REPRESENTATION LEARNING OF REMOTE SENSING KNOWLEDGE GRAPH FOR ZERO-SHOT REMOTE SENSING IMAGE SCENE CLASSIFICATION

*Yansheng Li [1], Deyu Kong [1], Yongjun Zhang [1, *], Ruixian Chen [1], and Jingdong Chen [2]*

[1] School of Remote Sensing and Information Engineering, Wuhan University, China
[2] Ant Group, China

IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-0369-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IGARSS47720.2021.9553667

ABSTRACT

Although deep learning has revolutionized remote sensing image scene classification, current deep learning-based approaches highly depend on the massive supervision of the predetermined scene categories and have disappointingly poor performance on new categories which go beyond the predetermined scene categories. In reality, the classification task often has to be extended along with the emergence of new applications that inevitably involve new categories of remote sensing image scenes, so how to make the deep learning model own the inference ability to recognize the remote sensing image scenes from unseen categories becomes incredibly important. By fully exploiting the remote sensing domain characteristic, this paper proposes a novel remote sensing knowledge graph-guided deep alignment network to address zero-shot remote sensing image scene classification. To improve the semantic representation ability of remote sensing-oriented scene categories, this paper, for the first time, tries to generate the semantic representations of remote sensing scene categories by representation learning of remote sensing knowledge graph (SR-RSKG). In addition, this paper proposes a novel deep alignment network with a series of constraints (DAN) to conduct robust cross-modal alignment between visual features and semantic representations. Extensive experiments on one merged remote sensing image scene dataset, which is the integration of multiple publicly open remote sensing image scene datasets, show that the presented SR-RSKG obviously outperforms the existing semantic representation methods (e.g., the natural language processing models and manually annotated attribute vectors), and our proposed DAN shows better performance compared with the state-of-the-art methods under different kinds of semantic representations.

Index Terms—Deep alignment network (DAN), remote sensing knowledge graph (RSKG), remote sensing image scene classification, zero-shot learning (ZSL).

1. INTRODUCTION

Due to the insurmountable limitation of pixel-level or object-level methods in high-resolution remote sensing image understanding, more attention has been paid to scene-level remote sensing image classification [1]. In addition, deep learning has greatly improved remote sensing image scene classification [2]. However, the current deep learning methods can achieve good classification performance only when each scene category has sufficient samples. In the era of remote sensing big data, the remote sensing categories are showing an explosive growth trend. It is unrealistic to collect sufficient remote sensing image samples for all categories. Hence, identifying remote sensing image scenes that never appear in the training stage has a vital meaning. Inspired by the humans' inference ability, how to introduce the prior knowledge into the deep learning process would be an ideal way to address this challenging task.

In the computer vision field, the development of zero-shot learning (ZSL) [3] has provided promising solutions to recognize samples from unseen categories. Specifically, ZSL aims to simulate the process of human learning and reasoning, and learns from samples of seen classes and leverages prior knowledge of categories as auxiliary information to identify samples from the unseen categories. By contrast, the following characteristics in the remote sensing domain limit the development of ZSL. On the one hand, the names of remote sensing scene categories have some domain specificities. If we generate the semantic representations of remote sensing scene categories by directly leveraging the natural language processing model (e.g., Word2Vec) to map the remote sensing scene category names, the name vectors cannot intrinsically reflect the semantic information of the remote sensing category. On the other hand, remote sensing image scenes, presenting large intra-class difference and inter-class similarity, have more complex appearances compared with natural images in the computer vision field. Apparently, it is not a straightforward work to extend the ZSL methods in the field of computer vision to the field of remote sensing. As a whole, it deserves massive exploration to promote zero-shot remote sensing image scene classification.

In this paper, we propose a novel remote sensing knowledge graph-guided deep alignment network to address zero-shot remote sensing image scene classification. To improve the semantic representation ability of remote sensing-oriented scene categories, we generate the semantic representations of remote sensing scene categories by representation learning of remote sensing knowledge graph (SR-RSKG). Specifically, this paper manually creates a

knowledge graph for the remote sensing domain based on the domain prior knowledge from human experts where the remote sensing knowledge graph fully considers the rich connections between remote sensing scene elements. And then, we can automatically extract the semantic representations of remote sensing scene categories by representation learning of the constructed remote sensing knowledge graph. To cope with zero-shot remote sensing image scene classification, this paper proposes a novel deep alignment network with a series of well-designed constraints (DAN), which learns the linking hyper-parameters between visual features and semantic representations. Experimental results on one integrated remote sensing image scene dataset show that our proposed SR-RSKG is superior to the traditional knowledge types (i.e., Word2Vec, Bert, and Attribute). In addition, the proposed DAN with SR-RSKG performs better than the state-of-the-art methods.

