2021 ICCV OCTOBER 11-17 VIRTUAL

Iterative label cleaning for transductive and semi-supervised few-shot learning

Michalis Lazarou[1], Tania Stathaki[1], Yannnis Avrithis[2]

[1]Imperial College London [2]Inria, Univ Rennes, CNRS, IRISA

Imperial College London

Inria INVENTEURS DU MONDE NUMÉRIQUE

Motivation

- **Problem:** Few-shot learning by using unlabelled data
- Limited labeled data make the problem fundamentally hard
- Available unlabeled data that can be used to improve accuracy performance
- **Solution:** Incorporate multiple ideas to design a novel algorithm
- Exploit the manifold structure of the data to classify the unlabelled data
- Select the most likely correctly classified predictions to augment labelled dataset and iteratively classify all available unlabeled data

Problem definition

Pre-training:

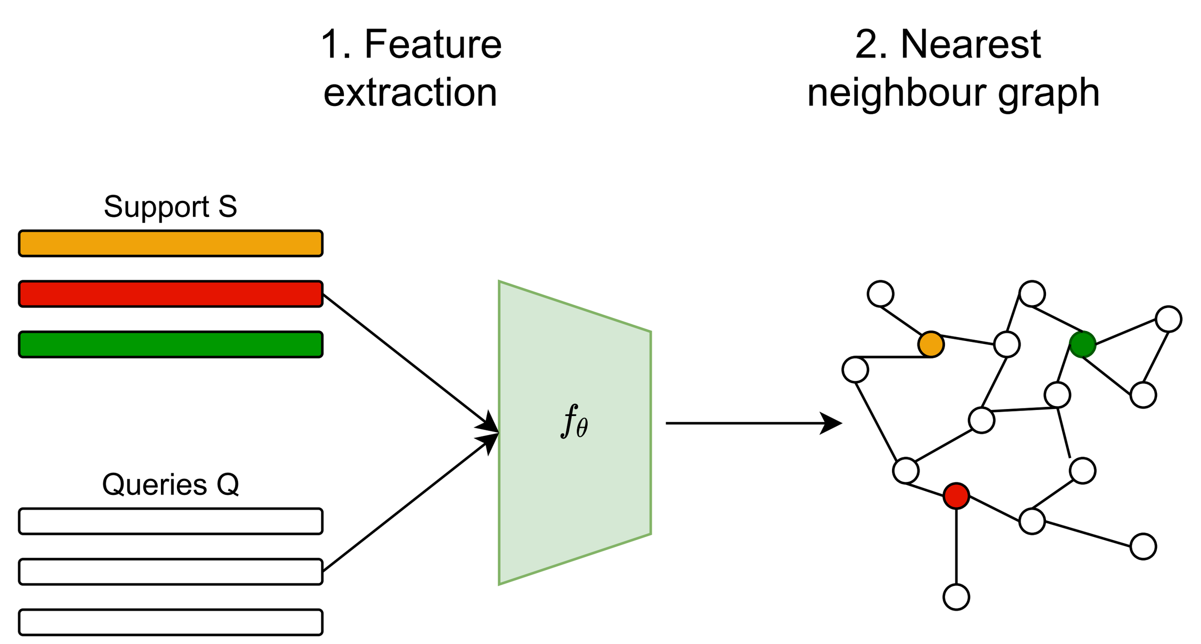
- We use publicly available pre-trained networks from published works
- Base class dataset: $D_{\text{base}} := \{(x_i, y_i)\}_{i=1}^{I}$ where $y_i \in C_{\text{base}}$
- Embedding network $f_\theta : \mathcal{X} \to \mathbb{R}^d$ is trained on D_{base}

Inference stage

- We focus on transductive and semi-supervised few-shot learning
- Novel class dataset D_{novel} with C_{novel} disjoint from C_{base}
- Assume access to f_θ, a support set, S, a query set, Q and in the semi-supervised setting also an unlabelled set, U

Nearest neighbour graph

- Embed all examples from S and Q into feature vectors and ℓ_1-normalize them, where T is the total amount of examples both labelled and unlabelled
- Construct a k-nearest neighbour graph using all the features from S and Q

