SPATIOTEMPORAL CONTRASTIVE REPRESENTATION LEARNING FOR BUILDING DAMAGE CLASSIFICATION

Bo Peng[1], *Qunying Huang*[1], *Jinmeng Rao*[2]

[1]Spatial Computing and Data Mining Lab, University of Wisconsin - Madison, WI, 53706, USA
[2]Geospatial Data Science Lab, University of Wisconsin - Madison, WI, 53706, USA

ABSTRACT

Automatic building damage assessment after natural disasters is important for emergency response. While existing supervised deep learning models achieved good performance on building damage classification, these models require massive human labels for training. Additionally, pre-trained models often fail to generalize well to new disaster events due to gaps between domains associated with training and testing data. In response, this study proposes a novel spatiotemporal contrastive representation learning model for learning features of building damages with big unlabeled data. Experimental results demonstrate superior performance of such features on classifying building damages resulting from various natural disasters (e.g., hurricanes, floods, wildfires, earthquakes, etc.) across different geographic locations worldwide, compared with the state-of-the-art supervised methods.

Index Terms— Spatiotemporal, contrastive, representation learning, building damage, natural disasters

1. INTRODUCTION

Natural disasters (e.g., floods and hurricanes) have posed a major threat to human lives and caused huge economic losses. Real-time building damage assessment during or after disasters is critical for quick delivery of accurate rescue and relief efforts and mitigation of economic losses [1]. With an increasing volume of labeled remote sensing (RS) data and advancements in artificial intelligence (AI), many deep learning and computer vision based methods have been developed for post-disaster building damage assessment, primarily by data-driven supervised learning to discover the underlying patterns of damaged buildings [1, 2].

Unfortunately, these methods are not applicable to near real-time disaster response due to the poor model generalizability and time-consuming human labeling of new training data [3]. Additionally, traditional machine learning and image processing models mainly rely on hand-crafted image features (e.g., textures, edges, and corners [4]), and therefore are case-by-case efforts with a lack of model generalizability for upcoming disasters.

This work was supported by the National Science Foundation under Grant 1940091, the Microsoft AI for Earth Grant, and Villas Associate Award, at the University of Wisconsin - Madison.

With the need for eliminating massive human labels for model training, self-supervised learning (SSL) has emerged as a new solution with its great potential of image representation learning with unlabeled data [3]. Among multiple SSL frameworks, contrastive learning (CL) of image features has shown promising results in recent studies [5, 6]. However, the crucial issues with CL remain unsolved, including intensive manual data augmentations, specialized neural network architectures, the image memory bank, etc.

This study proposes a novel spatiotemporal contrastive representation learning (ST-CRL) model for RS image representation learning. The ST-CRL model learns image features with unlabeled data by incorporating the spatial and temporal information of geospatial data. It is assumed that the temporally adjacent pair of RS images over the same geographic extent should have similar features while geographically distant pairs should have different features. Main contributions of this work include: (1) The ST-CRL model learns image features without massive human labels, contributing to the near real-time and automatic data processing in disaster response. (2) The ST-CRL model incorporates the natural spatiotemporal information to construct the input image samples for contrastive training without the need for intensive manual data augmentations and the image memory bank that are required in conventional CL. (3) The self-supervised ST-CRL model can be trained with new unlabeled data and thus provides a strong generalizability for different environments.

2. METHODOLOGY

2.1. Datasets

We created xBD-obj, a new dataset of building objects based on the xBD dataset [1] which consists of a large number of pre- and post-disaster image pairs and building footprints. Each post-disaster building is labeled as one of the four classes (i.e., no damage, minor damage, major damage, and destroyed). With xBD building footprints, we cropped image patches centered at each building of interest with a buffer of around 15 meters (Fig. 1). Image patches are used as building

(a) (b) (c) (d)

Fig. 1. Building object images. (a) and (b): pre- and post-hurricane. (c) and (d): pre- and post-flooding.

objects for object-based building damage classification. We use building objects from the xBD *train* and *tier3* subsets for model training and validation, the xBD *test* and *hold* subsets for evaluation.

2.2. Spatiotemporal Knowledge

The First Law of Geography [7] informs us that geographically near RS images should have similar features while distant RS images should have different features. In the global-scale RS dataset xBD-obj, it is easy to find distant pairs of RS images from two different areas. However, a good strategy is needed for the selection of neighboring image pairs for heterogeneous urban areas. Multiple hyperparameters (e.g., the image size and the distance of neighboring images) require further fine-tuning to fit the heterogeneous urban environment.

To address this issue, we leverage the assumption that temporally adjacent image pairs over the same geographic extent should have similar features despite their different acquisition conditions (e.g., weather, illumination, sensor viewing angles, etc.). Although disasters may result in changes between the corresponding bi-temporal RS images, the vast majority of corresponding image pairs in a large-scale RS dataset do not exhibit a major difference. Additionally, such a difference can be considered as a strong **natural** data augmentation, consistent with the strong **manual** data augmentation in contrastive learning to enhance the learning of major image patterns and to avoid the model collapse [5].

