MRA-Net for Semantic Segmentation of Remotely Sensed Images

Akshat Ramachandran[1], Sravan Chittupalli[2]**

1. akshatramachandran@gmail.com, 2. sravanchittupalli7@gmail.com

Abstract:

Segmentation of remotely sensed images lies at the intersection of the domains of remote sensing and computer vision. It is used to systematically extract information from data collected by various airborne and space-borne sensors, resulting in a simpler representation. This method is used in various applications which include change detection, land cover and land use classification, resource exploration, the study of natural hazards, and mapping. In this work, we will focus on the study of natural hazards, i.e., building a multi-class semantic segmentation model that categories the given post-disaster (earthquakes in particular) imagery based on damage.

Dataset:

To train our model, we used the Xview2 building damage assessment dataset as provided by the VBL team, which we have split into training and testing datasets. (Note: No additional satellite imagery or disaster type labels were used other than the ones already provided)

Proposed Methodology :

Pre-processing of Dataset:

There is usually a wide class disparity in classification problems involving remotely sensed images. The dataset for the current problem statement is no different either. Therefore, for alleviating the disparity and to provide the model with a more balanced view in terms of class labels we employed a variety of techniques as follows :

- Image Augmentation Techniques like rotation, flipping, variations in brightness, random zoom etc.
- Random crops of the image of the regions particularly consisting of damaged buildings were extracted In different orientations and positions using a simple sliding kernel (Fig.1)

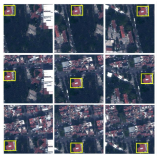

Fig.1

*Both members contributed equally

- CAP (Cut and Paste) Augmentation was employed on random cuttings of heavily damaged buildings and introduced in several different images in random orientations.

Additionally, we observed that the images were large (1024x1024) which is not feasible to train the network on the GPU due to memory constraints. Hence we propose splitting each image into 4 equal halves so that we can employ batch processing and also increase the training speed.

Network Architecture :

Motivation :

The intuition behind using multi-resolution analysis is that images contain features at different scales important for segmentation, therefore, a multi-resolution analysis (MRA) approach is useful for their extraction since this decomposition allows us to even segment structures of various dimensions and structures with ease.

Multi Resolution Analysis :

The central focus of this work is in the utility of MRA, and the method chosen for this is the 2-D Discrete Wavelet Transform (DWT). Transforms are projections of the input onto the space of the basis functions selected]. An image on decomposition with the 2-D DWT is broken down into one approximation and three detail sub-bands in horizontal, vertical and diagonal directions. The basis functions used to decompose are -

$$\phi(t) = \sum_{n} a[n]\phi(2^{m} t - n)$$

$$\psi(t) = \sum_{n} b[n]\phi(2^{m} t - n)$$

Where a[.] are approximation coefficients and b[.] are detail coefficients of a filter bank, m is scaling decomposition level index and n is translating index.

The approximation itself can be decomposed further into four sub-bands, one of which is the approximation of the approximation and the rest are details of the approximation. These two levels of decompositions are integrated with the layers of the U-Net to provide a multi-scale perspective along with directional details. This also supplements the network with decomposed low-frequency and high-frequency parts of data for feature extraction which proves to be pivotal for improvement in performance for segmentation.

Methodology :

For the given task we propose to go with the traditional U-Net architecture composed with an MRA (Multi-Resolution Analysis) framework. The U-Net architecture is a simple encoder-decoder fully convolutional pipeline consisting of contracting (encoder) and expanding/extracting (decoder) paths.

The MRA framework is interspersed into the U-Net Architecture in such a way that it pre-processes the inputs to the network at several stages to increase the contextual overview of the network as the same data on multiple scales is available for feature extraction and learning. It has been previously determined that the Daub basis (db7) is a suitable wavelet basis for the current application. We have employed 2 levels of wavelet decomposition such that the first decomposition is applied on the input image and the second level decomposition is applied on the output of the first level. And the outputs of both these levels are concatenated with suitable layers of the U-Net.

