Interpretable Convolutional Neural Networks for Building Damage Assessment in Satellite Imagery

Thomas Y. Chen[1] and Ethan Weber[2]

[1]Academy for Mathematics, Science, and Engineering, Morris Hills High School, 520 W Main St, Rockaway, NJ 07866
[2]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar St, Cambridge, MA 02139

***Abstract*— Natural disasters ravage the world's cities, valleys, and shores constantly. Having precise and efficient mechanisms for assessing infrastructure damage is essential to channel resources and minimize the loss of life. Using a dataset that includes labeled pre- and post- disaster satellite imagery, we train multiple convolutional neural networks to assess building damage on a per-building basis. We present a highly interpretable deep-learning methodology that seeks to explicitly convey the most useful information required to train an accurate classification model. Our findings include that ordinal-cross entropy loss is the most optimal loss function to train on and that including the type of disaster that caused the damage in combination with a pre- and post-disaster image best predicts the level of damage caused. Our research seeks to computationally aid in this ongoing humanitarian crisis.**

I. Introduction

Natural disasters devastate countless vulnerable communities and countries annually, killing on average 60,000 people each year worldwide [5]. The timely allocation of resources in the event of these tragedies is crucial to saving lives. The frequency and severity of these disasters will only continue to increase, exacerbated by climate change [6]. The catastrophic impact of natural disasters and their increasing prevalence motivates the problem addressed in this work. Deep neural networks (DNNs) have been used to locate and classify building damage within satellite imagery [2]. However, the current literature is limited in interpreting what exactly these neural networks are learning and identifying key predictors for assessing building damage. Thus, we present a novel analysis of the most important information that a deep learning model needs to assess building damage. We use a convolutional neural network (CNN) architecture called residual neural network (ResNet), pre-trained on ImageNet data. In our approach, we train multiple CNNs on xBD satellite imagery data [3] with different modalities of input, as well as with different loss functions, and compare accuracy on the validation set. We aim to explicitly provide insight into the most effective ways to train models to classify levels of building damage, maximizing the efficiency of the emergency response after a natural disaster, which has the potential to save lives and reduce economic strain.

II. Methods

For this work, we utilize the xBD dataset [3], which covers a wide range of disasters in fifteen countries. One of xBD's main purposes is to demonstrate changes between pre- and post-disaster satellite imagery to aid in detecting the damage caused. Therefore, each post-disaster building is labeled as one of the following: "unclassified," "no damage," "minor damage," "major damage," or "destroyed." The classification benchmark utilized is called the Joint Damage Scale (JDS) [3]. We use the xBD dataset because it incorporates a variety of disaster and building types, as well as geographical locations (for cross-region generalization), allowing for diversity in training the model. Additionally, the high-resolution imagery allows for detailed change detection between pre- and post-disaster images. These factors currently make xBD the leading dataset for building damage detection using labeled satellite imagery [3]. Using the segmentation ground truth masks (sets of coordinates constituting building polygons) provided, we extract individual building polygons to train on.

We train a baseline classification model to classify buildings by damage level, as defined by the JDS. Post-disaster images were the only model input. Notably, our baseline model does not use change detection because pre-disaster imagery is not taken as input. The model architecture is ResNet18, an 18-layer CNN [4]. This baseline model uses the cross-entropy loss function, which is defined as:

$$\sum_{c=1}^{4} y_{o,c} \log(p_{o,c}) \qquad (1)$$

where $y_{o,c}$ is a 0/1-binary indicator of whether c, as a label, correctly classifies observation o, and $p_{o,c}$ is the predicted probability that observation o is of the class c. The network is trained on 12,800 building crops with a batch size of 32. The crops are divided in a 0.8:0.2 ratio for training and validation data, respectively. The Adam optimizer, which is for adapted learning, is set at a learning rate of 0.001. The model trained for 100 epochs on NVIDIA Tesla K80 GPUs.

We train models that improve upon the performance of the baseline model. To do this, we introduce other model inputs, namely the pre-disaster image (in combination with the post-disaster image) and the type of disaster (e.g. volcano, wind) that caused the building damage. To train a model that takes in both pre-disaster images and their corresponding post-disaster images, we concatenate the RGB channels of the two and use that as input. To train a model that takes in the pre-disaster image, post-disaster image, and disaster type, we do the same, but also concatenate a one-hot encoded representation of the disaster type in one of the later layers of the CNN. Furthermore, we experiment with other loss functions, namely mean squared error loss (MSE) and ordinal cross-entropy loss to train these models. We define mean squared error as:

$$\frac{1}{b}\sum_{i=1}^{b}(y-\hat{y})^2 \qquad (2)$$

where b is the batch size, y is the ground truth (a class from 0 to 3 representing each damage level), and $\hat{y}$ is the prediction. Ordinal cross-entropy loss differs from cross-entropy loss in that it considers the distance between the ground truth and the predicted class (hence "ordinal"). This function is useful because the building damage classification problem involves different and increasing levels of damage from “no damage” to “destruction.” To implement ordinal cross-entropy loss as the loss function, we treat it as generic multi-class classification and encode the classes “no damage,” “minor damage,” “major damage,” and “destroyed” as [0, 0, 0], [1, 0, 0], [1, 1, 0], and [1, 1, 1], respectively [1].

