IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-0369-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IGARSS47720.2021.9553475

GENERALIZED SCALABLE NEIGHBORHOOD COMPONENT ANALYSIS FOR SINGLE AND MULTI-LABEL REMOTE SENSING IMAGE CHARACTERIZATION

Jian Kang[1], *Ruben Fernandez-Beltran*[2], *Antonio Plaza*[3]

1. School of Electronic and Information Engineering, Soochow University, Suzhou 215006, China
2. Institute of New Imaging Technologies, University Jaume I, E-12071 Castellón, Spain
3. Hyperspectral Computing Laboratory, University of Extremadura, E-10003 Cáceres, Spain

ABSTRACT

Deep metric learning has recently become a prominent technology for the semantic understanding of remote sensing (RS) scenes due to its great potential for characterizing visual semantics. However, state-of-the-art deep metric learning models are often constrained in RS by the use of single-label annotations, which eventually reduce their capacity to characterize complex aerial scenes. Additionally, many of the existing works are specialized in particular RS applications which constrains the study of their associated metric spaces from a multi-task perspective. In this paper, we propose a new unified deep metric learning approach for both single- and multi-label RS scene characterization while also taking into account different downstream RS applications. Specifically, we extend the Scalable Neighborhood Component Analysis (SNCA) to the multi-label case and propose its generalized version, i.e., GSNCA. Extensive experiments on single- and multi-label RS benchmark datasets have been conducted to evaluate the effectiveness of the proposed method for RS image classification, clustering and retrieval.

Index Terms— Deep metric learning, Neighbor embedding, Single- Multi-labels, Scene categorization

1. INTRODUCTION

Remote sensing (RS) scene images have been widely applied for numerous tasks, such as urban mapping, object detection, and scene retrieval [1–5]. How to sufficiently interpret the semantics of RS scenes is always an ongoing topic within the community. Recently, with the public availability of large-scale RS datasets, deep learning techniques have significantly facilitated the development of effective algorithms on semantic modeling of RS scenes. Most of the proposed methods can be categorized into two types: 1) those for characterizing single-label RS scenes [6–9]; and 2) others for RS scene characterization with multiple annotations [10, 11]. Normally, the approaches designed for encoding RS scenes with single labels cannot be easily adopted for the multi-label case and vice versa. For example, the well-known triplet loss [12] utilized for deep feature extraction based on modeling the semantic-similar relationship among image triplets, while the image triplets cannot be easily constructed through multi-labels. In this paper, we propose a unified deep metric learning framework for characterizing single-label and multi-label RS scene images based on the proposed generalized scalable neighborhood component analysis (GSNCA). Inspired by the scalable neighborhood component analysis (SNCA) [13], we first investigate its loss function in the single-label case and extend it for modeling the semantic-similarities among RS scene images with multiple annotations. The proposed framework includes two main parts: 1) a CNN architecture for learning the image features; and 2) the proposed GSNCA loss function for both single-label and multi-label images. A graphical illustration of the proposed framework is shown in Figure 1. This work extends our previous research presented in [14] in order to provide an unified single- and multi-label RS image characterization scheme, while taking into account different downstream applications.

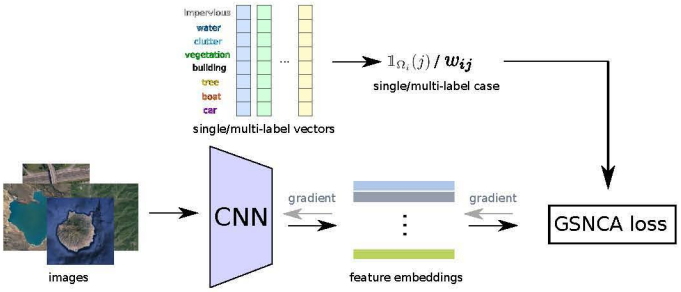

Fig. 1: The proposed unified RS scene characterization framework based on GSNCA loss.

2. METHODOLOGY

Let $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ be a set of N RS scenes and $\mathcal{Y} = \{\mathbf{y}_1, \cdots, \mathbf{y}_N\}$ their corresponding labels represented by single/multi-label hot encoding vectors. In the single-label case, the $\mathbf{y}_i$ label vector is represented by the single-label one-hot vector, i.e., $\mathbf{y}_i \in \{0, 1\}^C$, where C is the number of

classes. In the multi-label case, $\mathbf{y}_i$ is denoted by a multi-class hot encoding vector, i.e., $\mathbf{y}_i \in \{-1, 1\}^C$. $\mathcal{F}(\cdot;\theta)$ represents a particular backbone CNN model (with the parameter set θ), which performs a nonlinear mapping between the original RS image $\mathbf{x}_i$ and its corresponding feature embedding $\mathbf{f}_i \in \mathbb{R}^D$ on the unit sphere, i.e., $\|\mathbf{f}_i\|_2 = 1$. From $\mathcal{X}$, we assume that a subset $\mathcal{T}$ is extracted for training purposes.