Implementation :

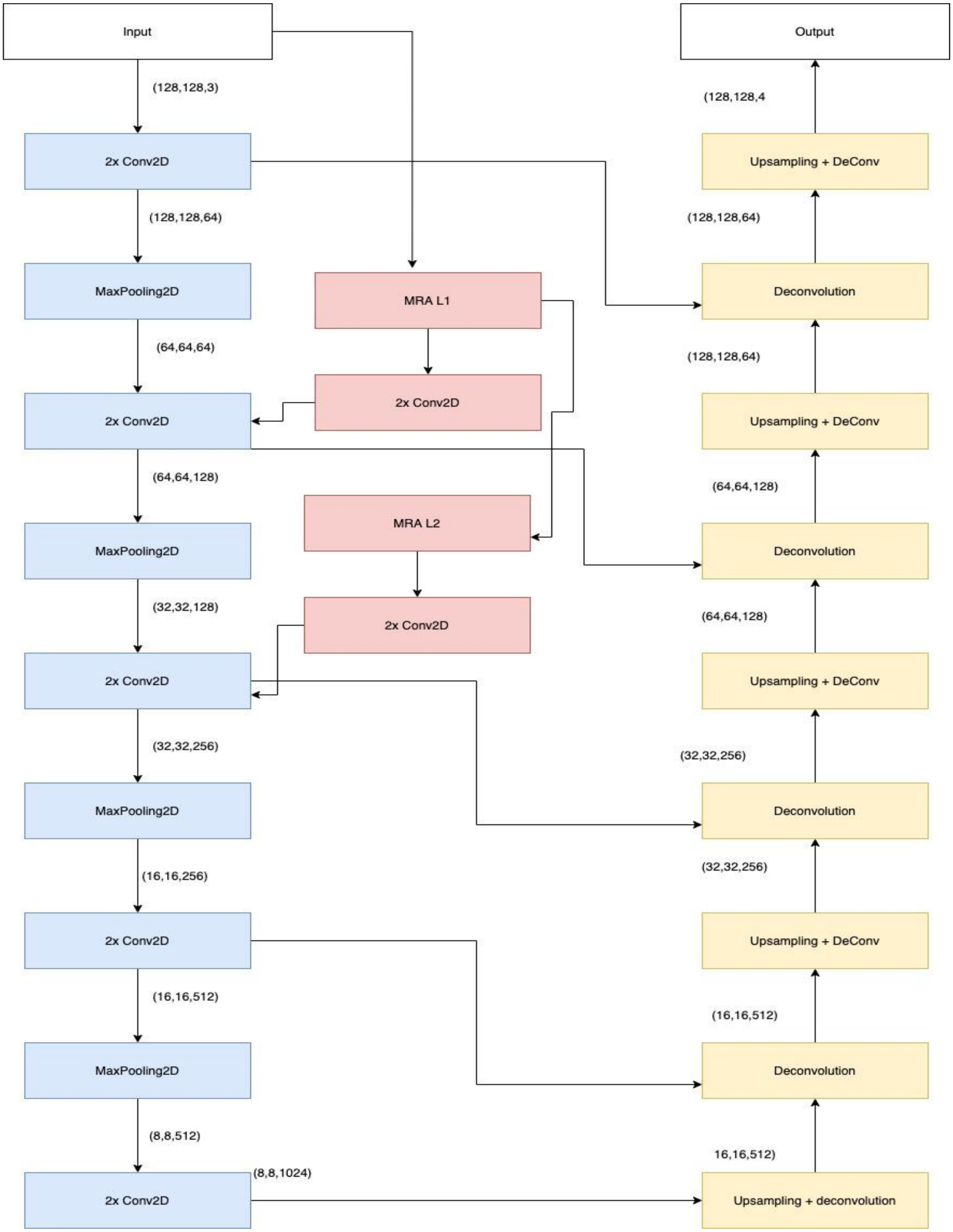

Fig.2.

Training:

For training, we propose a progressively growing network architecture trained in a coarse-to-fine fashion which predicts each next higher resolution i.e the weights used in training a lower resolution image were used to refine the predictions of a higher resolution image.

We trained for a total of 51 epochs which took approximately 72 hours to complete.
The loss at the end of training was on average 0.9.

