Network and Communication Technologies; Vol. 8, No. 1; 2023
ISSN 1927-064X E-ISSN 1927-0658
Published by Canadian Center of Science and Education

Analyzing Out-of-Domain Generalization Performance of Pre-Trained Segmentation Models

Johnson Zhong[1]

[1] Yew Chung International School Secondary Campus, Hong Kong

Correspondence: Johnson Zhong, Yew Chung International School Secondary Campus, Hong Kong. E-mail: johnson.wx.zh@gmail.com

Received: October 16, 2022 Accepted: February 3, 2023 Online Published: February 16, 2023

doi:10.5539/nct.v8n1p1 URL: https://doi.org/10.5539/nct.v8n1p1

Abstract

Artists illustrate objects to various degrees of complexity. As the amount of detail or the similarity to reality of a depiction decreases, the object tends to be reduced to its simplest, most relevant higher-level features (Harrison, 1981). One of the reasons Deep Neural Networks (DNN) may fail to identify objects in an image is that models are unable to recognize the order of importance of features such as shape, depth, or color within an image, which means even the most minute distortions of pixels within an image that would be imperceptible to humans would greatly impact the performance of the object detection models (Eykholt et al., 2018). However, by training DNN on artworks where the most prominent features defining specific objects are emphasized, perhaps a model can be made to be more resilient against small-scale changes in an image. In this paper, the correlation between the level of similarity to reality of images and artworks of an object and the accuracy of object detection models is investigated to test the ability of object detection models in identifying the most salient features of a particular object. The results of this report can help outline the efficacy of models only trained on real images in identifying increasingly abstract artworks that have simplified an object to its most prominent features. The experiment shows that the accuracies of models decrease as the images or illustrations provided become more abstract or simplified, which suggests the higher level features that identify a particular object are different in object detection models and humans.

Keywords: Neural Networks, Supervised Learning, Semantic Segmentation, Domain Generalization

1. Introduction

Machine Learning (ML) is a field of research that revolves around the goal of creating a computer system capable of self-improvement, thereby eliminating the need for large amounts of programming. Learning methods for ML models are generally divided into the categories of supervised, unsupervised, and semi-supervised algorithms depending on the amount of human intervention needed in each model. Supervised ML methods, such as neural networks or linear regression, use a set of training data that has already been labeled by humans to increase the accuracy of the outputs of a model. Tasks completed by ML trained through supervised methods can be further divided into classification, broadly defined as the task of separating different inputs into various discrete categories, and regression, which is the task of finding the relationship between independent variables within an input to predict a continuous dependent variable. For example, tasks defined as classification include image classification, character recognition, and medical diagnosis while regression may include, weather forecasting, stock predictions, and identifying the correlation between calories consumed and obesity levels.

In this paper, we focus on a specific classification task known as semantic segmentation where an ML model is trained to identify where each pixel within an image is classified under a particular semantic label, after which we provide an analysis of which pre-trained models have the best domain generalization. Semantic Segmentation provides high-level semantic information regarding the position of various objects in a scene relating to each other without the need for other human input, which can allow intelligent machines to achieve more complex tasks requiring the recognition of specific objects, people, or even patterns which exist as a part of multiple images. Current applications of semantic segmentation include autonomous driving (Kaymak & Uçar, 2019), medical image analysis (Asgari Taghanaki, Abhishek, Cohen, Cohen-Adad, & Hamarneh, 2021), video surveillance systems, and even forgery identification in artworks.

However, the fragility of many object detection models may be rectified through the use of adversarial examples

created through cropping (Yoshida & Okuda, 2022), freehand sketches (Kim, Nanni, & Süsstrunk, 2022), or other alterations to an image. In this paper, it is proposed that images or illustrations of greater abstraction can be used to train models in identifying specific objects because such images or illustrations are likely to have simplified the objects to their more salient features, which means using these illustrations as training data may result in more accurate and resilient models. To test the validity of this hypothesis, the accuracies of different object detection models in identifying illustrations of differing similarity to reality are recorded; these results will indicate whether these DNNs analyze the saliency of the features of objects in an image in a similar or entirely disparate manner to the human artists.

2. Related Works

Many supervised, unsupervised, and semi-supervised methods have been used to achieve Semantic Segmentation through the use of DNN. The task of Semantic Segmentation may be split into the tasks of classifying, localizing, and segmentation, which may be accomplished through context-based methods, Recurrent Neural Network (RNN)-based methods, Feature-enhancement-based methods, Deconvolution-based methods, etc.

