Department of Electrical and Computer Engineering

North South University

CSE498R: Directed Research

Out of Context Object Detection

Jubayer Hossain Arnob 1813124042

Fahima Noor 1912052042

Faculty Advisor:

Dr. Mohammad Ashrafuzzaman Khan

Assistant Professor

Department of ECE

Spring 2022

LETTER OF TRANSMITTAL

June, 2022

To

Dr. Rezaul Bari

Associate Professor & Chair

Department of Electrical and Computer Engineering,

North South University, Dhaka.

Subject: Submission of Directed Research Project on “Out of Context Object Detection”.

Dear Sir,

With due respect, we would like to submit our Directed Project Report on “Out of Context Object Detection” as a part of our BSc program. The report deals detection of objects that may be out of context to the other objects and the scene presented within an image. This project may prove to be useful in the field of autonomous driving and surveillance. We tried to make the report as informative as possible by including all the workflow of the project.

The Directed Research project proved to be quite useful for us to gain a clear concept and to have a more in depth knowledge about the practical aspects of the project. We carried out all the necessary task that was required by us for the completion of the Directed Research project.

We will be grateful if you are kind enough to accept this report. We hope that this report proves to be informative and beneficial to the field of research on which the project is done.

Sincerely Yours,

.....

Jubayer Hossain Arnob

1813124042

Department of ECE

North South University, Bangladesh

.....

Fahima Noor

1912052042

Department of ECE

North South University, Bangladesh

APPROVAL

The directed research project entitled “Out of Context Object Detection” by Jubayer Hossain Arnob (ID 1813124042) and Fahima Noor (ID 1912052042), is approved in partial fulfillment of the requirement of the Degree of Bachelor of Science in Computer Science and Engineering on June, 2022 and has been accepted as satisfactory.

Supervisor:

.....

Dr. Mohammad Ashrafuzzaman Khan

Assistant Professor

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh.

Department Chair:

.....

Dr. Rezaul Bari

Associate Professor & Chair

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh.

DECLARATION

This is our truthful declaration that the “Directed Research Project Report” we have prepared is not a copy of any “Directed Research Project Report” previously made by any other team. We also express our honest confirmation in support of the fact that the said “Directed Research Project Report” will not be submitted to any other team or authority in future.

.....

Jubayer Hossain Arnob

1813124042

Department of ECE

North South University, Bangladesh

.....

Fahima Noor

1912052042

Department of ECE

North South University, Bangladesh

ACKNOWLEDGEMENT

First of all, we wish to express our gratitude to the Almighty for giving us the strength to perform our responsibilities and to complete the report. We would like provide our uttermost gratitude to our supervisor Dr. Mohammad Ashrafuzzaman Khan for advising and guiding us throughout the project. This project would have not been possible without his guidance. We would like thank the ECE department of the North South University for providing us with the opportunity to gain knowledge and experience from the project. Finally, we would like to thank our families who have been giving us their unwavering support throughout the completion of our degree.

ABSTRACT

Image segmentation is a sub-domain of computer vision and digital image processing which aims at grouping similar regions or segments of an image under their respective class labels. Image segmentation forms the basis for object detection. Object detection is fairly common in computer vision as it is beneficial in fields of Robotics and object recognition. However, there is a lot of scope on working with Out of Context Objects Detection in images which is not being done. Therefore, in this project, we chose Detectron2, a segmentation model, and a highly annotated contextual dataset such as the COCO train17 dataset and used it to identify objects in contextual image. We implemented the image segmentation model and results show that the model was successful in identifying and segmenting different classes and objects from scenes containing many different types of objects. The object detection model is used to create a dataset of objects commonly found in contextual scenarios and the dataset is used to train the Word2Vec model. The trained Word2Vec holds the capability to identify any out of context objects from a list of objects presented to it.

Table of Contents

CHAPTER 1: INTRODUCTION

In this section, we will discuss about the image segmentation, overview the different types of image segmentation, image segmentation models and how we can use image segmentation and utilize the contents found in the image to identify the out of context object in an image.

