Performance Comparison of CNN-based Semantic Segmentation on Indoor and Outdoor Scenes

ANTHONY IROKOSU, University of Twente, The Netherlands

Semantic segmentation is an important task in computer vision. It involves the assignment of class labels to regions in an image. The use of Convolutional Neural Networks has been successful for semantic segmentation; however, the performance of CNN-based semantic segmentation models can vary depending on the properties of the input data. Such properties are found in the indoor or outdoor environment.

This paper presents a comprehensive performance comparison of CNN-based semantic segmentation models on indoor and outdoor datasets. The goal is to highlight the environment's impact on the effectiveness of CNN-based segmentation models.

To facilitate the conduct of the research, a DeepLabv3Plus based on the ResNet-101 CNN architecture was evaluated on datasets specific to indoor and outdoor environments. The datasets were from the popular benchmark dataset, Pascal Visual Object Classes 2012.

The performance of the CNN-based semantic segmentation model is assessed using well-known evaluation metrics such as mean over intersection union (mIoU), frequency-weighted mIoU, pixel accuracy, and class accuracy.

The evaluation shows better performance of the CNN-based semantic segmentation in outdoor environments compared to indoor environments across all metrics.

Additional Key Words and Phrases: Semantic Segmentation, Convolutional Neural Networks (CNN), Computer Vision, Indoor environment, Outdoor environment.

1 INTRODUCTION

Computer vision is a significant field with applications in self-driving cars, teleoperated robotics and medical research. Semantic segmentation is one of the tasks involved in scene understanding. It is the assignment of categorical labels to pixels in an image [8]. In recent years CNNs have positively impacted the field of computer vision. CNN-based semantic segmentation has gained considerable recognition due to its improved ability to segment objects in a visual scene. However, the effectiveness of CNN-based semantic segmentation models varies due to the characteristics of the input data, specifically in indoor and outdoor environments.

Indoor and outdoor environments pose distinct challenges for semantic segmentation due to variations in properties such as scene complexity and lighting conditions. Both indoor and outdoor environments require semantic segmentation for effective scene understanding. Therefore, understanding how these environmental factors impact the performance of CNN-based semantic segmentation models holds significant importance for their development.

This research aims to investigate the impact of indoor and outdoor scenes on the performance of CNN-based semantic segmentation models. Additionally, the following research questions are proposed:

- How does the semantic segmentation model perform in an indoor environment
- How does the semantic segmentation model perform in an outdoor environment

The research is conducted by evaluating a state-of-the-art CNN architecture, specifically the DeepLabv3Plus based on ResNet-101, on a benchmark dataset containing annotated images of indoor and outdoor scenes. This dataset captures a range of diverse indoor and outdoor environments reflective of visual characteristics encountered in the real world.

Furthermore, the dataset is curated specifically from the Pascal Visual Object Classes 2012 benchmark datasets [13]. The Pascal VOC benchmark dataset is a well-known dataset, and it has been used for other image segmentation tasks [13]. The dataset is sufficient for assessing the performance of the CNN model on challenging scenes.

Additionally, the performance of the CNN-based semantic segmentation model is measured using well-known evaluation metrics: mean over intersection union (mIoU), frequency-weighted mIoU, pixel accuracy, and class accuracy. These metrics provide a quantitative measure of the accuracy and consistency of the model.

The research reveals that the performance of CNN-based semantic segmentation models on outdoor datasets was approximately 12% better than the performance on indoor datasets when comparing the mIoUs. And, across the other metrics, the performance on outdoor datasets was better compared to the performance on indoor datasets.

The main contribution of this work is the advancement of CNN-based semantic segmentation algorithms by highlighting the role of environmental factors in the performance of semantic segmentation.

