Towards Multimodal Data Retrieval in Remote Sensing

Taghreed Abdullah[a*], *Lalitha Rangarajan*[b]

[a*,b]*Department of Studies in Computer Science, University of Mysore, India*

ABSTRACT

The world around us is multimodal in nature: seeing scenes, hearing voices, watching videos, and savoring flavors. Recently, multimodal applications, which deal with multiple modalities, especially image-text retrieval (matching) a topic of broad and current interest in the general literature of computer vision. Yet most of the existing remote sensing image retrieval approaches rely on the concept of image-image matching (unimodal). In this paper, we aim to draw the attention of researchers in the remote sensing community to a recent direction multimodal data retrieval (matching), particularly image-text matching, which is considered a recent research direction, due to its importance for human intelligence to grasp the relation between visual and textual content and to bridge the semantic gap between such different contents (modalities) in light of tremendous progress in deep learning techniques through highlighting the three main challenges (multimodal representation, similarity measurement, and dataset availability) that face researchers in this research line.

***Keywords**:* Remote Sensing; Multimodal; Image-text matching; Representation learning; Deep metric learning; deep learning**.**

1. Introduction

The world around us is multimodal in nature: seeing scenes, hearing voices, watching videos, and savoring flavors. In the literature the multimodal can be presented in several forms as shown in Fig.1

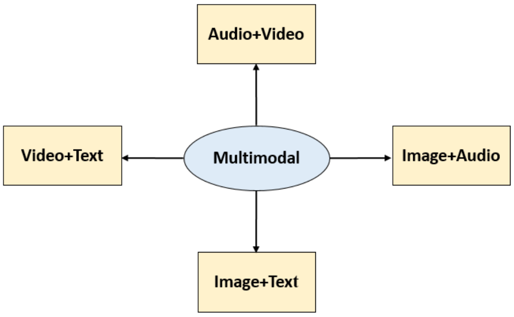


Fig.1 Different multimodal.

The amount of remote sensing data is growing exponentially, with the progress of earth observation technologies over the past few decades. Consequently, millions of high-resolution remote sensing images were collected and stored in huge archives. Wherefore, there is a pressing need to manage and interpret such data efficiently in order to leverage perfectly from such remote sensing data, in the light of the increasing demand for effective image retrieval systems that retrieve images depending on their semantic meaning from these huge archives. The challenge of finding a relevant image from such large archives is an essential challenging research problem For RS community. In this respect, the most popular topics are image classification and retrieval.

Narrowing the "semantic gap" is one of the first goals of remote sensing tasks like remote sensing image classification and retrieval. Connecting visual and textual data are mainly important for human intelligence to understand the relation between textual and visual content and to bridge the semantic gap between such different contents (modalities).

Lately, a new trend of multimodal applications protruded with focusing on natural language and visual data to explore and understand the relationship between them. In the literature of computer vision, this line of research direction became active, due to its significance in various applications. Image-text retrieval [1]–[3] aims to match the most related images for a specific query text. Image captioning [4], [5] is to produce a linguistic sentence for describing an image. Such applications are mainly relied on learning the image-text matching model effectively, which aims to establish a relationship between image and text, such that semantically related pairs (image, text) possess higher matching score than the unrelated ones. The mage-text retrieval task is challenging due to the existence of the heterogeneous gap between visual and textual modalities. In recent computer vision literature, lots of efforts have been performed for bridging such heterogeneous gap [6]–[8]. In remote sensing, the mainstream works focusing on unimodal data retrieval (image-image matching), which is known as remote sensing image retrieval (RSIR), due to that textual description for remote sensing images is too complex than natural image description, as well as remote sensing images, contain ambiguous semantic from the “view of God”. We use the term unimodal, referring to RSIR, which is an application of content-based image retrieval (CBIR). Basically, CBIR based on two main steps: 1) feature extraction for images in the archive and query image, 2) feature matching (similarity measurement) between query image features and archive images features. In the case of multimodal data retrieval (image-text matching), the key challenge is how to efficiently extracting discriminating features for an image and a text and how to measure the similarity between them, despite their difference, i.e., How to bridge the so-called heterogeneous gap between such different modalities. Fig.2 shown an overview of multimodal data retrieval.

