Detection of Recyclable Objects using MaskRCNN

Brennan Proudfoot
bwproud at cs.unc.edu

Jay Rao
jayrao at cs.unc.edu

November 2018

1 Introduction

Studies show that the amount of recyclable materials in the U.S. waste system could generate over $7 billion if they were properly recycled[1]. This poses the incredibly relevant challenge of accurately identifying recyclable material to help prevent the introduction of recyclable materials into the general waste system. We trained a Mask RCNN model with a ResNet-101 backbone to serve as an object detection system that detects, bounds, and segments glass bottles, plastic bottles, and metal cans. While this system could be used solely at a user-level, it is also conceivable that this could be deployed at an enterprise level in a recycling plant for fast sorting and detection of incoming objects.

2 Problem

For the aforementioned reasons, we have decided to develop a model that can detect and classify recyclable items. The classes of recyclables that we currently support are plastic bottles, glass bottles, and metal soda cans. The reasoning behind this was that these are three different materials that should ideally be separated as each has a different recycling strategy. For example, when plastic is brought to a recycling center it is shredded [2] while metal is melted [3]. In addition, the three classes we chose are similar in the service they provide, so it's likely that people will recycle all three of these items together. Ultimately, our computer vision model looks to cut down on human labor and increase the amount of material that gets properly recycled.

3 Data

3.1 Dataset

As training a model from scratch for this task would have taken weeks, we decided to use transfer learning to speed up the development process. To that end, we got the weights of an existing Mask RCNN model pre-trained on COCO and trained that model on our three classes. As there is no existing dataset of segmented plastic bottles, glass bottles, and metal soda cans with labels, we had to create our own dataset, using a mix of photos we took ourselves and photos we found online. We attempted to find photos that had interesting visual qualities with respect to the object that we were trying to detect. These characteristics included occlusions (other objects partially obscuring the object we are trying to detect), out of focus images, crushed bottle/can images, to name a few.

[1] https://www.byui.edu/university-operations/facilities-management/recycling-and-sustainability/recycling-statistics

[2] https://www.norcalcompactors.net/processes-stages-benefits-plastic-recycling/

[3] https://www.thebalancesmb.com/an-introduction-to-metal-recycling-4057469

Figure 1: Sample image that was segmented using VGG Image Annotator that is managed by the University of Oxford.

3.2 Preparation and Data Augmentation

Once we found all of the photos we wanted to use, we segmented them, creating a mask around the area of interest, so the Mask RCNN model could learn how to generate masks. We used an open source tool for segmenting images, in this case four metal cans as seen in figure 1, to provide training and validation data for our model. In the end, we had 100 training images for each class as well as 50 validation images and 50 test images to evaluate the trained model.

We also augmented our data slightly to prevent overfitting, randomly flipping images 50% of the time, as well as adding Gaussian noise. In addition to making our model more robust, these augmentations helped artificially increase the size of our image set, which was especially important in our case, as we didn't have a very large training set compared to most computer vision training sets.

4 Model

4.1 Mask RCNN

We chose to use a Mask RCNN model to accomplish our task. Mask RCNN is an extension of Faster RCNN which identifies Regions of Interest and determines bounding boxes and classifications for these candidate objects. Mask RCNN extends this output to also include a mask for each region of interest [2]. This mask allows us to perform accurate object segmentation in addition to the bounding and classification.

In more detail, Mask RCNN model is composed four parts: a feature extractor, the region proposal network, the classifier and bounding box regressor, and the segmentation masks. The original RCNN [1] that this model is based on has the feature extractor as a hyper parameter to the model, and Mask RCNN is no exception. In Mask RCNN's case, the feature extractor is Resnet-101. This current state-of-the-art feature extractor leverages learning residual functions which references layer inputs [3]. The feature extractor in Mask RCNN also includes a feature pyramid network (FPN) which helps detect objects at different scales. The FPN has a bottom-up pathway computes different feature maps with a scaling step of 2, and the top-down pathway creates higher resolution features by upsampling coarser

