Knowledge-Aware Cross-Modal Text-Image Retrieval for Remote Sensing Images

Li Mi[1], Siran Li[2], Christel Chappuis[1] and Devis Tuia[1]

[1]Environmental Computational Science and Earth Observation Laboratory, École Polytechnique Fédérale de Lausanne (EPFL), 1950 Sion, Switzerland.

[2]Section of Electrical and Electronics Engineering, École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland.

Abstract

Image-based retrieval in large Earth observation archives is difficult, because one needs to navigate across thousands of candidate matches only with the proposition image as a guide. By using text as a query language, the retrieval system gains in usability, but at the same time faces difficulties due to the diversity of visual signals that cannot be summarized by a short caption only. For this reason, as a matching-based task, cross-modal text-image retrieval often suffers from information asymmetry between texts and images. To address this challenge, we propose a Knowledge-aware Cross-modal Retrieval (KCR) method for remote sensing text-image retrieval. By mining relevant information from an external knowledge graph, KCR enriches the text scope available in the search query and alleviates the information gaps between texts and images for better matching. Experimental results on two commonly used remote sensing text-image retrieval benchmarks show that the proposed knowledge-aware method outperforms state-of-the-art methods.

Keywords

Cross-modal Retrieval, Knowledge Graph, Remote Sensing

1. Introduction

Recent advances in satellite data acquisition and storage have led to a rapid development of remote sensing image archives. To explore them, image retrieval has received increasing attention [1, 2]. However, retrieving images using example images would limit the versatility of the retrieval system, since with the query image only, one cannot specify which elements are essential for the query or what the retrieval objective is. As a solution, text-image retrieval [3, 4] has been introduced to explicit the retrieval targets in a semantic way. Text-image retrieval aims at recalling an image based on a text or, in reverse, retrieving a text according to an image. As a bridge between vision and language research, it provides a possibility to explore the growing amount of cross-modal remote sensing data.

When regarding text as the query, the prospective retrieval system gains in usability among cross-modal data, but at the same time faces the problem of information asymmetry between texts and images [5]. When dealing with very high-resolution remote sensing images, the image content can be very diverse, hence it is difficult to be comprehensively summarized by the natural language, especially by a short caption. On one hand, human captions can only describe the image from one or a few specific aspects, focusing on the most dominant information. For example, one image could receive the following caption text: *There is a lake.* Nevertheless, there might be trees and mountains around the lake which are ignored by humans or caption generators. In addition, different people will describe the image from subjective perspectives, resulting in a variety of text information for a single image, which may confuse the matching model. Therefore, strategies to handle lacunary captions, nuances and synonyms are needed for the task, and a balance between objectivity and completeness must be achieved.

Knowledge graphs [6] present relationships and proximities among concepts through graph structures. By providing the experience and commonsense from human understanding, knowledge graphs have been recognized as effective prior knowledge in many vision-and-language research [7, 8] to reveal commonsense and alleviate ambiguities. In this paper, we propose a Knowledge-aware Cross-modal Retrieval (KCR) method for remote sensing text-image retrieval. With the help of external Knowledge Graphs, KCR extends the text scope to obtain a more robust text representation. More specifically, based on the objects mentioned in a sentence as starting points, KCR proposes to mine the expanded nodes and edges in a knowledge graph and embeds them as features to enrich those extracted from the text content alone. As such, KCR integrates commonsense knowledge and leads to competitive performance on two commonly used remote sensing text-image retrieval benchmarks.