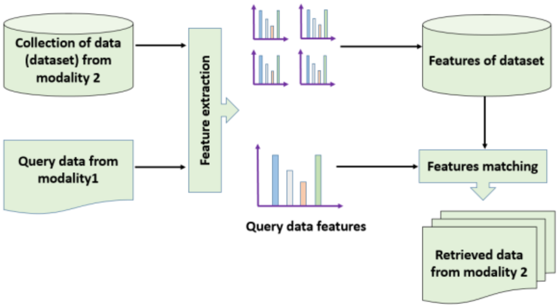

Collection of data (dataset) from modality 2
Query data from modality1
Feature extraction
Features of dataset
Features matching
Query data features
Retrieved data from modality 2

Fig.2 Overview of multimodal data retrieval.

To overcome the difficulties facing multimodal data retrieval, new technologies have emerged in deep learning, more efficient to extract powerful features from huge data, have accomplished salient success in many applications that deal with image and text together, specifically image-text retrieval (matching).

Due to the great advancement of deep learning techniques, especially Convolutional Neural Networks (CNNs) [9] and Recurrent Neural networks (RNNs), which demonstrate superior performance in both computer vision and remote sensing, the development of multimodal retrieval has taken a new turn. Pre-trained CNNs are commonly used to extract image features automatically. They can extract high-level semantic features due to their deep architectures. Such learned features by CNNs able to describe the remote sensing images effectively, and made great contributions in multimodal retrieval. In contrast, RNNs are models using for processing sequential data (e.g., text). Long-Short Term Memory networks (LSTM) [10] is a kind of RNNs, which has long-range dependencies that makes it more accurate than RNNs.

Learning from visual and textual sources helps to deeply grasp natural phenomena. Further, given the scarcity of studies related to remote sensing image-text retrieval as mentioned above as well as the importance of multimodal as a lifelike topic of increasing significance, there is a clear need to highlight the key challenges that are encountered by remote sensing image-text retrieval.

2. Challenges in image-text matching

The heterogeneity of visual and textual data brings some challenges for researchers in the remote sensing field. The three of the essential challenges, which are important to be tackled for progressing in this field are multimodal representation, similarity measurement, and data sets availability.

2.1. Multimodal Representation

Representation learning is an essential process for several tasks, such as classification, retrieval, recognition. Representation learning aims to learn discriminative features automatically using deep architectures. Multimodal representation learning is a particular representation learning, which automatically learns discriminative features from heterogeneous dependent modalities (e.g., image, text, video, audio, etc.) that have correlations among them.

In remote sensing, multimodal representation learning for image-text retrieval task faces many difficulties because of its characteristics: how to combine text and image?; how to commonly learning discriminating features from text and image?. Representing data in a good way is crucial to multimodal data retrieval performance, and the foundation for any model.

The architecture of the image-text matching framework consists of two different encoders one for an image, such as VGG16 [11], ResNet152 [12], and another one for a text such as GRU [13], LSTM. Robust representations are significant for the performance of the image-text retrieval model, as shown by the recent important improvement in performance. Wang et al. [1] proposed two-branch neural networks; a similarity network and an embedding network for image-text matching. The Similarity network and the Embedding Network each have two input divisions, one for images and one for text. Both divisions have a feature extractor and a fully connected layer. VGG network is used to extract image features, while text features are extracted by LSTM. One of the most recent work in multimodal data retrieval in remote sensing is matching text to remote sensing images [14]. In this work, the authors employed the most advance deep learning techniques to learn the discriminative representation for textual and visual data. To extract visual features, different pre-trained CNNs are used including, EfficientNet, inception-v3, VGG16, ResNet50 with some additional techniques like squeeze excitation (SE) [15] and global average pooling (GAP) to learn more informative features. For extracting textual features Long Short Term Memory network (LSTM) is adopted. This work is the first in the remote sensing literature that applied a retrieval task by using two different modalities (text and image). Different architectures for the image-text matching task are presented in Table 1.

Table 1- Different deep learning architectures used for representing image and text modalities.

Ref.	Architectures (Text and image encoders in image-text matching works)
[16]	LSTM+VGG19, ResNet-152
[17]	HGLMM+ResNet152
[2]	BI-LSTM+ResNet152
[18]	FV+VGG19, ResNet
[19]	BERT+ResNet101
[14]	LSTM+ResNet-50,VGG-16,Inception-v3, Efficient-B2

2.2. Similarity measurement (matching)

For bridging the heterogeneity gap, representation learning for image and text projects the data samples from image space and text space into a shared space (joint embedding space) as shown in Fig.3 i.e. to learn common representation, where the text and image feature vectors are comparable and similarity between them can be directly measured by using popular distance metrics such as Euclidean distance, Hamming distance, Cosine distance, Cityblock measure.

