Semi-Supervised Classification and Segmentation on High Resolution Aerial Images

Sahil Khose[*] [*] **Abhiraj Tiwari**[*] **Ankita Ghosh**[*]
Manipal Institute of Technology, Manipal
{sahil.khose, abhiraj.tiwari1, ankita.ghosh1}@learner.manipal.edu

Abstract

FloodNet is a high-resolution image dataset acquired by a small UAV platform, DJI Mavic Pro quadcopters, after Hurricane Harvey. The dataset presents a unique challenge of advancing the damage assessment process for post-disaster scenarios using unlabeled and limited labeled dataset. We propose a solution to address their classification and semantic segmentation challenge. We approach this problem by generating pseudo labels for both classification and segmentation during training and slowly incrementing the amount by which the pseudo label loss affects the final loss. Using this semi-supervised method of training helped us improve our baseline supervised loss by a huge margin for classification, allowing the model to generalize and perform better on the validation and test splits of the dataset. In this paper, we compare and contrast the various methods and models for image classification and semantic segmentation on the FloodNet dataset.

1 Introduction

The frequency and severity of natural disasters threaten human health, infrastructure and natural systems. It is extremely crucial to have accurate, timely and understandable information to improve our disaster management systems. Rapid data collection from remote areas can be easily facilitated using small unmanned aerial systems which provide high-resolution images. Visual scene understanding of these collected images is vital for quick response and large scale recovery post-natural disaster. Classification and segmentation tasks are fitting in such situations as they can provide scene information to help the task force make decisions.

One of the major challenges with generating a vision dataset is the cost of labeling the data, especially for semantic segmentation. This often leads to labels only for a small percentage of the data which gives rise to the need for semi-supervised methods that can produce results that are on par with supervised methods. Another challenge that we face apart from the lack of labeled dataset is the heavy class imbalance. Lack of labeled data coupled with class imbalance makes it a very challenging task to solve. Our contribution in this paper is two folds: semi-supervised classification and semi-supervised semantic segmentation. Through our approach, we try to tackle these problems and produce creditable results.

2 Related Works

Supervised Classification: Over the years, several architectures and methods have emerged which have leveraged extensive datasets like ImageNet Deng et al. (2009) to produce state of the art results

[*] Authors have contributed equally to this work and share first authorship. Link of the code: https://github.com/ankitaghosh9/FloodNet

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2021.

using supervised learning. The architectures used in our paper are ResNet He et al. (2016) and EfficientNet Tan and Le (2019).

ResNet proposes residual connection architecture which makes it feasible to train networks with a large number of layers without escalating the training error percentage. Using the technique of skip connections, it resolves the issue of vanishing gradient.

EfficientNet proposes a simple but highly effective scaling method that can be used to scale up any model architecture to any target resource constraints while maintaining model efficiency. They observed the effects of model scaling and identified that carefully balancing network depth, width and resolution can lead to better performance.

Supervised Semantic Segmentation: Segmentation aids in extracting the maximum amount of information from an image. Semantic segmentation associates every pixel of an image with a class label. Deep learning models like UNet Ronneberger et al. (2015), PSPNet Zhao et al. (2017) and DeepLabV3+ (DLV3+) Chen et al. (2018) have provided exceptional results for this task.

The architecture of UNet is divided into two parts: contracting path and expansive path. The contracting path follows the generic framework of a convolutional network while the expansive path undergoes deconvolution to reconstruct the segmented image.

PSPNet exploits the capability of global context information using different region-based context aggregation by introducing a pyramid pooling module with the proposed pyramid scene parsing.

DeepLabV3+ is a refinement of DeepLabV3 which uses atrous convolution. Atrous convolution is a powerful tool to explicitly adjust the filter's field-of-view as well as control the resolution of feature responses computed by Deep Convolution Neural Network.