III. Results and Discussion

In Table 1, we present model accuracy on the validation set across nine different models, which are differentiated by three different input combinations and three loss functions. The baseline model, which is trained with post-disaster data only and the cross-entropy loss function, has an accuracy of 59.5%, as shown, while the most accurate model has an accuracy of 74.6%. It is important to note that all models were trained and validated on data that is evenly split between building crops of each class (no damage, minor damage, major damage, and destroyed), so a purely blindly guessing model would achieve approximately 25% accuracy.

TABLE 1. COMPARISON OF VALIDATION ACCURACY ON 9 DIFFERENT MODELS

Model Input(s)	**Model Accuracy on Validation Set**		
	Mean Squared Error	***Cross-Entropy Loss***	***Ordinal Cross-Entropy Loss***
Post-Disaster Image	45.3%	59.5%	64.2%
Post-Disaster Image + Pre-Disaster Image	50.2%	68.3%	71.2%
Post-Disaster Image + Pre-Disaster Image + Disaster Type	49.7%	72.7%	74.6%

Percentages represent model’s validation accuracy given loss function and model input.

Much of our results confirm our hypotheses. Accuracy on the validation set improves when more modes of useful information are inputted into the model (accuracy generally increases moving down the rows of Table 1). This is justifiable given the intuitive assumption that the more information the model has to work with, the more accurate predictions it should make. A large part of our research addressed which types of input aid the convolutional neural networks in making accurate predictions. From the generated results, it seems that having the aspect of change detection (when the pre-disaster image is concatenated with the post-disaster image and inputted) is useful, along with the type of disaster. We also note that models using ordinal cross-entropy loss as their criterion for optimization perform the most accurately. As previously mentioned, ordinal cross-entropy loss is most specifically applicable for a classification problem that involves an ordinal scale (in this case, the JDS), as opposed to categories with no intrinsic ordering. MSE, not surprisingly, showed itself to be the least effective loss function to use for training. This result is justifiable because MSE is primarily used in regression problems, not classification problems. We find that cross-entropy loss models fall somewhere in between.

However, we note that none of the accuracy numbers are necessarily optimal. This can be explained by the fact that the differences between categories, particularly between minor-damage and major-damage, are largely difficult to discern for both humans and computers. This is a challenge that comes with non-binary classification tasks with building damage, and it has been acknowledged by many, including [3]. In addition, there is some noisy data in the dataset and cleaning it more thoroughly would most likely yield marginally more accurate predictions.

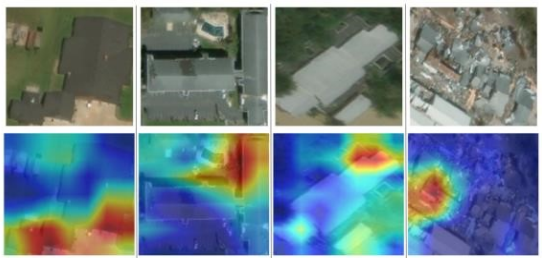

Figure 1: Gradient class activation maps depict which parts of the building crop lead the baseline model to predict a certain classification. On the top are the original images (crops) and on the bottom are the corresponding gradient class activation maps. The images included are only post-disaster images.

IV. Conclusion

The main insights that can be drawn from our work include using individualized building crops instead of semantic segmentation to train models and performing experiments with various combinations of model inputs and loss functions to explicitly examine their differences. Practically, our work (a novel, more interpretable approach) and others in the field advance methods for more robust emergency responses and more efficient allocation of resources, which saves lives.

Acknowledgment

I would like to thank my mentor, Mr. Ethan Weber, without whom this research would not have been possible.

References

[1] Cheng, Jianlin, Zheng Wang, and Gianluca Pollastri. "A neural network approach to ordinal regression." *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. IEEE, 2008.

[2] Gueguen, Lionel, and Raffay Hamid. "Large-scale damage detection using satellite imagery." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[3] Gupta, Ritwik, et al. "Creating xBD: A dataset for assessing building damage from satellite imagery." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2019.

[4] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[5] Ritchie, Hannah, and Max Roser. "Natural disasters." *Our World in Data* (2014).

[6] Van Aalst, Maarten K. "The impacts of climate change on the risk of natural disasters." *Disasters* 30.1 (2006): 5-18.