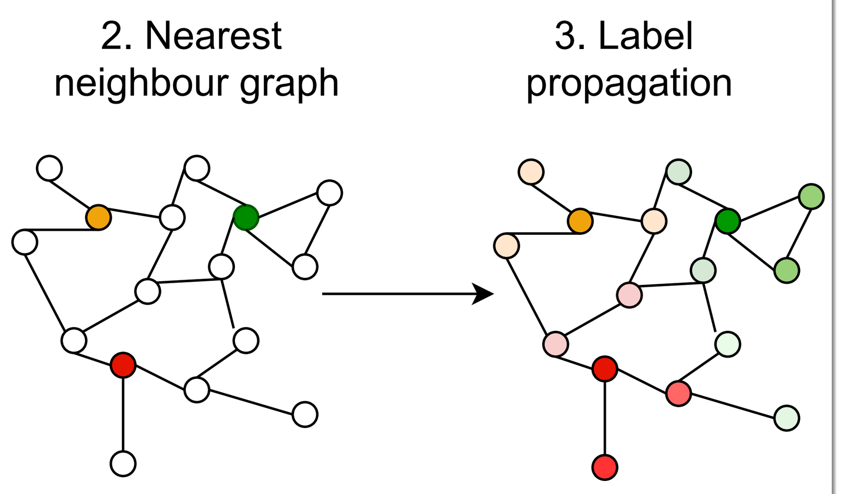


Label propagation

- Define the *label matrix* Y
- Label propagation to obtain a class probability distribution for every query

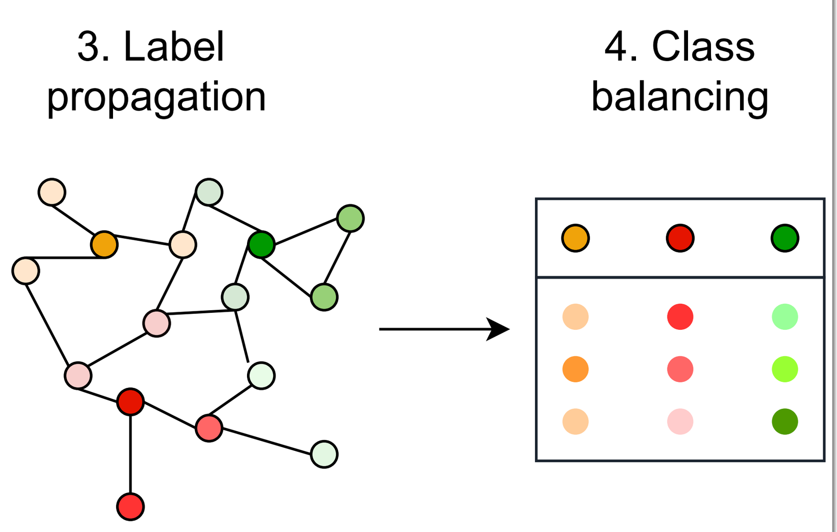
$$Z := (I - \alpha \mathcal{W})^{-1} Y, \quad (1)$$

where $\alpha \in [0, 1)$ is a hyperparameter, N is the number of classes and $\mathcal{W}$ is the adjacency matrix obtained from the construction of the nearest neighbour graph



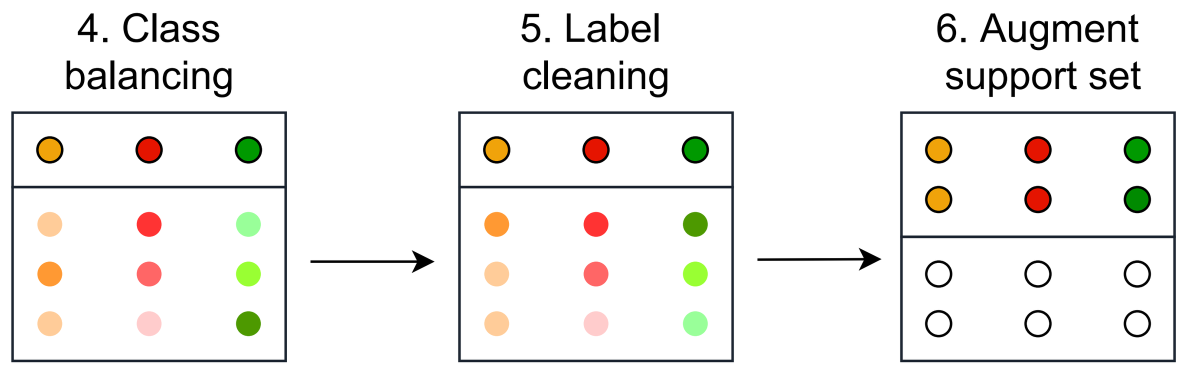
Class balancing

- Use the output matrix from the label propagation part
- We incorporate prior information in the class balancing regime by setting a row-wise sum $\mathbf{p}$ and a column-wise sum $\mathbf{q}$ as restrictions
- We search for a *transport plan* through the use of Sinkhorn-Knopp algorithm [2], i.e., iteratively normalizing rows and columns to the respective $\mathbf{q}$ and $\mathbf{q}$ until convergence



Label cleaning

- In order to identify the most likely correctly classified labels, we interpret the problem as *learning with noisy labels*
- We train an N-way linear classifier with large learning rate on both labeled and pseudo-labeled data
- For every class, we select the 3 pseudo-labels with the lowest loss value and treat them as labeled examples
- Iterate the procedure until all queries are classified



Implementation details

- We use PyTorch, FAISS and scikit-learn

Ablation study

Label propagation

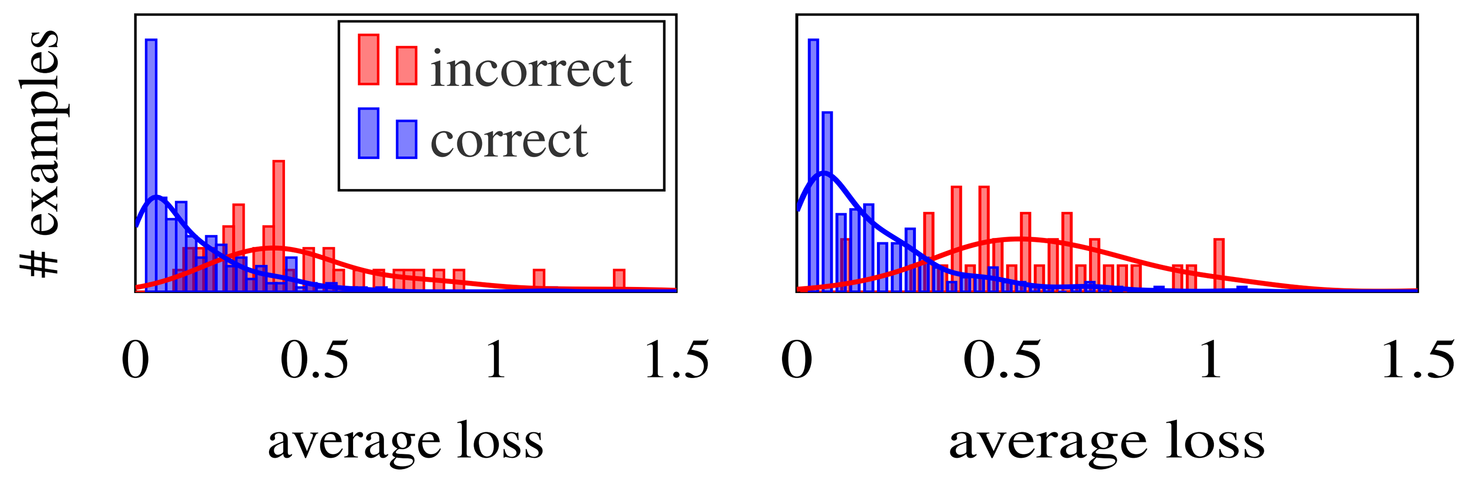
Inference	ResNet-12A 1-shot	ResNet-12A 5-shot	WRN-28-10 1-shot	WRN-28-10 5-shot
Inductive classifier	56.30$_{\pm0.62}$	**75.59**$_{\pm0.47}$	68.17$_{\pm0.60}$	**84.33**$_{\pm0.43}$
Label Propagation	**61.09**$_{\pm0.70}$	75.32$_{\pm0.50}$	**74.24**$_{\pm0.68}$	84.09$_{\pm0.42}$

Class balancing

Balancing	Network	*mini*ImageNet 1-shot	*mini*ImageNet 5-shot	*tiered*ImageNet 1-shot	*tiered*ImageNet 5-shot
None	WRN-28-10	78.06$_{\pm0.82}$	87.80$_{\pm0.42}$	86.04$_{\pm0.73}$	90.74$_{\pm0.46}$
True	WRN-28-10	**82.68**$_{\pm0.82}$	**89.07**$_{\pm0.41}$	**89.17**$_{\pm0.70}$	**92.67**$_{\pm0.44}$

Ablation study

Label cleaning



Iterative procedure

Inference	ResNet-12A 1-shot	ResNet-12A 5-shot	WRN-28-10 1-shot	WRN-28-10 5-shot
Non-iterative	65.04$_{\pm0.75}$	76.82$_{\pm0.50}$	79.42$_{\pm0.69}$	85.34$_{\pm0.43}$
iterative (iLPC)	**69.79**$_{\pm0.99}$	**79.82**$_{\pm0.55}$	**83.05**$_{\pm0.79}$	**88.82**$_{\pm0.42}$

State of the art comparisons

Transductive few-shot learning:

Method	Network	*mini*ImageNet 1-shot	*mini*ImageNet 5-shot	*tiered*ImageNet 1-shot	*tiered*ImageNet 5-shot	Cifar-FS 1-shot	Cifar-FS 5-shot
LR+ICI [3]*	ResNet-12A	66.85$_{\pm0.92}$	78.89$_{\pm0.55}$	82.40$_{\pm0.84}$	88.80$_{\pm0.50}$	75.36$_{\pm0.97}$	84.57$_{\pm0.57}$
iLPC (ours)	ResNet-12A	**69.79**$_{\pm0.99}$	**79.82**$_{\pm0.55}$	**83.49**$_{\pm0.88}$	**89.48**$_{\pm0.47}$	**89.00**$_{\pm0.70}$	**92.74**$_{\pm0.35}$
PT+MAP [1]*	WRN-28-10	82.88$_{\pm0.73}$	88.78$_{\pm0.40}$	88.15$_{\pm0.71}$	92.32$_{\pm0.40}$	**86.91**$_{\pm0.72}$	90.50$_{\pm0.49}$
LR+ICI [3]*	WRN-28-10	80.61$_{\pm0.80}$	87.93$_{\pm0.44}$	86.79$_{\pm0.76}$	91.73$_{\pm0.40}$	84.88$_{\pm0.79}$	89.75$_{\pm0.48}$
iLPC (ours)	WRN-28-10	**83.05**$_{\pm0.79}$	**88.82**$_{\pm0.42}$	**88.50**$_{\pm0.75}$	**92.46**$_{\pm0.42}$	86.51$_{\pm0.75}$	**90.60**$_{\pm0.48}$

Transductive inference, comparison with LR+ICI [3] and PT+MAP [1]. *: our reproduction with official code on our datasets.

Method	Network	*mini*ImageNet 1-shot	*mini*ImageNet 5-shot	*tiered*ImageNet 1-shot	*tiered*ImageNet 5-shot	Cifar-FS 1-shot	Cifar-FS 5-shot
LR+ICI [3]*	WRN-28-10	82.38$_{\pm0.86}$	88.78$_{\pm0.39}$	88.59$_{\pm0.74}$	92.11$_{\pm0.39}$	86.39$_{\pm0.79}$	90.02$_{\pm0.49}$
PT+MAP [1]*	WRN-28-10	83.79$_{\pm0.71}$	88.94$_{\pm0.33}$	88.87$_{\pm0.64}$	92.01$_{\pm0.36}$	87.63$_{\pm0.66}$	90.15$_{\pm0.46}$
iLPC (ours)	WRN-28-10	**85.98**$_{\pm0.74}$	**90.54**$_{\pm0.31}$	**90.02**$_{\pm0.70}$	**92.94**$_{\pm0.37}$	**88.54**$_{\pm0.68}$	**90.92**$_{\pm0.46}$

Transductive inference, 50 queries per class. *: our reproduction withofficial code on our datasets.

Semi-supervised few-shot learning:

Method	Network	Split	*mini*ImageNet 1-shot	*mini*ImageNet 5-shot	*tiered*ImageNet 1-shot	*tiered*ImageNet 5-shot	Cifar-FS 1-shot	Cifar-FS 5-shot
LR+ICI [3]*	ResNet-12A	30/50	67.57$_{\pm0.97}$	79.07$_{\pm0.56}$	83.32$_{\pm0.87}$	89.06$_{\pm0.51}$	75.99$_{\pm0.98}$	84.01$_{\pm0.62}$
iLPC (ours)	ResNet-12A	30/50	**70.99**$_{\pm0.91}$	**81.06**$_{\pm0.49}$	**85.04**$_{\pm0.79}$	**89.63**$_{\pm0.47}$	**78.57**$_{\pm0.80}$	**85.84**$_{\pm0.56}$
LR+ICI [3]*	WRN-28-10	30/50	81.31$_{\pm0.84}$	88.53$_{\pm0.43}$	88.48$_{\pm0.67}$	92.03$_{\pm0.43}$	86.03$_{\pm0.77}$	89.57$_{\pm0.53}$
PT+MAP [1]†	WRN-28-10	30/50	83.14$_{\pm0.72}$	88.95$_{\pm0.38}$	89.16$_{\pm0.61}$	92.30$_{\pm0.39}$	**87.05**$_{\pm0.69}$	89.98$_{\pm0.49}$
iLPC (ours)	WRN-28-10	30/50	**83.58**$_{\pm0.79}$	**89.68**$_{\pm0.37}$	**89.35**$_{\pm0.68}$	**92.61**$_{\pm0.39}$	87.03$_{\pm0.72}$	**90.34**$_{\pm0.50}$

Semi-supervised few-shot learning, comparison with [3, 1]. *: our reproduction with official code on our datasets. †: our adaptation to semi-supervised, based on official code.

[1] Y. Hu, V. Gripon, and S. Pateux. Leveraging the feature distribution in transfer-based few-shot learning. *arXiv preprint arXiv:2006.03806*, 2020.

[2] P. A. Knight. The Sinkhorn-Knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 2008.

[3] Y. Wang, C. Xu, C. Liu, L. Zhang, and Y. Fu. Instance credibility inference for few-shot learning. *CVPR*, 2020.

Contact: michalis.lazarou14@imperial.ac.uk