Semi-supervised Approach: Pseudo-Label Lee (2013) proposes a simple semi-supervised learning approach. The idea is to train the neural network in a supervised fashion with both labeled and unlabeled data simultaneously. For unlabeled data, pseudo labels are generated by selecting the class which has maximum predicted probability. This is in effect equivalent to Entropy Regularization Chapelle et al. (2006). It favors a low-density separation between classes, a commonly assumed prior for semi-supervised learning.

3 Classification

The dataset has 2343 images of dimensions $3000 \times 4000 \times 3$ divided into train(1445), valid(450) and test(448) splits. Out of the 1445 train images 398 are labeled and 1047 are unlabeled. In this section we describe our approach for classifying the FloodNet dataset Rahnemoonfar et al. (2020) into 2 classes, Flooded and Non-Flooded.

3.1 Data and Preprocessing

The labeled data consists of 51 flooded and 347 non-flooded samples. The large class imbalance prevents the model from achieving a good F1 score while training with the labeled dataset. To prevent this we used a weighted sampling strategy while loading the data in the model as inspired from R et al. (2021). Both the classes were sampled equally during batch generation.

Dataset was heavily augmented to get more images for training the model. The image samples were randomly cropped, shifted, resized and flipped along the horizontal and vertical axes.

We downsized the image to 300×400 dimensions to strike a balance between processing efficiency gained by the lower dimensional images and information retrieval of the high-resolution images.

3.2 Methodology

ResNet18 with a binary classification head was used for semi-supervised training on the dataset. The model was trained for E epochs out of which only the labeled samples were used for E_i^α epochs after which pseudo labels were used to further train the model. α has an initial value of α_i that increases up to α_f from epoch E_i^α to E_f^α as described in Algorithm 1.

Algorithm 1: Semi-supervised classification train loop

```
   Input: Sample image
   Output: Class of the given image
1  for $epoch \leftarrow 0$ to $E$ do
2      if $epoch < E_i^\alpha$ then
3          $\alpha \leftarrow \alpha_i$
4      else if $epoch < E_f^\alpha$ then
5          $\alpha \leftarrow \frac{\alpha_f - \alpha_i}{E_f^\alpha - E_i^\alpha} * (epoch - E_i^\alpha) + \alpha_i$
6      else
7          $\alpha \leftarrow \alpha_f$
8      end if
9      Run the model on train set
10     $loss \leftarrow BCE(l, \hat{l}) + \alpha * BCE(u_{epoch}, u_{epoch-1})$
11     Generate the pseudo labels for unlabeled data
12     Evaluate the model on validation set
13 end for
```

A modified form of Binary Cross-Entropy (BCE) was used as the loss function as shown in line 10 in Algorithm 1 where l is the label of a sample, $\hat{l}$ is the predicted class for labeled sample and u_{epoch} is the predicted class for an unlabeled sample in the current epoch. This loss function was optimized using Stochastic Gradient Descent (SGD) Robbins and Monro (1951).

3.3 Experiments

We used ResNet18 as it is computationally efficient. We experimented with Adam Kingma and Ba (2015) optimizer and SGD. Optimizing using SGD was much more stable as compared to Adam Kingma and Ba (2015) optimizer and it was less susceptible to overshooting. Different values of α were experimented with and it was found that a slow and gradual increase in alpha was better for training the model. Our best performing model uses $\alpha_i = 0$ and $\alpha_f = 1$. The value of α increases from epoch $E_i^\alpha = 10$ to $E_f^\alpha = 135$. The model was trained on batch size of 64.

3.4 Results

Our system performed significantly better than all the classification baseline results mentioned in the FloodNet paper while having a considerably smaller architecture (half the number of parameters) as shown in Table 1. Our best model achieves **98.10% F1** and **96.70% accuracy** on the test set.

4 Segmentation

In this section, we detail our approach for training a model which generates multi-class segmentation masks for given images. The semantic labels for the task is a 10 pixel-level class segmentation mask consisting of Background, Building-flooded, Building non-flooded(NF), Road-flooded, Road non-flooded(NF), Water, Tree, Vehicle, Pool and Grass classes. They are mapped from 0 to 9 respectively.

