Image to Image translation using GAN

Project Report | DS5500 | Fall 2019

Prasanna Challa, Nikhar Gaurav, Anish Narkar, Prathwish Shetty
Northeastern University, Boston, MA

1. Summary:

Image-to-Image translation is to learn a mapping between images from a source domain and images from a target domain. In this project, we investigate generative adversarial networks (GAN) to convert satellite images to google maps format. The original paper for Image to Image translation at UC Berkeley gives a good general-purpose solution that works well on image translation problems which were used our reference. But, the performance of the reference model was limited in detecting sharp edges in satellite images. The performance of traditional GAN system in detecting edges could be improved by introducing a superior edge detection architecture like Unet to detect edges in the input, before feeding it to GAN. This could potentially improve the process of translation as the edges are well defined in the modified input, generated by feeding the actual input to Unet.

Figure 1

2. Data Description:

We leveraged the Google Maps Static API (part of Google Cloud Platform) to obtain aerial satellite view along with corresponding processed map view to train our models. 1200 images were scrapped for areas around Boston. The images are in JPG and have a digit filename. Each image has a resolution of 1100x1100 pixels. Images obtained were of 3 types i.e Satellite Image, Street Level Image and Masked Image with outlines for roads and buildings as shown in Figure 2

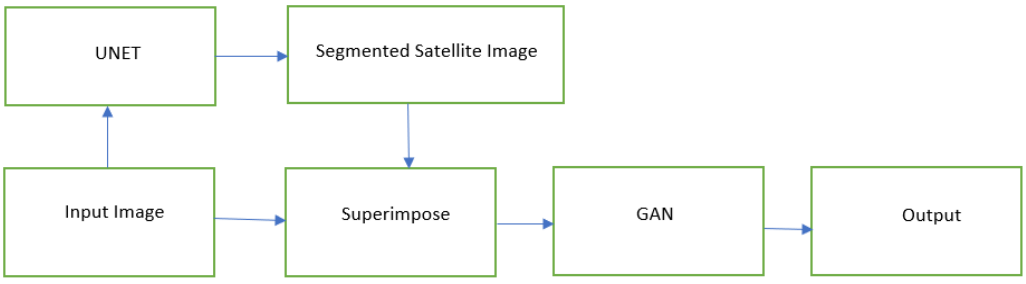
Figure 2

Satellite Image Masked Street Level

3. Methods:

The project aims to develop on top of an existing research paper which uses conditional adversarial networks as a general-purpose solution to image-to-image translation problems[1]. The satellite images will be segmented to detect the exact edges/blocks of each entity like roads, buildings, vegetation in the image. U-net was used for improved edge detection in the satellite image and consequent segmentation of entities. The obtained segmented image will be superimposed on top of the original satellite image to obtain a modified input with well-defined objects and sharp edges. The modified input will be used as the new input for GAN model to obtain the image map.

Figure 3

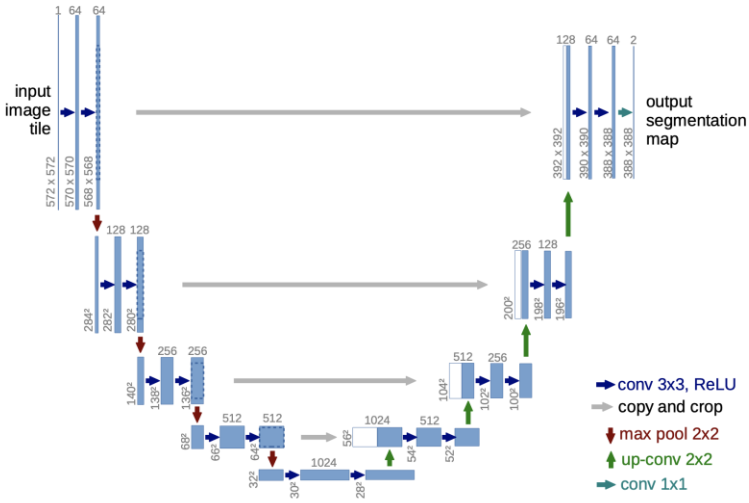


3.1 Unet:

Unet is one of the most widely used and successful edge detection algorithms, particularly in the field of biomedical research. We have used Unet to identify entities like roads in satellite images and

enhance the input provided to GAN. Roads are marked using a high contrast color and are superimposed on the original satellite image before feeding it as input to GAN.