2 RELATED WORKS

With the advancement in CNN architectures and the availability of large-scale datasets. Research has been conducted into the evaluation of the performance of different CNN architectures [1][3][4][8]. Garcia-Garcia et al's study into deep learning techniques applied to semantic segmentation evaluated popular CNN architectures on different datasets containing both indoor and outdoor datasets. This study provides a qualitative analysis of the performance of different architectures on compatible datasets. It showed promising results for DeepLabv3 on the PascalVOC2012 dataset. Although the study provided results on outdoor specific datasets, the architecture used differed from the architecture used for indoor specific datasets. Additionally, yeo et al proposed a scene classification scheme which measured a 92% accuracy on Microsoft COCO dataset, which contains both indoor and outdoor scenes.

Indoor scenes present a set of unique challenges for semantic segmentation due to the complex layout of objects in scene and varying lighting conditions. Researchers have conducted surveys into the performance of CNN architectures on indoor specific datasets [1][5]. In a survey done by Nasser et al, different architectures were

TScIT 39, July 7, 2023, Enschede, The Netherlands

evaluated on the NYUv2 dataset with varying numbers of classes. The models tested resulted in an mIoU in the range of 28% to 43%. The results seem to agree with results from Garcia-Garcia et al's evaluation of CNN model on NYUv2 dataset which resulted in an mIoU of 49%.

Outdoor scenes present a distinct set of challenges when it comes to image segmentation. Research into the performance of models on outdoor scenes seems promising. Garcia-Garcia et al's survey on deep learning technique, resulted in mIou scores in the range 78% to 80% on the Stanford background dataset [1]. On the cityscapes dataset DeepLab measured a 70% mIoU accuracy [1]. This indicates that current CNN based models perform on outdoor scenes than on indoor scenes.

Surveys have been carried out on dataset used for semantic segmentation. Garcia-Garcia et al's paper gives a concise summary of the popularly used datasets for indoor and outdoor environments such as Pascal VOC, NYUv2, cityscapes and Microsoft COCO datasets. These datasets contain a wide range of object classes, allowing for comprehensive evaluation of segmentation algorithms. Researchers have used these datasets to train and evaluate CNN-based models. Yeo et al made use of the Microsoft COCO datasets to conduct their analysis.

Metrics are necessary to quantify the performance of the model. A survey by Nasser et al provides a summary of standard metrics used extensively by other researchers, such as Garcia-Garcia et al and Yeo et al. In their survey Nasser et al, highlights mean intersection over union, frequency weighted intersection over union and pixel accuracy as standard metrics for quantifying performance of semantic segmentation.

The related works in CNN-based semantic segmentation show the significant research and progress made in the field. However, a comprehensive performance comparison focusing on indoor and outdoor scenes is essential to improving semantic segmentation on indoor and outdoor scenes.

3 METHODS

3.1 Datasets

A total of 2913 images were curated from The Pascal VOC2012 benchmark dataset [13] specifically for semantic segmentation analysis. The dataset contained 20 unique object classes. The dataset contained indoor and outdoor scenes. Additionally, the datasets had annotation files and segmentation masks for the object classes present in respective images. With annotation files and object classes, the dataset was divided into sets of indoor and outdoor scenes. The division was approximated using the object classes. The new image sets are further divided into train, validation, test split, on a 70-15-15 split.

3.2 Framework

The Vedaseg framework, which is available on GitHub [7], was used to perform semantic segmentation using CNN. The framework supports multiple architectures such as Deeplab,U-net and, PSPNet, nevertheless DeepLabv3plus was used as the architecture due to its performance on Pascal VOC in Garcia-Garcia et al's survey [1] on CNN-based semantic segmentation techniques and its performance in the framework benchmark [7]. Additionally, the framework used Resnet-101 as the backbone architecture. The framework was configured for 21 classes, the number classes in the Pascal VOC dataset including background. The framework was configured with evaluation metrics: mean over intersection union (mIoU), frequency-weighted mIoU, pixel accuracy, and class accuracy. A cross entropy loss function was used for the loss function. After testing different epoch values, the framework was configured on 100 epochs as it resulted in the lowest loss value with no notable change in the loss value between iterations.

3.3 Metrics

To evaluate the model's performance, the following standard metrics are used:

Mean Intersection over union(mIoU): It is the sum mean of the ratio of predicted segmentation to the union of the predicted segmentation and the ground truth. A higher value indicates a stronger performance.