The remainder of this paper is organized as follows. Section 2 depicts the method including representation learning of remote sensing knowledge graph and the proposed DAN. Section 3 summarizes the experimental results. Finally, the conclusion is detailed in Section 4.

2. METHODOLOGY

In the following, Section 2.1 depicts the semantic representation generation process of remote sensing scene categories by representation learning of remote sensing knowledge graph. Furthermore, Section 2.2 details the proposed DAN for zero-shot remote sensing image scene classification.

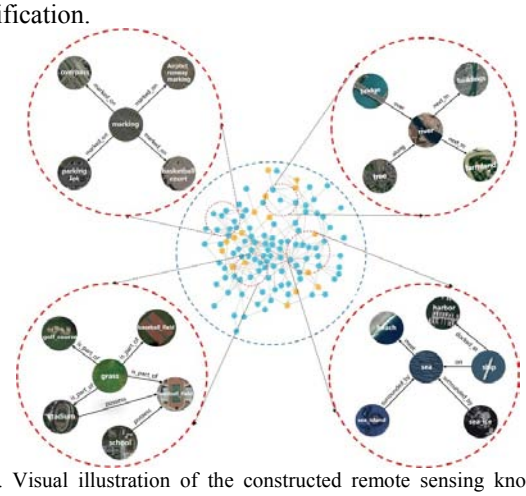


Fig. 1. Visual illustration of the constructed remote sensing knowledge graph.

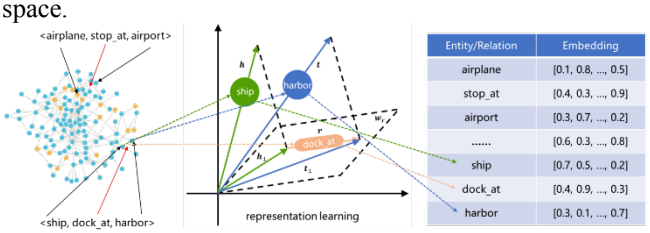
2.1. Representation learning of remote sensing knowledge graph

To support representation learning of remote sensing knowledge graph, we construct one new remote sensing knowledge graph from scratch. In the implementation of the first version of remote sensing knowledge graph, we construct the remote sensing knowledge graph category by category. Specifically, we model the relationship between the ground objects in the remote sensing images from one specific category, and store it in the way of <entity, relation, entity> or <entity, attribute, attribute value> to construct the category-specific remote sensing knowledge graph. And then, we analyze the relationship between different remote sensing scene categories, and aggregate the category-specific knowledge graph to form a unified remote sensing knowledge graph as shown in Fig. 1. The blue represents the entity, and the yellow represents the attribute value. The current remote sensing knowledge graph has 117 entities, 26 relationships, and 191 triples.

As well known, TransE is one traditional representation learning approach [4]. Specifically, for each triple (h, r, t) in the knowledge graph, TransE assumes $c_h + c_r \approx c_t$, that is, the head entity vector plus the relationship vector is approximately equal to the tail entity vector. However, TransE cannot cope with the complex relationships such as 1-N or N-1 appearing in the knowledge graph. In reality, this situation is unavoidable in the remote sensing knowledge graph. As far as the triples <commercial_area, have, trees> and <commercial_area, have, buildings>, if TransE is used to conduct representation learning, the corresponding vectors of trees and buildings can not be effectively distinguished. Considering this limitation of TransE, we adopt TransH [5] to conduct representation learning of remote sensing knowledge graph as TransH models the relation as a hyperplane together with a translation operation on it and improves the model's ability to handle the complex relationships. As shown in Fig. 2, given a triple (h, r, t), TransH maps h and t to the hyperplane, let $h_\perp = h - w_r^{\mathrm{T}} h w_r$ and $t_\perp = t - w_r^{\mathrm{T}} t w_r$ where w_r is the norm vector of the hyperplane, then let $c_{h_\perp} + c_r \approx c_{t_\perp}$. In this way, TransH can learn complex relationships, and the loss function of TransH is formulated by Eq. (1):

$$f_r(h,t) = \|h_\perp + r - t_\perp\|_2^2 \quad (1)$$

After the sufficient optimization of Eq. (1), each entity and relationship in the remote sensing knowledge graph are mapped into semantic representations in one unified feature space.