CDCEO 2022: 2nd Workshop on Complex Data Challenges in Earth Observation, July 25, 2022, Vienna, Austria

li.mi@epfl.ch (L. Mi); siran.li@epfl.ch (S. Li); christel.chappuis@epfl.ch (C. Chappuis); devis.tuia@epfl.ch (D. Tuia)

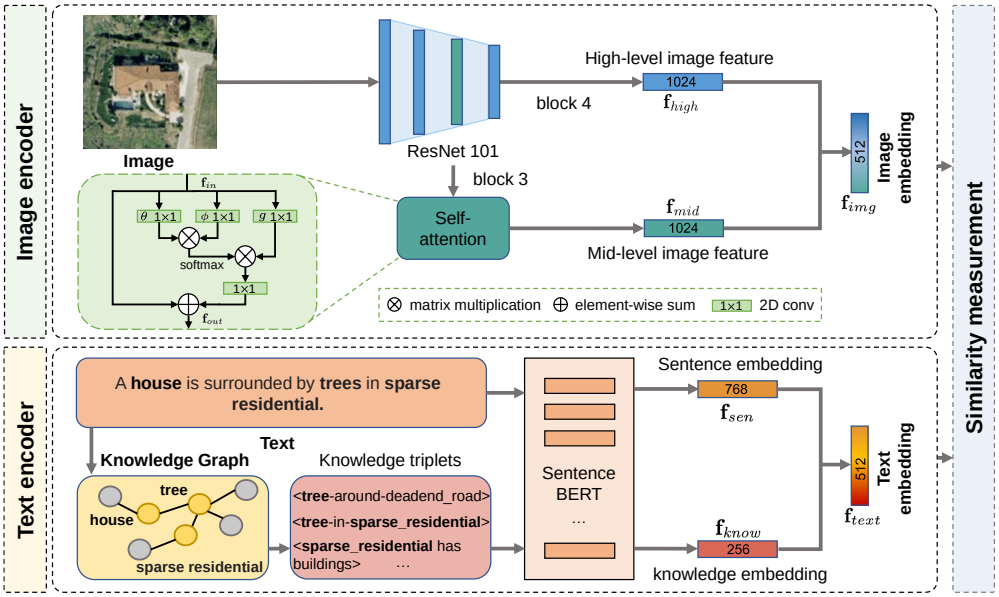

Figure 1: The proposed KCR method. KCR consists of three main components: an image encoder with self-attention block, a knowledge-aware text encoder and a similarity measurement module. During the training process, the parameters of ResNet (except the final FC layer) and Sentence BERT are fixed.

2. Related Work

Due to the emergence of multi-modal remote sensing data, vision-language research, such as image captioning [9], visual question answering [10], and cross-modal retrieval [4] has attracted increasing attention [11]. Recent advances in remote sensing text-image retrieval mainly focused on: 1) learning a more representative image feature by fusing local and global image feature [12] or features from different feature extractors [4]. 2) learning a distinguishable joint embedding space by fusing the cross-modal features and leveraging a ranking-based loss function [4]. Departing from previous efforts that based the retrieval on the image characteristic and the caption only, we propose to enrich the latter with a knowledge graph that would extend the text content and alleviate ambiguities for a more robust text representation.

Consisting of various nodes as concepts, knowledge graphs encode commonsense knowledge about the world [6, 13, 14]. By exploring the knowledge graphs, vision-and-language research has been promoted due to the priors for visual understanding [7, 8]. In remote sensing research, Li *et al.* [15] constructed a remote sensing knowledge graph to support zero-shot remote sensing image scene classification. Their efforts in exploiting a remote sensing knowledge graph to image understanding focused on using graph embedding as an overall representation of an image. Different from this work, the proposed KCR explores the fine-grained object-level connections between nodes in the graph and words in sentences.

3. Knowledge-aware Cross-modal Retrieval Method

The proposed text-image retrieval system comprises three main components: an image encoder, a text encoder and a similarity measurement module (Figure 1). The image encoder is designed to extract image features by a pre-trained feature extractor and a self-attention block. The text encoder embeds a sentence and its related external knowledge extracted from a knowledge graph into a joint feature space. Finally, the image and text features are both used within the similarity measurement module to compute the similarity score between text queries and candidate images, which are then ranked according to their relevance. The model can also be applied in reverse, where the best captions to summarize an image are retrieved.