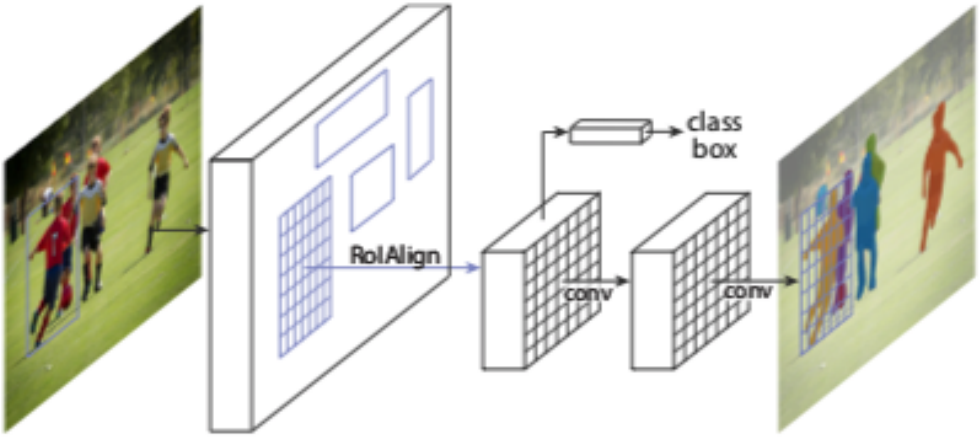


Figure 2: Image from the original MaskRCNN research paper released by Facebook Research. The image shows how regions of interest are both bounded in a box and masked[2].

feature maps. The two pathways are connected laterally to merge feature maps at the same scale created by both [6].

Once features are extracted, then the region proposal network (RPN), introduced in Faster RCNN [7], scans the anchor boxes in the image (more precisely the feature map of the image) and proposes either a foreground object or background for each anchor box. If a foreground object is proposed, then the RPN also creates a bounding box for the object. A reduced number of likely proposals is achieved by only choosing the foreground object proposals with highest likelihood and using non-max suppression to reduce proposals near each other.

Mask RCNN primarily expands on Faster RCNN after the RPN. Mask RCNN still predicts the class and performs bounding box regression, but it also produces a binary segmentation mask in parallel. The class predictor can choose from the number of classes predefined along with a background class that will automatically discard the region proposal if selected. The bounding box refinement allows for slight adjustment in shift and scale for the bounding box to more precisely fit the object. An additional adjustment from Faster RCNN is the RoIAlign step which aims to reduce the misalignments between regions of interest and features that could affect the prediction of pixel accurate masks.

From the Mask RCNN paper: the segmentation encodes the objects spatial layout [2]. This differs from the classification and bounding box because the mask has corresponding pixel values without reducing it to vector representations. While the masks are usually small in size, the ground truth masks during train are scaled down while the mask predictions are scaled up to the bounding box. These considerations keep the model from being as computationally intense. The loss function is a sum of the class loss, bounding box loss and mask loss.

4.2 Transfer Learning

The basic idea of Transfer Learning is to take some pre-trained model and provide it different data to accomplish a task other than the one it was originally trained for. This allows us to take advantage of the training that has already happened and drastically reduces the amount of time we would have had to spend training a model from scratch [8]. In our case we used a Mask-RCNN model pre-trained on the COCO dataset as a base for our model.

We attempted two different types of transfer learning for this model - fine tuning and fixed feature. In fine tuning, we trained all layers in the existing model with the new training data, while in fixed feature, we froze all layers in the model besides the heads of the networks.

