COMPARING DEEP LEARNING METHODS FOR FLOOD MAPPING WITH SAR DATA

Rodrigo BRUST SANTOS, Christina CARTY, Rama PARASA, Dhia TURKI

Université de Bretagne Sud - Copernicus Digital Earth Erasmus Mundus

ABSTRACT

In this paper, three distinct deep-learning models are presented for the participation in Track 1 of the IEEE GRSS Data Fusion Contest 2024. The objective of this challenge was to predict flooded areas using SAR imagery of the flood events, along with additional data such as, land cover, and more. The first model employed in the study is a naive CNN designed without reliance on established benchmark architectures. The second model utilizes a U-NET architecture and is trained from scratch. Finally, the third model is built by fine-tuning a pre-trained SegFormer, that uses a hierarchical transformer architecture. We conclude that the SegFormer model gave the best performance for the task, with an F1-Score accuracy of 83.8%.

1. INTRODUCTION

The increased onslaught of natural disasters brought about by climate change has been well-noted in recent years [1] and improving the ability to quickly deal with the damage brought on by these events has become a shared focal point across many sectors such as governmental, private, humanitarian and scientific. Remote sensing imagery has historically played a valuable role in extreme weather event management and response, but recent advancements in deep learning technologies have allowed the leveraging of this data in an especially promising and powerful way that allows faster and more efficient prediction of damage extent and emergency response.

The goal of the present study is to explore this synergy between remote sensing and natural disasters, specifically in the context of floods and flood mapping. We compare the performance of three different neural network models in predicting flooded areas from various SAR inputs. These data, as well as the motivation behind this analysis, were provided by the 2024 Data Fusion Contest (DFC) in Flood Rapid Mapping put on by the Institute of Electrical and Electronics Engineers (IEEE) in collaboration with the Geoscience and Remote Sensing Society (GRSS).

The models used, namely a Naive CNN implementation, a U-Net model, and a fine-tuned Vision Transformer model using Segformer as a backbone, are trained to map not only the pre-existing water bodies but also the predicted flooded areas. The results here highlight the struggle of all three models to learn the flooded areas specifically, despite general success at mapping the persistent water bodies. While all models performed well relative to each other (within 0.1 F1 score points with the Segformer model holding the slight advantage in performance), the learning paths of each model, and thus the prediction maps, showed characteristic differences.

The rest of the paper is organized as data, models, results, discussion, and conclusion. to five sections. The first section explains the dataset and the pre-processing steps carried out on the data. The second section describes the architectures and configurations of the deep learning models deployed in the study. The third section discusses the results of our studies. The fourth section focuses on the implications of our analyses. Finally, the last section makes concluding remarks and briefly discusses potential next steps.

2. DATA

The dataset provided for Track 1 or the SAR Track of the DFC2024 challenge comprises 6 data layers. These are - Sentinel 1 SAR imagery as VV and VH products, MERIT DEM, Copernicus DEM (Cop. DEM), ESA Landcover Map (LCVM), and, Water Occurrence Probability (Water Prob.). A total of 1631 image tiles of size 512 x 512 pixels were provided for the training purpose along with the corresponding label images containing binary information per pixel - "water" or "non-water". Additionally, there is a validation set comprising 349 image patches with the same data layers, but provided without label information, treated as the 'test' data in our studies. The datasets provided in the challenge mainly come from the Copernicus Emergency Management Service and a hydrodynamics modeling exercise.

We incorporated domain knowledge into the models by deriving four additional data layers from the originally provided layers and stacking both sets together. These additionally derived data layers are: **1.** VV/VH ratio (Referred to as VV/VH); **2.** water binary mask (WBM) generated from the LCVM; **3.** distance transform (Distance) generated from the WBM ; and **4.** a sum of WBM with the first percentile of heights from DEM Merit, denominated DEM Waster Mask (DEM-WM). In all models, the input data layers, both original and derived, were normalized using the mean-centering approach as a pre-processing step and batch-normalized during the training process.