3.1. Image encoder

The image encoder is a pre-trained feature extractor with a self-attention block [16, 17]. Two sets of image features are extracted from the image encoder:

High-level image feature. We use ResNet-101 [18] as a backbone and the last Fully Connected (FC) layer is retrained. For an image i, the output of the retrained FC layer is regarded as the high-level image feature $\mathbf{f}_{high}$.

Mid-level image feature. The output of ResNet block 3, denoted as $\mathbf{f}_{in}$, is sent to an additional self-attention block to further capture the long-range dependencies among

pixels and provide more detailed information at relatively mid-level. The self-attention block can be defined as Eq. (1) [16].

$$\mathbf{f}_{out} = \text{softmax}\left(\mathbf{f}_{in}^T W_\theta^T W_\phi \mathbf{f}_{in}\right) g(\mathbf{f}_{in}). \quad (1)$$

Here $\mathbf{f}_{out}$ is the output feature of the same size as $\mathbf{f}_{in}$. g is a linear embedding of the input feature: $g(\mathbf{f}_{in}) = W_g \mathbf{f}_{in}$. W_g, W_θ, and W_ϕ are weight matrices of 1×1 convolutions to embed the feature. Then followed by a 2D pooling layer and a flattening operation, the mid-level image feature is extracted as a vector:

$$\mathbf{f}_{mid} = \text{pool2d}(\mathbf{f}_{out}). \quad (2)$$

To project the image feature and text feature into a same dimension, the concatenated high-level and mid-level image feature are sent to a final FC layer to obtain the overall image presentation:

$$\mathbf{f}_{img} = \text{FC}_{img}(\text{concat}(\mathbf{f}_{high}, \mathbf{f}_{mid})). \quad (3)$$

3.2. Knowledge-aware text encoder

Knowledge representation. For a sentence s with n words: $s = \{w_1, w_2, ..., w_n\}$ $(n \geq 1)$, a tokenizer is used to separate every word and divide the part-of-speech (e.g. noun, verb, adjective, adverb, etc.) for them. Based on the part-of-speech tags, all the nouns can be appended into a word list. Then we extract a sentence graph G_s based on the word list only. The sentence graph can be regarded as a subgraph of the existing remote sensing knowledge graph [15], G. More specifically, the nouns in the word list are regarded as the initial nodes. Starting from those nodes, all the one-step neighbours with the connected edges in G are included in G_s. Note that the sentence graph is a directed graph, which means the edge between two nodes is a one-way relationship. In the sentence graph, each edge can be represented as a relationship triplet $r(s, p, o)$, shown as $<subject-predicate-object>$, which can be regarded as a short sentence with three words. Mining all the edges of the sentence graph might be redundant, so we decide to randomly select q triplets from all the available ones.
Text encoder. We use Sentence-Transformer [19] as the text encoder for the sentence features $\mathbf{f}_{sen}$, as well as the external knowledge representation $\mathbf{f}_{know}$. Sentence-Transformer is a modification of the pretrained BERT network using siamese and triplet network structures to derive semantically meaningful sentence embeddings. The encoding process of a sentence and the corresponding knowledge can be formulated as:

$$\begin{aligned} \mathbf{f}_{sen} &= \text{SenTrans}(s\{w_1, w_2, ..., w_n\}) \\ \mathbf{f}_{know} &= \text{FC}_{know}(\frac{1}{q}\sum_{i=0}^{q} \text{SenTrans}(r_i\{s_i, p_i, o_i\})). \end{aligned} \quad (4)$$

After being embedded in the feature space, sentence representation and external knowledge representation are concatenated and sent to the finial FC layer to obtain the overall representation of a text:

$$\mathbf{f}_{text} = \text{FC}_{text}(\text{concat}(\mathbf{f}_{sen}, \mathbf{f}_{know})). \quad (5)$$

