Adaptive Fine-Grained Sketch-Based Image Retrieval

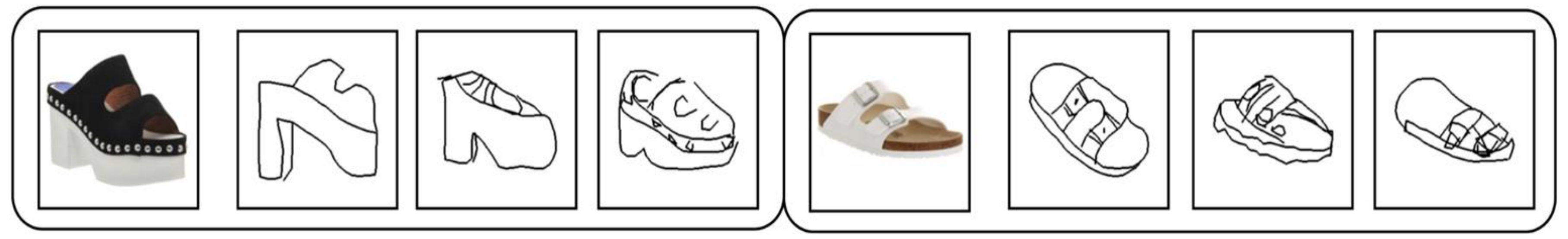
Ayan Kumar Bhunia[1], Aneeshan Sain[1,2], Parth Hiren Shah*, Animesh Gupta*,
Pinaki Nath Chowdhury[1,2], Tao Xiang[1,2], Yi-Zhe Song[1,2]

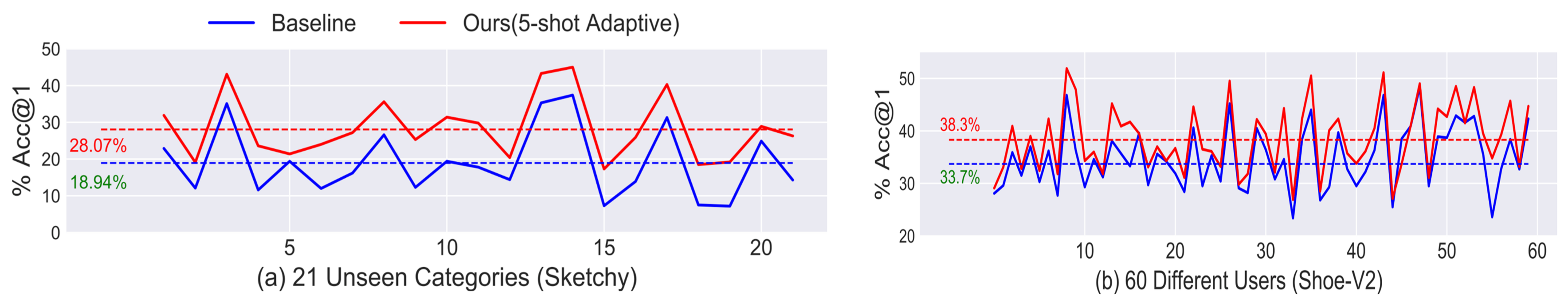
[1]SketchX, CVSSP, University of Surrey, UK; [2]iFlyTek-Surrey Joint Research Centre on Artificial Intelligence; *Intern at SketchX

EUROPEAN CONFERENCE ON COMPUTER VISION
TEL AVIV 2022
October 23-27, 2022

Overview :

- The recent focus in Fine-Grained Sketch-Based Image Retrieval, has been shifted to generalize, a model to new categories without any new training data.
- However, a trained retrieval model faces issues in real-world applications:
  - New categories with no sketch photo pairs.
  - Different drawing styles of different.
- Model Agnostic Meta Learning is a suitable option; and quite realistic as it leverages only a few examples to quickly adapt to new categories and drawing styles.
- A major issue in here is to solve heavy computation which occurs due to second order gradients.
- Also, optimal margin value in triplet loss varies for different categories.
- Can we learn the margin value on the fly for different categories?