Frequency Weighted IoU: Like IoU, frequency weighted IoU also takes into consideration the frequency of occurrence of an object class. It assigns weights to each class, with frequently appearing classes giving higher weight values. A higher value indicates a stronger performance.

Pixel Accuracy: It is the ratio of correctly classified pixels in the predicted segmentation mask to the total number of pixels in the ground truth mask. A higher value indicates a stronger performance.

Class Accuracy: It measures the accuracy of the model in correctly predicting the class label. A higher value indicates a stronger performance.

3.4 Procedure

The dataset is divided into indoor and outdoor. The division is approximate using the object classes, by placing obvious classes such as trains and airplanes in outdoor sets and indoor objects such as monitors and sofas in indoor environments. The indoor and outdoor sets are split into train, validation, and test on a 70-15-15 split. The model is then trained on the indoor and outdoor datasets. Upon completion the framework produces a prediction model. The prediction model is used to test the performance on the test split. The metrics are saved in a log file, Additionally the prediction and labeled masks are saved to have visual comparison of the output.

4 RESULTS

The results show that the model achieved good segmentation accuracy, with the model measuring an mIoU of 59.24% on indoor scenes and 71.18% on outdoor scenes. An mIoU of > 50% is considered good segmentation. The performance of the model can be seen in Table 1 and Table 2.

Table 1. Results on Indoor dataset

Metric	Value (in percent)
mIoU	59.24
FwIoU	81.45
Pixel Accuracy	88.9
Class Accuracy	71.92

Table 2. Results on Outdoor dataset

Metric	Value (in percent)
mIoU	71.18
FwIoU	91.82
Pixel Accuracy	95.62
Class Accuracy	78.05

4.1 Findings

The datasets were further analyzed on the average number of objects per image and the average number of pixels per image. In both categories the value was greater in Indoor scenes. See Table 3 and Table 4.

Table 3. Results on average number of objects

Scene	Avg num of objects/image
Indoor	**2.49**
Outdoor	2.28

Table 4. Results on average density

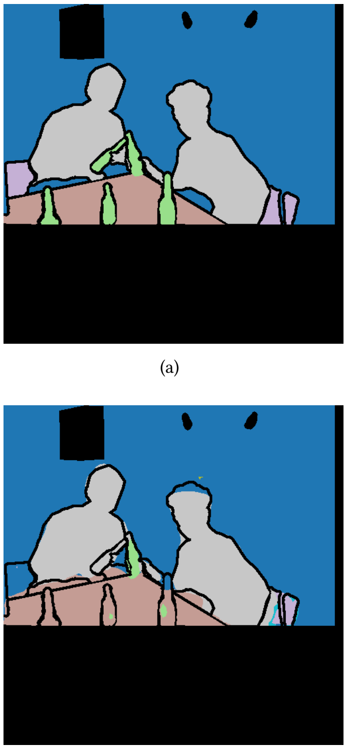
Scene	Avg num of pixels/image
Indoor	**53,546.66**
Outdoor	41,525.43

5 DISCUSSION

Looking at the results, the model performed better across all metrics on outdoor scenes than on indoor scenes, indicating a significant impact of environment on the performance of CNN-based sematic segmentation models. This aligns with my initial hypotheses. Outdoor scenes tend to be better lit and contain less clustered objects, and I would expect this characteristic to impact the performance of the model. It also falls in line with the pattern of results gotten from evaluations done by previous researchers.

Analyzing the complexity evaluation shows that the average number of objects per image in the indoor dataset was 2.49, 9% higher than in the outdoor dataset which averaged 2.28 objects per image. This shows that the outdoor scenes were less complex than indoor scenes.

This theoretical means that the model would perform better on indoor scenes with less complexity. This can be seen by taking an example output. The predicted mask (Figure 1) contains 4 unique classes. The predicted segmentation is good, but it is clear the model struggled to get it right as it labeled the bottles on the table as a table. This could be due to the lighting of the environment.

