SLADE: A Self-Training Framework For Distance Metric Learning

Jiali Duan[1,2], Yen-liang Lin[2], Son Tran[2], Larry S. Davis[2], C.-C. Jay Kuo[1]

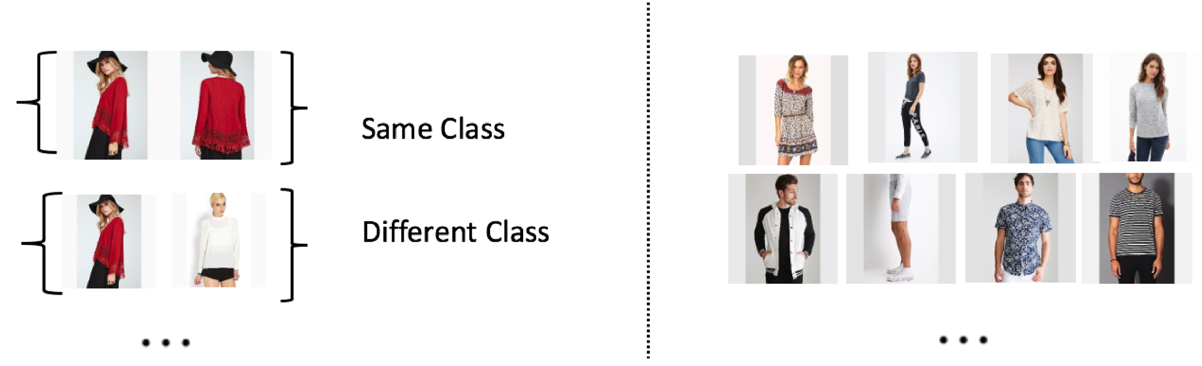
[1]University of Southern California, [2]Amazon

CVPR VIRTUAL JUNE 19-25

Problem Definition

Goal:
- Propose and investigate a semi-supervised framework for deep metric learning.
- Leverage unlabeled data to further improve the fully-supervised metric learning approaches.

Motivation: A recent study [1] shows that most losses perform similarly when properly tuned. We explore another direction that leverages un-annotated data.



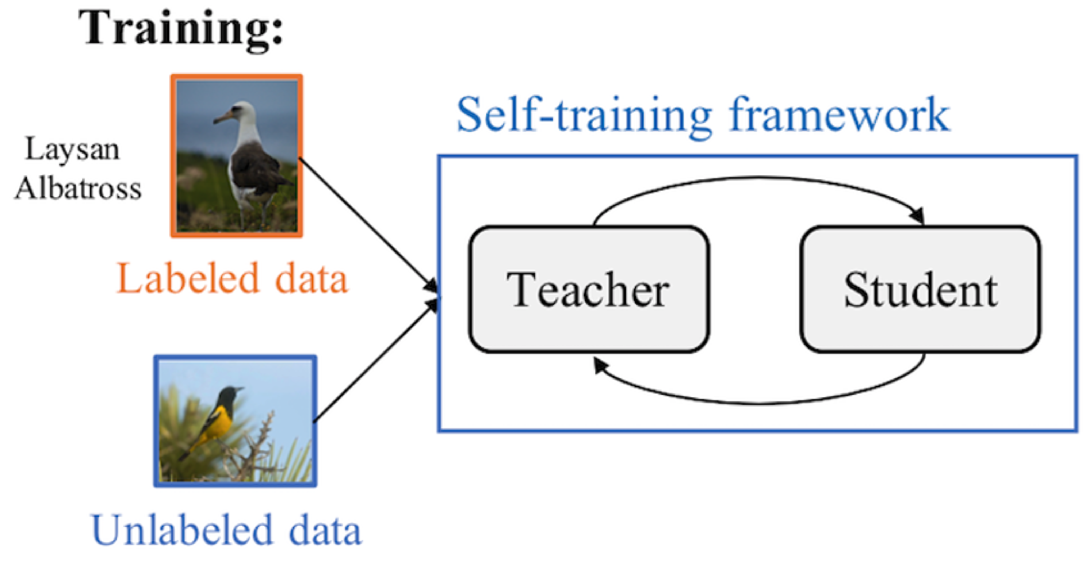
Contributions:
- A novel self-training framework to improve retrieval performance with unlabeled data.
- A feature basis learning approach to deal with noisy pseudo-labels during self-training.