1.1 IMAGE SEGMENTATION

Image Segmentation is considered one of the main steps in image processing. Image segmentation allows identifying objects from the image. It does so by dividing regions of images into classes by grouping pixels that have similar attributes under one category. Image segmentation is highly valuable in fields like robotics, medical imaging, and image recognition. As image segmentation allows the contents within an image can be identified and analyzed, it is highly valuable in medical image processing fields such as the identification of cancerous cells from healthy cells or surveillance. It is also beneficial in self-driving cars where identifying nearby objects on the road is crucial for autonomous driving.

1.2 IMAGE SEGMENTATION TYPE

There are many types of image segmentation. The most common form of image segmentation is semantic image segmentation, instance segmentation and panoptic segmentation. The segmentation style of these are shown in Fig. 1 [1]. In semantic image segmentation, all the objects are segmented into different classes. All the instances of a class will be treated same. However, in instance segmentation, even if all the instances may fall under one class, it would still be treated as different objects as shown in the figure. However, panoptic segmentation is a mixture of both semantic and instance segmentation. Since our project will utilize the contextual understanding using image segmentation, we have chosen to use the panoptic segmentation for our image processing.

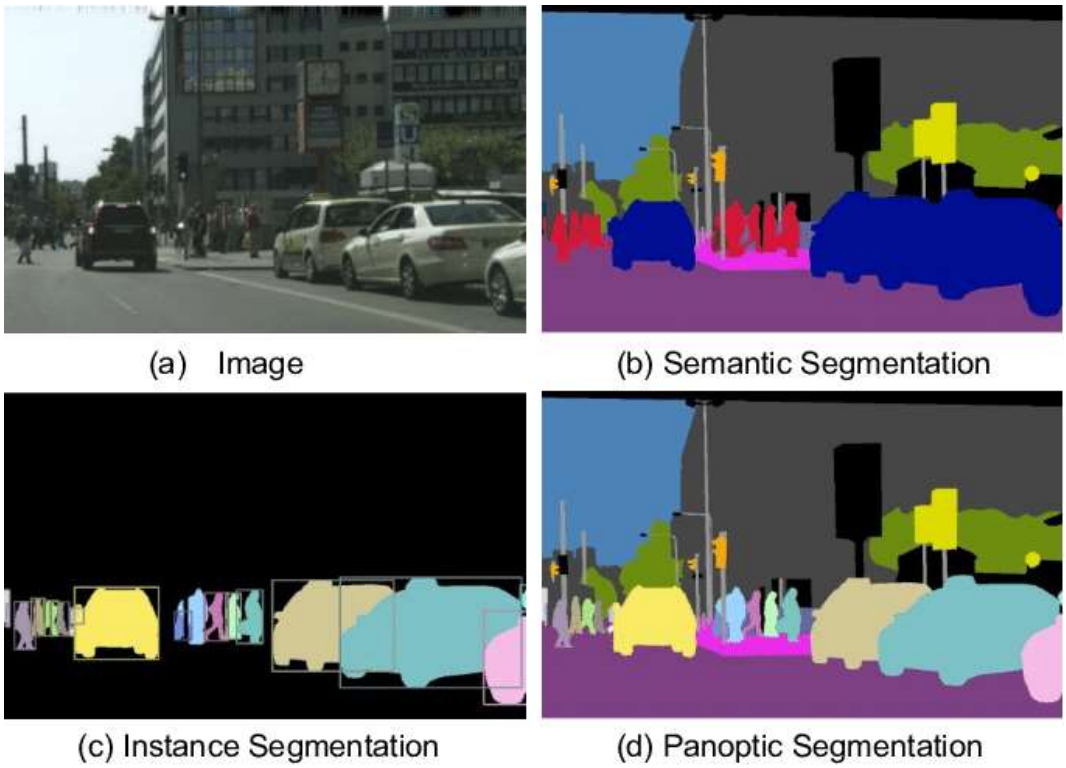

Figure 1: (a) represents an image and (b)-(d) represents different type of segmentation for the images [1]

1.3 Out of Context Object

When objects are said to be in-context, it refers to the fact that all the objects images are in context to each other and with the environment they are situated in. Thus, contextual images are comprised of objects that are normally found in the natural setting of the environment. Example of contextual images are shown in Fig. 2.