Table 1: Classification models comparison

Metrics	InceptionNetv3	ResNet50	Xception	**ResNet18 (our)**
Training Accuracy	99.03%	97.37%	**99.84%**	96.69%
Test Accuracy	84.38%	93.69%	90.62%	**96.70%**
#params	23.8M	25.6M	22.9M	**11.6M**

4.1 Data and Preprocessing

To expedite the process of feature extraction we apply bilateral filter to the image, followed by two iterations of dilation and one iteration of erosion. For image augmentation we perform shuffling, rotation, scaling, shifting and brightness contrast. The images and masks are resized to 512×512 dimensions while training to preserve useful information.

4.2 Methodology

The dataset contains labeled masks of dimension $3000 \times 4000 \times 3$ with pixel values ranging from 0 to 9, each denoting a particular semantic label. The pixel values of the labeled masks are one-hot encoded to generate labels with 10 channels, where i^{th} channel contains information about i^{th} class.

We experiment with various encoder-decoder and pyramid pooling based architectures to train our model, the details of which are mentioned in Section 4.3. The loss function used is a weighted combination of Binary Cross-Entropy loss (BCE) and Dice loss as it provides visually cleaner results.

We apply semi-supervised learning (SSL) and generate pseudo masks for the unlabeled images. While training the model for E epochs, the labeled samples were used for E_i^α epochs where Adam is used as an optimizer. After that pseudo masks were used to further train the model as described in Algorithm 1. α has an initial value of α_i that increases upto α_f from epoch E_i^α to E_f^α. SGD optimizer with 0.01 LR is used when pseudo masks are introduced to the model.

4.3 Experiments

We adopt one encoder-decoder based network named UNet, one pyramid pooling module based network PSPNet and the last network model DeepLabV3+ employs both encoder-decoder and pyramid pooling based module. We train all of them in a supervised fashion. For UNet, PSPNet and DeepLabV3+ the backbones used were ResNet34 , ResNet101 and EfficientNet-B3 respectively.

For UNet the learning rate was 0.01 with step LR scheduler set at intervals [10,30,50] and decay factor γ set to 0.1. For PSPNet the learning rate was 0.001 without any LR decay. For DeepLabV3+ the learning rate was 0.001 with step LR scheduler set at intervals [7,20] and γ set to 0.1.

Adam optimizer and batch size of 24 was used for all the models with MIoU as the evaluation metric. We observed the best results when we weighed the BCE loss and Dice loss equally.

Once we recognized the best performing model on the task, we trained a DeepLabV3+ using SGD optimizer in a semi-supervised fashion. Due to resource constraints we randomly sampled unlabeled data with the ratio of $1 : 10$ for generating the pseudo masks.

