A Simplified Framework for Zero-shot Cross-Modal Sketch Data Retrieval

Ushasi Chaudhuri, Biplab Banerjee, Avik Bhattacharya, & Mihai Datcu

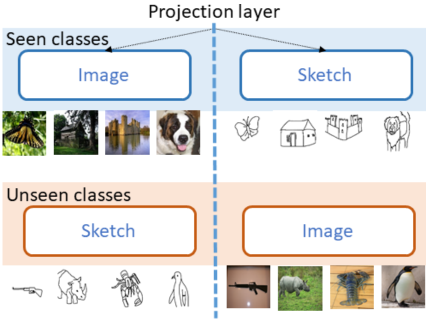
Center of Studies in Resources Engineering (CSRE)
IIT Bombay, India
&
German Aerospace Center (DLR),
Germany

June 2020

Introduction

Definition (Zero-shot learning)

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



No training samples of a few class. Only tested on these classes.

Motivation

Why do we need Zero-Shot Learning?

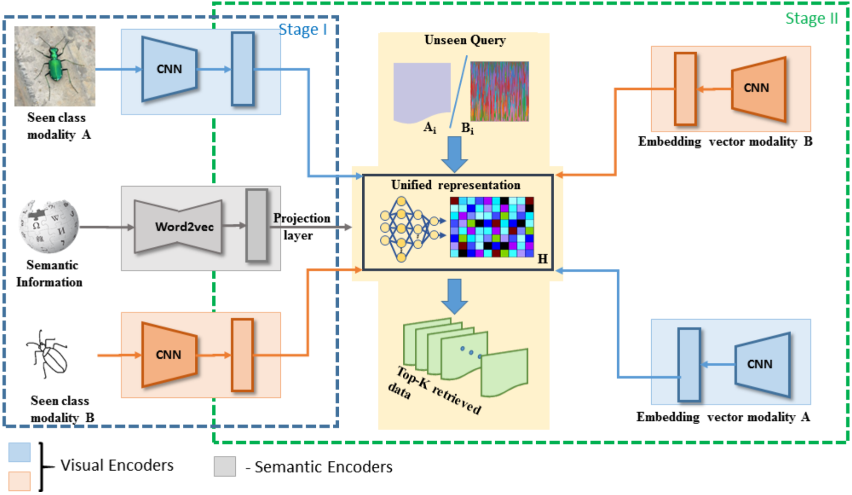
- Necessary to collect as many sample images as possible for object classes.
- Exists cases that we are not always so lucky.
- In such cases, we can quickly draw a sketch as a query ($\therefore$ SBIR sketch-based image retrieval).

Source: https://medium.com/@cetinsamet/zero-shot-learning-53080995d45f

This Ili pika was seen last summer in China's Tianshan Mountains.

Methodology

The Overall pipeline of the proposed **ZSCMR** network.



Loss Functions

1. **Cross-modal latent loss ($\mathcal{L}_{cmd}$):**
   - Bring closer Image (Img) and Sketch (Skc), w.r.t Semantic vector (Sem).
   - Reduces the cross-modal intra-class variance

$$\mathcal{L}_{cmd} = ||\text{Img} - \text{Sem}||_{\mathbf{F}}^2 + ||\text{Skc} - Sem||_{\mathbf{F}}^2$$

2. **Cross-modal triplet loss ($\mathcal{L}_{3lt}$):**
   - Reduce intra-class distances, & increase inter-class distances.
   - Image-anchored triplets:

$$\mathcal{L}_{si} = \max\left(d(\text{Img}, \text{Skc}) - d(\text{Img}, \tilde{\text{Skc}}) + \alpha, 0\right)$$

   - Sketch-anchored triplets:

$$\mathcal{L}_{is} = \max\left(d(\text{Skc}, \text{Img}) - d(\text{Skc}, \tilde{\text{Img}}) + \alpha, 0\right)$$

Overall Objective Functions

3. **Decoder loss ($\mathcal{L}_{rcs}$):**
   - Reduces distributions-gap between Img and Skc.
   - Helps achieve domain-independence.
   - Given Img instance, we reconstruct corresponding Skc.

$$\mathcal{L}_{rcs} = ||g_{is}(\mathrm{Img}) - \mathrm{Skc}||_{\mathbf{F}}^2 + ||g_{si}(\mathrm{Skc}) - \mathrm{Img}||_{\mathbf{F}}^2$$

