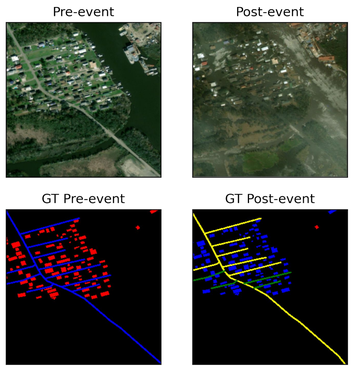
Comparing Transformers and CNNs on the SpaceNet Flood Detection Challenge

Adrian Stoll, Naijing Guo, Nenad Bozinovic
{adrs, njguo, nesa}@stanford.edu

Abstract

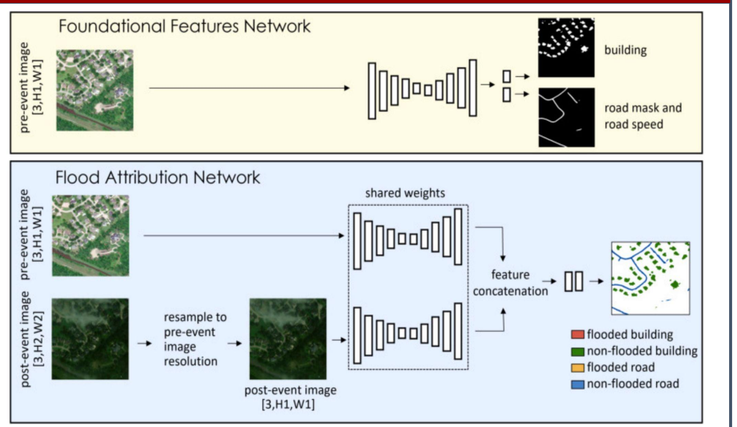
This project explores the SpaceNet8 Challenge, which aims to detect floods caused by hurricanes and heavy rains. We compared a variety of transformer and CNN segmentation architectures. We found that large pre-trained Segformer models had better performance than Resnet and U-net based models despite consuming more computational resources. The highest IoU for building detection was 61% for Segformer, which indicated that attention is better suited for detecting building-footprints than convolutions. We noticed that flooded road detection was particularly hard with highest IoU of 40%. We observed pre-training on ImageNet and Cityscapes datasets provided a moderate improvement compared to pre-training on the ADE20k dataset and a significant improvement compared to training a model from scratch.

Datasets



Examples of SpaceNet8 raw images pre- and post-event (top row) and respective ground truth segmentation masks (bottom row)[1]. Colors indicate classes (buildings, road, flooded buildings and flooded roads). In this example blue and yellow colors refer to flooded buildings and roads. For pre-event labels there is 1 building class and 8 road classes denoting speed (10 mph per class). For post event there are 4 classes (non-flooded building, flooded building, flooded road, non-flooded road). Dataset was imbalanced and favored non-flooded areas (~85% of labels). We performed a split of 679 training images and 122 (15%) validation ones (used for all experiments).

Model Architecture

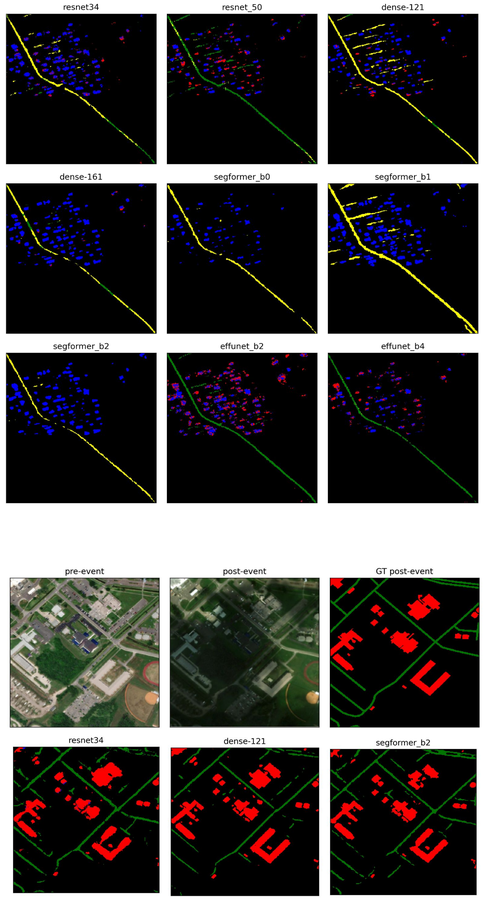


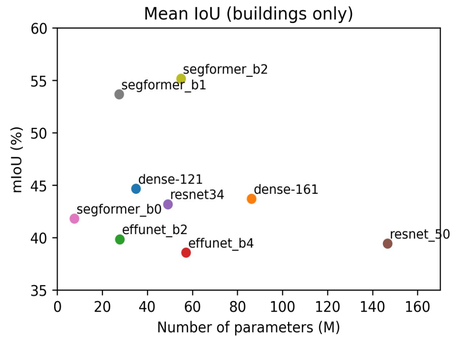
Pipeline provided by SpaceNet8 containing foundation features network and flood attribution networks. The pipeline design is modular and allows different backbone models to be swapped in while maintaining the same data-loading, training, and evaluation code.