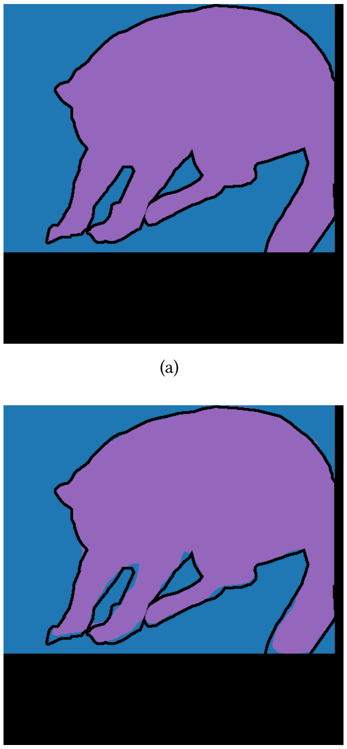


(a)

(b)

Fig. 1. (a): Labeled. (b): Predicted.

Comparatively, with less objects in the scene the model performs better as seen in figure 2. In the scene there is a single unique class, and the model performs better on this image file.



(a)

(b)

Fig. 2. (a): Labeled. (b): Predicted.

Because the average complexity differed by approximately 9% it is hard to say with certainty if the number of objects in the scene is the only factor playing a part in impacting the performance of the model, but it does warrant further research.

The example images selected highlight the observation. In further research, it would be beneficial to investigate the impact of scene complexity on the performance of CNN-based semantic segmentation models

Another factor noticed was the average density. Like the number of objects.it was the case that the indoor scenes contained more pixels than outdoor scenes. Indoor scenes averaged approximately 53,546.66 pixels while outdoor scenes averaged 41,525.43 pixels per file. The impact this has on the result is unclear, intuitively I would expect better performance with more pixels. But with outdoor scenes performing better than indoor scenes. It is also possible that the number of pixels correlate with the number of objects present in the scene, but this is uncertain and an isolated investigation of the impact of pixel density, if any, would be important.

The procedure of curating the dataset was not perfect. Because Pascal VOC 2012 is not separated into indoor and outdoor scenes, it had to be separated into these categories by approximating using the object classes. This could lead to having images that are indoors in the outdoor datasets and vice versa. This makes the system biased.

5.1 Limitaions

During the research period, problems were encountered. Specifically with adapting the framework to fit the needs of the research, coupled with the limited time and limited hardware, it made it difficult to expand into testing different architectures and using larger datasets. Initially, the Stanford background dataset was going to be used for outdoor scenarios, but due to challenges encountered while trying to adapt it for the model, it had to be dropped, instead opting for a generic dataset, and manually performing a separation into indoor and outdoor scenes. There was a lot of trial and error when it came to configuring the framework, the documentation was lacking for adapting the model for other datasets outside of the already supported ones which all happen to be indoor exclusive datasets. Using much larger datasets with a larger number of classes was not feasible as it caused the computer to run out of memory. Whilst the evaluation shows that outdoor scenes performed better than indoor scenes, the result should be taken with caution. The dataset chosen could be biased towards outdoor scenes. This can happen if the outdoor scenes feature better lighting and less clustered objects compared to indoor scenes. It is possible that better lit up indoor scenes with less clustering would perform just as well or better than outdoor scenes. It is also possible that outdoor scenes with poor lighting, for example nighttime images, would perform far worse than indoor scenes.

6 FUTURE WORK

The next step to be taken to further the research would be to investigate the impact of complexity and pixel density on the performance of CNN-based segmentation models. Understanding the relationship between these factors could be important and crucial for developing better CNN-based segmentation models for both indoor and outdoor environments. Furthermore, further research should also include testing on multiple different architectures and should include 2.5D/3D datasets.

7 CONCLUSION

Based on the results, the model performed well on both indoor and outdoor scenes.