Figure 4



The network architecture is illustrated in Figure 4. It consists of a contracting path (left side) and an expansive path (right side). The contracting path follows the typical architecture of a convolutional network. It consists of the repeated application of two 3x3 convolutions (unpadded convolutions), each followed by a rectified linear unit (ReLU) and a 2x2 max pooling operation with stride 2 for downsampling. At each downsampling step, we double the number of feature channels. Every step in the expansive path consists of an upsampling of the feature map followed by a 2x2 convolution (“up-convolution”) that halves the number of feature channels, a concatenation with the correspondingly cropped feature map from the contracting path, and two 3x3 convolutions, each fol- lowed by a ReLU. The cropping is necessary due to the loss of border pixels in every convolution. At the final layer, a 1x1 convolution is used to map each 64-component feature vector to the desired number of classes. In total the network has 23 convolutional layers.

3.1.1 UNet Experimental framework:

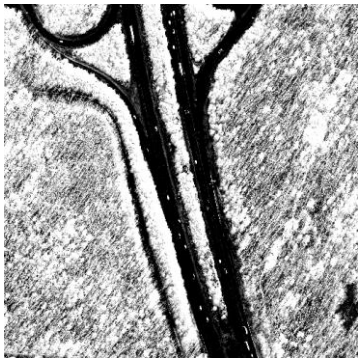
Iteration 1: Satellite images from google maps were fed as the input(source) and the masked images were fed as the output(target) to the U-net. The model was not able to detect sharp edges with the current setting and the results were not satisfactory.

Iteration 2: To better extract contours and roads, the input image was processed by inverting color mapping and increasing contrast. The masked image (target) was also inverted and the model was trained with modified input and output. The results were better than the previous iteration but still has a scope for improvement.

Iteration 3: We employed transfer learning technique by using pre-trained weights from the cell membrane data to predict the roads in satellite images, which produced our best results.

Denoising: The final output image still has a noise issue with random gray areas. The darkest pixel method was used to filter the output and retain only 15% of the darkest pixels to get a good mask.


Input Image to UNET


Contrasted image


Output from UNET with edges

3.2 Superimpose:

The output of Unet has roads segmented with a high contrast color, which is superimposed on top of the original satellite image to accentuate the boundary of demographics.


Denoised Superimposed Image

3.3 GAN:

The superimposed images were used as the source and street-level images were used as the target for training GAN (Generative Adversarial Network) model. GAN architecture has two important components i.e generator and discriminator. Generator model is responsible for outputting new plausible synthetic images and discriminator model classifies the image as real or fake. Both generator and discriminator train simultaneously in an adversarial manner where the generator tries to fool discriminator and the discriminator seeks to better identify the counterfeit image. For this project, the GAN model being used is Pix2Pix which is a conditional GAN where the generation of the output image is conditional on input. In Pix2Pix, discriminator defines a relationship between the output of the model to the number of pixels in the input image such that each output could be mapped to 70x70 patch of the input image. The generator is an encoder-decoder model that has the architecture similar to U-net. This model takes the source and generates a target image which is done by downsampling or encoding the