The training set of 1631 images, was split into training

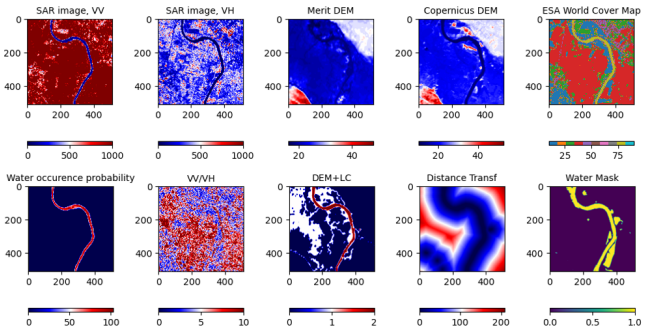

Fig. 1. Sample of input data layers

and validation sets with an 80-20 distribution.

3. MODELS

All three models were trained on a GPU via Kaggle. The U-Net and Segformer were implemented in Tensorflow, while the Naive CNN used PyTorch implementation.

3.1. Naive CNN

In the initial attempt to construct a model from scratch, a basic fully convolutional network with six layers was created, corresponding to the six input channels. These input channels are combined to form the model's input tensor, passing through the convolutional layers for binary water mask classification. The output layer involves a convolution followed by a sigmoid activation function, generating a probabilistic map indicating water presence in the input image pixels.

Fundamentally, the model comprises of six convolutional layers followed by spatial dropout operations. The convolutional layers progressively increase in channels from 32 to 128, enabling the model to capture diverse features in the input images.

The architecture utilizes rectified linear unit (ReLU) activation functions to introduce non-linearity, enhancing the model's ability to comprehend complex relationships. The final layer consists of a single neuron with a sigmoid activation function, transforming the output into probabilities between 0 and 1, representing the likelihood of the positive class (water presence). This choice aligns with the binary classification task, allowing a probabilistic interpretation of the model's predictions.

3.2. U-NET

U-NET is a convolution neural network designed to perform image segmentation, developed in 2015 by [2], it gained popularity due to its effectiveness to accurately segment images, and was the second model used for the comparison.

U-NET consists of two key components: the encoder and decoder. The Encoder's primary goal is to capture high-level features during down-sampling operations. Subsequently, the Decoder up-samples images to the original resolution via a series of up-convolutions, maintaining spatial resolution by concatenating feature maps.

A distinctive feature of U-NET is the Skip Connections, addressing the vanishing gradient problem and promoting information propagation. The final layer employs a 1x1 convolution followed by a sigmoid function, suitable for binary classification.

Weighted Adam was selected for its adaptive momentum characteristics, with embedded L2 regularization to prevent model overfitting. The chosen loss function is binary cross-entropy, quantifying the disparity between ground-truth and predicted images.

3.3. Segformer

The third and final model tested was a fine-tuned version of the SegFormer model, originally introduced by [3]. Segformer is a semantic segmentation model that relies on Transformer mechanisms designed by the viral 2017 research paper "Attention is All you Need" [4] and utilized by popular Natural Language Processing models such as Chat-GPT. The 'Attention' referenced in the aforementioned paper refers to the Transformers' ability to capture global relationships in input data, as opposed to the local relationships emphasized in CNN's via the convolutional kernel. In other words, the spatial relationship between elements of an image can be learned regardless of their proximity in the original input. The trade-off, however, is that Vision Transformer models such as SegFormer tend to be more complex and computationally intensive than CNN's. Indeed, the Segformer model used boasted 3,719,362 trainable parameters.

Thus, the justification behind the employment of a Segformer model was twofold: **1**. Assess if a model capable of learning global spatial relationships has an improved ability to identify floods, especially outside of pre-existing water bodies; and **2**. Leverage Fine-Tuning to implement a more complex model to assess if the increased complexity of SegFormer results in a 'smarter' model.

The Segformer model backbone used in this study was the nvidia/mit-B0 taken from HuggingFace, and pre-trained on ImageNet-1k inputs. As a result, this particular model only had trained weights for 3 band images (RGB). Thus, all extra bands implemented in the present study were done so with randomly initialized weights, the consequences of which will be further discussed later on. An additional limitation of this model came from the fact that outputs are restricted to 128 x 128 dimensionality, and thus required up sampling.