2.1. Scalable Neighborhood Component Analysis

SNCA [13] was presented to uncover CNN-based characterizations where the semantic relationships among the input images can be preserved in the embedding space. That is, semantically similar images are projected onto nearby locations in the corresponding metric space, whereas dissimilar images are separated. From a training set $\mathcal{T}$, the similarity s_{ij} between two images $(\mathbf{x}_i, \mathbf{x}_j)$ can be computed by means of the cosine similarity function as follows:

$$s_{ij} = \mathbf{f}_i^T \mathbf{f}_j, \tag{1}$$

where $s_{ij} \in [-1, 1]$ and larger values indicate a higher similarity. Under SNCA assumptions, the probability p_{ij} that, given the $\mathbf{x}_i$ image, $\mathbf{x}_j$ is projected onto $\mathbf{x}_i$ neighborhood can be defined as:

$$p_{ij} = \frac{\exp(s_{ij}/\sigma)}{\sum_{k \neq i} \exp(s_{ik}/\sigma)}, \quad p_{ii} = 0, \tag{2}$$

being σ a temperature parameter that controls the sample distribution concentration. Note that $\mathbf{x}_j$ can be chosen as $\mathbf{x}_i$ neighbor when s_{ij} takes larger values, in contrast to another $\mathbf{x}_k$ image. Besides, the $p_{ii} = 0$ condition forces that images cannot select themselves as neighbors. Accordingly, the correct classification probability of $\mathbf{x}_i$ can be represented by:

$$p_i = \sum_{j \in \Omega_i} p_{ij}, \tag{3}$$

where $\Omega_i = \{j | \mathbf{y}_i = \mathbf{y}_j\}$ indicates those training samples that belong to the $\mathbf{x}_i$ class. Essentially, p_i is the probability that $\mathbf{x}_i$ is correctly classified and it increases with the number of $\mathbf{x}_j$ neighbouring images that share the same class. To achieve this goal, SNCA minimizes the expected negative log-likelihood over $\mathcal{T}$ as:

$$L_{\text{SNCA}} = -\frac{1}{|\mathcal{T}|} \sum_i \log(p_i), \tag{4}$$

with $|\mathcal{T}|$ being the number of training samples. In order to optimize Equation (4), the similarities between $\mathbf{x}_i$ and other images in the dataset should be calculated. For stochastically training a CNN model via L_{SNCA}, an off-line memory bank $\mathcal{B}$ is constructed and updated to store the normalized features of $\mathcal{T}$ during training. From a practical perspective, SNCA pursues to learn each image nearest neighbors in the metric space using a supervised scheme. Precisely, this fact allows SNCA to uncover inherent image semantic relationships, especially with high intra-class variations, as it is often the case in RS. Nonetheless, the SNCA loss (Equation (4)) is limited to single-label annotations, and this feature strongly constrains the capacity of the model to effectively characterize complex RS scenes for classification, clustering and retrieval.

2.2. Generalized Scalable Neighborhood Component Analysis

For extending SNCA to the use of multiple RS image annotations, we initially need to reformulate the aforementioned p_i probability as follows:

$$p_i = \sum_j \mathbb{1}_{\Omega_i}(j) p_{ij}, \tag{5}$$

where the $\mathbb{1}_{\Omega_i}(j)$ function is given by:

$$\mathbb{1}_{\Omega_i}(j) := \begin{cases} 1 & \text{if } j \in \Omega_i, \\ 0 & \text{if } j \notin \Omega_i. \end{cases} \tag{6}$$

Note that, given the index set (Ω_i), the indicator function ($\mathbb{1}_{\Omega_i}(j)$) controls the images that can be located as neighbors of $\mathbf{x}_i$ in the metric space. Then, p_i can be computed as a weighted sum of p_{ij} probabilities over the training set, where the final class decision for $\mathbf{x}_i$ depends on those RS scenes that belong to the same semantic category. Following these intuitions, we define the probability that $\mathbf{x}_i$ can be correctly classified within a multi-label scheme as follows:

$$p_i = \sum_j w_{ij} p_{ij}, \tag{7}$$

where w_{ij} represents the weight of p_{ij}. From $\mathbf{x}_i$ and its corresponding multi-label annotations, our objective is based on pulling in RS images that share more common labels and pushing away other scenes. To meet this goal, we propose using the following expression to balance w_{ij} weights in Equation (7) according to the number of common labels:

$$w_{ij} = \frac{\langle \mathbf{y}_i, \mathbf{y}_j \rangle + C}{2C}, \quad w_{ij} \in [0, 1]. \tag{8}$$

