Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2019) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

Third International Conference on Computing and Network Communications (CoCoNet'19)

COCO (Creating Common Object in Context) Dataset for Chemistry Apparatus

Silvia Rostianingsih, Alexander Setiawan, Christopher Imantaka Halim

Informatics Department, Petra Christian University, Siwalankerto 121-131, Surabaya 60236, Indonesia

Abstract

In order to create machine learning, we need to build a model. The model is created from a process called training. The goal of training is to develop an accurate model that answers some questions and in order to train a model, we need to collect a dataset. The quality and quantity of the data gathered will determine how good the predictive model can be. Helping the model to understand datasets like humans do is one of the important processes of machine learning. Datasets need to be constructed and transformed correctly. In this research, we compare the difference between creating a COCO dataset manually and creating a synthetic COCO dataset. Creating datasets for chemistry apparatus is not as difficult as creating a human object. The apparatus has a specific shape and form, thus the dataset had to have a limited number. As a result, we create both a dataset both manually and synthetically. The synthetic dataset helps to gain more datasets by combining some objects with different backgrounds.

Keywords: COCO dataset; synthetic dataset, annotate, chemistry apparatus

1. Introduction

One important process of machine learning is helping the model to understand a dataset like humans do. Dataset are required to be constructed and transform correctly. However, some problems arise when dealing with data. First, one limitation to developing a good model is limited data [1]. Second, an imbalanced dataset presents difficulties for learning models to achieve high performances [2]. Third, some privacy issues occur when dealing with information about the individual [3] [4]. To balance the dataset, we need to collect data equally. Various techniques can be used

to collect datasets such as manually, using API, doing web scraping, using a dataset platform or generating synthetic dataset.

To generate a synthetic dataset, many algorithms are used depending on the data type. For an online social network dataset, graph topology can be used [3]. If it comes with images, some methods include UCSD Anomaly Detection [1] [5], procedural randomization and manual modelling [6] or image masking and vertices [7].

In this research, we are going to create COCO dataset of chemistry apparatus, which COCO has not provided it yet. The chemistry apparatus has a certain forms, thus most of them have a transparent form. In order to enrich the dataset, we also create synthetics dataset. We are going to make a comparison between creating the COCO dataset manually and creating the synthetic COCO dataset.

2. Literature

2.1. Common Objects in Context (COCO) dataset

COCO is a large-scale object detection, segmentation and captioning dataset [8]. They use several sources to collect object categories, which are grouped into three types, namely iconic-object images, iconic-scene images and non-iconic images. Iconic-object images have a single large object that appeals to the image. Iconic-scene images are recognized as a scene without any object being obvious. A non-iconic image is an image with some objects. It showed that datasets with more non-iconic images are better to use for generalizing. Three stages are used to annotate the image collections. First, category labelling is used to determine which object categories are present in each image. Second, instance spotting is used to label each instance of a specific category found in the previous stage. The final stage is segmenting each object instance.

COCO is a common dataset format used by Microsoft, Google, and Facebook. It is used as a benchmark to measure machine learning algorithm performance. The other common datasets are PASCAL VOC and ImageNet.

2.2. Synthetic dataset

Synthetic data generation is an alternative method for creating a large number of datasets. There are two techniques to create synthetic dataset [9]. First, we create fully synthetic data by making data synthetically. Second, we can combine the synthetic dataset with real data, by replacing some sensitive attribute values. Beside the privacy issue, synthetic dataset is also effective to evaluate algorithms and their usefulness. Mayer found points that need to be considered in order to create a synthetic dataset, namely (1) diversity is important, (2) realism is overrated, (3) learning schedules matters and (4) camera knowledge helps [6].

2.3. Chemistry apparatus

For our collections, we are creating datasets for chemistry apparatus such as a beaker, Erlenmeyer flask, Florence flask, graduated cylinder and petri dish. Each contains 200-300 datasets. We gathered the data from taking pictures at chemistry laboratories located at Sepuluh Nopember Institute of Technology, Surabaya. We are focusing on chemistry laboratories for first to second-year students, because they often use basic chemistry apparatus. We were visiting four laboratories from three faculties, which are Department of Chemistry, Department of Chemical Engineering and Department of Industrial Chemical Engineering. The example of images can be seen in Fig. 1.

