VisionKG: Towards A Unified Vision Knowledge Graph

Anh Le-Tuan[1], Trung-Kien Tran[2], Manh Nguyen-Duc[1], Jicheng Yuan[1], Manfred Hauswirth[1,3], and Danh Le-Phuoc[1,3]

[1] Open Distributed Systems, Technical University of Berlin
[2] Bosch Center for Artificial Intelligence, Renningen, Germany
[3] Fraunhofer Institute for Open Communication Systems, Berlin, Germany

Abstract. Computer Vision (CV) has recently achieved significant improvements, thanks to the evolution of deep learning. Along with advanced architectures and optimisations of deep neural networks, CV data for (cross-datasets) training, validating, and testing contributes greatly to the performance of CV models. Many CV datasets have been created for different tasks, but they are available in heterogeneous data formats and semantic representations. Therefore, it is challenging when one needs to combine different datasets either for training or testing purposes. This paper proposes a unified framework using the Semantic Web technology that provides a novel way to interlink and integrate labelled data across different data sources. We demonstrate its advantages via various scenarios with the system framework accessible both online and via APIs.[4]

Keywords: Semantic Web · Knowledge Graph · Computer Vision Dataset

1 Motivation and Contributions

Image datasets (e.g., ImageNet [6], COCO [8], etc.) contribute greatly to the current success of deep learning in computer vision (CV). The quality of a trained deep neural network (DNN) is influenced by not only the advanced architecture and optimisation of the DNN but also the annotations and images used for training, validating and testing [12]. The number of labelled datasets has been rapidly growing, and working with different datasets is desirable (e.g., to resolve the out-of-distribution problem and to increase the robustness of CV models [4, 7]). However, the labels are available in heterogeneous formats and are not consistent across datasets. As illustrated in Figure 1, the `pedestrian` in KITTI dataset [3] or the `man` in Visual Genome dataset [5]) are annotated as `person` in

[4] https://vision.semkg.org

COCO dataset. Therefore, it is challenging when one needs to combine different datasets for training or testing purposes.

Recently, the Semantic Web technologies have offered a flexible and powerful mechanism to integrate data from different sources [1]. However, such technologies have been used in very limited settings to manage CV datasets. Prominent CV datasets, such as ImageNet or Visual Genome only use a light form of taxonomy (e.g., WordNet[5]) to label their images. Even when these datasets can be queried, e.g., using SPARQL, the lack of interoperability leads to complex queries that cover all possible cases to unify the labels across different datasets (the left query in Figure 1).

Such shortcomings motivate us to build a unified knowledge graph (KG) to realise the FAIR principles [10] for CV datasets. Our vision is that the ability to interlink labels across label spaces under shared semantic understanding will not only enable a more convenient way to organise training data (e.g., see the right query in Figure 1) but also enable a more robust way to analyse and test trained DNNs. Moreover, this KG can pay the way to enable the interpretability and explainability of the resulting models, e.g. [11, 9].

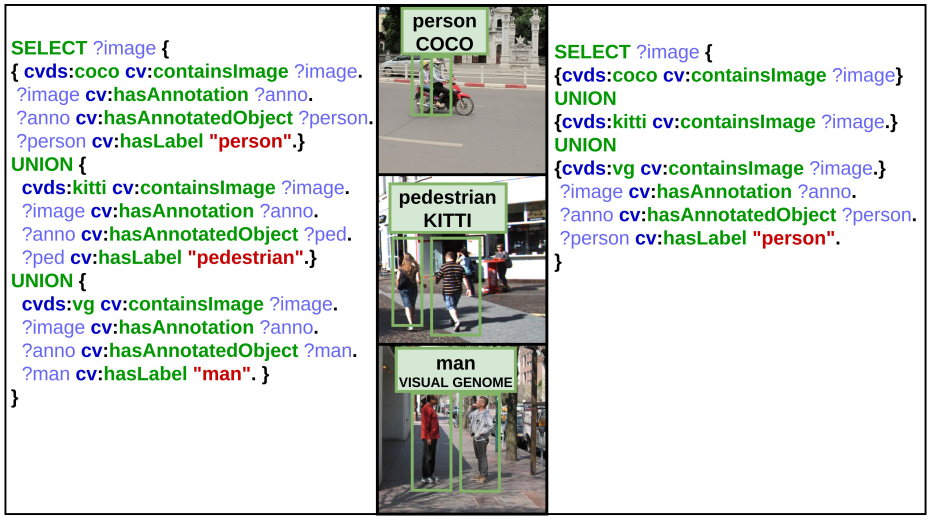

Fig. 1: An example of two equivalent queries to obtain images that contain `Person` from COCO, KITTI, and Visual Genome datasets.