2.1 Semantic Segmentation

Lucchi et al. (2011) showed that the accuracy of Semantic Segmentation algorithms may be improved by incorporating local and global contexts of the image in the classification process. Information about the various contexts within an image provides the model with valuable information about the pixels' relationship to each other in an image, this allows for more accurate classification of objects within an image. Yu and Koltun (2015) introduced DilatedNet which aggregates multi-scale contextual information using dilated convolutions to do so without losing resolution or coverage. Another work (Liu, Rabinovich, & Berg, 2015) proposed using global pooling to find a summary of an entire image, providing the network with a higher awareness of the global context in an image and boosting accuracy.

Deeper Layers in a CNN are more "semantic-aware", analyzing higher-level features, but less aware of finer details due to pooling or a larger stride being used, while the opposite is true for shallower layers with smaller details. Hence the combined use of higher-level features and lower-level features in the final prediction is likely to boost the performance of the model. Long, Shelhamer, and Darrell (2015) shows that Fully Convolutional Networks (FCN/Completely Convolutional Neural Networks) apply the skip-connection strategy.

Noh, Hong, and Han (2015) shows that CNN combined with a deconvolution network to allow for analysis of more general shapes and higher-level features by starting with the CNN pooling and convolution operations and ending with unpooling and deconvolution operations. This allows networks that use Deconvolution-based methods to be capable of identifying finer details in the entire image, thereby improving the performance of the model.

Although the Recurrent Neural Network (RNN) is generally used in processing sequential signals such as text-to-speech, there are also RNN-based semantic segmentation methods. Pinheiro and Collobert (2014) proposed the Recurrent Convolutional Neural Network (RCNN), where the output of each CNN segment is inputted into the following segment sequentially. The horizontal and vertical coverage of the RNN algorithm provides relevant global information, allowing a more accurate overall performance of the model.

3. Methodology

In this experiment, images of five classes – Apples, Books, Elephants, Laptops, and Planes – were collected. In each class, the images were further separated into levels according to their likeness to real-life images – Real Images, Realistic Illustrations, Cartoon Images, and Abstract Illustrations. Each image was passed through Faster R-CNN and RetinaNet Object Detection models, both of which were trained on real-life images, and the accuracy of each output was analyzed. Intersection over Union (IoU) was used as the metric of performance to measure the accuracy of the object detection model in identifying a particular object and was defined as the following equation.

$$IoU = TP / (TP + FP + FN) \quad (1)$$

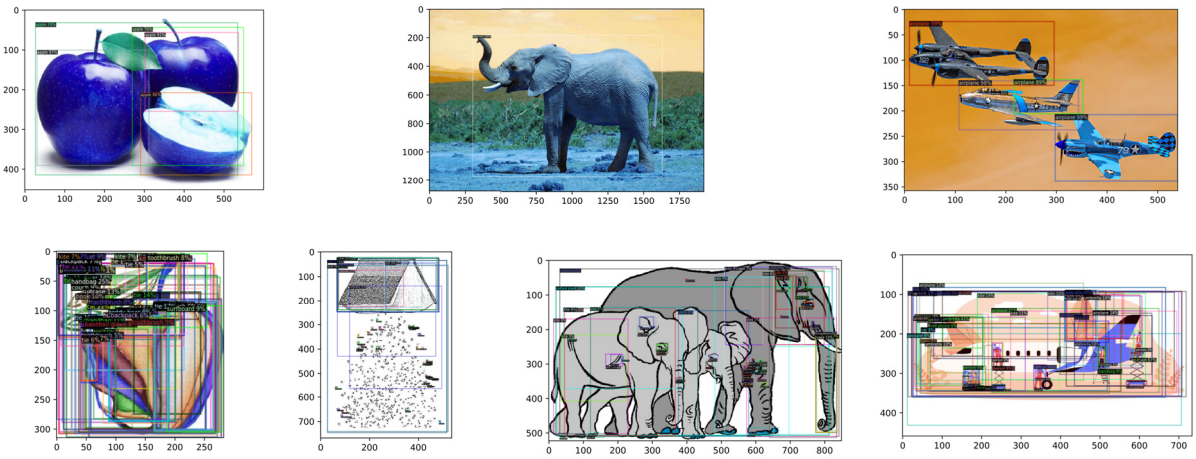

Figure 1. A Comparison of Real Images or Realistic Illustrations (Top Row) to Cartoon Images or Abstract Illustrations (Bottom Row)

True Positive (TP) is the number of intersecting pixels between the Ground Truth and bounding box predicted by the model; False Positive (FP) is the number of intersecting pixels between the Non-Ground Truth and bounding box predicted by the model; False Negative (FN) is the number of intersecting pixels between the Ground Truth and the parts of the image that are not a section of the bounding box.