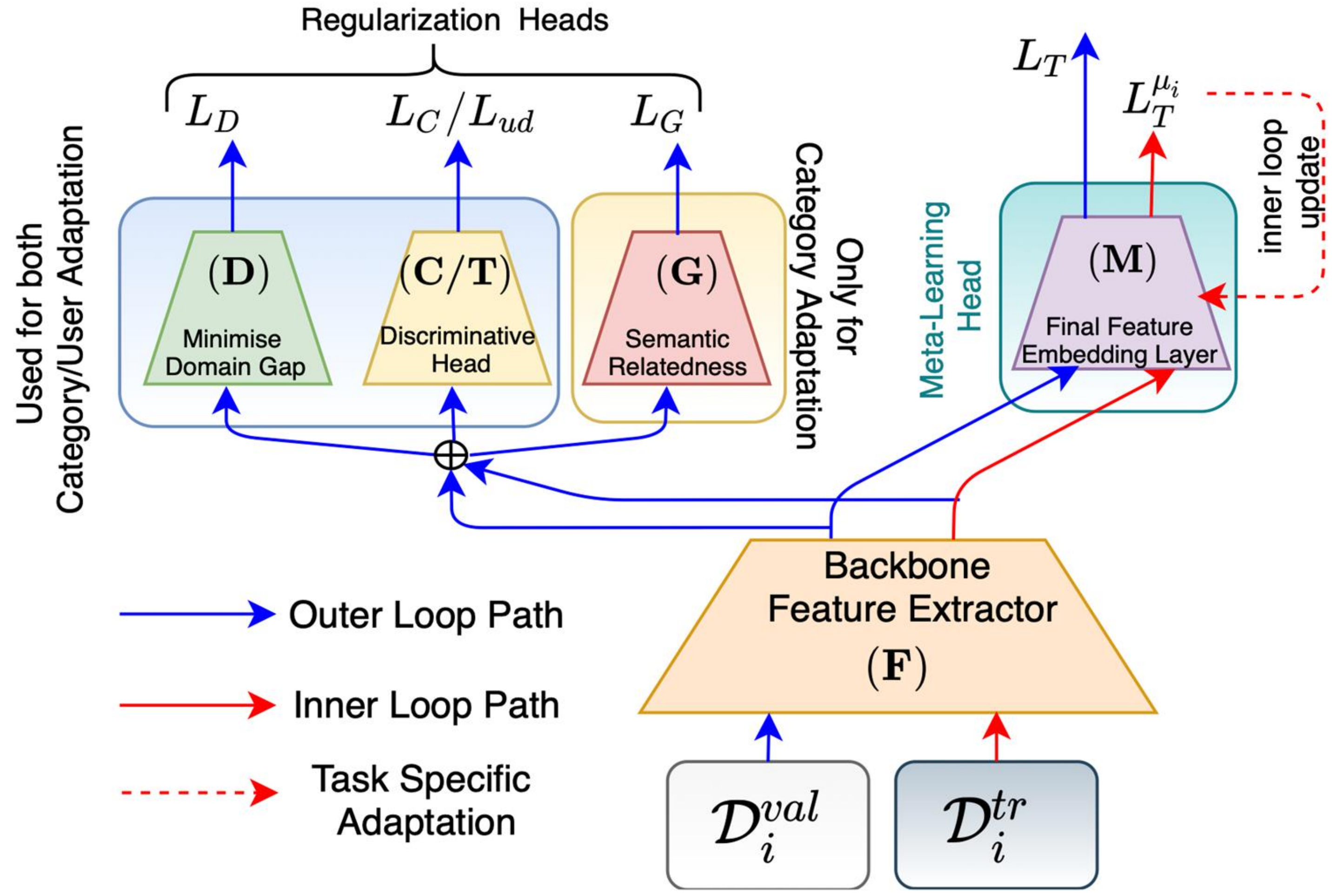
(a) 21 Unseen Categories (Sketchy)
(b) 60 Different Users (Shoe-V2)

- We therefore extend meta learning research even further towards practicality and human-likeness.
- We solved the heavy computation problem by only doing inner loop update on the final joint-feature embedding layer.
- We introduce the learning to learn concept, and used meta learning to learn the margin value used in the triplet loss on the fly.

Please visit https://ayankumarbhunia.github.io for more.

Proposed Model:

- **Objectives:**
  - Quickly adapt the model to new categories and different drawing styles.
  - Meta learn the margin value.
  - Reduce domain gap between sketch and photo images.