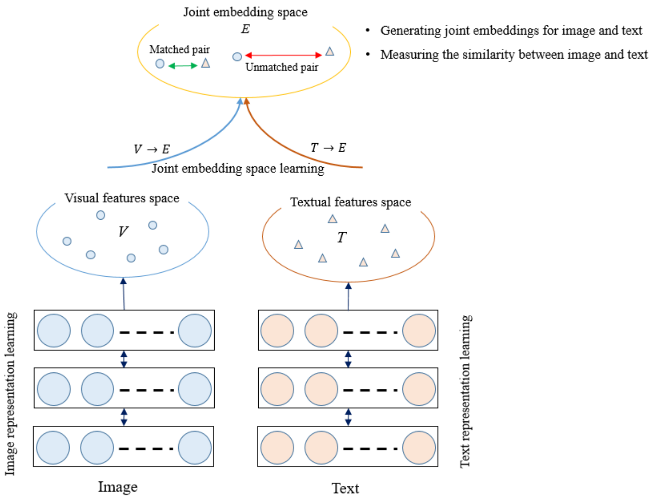

Joint embedding space
E
Matched pair
Unmatched pair
• Generating joint embeddings for image and text
• Measuring the similarity between image and text
$V \rightarrow E$
$T \rightarrow E$
Joint embedding space learning
Visual features space
V
Textual features space
T
Image representation learning
Text representation learning
Image
Text

Fig.3 Joint representation.

He et al. [20] used pre-trained CNNs to project images and texts into multimodal space (joint space), the image-text similarity is calculated by using cosine distance, then, a bidirectional loss including the matched image-text pairs and unmatched ones is optimized.

Recent developments in deep learning and metric learning have helped to learn a similarity measure (a distance function) for a set of samples by employing so-called deep metric learning (DML) [21], which used to measure similarity or dissimilarity (distance) between objects automatically for a task of interest. DML aims to learn a new metric where the distance between similar objects (samples) is reduced, meanwhile the distance between dissimilar objects is increased through learning a distance function. The learned distance can be adopted in many applications, such as image retrieval [22], [23], classification [24],[25], clustering [26].

The loss function is an important part of deep metric learning networks and thus a large variety of loss functions are designed for DML recently such as triplet ranking loss [27], quadruplet loss [28], lifted structured [29], multi-class N-pairs loss [22] and angular loss [30]. The proposed network in [14] used a bidirectional triplet loss for training to enable the task of text-image retrieval and image-text retrieval. Therefore, measuring the similarity between text and image is done by deep metric learning. Table 2 shows several loss functions that are used in image-text matching works.

Table 2- Different types of loss functions using in image-text matching works.

Ref.	**Loss function**
[31],[32]	Ranking loss
[33],[34],[16]	Triplet loss
[14]	Bi-directional triplet loss

[2]	CMPM loss
[35]	Adaptive loss
[36]	Instance loss
[37]	Quintuplet loss

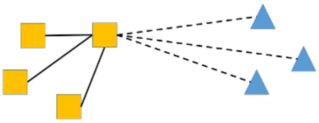


Fig.4 Metric leanring

Wu at al. [38] proposed image-text online similarity function learning approach named (CMOS) to learn automatically the metric that determines the similarity between image and text. In [39], deep coupled metric learning is introduced for text-image cross-modal matching. The samples of two modalities are transformed nonlinearly into common space by using a metric learning technique. A key challenge in image-text retrieval is how to decrease the heterogeneous gap effectively, and exploit the discriminating features across text and image modalities where the similarity between text and image can be measured. In other words, how to measure the similarity (matching) between text and image efficiently?

Most of the image-text matching works used Recall@K (R@K) for different values of (K=1, 5, 10) as matching scores of an image with query sentence and vice versa which is calculated as:

$$\text{Recall@k} = \frac{true\ positives\ @k}{(true\ positives\ @k) + (false\ negative@k)}$$

2.3. Dataset availability

The mainline of remote sensing works concentrate mainly on the scenario of unimodal data retrieval (matching) (image-image). For unimodal data retrieval, there is an abundance of remote sensing datasets publically available, which assign a single label for each image, such as UC Merced [40], AID [41], Patterned, etc. The example images with their single labels of each class in the UC Merced dataset are shown in Fig.4

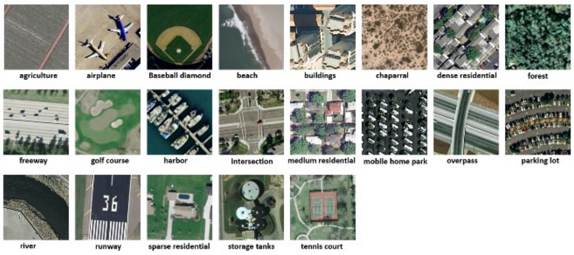


Fig.5 Example images from UC Merced dataset.

