research*, R. A. Buchmann, A. Polini, B. Johansson, and D. Karagiannis, Eds., Cham: Springer International Publishing, 2020, pp. 111–127.

[8] J. Park, D. H. Kim, Y. S. Shin, and S. Lee, "A comparison of convolutional object detectors for real-time drone tracking using a PTZ camera," in *2017 17th International Conference on Control, Automation and Systems (ICCAS)*, Jeju, Oct. 2017 - Oct. 2017, pp. 696–699.