- **Training Methods:**
  - *Feature Extractor F*:
    - Train a feature extractor which uses a Siamese network with spatial attention.
  - *Meta Learning Head M*:
    - Features are passed to a fully connected layer, followed by a l2 normalization to embed the photo and sketch images into a shared embedding space.

$$L_T = \frac{1}{N}\sum_{i=1}^{N} max\{0, \mu + \beta_i^{+} - \beta_i^{-}\}.$$

$$L_D = t\cdot\log(D(F(I))) + (1-t)\cdot\log(1-D(F(I)))$$

$$L_C = \mathtt{Cross\_Entropy}(c_I, \mathtt{softmax}(C(F(I))))$$

$$L_{ud} = \max\{0, \beta'^{+} - \beta'^{-} + \mu'\}.$$

$$L_S = \frac{1}{2}\left(1 - \frac{\langle G(F(I)),\ S_w\rangle}{\|G(F(I))\|_2\cdot\|S_w\|_2}\right)$$

- **Three regularizers to handle fine grained SBIR**:
  - *Minimize sketch-photo domain gap*
    - We used a discriminator to predict the domain of the input in the intermediate latent space.
  - *Discriminative intermediate latent space*
    - We used a classification loss to discriminate different categories and a triplet loss if there exists only 1 category for intra sample discrimination.
  - *Transfer of semantic knowledge to unseen categories*
    - Semantic decoder head over F to reconstruct embedding representation of the category label with respect to either sketch or photo.

Experiments & Results:

- Dataset: Sketchy[1] and QMUL-Shoe-V2[2]
- Evaluated against a few designed baselines on 4 setups to judge a model's adaptability, its generalizing potential and impact of meta-learn margin value. Further details in paper.

Datasets	Baseline		Fine-Tuning		Generalisation [36]		**Proposed** (k=5)			
	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5	GAP_B	GAP_G
Sketchy (Category Level)	18.4%	37.3%	18.5%	37.5%	22.7%	42.1%	28.1%	51.8%	9.7↑	5.4↑
Shoe-V2 (User Level)	33.7%	70.2%	33.8%	70.2%	33.8%	70.4%	38.3%	76.6%	4.6↑	4.5↑

Quantitative evaluation showing average classification accuracy.

		Sketchy (Category)		Shoe-V2 (User)	
		Acc@1	Acc@5	Acc@1	Acc@5
	Our Baseline	18.4%	37.3%	33.7%	70.2%
	Our Baseline + Reg.	19.2%	39.6%	33.9%	71.3%
	Upper-Bound	29.8%	53.7%	–	–
SOTA	Triplet-SN [60]	15.3 %	34.0%	28.5%	67.3%
SOTA	Triplet-HOLEF [53]	16.7%	35.9%	31.4%	69.1%
SOTA	Triplet-RL [7]	4.7%	7.8%	34.1%	70.2%
SOTA	Mixed-Jigsaw [36]	16.7%	34.3%	33.5%	71.4%
SOTA	StyleMeUp [46]	19.6%	39.7%	36.4%	81.8%
GA	CC-DG [36]	22.7%	42.1%	33.8%	70.4%
GA	Distill(non-MAML) [36]	18.9%	38.1%	33.9%	70.9%
ZS-SBIR	CVAE-Regress [57]	2.4%	9.5%	1.8%	3.1%
ZS-SBIR	Sem-Pyc [16]	4.9%	17.3%	2.1%	4.7%
ZS-SBIR	Doodle2Search [14]	14.8%	34.5%	28.1%	66.9%
ZS-SBIR	SAKE [33]	6.4%	20.3%	3.6%	5.7%