In reality, images can contain several categories at the same time. In seeking for tackling this issue, a little few recent works tried to assign multiple labels to each image. In [42] the first multi-label dataset is released, where each image in such dataset is manually labeled. Shao et al. [43] introduced a new multi-label dataset, which assigns multi-label to each image, named MLRSIR.

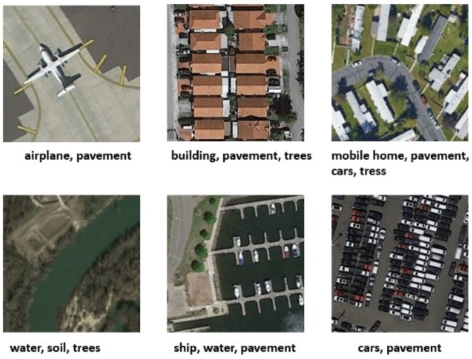
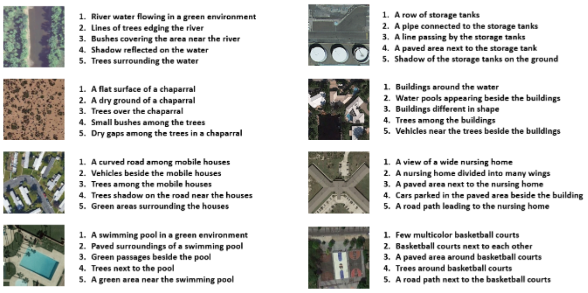


Fig.6 Example to assign multi-label to remote sensing images.

However, when multi labels are given to an image irrespective of the relativity of the presence of each designation in the image, coherence between the labels in such cases remains doubtful. Therefore, these approaches do not specifically define the relationship between the various objects in a query image to better understand its content. Motivated by the successes of computer vision studies in image captioning [44],[45], researchers in the remote sensing community tended to this topic trying to produce a natural language for describing the content of remote sensing images. Lu et al. [4] used CNNs to extract image features and the LSTM network for generating a textual description for the image. Shi et al. [46] proposed a new model by leveraging attribute attention techniques for generating remote sensing images description. Looking at the remote sensing literature, it is evident that remote sensing image captioning has received very limited coverage in this context. There are several datasets for the task of image captioning, such as MS COCO [47], Flicker30K [48]. Yet the remote sensing community lack behind due to the unavailability of suitable datasets that contains remote sensing images and their textual description, exception the first work in [14] that introduced the first dataset for text-image matching in the remote sensing community, which gives five sentences to describe a single image as shown in Fig.7. The lack of a suitable dataset for text-image retrieval task is one of the most important challenges facing researchers in the remote sensing community.



1. River water flowing in a green environment
2. Lines of trees edging the river
3. Bushes covering the area near the river
4. Shadow reflected on the water
5. Trees surrounding the water

1. A row of storage tanks
2. A pipe connected to the storage tanks
3. A line passing by the storage tanks
4. A paved area next to the storage tank
5. Shadow of the storage tanks on the ground

1. A flat surface of a chaparral
2. A dry ground of a chaparral
3. Trees over the chaparral
4. Small bushes among the trees
5. Dry gaps among the trees in a chaparral

1. Buildings around the water
2. Water pools appearing beside the buildings
3. Buildings different in shape
4. Trees among the buildings
5. Vehicles near the trees beside the buildings

1. A curved road among mobile houses
2. Vehicles beside the mobile houses
3. Trees among the mobile houses
4. Trees shadow on the road near the houses
5. Green areas surrounding the houses

1. A view of a wide nursing home
2. A nursing home divided into many wings
3. A paved area next to the nursing home
4. Cars parked in the paved area beside the building
5. A road path leading to the nursing home

1. A swimming pool in a green environment
2. Paved surroundings of a swimming pool
3. Green passages beside the pool
4. Trees next to the pool
5. A green area near the swimming pool

1. Few multicolor basketball courts
2. Basketball courts next to each other
3. A paved area around basketball courts
4. Trees around basketball courts
5. A road path next to the basketball courts

Fig.7 Example images from TextRs dataset.