As a step towards the above vision, we propose a unified framework, called VisionKG, that facilitates a novel way to organise CV datasets. Our on-going implementation of VisionKG employs the Semantic Web technology to interlink labelled data across different datasouces. This demo paper will show its advantages via three scenarios: (i) exploring labelled images across datasets, (ii) building training pipelines with mixed datasets , and (iii) validating and testing a trained DNN.

[5] https://wordnet.princeton.edu/

2 Vision Semantic Knowledge Graph

Figure 2 illustrates the overview of our VisionKG framework and the process of creation and enrichment of our unified KG for CV datasets. In step ①, we analyse the structure of the collected CV datasets and propose a unified data model to integrate these datasets. Following the FAIR principles, to make the data ***f****indable*, in step ②, we add metadata and semantic annotations for the data, e.g., what it is about and where it comes from. To make the data ***a****ccessible*, we use RDF and provide a query interface with SPARQL as in step ④. To enhance the ***i****nteroperability*, in step ②, we also link the data with WordNet and Wikidata[6] to reuse the taxonomy of the labels defined by the original sources; and, in step ③, we utilise a reasoner to expand the taxonomy by materialising the labels in each dataset using the ontology hierarchy, e.g., `pedestrian` or `man` is `SubClassOf` `person`. This makes the two queries in Figure 1 equivalent, and thus, helps users to avoid complex queries such as the one on the left. Additionally, in our data model, we utilise existing standardised ontologies whenever it is possible, which makes it ***r****eusable*, i.e., the metadata and data are well described and are ready to be used in different settings.

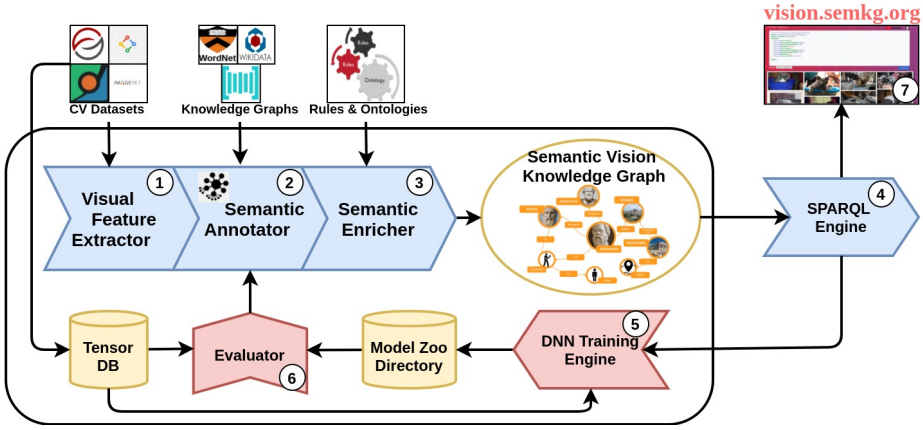

Fig. 2: The overview of VisionKG

Additionally, VisionKG system includes a front-end web interface ⑦ that allows users to explore the KG as shown in the first scenario of our demonstration. Furthermore, our framework contains a DNN Training Engine ⑤ and an Evaluator ⑥. The image data as the tensor inputs for the Evaluator and the DNN Training Engine can be stored in a tensor storage (i.e. TensorDB). And the labels can be retrieved with SPARQL as demonstrated in the second scenario. The trained models are stored in our Model Zoo Directory and are evaluated by the Evaluator.

The tutorials for the training pipelines based on [2] are available online at **https://github.com/cqels/vision**. Most of the data preparation and con-

[6] https://www.wikidata.org/

figuration steps are automated so that Semantic Web developers familiar with SPARQL can easily try out the pipelines.

In the current version (by August, 2021), VisionKG has 67 million triples which cover Visual Genome, COCO, and KITTI datasets with the total number of 239k images, a million of labels (including ones for bounding-box), and hundreds of object categories. These categories are reused but aligned with Wikidata concepts/classes. VisionKG also contains millions of detection results that are evaluated with popular pretrained models such as Yolov3, Yolov4, EfficientDet, FRCNN, etc.