Figure 2: Contextual Objects images from COCO17. [2]

Since the main idea was to utilize the information of objects, for out of context images, we choose to work with those images where an object is present in situation where it should not be. For example, fruit-basket inside a washroom or a giraffe inside a bedroom. These scenes are unnatural and they violate the natural scenario. As such, these objects are taken to be out of context. Examples of different out of context images used for the project are shown in Fig. 3.

Figure 3: Images containing out of context images. [3]

1.4 Project Aim and Objective

The objective and project aim was to utilize the content in an image to make sense of contextual information of the scene and use it to identify out of context object present in a scene. To do so, a dataset had to be

created of objects in different contextual setting. Using this information, out of context object will be identified by appropriate model.

1.5 MOTIVATION

The idea of the project can be useful in fields where constant monitoring of a scene is needed such as autonomous driving or surveillance.

1.6 THESIS OUTLINE

In this project, we implement PSPNET [4] on the ADE20K [5] dataset and Detectron2 [6] on COCO17 dataset [2]. However, comparing the performance of object detection model, Detectron2 was selected as the object detection model. The model was used to identify the contents from the COCO17 dataset images, which was then used to build a dataset of contextual objects present for different scenes. This dataset is used to train the Word2Vec [7]. Then, with the help of Detectron2, the Word2Vec model is used to identify the out of context object from out of context images.

CHAPTER 2: LITERATURE REVIEW

This section represents the recent related works that has been on out-of-context image detection as well as how content can be used to make sense of images.

2.1 Existing Literature Explanation

There are studies being done that uses context of an image to make prediction. A study was done by Choi et al [8] where, they used context models to understand out of context objects. They proposed a new model which used the content found in the images from the SUN dataset [9] to verify how much the images where in context to each other. Their context model could be used for object recognition as well as out-of-context object detection. Another study used content to make sense of the context in images [10]. The proposed new techniques to compute content and context features, and use them together for classification. Lastly, a study proposed a new context aware deep learning network that can be used for object detection [11]. The model is able to identify and utilize the contextual clues presented in the images.

CHAPTER 3: METHODOLOGY

This section represents the workflow of our project for the course.

3.1 WORKFLOW

The workflow of the project is shown in Fig. 4.