2.3. ST-CRL

To incorporate the spatiotemporal knowledge into contrastive learning, we use the pre- and post-disaster temporally adjacent building object image pairs as positive pairs and the geographically distant pairs as negative pairs inspired by the recent contrastive learning framework [5] (see Fig. 2).

Let $b_i (i = 1, 2, \ldots, N)$ denotes the building object image, where N is the total number of buildings. Given the pre- and post-disaster building object image pair denoted as $(b_i^{\text{pre}}, b_i^{\text{post}})$, ST-CRL learns building object features h_i by maximizing the similarity between positive pairs $(b^{\text{pre}}, b^{\text{post}})$ and the dissimilarity between negative pairs including $(b^{\text{pre}}, b^{\text{pre}})$, $(b^{\text{pre}}, b^{\text{post}})$, $(b^{\text{post}}, b^{\text{pre}})$, and $(b^{\text{post}}, b^{\text{post}})$. It is assumed that, for the global large-scale RS dataset xBD-obj, there is a very high probability that a randomly selected pair of different building objects constructs a negative pair.

Fig. 2. The pre- and post-disaster building object images $(b^{\text{pre}}, b^{\text{post}})$ are geographically distant from $(b^{\text{pre}}, b^{\text{post}})$.

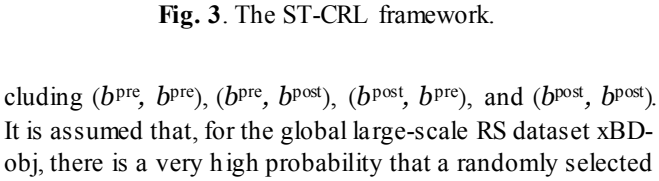

Fig. 3. The ST-CRL framework.

As illustrated in Fig. 3, ST-CRL first encodes building object images, including the positive pair $(b_i^{\text{pre}}, b_i^{\text{post}})$ for the building b_i and the negative pairs between $b^{\text{pre/post}}$ and $b^{\text{pre/post}}$, into low-dimensional features h_i via the same encoding module $f_\theta(\cdot)$. In the paper, we use a variant of the ResNet architecture [8] without the last output layer for $f_\theta(\cdot)$. Then a multi-layer perceptron (MLP) $p_\gamma(\cdot)$ with one hidden layer maps building features h_i to another space z_i for computing the contrastive loss. Finally, the contrastive loss function encourages the similarity between positive building pairs while discourages the similarity between negative pairs. Similar to [5], we only keep the pre-trained image encoding module $f_\theta(\cdot)$ for image feature extraction and classification.

During training, we randomly sample a batch of N pairs of pre- and post-disaster building object images $(b^{\text{pre}}, b^{\text{post}})$, $i = 1, 2, \ldots, N$. For the building b_i within this batch, we compute the cosine similarity between the only positive pair $(b_i^{\text{pre}}, b_i^{\text{post}})$, and the cosine similarity between multiple negative pairs $(b_i^{\text{pre}}, b_j^{\text{pre}})$, $(b_i^{\text{pre}}, b_j^{\text{post}})$, $(b_i^{\text{post}}, b_j^{\text{pre}})$, and $(b_i^{\text{post}}, b_j^{\text{post}})$, $j = 1, 2, \ldots, N$ and $j \neq i$. The contrastive loss for the building b_i is defined similar to that in [5] as

$$l_i = -\log \frac{\exp(\text{sim}(z_i^{\text{pre}}, z_i^{\text{post}})/\tau)}{} \quad (1)$$

$$B = \sum_{i,j} \Big[\exp(\mathrm{sim}(z_i^{\mathrm{pre}}, z_j^{\mathrm{pre}})/\tau) + \exp(\mathrm{sim}(z_i^{\mathrm{pre}}, z_j^{\mathrm{post}})/\tau) + \exp(\mathrm{sim}(z_i^{\mathrm{post}}, z_j^{\mathrm{pre}})/\tau) + \exp(\mathrm{sim}(z_i^{\mathrm{post}}, z_j^{\mathrm{post}})/\tau) \Big] \quad (2)$$

where $\mathrm{sim}(\cdot)$ is the cosine similarity, τ is the temperature parameter. The final loss for this batch is given by $(1/N) \sum_{i=1}^{N} l_i$.

3. EXPERIMENTS AND RESULTS

Good image features would not require massive labels for image classification. To evaluate the quality of self-supervised SR-CRL features, we train an image classifier on top of image features encoded by the pre-trained $f_\theta(\cdot)$. The logistic regression (LR) and MLP classifiers are chosen. We also fine-tune the encoding module $f_\theta(\cdot)$ when training the classifiers. For comparative analysis, we train a variant of the ResNet in a completely supervised manner for building damage classification.