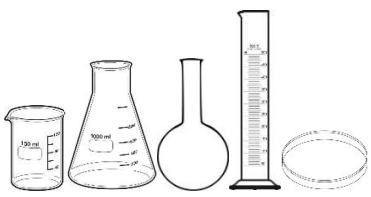

Fig. 1. Beaker, erlenmeyer flask, florence flask, graduated cylinder and petri dish

The beaker, erlenmeyer flask, florence flask, graduated cylinder and petri dish are some examples of the chemistry apparatus. The beaker can be made of plastic or glass. We only gathered data from a beaker glass, which can be heated. The common sizes are 50 mL, 100 mL, 250 mL and 400 mL. The erlenmeyer flask is made of glass, which can be heated and used in titration. The common sizes are 100 mL or 250 mL. The florence flask is made of glass, which can be heated and used in making or storing solution. The common sizes are 125 mL, 250 mL and 500 mL. A graduated cylinder can be made of glass or plastic, which is used to measure approximate values; however, it must not be heated. The common sizes are 10 mL, 50 mL and 100 mL used to measure approximate volumes, must not be heated.

3. Methodology

3.1. Creating COCO dataset manually

In order to create the COCO dataset manually (Fig. 2), first we specify the chemistry apparatus labels. As mentioned previously, labels would be Erlenmeyer flask, Florence flask, graduated cylinder and petri dish.

Second, we collect the images from laboratories at Sepuluh Nopember Institute of Technology, Surabaya. For each category, there are approximately 200-300 images. The images are taken with the original background. Some of the objects had an overlap position. We also scraped the Google Image Search [10], using a python helper software. Herewith the script:

googleimagesdownload --keywords " beaker, erlenmeyer flask, florence flask, graduated cylinder, petri dish"

Last, we annotate the images as COCO format. We are using an open-source annotation software, which can automatically produce a COCO formatted data [11]. This step is the most time-consuming.

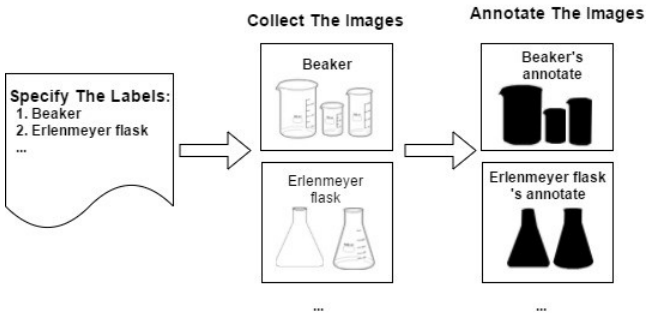

Fig. 2. Methodology of creating COCO dataset manually

3.2. Creating synthetic COCO dataset

In order to create a synthetic COCO dataset (Fig. 3), first we specify the chemistry apparatus labels. As mentioned earlier, labels would be erlenmeyer flask, florence flask, graduated cylinder and petri dish. Second, we choose the foreground and background image, and then we combine with the chemistry apparatus which had already been annotated.

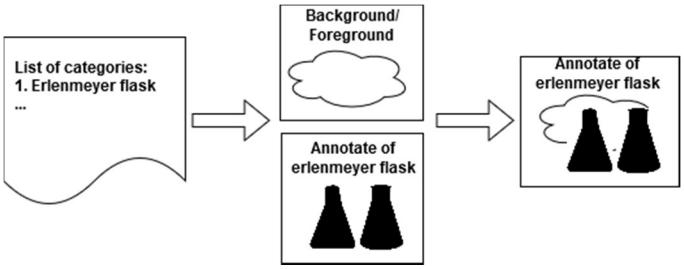

Fig. 3. Methodology of creating synthetic COCO dataset

4. Implementation

4.1. Creating COCO dataset manually

COCO is a standard dataset format for annotating the image collection, which is used to for data preparation in machine learning. Annotate means to create metadata for an image. From Fig. 2, we know there is an image named beaker and we know the position (x and y coordinate, width, length, and polygon area).

The labels from Fig. 2 will be the category of the COCO dataset (Fig. 4). The number described in each category is the number of objects already annotated.