In indoor scenes, the model demonstrated moderately satisfactory performance in segmentation task. With an mIoU score of 59.24%, the model effectively classifies and segments objects with low complexity but struggles with images with more complex scenes. The challenges of the model

In outdoor scenes, the performance of the model was better than the performance on indoor scenes. With an mIoU score of 71.18%, the model better classifies and segments objects. The performance increase correlates with the reduced complexity

In general the performance disparity between indoor and outdoor environments can be explained by the differences in environmental complexity. The indoor environment contained more objects

This work showed the impact of indoor and outdoor environments on the performance of CNN-based semantic segmentation models. This work contributed to the advancement of the development of CNN-based semantic segmentation models. This work answers the research question posed, and it covered related works done in the same field. It provided a comparative summary of the performance of the DeepLabv3 CNN (Convolutional Neural Networks) architecture on indoor and outdoor datasets curated from the Pascal VOC 2012 datasets. It covered interesting findings about the relationship between performance, complexity, and pixel density. In conclusion, CNN-based semantic segmentation worked well on both indoor and outdoor environments but performed better on outdoor environments.

8 REFERENCES

[1] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A Review on Deep Learning Techniques Applied to Semantic Segmentation." arXiv, Apr. 22, 2017. Accessed: May 04, 2023. [Online]. Available: http://arxiv.org/abs/1704.06857

[2] R. Mo, "A Survey of Image Classification Algorithms based on Convolution Neural Network," Highlights in Science, Engineering and Technology, vol. 15, pp. 191–198, Nov. 2022, doi: 10.54097/hset.v15i.2222.

[3] F. Cao and Q. Bao, "A Survey On Image Semantic Segmentation Methods With Convolutional Neural Network," in 2020 International Conference on Communications, Information System and Computer Engineering (CISCE), Jul. 2020, pp. 458–462. doi: 10.1109/CISCE50729.2020.00103.

[4] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 7, pp. 3523–3542, Jul. 2022, doi: 10.1109/TPAMI.2021.3059968.

[5] M. Naseer, S. Khan, and F. Porikli, "Indoor Scene Understanding in 2.5/3D for Autonomous Agents: A Survey," IEEE Access, vol. 7, pp. 1859–1887, 2019, doi: 10.1109/ACCESS.2018.2886133.

[6] "Papers with Code - Semantic Segmentation." https://paperswithcode.com/task/semantic-segmentation (accessed May 04, 2023).

[7] Media-Smart. (2023). Media-Smart/vedaseg [Python]. https://github.com/Media-Smart/vedaseg (Original work published 2019)

[8] M. Gao, J. Jiang, G. Zou, V. John, and Z. Liu, "RGB-D-Based Object Recognition Using Multimodal Convolutional Neural Networks: A Survey," IEEE Access, vol. 7, pp. 43110–43136, 2019, doi: 10.1109/ACCESS.2019.2907071.

[9] W.-H. Yeo, Y.-J. Heo, Y.-J. Choi, S.-J. Park, and B.-G. Kim, "Scene Classification Algorithm Based on Semantic Segmented Objects," in 2021 IEEE International Conference on Consumer Electronics (ICCE), Jan. 2021, pp. 1–4. doi: 10.1109/ICCE50685.2021.9427672.

[10] M. A. A. Tahir, X. Feng, and Z. Shaker, "SIS-CNN: Semantic Image Segmentation Using Convolutional Neural Networks," International Journal of Advanced Network, Monitoring and Controls, vol. 6, no. 3, pp. 9–17, Jan. 2021, doi: 10.21307/ijanmc-2021-022.

[11] T. A. Patel, V. K. Dabhi, and H. B. Prajapati, "Survey on Scene Classification techniques," in 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Mar. 2020, pp. 452–458. doi: 10.1109/ICACCS48705.2020.9074460.

[12] P. Wang et al., "Understanding Convolution for Semantic Segmentation," in 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Mar. 2018, pp. 1451–1460. doi: 10.1109/WACV.2018.00163.

[13] The PASCAL Visual Object Classes Challenge 2012 (VOC2012). (n.d.). Retrieved June 19, 2023, from http://host.robots.ox.ac.uk/pascal/VOC/voc2012/index.html