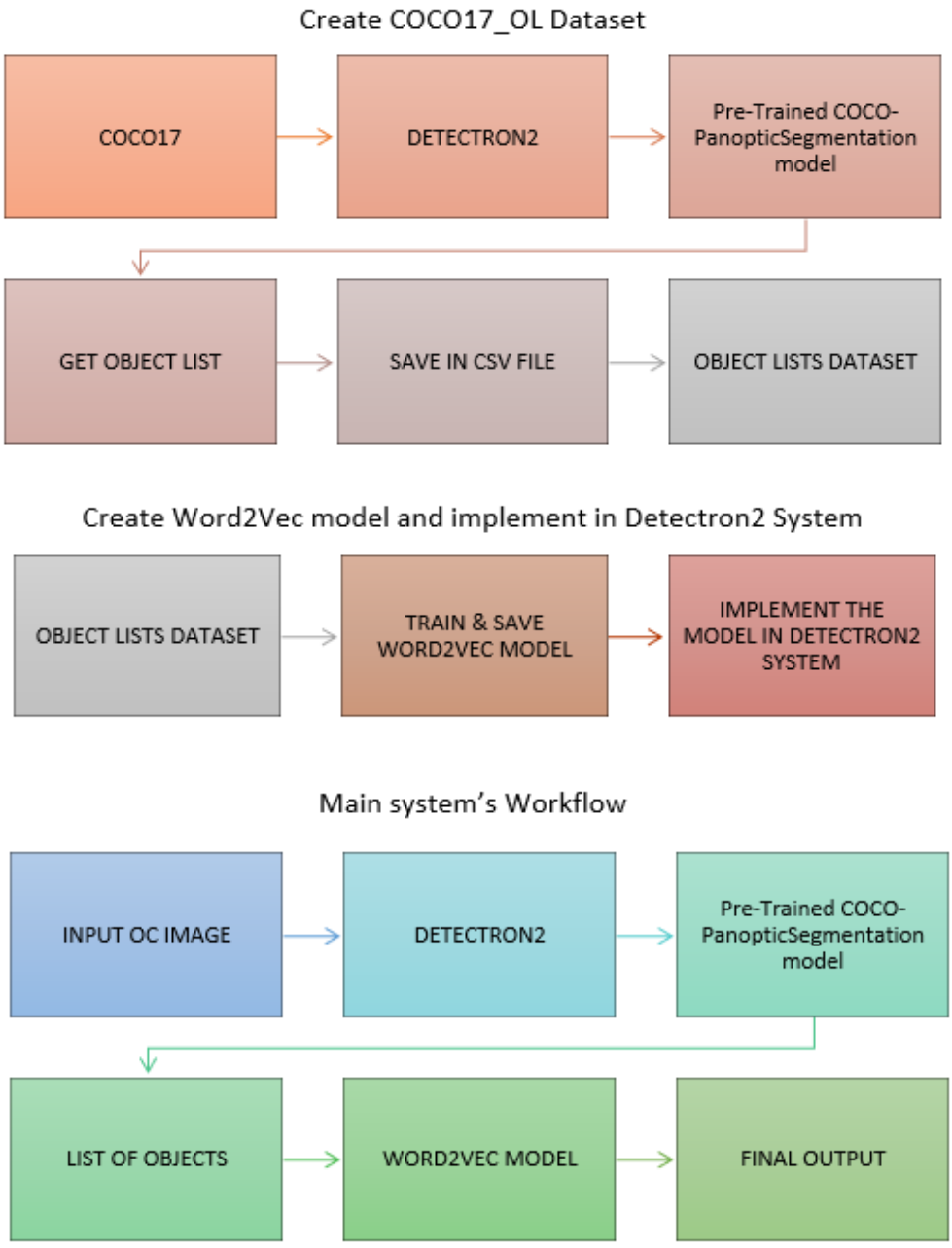

Figure 4: Workflow diagram of the project

The workflow diagram gives an overall idea of the sequence of process that is carried out throughout the project. Firstly, a dataset filled with contextual images were needed. For that, the COCO17 dataset was chosen. COCO17 dataset is comprised of images where all the objects are in-context to each other as well as with the scene it is in. Next, an accurate object detection model was to be selected. PSPNET was chosen at first since it was a more advanced and more accurate segmentation model. However, since the accuracy of the PSPNET depended on its ability to detect objects based on semantic information, it performed poorly on out of context image dataset, where it would detect the out of context object as a wrong object based on its semantic understanding of the scene. As such, PSPNET performed poorly on the detection of out of context images and so a new object detection model was chosen for the project.

PSPNET was replaced with Detectron2. In Detectron2, its Pre-Trained COCO-Panoptic Segmentation model was used for object detection. Unlike PSPNET, it was able to correctly detect the objects specially detecting out of context object just as it is.

The Detectron2 was used to detect images in COCO17 dataset. The objects detected for each of the images are stored in a file. This file is the contextual objects dataset. This dataset is used to train the Word2Vec model. The Word2Vec model has the ability to understand the word association from the corpus of text. It will group together objects that frequently occur together in a tight cluster.

To verify the ability of Word2Vec model on the detection of out of context image, we needed to create a similar type of dataset with which Word2Vec model was trained. Images containing out of context object was taken and Detectron2 was used to identify objects that are found there. Detectron2 was used to create a dataset of list of objects found in these out of context images. For each of the images, the out of context object is manually identified to check the accuracy of Word2Vec model. This dataset is then inputted to the trained Word2Vec model which identifies the object is it thinks is not associated to the other objects in that. Then the selected objects are cross-checked with the ground truth to see whether the detection is correct. Around 88% of accuracy is achieved on using Word2Vec to detect out of context objects.

3.2 DATASET

To build a dataset of contextual objects in different scene, COCO17 [2] is used. COCO train17 dataset has more than a million images that presents common objects in context. As the dataset is built with contextual objects in different scenes, it is used for the project. As we have used this dataset to create the object lists dataset, the dataset has more than a million lists of objects.

To build a dataset of contextual objects with a particular out of context object in different scene, we have mainly scrapped a website called "Getty Images" [3] with particular search tags. Moreover, different images that fit our vision of out of context object images has been either scraped from other websites manually or build to increase the size of the dataset.