3 Feature Demonstrations

The demonstration session will consist of three scenarios with a demonstration video at **https://vision.semkg.org/iswc2021-demo.html**. For all mentioned scenarios, we provide both Python APIs for developers and a Web interface for end-users in the training/testing phase and data exploration phase respectively.

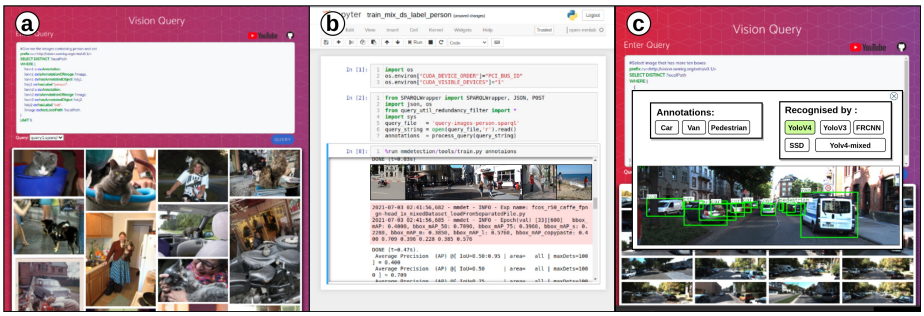

Fig. 3: The screenshots of the demonstation with VisionKG

ⓐ **Graph-based Exploration Across Visual Label Spaces:** In this scenario, we demonstrate our web-based image explorer that is used for retrieving images using SPARQL (Figure 3 ⓐ). The demonstration shows that with VisionKG, users will be able to search for images containing different labels, e.g., images that contain a cat and a person; or images that have 10 cars.

ⓑ **Building Training Pipelines with Mixed datasets:** In the second part, we demonstrate the scenario to obtain a mixed dataset for training purposes (Figure 3 ⓑ). A user starts a training pipeline by writing a SPARQL to retrieve the images and labels as desired. This includes an advanced setting like merging training data with the same label, e.g. `Person`, from different datasets (as shown in the right query of Figure 1).

ⓒ **Cross-dataset Validation and Testing:** The third part demonstrates the scenario of getting the mixed dataset for validating purposes (Figure 3 ⓒ).

Similar to the scenario for training data, this includes the case where test data with the same labels are combined from different datasets. In advanced settings, one can test different models on specific labels in one specific dataset or over different datasets. This scenario is particularly useful in the case that developers want to target specific applications. For example, one can test the trained models to detect `Car` on images of `Car` in crowded traffic scenes or in mountain areas.

4 Next Steps

Our proposed framework opens various new research venues. First, we plan to build DNNs with a unified label space powered by VisionKG. Next, we will extend VisionKG to analyse the robustness of DNNs models using testing samples with semantic similarities via KG embeddings. Such KG embeddings can be combined with visual features of DNN-based visual models to investigate the interpretability and explainability of these models, e.g. [9, 11].

Acknowledgments This work was funded by the German Research Foundation (DFG) under the COSMO project (ref. 453130567), the German Ministry for Education and Research via The Berlin Institute for the Foundations of Learning and Data (BIFOLD, ref. 01IS18025A and ref. 01IS18037A), and the German Academic Exchange Service (DAAD, ref. 57440921).

References

1. Benjelloun, O., Chen, S., Noy, N.F.: Google dataset search by the numbers. In: 19th International Semantic Web Conference Proceedings. Springer (2020)
2. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al.: Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
3. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research (2013)
4. Huang, R., Li, Y.: Mos: Towards scaling out-of-distribution detection for large semantic space. In: Proceedings of the IEEE/CVF (2021)
5. Krishna, R., et la.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV (2017)
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. NeurIPS (2012)
7. Lambert, J., Liu, Z., Sener, O., Hays, J., Koltun, V.: MSeg: A composite dataset for multi-domain semantic segmentation. In: CVPR (2020)
8. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
9. Park, J.S., Bhagavatula, C., Mottaghi, R., Farhadi, A., Choi, Y.: Visualcomet: Reasoning about the dynamic context of a still image. In: ECCV (2020)
10. Wilkinson, M.D., et la.: The fair guiding principles for scientific data management and stewardship. Scientific data **3**(1), 1–9 (2016)
11. Zareian, A., Karaman, S., Chang, S.: Bridging knowledge graphs to generate scene graphs. In: ECCV 2020 (2020)
12. Zhu, X., et la.: Do we need more training data? IJCV (2016)