Entity/Relation	Embedding
airplane	[0.1, 0.8, ..., 0.5]
stop_at	[0.4, 0.3, ..., 0.9]
airport	[0.3, 0.7, ..., 0.2]
.....	[0.6, 0.3, ..., 0.8]
ship	[0.7, 0.5, ..., 0.2]
dock_at	[0.4, 0.9, ..., 0.3]
harbor	[0.3, 0.1, ..., 0.7]

Fig. 2. Representation learning of remote sensing knowledge graph.

2.2. Deep alignment network for zero-shot remote sensing image scene classification

In the following, we firstly give a formulated definition of zero-shot remote sensing image scene classification. And then, the DAN for zero-shot remote sensing image scene classification is detailed.

2.2.1. Formulated definition of zero-shot remote sensing

image scene classification

To facilitate understanding, we firstly give a formulation of ZSL as follows. Let $D^s = \{(x_i^s, y_i^s, c(y_i^s)): i = 1, \dots, N\}$ stands for a set of training examples, where $x_i^s \in X^s$ denotes the visual image feature, extracted by CNN, from the seen scene category. In addition, $y_i^s \in Y^s$ denotes the label of the seen scene category, which are available during the training stage and $c(y_i^s) \in C^s$ stands for the semantic representation (e.g., SR-RSKG) of seen scene categories. In the same way, we define X^u, Y^u, C^u as unseen image features, their corresponding labels and their semantic representations, respectively. It is noted that the seen classes and unseen classes are disjoint, i.e., $Y^s \cap Y^u = \emptyset$. More specifically, X^u does not appear in the training phase, but C^u is always available during both the training and testing phases as it is the prior domain knowledge.

Given the training set D^s and $\{Y^u, C^u\}$, ZSL aims to learn a classifier $F_{ZSL}: X^u \rightarrow Y^u$ to recognize the samples from the unseen categories.

2.2.2. DAN for zero-shot remote sensing image scene classification

Instead of learning the mapping from visual space to semantic space or from semantic space to visual space, we learn the alignment of visual features and semantic representations in latent space. Considering the visual features and semantic representations alignment in latent space and multi-category distribution dispersion, the overall loss of the proposed model is defined as Eq. (2):

$$\mathcal{L} = \mathcal{L}_{VSFR} + \alpha\mathcal{L}_{CMFR} + \beta\mathcal{L}_{VSDM} + \gamma\mathcal{L}_{MCDD} \quad (2)$$

where α, β and γ are the weighting factors of the Cross-modal feature reconstruction loss, Visual and semantic distribution matching loss, Multi-category distribution dispersion loss, respectively.

More specifically, the loss function of visual features and semantic representations reconstruction can be defined by Eq. (3).

$$\begin{aligned}\mathcal{L}_{VSFR} = {} & \mathbb{E}_{q_{\phi^{(v)}}(z^{(v)}|x)}[\log p_{\theta^{(v)}}(x|z^{(v)})] \\ & - D_{KL}(q_{\phi^{(v)}}(z^{(v)}|x)||p_{\theta^{(v)}}(z^{(v)})) \\ & + \mathbb{E}_{q_{\phi^{(a)}}(z^{(a)}|c)}[\log p_{\theta^{(a)}}(x|z^{(a)})] \\ & - D_{KL}(q_{\phi^{(a)}}(z^{(a)}|c)||p_{\theta^{(a)}}(z^{(a)}))\end{aligned} \quad (3)$$

where x represents the original input data of the visual feature, $q_{\phi^{(v)}}$ corresponds to the encoder of visual features, $p_{\theta^{(v)}}$ corresponds to the decoder of visual features, c represents the original input data of visual features, $q_{\phi^{(a)}}$ corresponds to the encoder of semantic representations, and $p_{\theta^{(a)}}$corresponds to the decoder of semantic representations.

The loss function of cross-modal feature reconstruction can be defined by Eq. (4).

$$\begin{aligned}\mathcal{L}_{CMFR} = \Sigma_i^N |x_i - p_{\theta^{(v)}}(q_{\phi^{(a)}}(c_i))| + |c_i \\ - p_{\theta^{(a)}}(q_{\phi^{(v)}}(x_i))|\end{aligned} \quad (4)$$

where x_i and c_i represent the visual features and semantic representations of the same category.