3.3 IMAGE SEGMENTATION MODELS

3.3.1 PSPNET

The architecture of the Pyramid Parsing Network (PSPNet) is shown in Fig. 5 [4]. The Encoder part of PSPNET contains CNN backbone with dilated convolutions and a pyramid pooling module. The use of dilated convolution layer increases the receptive fields of the model and thus improves its performance.

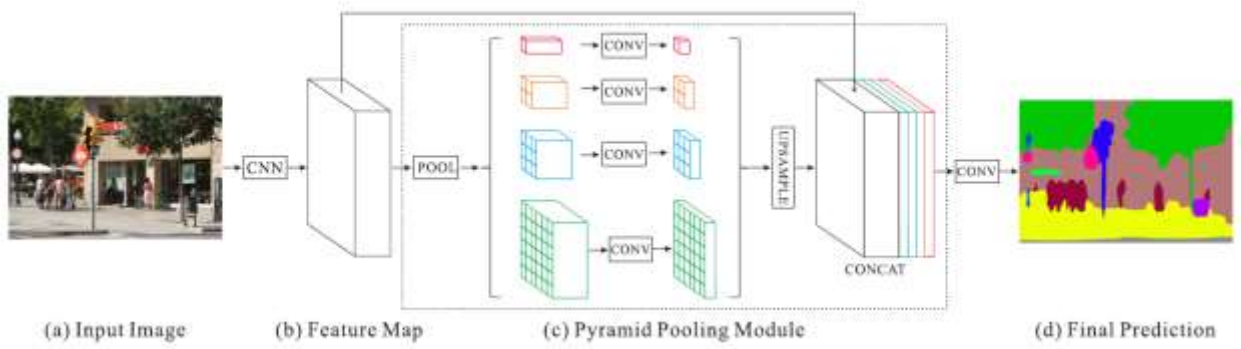

Figure 5: PSPNET Architecture [4]

The way the PSPNET architecture works is that it takes in the global context of the image to predict the local level predictions. This allows the model to make better predictions. The pyramid pooling module is the main part of this model as it the one which capture the global context in the image. This helps the model to classify the pixels based on the global information present in the image.

However, for the project, due to the poor performance of PSPNET on the detection of out of context objects in an image, this object detection model is replaced by Detectron2 [6].

3.3.2 DETECTRON2

Detectron2 is computer vision model next generation library developed by Facebook AI Research (FAIR), which provides state of the art object detection [6]. For the project, we used Detectron2 with its pre-trained COCO panoptic segmentation mode [6]. It's ability to carry out panoptic segmentation makes it a suitable system for the project. As there's no pre-trained model on semantic segmentation, panoptic segmentation model is the closest to detect almost all the objects (both stuff class and thing class) in an image.

3.4 CAPTURING SEMANTIC RELATIONSHIP OF WORDS

3.4.1 WORD2VEC

As the name implies, the Word2Vec model is a two-layer neural net that represents words inside corpus in vector form [7]. It takes in corpus of texts and outputs all the individual words present inside the corpus as vectors. It is mostly useful for task where the patterns within a corpus is needed to be understood. The way Word2Vec works is that it groups together similar type of words together in a vector space. Using enough data and context, Word2Vec model can make accurate word association.

For our project, the word2vec model was trained using 100 epochs. We set the minimum word count to word as we want to work with all the words present in the dataset. The parameters used during the training is as follows: min_count=1, window=3, size=500, sample=6e-05 and alpha=0.03.

CHAPTER 4: RESULT

4.1 IMAGE SEGMENTATION

The Fig. 6 is an example provided on the performance of Detectron2 on image detection of an out of context image. Fig. 5 is a scene which represents a usual street-view [3]. However, there is seen to be a bear crossing a road. Here, it is usual to see bear in a typical street-view and hence the bear is taken to be out of context relating to the scene and the objects that are present within the scene. The Detectron2 model identifies the objects that it sees inside the image.

Figure 5: An image containing an out of context object.

Figure 6: Objects detected by the Detectron2 from the image shown in Fig 5.