Dataset

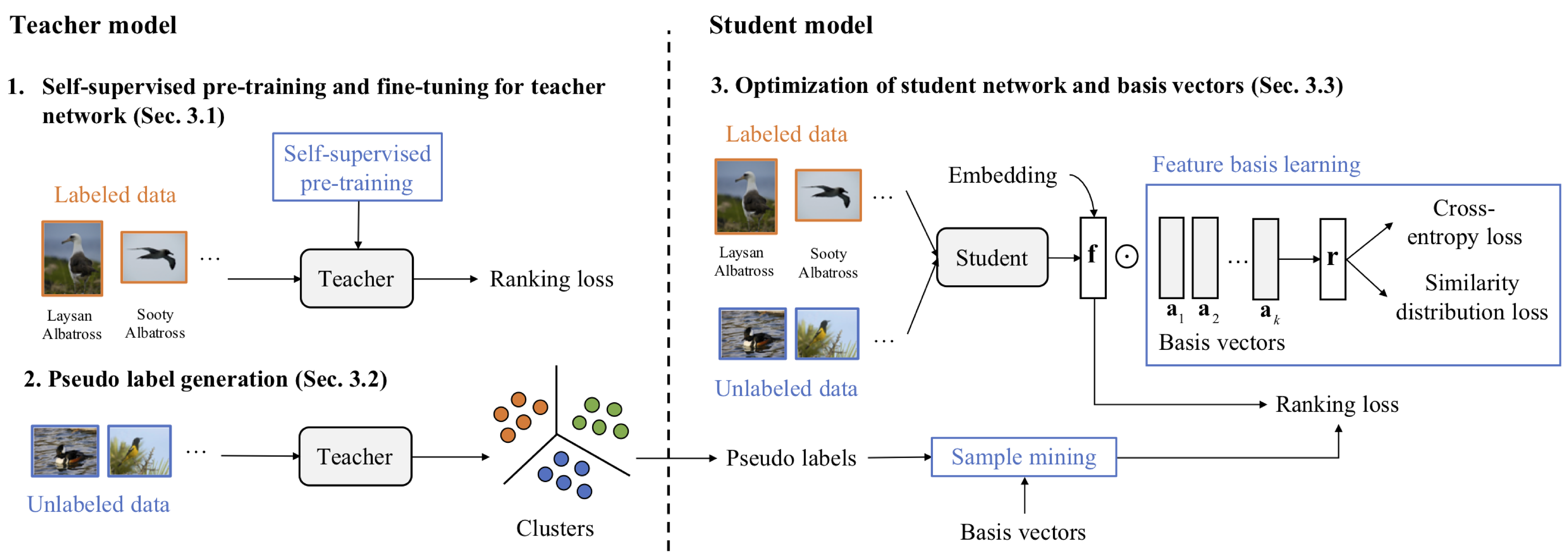
CUB-200/NABirds:
- CUB-200: 200 fine-grained species / 11,788 images
- NABirds: 743 categories birds / 48,000 images

Cars-196/CompCars:
- Cars-196: 196 classes / 16,185 images
- CompCars: filtered 145 classes / 16,537 images

In-shop/Fashion200k:
- In-shop: 7,982 clothing items / 52,712 images
- Fashion200k: filtered 1,045 items / 14,635 images



Method



Embedding learning:

$$\mathcal{L}_{\text{rank}} = [d_{\text{p}} - m_{\text{pos}}]_+ + [m_{\text{neg}} - d_{\text{n}}]_+$$

Feature basis learning:

$$\mathcal{L}_{\text{Basis}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{SD}}$$

Sample mining:

$$P = \{(\hat{x}_i, \hat{x}_j) | s(\hat{x}_i, \hat{x}_j) \geq T_1\}$$
$$N = \{(\hat{x}_i, \hat{x}_j) | s(\hat{x}_i, \hat{x}_j) \leq T_2\}$$

Joint training of student and feature basis:

$$\min_{\theta^s, \mathbf{W}_a} \mathcal{L}(\theta^s, \mathbf{W}_a) = \mathcal{L}_{\text{rank}}(D^l; \theta^s) + \lambda_1 \mathcal{L}_{\text{rank}}(D^u; \theta^s) + \lambda_2 \mathcal{L}_{\text{Basis}}(D^l, D^u; \theta^s, \mathbf{W}_a)$$