We train all classifiers with different sizes of the training data. The best model is used for evaluating the classification performance on the testing data. The precision, recall, and F1 score are evaluated for all building damage classes including no damage (no), minor damage (mi), major damage (ma), and destroyed (de). The overall F1 score is computed as

$$F1 = \frac{4}{\frac{1}{F1_{\mathrm{no}}} + \frac{1}{F1_{\mathrm{mi}}} + \frac{1}{F1_{\mathrm{ma}}} + \frac{1}{F1_{\mathrm{de}}}} \quad (3)$$

For each size, we train the classifiers with 10 trials of randomly sampled training subsets, in which all damage classes have the same class frequency to avoid the impact of class imbalance. Fig. 4 shows the learning curves of the mean overall F1 score and the standard deviation with respect to the training subset size. It is worth noting that the pre-trained $f_\theta(\cdot)$ learned in ST-CRL with the LR classifier has the same architecture as the fully supervised model. As such, Fig. 4 demonstrates that self-supervised ST-CRL features significantly boosted the performance of building damage classification.

Regarding the classification performance for different damages, we report the F1 scores with the mean and standard deviation for all damage classes corresponding to 10 trials of randomly sampled training subsets of size $n = 1000$ and $n = 9000$ in Table 1. We observe that it is more challenging to identify buildings with minor and major damages compared to those destroyed or with no damage. It is also worth noting that self-supervised ST-CRL image features provide better representations of different building damages compared to the fully supervised model. Fig. 5 visualizes some examples of building damages predicted by a LR classifier on self-supervised ST-CRL features and 10,000 labels.

Fig. 4. LR and MLP trained with self-supervised ST-CRL features outperform the fully supervised model.

4. CONCLUSION

In this study, we propose a novel self-supervised learning framework ST-CRL that learns RS image features by incorporating the spatiotemporal knowledge. ST-CRL offers the capability of learning RS image features without human labels. Experimental results show that self-supervised ST-CRL features significantly boost the performance of building damage classification compared to supervised classifiers trained from scratch. Since deep neural networks for image classification (e.g., ResNet) require massive human labels for image feature learning, the proposed ST-CRL paves the way to learn informative image features in the geospatial domain with unlabeled geospatial big data. In the future, we plan to investigate the spatial statistics of RS images for an improved construction of image samples for ST-CRL such that the self-supervised ST-CRL image features further inform building damage classification.

5. REFERENCES

[1] Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston, "xBD: A Dataset for Assessing Building Damage from Satellite Imagery," *arXiv preprint arXiv:1911.09296*, nov 2019.

[2] Ethan Weber and Hassan Kané, "Building Disaster Damage Assessment in Satellite Imagery with Multi-Temporal Fusion," in *International Conference on Learning Representations (ICLR) AI For Earth Sciences Workshop*, apr 2020.

[3] B Peng, Q Huang, J Vongkusolkit, S Gao, D B Wright, Z N Fang, and Y Qiang, "Urban Flood Mapping with Bi-temporal Multispectral Imagery via a Self-supervised Learning Framework," *IEEE Journal of Selected Topics*

Table 1. F1 scores for each type of building damages

Damages	n=1000			n=9000		
	ST-CRL (MLP)	ST-CRL (LR)	Supervised	ST-CRL (MLP)	ST-CRL (LR)	Supervised
No	0.830 ± 0.010	0.833 ± 0.016	0.553±0.068	0.858 ± 0.009	0.857 ± 0.003	0.717±0.024
Minor	0.506 ± 0.021	0.509 ± 0.008	0.190±0.031	0.518 ± 0.009	0.514 ± 0.009	0.334±0.017
Major	0.436 ± 0.018	0.441 ± 0.017	0.250±0.017	0.518 ± 0.009	0.512 ± 0.013	0.361±0.018
Destroyed	0.649 ± 0.023	0.635 ± 0.025	0.332±0.031	0.713 ± 0.019	0.720 ± 0.014	0.508±0.024

(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 5. Classification of building damages caused by hurricanes (first row) and floods (second row). Left two columns: pre- and post-disaster images. Right two columns: labels and predictions of building damages. (Green: no damage, Blue: minor damage, Orange: major damage, Red: destroyed)

in Applied Earth Observations and Remote Sensing, p. 1, 2020.

[4] Xue Wang and Peijun Li, "Extraction of urban building damage using spectral, height and corner information from VHR satellite images and airborne LiDAR data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 159, pp. 322–336, jan 2020.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in *Proceedings of the 37th International Conference on Machine Learning*, Hal Daume´ III and Aarti Singh, Eds., Virtual, 2020, vol. 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607, PMLR.

[6] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[7] Waldo R Tobler, "A computer movie simulating urban growth in the Detroit region," *Economic geography*, vol. 46, no. sup1, pp. 234–240, 1970.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.