3.3. Similarity Measurement

Similarity score. The similarity score S_{is} is defined as the negative pairwise euclidean distance between two features: $S_{is} = -\text{dis}(\mathbf{f}_{img}, \mathbf{f}_{text})$. With smaller distance to the query feature, the similarity score is larger and the target ranks higher.
Loss function. Triplet loss is commonly used in the text-image retrieval task [3, 20, 4]. It constrains the similarity score of the matched image-text pairs to be larger than the similarity score of the unmatched ones by a margin. Meanwhile, previous research [3] discovered that using the hardest negative in a batch during training rather than all negatives samples can boost performance. Therefore, the loss function can be formulated as:

$$\begin{aligned} L(i, s) = &\max(0, m - S_{is} + S_{is'}) \\ &+ \max(0, m - S_{si} + S_{si'}), \end{aligned} \quad (6)$$

where m is a margin parameter, image i and sentence s are the corresponding pair. Sentence s' is the top-1 text retrieval result with query image i and image i' is the top-1 image retrieval result with query text s.

4. Experiments

4.1. Experimental details

Datasets. We perform experiments on two commonly used RS text-image datasets: the RSICD dataset and UCM-Caption dataset. The RSICD dataset [9] contains 10921 images with the size 224×224 pixels. The UCM-Captions dataset [23], which is based on the UC Merced Land Use dataset [24], contains remote sensing images categorized into 21 land use classes, with 100 samples for each class. For each sample in both datasets, there are 5 sentences describing the image content. We follow the train-test split in previous work [4], randomly selecting 80%, 10% and 10% for the dataset as the training set, validation set and test set, respectively.
Metrics. To evaluate the model performance, we exploit the standard evaluation metrics in retrieval tasks and measure the rank-based performance by Recall@k (R@k) and mR [3, 25]. With different values of k, R@k means the fraction of queries for which the most relevant item is ranked among the top-k retrievals. mR represents the average of all R@k in both text-image retrieval and

Table 1
Experimental results on the RSICD dataset. **KCR w/o KG** denotes the proposed model without knowledge triplet embedding and **KCR w/o KG Att** represents the proposed model without knowledge triplet embedding and self-attention block.

	Backbone	Text - Image Retrieval			Image - Text Retrieval			mR
		R@1	R@5	R@10	R@1	R@5	R@10	
VSE++ [3]	ResNet18	2.82	11.32	18.10	3.38	9.51	17.46	10.43
SCAN [20]		3.71	16.40	26.73	5.85	12.89	19.84	14.24
MTFN [21]		4.90	17.17	29.49	5.02	12.52	19.74	14.81
AMFMN [12]		4.90	18.28	31.44	5.39	15.08	23.40	16.42
GaLR [4]		4.69	19.48	32.13	**6.59**	**19.85**	**31.04**	18.96
KCR		**5.84**	**22.31**	**36.12**	4.76	18.59	27.20	**19.14**
CAMP [22]	ResNet101	4.15	15.23	27.81	5.12	12.89	21.12	14.39
KCR w/o KG Att		4.47	20.64	33.68	3.94	12.36	24.08	16.53
KCR w/o KG		4.63	20.11	34.77	4.12	18.40	29.30	18.56
KCR		**5.40**	**22.44**	**37.36**	**5.95**	**18.59**	**29.58**	**19.89**

Table 2
Experimental results on the UCM-Caption dataset.

	Backbone	Text - Image Retrieval			Image - Text Retrieval			mR
		R@1	R@5	R@10	R@1	R@5	R@10	
VSE++ [3]	ResNet18	10.10	31.80	56.85	12.38	44.76	65.71	36.93
SCAN [20]		12.76	50.38	77.24	14.29	45.71	67.62	44.67
MTFN [21]		14.19	52.38	78.95	10.47	47.62	64.29	44.65
AMFMN [12]		12.86	53.24	79.43	**16.67**	45.71	68.57	46.08
KCR		**17.24**	**56.95**	**81.14**	11.90	**48.57**	**71.43**	**47.87**
CAMP [22]	ResNet101	11.71	47.24	76.00	14.76	46.19	67.62	43.92
KCR w/o KG Att		16.67	54.19	81.52	7.14	41.42	62.38	43.89
KCR w/o KG		16.00	52.90	**81.81**	10.00	44.29	69.05	45.68
KCR		**17.43**	**57.52**	80.38	**15.24**	**50.95**	**73.33**	**49.14**

image-text retrieval. In our experiment, we report the results of $k = 1$, $k = 5$, and $k = 10$.