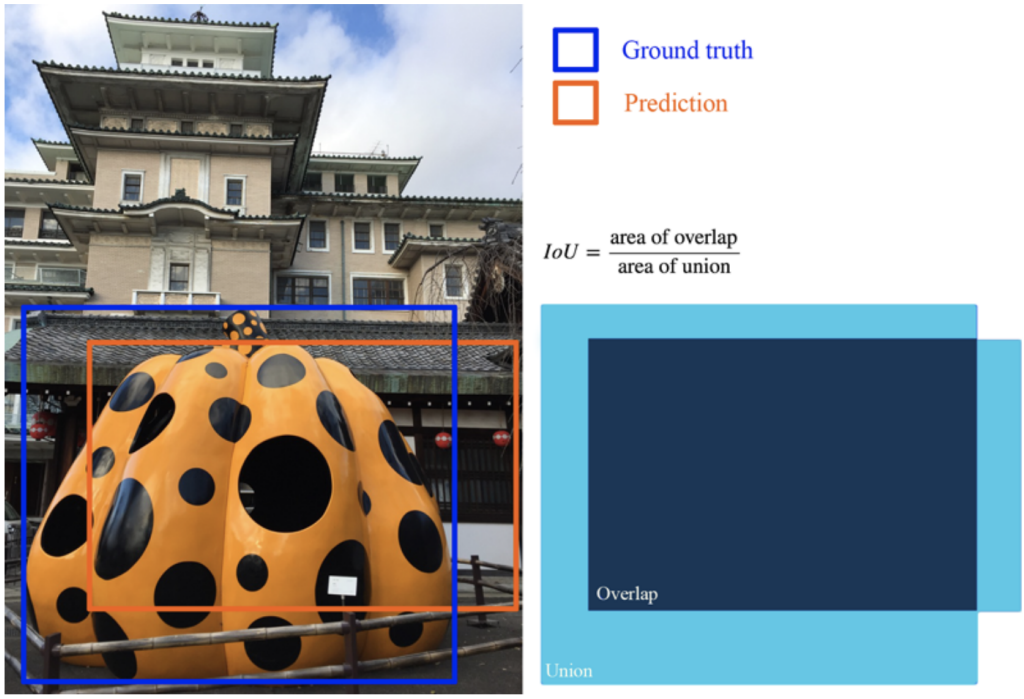


Figure 3: Intersection over Union from Jonathan Hui's explanation of mean average precision [4]

The latter is called fixed feature because the pre-trained model effectively acts as a feature extractor as every layer but the last is frozen, so the pretrained model finds the features, and the trained model then uses those features to classify the objects into our task-specific classes. We chose these two different styles of transfer learning in order to compare and contrast their relative performance.

4.3 Models Trained

We trained four Mask RCNN models to see how altering the number of layers trained and the amount of training data affected accuracy. We had two main classes of models: those that were trained through fine tuning and those that used transfer learning as a fixed feature extractor. As explained previously, that means that half of the models trained every layer in each part of the model (region proposal network, feature pyramid network, ResNet) while the other half (fixed feature) only trained the heads of these networks, using the weights of the pre-trained model for every other layer. In addition, we trained both of these types with a small training set (15 images in each class) and a large training set(100 images in each class) to see what effect the training size had on accuracy. As a control, we also included the pretrained model on the coco dateset without any transfer learning. Although the model does not classify metal cans, bottles is one of the classes it is trained on.

4.4 Metrics Used in Evaluation

We used mean average precision (mAP) with an intersection over union (IoU) threshold of .5 to evaluate the performance of the different models. When we ran our model on our validation data we calculated the average precision value on each run by looking at the precision of the previous run and recalculating its value based on the current run. At

Figure 4: Examples of training data. On the left is an example of training data for the model and on the right is an example of what the model should return

the end of the testing of our model on the validation set we had a list of running average precisions that we simply took the mean of to calculate mean average precision [4].

Intersection over Union was used to compare the bounding boxes that the model predicts with the ground truth bounding boxes that we define for the images. As can be seen in Figure 3, intersection over union simply quantifies the area of overlap of the ground truth with the predicted bounding box over the total area bounded by both boxes. As the ground truth and predicted bounding boxes start to converge over subsequent epochs, the value of IoU converges to 1 [4].

5 Training

To train the model we used 30 epochs of 100 steps each, where a batch of images was evaluated and loss calculated for the batch at each step. We chose 30 epochs as empirically we did not see much improvement in any successive epochs after 30. At each step, the loss calculated is a composite of many different categories of loss, including rpn_class_loss, rpn_bbox_loss, mrcnn_class_loss, mrcnn_bbox_loss, and mrcnn_mask_loss. RPN class loss and bounding box loss is the loss associated with the predicted bounding boxes and class prediction within the bounding box by the Region Proposal Network (RPN). MRCNN bounding box loss, mask loss, and class loss are the losses associated with the final refined bounding boxes, masks, and class predictions associated with each bounding box. On Google Colab's 12 core GPU, the models we created took about 3-4 hours to train, depending on whether we trained just the heads of the networks or on all layers of the model.

6 Results

Table 1: Results

Mask RCNN Model	mAP on Val Set	mAP on Test Set
Fine Tuning(Small training set)	.873	.885
Fine Tuning(Large training set)	.845	.905
Fixed Feature(Small training set)	.766	.774
Fixed Feature(Large training set)	.842	.781
Pretrained coco model	.210	.223

The results of calculating the mean average precision of each model on our validation and test image sets are contained in Table 1. These results suggest that fine tuning the

model definitely improved accuracy over just using the pre-trained model as a fixed feature extractor. The best fixed feature mAP results (those associated with model trained on the large image set) still don't match the worst mAP results of the fine tuned models. It does appear that increasing the size of the training image set increased the accuracy of the fixed feature model, however the same doesn't appear to be true for the fine tuned models. In fact, the fine tuned model on the small dataset actually had the highest overall average mAP between the validation and test image sets at .879.