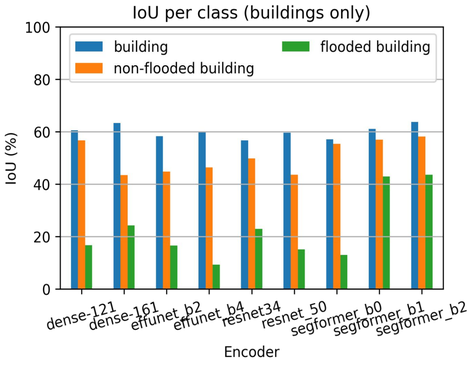
Experiments & Results

We compared several backbones for Intersection over Union (IoU) metric. Each model was trained on the entire training set for typically 10 epochs (45 min on A6000) after which we would typically see validation loss convergence and overfitting on the training dataset. Each training run used the Adam optimizer with a learning rate of 0.0001. Loss functions were binary-cross-entropy (building detection), combined dice loss and focal loss (roads) and cross-entropy (flooded vs non flooded buildings and roads).

Resnet34/50 is a baseline model. Segformer uses a hierarchical transformer encoder combining multiscale features using MLP. Unlike ViT, Segformer does not use positional encoding which aids transfer learning between datasets with different image resolutions [2]. DenseNet's main feature is that it connects each layer to every other layer in a feed-forward fashion for a total of L(L+1)/2 connections, comparing to traditional CNN's L layers [3]. EfficientNet is a CNN that uses so-called compound scaling with optimal layer width, depth, and resolution [4].



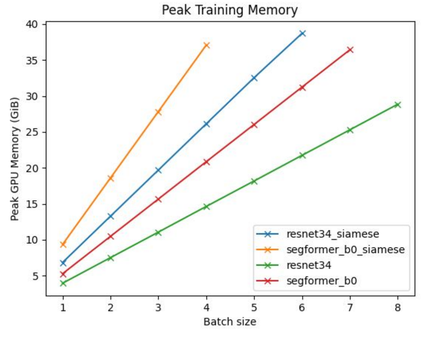


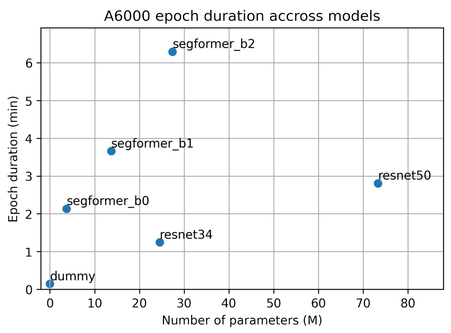


Model (Pre-training)	NF Bldg	F Bldg	NF Road	F Road
Resnet (None)	0.28	0.03	0.10	0.00
Resnet (Imagenet)	0.46	0.12	0.30	0.18
Segformer (None)	0.22	0.01	0.13	0.00
Segformer (Imagenet)	0.58	0.38	**0.43**	**0.37**
Segformer (ADE20k)	0.46	0.0	0.31	0.03
Segformer (Cityscapes)	**0.61**	**0.40**	0.41	0.34

- Overall Segformer performs better than CNN models on all metrics, showing that **attention is better suited for detecting building and road-like polygonal objects**.
- **Pre-training with ImageNet and Cityscapes datasets yields an improvement over ADE20k**, likely due to the fact that ADE20k images contain mostly indoor images of objects, unlike Spacenet8 satellite imagery. This is also likely the reason why ADE20k pretrained model did not yield any improvement on flooded objects, that are already less represented due to the imbalances dataset.
- We also did not see improvement by using **larger CNN models** (resnet 50, densenet161, efficientnet_b4) indicating that these models **are overfitting** on a relatively small SpaceNet8 dataset.

- **Segformer b0 consumes more memory and has longer epoch durations that Resnet34 despite having fewer parameters**. (Attention has quadratic time and space complexity in comparison to linear complexity for CNN models)
- The y-intercept of the memory vs batch-size best fit line is the memory from model weights, their gradients, and overhead. The slope represents the memory for training examples, activations, and gradients of activations.
- We noticed **loading data from shared network drive increased epoch time 6-fold** (~4 minutes) in comparison to loading data locally from hard drive.





Conclusion:
- Transformer models have better overall performance than CNN based models, although they are more resource demanding
- Pre-training makes a positive contribution to the performance, and models pre-trained on different dataset could have slightly different performance in specific tasks

Future Work:
- Explore other model architectures, eg. Segment Anything
- Image pre-processing, including normalization
- Pre-train model on additional dataset
- Add weights to the CrossEntropy loss function

References:
[1] Ronny Hansch et al, Spacenet 8 - the detection of flooded roads and buildings. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 1471–1479, 2022
[2] Enze Xie et al, Segformer: Simple and efficient design for semantic segmentation with transformers, 2021.
[3] Gao Huang et al. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2261–2269, 2017. 3, 4 [7]
[4] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning. PMLR, 2019.