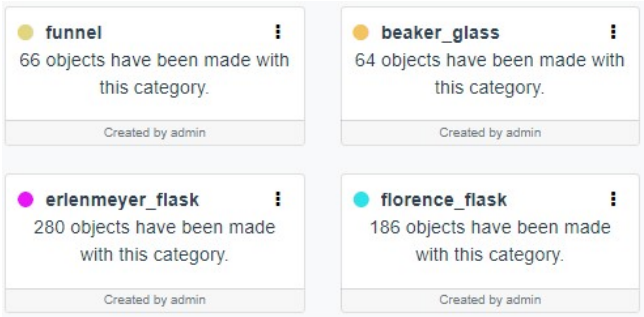

Fig. 4. Category of the dataset

Fig. 5. The images of chemistry apparatus

Fig. 6. The objects with their annotates

Fig. 7. The categories and their objects

Fig. 5 contains many chemistry apparatus objects, where more than 10 objects are displayed. For this example, we are only annotating seven objects with three categories (Fig. 6). The categories are the beaker glass, Erlenmeyer flask and Florence flask (Fig. 7). For the beaker glass, there are four objects that represent with yellow color. For the Erlenmeyer flask, there are two objects that represent with the magenta color, For the Florence flask, there is one object that represents with cyan color.

After creating the COCO dataset by using the tools from Justin Brooks, we download the json. The json structure contains images, categories and annotations. The images contain an id (auto increment), dataset_id (the folder id), file name, width, height, and file location (Fig. 8).

Categories are the information of the labels and the color (Fig. 9). Annotations contain segmentation, bounding box and color (Fig. 10). Segmentation contains the position of each vertex from the annotation. The bounding box is the maximum x and y of the image. The result from this manually annotate the objects can be seen in Fig. 11. We had a set of datasets that annotate manually. Average time to manually annotate per objects is three and a half minutes.

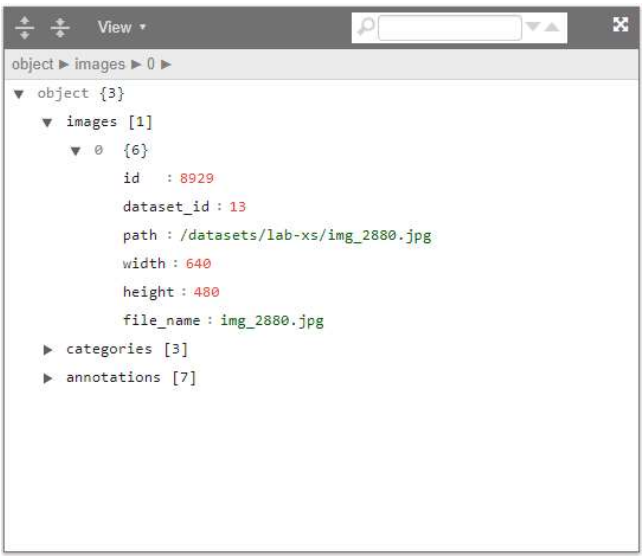

Fig. 8. The structure of json (1)

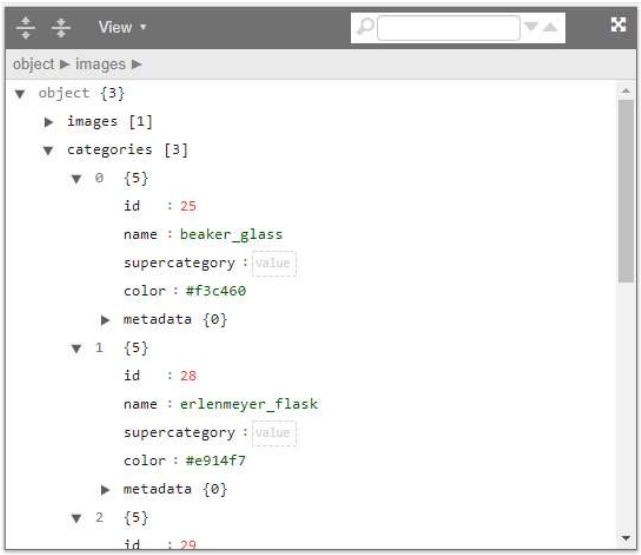

Fig. 9. The structure of json (2)

4.2. Creating Synthetic COCO dataset

Creating a synthetic dataset is challenging; we need to combine many backgrounds and objects. The background for synthetic dataset would be the laboratory itself. We choose ten backgrounds for five objects and we generate 500 annotations for each objects. The json format will be same as Fig. 7 – Fig. 11. We also used the tools from Brooks [11]. The result from the synthetic dataset can be seen in Fig. 12. With the same background, we put the funnel, beaker, Erlenmeyer flask and Florence flask in a different place. The average time to create a synthetic dataset per object is 0.05 minute.