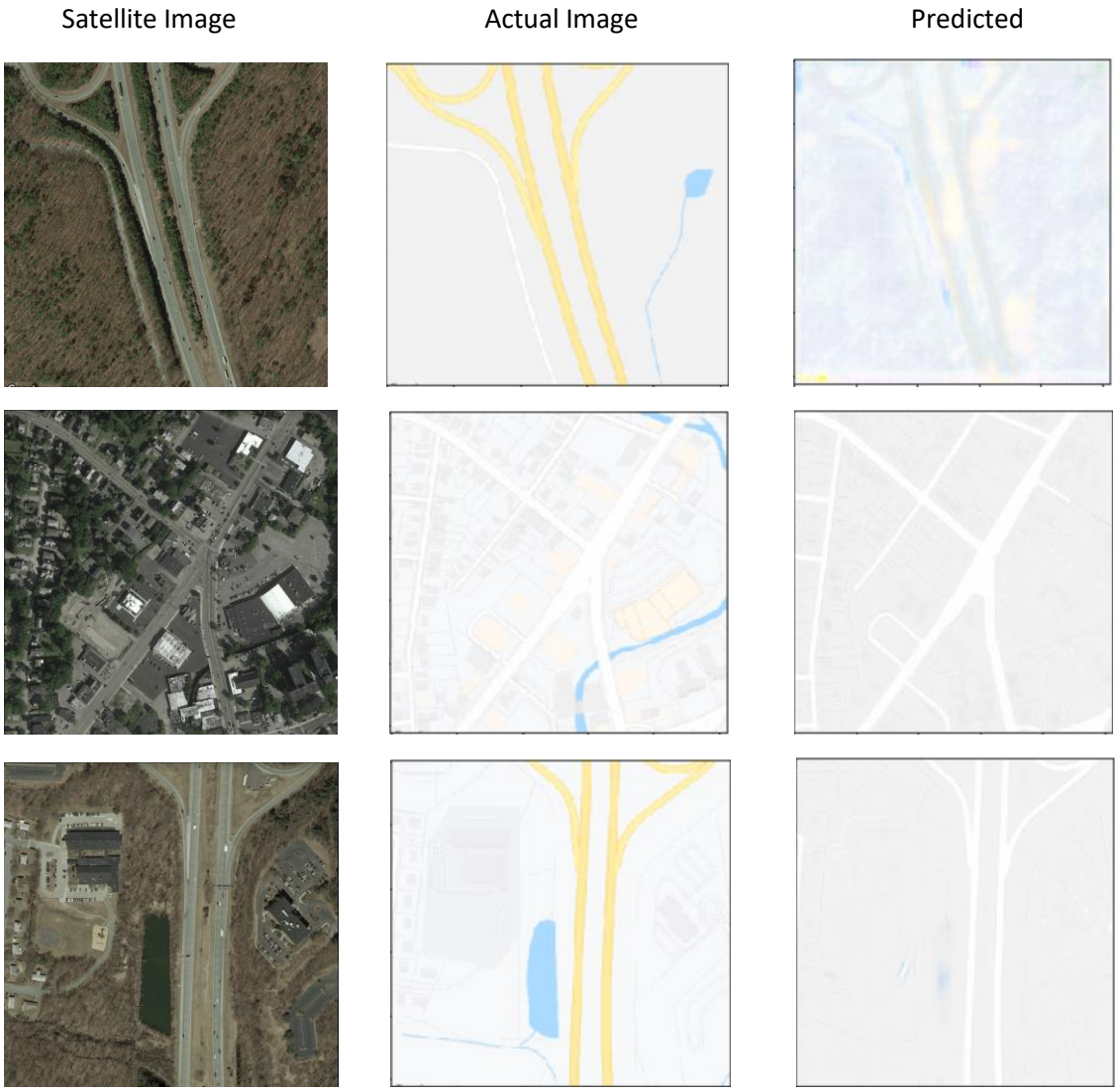
input image to a bottleneck layer, then upsampling or decoding the bottleneck to the size of the output image. For every 10 epochs in the training phase, the current model and the predicted result for a random image is stored and later used for selecting the best model by comparing the results for each model saved.

Initially, we used 100 superimposed images with size 256*256 as the source to train the GAN model with map images as the target to assess the model performance. But, the model tends to produce broken roads and blurry edges with the selected sample size. The resolution was changed to 512*512 and the sample size was increased to 500 images which has significantly improved the model performance. The loss function used in all the models is Binary Cross Entropy from the original paper[1].