4. RESULTS

4.1. Metric Performances & Final Architectures

While all models underwent some level of prototyping and versioning, metrics for the best version of each model can be seen in Table 1 and Details on the final architectures used to generate these metrics are stored in Fig. 2.

The F1 scores for all models were relatively high, consistently staying above 0.80 and within 0.1 points of each other. Validation Losses were similarly strong for all three models, however the Segformer model achieves almost half of the the loss of the convolution-based models. Batch sizes varied due to differing memory constraints of the models and the effect is reflected in the difference in number of training epochs (All models utilized early stopping). The NaiveCNN converged the fastest in only 8 epochs with a batch size of 16 (although the simplicity of the model likely also contributed), while U-Net took the most epochs to converge (21) and also utilized the smallest batch size (4).

With respect to the individual models, the Segformer displayed strong 'out-of-the-box' performance (that is, using all original input bands and with- out standardization) with an F1 score of 0.80 and a validation loss of 0.07. However, upon further versioning, the model benefited from an extra normalization step as well as the de- rived bands as described in Section 2, save for DEM-WM. These changes boosted the F1 score by 3%, as shown in Table 1.

The U-Net similarly displayed strong performance when trained with little manipulation using all six original bands, achieving an F1 Score of 0.82. After many experiments with different band combinations, the best U-Net model (U-NET-019) utilized all four derived bands (including DEM-WM) and, again similar to Segformer, trained without the Land Cover layer in addition to removing the MERIT DEM layer. However, these changes only resulted in an F1 Score increase of 0.01. U-NET-019 also leverage regularization i the form of a weight decay of 0.005, β1=0.9, β2=0.99.

Finally, despite being simple and straightforward, the fully convolutional model performs better than expected on the validation set with a final F1-Score of 0.809. The model associated with this score implemented a 2d dropout with p=0.2 along with a weight decay of 0.0001. As this was the 'baseline' model, the Naive CNN also experienced the most improvement in F1 score (from a starting point of 0.752), emphasizing the fact that simple models can be effective, but more care must be taken to their implementation.

4.2. Visual Assessment of Outputs

While the performance metrics for each model were quite similar, a visual assessment of the predicted masks by each model proved to offer much more valuabele insight into how and why the models were performing the way they were. Figure 3 shows a comparison of the Water Probability layer (representing persistent water bodies), the Ground truth mask and the predicted mask for all three outputs. Notably, the Fig 3A and Fig3C represent the same location but different flooding events, represented by a difference in the Sentinel 1 SAR bands (VV, VH, and the related derive band VV/VH).

Feature	Naive CNN	U-NET	Segformer
Original Bands Used	All	VV, VH, Cop. DEM, Water Prob.	VV, VH, MERIT DEM, Cop. DEM, Water Prob.
Derived Bands Used	None	WBM, DEM-WB, Distance, VV/VH	WBM, Distance, VV/VH
Normalization	Batch Norm	Batch Norm	Standardization
Regularization	Dropout, Weight Decay	Weight Decay	None

Fig. 2. Final Model Architectures

The Naive CNN, serving as the earliest working model, highlighted from the start a common issue of models predicting only the persistent water bodies and not the floods. As such, improvement both within and across models could often visually be marked by an improved ability to predict the entire flood extent, rather than just the pre-existing water body. Figure 3A serves as a representative example at the increasing ability of the model to predict the entire flood extent with increasing F1 Score, (i.e., Naive CNN →U-NET →Segformer).

However, this trend is not an absolute, while the Segformer was more often able to capture flooding outside of water bodies, it also tended to be incredibly noisy, resulting in an overestimation of the flooded area (Fig. 3B). While there was some initial concern that this noisiness could be due to the upsampling required of the segformer outputs (See Section 3.3), plotting of the original 128x128 outputs should the same noisy behavior.

Finally, the U-Net's displayed interesting behavior when analyzed visually. There were many inconsistencies with the way the model predicted mask, where in some instances the prediction was the best out of all three (Fig 3C), in some instances it performed the same as the CNN (Fig 3A) and in some instances it completely overestimated or mispredicted the mask.

