Intercontinental Geoinformation Days

http://igd.mersin.edu.tr/2020/

Training data development strategy for applying deep learning in remote sensing applications

Vaibhav Katiyar[*1], **Masahiko Nagai** [1,2]

[1]Yamaguchi University, Environmental Engineering, Graduate School of Sciences and Technology for Innovations, Ube, Yamaguchi Japan
[2] Yamaguchi University, Center for Research and Application of Satellite Remote Sensing, Ube, Yamaguchi Japan

Keywords
DCNNs
Training Data preparation
Remote sensing
U-Net
SegNet

ABSTRACT

Deep Convolution Neural Networks (DCNNs) are playing a very important role in remote sensing applications. However, one of the major challenges in utilizing DCNNs is, of access to the training data. Firstly, there are very few training data available in various fields such as in natural disaster area, secondly, even if it's available it may not be suited to the area we are planning to implement. In such a case creating training data by oneself becomes very important. However, we need to understand that there is a big difference between computer vision dataset and remote sensing dataset. As in the latter case, one scene may cover thousands of Kilometers and the total number of scenes are limited. This is why there is a concept of 'chips' used in remote sensing domain which means a subset of the satellite scene to be used as an 'image' in computer vision sense. This study is comparing the various possible strategies to make the chips from the ALOS-2 scenes and recommending the best after utilizing these chips with popular segmentation network U-Net and SegNet.

1. INTRODUCTION

In recent time, we have seen lots of applications of deep learning in the remote sensing domain. DCNNs have achieved significantly higher accuracy in comparison to other image processing methods especially in cases like road segmentation (Li, Comer and Zerubia 2019), building detection (Li et al. 2019), land cover classification (Zhang et al. 2019) etc. However, we need to understand that these higher accuracies were achieved due to well established public datasets, provided through SpaceNet challenges (Etten et al. 2018), ISPRS labelling contest (Gerke et al, 2014), DeepGlobe challenge (Demir et al. 2018) etc. These kinds of public datasets are not available in many other areas such as in more dynamic cases of natural disasters. Moreover, most of the datasets are available for optical high-resolution images. Synthetic Aperture Radar (SAR) dataset are very scarce, which created the need to develop our own datasets. This study has used ALOS-2 level 2.1 image scenes of HH polarization.

As satellite scenes are too big that is why we need to create image-chips out of it (Han et al. 2017), which can be feed to DCNNs that can run efficiently on the GPUs memory. As per Ning et al. (2020), training a network with a higher number of image chips normally leads to greater accuracy. Also, data augmentation has proved a successful mechanism to increase the variability from limited data and in-turn improves the performance of the networks.

Our main objectives in this study were, studying different strategies to make image-chips for training and then cross-comparing them on two very popular segmentation network, U-Net (Ronneberger et al. 2015) and SegNet (Badrinarayanan et al. 2017).

2. METHOD

For simplicity, this study has selected binary class situation i.e. images with the flooded area and non-flooded area.

Three different satellite scenes subsets have been used to create the training data and the fourth scene subset has been used for testing. U-Net and SegNet have been used as the network for the segmentation.

In the paper, Methodology follows two steps that are creating data and training & testing.

* **Corresponding Author**

*(i502wd@yamaguchi-u.ac.jp) ORCID ID 0000-0003-3453-3158
(nagaim@yamaguchi-u.ac.jp) ORCID ID 0000-0002-0625-762X

Cite this study

Katiyar V & Nagai M (2020). Training data development strategy for applying deep learning in remote sensing applications. Intercontinental Geoinformation Days (IGD), 144-147, Mersin, Turkey

2.1. Training Data Preparation

The chip size for the training data has been selected as 512x512 pixels. This size has been chosen with keeping in mind that flood is the phenomenon which affects a larger area, so to better capture the context the bigger chips size has been chosen. However, we also need to take care of the GPU memory as bigger chips mean smaller batch size and larger training time. Another point to consider is that flooded region is much smaller in comparison to the non-flooded region, this creates an imbalance in the dataset. To reduce this imbalance, the study has selected only those image-chips which has at least 10% of total pixels belonging to a flooded area, called in this paper as valid chips. 10% pixels from ALOS-2 image with 3m spatial resolution means approx. 8-hectare area, over which very few surface water bodies occur in that area. This helps to remove the noise from a smaller lake/ponds. Following methods have been used to extract image chips- 1) Sliding Window 2) Randomized sampling.

2.1.1. Sliding window method

Under this method, a sliding window has been used to slide over the scene and create the chips with the different overlap of successive steps. Four sets of overlap have been used, no overlap, 30% of overlap, 50% overlap and 70% overlap as shown in Fig.1.

