ATTENTION-DRIVEN CROSS-MODAL REMOTE SENSING IMAGE RETRIEVAL

Ushasi Chaudhuri[1], Biplab Banerjee[1], Avik Bhattacharya[1], Mihai Datcu[2]

[1] Indian Institute of Technology Bombay, Mumbai, India
[2] German Aerospace Center (DLR), Oberpfaffenhofen, Germany.

IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-0369-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IGARSS47720.2021.9554838

ABSTRACT

In this work, we address a cross-modal retrieval problem in remote sensing (RS) data. A cross-modal retrieval problem is more challenging than the conventional uni-modal data retrieval frameworks as it requires learning of two completely different data representations to map onto a shared feature space. For this purpose, we chose a photo-sketch RS database. We exploit the data modality comprising more spatial information (sketch) to extract the other modality features (photo) with cross-attention networks. This sketch-attended photo features are more robust and yield better retrieval results. We validate our proposal by performing experiments on the benchmarked Earth on Canvas dataset. We show a boost in the overall performance in comparison to the existing literature. Besides, we also display the Grad-CAM visualizations of the trained model's weights to highlight the framework's efficacy.

Index Terms— Cross-modal retrieval, Remote Sensing, Sketch-based image retrieval, Attention network, Deep learning.

1. INTRODUCTION

Various sensors scanning over the same region of the earth often captures a variety of information. To obtain this overwhelming and varied information from different imaging sources, we need to have robust cross-modal information mining/data retrieval frameworks. In this regard, recent studies have recognized the importance of developing a cross-modal retrieval framework and have proposed different datasets and framework [1, 2, 3].

Designing a uni-modal retrieval algorithm is a relatively more straightforward problem as it requires fitting the data obtained from only a single sensor. A cross-modal framework requires learning the data distribution from both modalities. A model fit for one particular set of data necessarily pays off in performance for all other sets of data/problems. Therefore arises the need for having a unified framework that is optimized for handling multiple data modalities. Designing such a framework is challenging.

Often there is a lack of available images for querying. The target query image just remains in the user's perception, and no particular image sample of that class is readily available at hand [2]. For such cases, researchers have come up with sketch-based image retrieval wherein the user can instantly picture a rough sketch of the target image and use it as a query. These type of problems also fall under the category of cross-modal image retrieval.

Recent works on cross-modal retrieval in RS includes various attempts by researchers to SAR-optical data [4], image-text [5], image-speech [1], panchromatic-multispectral [6], RGB image-depth [7], photo-sketch [2, 3], etc. Both hand-crafted and deep learning-based feature extractors are applied in conjunction with the RS image retrieval. The ad-hoc image features range from the basic colour moments, texture, shape, and morphological descriptors and combinations of them. The data-driven deep learning techniques have shown excellent performance for retrieval tasks. Several endeavours have directly used the Imagenet pre-trained deep Convnet models like the VGG-16 [8] for RS image retrieval, while others depend on fine-tuning the pre-trained models, for the task at hand [9].

In this work, we propose an extension to the CMIR-Net architecture [1] wherein the framework learns a discriminative shared feature space from the different data modalities. We add an attention network [10] on this to align the sketch and images from two other modalities. The sketch data lacks texture information and just constitutes a minimally representative outline for the target query. So essentially the sketch field consists of only the spatial information. The image data comprises high texture information, which makes the CNNs learn from the micro-textures. However, we also want to enforce the CNNs to learn the images' spatial information for better alignment. We use a sketch attended cross attention network to extract the meaningful features from the images to achieve this. Exploiting the sketch images' spatial information content helps us efficiently capture the important constructs from the photo data and align the two modalities on the embedding space. The results obtained show superior results than a model without attention. We use this motivation to design the proposed framework. Further, we use a Grad-CAM visualization to show the region that receives the most attention for a given class.