Overall, however, the results were very good across the board. The model performed very well at detection, placing tight bounding boxes, classification of the object within the bounding box, and masking of the object. We've included many examples of the results of the model in the Appendix. We initially thought the model was going to have a tough time telling plastic bottles from glass bottles due to the similarity in their shape, but that didn't seem to be an issue. In addition, the model seemed to deal with occlusion relatively well, which was one of our main goals before we started training. When compared to the model pretrained on coco only, our models performed much better. This is primarily because the pretrained model only classifies bottles among the 80 classes it is trained on, so it was unable to identify cans. While our validation and test sets were evenly distributed between the three classes, any imbalance would affect the pretrained model's scores more than our models.

6.1 Limitations

The result that more training images didn't lead to an in increase in mAP seems counter-intuitive, but there could be a few confounding variables that caused this result. For example, in our large image set, we diversified and expanded the definition of what it meant to be part of each class. Where before we were just using water/beer bottles and wine bottles as examples of glass bottles, we expanded this to include glass liquor bottles which definitely have a different shape to our previous definition of what it meant to be a glass bottle. Adding to this, it's possible we didn't update our validation/test set enough to reflect this expansion of the classes. Another possible reason could be that we effectively increased the size of the training set by a factor of 10 but kept the same number of epochs as we had with the small image set.

Figure 5: Mask RCNN False Positive

In addition, while the model was very good at detecting and classifying our intended targets when they existed, we did find that it had trouble with false positives, especially in the context of objects it had never seen before. Specifically, we found that our model tended to classify dogs and faces as examples of metal cans. This is likely because we didn't include many negative examples in our training set, and no images in our training set included dogs. This might be a problem with over-fitting and could potentially be resolved by adding in negative examples and further increasing the size of our training set. We also found that the model performed worse at detection of objects in context which is likely due to the images we used in training (normally a large centered objects with little to no background).

7 Resources

We used a third party number of tools to complete this project. In order to gather the images, we used Google image search to find images we believed would increase our precision (like 'crushed cans' or 'person drinking glass bottle'). Once we had our images, we used VGG Image Annotator to segment each image and export the masks in a format we could use in training the model.

The base Mask RCNN implementation we chose was Matterport's Mask RCNN implementation as it was incredibly well documented and had samples/demos available which showed how to train on your own data [5]. The other option was Facebook's new PyTorch Mask RCNN implementation, however this implementation had worse documentation and no examples about how to extend the model to our use case. While Facebook's implementation is very new (it was released within the past month or so), there have been some preliminary results which say that Facebook's implementation is slightly more accurate. With that said, it takes longer to train and is slightly slower in the inference stage than Matterport's implementation. As our starting point in transfer learning, we used a pretrained COCO model.

We didn't have a GPU at our disposal, which was necessary to train our model, so we used Google's Colab software, which is essentially an online jupyter notebook that allows users to use a 12 core GPU for 12 hours each day. This allowed us to train our model in about 6 hours instead of the weeks to months it would have taken on a CPU.

8 Conclusion

In this project we trained a Mask RCNN model to detect different classes of recyclables and were able to achieve a maximum mAP score of .905. We believe that this is a good first step in creating a useful consumer and enterprise tool for detecting recyclable material through computer vision. While our results were promising, there is quite a lot of room for improvement, especially with regards to our gathered data in terms of variety as we did have some false positive identifications that could be eliminated with a stronger variety of training images as well as adding in negative examples.

References

[1] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(1):142–158, 2016.

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016.

[4] Jonathan Hui. mAP (mean average precision) for Object Detection. *Medium*.

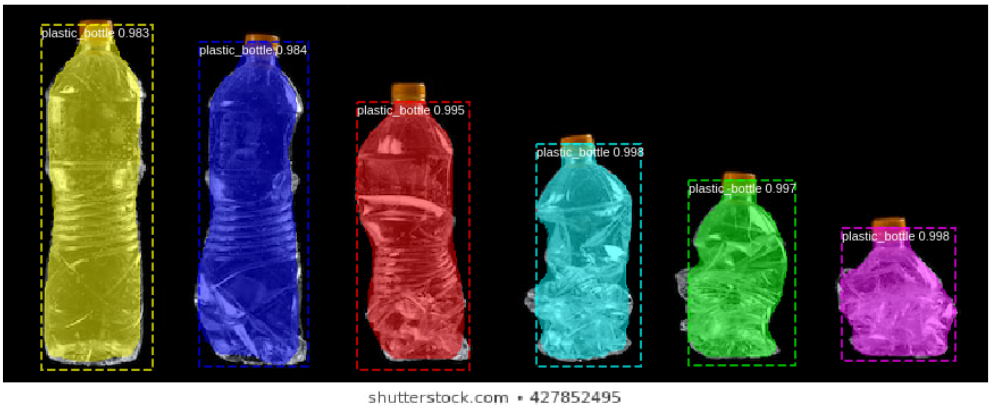
[5] Matterport Inc. Mask-RCNN. *Github*, 2018.

