CMIR-NET : A Deep Learning Based Model For Cross-Modal Data Retrieval

Ushasi Chaudhuri

Ph.D. student, Under the guidance of
Avik Bhattacharya, Biplab Banerjee & Mihai Datcu (DLR, Germany)

Indian Institute of Technology, Bombay (IIT-B)

December 2020
IEEE GRSS Young Researchers Conclave.

Content Based Image Retrieval:

- **Availability** of wide range of satellite sensors: accumulation of an unprecedented volume of data.
- Necessity of sophisticated **information extraction strategies**.
- **Image retrieval** aims to retrieve a number of visually coherent images from a query.
- **Uni-modal** retrieval is a relatively easy work. **Cross-modal** is challenging.

A network learned for a particular modality may not give good performance on different data.

Definition (No free lunch theorem)

[a] If an algorithm performs well on a certain class of problems then it necessarily pays for that with degraded performance on the set of all remaining problems.

[a] Wolpert, D.H., Macready, W.G. (1997), "No Free Lunch Theorems for Optimization", IEEE Transactions on Evolutionary Computation 1, 67.

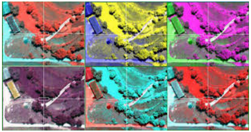
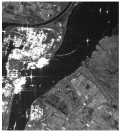
Motivation - TPAMI 2020

Why limit CBIR to just one data stream?

Advantages of each modality:

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

Can be an ill-posed problem.

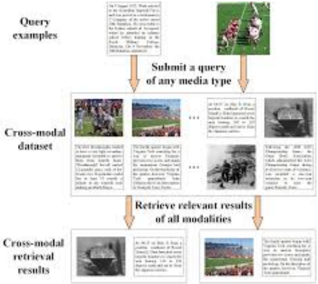

1. 1m pan IKONOS image of Valparaiso, Chile, 2. multispectral-imaging-sensors
3. Sentinel-1 SAR dataset with C band, 4. `sparseresidentialarea` from PatternNet dataset.

Motivation

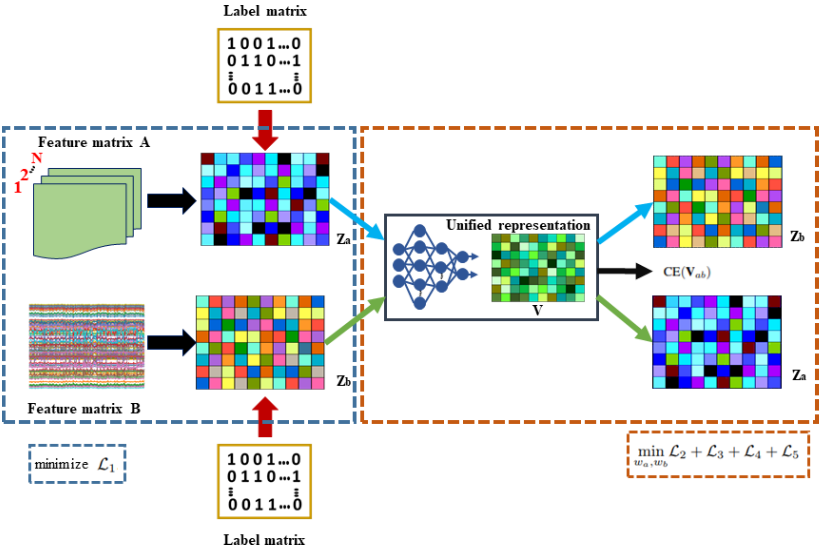
Why not try this in RS with so many available sensors?

- Important problems:
  1. Panchromatic ⇔ Multispectral.
  2. Optical ⇔ Synthetic Aperture Radar (SAR).
  3. RGB ⇔ Digital elevation maps/LiDAR point clouds.
  4. Image ⇔ text.
  5. Image ⇔ audio.

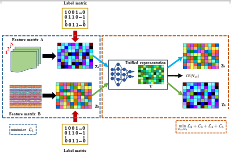
Upcoming: VideoSAR: text and audio annotation

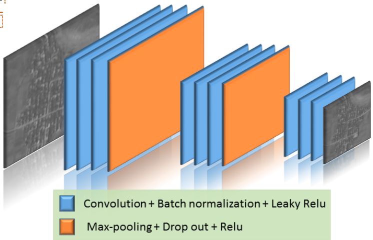


Overall Block Diagram



Pre-training

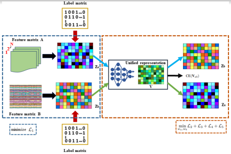




- Train 2 separate classification n/w for **A** and **B** data.
- Extracted features ($\mathbf{Z}_a$ and $\mathbf{Z}_b$) are made highly non-redundant by adding a soft orthogonality constrained.