The loss function of visual and semantic distribution matching can be defined by Eq. (5).

$$\mathcal{L}_{VSDM} = \Sigma_i^N \sqrt{\left\|\mu_i^{(v)} - \mu_i^{(a)}\right\|_2^2 + \left\|\left(\Sigma_i^{(v)}\right)^{\frac{1}{2}} - \left(\Sigma_i^{(a)}\right)^{\frac{1}{2}}\right\|_F^2} \quad (5)$$

where $\mu_i^{(v)}$ and $\sqrt{\Sigma_i^{(v)}}$ represent the mean and standard deviation of the visual feature distribution in latent space, $\mu_i^{(a)}$ and $\sqrt{\Sigma_i^{(a)}}$ represent the mean and standard deviation of the visual feature distribution in latent space.

Finally, the loss function of multi-category distribution dispersion can be defined by Eq. (6).

$$\mathcal{L}_{MCDD} = \| VHV^T - I \|_F^2 \quad (6)$$

where $V = \left\{\mu_1^{(a)}, \mu_2^{(a)}, \dots, \mu_N^{(a)}\right\} \in \mathbb{R}^{d \times N}$.

2.2.3. Classification of remote sensing image scenes from unseen categories

Let $X \in X^u$ denote the visual feature of image scenes from unseen categories. We use the semantic representation encoder to map the semantic representation C to $Z^{(a)}$ in the latent space, and leverage the visual feature encoder to map the visual feature X to $Z^{(v)}$ in the latent space. Then we train a classifier F_{ZSL} using $Z^{(a)}$ as the training sample. Finally, we use F_{ZSL} to classify $Z^{(v)}$ to complete the classification of remote sensing image scenes from unseen categories.

3. EXPERIMENTAL RESULTS

In the following, Section 3.1 introduces the evaluation dataset and metric. In addition, Section 3.2 gives a comparison with the state-of-the-art methods.

3.1. Evaluation dataset and metric

As depicted in [6], we adopt a combined dataset by integrating several public datasets. The merged image scene dataset is composed of 70 scene categories and each category contains 800 image scenes with the size of 256× 256. About the knowledge types, we use four kinds of semantic representations: 1) Word2Vec: we use a Word2Vec model [7] which is trained on the Wikipedia corpus to obtain the 300-D vector for each class name. 2) Bert: we depict each remote sensing scene category by one summarized sentence after checking over 10 random remote sensing image scenes from one given category, then the Bert model [8] maps the sentence description of each remote sensing scene category to one different semantic representation with 1024-dimensional. More details about Bert can refer to our previous research [6]. 3) Attribute: Considering the color, shape and objects contained in each remote sensing scene category, we manually design each dimension of the vector. If the remote sensing scene category has a certain attribute, the corresponding dimension is 1, otherwise it is marked as 0. We create the 59-dimensional vector in this way. 4) SR-RSKG: We use the remote sensing knowledge graph created in Section 3.1

and use the representation learning method to obtain the 50-dimensional vector of each category.

The overall accuracy (OA) is taken as the quantitative metric to evaluate the classification performance of ZSL.

3.2. Comparison with the state-of-the-art methods

To give a full analysis of ZSL methods, we report the classification results under different seen/unseen ratios (e.g., 60/10, 50/20, and 40/30). More specifically, in each given seen/unseen ratio, we evaluate each method over 5 random seen/unseen splits. As aforementioned, four kinds of semantic representations including the Word2vec, Bert, Attribute and SR-RSKG are evaluated, respectively. To show the superiority of our proposed method, we consider the following baselines SAE [9], CIZSL [10], CADA-VAE [11], ZSC-SA [12] in ZSL, the ZSC-SA is a ZSL method for remote sensing field and other methods are for computer vision field. As far as the visual feature space, we extracted a 512-dimensional CNN features for images using ResNet-18 [13] as the visual features of our DAN.

As shown in Table.1, our method can obviously outperform the state-of-the-art methods in terms of under different seen/unseen ratios and with different knowledge types. In addition, it is worth noting that, comparing the different knowledge types used in the same method, the SR-RSKG achieves the best performance in most cases. This proves that vectors obtained based on remote sensing knowledge graph representation learning are superior to embedding vectors extracted by natural language models and manual annotation attribute vectors in describing remote sensing scenes.

