ProtoSEED: Prototypical Self-Supervised Representation Distillation

Kyungmin Lee
Agency for Defense Development

국 방 과 학 연 구 소
Agency for Defense Development

Backgrounds

Self-supervised representation distillation (SEED)

- ✓ SSL works well on large networks (e.g. ResNet-50), but are not good on small networks (e.g. ResNet-18)
- ✓ Student networks can learn better representation by distilling the large SSL pre-trained models [1].

Knowledge Distillation (KD)

- ✓ Supervised learning: minimize the probability output of teacher and student models .
- ✓ Self-supervised learning: use prototypes to generate probability score and learn by self-distillation [2, 3].

Contrastive representation distillation

- ✓ Maximize the mutual information between teacher and student representations (e.g. CPC)
- ✓ Contrastive objective helps distillation in supervised learning [4]

Motivation

Our approach: prototypical KD + contrastive learning

- ✓ We adopt prototypes to generate probability score and distill the teacher's score to student.
- ✓ We propose novel Prototypical CPC objective where the critic measures the probabilistic discrepancy of teacher and student.

Methods

Prototypical Contrastive Predictive Coding

- ✓ Given features z_t and z_s, generate teacher probability p_t and student probability p_s by softmax operator with prototypes.
- ✓ Set the critic by

$$e^{-H(p_t,p_s)} = e^{\sum_{k=1}^{K} p_t^{(k)} \log p_s^{(k)}}$$

- ✓ Plug in to contrastive objective

$$I(T;S) \geq \mathbb{E}\left[\log \frac{e^{-H(p_t,p_s)}}{\frac{1}{N}\sum_{j=0}^{N-1} e^{-H(p_{tj},p_s)}}\right] = \mathbb{E}\left[\log \frac{\exp(p_t \cdot \tilde{z}_s/\tau_s)}{\frac{1}{N}\sum_{j=0}^{N-1} \exp\left(p_{tj} \cdot \tilde{z}_s/\tau_s\right)}\right]$$

$$\geq \mathbb{E}\left[\log \frac{\exp(p_t \cdot \tilde{z}_s/\tau_s)}{\frac{1}{N}\sum_{j=0}^{N-1}\sum_{k=1}^{K} p_{tj}^{(k)} \exp\left(\tilde{z}_s^{(k)}/\tau_s\right)}\right] = \mathbb{E}\left[\log \frac{\exp(p_t \cdot \tilde{z}_s/\tau_s)}{\sum_{k=1}^{K} q^{(k)} \exp\left(\tilde{z}_s^{(k)}/\tau_s\right)}\right]$$

- ✓ $q^{(k)} = \frac{1}{N}\sum_{j=0}^{N-1} p_{tj}^{(k)}$ acts as the prior of each prototype

Prior momentum

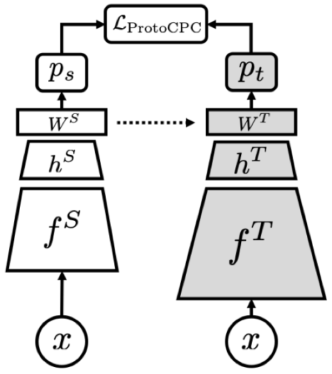
- ✓ Prior term contains information about negative samples.
- ✓ More accurate estimation of prior terms require large negative samples.
- ✓ We use momentum for prior term to avoid large batch.

Prototypical SEED

- ✓ Set student same as teacher.
- ✓ Train the student by following:

$$\min_{g^S, W^S} \mathbb{E}_{x\sim X}\left[\mathcal{L}_{\mathrm{ProtoCPC}}(p_t(x), p_s(x))\right]$$

- ✓ **The teacher prototypes are copied from student's.**



Experiments

Main results

- ✓ Various teacher SSL representations to ResNet-18

	MoCo ResNet-50		SwAV ResNet-50		DINO ResNet-50		DINO DeiT-S/16	
	Linear	*k-NN*	*Linear*	*k-NN*	*Linear*	*k-NN*	*Linear*	*k-NN*
Teacher	71.1	61.9	75.3	65.7	75.3	67.5	77.0	74.3
Supervised	69.5	69.5	69.5	69.5	69.5	69.5	69.5	69.5
SSL	52.5	36.7	57.5	48.2	58.2	50.3	58.2	50.3
ProtoSEED	**61.1**(+8.6)	**55.6**(+18.9)	**63.1**(+5.6)	**57.7**(+9.4)	**63.9**(+5.7)	**60.3**(+10.0)	**65.5**(+7.3)	**63.2**(+12.9)

- ✓ ProtoSEED outperforms SSL and works well across different architectures

Comparison with SEED

- ✓ ProtoSEED outperforms SEED

Teacher	Method	Epochs	*Linear*	*k-NN*
MoCo	SEED	200	60.5	49.1
	ProtoSEED	100	**61.1**	**55.6**
SwAV	SEED	100	61.1	-
	SEED	200*	62.6	-
	ProtoSEED	100	63.1	**57.7**
	ProtoSEED	100*	**63.9**	57.0
DINO	ProtoSEED	100	63.9	60.3
	ProtoSEED	100*	**65.3**	**60.7**

- ✓ With multi-crops data augmentation (denoted by *), we achieve state-of-the-performance on ResNet-18.
- ✓ ProtoSEED outperforms SSL and works well across different architectures

References

[1]. Fang, Zhiyuan, et al. "Seed: Self-supervised distillation for visual representation." arXiv preprint arXiv:2101.04731 (2021).
[2]. Caron, Mathilde, et al. "Unsupervised learning of visual features by contrasting cluster assignments." arXiv preprint arXiv:2006.09882 (2020).
[3]. Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." arXiv preprint arXiv:2104.14294 (2021).
[4]. Tian, Yonglong, Dilip Krishnan, and Phillip Isola. "Contrastive representation distillation." arXiv preprint arXiv:1910.10699 (2019).