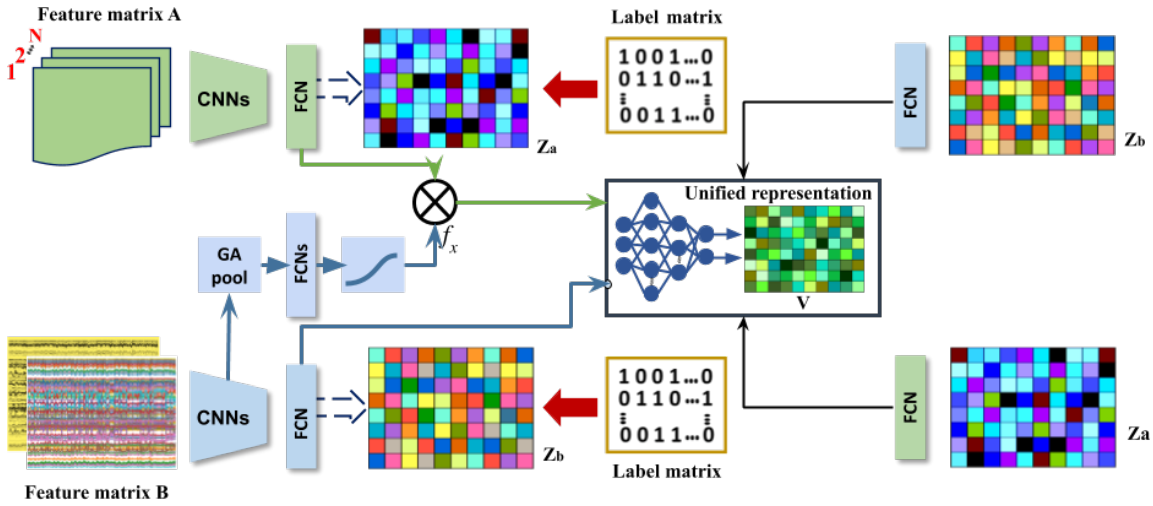


Fig. 1. Overall pipeline of the proposed framework at training phase.

2. METHODOLOGY

Let us consider **A** and **B** as two separate input data modalities, i.e., photo and sketches, having shared labels from **L**. To design a cross-modal retrieval framework, we construct triads comprising of $\mathbf{X} = \{(a_i, b_i, l_i)\}$, where $a_i \in \mathbf{A}$, $b_i \in \mathbf{B}$, and $l_i \in \mathbf{L}$. We aim to learn an unified representation of features of both the modalities in the latent space V wherein we can perform cross-retrieval. All other notations remain consistent with [1].

We aim to project the samples from each modality onto the shared feature space to make them class-wise discriminative while reducing their domain gap. We refer to this shared space features of modalities **A** and **B** as $\mathbf{V}_{a_i}$ and $\mathbf{V}_{b_i}$, corresponding to their instances a_i and b_i, respectively. We train the framework to obtain the shared feature embeddings $\mathbf{V}_{a_i}$ and $\mathbf{V}_{b_i}$ and during inference, use a simple k-nearest neighbour approach to find the top retrieved instances from the concerned modality features.

The proposed network is broadly similar to the architecture in [1] with two encoder streams and two cross-decoder networks. In this framework, we use an end-to-end framework, unlike a two-stage framework in [1]. We use a VGG-16 Imagenet pre-trained network to extract the features to learn the preliminary features from these images. Besides, we also add an attention network to highlight the important constructs of each photo instance. From the output of the last convolution layer of the sketch framework which yields a $7 \times 7 \times 512$ dimensional encoded image feature, we perform a global average pooling, followed by a fully connected layer and then a `Sigmoid(·)` non-linear activation to get the sketch-attention output. The remainder of the framework remains similar to the CMIR-Net architecture. We use similar loss functions as introduced in the CMIR-Net. Figure 1 illustrates the overall pipeline of the proposed architecture.