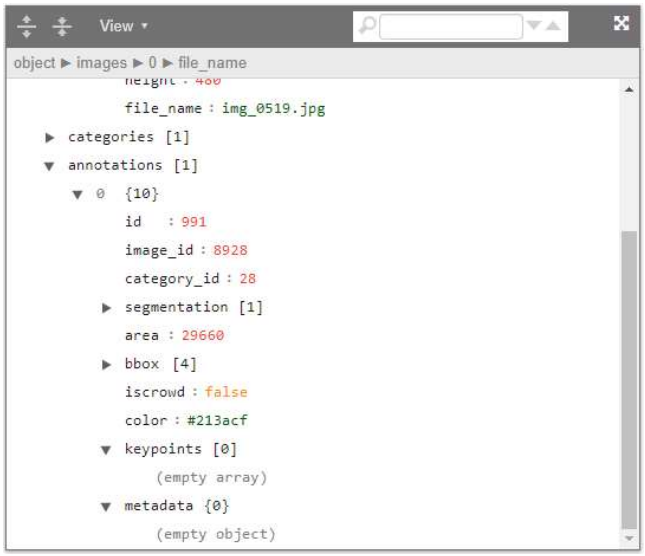

Fig. 10. The structure of json (3)

Fig. 11. The result of manual annotate

Fig. 12. Synthetic dataset with laboratory background

5. Conclusion

Chemistry apparatus has a specific shape and form; therefore, the dataset had a limited number. In order to enrich the dataset, we need to collect images with many backgrounds or foregrounds. A synthetic dataset helps to collect more datasets by combining background and object. We placed some overlap objects in one background to get different viewpoints. For further research, the COCO dataset will be used for recognize chemistry apparatus using machine learning.

Acknowledgements

Special thanks to our Ministry of Research, Technology and Higher Education for supporting this research financially. We also thank Sepuluh Nopember Institute of Technology, Surabaya for allowing us to gather dataset information from a chemistry laboratory.

References

[1] Ekbatani, Hadi Keivan, Oriol Pujol, and Santi Segui. (2017). "Synthetic data generation for deep learning in counting pedestrians." *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*.

[2] Der-Chiang, Li, Hu Susan C., Lin Liang-Sian, and Yeh Chun-Wu. (2017). "Detecting representative data and generating synthetic samples to improve learning accuracy with imbalanced data sets." *PLoS One* **12** (**8**): e0181853.

[3] Nettleton, David. (2016). "A synthetic data generator for online social network graphs." *Social Network Analysis and Mining* **6**, 44.

[4] H, Surendra and Mohan H. S. (2017). "A review of synthetic data generation methods for privacy preserving data publishing." *International Journal of Scientific & Technology Research* **6 (3)**: 95–101.

[5] Li, Weixin, Vijay Mahadevan, and Nuno Vasconcelos. (2014). "Anomaly Detection and Localization in Crowded Scenes." *IEEE Transactions on Pattern Analysis and Machine Intelligence* **36** (**1**): 18–32.

[6] Mayer, Nikolaus, Eddy Ilg, Philipp Fischer, Caner Hazirbas, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. (2018). "What makes good synthetic training data for learning disparity and optical flow estimation?" *International Journal of Computer Vision* **126** (**9**): 942–960.

[7] Remez, Tal, Jonathan Huang, and Matthew Brown. (2018) "Learning to segment via cut-and-paste*." Computer Vision – ECCV 2018* **11211**: 39-54.

[8] Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, C. Lawrence Zitnick. (2014) "Microsoft COCO: common objects in context." *Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science* **8693**: 740-755.

[9] Dandekar, Ashish, Remmy A. M. Zen, and Stéphane Bressan. (2018) "A comparative study of synthetic dataset generation techniques." *Database and Expert Systems Applications. DEXA 2018. Lecture Notes in Computer Science* **11030**: 387-395.

[10] Vasa, Hardik. (2019) "Google images download." *Online:* https://github.com/hardikvasa/google-images-download.

[11] Brooks, Justin. (2019) "COCO Annotator." *Online:* https://github.com/jsbroks/coco-annotator/.