Source Image

Target Image

4. Results:



5. Evaluation:

Evaluating accuracy of the final model is similar to the evaluation of segmentation models. These techniques function on pixel-level confusion matrix and these results are aggregated for the entire dataset. In this case map section there are few terminologies which feature prominently, these being Ground Truth and Prediction, ground truth is view of the satellite image which prediction is the output of GAN.

Overall Pixel (OP) accuracy measures the proportion of pixels correctly labeled pixel. Limitation of this measure this the bias in favor of majority class when the data is highly imbalanced. The values we obtained were ranged from 0.8-0.9, this wasn't the true reflection of our model as these values were inflated due to class imbalance Accuracy for our reference paper was also in the same range.

The second metric which we used was Jaccard index(also referred to as Intersection over Union), this metric takes into account false positives in calculation thus penalizing any imbalance in prediction. We also used F1 scores(also referred as Dice), this is similar to IoU but IoU penalizes false alarms more severely. We observed that these metrics indicated a better reflection of the final generated images. The mean IoU which we obtained was 0.53 with a variance of 0.11, mean of Dice scores was 0.56 with a variance of 0.11. These metrics were affected by differences in RGB value of pixels and minor shifts in the position of boundaries were classified as errors, so we came up with metrics which took both these factors into account. Mean IoU and F1 for our reference was 0.52 and 0.55 respectively with variance of 0.12 and 0.08 each..

The fourth metric which we used was Trimap, this was an extension to existing metrics and focussed on boundaries. Trimap evaluates accuracy in a narrow band around each boundary. This metric was less severe on minor shifts in the images and considered average score over the band. The Dice score which we obtained had a mean of 0.41 with a variance of 0.13. The width of the band affects the metric, with a narrow band ignoring important foreground-background information while a large band will eventually converge to OP/PC. The mean trimap value for our reference paper was 0.45 with variance being 0.07.

Visual effectiveness for our results were gravely affected by the efficiency of separation of boundaries between streets and landscape, so there is another class of metrics called contour-based metrics which focus on the closest match between boundary points in the source and target segmentation maps. We use a boundary metric that computes the F1-measure from precision and recall values with a distance error tolerance θ to decide whether a boundary point has a match or not. This metric (BF) ignores any content of segmentation beyond the threshold. We adopted this metric and modified it slightly to account for issues that arise with minor differences in colors. A shade of white color which is indicated as 240,241,239 in ground truth can be indicated as 238,237,237, visually they appear similar to us but this will be identified as a miss by the metrics which we used thus far, but the theta value i.e threshold takes care of such minute differences and gives a better indication of results. The mean value which we obtained was 0.57 with a variance of 0.06, while our reference paper managed a mean of 0.54 with a variance of 0.12 for this metric.

6. Discussion:

We made a few observations about data and methods that could improve the model from our analysis. The satellite images used as input have a high diversity of content including images from urban (more buildings) and rural (mostly sparse and has more vegetation) that differ in a lot of aspects. Generalizing such varying inputs is a hard task for the model, which aims to reduce the error on average. Reduction in such diversity could be a possible solution for this case. And, other edge detection algorithms that are more related and customized to suit the satellite image data could improve the segmentation of input images. The sample size to train the model was limited due to computational difficulties.

7. Statement of contributions:

- Prasanna Kumar Challa was responsible for improving Unet Model working alongside Prathwish Shetty.
- Nikhar Gaurav was responsible for training GAN models and generate results.
- Anish Narkar was responsible for the evaluation of the results by identifying appropriate metrics.
- Prathwish Shetty was responsible for scraping data through Google Api and training Unet Model.

8. References:

[1] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros. Image-to-Image Translation with Conditional Adversarial Networks

[2] Olaf Ronneberger, Philipp Fischer, Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation

[3] Jason Brownlee. How to Develop a Pix2Pix GAN for Image-to-Image Translation

[4] Implementation of deep learning framework — Unet, using keras