2.1. Training

We propose an encoder-decoder based framework which derives attention from the sketch network. From the final layer of convolution of the VGG-16 encoder from the sketch network, we obtain a $7 \times 7 \times 512$ dimensional intermediate output corresponding to each sketch instance. We perform a global average pooling on this layer and pass this through a fully-connected network. We finally use a Sigmoid activation layer and multiply the output with the fully connected features obtained from the photo encoder network. This is the attended photo feature that highlights the important constructs in its extracted features, obtained with the sketch network's help.

To bring the sketch and photo modalities closer in the feature space, we minimize the difference between the samples in $\mathbf{V}_a$ and $\mathbf{V}_b$. This helps us make the features of both the modalities analogous, making it domain-agnostic. Let **F** denote the Frobenius norm, then we define this loss simply as:

$$\mathcal{L}_1 = ||\mathbf{V}_a - \mathbf{V}_b||_{\mathbf{F}}^2 \tag{1}$$

To make these embeddings of the sketch and the photo modalities discriminative in space, we use a simple cross-entropy loss function on $\mathbf{V}_{ab} = [\mathbf{V}_a, \mathbf{V}_b]$. We define this loss as:

$$\mathcal{L}_2 = \text{CE}(\mathbf{V}_{ab}) \tag{2}$$

Since we are dealing with hand-drawn sketch instances, the features may vary substantially depending on the data instances. To put a check on any unbounded modulation on the features, we use norm loss on the shared space features of both the modalities. This is represented as:

$$\mathcal{L}_3 = ||\mathbf{V}_a||_{\mathbf{F}}^2 + ||\mathbf{V}_b||_{\mathbf{F}}^2 \tag{3}$$

Finally, a decoder loss from the decoder network to align the features of the two modalities better in the shared feature

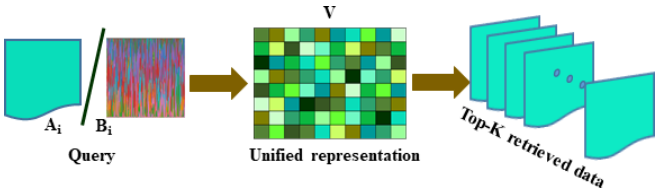
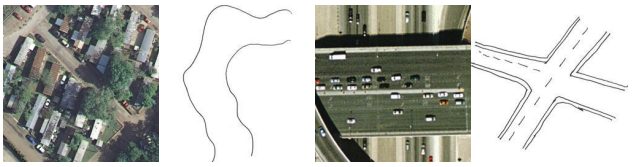


Fig. 2. Retrieval at the inference phase.



(a) Mobile home (b) River (c) Overpass (d) Intersection

Fig. 3. Sample instances of photo and sketch images from the Earth on Canvas dataset.

space and ensure domain-invariance. This loss reconstructs the cross domain samples from their latent feature representations. Let w_{ab} and w_{ba} denote the learnable weights for the decoder blocks. We denote this loss as:

$$\mathcal{L}_4 = ||w_{ab}\mathbf{V}_a - \mathbf{V}_b||_{\mathbf{F}}^2 + ||w_{ba}\mathbf{V}_b - \mathbf{V}_a||_{\mathbf{F}}^2 \quad (4)$$

The overall objective function is given as the sum of the above mentioned loss $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \mathcal{L}_4$. We optimize on the overall loss $\mathcal{L}$ using the standard mini-batch stochastic gradient descent optimizer.

2.2. Inference

After optimizing the loss $\mathcal{L}$, we obtain the shared space feature embeddings of both the sketch and the photo modalities as $V_{(a/b)}$. From this, we use a simple Euclidean distance to find the simple k-nearest neighbours from the embedding of any query photo or sketch. We can retrieve within a modality as well as retrieve from a cross-modal data source.

3. EXPERIMENTS AND RESULTS

In this work, we chose sketch and photo images as two different modalities. For this purpose, we use the Earth on Canvas dataset [3] with a 70:30 train:test split. This dataset consists of 14 classes for object retrieval from photo and sketch modalities. Each modality consists of 100 images per class. Figure 3 shows a few sample images from both the modalities of this dataset.