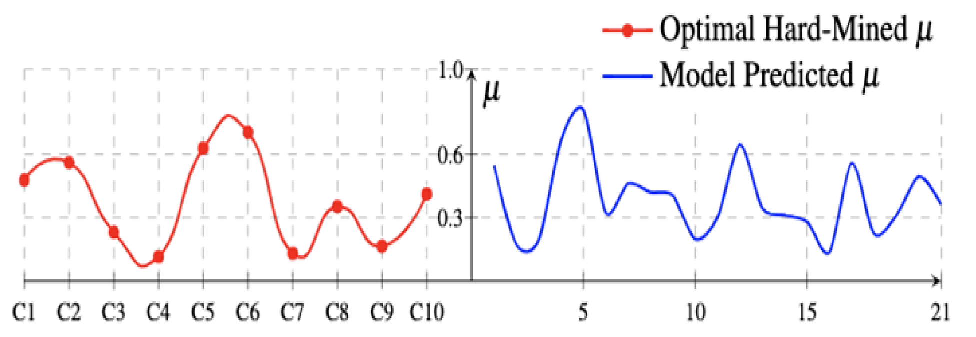
			Sketchy (Category)		Shoe-V2 (User)	
			Acc@1	Acc@5	Acc@1	Acc@5
Adaptation Based Approaches	Fine-Tuning	k=1	18.4%	37.3%	33.7%	70.2%
		k=5	18.5%	37.5%	33.8%	70.2%
		k=10	18.6%	37.5%	–	–
	MAML [20]	k=1	19.5%	38.7%	34.2%	70.7%
		k=5	22.8%	42.3%	35.5%	74.6%
		k=10	26.4%	48.9%	–	–
	sign-MAML [19]	k=1	19.1%	38.2%	33.8%	69.6%
		k=5	20.5%	39.6%	34.1%	70.8%
	ANIL [41]	k=1	19.7%	38.9%	34.5%	70.9%
		k=5	23.2%	42.8%	35.7%	75.3%
		k=10	26.9%	48.3%	–	–
	Ours	k=1	21.8%	42.5%	34.9%	71.4%
		k=5	28.1%	51.8%	38.3%	76.6%
		k=10	32.7%	53.5%	–	–

Quantitative evaluation showing average classification accuracy on 4 different competitors.

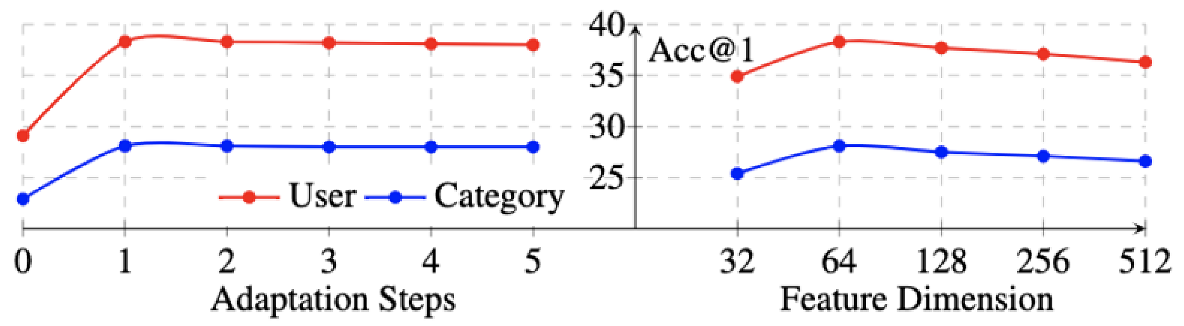
L_D	L_S	L_C	Sketchy Category Level	L_D	L_{ud}	Shoe-V2 User Level
✓	✓	✓	28.1%	✓	✓	38.3%
×	✓	✓	26.3%	×	✓	37.1%
×	×	✓	23.7%	×	×	35.8%
×	×	×	16.5%	-	-	-

Acc. falls

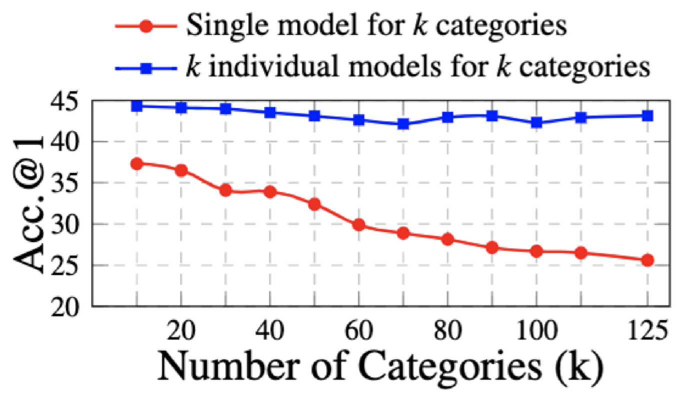
Ablative study judging design choice



Ablation on predicted margin value



Ablative study varying adaptation steps and feature dimension



Ablative study on single model with k individual models