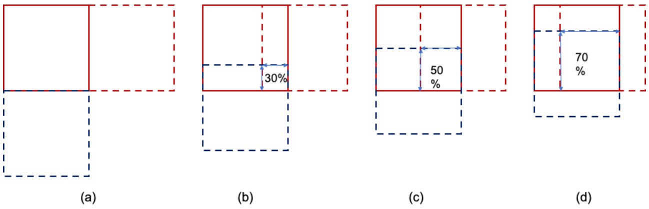

Figure 1. Overlap in successive steps in the case of a sliding window. (a), (b), (c) and (d) are showing the different scenario- 0%, 30%, 50% and 70% overlap for the chips.

As different overlap will result in a different number of the chips, and a greater number of chips may give a better result (Ning et al. 2020). So, in this study, we have decided to select an equal number of chips for each scenario. To decide that how many chips should be selected, we have used 70% overlap, as in this scenario the maximum number of chips will be created to cover the whole scene (Fig. 2), the total valid image-chips have been found around 200 (this number depends upon the size of the image as well as the abundance of foreground pixels). For this reason, in all scenario we have created 200 image-chips, if the total number of valid chips is less than this limit such as in no-overlap scenario, then we just duplicated the valid chips to satisfy the condition.

2.1.2. Randomized sampling method

In this method, the 200 random patches have been selected using a 'sampling' method. The validity of these patches was calculated and recursively sampling has been done till the time the total number of valid image-chips reach to 200 or more (shown by Red square in Fig. 2). In the end, only 200 of valid-chips has been saved for the training step (Yellow square in Fig. 2).

This way we have created the training set of total 600 image chips from three different satellite scenes for each method and scenario.

2.2. Training and Testing

The training datasets have been used, to train U-Net and SegNet for each method and scenario. Each of these networks has five encoder and five decoder blocks with one bottleneck block, each block having two convolutions. Each kernel of an encoder is of 3x3 size with 'same' padding while decoder kernel size has been chosen as 2x2. The network has been trained for 50 epochs with the batch size of 10 and validation split of 20%, along with the binary-cross-entropy as loss function and the optimizer Adam.

All the trained models have been tested on the same test set and the result has been compared based on F1-Score, Accuracy and Jaccard Score.

$$F1\ Score = \frac{2 * Precision * Recall}{(Precision + Recall)}$$

$$Accuracy = \frac{Correctly\ identified\ pixels}{Total\ Pixels\ in\ the\ Image}$$

$$Jaccard\ Score = \frac{GroundTruth\ \cap Predicted}{GroundTruth\ \cup Predicted}$$

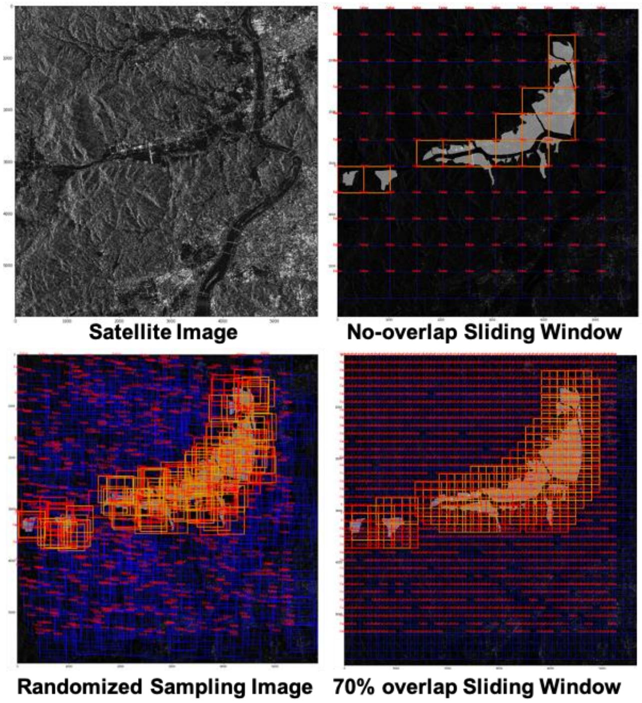
Satellite Image
No-overlap Sliding Window
Randomized Sampling Image
70% overlap Sliding Window

Figure 2. Training data creation. Image-chips were created concerning the flood mask as can be seen in all the images above leaving satellite image aside. Blue colour bounding boxes (BBs) in the images are invalid due to not meeting the condition of 10% water pixels and Red colour BBs are valid while yellow BBs are the selected chips from total valid ones.

