Zero-Shot Sketch-Based Object Retrieval for Remote Sensing Images

Ushasi Chaudhuri

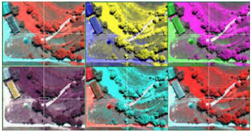
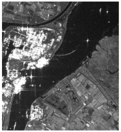
IEEE GRSS Talkathon 2021

Center of Studies in Resources Engineering
Indian Institute of Technology Bombay

28 Aug 2021

Cross-modal Retreival

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


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 band
4. `sparseresidentialarea` from PatternNet dataset.

Motivation

Why do we need Zero-Shot Learning?

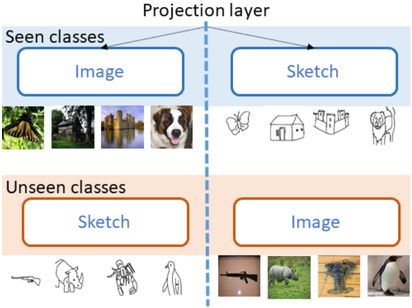
- Necessary to determine a certain number of object classes for object recognition with high success.
- Necessary to collect as many sample images as possible for object classes.
- Exists lots of object classes that we can gather sample images.
- Also exists cases that we are not always so lucky.

For the first time in history Tornado Hits India in State of West Bengal On 25 May 2021.

Zero-shot Learning?

Definition (Zero-shot learning)

Zero-Shot learning method aims to solve a task without receiving any example of that task at training phase.



- **Upon the unavailability of a query sample, can I use a quick handmade sketch query?**

Semantic information

No samples from ZS classes to be used during training.

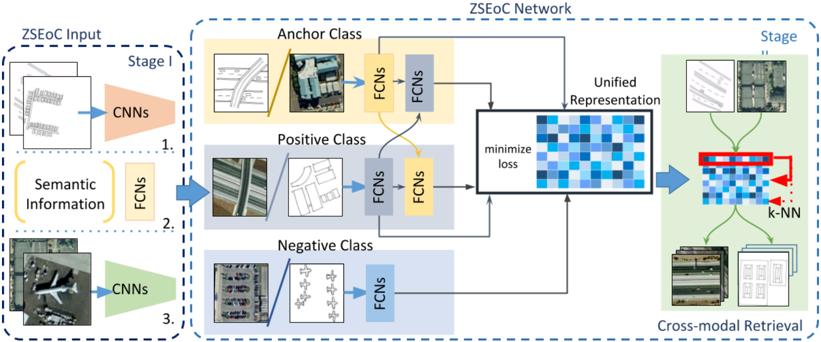
How does a Indian kid, who has never seen a zebra, recognize one in the zoo?

Class Description!!

Therefore, We need 2 data representations:

1. Image: Photos and sketches.
2. Class descriptions as auxiliary semantic information.

Proposed ZS-SBIR Framework



- Images - $\mathcal{A}$, Sketches - $\mathcal{B}$; Constraint: $\mathcal{Y}^s \cap \mathcal{Y}^u = \varnothing$
- Modality-specific classifiers (fine-tune Imagenet pre-trained CNN).
- Multi-stream encoder-decoder n/w for visual-semantic mapping.

ZS-SBIR

- **a) Visual Encoders:** $f_A(;,\theta_A)$ and $f_B(;,\theta_B)$
- **b) Semantic Encoders:** $f_Z(;,\theta_Z)$
- **c) Cross-Modal Decoders:** $g_{AB}(;\theta_{AB})$ and $g_{BA}(;,\theta_{BA})$ which reconstructs $\mathcal{A}$ given $f_B(\mathcal{B})$ and vice-versa.

★ **1. Cross-modal latent loss:** Reduces intra-class variance.

$$\mathcal{L}_{cmd} = ||f_A(\mathcal{A}_c) - f_Z(\mathcal{Z}_c)||^2 + ||f_B(\mathcal{B}_c) - f_Z(\mathcal{Z}_c)||^2$$

★ **2. Cross-modal triplet loss:** Further reduce the intra-class distances, and increase the inter-class distances.