4.2 Visualization of Word Embedding Vectors

Fig. 7. the represents the visualization of Word2Vec model trained on the contextual objects dataset. As it can be seen, the Word2Vec clusters associate words closely. Words that represent common objects in street are closer to each other. Similarly, all the other items which are highly related form a close cluster. Another thing to consider is that the objects that are commonly found out-doors are clustered together at the 3rd quadrant whereas common in-door objects are clustered at the 1st quadrant.

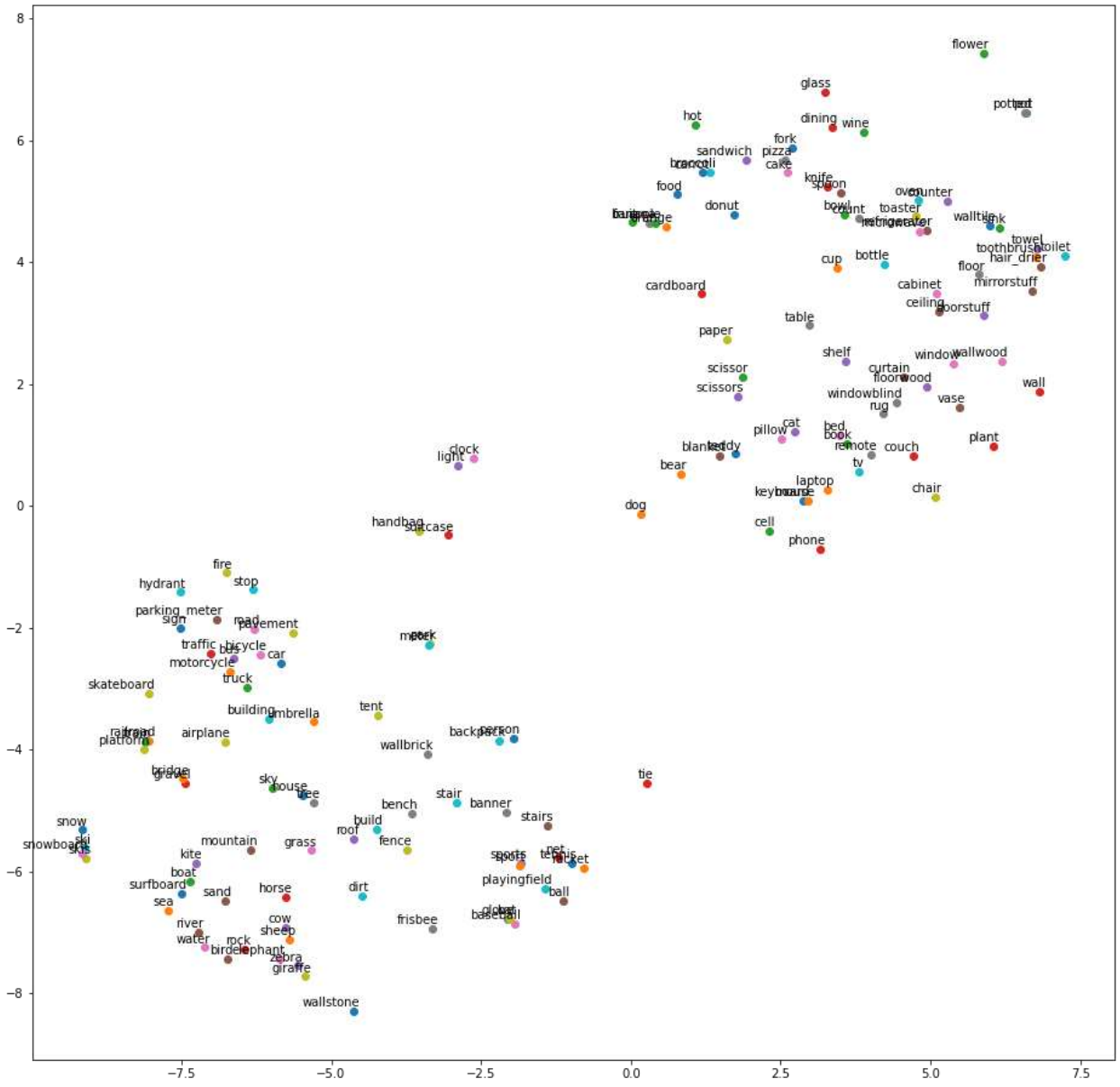

Figure 7: Visualization of Word2Vec model trained on contextual dataset

4.3 PERFORMANCE

The object list that is found by the Detectron2 for the image shown in Fig.5 is shown in Fig. 8. This object list is passed to the Word2Vec model. The Word2Vec model is previously build using contextual object dataset. As such, it identifies the object that does not closely associates to the other objects present within the scene and outputs it as out of context object for that particular image. Since bear is not related to typically