After training the network, we learn the shared space feature embeddings of both the photo and sketch modalities $\mathbf{V}_a$ and $\mathbf{V}_b$. We obtain the dimensions of $\mathbf{Z}_{a/b}$ in 128-d, and refer to this feature dimension as d_v. For minimizing the overall objective function, we use the stochastic gradient descent optimizer with a learning rate of 0.001 and a batch size of 128. We found a high learning rate to have learning difficulties on the sketch data. The network converges in about 40-50 epochs. For the evaluation, we use the standard mean average precision (mAP) and precision at X (P@X) values.

In table 1 we show the retrieval performance of our model in all the four possible experimental setups. We compare our method to the existing CMIR-Net [1] as this is the only framework which handles cross-modal as well as the uni-modal retrieval in RS to the best of our knowledge (table 2). Comparing the proposed methodology with the existing CMIR-Net

Table 1. Performance of the proposed framework on the Earth on Canvas dataset in terms of mAP (%) and precision at top-10 (P@10) (%) values.

Task	d_v = 128			
	mAP	P@10	P@20	P@50
Sketch→Photo	0.753	0.784	0.765	0.721
Photo→Sketch	0.723	0.745	0.734	0.692
Sketch→Sketch	0.775	0.788	0.765	0.724
Photo→Photo	0.804	0.823	0.805	0.793

helps us clearly highlight the exact contribution of the attention network module compared to the baseline method. Besides, we also show the class-wise average precision for retrieving the top-50 samples for both CMIR-Net and the proposed Att-CMIR-Net in the form of a bar plot in figure 4. From table 2 and figure 4 we can easily see that appending the network with an attention module helps align the two domains better and yields better results.

At a broad level, making any changes to a network might seem random and difficult to understand its exact contribution of the baseline framework.For this purpose, we recognize that the visualization of weights on the image is necessary for interpreting the working of the framework. In this regard, we use the Grad-CAM visualization to highlight the image regions that receive the most attention for a target class label. The Grad-CAM plots use the gradients of any target label flowing into the final convolution layer to produce a gradient mask of the image's size. This mask makes the localization map highlighting the most important regions of the image pertaining to that label. We show the Grad-CAM plots of both photo and the sketch image instances in Fig. 5 and show that after using the attention network, we can highlight the correct object label in the image. The top row shows the photo and the sketch instances from a few random classes. The bottom row shows their corresponding Grad-CAM plots. The red regions denote higher importance in those regions, while the blue signifies the lesser extent to those areas.

While generating the CAM plots for the sketch images, we noted that there was a tendency to show high importance to the borders of the image. This is possibly caused because the sketches were hand-drawn and scanned. Scanning often creates edge along the ends of the pages. These borders in-

Table 2. Comparison of the proposed framework and with the existing literature.

Task	Model	d_v=128	
		mAP	P@10
Sketch→Photo	CMIR-Net [1]	0.732	0.756
	Proposed	**0.753**	**0.784**
Photo→Sketch	CMIR-Net [1]	0.696	0.708
	Proposed	**0.723**	**0.745**