Hyper-parameters. In all experiments, the margin of the triplet loss function is set to 0.2 following the previous work [4]. For the image encoder, the input and intermediate dimensions of the self-attention block are respectively set to 1024 and 512, according to [16]. In terms of text encoder, the number of selected triplets are set to 10 in our experiments. Other feature dimensions are annotated in Figure 1. In addition, to achieve fair comparison with the competing methods, results with the ResNet18 backbone are also reported. For ResNet18 backbone, the dimension of the mid-level feature is 256 and other parameters are as the same of the model with ResNet101 backbone.

Implementation details. For the training process, we train and evaluate the model in mini-batch with a batch-size of 100. The optimizer is Adam optimizer with a weight decay of 5e-4 and initial learning rate of 0.001. For every 10 epochs, the learning rate drops 10%. All the experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU. The max training epoch are 150 and 200 for the UCM-Caption and RSICD dataset, respectively.

4.2. Experimental results

Comparison methods. We compare the proposed method with the following state-of-the-art methods in text-image retrieval, especially those for remote sensing text-image retrieval.

- **VSE++** [3] uses a CNN and a Gated Recurrent Unit (GRU) [26] to capture image and text features, respectively.
- **SCAN** [20] exploits fine-grained interplay between images and texts by inferring the semantic alignment between them.
- **CAMP** [22] proposes a cross-modal message passing method to explore the image-text interactions before calculating similarities.
- **MTFN** [21] introduces a rank-based fusion model to avoid finding the common embedding space for cross-modal data.
- **AMFMN** [12] employs multiscale visual self-attention module to extract the visual features and guide the text representation.
- **GaLR** [4] utilizes an attention-based multi-level

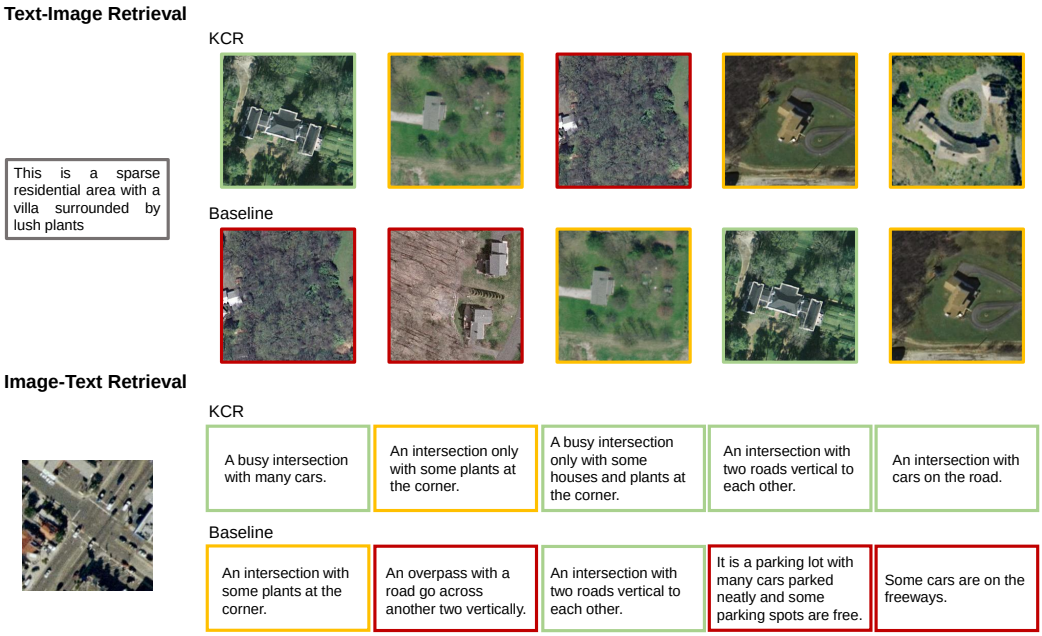

Figure 2: Examples of KCR and the baseline model (KCR w/o KG) top 5 retrieval results. Texts or images in green boxes are predictions that correspond to the ground truth. If the Texts or images are in orange boxes, it means that these results are not annotated as the ground truth but regarded as reasonable predictions according to human evaluation. Texts or images in red are semantically incorrect retrieval results.