Table 3- Recent research works in image-text matching.

Challenges	Ref.	Method/Technique
Image-text Representation	[19]	TIMAM
	[32]	A semantic-enhanced image and

		sentence matching model
	[49]	Position focused attention network
Similarity measurement	[50]	KNN-margin loss, Inverted Softmax (I S), and Cross-modal Local Scaling (CSLS)
	[18]	A neighbor-aware network, Intra-attention
	[51]	Bi-directional Spatial-Semantic attention Networks
Building data set	[14]	Building a new dataset for text-image matching task, Average fusion technique

Table 3 shows the recent works in image-text matching which are classified based on the main three challenges image and text representation, similarity measurement, and building data set.

3. Conclusion

Multimodal data retrieval (matching) is a hot topic in recent literature due to its importance in understanding the association between visual and textual data. In the remote sensing community, the mainstream of data retrieval works focuses on unimodal data retrieval (image-image) because dealing with multimodal data retrieval carries out more challenging problems. In this paper, we presented the main three challenges that can be faced by the researchers in the image-text retrieval task.

REFERENCES

[1] Wang, L., Li, Y., & Lazebnik, S. (2018). Learning Two-Branch Neural Networks for Image-Text Matching Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence. 41(2)*, 394–407.

[2] Zhang, Y., & Lu, H. (2018). Deep Cross-Modal Projection Learning for Image-Text Matching. *In Proceedings of the European Conference on Computer Vision (ECCV) (*pp.686-701).

[3] Zhang, F., Du, B., & Zhang, L. (2015). Saliency-Guided Unsupervised Feature Learning for Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing. 53,* 2175–2184.

[4] Lu, X., Wang, B., Zheng, X., & Li, X. (2018). Exploring Models and Data for Remote Sensing Image Caption Generation. *IEEE Transactions on Geoscience and Remote Sensing. 56*, 2183–2195.

[5] Yao, T., Pan, Y., Li, Y., Qiu, Z., & Mei, T. Boosting Image Captioning with Attributes. In Proceedings of the IEEE International Conference on Computer Vision. 4894-4902.

[6] Hua, Y., Yang, Y., & Du, J. (2020). Deep Multi-Modal Metric Learning with Multi-Scale Correlation for Image-Text Retrieval. Electronics, 9(3), 466.

[7] Fu, X., Zhao, Y., Wei, Y., Zhao, Y., & Wei, S. (2019). Rich Features Embedding for Cross-modal Retrieval: A Simple Baseline. IEEE Transactions on Multimedia. 1-1.

[8] Liu, R., Zhao, Y., Wei, S., Zheng, L., & Yang, Y. (2019). Modality-invariant image-text embedding for image-sentence matching. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 15(1), 1-19.

[9] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90.

[10] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[11] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[12] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. 770-778.

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

[14] Abdullah, T., Bazi, Y., Al Rahhal, M. M., Mekhalfi, M. L., Rangarajan, L., & Zuair, M. (2020). TextRS: Deep Bidirectional Triplet Network for Matching Text to Remote Sensing Images. Remote Sensing, 12(3), 405.

[15] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition.7132-7141.

[16] Nam, H., Ha, J. W., & Kim, J. (2017). Dual attention networks for multimodal reasoning and matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 299-307.

[17] Liu, Y., Guo, Y., Bakker, E. M., & Lew, M. S. (2017). Learning a recurrent residual fusion network for multimodal matching. In Proceedings of the IEEE International Conference on Computer Vision (pp.4107-4116).

[18] Liu, C., Mao, Z., Zang, W., & Wang, B. (2019). A neighbor-aware approach for image-text matching. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3970-3974).

[19] Sarafianos, N., Xu, X., & Kakadiaris, I. A. (2019). Adversarial representation learning for text-to-image matching. In Proceedings of the IEEE International Conference on Computer Vision (pp. 5814-5824).

[20] He, Y., Xiang, S., Kang, C., Wang, J., & Pan, C. (2016). Cross-modal retrieval via deep and bidirectional representation learning. IEEE Transactions on Multimedia, 18(7), 1363-1377.

[21] Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 815-823).

[22] Sohn, K. (2016). Improved deep metric learning with multi-class n-pair loss objective. In Advances in neural information processing systems (pp. 1857-1865).

[23] Movshovitz-Attias, Y., Toshev, A., Leung, T. K., Ioffe, S., & Singh, S. (2017). No fuss distance metric learning using proxies. In Proceedings of the IEEE International Conference on Computer Vision (pp. 360-368).