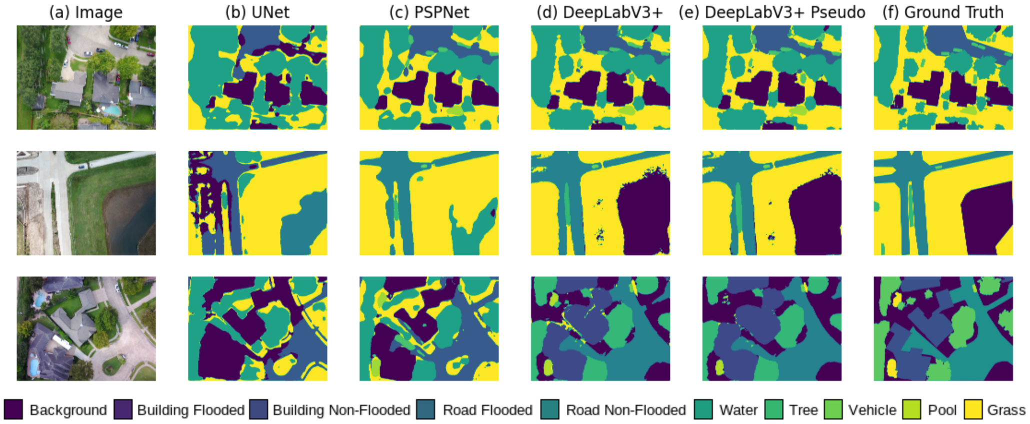

Figure 1: Visual comparison on FloodNet dataset for semantic segmentation

Table 2: Classwise segmentation results on FloodNet testing set

Method	Back-ground	Building NF	Building Flooded	Road NF	Road Flooded	Water	Tree	Vehicle	Pool	Grass	**mIoU**
UNet	0.	0.	0.34	0.	0.45	0.49	0.47	0.	0.	0.64	0.239
PSPNet	0.04	0.45	0.66	0.32	0.73	0.61	0.71	0.14	0.18	0.82	0.4665
DLV3+	0.16	**0.49**	**0.69**	0.45	**0.76**	**0.72**	**0.76**	0.14	**0.18**	**0.85**	0.5204
DLV3+ (SSL)	**0.17**	0.48	**0.69**	**0.48**	0.75	**0.72**	**0.76**	**0.15**	**0.18**	**0.85**	**0.5223**

4.4 Results

Table 2 showcases the comparison of the best models we achieved for each architecture. The best test set result was achieved by a DeepLabV3+ architecture with EfficientNet-B3 backbone. A few examples of the predictions of the model against the ground truth are provided in Figure 1.

5 Application Context

A good assessment of the damages is critical during and even after a natural disaster to aid rescue efforts. After a natural disaster, the response team must first locate the affected areas and distinguish flooded areas from non-flooded neighbourhoods. This can be achieved through the classification task. Classification helps in immediate identification and a quick response to manage the calamity at hand.

Further, flooded structures and roadways need to be identified in each neighbourhood so that a rescue team can be dispatched to the impacted areas. The segmentation in FloodNet allows us to identify flooded buildings and roads from non-flooded ones. Additionally, it segments other water bodies including pools, which help in avoiding misinformation about the flooding. We can also identify objects like trees and vehicles which can help in recognizing the exact location in which the emergency team should be deployed to rescue any individuals in peril.

Classification and segmentation can immensely improve the utility of an application for disaster assessment created to instruct or provide prompts to the rescue operatives.

6 Conclusion

In this work, we have explored methods to approach semi-supervised classification and segmentation along with handling the class imbalance problem on high-resolution images. We have conducted a range of experiments to obtain the best possible technique and models to optimize for the tasks.

Our classification framework achieves laudable results with just 398 labeled images. Our segmentation framework shows an increase of 0.19% on using pseudo labels. This provides a wide scope of improvement as the amount of unlabeled data is three times the amount of labeled data which if employed efficiently can produce superior results.

We foresee multiple opportunities for future research. Self-supervised pretraining, attention based models, addition of discriminative loss and vision transformers can be explored as future work.

References

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien, editors. *Semi-Supervised Learning*. The MIT Press, September 2006. doi: 10.7551/mitpress/9780262033589.001.0001. URL `https://doi.org/10.7551/mitpress/9780262033589.001.0001`. 2

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848. 1

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)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90. 2

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1412.6980`. 3

Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 07 2013. 2

Sidharth R, Abhiraj Tiwari, Parthivi Choubey, Saisha Kashyap, Sahil Khose, Kumud Lakara, Nishesh Singh, and Ujjwal Verma. Bert based transformers lead the way in extraction of health information from social media, 2021. 2

Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin R. Murphy. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *CoRR*, abs/2012.02951, 2020. URL `https://arxiv.org/abs/2012.02951`. 2

Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951. doi: 10.1214/aoms/1177729586. URL `https://doi.org/10.1214/aoms/1177729586`. 3

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 2

Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019. URL `http://proceedings.mlr.press/v97/tan19a.html`. 2

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2