4. Results and Discussion

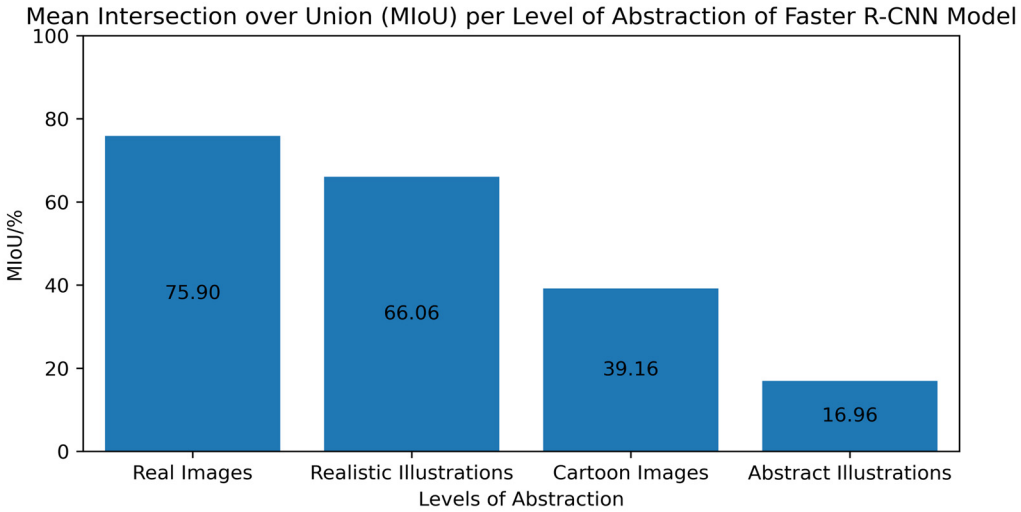

Figure 2. Comparison of the accuracy (using MIoU as a measure) of the Faster R-CNN Model in classifying images of different levels of abstraction

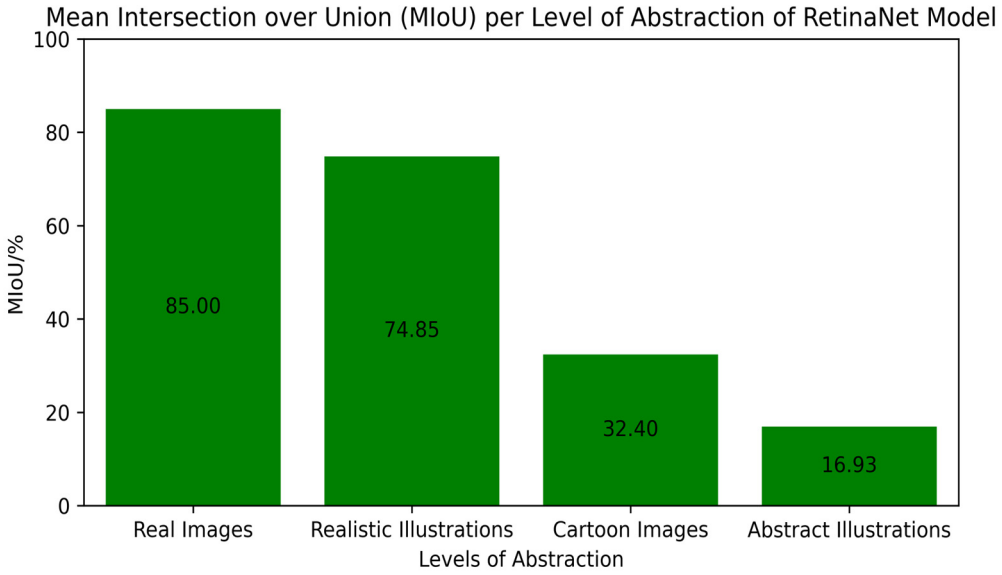

Figure 3. Comparison of the accuracy (using MIoU as a measure) of the RetinaNet Model in classifying images of different levels of abstraction