Table 1. Comparison of different methods under the OA indicator.

Knowledge type	Seen/Unseen ratio	SAE	CIZSL	CADA-VAE	ZSC-SA	Our DAN
Word2Vec	60/10	23.5±4.2	20.6±0.4	41.4±2.3	26.7±5.3	44.3±2.6
	50/20	13.7±1.7	10.6±3.7	30.3±2.7	15.2±1.0	34.7±1.7
	40/30	9.6±1.4	6.0±1.2	21.2±2.9	12.1±0.8	24.3±3.7
Bert	60/10	22.0±1.7	20.4±4.1	48.1±2.9	29.3±3.8	50.2±2.5
	50/20	12.4±1.9	10.3±1.9	37.1±3.5	18.3±1.3	43.4±2.7
	40/30	8.8±1.3	6.2±2.1	26.3±2.2	13.1±3.0	31.5±2.0
Attribute	60/10	23.6.±2.8	16.4±3.1	47.1±2.9	28.5±3.2	50.1±3.3
	50/20	12.1±1.7	7.5±3.2	35.2±2.1	19.4±2.8	43.1±1.5
	40/30	8.6±1.0	6.2±2.9	26.1±2.6	12.7±2.1	30.1±1.9
Our SR-RSKG	60/10	22.1.±2.3	18.2±2.6	50.5±2.6	31.3±2.5	53.3±3.8
	50/20	12.8±2.3	8.9±2.5	39.6±3.1	19.1±1.7	45.2±1.3
	40/30	9.2±1.5	7.1±1.5	28.2±2.6	13.6±2.5	33.4±3.0

4. CONCLUSION

Driven by increasing practical demands of ZSL in the remote sensing field, this paper focuses on exploiting zero-shot remote sensing image scene classification. In this work, we generate semantic representations of remote sensing scene categories as the calculable prior knowledge by representation learning of remote sensing knowledge graph and apply them to the zero-shot remote sensing image scene classification. In addition, we propose a robust deep alignment network with a series of constraint to complete zero-shot remote sensing image scene classification. In order to evaluate our method, we conduct extensive comparative experiments under different seen/unseen ratios using a large remote sensing image scene dataset. By comparing with the traditional knowledge vectors, we verify the superiority of SR-RSKG from the intuitive illustration and quantitative analysis perspectives. Our proposed DAN outperforms the state-of-the-art ZSL methods.

5. ACKNOWLEDGEMENTS

This work was supported by the National Key Research and Development Program of China under grant 2018YFB0505003; the National Natural Science Foundation of China under Grant 41971284; the China Postdoctoral Science Foundation under Grant 2016M590716 and 2017T100581.

6. REFERENCES

[1] G. Cheng, and et al. Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE, 105(10): 1865-1883, 2017.

[2] Y. Li, and et al. Error-Tolerant Deep Learning for Remote Sensing Image Scene Classification. IEEE Transactions on Cybernetic, in press, 2020.

[3] H. Larochelle, and et al. Zero-data learning of new tasks, In: Proceedings of AAAI, 1(2): 3, 2008.

[4] A. Bordes, and et al. Translating embeddings for modeling multi-relational data. In: Proceedings of Neural Information Processing, 26: 2787-2795, 2013.

[5] Z. Wang, and et al. Knowledge graph embedding by translating on hyperplanes, Proceedings of the AAAI Conference on Artificial Intelligence. 28(1), 2014.

[6] Y. Li, and et al. Learning deep cross-modal embedding networks for zero-shot remote sensing image scene classification, IEEE Transactions on Geoscience and Remote Sensing, in press, 2021.

[7] P. Bojanowski, and et al. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5: 135-146, 2017.

[8] J. Devlin, and et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805v2, 2018.

[9] E. Kodirov, and et al. Semantic autoencoder for zero-shot learning, In: CVPR, pp. 3174-3183, 2017.

[10] M. Elhoseiny, M. Elfeki. Creativity inspired zero-shot learning, In: ICCV, pp. 5784-5793, 2019.

[11] E. Schonfeld, and et al. Generalized zero- and few-shot learning via aligned variational autoencoders, In: CVPR, pp. 8247-8255, 2019.

[12] J. Quan, and et al. Structural alignment based zero-shot classification for remote sensing scenes, In: ICECE, pp. 17-21, 2018.

[13] K. He, and et al. Deep residual learning for image recognition, In: CVPR, pp. 770-778, 2016.