[6] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 936–944, 2017.

[7] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2017.

[8] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, pages 242–264. IGI Global, 2010.

9 Appendix



Predictions

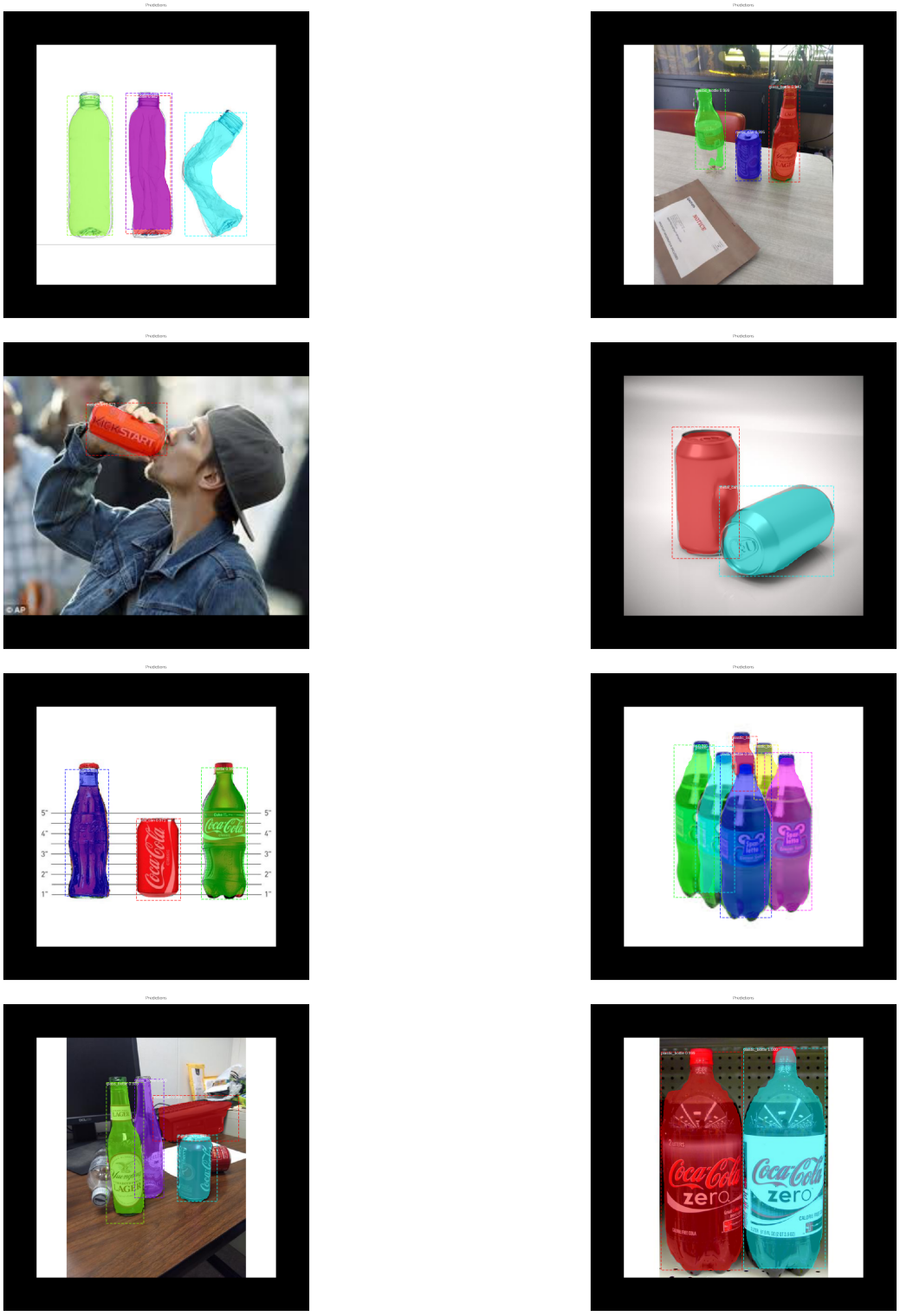


Figure 6: Examples of model output

Figure 7: More examples of model output