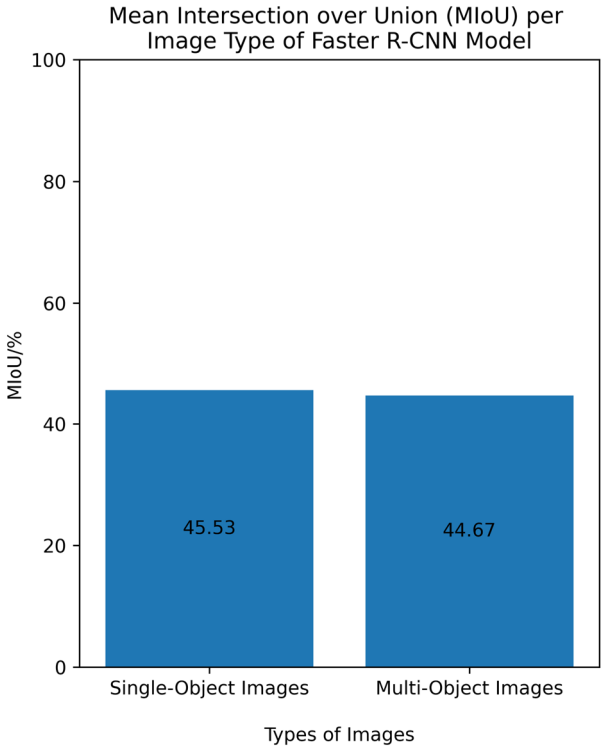

Figure 4. Comparison of the accuracy (using MIoU as a measure) of Faster R-CNN in classifying images of with a single or multiple objects

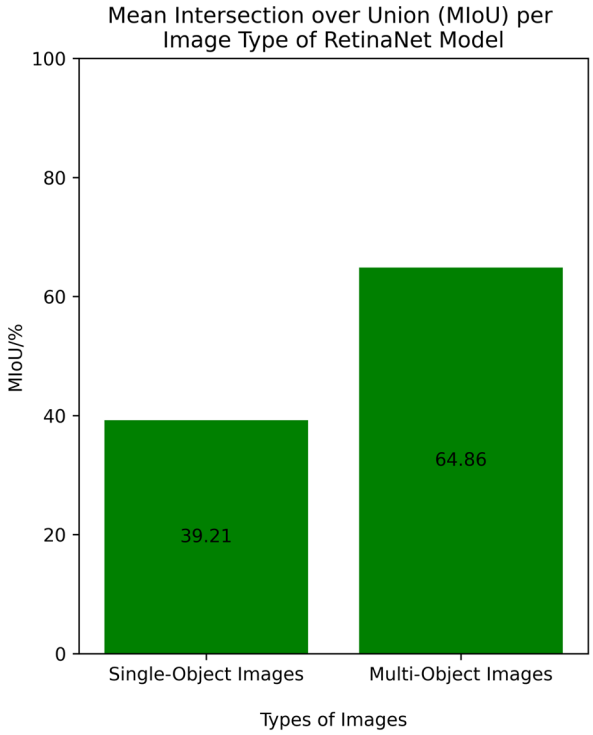

Figure 5. Comparison of the accuracy (using MIoU as a measure) of RetinaNet in classifying images of with a single or multiple objects

There exists a trend in both models where the average MIoU is highest in images closest to real-life images and decreases as images become more abstract. Perhaps due to the RetinaNet being a one-stage detection model which samples a much larger amount of possible object locations and the probability scores of the model predictions being ignored, RetinaNet received an 8-10% higher average accuracy compared to Faster R-CNN for both real images and realistic illustrations. The much higher sampling rate of the RetinaNet may have also caused the lower percentage accuracy of the RetinaNet model for both the cartoon images and abstract illustrations, possibly due to the RetinaNet model increasing the number of bounding boxes significantly as the model becomes more uncertain. It is ambiguous as to whether this could also be due to the greater difference in abstractness between cartoon images or the difference in the artists' aim in creating the artwork (as realistic illustrations are generally made to emulate the abstractness of real images while cartoons display more symbolic depictions).

Another possibility would be that the RetinaNet model is slightly overfitted with its original training data. The higher accuracy in identifying real images, the original training data for the models, has compromised the accuracy of the RetinaNet model in classifying more abstract or symbolic representations of the same objects.