3. RESULTS

In the step of the creation of the training data, maximum time was spent in randomized sampling method and in case of sliding window method, time is decreasing with decreasing overlap. So, the fastest method for image-chip creation was sliding window with no overlap.

After training data preparation, U-Net and SegNet were trained on each training set i.e. five different training sets. Each training took between 15-20 mins for finishing 50 epochs. Here it needs to be focused that, hyper tuning of the network has not been done and rather than saving the best model, the model has been saved after 50 epochs. As the main aim was to do the cross-method comparison for different methods of training data preparation.

Training accuracy and binary cross-entropy loss during the training have also been plotted (Fig. 3). As per the plot, the no-overlap scenario was converging fastest, while randomized sampling one was more versatile and showing less sudden peaks with increasing epochs.

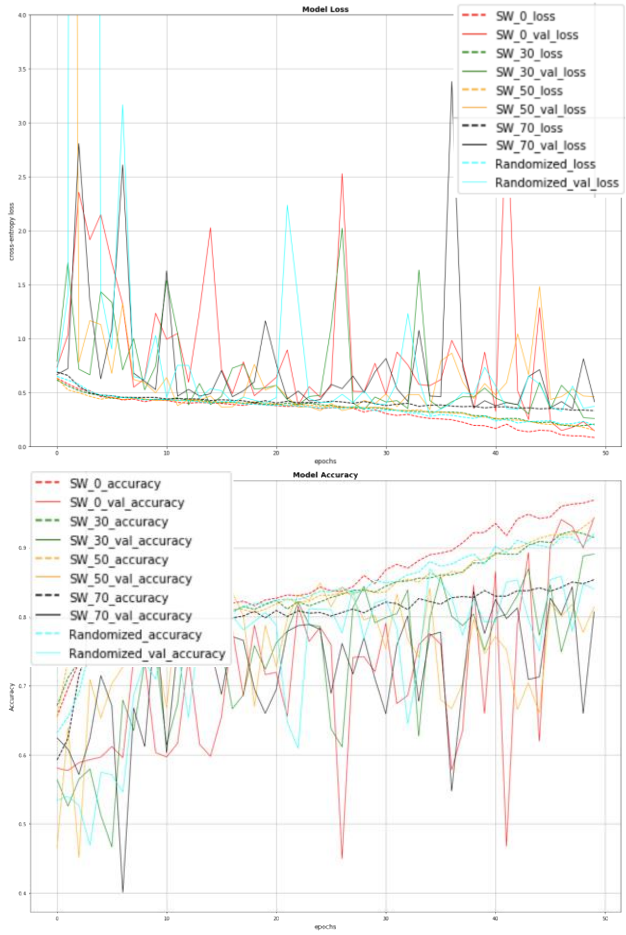

Figure 3. Cross-entropy loss and accuracy plot during the SegNet training.

The accuracy of the trained model has been measured on the testing scene which is a completely different flood event happened at a different place. The comparison based on Accuracy, Jaccard Score and F1 Score have been charted as shown in Fig. 4.

4. DISCUSSION

As we can see by our result that the F1 score and Jaccard score is very low throughout the different networks as well as different method or scenarios. This is the case because we have selected one of the most difficult problems, first SAR images segmentation already a difficult problem and on top of that flooded area segmentation which has included urban flood, mountainous region, paddy field area etc. making this a very difficult case. However, we need to focus on our main objective and that is the cross-method comparison for the training data development.

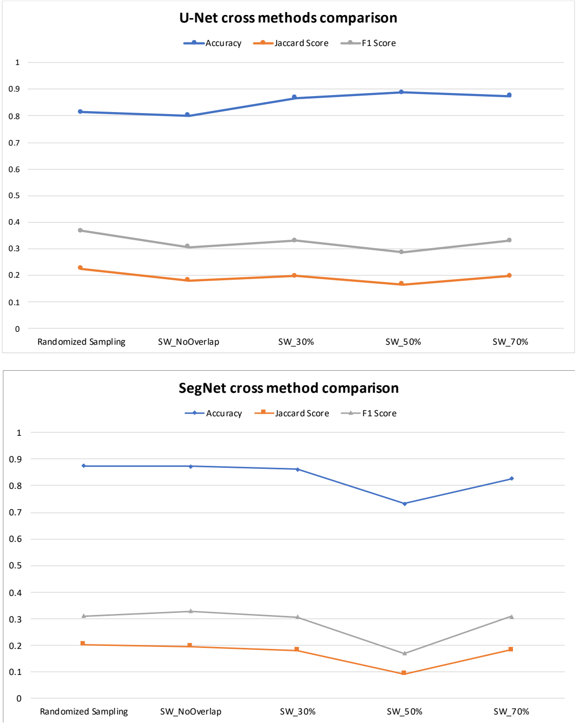

Figure 4. Image-chips creation. SW here represents a sliding-window method and different % in number shows the overlap per cent of consecutive windows.