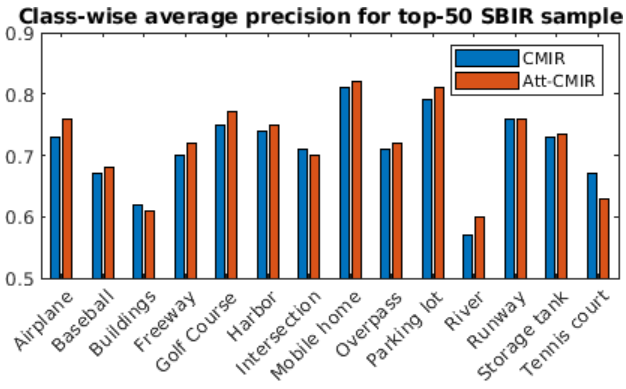
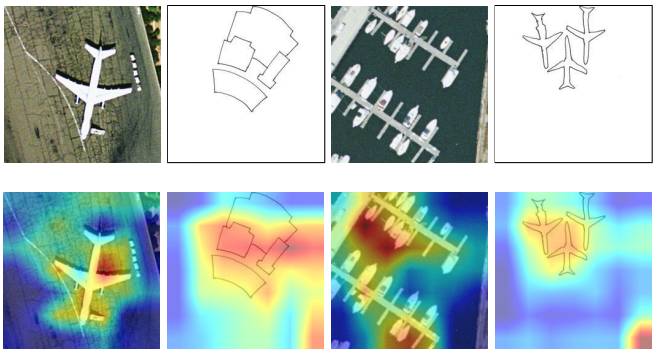


Fig. 4. Class-wise average precision for top-50 retrieval for sketch-based image retrieval.

terfere with the object class and also captures attention. Another important observation is the class-wise precision of 2-3 classes fall after applying the attention module. These are the classes which occupy a large part of the image, and hence the attention is not localized.

4. CONCLUSION

We propose an improved framework over the existing CMIR-Net for cross-modal retrieval of data. The framework mainly exploits one modality's attention and uses it to capture the other modality's important construct. The proposed framework beats the existing state-of-the-art, in addition to supporting a uni-modal retrieval framework too. We validate our claim by providing experimental results on aerial photo-sketch data. We further show a Grad-CAM visualization to show more insights to understand what is happening inside the network. We are currently excited to use this motivation and see how it works for SAR - optical data. One of SAR data's main challenges in machine learning is its lack of similarity with the natural image statistics. Therefore we do not get any benefit by using any of the Imagenet pre-trained encoder models. However, if we can exploit the attention from its corresponding optical data, we might get decent performance.



(a) Airplane (b) Building (c) Harbour (d) Airplane

Fig. 5. Grad-CAM visualization of photo and sketch image instances.

5. REFERENCES

[1] U Chaudhuri, B Banerjee, A Bhattacharya, and M Datcu, "Cmir-net: A deep learning based model for cross-modal retrieval in remote sensing," *Pattern Recognit. Lett.*, vol. 131, pp. 456–462, 2020.

[2] F. Xu, W. Yang, T. Jiang, S. Lin, H. Luo, and G. Xia, "Mental retrieval of remote sensing images via adversarial sketch-image feature learning," *IEEE Trans. Geosci. Remote Sens.*, pp. 1–14, 2020.

[3] U Chaudhuri, B Banerjee, A Bhattacharya, and M Datcu, "A zero-shot sketch-based inter-modal object retrieval scheme for remote sensing images," *arXiv:2008.05225*, 2020.

[4] M Schmitt, C Hughes, Land Qiu, and X Zhu, "Sen12ms- a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion," *arXiv:1906.07789*, 2019.

[5] T Abdullah, Y Bazi, M Al Rahhal, M Mekhalfi, and M Rangarajan, Land Zuair, "Textrs: Deep bidirectional triplet network for matching text to remote sensing images," *Remote Sensing*, vol. 12, no. 3, pp. 405, 2020.

[6] Y Li, Yand Zhang, X Huang, and J Ma, "Learning source-invariant deep hashing convolutional neural networks for cross-source remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, , no. 99, pp. 1–16, 2018.

[7] D Eigen, C Puhrsch, and R Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *NIPS*, 2014, pp. 2366–2374.

[8] K Simonyan and A Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014.

[9] D Mandal, K Chaudhury, and S Biswas, "Generalized semantic preserving hashing for n-label cross-modal retrieval," in *CVPR*, 2017, pp. 4076–4084.

[10] S. Mohla, S. Pande, B. Banerjee, and S. Chaudhuri, "Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification," in *CVPRW*, 2020, pp. 416–425.