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

Construction of V



To construct $\{\mathbf{V}_{a_i}\}$ and $\{\mathbf{V}_{b_i}\}$ from $\{\mathbf{Z}_{a_i}\}$ and $\{\mathbf{Z}_{b_i}\}$, we use a neural network based discriminative encoder-decoder architecture which minimizes:

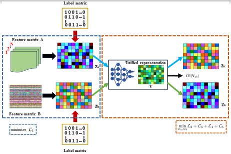
- 1. **Difference** between each pair of corresponding i^{th} samples in $\mathbf{V}_a$ and $\mathbf{V}_b$:

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

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

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

Construction of V



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

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

- 4. **Decoder** loss which is deemed to reconstruct cross-domain samples given the latent representations:

$$\mathcal{L}_5 = ||w_{ab}\mathbf{V}_a - \mathbf{Z}_b||_\mathrm{F}^2 + ||w_{ba}\mathbf{V}_b - \mathbf{Z}_a||_\mathrm{F}^2$$

where $\mathbf{Z}_{a/b} = \{Z_{a_i/b_i}\}$.

Loss Functions

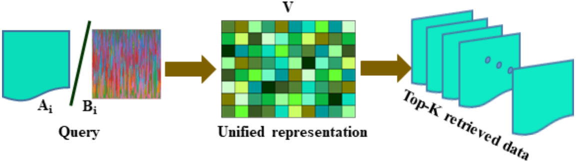
Overall equation:

$$\mathcal{L} = \lambda_1 \mathcal{L}_2 + \lambda_2 \mathcal{L}_3 + \lambda_3 \mathcal{L}_4 + \lambda_4 \mathcal{L}_5 + \lambda_5 \mathcal{R}$$

Where, given the non-negative weights λ_{1-5} and $\mathcal{R}$ defines the standard ℓ_2 regularizer on w_a and w_b.

$$\mathcal{R} = ||w_a - \alpha||_F^2 + ||w_b - \alpha||_F^2$$

for $\alpha \geq 0$. (**To avoid:** Trivial solution)



Algorithm

Algorithm 1 The proposed training and inference stage

Input: $\{(a_k, l_k)\}$, $\{(b_j, l_j)\}$, and $\mathbf{X}$

Output: Unified representations $\mathbf{V}_{a/b}$ ($w_a\mathbf{Z}_a$ and $w_b\mathbf{Z}_b$).

1: Normalize $\mathbf{A}$ and $\mathbf{B}$.

2: Generate intermediate representations $\{\mathbf{Z}_{a_k}\}$ and $\{\mathbf{Z}_{b_j}\}$ by minimizing $\mathcal{L}_{\mathbf{A}/\mathbf{B}}$.

3: Train the network to obtain V by optimizing $\mathcal{L}$. The optimization follows the following stages:

4: **do**

5:

$$\min_{w_a, w_b} \lambda_1 \mathcal{L}_2 + \lambda_2 \mathcal{L}_3 + \lambda_3 \mathcal{L}_4 + \lambda_4 \mathcal{L}_5 + \lambda_5 \mathcal{R} \tag{8}$$

6: **while** until convergence

7: **return** w_a and w_b (for projecting data onto V)

Input: $a \in \mathbf{A}$ or $b \in \mathbf{B}$

Output: Top-K retrieved data.

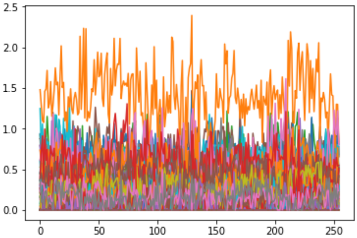
8: Uni-modal retrieval using K-NN from $w_a\mathbf{Z}_a$ **or** $w_b\mathbf{Z}_b$.

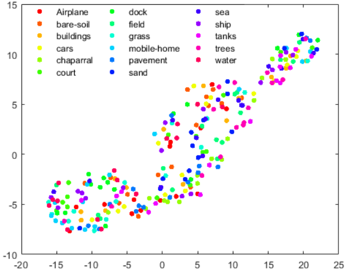
9: Cross-modal retrieval using K-NN from $w_a\mathbf{Z}_a$ **and** $w_b\mathbf{Z}_b$.