The study has found that 50% overlap has poorest scores throughout, one of the possible reasons can be that it has learned more on the negative samples, this can be seen in Fig. 4, U-Net cross method comparison. In the figure, 50% overlap scenario is showing the highest accuracy which is just a measure of total correct pixels predicted and as non-flooded pixels are much higher in the scene, so predicting most of the pixels as non-flooded area leads to increase the accuracy. However, F1 Score and Jaccard score are lowest showing the failure of predicting the right class for foreground i.e. flooded area.

Overall Randomized sampling shows best or approx. equal score around all the parameters. This seems logical too as all the methods in the study can also be seen as image-chips creation along with data augmentation in the manner of ‘translation’ (moving image to X and Y direction). Randomized sampling can be seen as translation of the image with an arbitrary factor within (0,512), whereas other method and scenarios have a fixed-step translation.

As per the time concerned for image-chips creation, it was the fastest in the case of the sliding window with no overlap. It was the case due to a smaller number of chips possible with this condition and then for making the desired number of chips, we have just duplicated the valid chips. This duplication is the reason why Fig. 3 shows sharp convergence of SW_O (in Red colour) but with multiple bigger peaks at each interval. This behaviour may be better monitored when training for a greater number of epochs.

Although we have found that randomized sampling is best suited in this case, still it needs a greater number of test cases, for checking that either this observation holds in other cases or not. Moreover, we need to emphasize here that these strategies are not an alternative to more data. As in this case, data information remains the same and the only augmentation happens but augmentation has its limit.

5. CONCLUSION

Training data is a very important first step in implementing the DCNNs in the remote sensing areas. We have presented the different strategies to create the training data especially when we have limited satellite scenes. These training datasets has been tested on two very popular segmentation network U-Net and SegNet. Overall both networks show better learning capability with the randomized sampling method. This is justifiable when we compare randomized sampling as a variant of translation of the image chips with an arbitrary data augmentation factor. However, randomized sampling may not give a better result when total selected image-chips are few, as random image chips possibly not be distributed in the whole image but this problem reduces with the increasing number of tiles chosen.

ACKNOWLEDGEMENT

Authors are thankful to JAXA (Japan Aerospace Exploration Agency) for giving access to the ALOS-2 images as well as to the GSI (Geospatial Information Authority of Japan) for proving the flood mask.

REFERENCES

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39, 2481-2495.

Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., & Raskar, R. (2018). DeepGlobe 2018: A Challenge to Parse the Earth through Satellite Images. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 172-17209.

Etten, A.V., Lindenbaum, D., & Bacastow, T.M. (2018). SpaceNet: A Remote Sensing Dataset and Challenge Series. ArXiv, abs/1807.01232.

Gerke, M., Rottensteiner, F., D Wegner, J., Sohn, G.: Isprs semantic labeling contest (09 2014)

Han, S., Fafard, A., Kerekes, J., Gartley, M.G., Ientilucci, E., Savakis, A., Law, C., Parhan, J., Turek, M., Fieldhouse, K., & Rovito, T. (2017). Efficient generation of image chips for training deep learning algorithms. Defense + Security.

Li, T., Comer, M., & Zerubia, J. (2019). Feature Extraction and Tracking of CNN Segmentations for Improved Road Detection from Satellite Imagery. 2019 IEEE International Conference on Image Processing (ICIP), 2641-2645.

Li, W., He, C., Fang, J., Zheng, J., Fu, H., & Yu, L. (2019). Semantic Segmentation-Based Building Footprint Extraction Using Very High-Resolution Satellite Images and Multi-Source GIS Data. Remote. Sens., 11, 403.

Ning, H., Li, Z., Wang, C., & Yang, L. (2020). Choosing an appropriate training set size when using existing data to train neural networks for land cover segmentation. Annals of Gis: Geographic Information Sciences, 1-14.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention , 234–241, Munich Germany.

Zhang, C., Wei, S., Ji, S., & Lu, M. (2019). Detecting Large-Scale Urban Land Cover Changes from Very High-Resolution Remote Sensing Images Using CNN-Based Classification. ISPRS Int. J. Geo Inf., 8, 189.