information dynamic module to fuse global and local feature extracted by a CNN and a Graph Neural Network (GCN), respectively. In addition, GaLR involves a post-processing stage based on a plug-and-play multivariate rerank algorithm.

Results on the RSICD dataset (Table 1). KCR achieves the best performance with the exception of image-text retrieval, where GaLR is the best performing. With ResNet18 backbone, KCR outperforms GaLR on mR by 0.18%. In terms of text-image retrieval, the improvements are 1.15%, 2.83% and 3.99% for R@1, R@5, and R@10, respectively. For image-text retrieval, KCR achieves close performance compared to GaLR and outperforms other competitors. Note that compared to GaLR, which has multiple image feature extractors and post-processing stage, the structure of KCR is less conceptually heavy. Experimental results on the sub-component analysis of KCR (e.g. running the model without knowledge embedding and self-attention module) show that incorporating commonsense knowledge can extend sentence content and alleviate the information gap, since the model performance is significantly improved. The combination of external knowledge brings an extra 0.77%, 2.33%, and 2.59% for the three metrics in text-image retrieval. For image-text retrieval, external knowledge improves the model performance by 1.83%, 0.19%, and 0.28% on R@1, R@5, and R@10, respectively. Removing the self-attention module and the mid-level feature degrade the results by 2.03% on mR, which indicates the importance of representative mid-level image feature.

Results on the UCM-Caption dataset (Table 2). For text-image retrieval, KCR significantly outperforms state-of-the-art methods. For R@1, R@5, and R@10, With ResNet18 backbone, KCR achieves the best performance with an average improvement of 3.27% compared to AMFMN. For image-text retrieval, KCR gains 2.86% on both R@5 and R@10. The overall improvement on mR is 1.79%. As is shown in the sub-component analysis, self-attention block and mid-level feature improve the model performance on mR by 1.79%. External knowledge improves the model performance, especially for image-text Retrieval. The average improvements on the three metrics are 5.24%, 6.66%, and 4.28% respectively. mR gains a 3.46% increase because of introducing relevant knowledge from knowledge graph. Meanwhile, compared with RSICD dataset, knowledge embedding has a more obvious improvement on the UCM-Caption dataset, indicating that the information gap might be larger on the smaller dataset. Examples of the top-5 retrieval results of KCR and KCR w/o KG are shown in Figure 2. In addition, we observe that ResNet101 is slightly more effective than ResNet18, with observed improvements of 0.75% on average for RSICD dataset and 1.27% on average for UCM-Caption dataset.

5. Conclusion

Retrieving remote sensing images from text queries is appealing but complex, since retrieval needs to be both visual and semantic. To address the information asymmetry between images and texts, we propose a Knowledge-aware Cross-modal Retrieval (KCR) method. By integrating relevant information from external knowledge graph, the model enriches the text scope to better match texts and images. Despite its conceptual simplicity, KCR shows improved performance with respect to all competitors, which indicates potential generalization capabilities of the knowledge-aware method.

References

[1] W. Zhou, S. Newsam, C. Li, Z. Shao, PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval, ISPRS Journal of Photogrammetry and Remote Sensing 145 (2018) 197–209.

[2] G. Hoxha, F. Melgani, B. Demir, Toward remote sensing image retrieval under a deep image captioning perspective, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13 (2020) 4462–4475.