Cross-modal data

Datasets used:

1. DSRSID: Panchromatic - Multi-spectral (GF-1 satellite). **(Single-label, paired)**.
2. UCMerced: VHR RS images (RGB) - speech. **(Multi-label, un-paired)**.





Results

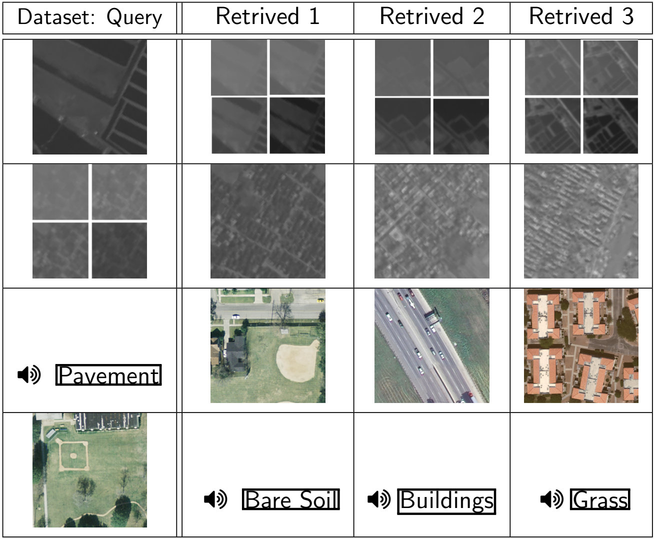


Figure: Top-3 retrieval instances from cross-modal query samples.

1. Results on DSRSID dataset

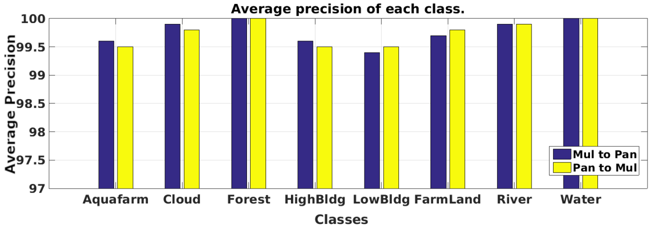


Table: Performance on the DSRSID dataset, under different embedding vector code lengths (d_v).

Task	d_v=16		$d_v = 32$		$d_v = 64$	
	mAP	P@10	mAP	P@10	mAP	P@10
Pan→Mul	95.52	97.10	98.96	98.99	99.05	99.40
Mul→Pan	98.77	99.00	97.95	97.99	98.93	98.60
Pan→Pan	99.41	99.82	98.11	98.40	98.69	99.40
Mul→Mul	99.55	99.69	98.18	98.60	98.25	98.40

2. Results on UC Merced-Speech dataset

Table: Performance of the CMIR-NET framework on UC Merced-Audio dataset, with variation in embedding vector code length (d_v).

Model	d_v=32		$d_v = 64$		$d_v = 128$	
	mAP	P@10	mAP	P@10	mAP	P@10
Img→Aud	29.67	60.91	41.60	63.15	**62.11**	**64.81**
Aud→Img	21.60	40.11	42.36	51.29	**54.21**	**56.00**

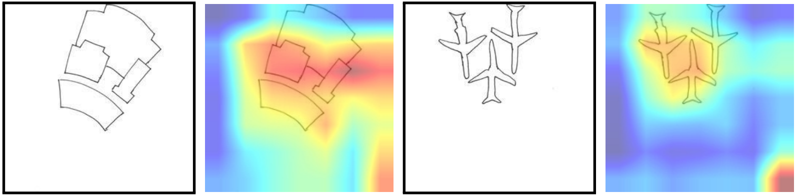
$$\mathcal{R} = ||w_a - \alpha||_F^2 + ||w_b - \alpha||_F^2$$

Table: Sensitivity to critical parameter α for the UC-Merced↔ Audio.

Model	$\alpha = 0$		$\alpha = 1$		$\alpha = 2$	
	mAP	P@10	mAP	P@10	mAP	P@10
Img→Aud	0	0	**62.11**	**64.81**	32.09	52.77
Aud→Img	0	0	**54.21**	**56.00**	33.64	45.01

Conclusions

- Novel framework for cross-modal information retrieval.
- Framework focuses on learning a **unified and discriminative embedding space** from different input modalities.
- Generic enough to handle both uni-modal and cross-modal retrieval.
- **Future work:** Self-supervised CMIR or zero-shot CMIR or Sketch-based IR?



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* 131 (2020): 456-462.


THANK YOU!