- D^l: labeled data
- D^u: unlabeled (pseudo-labeled) data
- θ^s: learnable student parameters
- $\mathbf{W}_\mathbf{a}$: learnable feature-basis parameters

Experiments & Results

Main Result: Comparison against state-of-the-art fully supervised approaches on CUB-200 and Cars-196

Methods	Frwk	Init	Arc / Dim	CUB-200-2011 MAP@R	CUB-200-2011 RP	CUB-200-2011 P@1	Cars-196 MAP@R	Cars-196 RP	Cars-196 P@1
Contrastive	[1]	ImageNet	BN / 512	26.53	37.24	68.13	24.89	35.11	81.78
Triplet	[1]	ImageNet	BN / 512	23.69	34.55	64.24	23.02	33.71	79.13
ProxyNCA	[1]	ImageNet	BN / 512	24.21	35.14	65.69	25.38	35.62	83.56
N. Softmax	[1]	ImageNet	BN / 512	25.25	35.99	65.65	26.00	36.20	83.16
CosFace	[1]	ImageNet	BN / 512	26.70	37.49	67.32	27.57	37.32	85.52
FastAP	[1]	ImageNet	BN / 512	23.53	34.20	63.17	23.14	33.61	78.45
MS+Miner	[1]	ImageNet	BN / 512	26.52	37.37	67.73	27.01	37.08	83.67
Proxy-Anchor[1]	[2]	ImageNet	R50 / 512	-	-	69.9	-	-	87.7
Proxy-Anchor[2]	[1]	ImageNet	R50 / 512	25.56	36.38	66.04	30.70	40.52	86.84
ProxyNCA++	[3]	ImageNet	R50 / 2048	-	-	72.2	-	-	90.1
Mutual-Info	[4]	ImageNet	R50 / 2048	-	-	69.2	-	-	89.3
Contrastive (T_1)	[1]	ImageNet	R50 / 512	25.02	35.83	65.28	25.97	36.40	81.22
Contrastive (T_2)	[1]	SwAV	R50 / 512	29.29	39.81	71.15	31.73	41.15	88.07
SLADE (Ours) (S_1)	[1]	ImageNet	R50 / 512	29.38	40.16	68.92	31.38	40.96	85.8
SLADE (Ours) (S_2)	[1]	SwAV	R50 / 512	**33.59**	**44.01**	**73.19**	**36.24**	**44.82**	**91.06**
MS (T_3)	[1]	ImageNet	R50 / 512	26.38	37.51	66.31	28.33	38.29	85.16
MS (T_4)	[1]	SwAV	R50 / 512	29.22	40.15	70.81	33.42	42.66	89.33
SLADE (Ours) (S_3)	[1]	ImageNet	R50 / 512	30.90	41.85	69.58	32.05	41.50	87.38
SLADE (Ours) (S_4)	[1]	SwAV	R50 / 512	**33.90**	**44.36**	**74.09**	**37.98**	**46.92**	**91.53**

Note: The teacher networks (T_1, T_2, T_3, T_4) are trained with the different losses, and then used to train the student networks (S_1, S_2, S_3, S_4).

Ablation study 1: Initialization of Teacher Network

Pre-trained weight	MAP@R CUB-200	MAP@R Cars-196
ImageNet	29.38	31.38
Pre-trained SwAV	32.79	35.54
Fine-tuned SwAV	33.59	36.24

Ablation study 3: Pairwise Similarity Loss

Regularization	CUB-200 MAP@R	CUB-200 RP	CUB-200 P@1
Local-CE	32.69	43.20	72.64
Global-CE	32.23	42.68	72.45
SD (Ours)	33.59	44.01	73.19

Ablation study 2: Components in Student Network

Components	MAP@R CUB-200	MAP@R Cars-196
Teacher (contrastive)	29.29	31.73
Student (pseudo label)	30.81	31.99
+ Basis	32.45	35.78
+ Basis + Mining	33.59	36.24

Ablation study 4: Number of Clusters

k	NABirds MAP@R	NABirds RP	NABirds P@1
100	31.83	42.25	72.19
200	32.61	43.02	72.75
300	32.81	43.18	72.21
400	33.59	44.01	73.19
500	33.26	43.69	73.26

Qualitative Results:

CUB-200 Cars-196

Reference:

1. Musgrave Kevin, et al. "A metric learning reality check." ECCV 2020.

Paper ID: 3886
Paper Link: https://arxiv.org/abs/2011.10269