[3] F. Faghri, D. J. Fleet, J. R. Kiros, S. Fidler, VSE++: Improving visual-semantic embeddings with hard negatives, arXiv preprint arXiv:1707.05612 (2017).

[4] Z. Yuan, W. Zhang, C. Tian, X. Rong, Z. Zhang, H. Wang, K. Fu, X. Sun, Remote sensing cross-modal text-image retrieval based on global and local information, IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 1–16.

[5] K. Wang, Q. Yin, W. Wang, S. Wu, L. Wang, A comprehensive survey on cross-modal retrieval, arXiv preprint arXiv:1607.06215 (2016).

[6] F. Ilievski, P. Szekely, B. Zhang, CSKG: The commonsense knowledge graph, in: ESWC, 2021, pp. 680–696.

[7] W. Yang, X. Wang, A. Farhadi, A. Gupta, R. Mottaghi, Visual semantic navigation using scene priors, arXiv preprint arXiv:1810.06543 (2018).

[8] Y. Fang, K. Kuan, J. Lin, C. Tan, V. Chandrasekhar, Object detection meets knowledge graphs, in: IJCAI, 2017, pp. 1661–1667.

[9] X. Lu, B. Wang, X. Zheng, X. Li, Exploring models and data for remote sensing image caption generation, IEEE Transactions on Geoscience and Remote Sensing 56 (2017) 2183–2195.

[10] S. Lobry, D. Marcos, J. Murray, D. Tuia, RSVQA: Visual question answering for remote sensing data, IEEE Transactions on Geoscience and Remote Sensing 58 (2020) 8555–8566.

[11] D. Tuia, R. Roscher, J. D. Wegner, N. Jacobs, X. Zhu, G. Camps-Valls, Toward a collective agenda on AI for earth science data analysis, IEEE Geoscience and Remote Sensing Magazine 9 (2021) 88–104.

[12] Z. Yuan, W. Zhang, K. Fu, X. Li, C. Deng, H. Wang, X. Sun, Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval, arXiv preprint arXiv:2204.09868 (2022).

[13] M. Sap, R. Le Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, Y. Choi, ATOMIC: An atlas of machine commonsense for if-then reasoning, in: AAAI, 2019, pp. 3027–3035.

[14] R. Speer, J. Chin, C. Havasi, ConceptNet 5.5: An open multilingual graph of general knowledge, in: AAAI, 2017, pp. 4444–4451.

[15] Y. Li, D. Kong, Y. Zhang, Y. Tan, L. Chen, Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification, ISPRS Journal of Photogrammetry and Remote Sensing 179 (2021) 145–158.

[16] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: CVPR, 2018, pp. 7794–7803.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: NIPS, volume 30, 2017.

[18] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016, pp. 770–778.

[19] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: EMNLP, 2019, pp. 671–688.

[20] K.-H. Lee, X. Chen, G. Hua, H. Hu, X. He, Stacked cross attention for image-text matching, in: ECCV, 2018, pp. 201–216.

[21] T. Wang, X. Xu, Y. Yang, A. Hanjalic, H. T. Shen, J. Song, Matching images and text with multi-modal tensor fusion and re-ranking, in: ACM MM, 2019, pp. 12–20.

[22] Z. Wang, X. Liu, H. Li, L. Sheng, J. Yan, X. Wang, J. Shao, CAMP: Cross-modal adaptive message passing for text-image retrieval, in: CVPR, 2019, pp. 5764–5773.

[23] B. Qu, X. Li, D. Tao, X. Lu, Deep semantic understanding of high resolution remote sensing image, in: CITS, 2016, pp. 1–5.

[24] Y. Yang, S. Newsam, Bag-of-visual-words and spatial extensions for land-use classification, in: SIGSPATIAL, 2010, pp. 270–279.

[25] X. Huang, Y. Peng, Deep cross-media knowledge transfer, in: CVPR, 2018, pp. 8837–8846.

[26] K. Cho, B. Van Merriënboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: Encoder-decoder approaches, arXiv preprint arXiv:1409.1259 (2014).