The much greater difference in the MIoU of realistic illustrations and cartoon images than the other consecutive levels of abstraction in both models clearly indicates that there is a fundamental difference in the two types of art which causes the models' accuracy to drop significantly. Many possible factors may be a part of the difference in accuracy: artistic intention, whether the illustration was drawn in two or three dimensions, or even the amount of detail required to make a specific object recognizable by humans. Regardless of this, the amount of personal interpretation concerning the level of abstraction of each artwork makes it difficult to quantify a specific metric for the abstraction of each work of art.

While the Faster R-CNN Model has similar accuracy in classifying images with one object of its class and images with multiple objects of their class, the RetinaNet Model appears to have a significant disparity between its accuracy in classifying single-object images and multi-object images of 25.65%. This large discrepancy may be attributed to the greater sampling rate and therefore larger number of pixels within the predicted bounding boxes of the RetinaNet model, which would have performed more poorly for single-object images whose ground truths would naturally take up a smaller number of pixels while performing with much higher accuracy in multi-object images due to the larger number of possible object locations considered by the RetinaNet Model that allows it to

identify more of the multiple objects within an image than the Faster R-CNN Model.

5. Conclusion

In this paper, the accuracy of object detection models in detecting objects in various artworks was tested and shown to have a negative correlation with the level of abstraction of an object. This demonstrates how object detection models may identify and prioritize different features of salience in comparison to the features chosen by human artists when depicting a particular object to a specific degree of simplification, which further suggests that it may be helpful for object detection models to be trained with artworks of objects detailing the most relevant features of an object. Although the subjective nature of human perception may hinder the performance of a model tested on such images and illustrations, the process of training the model should allow it to find objective common features across a high enough number of artworks to allow it to identify objects whether abstract or realistic to a high degree of accuracy.

The experiment could also have been improved further by using a larger sample size of images, a more precise method of measuring the number of pixels in each bounding box than simply counting by hand, and perhaps even the creation of a model which uses training data that consists of both real images and abstract illustrations to test my hypothesis. However, the two-month period of this project and the limited resources meant that I was not able to realistically make all of these improvements to my experiment.

References

Asgari Taghanaki, S., Abhishek, K., Cohen, J. P., Cohen-Adad, J., & Hamarneh, G. (2021). Deep semantic segmentation of natural and medical images: a review. *Artificial Intelligence Review, 54*(1), 137-178. https://doi.org/10.1007/s10462-020-09854-1

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., … & Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1625-1634). https://doi.org/10.1109/CVPR.2018.00175

Harrison, R. (1981). *How Cartoons Work: The Cartoon Code.* [online] Medialit.org. Retrieved from https://www.medialit.org/reading-room/how-cartoons-work-cartoon-code

Kaymak, Ç., & Uçar, A. (2019). A brief survey and an application of semantic image segmentation for autonomous driving. *Handbook of Deep Learning Applications* (pp. 161-200). https://doi.org/10.1007/978-3-030-11479-4_9

Kim, H. G., Nanni, D., & Süsstrunk, S. (2022). Natural-Looking Adversarial Examples from Freehand Sketches. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP) (pp. 3723-3727). IEEE. https://doi.org/10.1109/ICASSP43922.2022.9747480

Liu, W., Rabinovich, A., & Berg, A. C. (2015). *Parsenet: Looking wider to see better*. arXiv preprint arXiv:1506.04579.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431-3440). https://doi.org/10.1109/CVPR.2015.7298965

Lucchi, A., Li, Y., Boix, X., Smith, K., & Fua, P. (2011). Are spatial and global constraints really necessary for segmentation?. In *2011 international conference on computer vision* (pp. 9-16). IEEE. https://doi.org/10.1109/ICCV.2011.6126219

Noh, H., Hong, S., & Han, B. (2015). Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision* (pp. 1520-1528). https://doi.org/10.1109/ICCV.2015.178

Pinheiro, P., & Collobert, R. (2014). January. Recurrent convolutional neural networks for scene labeling. In *International conference on machine learning* (pp. 82-90). PMLR.

Yoshida, M., & Okuda, M. (2022). Adversarial Examples for Image Cropping in Social Media. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP) (pp. 4898-4902). IEEE. https://doi.org/10.1109/ICASSP43922.2022.9746949

Yu, F., & Koltun, V. (2015). *Multi-scale context aggregation by dilated convolutions*. arXiv preprint arXiv:1511.07122.