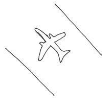
$$\mathcal{L}_{si} = \max\left(d(\mathrm{f}_A\left(\mathcal{A}_c\right), \mathrm{f}_B\left(\mathcal{B}_c\right)) - d(\mathrm{f}_A\left(\mathcal{A}_c\right), \mathrm{f}_B\left(\tilde{\mathcal{B}}_c\right)) + \alpha, 0\right)$$

★ **3. Classification loss:** Class-wise discernibility.

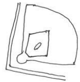
$$\mathcal{L}_{class} = \mathrm{CE}\left(f_A(\mathcal{A})\right) + \mathrm{CE}\left(f_B(\mathcal{B})\right)$$

Experiments — Datasets

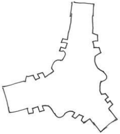
- **Earth on Canvas:** 14 classes $\times$ 100 samples. 10:4 train:test.



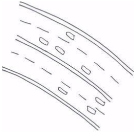
(a) Airplane



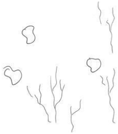
(b) Baseball diamond court



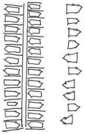
(c) Buildings



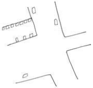
(d) Freeway



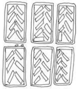
(e) Golf course



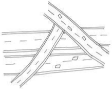
(f) Harbour



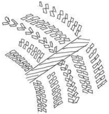
(g) Intersection



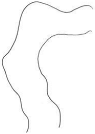
(h) Mobile home park



(i) Overpass

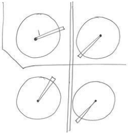


(j) Parking lot

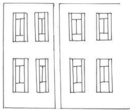


(k) River

(l) Runway



(m) Storage



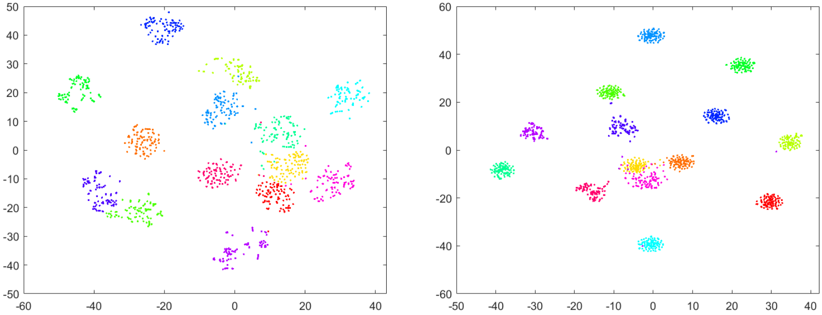
(n) Tennis

Results

Table: SBIR performance of the proposed ZSEoC framework on the **EoC** dataset in terms of mAP (%) and precision at top-100 (P@100) (%) values.

Task	**EoC**		
	mAP	**P@100**	**Feature dimension**
Baseline-I (VggNet-16)	0.221	0.234	4096
Baseline-II (ResNet-50)	0.236	0.254	2048
Baseline-III (ResNet-101)	0.269	0.284	2048
Baseline-IV (CNN)	0.30	0.284	128
Baseline-V (Pre-train + CNN)	0.196	0.284	128
ZS-SBIR [**kiran2018zero**]	0.395	0.421	1024
ZSIH (binary) [**shen2018zero**]	0.452	0.487	64
ZSEoC-300 (fixed semantic vector)	**0.686**	**0.698**	300
ZSEoC-128 (latent semantic vector)	**0.674**	**0.732**	128

Results — *t*-SNE Plots



(a) Image (b) Sketch

Figure: Two-dimensional scatter plots of high-dimensional features generated with *t*-SNE of image and sketch features, in the shared latent space, trained with a fixed-semantic vector. Clusters with distinct colours denote separate classes in the dataset.

Results — Retrieval on Unseen Classes

Inter-modal	EoC		Uni-modal	EoC	
	mAP	P@100		mAP	P@100
Sketch→Image	0.686	0.698	Sketch→Sketch	0.719	0.737
Image→Sketch	0.612	0.632	Image→Image	0.839	0.855

Thank You