[24] Prabhu, Y., & Varma, M. (2014, August). Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 263-272).

[25] Yen, I. E. H., Huang, X., Ravikumar, P., Zhong, K., & Dhillon, I. (2016, June). Pd-sparse: A primal and dual sparse approach to extreme multiclass and multilabel classification. In International Conference on Machine Learning (pp. 3069-3077).

[26] Hershey, J. R., Chen, Z., Le Roux, J., & Watanabe, S. (2016, March). Deep clustering: Discriminative embeddings for segmentation and separation. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 31-35).

[27] Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 815-823).

[28] Law, M. T., Thome, N., & Cord, M. (2013). Quadruplet-wise image similarity learning. In Proceedings of the IEEE International Conference on Computer Vision (pp. 249-256).

[29] Oh Song, H., Xiang, Y., Jegelka, S., & Savarese, S. (2016). Deep metric learning via lifted structured feature embedding. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4004-4012).

[30] Wang, J., Zhou, F., Wen, S., Liu, X., & Lin, Y. (2017). Deep metric learning with angular loss. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2593-2601).

[31] Huang, Y., Wu, Q., Song, C., & Wang, L. (2018). Learning semantic concepts and order for image and sentence matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6163-6171).

[32] Ma, L., Lu, Z., Shang, L., & Li, H. (2015). Multimodal convolutional neural networks for matching image and sentence. In Proceedings of the IEEE international conference on computer vision (pp. 2623-2631).

[33] Lee, K. H., Chen, X., Hua, G., Hu, H., & He, X. (2018). Stacked cross attention for image-text matching. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 201-216).

[34] Li, Z., Ling, F., Zhang, C., & Ma, H. (2020). Combining Global and Local Similarity for Cross-Media Retrieval. IEEE Access, 8, 21847-21856.

[35] Gao, D., Jin, L., Chen, B., Qiu, M., Li, P., Wei, Y., & Wang, H. (2020, July). Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 2251-2260).

[36] Zheng, Z., Zheng, L., Garrett, M., Yang, Y., Xu, M., & Shen, Y. D. (2020). Dual-Path Convolutional Image-Text Embeddings with Instance Loss. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 16(2), 1-23.

[37] Chen, T., Deng, J., & Luo, J. (2020). Adaptive Offline Quintuplet Loss for Image-Text Matching. arXiv preprint arXiv:2003.03669.

[38] Wu, Y., Wang, S., & Huang, Q. (2017). Online asymmetric similarity learning for cross-modal retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4269-4278).

[39] Liong, V. E., Lu, J., Tan, Y. P., & Zhou, J. (2016). Deep coupled metric learning for cross-modal matching. IEEE Transactions on Multimedia, 19(6), 1234-1244.

[40] Yang, Y., & Newsam, S. (2010, November). 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 (pp. 270-279).

[41] Xia, G. S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., ... & Lu, X. (2017). AID: A benchmark data set for performance evaluation of aerial scene classification. IEEE Transactions on Geoscience and Remote Sensing, 55(7), 3965-3981.

[42] Chaudhuri, B., Demir, B., Chaudhuri, S., & Bruzzone, L. (2017). Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method. IEEE Transactions on Geoscience and Remote Sensing, 56(2), 1144-1158.

[43] Shao, Z., Yang, K., & Zhou, W. (2018). A Benchmark Dataset for Performance Evaluation of Multi-Label Remote Sensing Image Retrieval. Remote Sensing, 10(6).

[44] You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016). Image captioning with semantic attention. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4651-4659).

[45] Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., & Chua, T. S. (2017). Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5659-5667).

[46] Shi, Z., & Zou, Z. (2017). Can a machine generate humanlike language descriptions for a remote sensing image?. IEEE Transactions on Geoscience and Remote Sensing, 55(6), 3623-3634.

[47] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Springer, Cham.

[48] Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics, 2, 67-78.

[49] Wang, Y., Yang, H., Qian, X., Ma, L., Lu, J., Li, B., & Fan, X. (2019). Position focused attention network for image-text matching. arXiv preprint arXiv:1907.09748.

[50] Liu, F., & Ye, R. (2019). A strong and robust baseline for text-image matching. arXiv preprint arXiv:1906.01205.

[51] Huang, F., Zhang, X., Zhao, Z., & Li, Z. (2018). Bi-directional spatial-semantic attention networks for image-text matching. IEEE Transactions on Image Processing, 28(4), 2008-2020.