4. **Classification loss ($\mathcal{L}_{class}$):**
   - Preserves the class information in the shared space
   - Reduces the cross-modal intra-class variance

$$\mathcal{L}_{class} = \mathrm{CE}\,(\mathrm{Img}) + \mathrm{CE}\,(\mathrm{Skc})$$

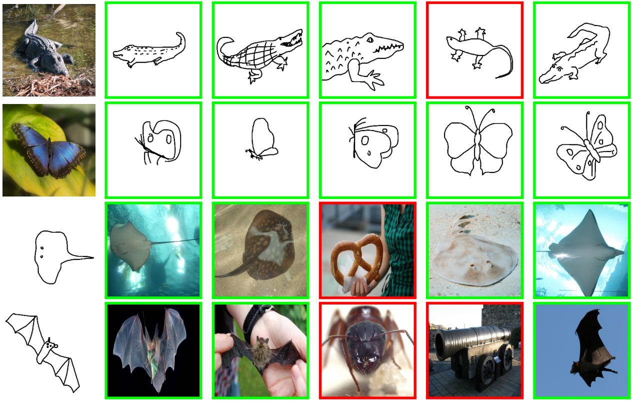
Overall Objective function ($\mathcal{L}$):

$$\mathcal{L} = \mathcal{L}_{cmd} + \mathcal{L}_{rcs} + \mathcal{L}_{3lt} + \mathcal{L}_{class}$$

Results

Datasets used:

- **Sketchy:** 12,500 photos; 75,471 sketches; Train: 100 classes; Test: 25 classes.
- **TU-Berlin:** 204,489 photos; 20,000 sketches; Train: 220 classes; Test: 30 classes.

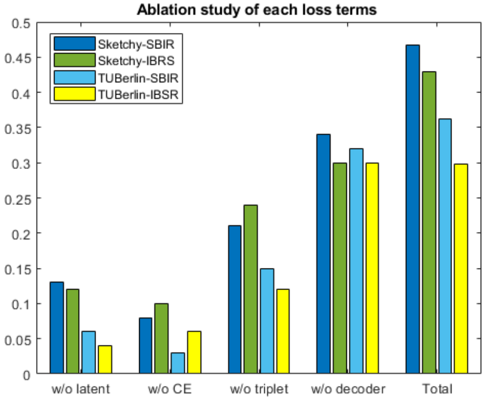


Comparison with the SOTA (Sketch→Image)

	Task	Sketchy		TU-Berlin		size
		mAP	P@100	MAP	P@100	
SBIR	Siamese CNN	0.183	0.143	0.153	0.122	64
	SaN	0.129	0.104	0.112	0.096	512
	3D Shape	0.070	0.062	0.063	0.057	64
	DSH (Binary)	0.171	0.231	0.129	0.189	64
	GDH (Binary)	0.187	0.295	0.135	0.212	64
	GN Triplet	0.204	0.296	0.175	0.253	1024
ZSL	SSE	0.154	0.108	0.133	0.096	100
	JLSE	0.131	0.185	0.109	0.155	220
	ZSH	0.159	0.214	0.141	0.177	64
	SAE	0.216	0.293	0.167	0.221	300
ZSL:SBIR	ZS-SBIR	0.196	0.284	0.005	0.001	1024
	ZSIH (Binary)	0.258	0.342	0.223	0.294	64
	EMS	-	-	0.259	0.369	512
	EMS (Binary)	-	-	0.165	0.252	64
	CAAE	0.196	0.284	-	-	4096
	CVAE	0.225	0.333	-	-	4096
	SEM-PCYC	0.349	0.463	0.297	0.426	64
	SAKE	0.364	0.487	0.359	**0.481**	64
	ZSCMR	**0.467**	**0.510**	**0.362**	0.429	64

Cross-modal retrieval & Ablation study

Task	Sketchy		TU-Berlin	
	mAP	P@100	MAP	P@100
Sketch→Image	0.467	0.510	0.362	0.429
Image→Sketch	0.429	0.451	0.298	0.353
Sketch→Sketch	0.444	0.559	0.318	0.397
Image→Image	0.686	0.718	0.605	0.653



Conclusion

- The main motive of our problem statement is to project different domain data onto a common discriminative embedding space.
- Model has been extended for zero-shot architecture.
- Simple encoder-decoder based architecture in place of generative frameworks.
- The proposed framework shows a significant boost to the current state-of-the-art in ZS:SBIR.
- Additionally, allows image-based sketch retrieval (IBSR) & uni-modal data retrieval.

Thank You