Attention-driven Cross-Modal Remote Sensing Image Retrieval

Ushasi Chaudhuri[a], Biplab Banerjee[a], Avik Bhattacharya[a],
&
Mihai Datcu[b]

[a]Indian Institute of Technology Bombay, India
[b]German Aerospace Center (DLR), Germany.

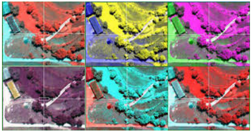
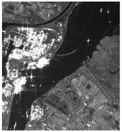
July, 2021

Objective

- Due to availability of wide range of satellite sensors, accumulation of an unprecedented volume of remote sensing images.
- Information to describe X can come from multiple modalities (image, speech, sketch, information from various specral bands, etc.).
- Develop a retrieval technique that can handle cross-modal information.
- Within the same dataset, retrieval is a relatively easy work. If we have different modalities of data, ex Pan images, multi-spectral images, SAR images, optical images, or texts, retrieval amongst cross domain becomes more challenging.

Motivation

- **PAN images:** High spatial resolution.
- **Multi-spectral images:** high spectral resolution.
- **SAR images:** Polarization information. No cloud clutter.
- **VHR optical image:** High spatial res. + 3 spectral channel.

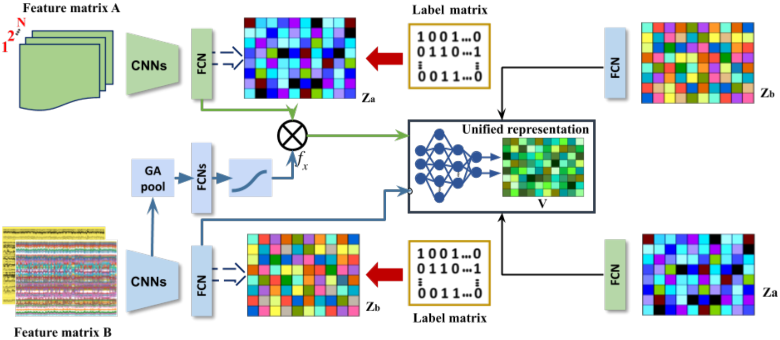

1. 1m pan IKONOS image of Valparaiso, Chile 2. source: https://www.sensorsmag.com/components/hyperspectral-and-multispectral-imaging-sensors-find-hyper-flexible-applications 3. Sentinel-1 SAR dataset with C band4. PatternNet dataset.

- Important problems:
  1. PAN $\Leftrightarrow$ Multispectral.
  2. Optical $\Leftrightarrow$ SAR.
  3. RGB $\Leftrightarrow$ DEM.
  4. Image $\Leftrightarrow$ text.
  5. **Image $\Leftrightarrow$ sketch.**

Cross-modal Retreival

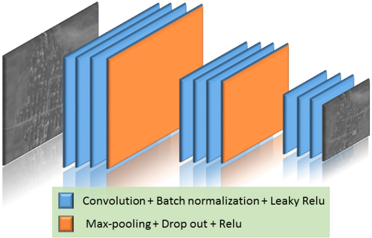
Datasets used:

1. **Earth on Canvas*:** 14 image and sketch classes.



U. Chaudhuri, B. Banerjee, A. Bhattacharya, M. Datcu, "CMIR-NET : A deep learning based model for cross-modal retrieval in remote sensing", *Pattern Recognition Letters* (PRL), volume 131, pp 456-462, 2020.

Cross-modal Retreival



- Train 2 separate classification networks $\{(a_k, l_k)\}_{k=1}^{|\mathbf{A}|}$ and $\{(b_j, l_j)\}_{j=1}^{|\mathbf{B}|}$.
- Extracted features ($\mathbf{Z}_{a_k}$ and $\mathbf{Z}_{b_j}$) are made highly non-redundant by adding a soft orthogonality constrained.

$$\mathcal{L}_{\mathbf{A}/\mathbf{B}} = \mathrm{CE}(\tilde{\mathbf{Z}}_{a/b}) + ||\tilde{\mathbf{Z}}_{a/b}^{T}\tilde{\mathbf{Z}}_{a/b} - \mathbf{I}||_{\mathbf{F}}^{2}$$

- $\{\mathbf{Z}_{a_i}\}$ and $\{\mathbf{Z}_{b_i}\}$ are considered as inputs for obtaining V.

Objective Function

- 1. **Difference** between each pair of corresponding i^{th} samples in $\mathbf{V}_a$ and $\mathbf{V}_b$ ($\mathcal{L}_2$):

$$\mathcal{L}_2 = ||\mathbf{V}_a - \mathbf{V}_b||_{\mathbf{F}}^2$$

- 2. **Classification** loss on $\mathbf{V}_{ab} = [\mathbf{V}_a, \mathbf{V}_b]$ ($\mathcal{L}_3$):

$$\mathcal{L}_3 = \mathrm{CE}(\mathbf{V}_{ab})$$

- 3. Separate **feature norm** loss measures on both $\mathbf{V}_a$ and $\mathbf{V}_b$, ($\mathcal{L}_4$) (Since the range of values of raw data features varies widely):

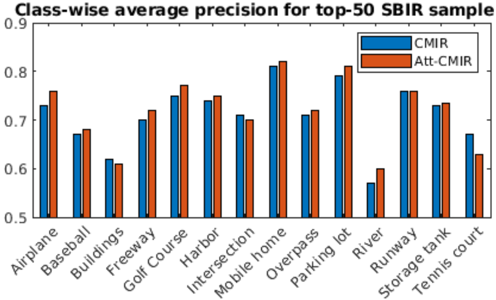
$$\mathcal{L}_4 = ||\mathbf{V}_a||_{\mathbf{F}}^2 + ||\mathbf{V}_b||_{\mathbf{F}}^2$$

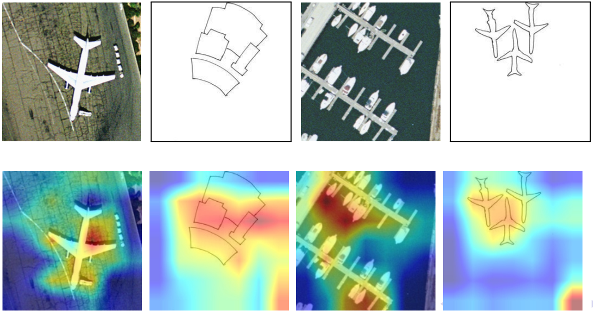
- 4. **Decoder** loss which is deemed to reconstruct cross-domain samples given the latent representations: ($\mathcal{L}_5$):

Comparison with the existing literature.

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

Attention-aware Cross-modal Retreival





Conclusions

- Developed an encoder-decoder based cross-modal retrieval framework, robust over various types of input data.
- Appended it with cross-attention network to extract more representative feature embeddings.
- Proposed framework outperformed the existing state-of-the-art method.
- Verified with the newly proposed Earth on Canvas dataset.

Thank You