Table 1. Comparison of Model Training Metrics & Parameters

Metric	Naive CNN	U-NET	Segformer
Optimizer	Adam	Adam	Adam
Batch Size	16	4	8
Epochs	8	21	15
Validation Loss	0.135	0.112	0.067
F1 Score	0.809	0.824	0.838

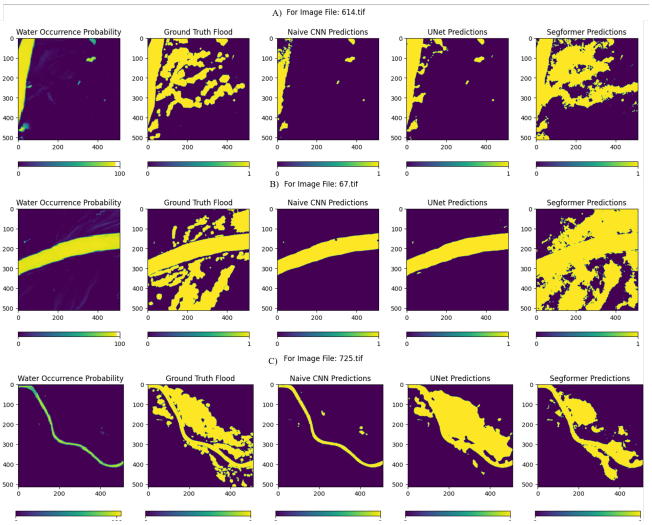

Fig. 3. Samples of inferences (last three columns) for comparative analysis

5. DISCUSSION

The results emphasize some key points that bear further discussion. The first being the emphasis of visual assessment to understand model performance. While all three models achieved fairly high F1 scores, in the case of the convolutional models (Naive CNN and U-NET) this was because the models were very easily mapping the pre-existing water bodies but struggling to capture the extra flood extent. In the context of the DFC challenge, these high F1 scores are not enough as they don't actually achieve the goal set out to capture flood damage. On the other hand, while the Segformer appeared to have the best F1 Score and did appear to successfully map floods, the visual inspection also allowed us to recognize that the model was suffering from extreme over-estimation of the flood extent.

Another interesting finding throughout the course of the study was the role of the input data. The LCVM, for example was an interesting testament to the value of domain knowledge. Both the U-Net and the Segformer found an improvement when removing this Land Cover map, and the drawback of this layer are posited to be twofold: 1. This band contains class-type data, and is thus harder to harmonize with the other inputs in steps such as normalization, 2. Because there is a strictly 'water' class, the models could over fit to those water areas and struggle to recognize the other potentially floodable areas that are not in the 'water' class. Given this, the WBM derived from the land cover class (which included the areas classified as water but also other areas we deemed 'floodable' such as wetlands) proved to be a useful repurposing of the original input data, as the best versions of both U-Net and Segformer improved when this layer was used over the Landcover.

6. CONCLUSION

This project aimed to tackle the challenge set out by the IEEE GRSS Data Fusion Contest on Rapid Flood Mapping using SAR imagery. We presented three different deep learning methods to tackle this problem; a naive CNN, a U-Net, and a fine-tuned Segformer. Overall, the models performed acceptably with an F1 Score between 0.80 and 0.85, however visual analysis showed very differing behaviors between the models. The convolutional models struggled to capture the flooded areas in addition to the pre-existing water body, while the fine-tuned Segformer had a high tendency to produce noisy outputs that overestimated the flood extent.

In general, however, this project suffered greatly from the constraints on time and resources. In an ideal world, more iterations of each model would have been run under similar constraints in order to better facilitate model comparison, but timing restricted the extent to which the models could be harmonized. Additionally, the successful implementation of HPC resources could have allowed for improvement measures such as a data augmentation. Future steps on this project would certainly center in these two aspects.

In summary, the presented models are not quite ready to be used as the cutting edge response to flood event emergency response. However as an exploratory and academic endeavor and with respect to their personal and professional developments throughout the course of this study, the authors of this paper find the outcomes to be a great success.

7. REFERENCES

[1] Intergovernmental Panel on Climate Change (IPCC), *Climate Change 2014 – Impacts, Adaptation and Vulnerability: Part A: Global and Sectoral Aspects: Working Group II Contribution to the IPCC Fifth Assessment Report*, Cambridge University Press, 2014.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[3] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12077–12090, 2021.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.