Loss Function

For the loss function we used a categorical cross-entropy loss function (Fig.4)

$$\text{Loss} = -\sum_{i=1}^{\substack{\text{output}\\\text{size}}} y_i \cdot \log \hat{y}_i$$

Fig.4

We initially tried a Focal loss inspired loss function but the final semantically segmented result was found to be better with the cross-entropy based loss function.

Results :

Fig.3 below shows model predictions of a few images, with the output mask concatenated on the input image.

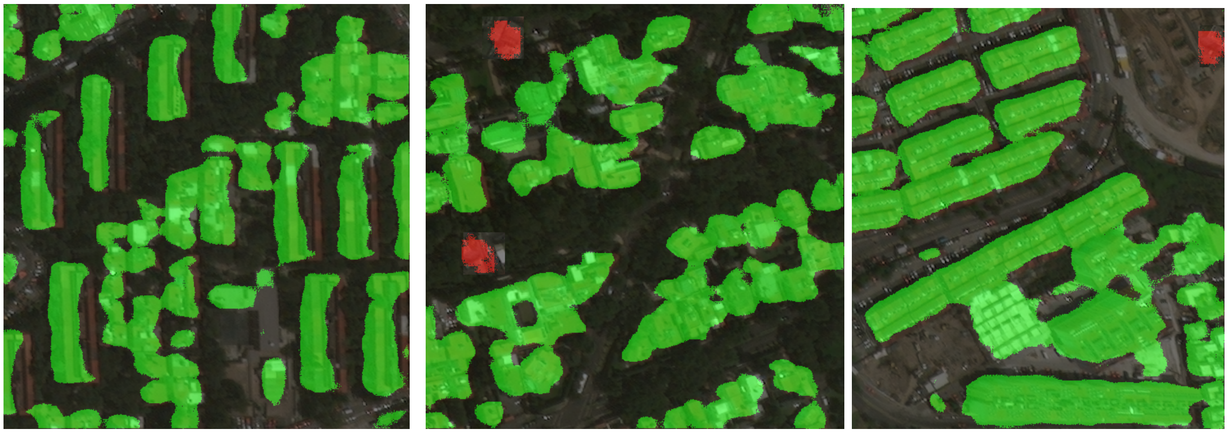

Fig.3.

Quantitative values of Metrics

1. **Pixel Accuracy:**
   It is the percent of pixels in your image that are classified correctly.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Observed Pixel Accuracy: 89.9% ~ 90%

2. **Intersection over Union and Dice:**
   The Intersection over Union (IoU) metric, also referred to as the Jaccard index, is essentially a method to quantify the percent overlap between the target mask and our prediction output. This metric is closely related to the Dice coefficient which is often used as a loss function during training.

$$IoU = \frac{target \cap prediction}{target \cup prediction}$$

Observed MeanIOU = 0.8817
Observed meanDice = 0.9255

3. **Precision :**
   Precision effectively describes the purity of our positive detections relative to the ground truth.

$$Precision = \frac{TP}{TP + FP}$$

Observed meanPrecision = 0.9027

4. **Recall :**
   Recall effectively describes the completeness of our positive predictions relative to the ground truth.

$$Recall = \frac{TP}{TP + FN}$$

Observed meanRecall = 0.9298

Project Repository :

All the codes that have been implemented can be found here :
https://github.com/ARamachandran2000/MRANet

References :

[1] P. Ranjan, S. Patil and R. A. Ansari, "U-net based MRA framework for segmentation of remotely sensed images," 2020 International Conference on Artificial Intelligence and Signal Processing (AISP), 2020, pp. 1-4, doi: 10.1109/AISP48273.2020.9073131.

[2] H. Gao, H. Yuan, Z. Wang and S. Ji, "Pixel Transposed Convolutional Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 5, pp. 1218-1227, 1 May 2020, doi: 10.1109/TPAMI.2019.2893965.

[3] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, Hanqing Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3146-3154

[4] Ghiasi, Golnaz & Cui, Yin & Srinivas, Aravind & Qian, Rui & Lin, Tsung-Yi & Cubuk, Ekin & Le, Quoc & Zoph, Barret. (2020). Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation