Semi-Supervised Few-Shot Learning with Prototypical Random Walks

Formerly: Few-shot learning with local and global consistency

Ahmad Ayad[1], Nassir Navab[1,3], Mohamed Elhoseiny[2*], Shadi Albarqouni[1*]

* shared senior authorship

1. Computer Aided Medical Procedures (CAMP), Technical University of Munich, Munich, Germany
2. Facebook AI Research /KAUST
3. Computer Aided Medical Procedures (CAMP), Johns Hopkins University, USA

Introduction

There remains a huge performance gap between human and artificial learners when it comes to sample efficiency, and there has recently been a lot of work in few-shot learning(FSL) motivated to bridge that gap.

However, most work is focused on fully supervised FSL. On the other hand, most data collected or observable in the world, does not come with a label. We tackle semi-supervised few-shot learning(SS-FSL), where few-shot learners are expected to improve their performance by leveraging unlabeled data.

We show that state-of-the-art performance in the various SS-FSL tasks can be obtained by enforcing local and global consistency.

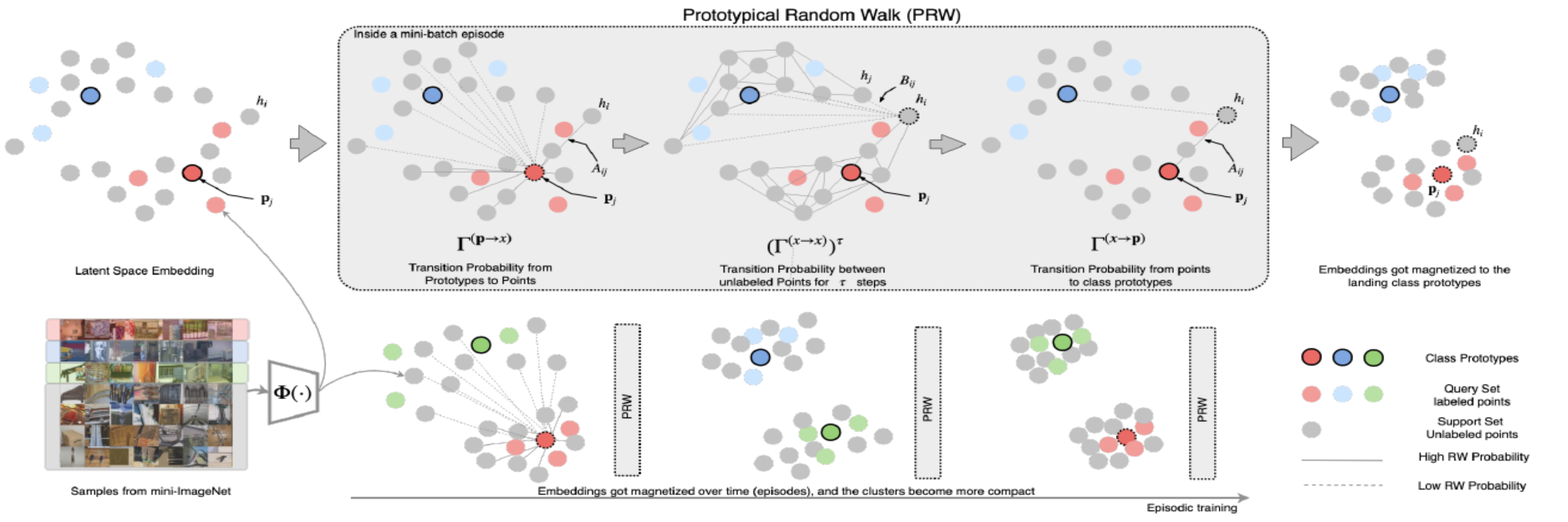
Prototypical Random Walk (PRW)
Inside a mini-batch episode
h_i
h_j
B_{ij}
A_{ij}
$\mathbf{p}_j$
Latent Space Embedding
$\Gamma^{(\mathbf{p}\to x)}$
Transition Probability from Prototypes to Points
$(\Gamma^{(x\to x)})^\tau$
Transition Probability between unlabeled Points for τ steps
$\Gamma^{(x\to \mathbf{p})}$
Transition Probability from points to class prototypes
Embeddings got magnetized to the landing class prototypes
$\Phi(\cdot)$
Samples from mini-ImageNet
PRW
PRW
PRW
Embeddings got magnetized over time (episodes), and the clusters become more compact
Episodic training
Class Prototypes
Query Set labeled points
Support Set Unlabeled points
High RW Probability
Low RW Probability

Figure 1: Our PRW aims at maximizing the probability of a random walk begins at the class prototype $\mathbf{p}_j$, taking τ steps among the unlabeled data, before it lands to the same class prototype. This results in a more discriminative representation, where the embedding of the unlabeled data of a particular class got magnetized to its corresponding class prototype, denoted as *prototypical magnetization*.

Methodology

Few-Shot Learning:

- Based on Prototypical Networks(PN)
- Goal is to learn a representation where points of the same class form tight, well-separated clusters around the **class prototype**.
- During training, labelled points' representations are moved to be closer to their prototypes, but what do we do with unlabelled points?
- **Local & Global consistency**
- We add two loss terms to the original PN enforcing local and global consistency.
- Our loss terms work for semi-supervised meta-training, when additional unlabeled data is available at test time, we can combine our model with a semi-supervised adaptation step from [1], and benefit from unlabeled data for both meta-training and adaptation.

Local Consistency:

- Enforce that points do not fall close to decision boundaries i.e. in the spaces separating our tight clusters.
- Based on a Virtual Adversarial Training Loss. We require points close together to have similar predictions.
- Minimize $\mathrm{KL}(\mathrm{f}_\theta(x),\ f_\theta(x+\epsilon_{adv}))$

Global Consistency:

- Well aligned with PN: enforces a representation where points of the same class form well separated clusters around the class prototype.
- Takes a global view of the latent space structure, by constructing a similarity graph for the prototypes and unlabeled points and simulating a random walk.
- **Prototypical Random Walks(PRW)** start at some class prototype, walk over to the unlabeled points, and eventually go back to a prototype. The objective is to maximize the probability that the walk ends at the starting prototype: **The Landing Probability.**
- Our PRW loss *magnetizes* points of a class around the prototype, while repelling other points and prototypes away.
- We minimize the PRW loss as computed in Algorithm 1.

Algorithm 1: Prototypical Random Walk loss

```
Data: $A^{M\times N}$ s.t. $A_{i,j} = -\|h_i - \mathbf{p}_j\|^2$.
$B^{M\times M}$ s.t. $B_{i,j} = -\|h_i - h_j\|^2$.
$h$ are point representations, and $\mathbf{p}$ are prototypes
Result: PRW loss

// Compute RW transition probabilities
$\Gamma^{(\mathbf{p}\to x)} = \mathrm{softmax}(A^T)$,
$\Gamma^{(x\to \mathbf{p})} = \mathrm{softmax}(A)$,
$\Gamma^{(x\to x)} = \mathrm{softmax}(B)$,

// Simulate random walk
$T^{(\tau)} = \Gamma^{(\mathbf{p}\to x)} \cdot (\Gamma^{(x\to x)})^\tau \cdot \Gamma^{(x\to \mathbf{p})}$

// Compute PRW loss
$\mathcal{L}_{walker} = \sum_{i=0}^{\tau} \alpha_i \cdot -\frac{1}{N_c} \sum_{i=0}^{N_c} \log T_{i,i}$

RETURN $\mathcal{L}_{walker}$
```

Experiments

- We run experiments on two well know FSL datasets: Omniglot, and Mini-Imagenet.
- We follow the labelled/unlabeled split from previous works[1,2], with 10% of the labels for Omniglot, and 40% of the labels for Mini-Imagenet.
- We use the same Conv-4 network popular in FSL, and used the previous state-of-the-art in SS-FSL [1,2] for fair comparison.
- We perform experiments with/without additional unlabeled data at test time: semi-supervised adaptation.
- We perform experiments with and without the presence of **distractor** points. Distractors are unlabeled points which do not belong to the classes of the labelled data. This setting is more realistic but more challenging.

Ablation study:

Model	Omniglot	Mini-Imagenet	
	1-shot	1-shot	5-shot
PN (Ren et al., 2018)	94.62 ± 0.09	43.61 ± 0.27	59.08 ± 0.22
PN+VAT	95.66 ± 0.21	44.63 ± 0.21	64.02 ± 0.20
PN+VAT+ENT	97.14 ± 0.16	44.48 ± 0.22	66.94 ± 0.20
PN+PRW	97.96 ± 0.07	50.33 ± 0.27	66.99 ± 0.24
CPN: PN+PRW+VAT	**98.03 ± 0.11**	**51.03 ± 0.23**	**67.78 ± 0.20**

- Here we compare with the vanilla PN (our baseline), denoted PN in the table, and PN with individual components of our loss.
- ENT refers to Shannon entropy minimization, it has been found the work well with VAT.
- We can see that all our models improve on the baseline, with CPN (our full loss) performing best on all benchmarks
- Following CPN, we can see that PRW on its own is also effective, and improves significantly on the baseline in all tests. It is also much faster to train than VAT, as VAT requires additional forward and backward passes through the network.

Benchmark results:

Without distractors:

Model	Omniglot	Mini-Imagenet	
	1-shot	1-shot	5-shot
PN$_{all}$(Snell et al., 2017)	98.8	49.4	68.2
PN(Ren et al., 2018)	94.62 ± 0.09	43.61 ± 0.27	59.08 ± 0.22
MetaGAN	97.58 ± 0.07	50.35 ± 0.23	64.43 ± 0.27
Ours: CPN	**98.03 ± 0.11**	**51.03 ± 0.23**	**67.78 ± 0.20**
PN+ Semi-supervised inference	97.45 ± 0.05	49.98 ± 0.34	63.77 ± 0.20
PN+ Soft K-means	97.25 ± 0.10	50.09 ± 0.45	64.59 ± 0.28
PN+ Soft K-means + cluster	97.68 ± 0.07	49.03 ± 0.24	63.08 ± 0.18
PN+ Masked soft K-means	97.52 ± 0.07	50.41 ± 0.24	64.39 ± 0.24
Ours: CPN + semi-supervised inference	**99.30 ± 0.04**	**56.91 ± 0.25**	**70.11 ± 0.19**

- The first row denotes a PN trained on **100% of the labels**, the middle section is results without semi-supervised adaptation. The bottom section is results with semi-supervised adaptation
- Remarkably, our CPN outperforms the **fully supervised** PN$_{all}$ in the 1-shot Mini-Imagenet with 51.03% to 49.4%.
- For all experiments, our model improves greatly on the state-of-the-art, with dramatic increases in some cases
- 70.11% to 64.49% in the 5-shot Mini-Imagenet with adaptation.
- 67.78% to 64.43% in the 5-shot Mini-Imagenet without adaptation.

With distractors:

Model	Omniglot	Mini-Imagenet	
	1-shot	1-shot	5-shot
PN	94.62 ±0.09	43.61 ±0.27	59.08 ±0.22
Ours: PN+PRW	**97.76 ± 0.11**	**50.96 ± 0.23**	**67.64 ± 0.18**
Ours: CPN	96.44 ±0.11	50.2 ±0.23	64.1 ±0.26
PN+ Semi-supervised inference	95.08 ±0.09	47.42 ±0.33	62.62 ±0.24
PN+ Soft K-means	95.01 ±0.09	48.70 ±0.32	63.55 ±0.28
PN+ Soft K-means + cluster	97.17 ±0.04	48.86 ±0.32	61.27 ±0.24
PN+ Masked soft K-means	**97.30 ±0.30**	49.04 ±0.31	62.96 ±0.14
Ours: CPN + semi-supervised inference	96.76 ±0.09	**53.76 ±0.23**	**66.17 ±0.21**

- Here we present a comparison with the state-of-the-art in the challenging distractor case.
- The first section denotes experiments without semi-supervised adaptation.
- Even with distractors, CPN gets significant improvements over the baseline, with 50.2% to 43.6% in the 1-shot Mini-Imagenet for example.
- However, without the VAT loss (PN+PRW), performs better than CPN. This is due to the **global** character of the loss, and its ability to *ignore,* to an extent, distractor points.
- In the case of semi-supervised adaptation, our model still performs strongly, giving remarkable state-of-the-art improvements on Mini-Imagenet.

Conclusion

- We presented Consistent Prototypical Networks, which obtains state-of-the-art results in a wide range of SS-FSL tasks.
- Both local and global consistency benefit semi-supervised learning in the few-shot setting, with global consistency having an edge.
- Our loss provides substantial benefit over the baselines, even when distractors are present.
- PRW is particularly robust to distractor points.
- Unlike previous works which tend to be effective for either meta-training **or** adaptation, we show that our model can perform in both settings.

References

[1] Ren, M., Ravi, S., Triantafillou, E., Snell, J., Swersky, K., Tenenbaum, J. B., Larochelle, H., and Zemel, R. S. Meta-learning for semi-supervised few-shot classification. In International Conference on Learning Representations, 2018.

[2] Zhang, R., Che, T., Ghahramani, Z., Bengio, Y., and Song, Y.Metagan: An adversarial approach to few-shot learning.InAdvances in Neural Information Processing Systems,pp. 2371–2380, 2018.

Scan the QR code to have access to the github repository.