associated to any objects that are present in a normal street-view, it is detected as out of context. Fig. 9. shows the results obtained from Word2Vec for two different images where traffic light is seen to be out of context in one of those images and cat in the other. Around 250 images containing out of context objects are used for the project. The accuracy of out of context object detection by Word2Vec is found to be 88% as shown in Fig. 10. However, there is scope of improving the accuracy by increasing the size of out of context dataset. We have used only 134 objects.

```
['car', 'fire hydrant', 'person', 'bear', 'traffic light', 'road', 'pavement', 'building']
Out of context object: bear
```

Figure 8: Object list for bear and the out of context object detected by Word2Vec model

```
w2v_model.wv.doesnt_match(['traffic light', 'sand', 'sky', 'grass', 'dirt'])
```

```
'traffic light'
```

```
w2v_model.wv.doesnt_match(['cat', 'wall', 'refrigerator'])
```

```
'cat'
```

Figure 9: Out of Context Object detected by Word2Vec model

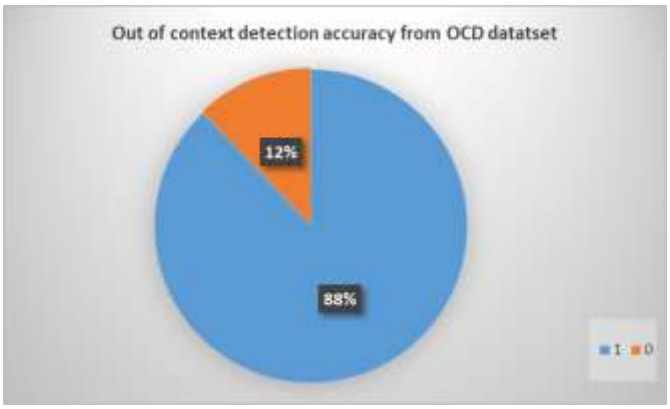

Figure 10: Accuracy of our Word2Vec model

CHAPTER 5: CONCLUSION

5.1 DISCUSSION

Image Segmentation is a subtopic of image processing where a lot of work is yet to be done. Image segmentation has proven to be highly beneficial in the fields of image recognition, image content analysis, and medical image segmentation. The application and workings of Detectron2 and Word2Vec made is easier to identify out of context objects in an image. All the files used for the project can be found in the GitHub repository [12]. One limitation we faced during the project was that the Detectron2 trained on the COCO17 could categorize objects within a list of 134 objects. This meant that objects within an image which is not same as those in the object list would be wrongly detected. The object list is as follows: ‘person’, ‘umbrella’, ‘boat’, 'traffic light', 'river', 'water', 'tree', 'pavement', 'building', 'sky', 'street', 'zebra', 'mountain', 'grass', 'banana', 'handbag', 'dirt', 'food', 'motorcycle', 'truck', 'suitcase', 'bus', 'parking meter', 'road', 'wallbrick', 'fire hydrant', 'car', 'toothbrush', 'book', 'cardboard', 'shelf', 'stairs', 'wallwood', 'floor', 'wall', 'giraffe', 'stop sign', 'railroad', 'train', 'gravel', 'cow', 'bird', 'elephant', 'bench', 'fence', 'surfboard', 'sea', 'potted plant', 'cat', 'curtain', 'window', 'table', 'paper', 'skis', 'snow', 'cup', 'apple', 'dining table', 'sports ball', 'baseball glove', 'baseball bat', 'playingfield', 'house', 'dog', 'bowl', 'couch', 'floorwood', 'ceiling', 'cabinet', 'rug', 'chair', 'tv', 'keyboard', 'mouse', 'cell phone', 'kite', 'sand', 'bed', 'remote', 'toilet', 'walltile', 'orange', 'fruit', 'horse', 'roof', 'refrigerator', 'microwave', 'pizza', 'skateboard', 'bottle', 'sheep', 'fork', 'wine glass', 'sandwich', 'vase', 'clock', 'frisbee', 'platform', 'broccoli', 'carrot', 'laptop', 'sink', 'knife', 'counter', 'blanket', 'bicycle', 'banner', 'tent', 'backpack', 'tennis racket', 'flower', 'rock', 'doorstuff', 'mirrorstuff', 'cake', 'teddy bear', 'airplane', 'oven', 'donut', 'bear', 'bridge', 'light', 'hot dog', 'net', 'towel', 'spoon', 'snowboard', 'windowblind', 'toaster', 'wallstone', 'scissors', 'tie', 'hair drier', 'pillow'.