As it is possible to observe, w_{ij} considers the inner product between $\mathbf{y}_i$ and $\mathbf{y}_j$ and this relation generates that heavier weights will be assigned to the s_{ij} similarity term when $\mathbf{y}_i$ and $\mathbf{y}_j$ become more similar. Note that we also introduce in Equation (8) a normalization factor to adjust the value range of $\langle \mathbf{y}_i, \mathbf{y}_j \rangle$ between 0 and 1. With all these considerations, our overall objective function can be formulated as:

$$L_{\text{GSNCA}} = -\frac{1}{|\mathcal{T}|} \sum_i \log(p_i) = -\frac{1}{|\mathcal{T}|} \sum_i \log(\sum_j w_{ij} p_{ij}). \tag{9}$$

Table 1: K-NN classification accuracies (%) obtained by using different learning methods on single-label datasets, for $K = 10$.

	AID	NWPU-RESISC45
D-CNN	93.75	91.48
Triplet	93.25	91.43
GSNCA	94.60	92.14

When compared to L_{SNCA}, w_{ij} can be replaced by $\mathbb{1}_{\Omega_i}(j)$ in the single-label case, and the proposed L_{GSNCA} leads to the original L_{SNCA}. In other words, the proposed GSNCA method provides a general framework that contains SNCA while allowing the use of single- and multi-label annotations.

3. EXPERIMENTS

In this paper, three RS benchmark datasets, including UCM, AID and NWPU-RESISC45 [15–17], are utilized to validate the performance of the proposed method. More specifically, we use the single-label annotations of AID and NWPU-RESISC45 and the multi-label versions of UCM and AID [18, 19]. In the single-label case, we consider two different downstream applications for evaluating the effectiveness of the proposed method: 1) RS scene classification based on the K-NN classifier; and 2) RS image clustering. In the multi-label case, we conduct experiments including: 1) K-NN RS image classification; and 3) RS image retrieval. For these experiments, we randomly select 70% of the data for training, 10% for validation, and 20% for testing purposes. The clustering task is performed over the test feature embeddings generated by the learned CNN model. In the case of the image retrieval task, training and test sets serve as archive and query samples, respectively. The proposed method is implemented in PyTorch. For the sake of simplicity, we use ResNet18 [20] as the backbone CNN architecture for all the experiments. Nonetheless, other CNN architectures (e.g., ResNet50) could also be applied to the proposed approach. All the images are resized to 256×256 pixels. Besides, three data augmentation mechanisms are adopted during training: 1) *RandomGrayscale*, 2) *ColorJitter*, and 3) *RandomHorizontalFlip*. Regarding the parameter selection, we set D and σ to 128 and 0.1, respectively. The Stochastic Gradient Descent (SGD) optimizer is used for training, with an initial learning rate of 0.01 together with a 30-epoch decay. Finally, the batch size is set to 256 and the model is trained for 100 epochs. Regarding the experimental comparison, we compare the proposed approach to three different state-of-the-art deep metric learning methods: 1) D-CNN [6] (single-label case); 2) deep metric learning based on triplet loss (single-label case); and 3) BCE loss (multi-label case).

We report the overall accuracy of all the methods on the considered single/multi-label test sets in Table 1 and Table 2, respectively. Compared with the other methods, the proposed method can achieve the best classification performance on the considered benchmark archives. For example, in the single-label case, GSNCA can improve the K-NN accuracy in about 1% more than D-CNN and Triplet. The proposed approach allows finding image similarities beyond the current mini-batch which eventually enhances the model generalization capability along the training process. In contrast, D-CNN samples the required negative and positive image pairs from each mini-batch, leading to an under-complete model optimization, especially with highly complex RS data. In the case of the triplet loss, we can also find additional problems, since building sufficient similarity and dissimilarity image triplets for training may be impossible in scalable RS datasets. Note that data complexity and volume are both important factors in RS and building a representative training set with about $\mathcal{O}(|\mathcal{T}|^3)$ may easily become unaffordable. Precisely, these limitations on the contrastive and triplet loss schemes may lead to the fact that some complex RS scenes may not be well separated in the resulting metric space. Table 3 presents the corresponding NMI scores which are obtained by applying the K-means clustering algorithm to the feature embeddings of the test sets. As it is possible to observe, GSNCA provides the best matching between ground-truth labels and clusters, being such improvement higher than 3% with respect to D-CNN and Triplet. To better analyze the generated feature em-

Table 2: K-NN classification micro F1 scores (%) obtained by using different learning methods on multi-label datasets, for $K = 10$.

	UCM	AID
BCE	87.76	88.31
GSNCA	88.47	89.13

Table 3: NMI scores of the feature embeddings of the test sets produced by different learning methods.

