Building Detection in Satellite Images:
Improving Resource Distribution in Rohingya Muslim Refugee Camps

Aprotim C. Bhowmik, Nichelle Hall, Minh-An Quinn, {abhowmik, nhall2, minhan}@stanford.edu
Stanford University, CS230, Spring 2018

Introduction

In an effort to support the maintenance of refugee camps, we worked with UNICEF to create an open-source algorithm to estimate the size and population of refugee camps housing Rohingya Muslims. For this project, we implemented the first step of this task: a building detection algorithm to help with mapping refugee camps. Given a satellite image of a Rohingya refugee camp, our model annotates all buildings within the image. Our model is able to get a precision of 93.2% and recall of 87.9% on the refugee camp images.

Data

We used two different datasets to train our model:

- We first trained our model on a dataset from CrowdAI, which consists of over 300,000 RGB satellite images (train/dev/test: 280,000/60,000/60,000) that were annotated with bounding box dimensions of buildings in the image.
- After training our model, we fine-tuned our model using satellite images of Rohingya refugee camps from OpenAerialMap.org. Since these images were not annotated, we had manually crop, resize, and hand-annotate these images. Due to a limited amount of time and resources, we were only able to hand-process 800 of these images.

Features

The input for our model is a raw, RGB satellite image that is 300x300 pixels. We use satellite images for our task because satellite images are widely available throughout the world, particularly in areas with refugee camps. In addition, satellite images are useful for other tasks that might follow the the task of building detection, such as mapping roads and estimating the size of refugee camps.

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i,p_i^*) + \lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i,t_i^*)$$

$$L_{cls}(p,u) = -\log p_u$$

$$L_{reg}(t_u,v) = \sum_i smooth_{L_1}(t_i^u - v_i)$$

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2 & |x|<1 \\ |x|-0.5 & otherwise \end{cases}$$

Equation 1: Mask R-CNN loss functions

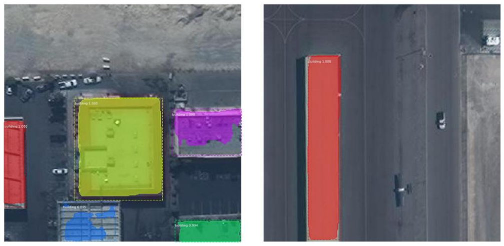

Figure 1: Annotated images outputted by our model of real satellite images of Rohingya refugee camps. Our model is quite robust to more densely packed buildings, as shown on the left image, as well as more sparse images with buildings on the periphery, as shown on the right image.

Model

Our model uses a Mask R-CNN architecture, which enables instance segmentation for pixel-level accuracy of building detection. Mask R-CNN uses a CNN to get proposed regions for an image and then produces a binary mask for each region of interest. This binary mask indicates whether each pixel belongs to an object. When training our model, we started by training the network head; this part is used for bounding box classification and regression. We then trained the ResNet backbone. Finally, we fine-tuned all layers together, attaining pixel-level accuracy for building detection.

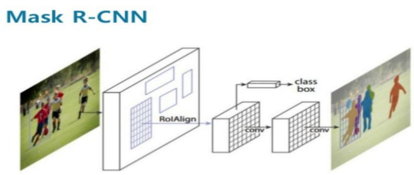

Figure 2: Visualization of the Mask R-CNN architecture.

Results

	# Images	Precision	Recall
CrowdAI Test Set	60,697	69.7%	47.9%
Rohingya Test Set	800 *	93.2%	87.9%

We believe our model performed better on the Rohingya test set because images of the Rohingya camp contained more uniform, box-shaped buildings, making the buildings more easy to identify. However, our model struggled to identify Rohingya camp buildings that were (1) similar in color to the surrounding grass/environment, (2) seemed to have less distinguishable boundaries with the naked eye, and (3) were either small and between more distinguishable buildings or were at the edges of the image.

* Due to lack of availability of annotated refugee camp images, we hand-labeled these 800 images.

Discussion

Though we did not have much labeled data specifically for Rohingya refugee camps, we were still able to obtain good results by using the annotated data from CrowdAI. Due to the similarities between the real satellite images of Rohingya refugee camps and the images from the CrowdAI databases, we were able to transfer the learning of the more general model to the more specific refugee satellite images. To transfer our more general model to the more specific images, we took segments of our large dataset that specifically catered to the Rohingya refugee camps (e.g. images with buildings of similar roof tiles and similar building shape) to fine-tune our model. We believe that if given more annotated data of Rohingya refugee camps, our model would be able to more robustly detect buildings.

Future Work

If given more time, we plan to:

- Annotate more images of Rohingya camps to allow for more fine-tuning, as the availability of such images is low.
- Make model our accessible and open-source to allow UNICEF and affiliated groups to use.
- Use our base architecture to contribute to UNICEF's projects of estimating refugee camp size and population size.

References

[1] Girshick, R. B. (2015). Fast R-CNN. *Computing Research Repository.*

[2] Girshick, R. B., Donahue, J., Darrell, T., \& Malik, J. (2013). Rich feature hierarchies for accurate object detection and semantic segmentation. *Computing Research Repository*.

[3] He, K., Gkioxari, G., Doll, P., \& Girshick, R. B. (2017). Mask R-CNN. *Computing Research Repository*.

[4] Parthasarathy, D. (2017). A brief history of CNNs in image segmentation: From R-CNN to Mask R-CNN. *Computing Research Repository.*

[5] Ren, S., He, K., Girshick, R. B. \& Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Medium - Athelas.*