5.2 SUMMARY

Image segmentation models proves to be highly valuable. These model uses extensive computation and demands GPU. Training a highly accurate image segmentation model on a detailed dataset such as COCO17 dataset proves to be very useful as it can identify more objects in an unseen image with higher accuracy. Word2Vec is an appropriate model that successfully utilized the contextual information of objects.

5.3 FUTURE WORKS

In this project, we identified the out of context object from images using Word2Vec, by utilizing the contextual object dataset created using Detectron2, trained on COCO17 dataset. In the future, we would try to increase the out of context images dataset. Increasing the dataset of common object lists will achieve more accuracy in Word2Vec training. Since we are limited to single out of context object detection, future work can develop a system that can detect multiple out of context object.

REFERENCE

[1] Chen, Changhao & Wang, Bing & Lu, Chris & Trigoni, Niki & Markham, Andrew. (2020). A Survey on Deep Learning for Localization and Mapping: Towards the Age of Spatial Machine Intelligence. Available at: https://doi.org/10.48550/arXiv.2006.12567

[2] Lin, Tsung-Yi & Maire, Michael & Belongie, Serge & Hays, James & Perona, Pietro & Ramanan, Deva & Dollár, Piotr & Zitnick. (2014). Microsoft COCO: Common Objects in Context. Available at: https://arxiv.org/abs/1405.0312

[3] GettyImages, [online]. Available at: https://www.gettyimages.com/photos/out-of-context

[4] Zhao Hengshuang, Shi Jianping, 'Pyramid Scene Parsing Network', 2017. [Online]. Available at: https://arxiv.org/pdf/1612.01105v2.pdf

[5] Zhou Bolei, Zhao Hang, 'Scene Parsing through ADE20K Dataset', 2017. [Online]. Available at: http://people.csail.mit.edu/bzhou/publication/scene-parse-camera-ready.pdf

[6] Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., & Girshick, R. (2019). Detectron2. https://github.com/facebookresearch/detectron2

[7] Mikolov, Tomas & Chen, Kai & Corrado, G.s & Dean, Jeffrey. (2013). Efficient Estimation of Word Representations in Vector Space. Proceedings of Workshop at ICLR. 2013. Available at: https://doi.org/10.48550/arXiv.1301.3781

[8] Choi MJ, 'Context Models and Out-of-context Objects', 2011. [Online] Available at: http://people.csail.mit.edu/myungjin/publications/outOfContext.pdf?fbclid=IwAR1Jb0-%20miKL9AD1zW3TJOumpSrkP5iVBA5nnJ9-q-Jz7ChSwHQ1GYYyHjTk

[9] J. Xiao, K. A. Ehinger, J. Hays, A. Torralba, and A. Oliva, 'SUN Database: Exploring a Large Collection of Scene Categories',2013. [Online]. Available at: http://3dvision.princeton.edu/projects/2010/SUN/paperIJCV.pdf

[10] Sitaula Chiranjibi, Aryal Sunil, Xiang Yong, 'Content and Context Features for Scene Image Representation', 2021. [Online]. Available at: https://arxiv.org/pdf/2006.03217.pdf

[11] Bardool Kevin, Tuytelaars Tinne, Oramas Joe, 'A Context Aware Deep Learning Architecture for Object Detection', 2019 [Online]. Available at http://ceur-ws.org/Vol-2491/abstract92.pdf

[12] Arnob, J.& Noor, F. (2022). Out of Context Image Detection. https://github.com/FahimaNoor/CSE499