	AID	NWPU-RESISC45
D-CNN	88.83	85.30
Triplet	89.87	88.14
GSNCA	92.96	90.20

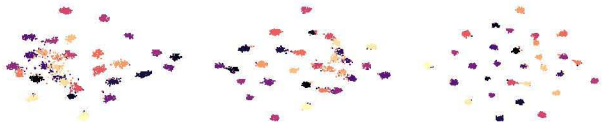

Fig. 2: 2D projection of the feature embeddings on the AID test set using t-SNE: (a) D-CNN; (b) Triplet; and (c) GSNCA.

Table 4: Image retrieval performances, evaluated with MAP (%), by the BCE and GSNCA on the test sets.

	UCM	AID
BCE	97.70	97.36
GSNCA	99.64	99.67

beddings, we use the t-distributed stochastic neighbor embedding (t-SNE) to show their corresponding projections in a 2-D plane. As shown in Figure 2, the proposed approach provides the most compact intra-class features and the most isolated inter-class features. Finally, Table 4 presents the quantitative retrieval results obtained by BCE and GSNCA. The proposed method can achieve higher retrieval accuracy when compared with BCE. This fact indicates that GSNCA is able to learn a metric space that allows for the retrieval of more semantically relevant images.

4. CONCLUSION

In this paper, we propose a novel deep metric learning method based on GSNCA for characterizing both single- and multi-label RS scene images. The proposed approach pursues to pull the most semantically similar RS images closer in the metric space when they share more classes in common, from a multi-label perspective. Extensive experiments on single- and multi-label RS benchmark datasets have been conducted to evaluate the effectiveness of our method, which outperforms other state-of-the-art approaches.

5. REFERENCES

[1] X. X. Zhu, J. Hu, C. Qiu, Y. Shi, J. Kang, L. Mou, H. Bagheri, M. Häberle, Y. Hua, R. Huang *et al.*, "So2sat lcz42: A benchmark dataset for global local climate zones classification," *arXiv preprint arXiv:1912.12171*, 2019.

[2] Y. Li, J. Ma, and Y. Zhang, "Image retrieval from remote sensing big data: A survey," *Information Fusion*, 2020.

[3] J. Kang, M. Körner, Y. Wang, H. Taubenböck, and X. X. Zhu, "Building instance classification using street view images," *ISPRS journal of photogrammetry and remote sensing*, vol. 145, pp. 44–59, 2018.

[4] R. Fernandez-Beltran, B. Demir, F. Pla, and A. Plaza, "Unsupervised remote sensing image retrieval using probabilistic latent semantic hashing," *IEEE Geosci. Remote Sens. Lett.*, 2020, dOI:10.1109/LGRS.2020.2969491.

[5] J. Kang, R. Fernandez-Beltran, Z. Ye, X. Tong, and A. Plaza, "Deep hashing based on class-discriminated neighborhood embedding," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 5998–6007, 2020.

[6] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, 2018.

[7] J. Kang, R. Fernandez-Beltran, Z. Ye, X. Tong, P. Ghamisi, and A. Plaza, "Deep metric learning based on scalable neighborhood components for remote sensing scene characterization," *IEEE Trans. Geosci. Remote Sens.*, 2020.

[8] X.-Y. Tong, G.-S. Xia, F. Hu, Y. Zhong, M. Datcu, and L. Zhang, "Exploiting deep features for remote sensing image retrieval: A systematic investigation," *IEEE Transactions on Big Data*, 2019.

[9] J. Kang, R. Fernandez-Beltran, Z. Ye, X. Tong, P. Ghamisi, and A. Plaza, "High-rankness regularized semi-supervised deep metric learning for remote sensing imagery," *Remote Sensing*, vol. 12, no. 16, p. 2603, 2020.

[10] Z. Shao, W. Zhou, X. Deng, M. Zhang, and Q. Cheng, "Multilabel remote sensing image retrieval based on fully convolutional network," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 318–328, 2020.

[11] J. Kang, R. Fernandez-Beltran, D. Hong, J. Chanussot, and A. Plaza, "Graph relation network: Modeling relations between scenes for multilabel remote-sensing image classification and retrieval," *IEEE Trans. Geosci. Remote Sens.*, 2020.

[12] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[13] Z. Wu, A. A. Efros, and S. X. Yu, "Improving generalization via scalable neighborhood component analysis," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 685–701.

[14] J. Kang, R. Fernandez-Beltran, Z. Ye, X. Tong, P. Ghamisi, and A. Plaza, "Deep Metric Learning based on Scalable Neighborhood Components for Remote Sensing Scene Characterization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8905–8918, 2020, dOI:10.1109/TGRS.2020.2991657.

[15] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, 2010, pp. 270–279.

[16] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, 2017.

[17] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.

[18] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, "Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1144–1158, 2017.

[19] Y. Hua, L. Mou, and X. X. Zhu, "Label relation inference for multi-label aerial image classification," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2019, pp. 5244–5247.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.