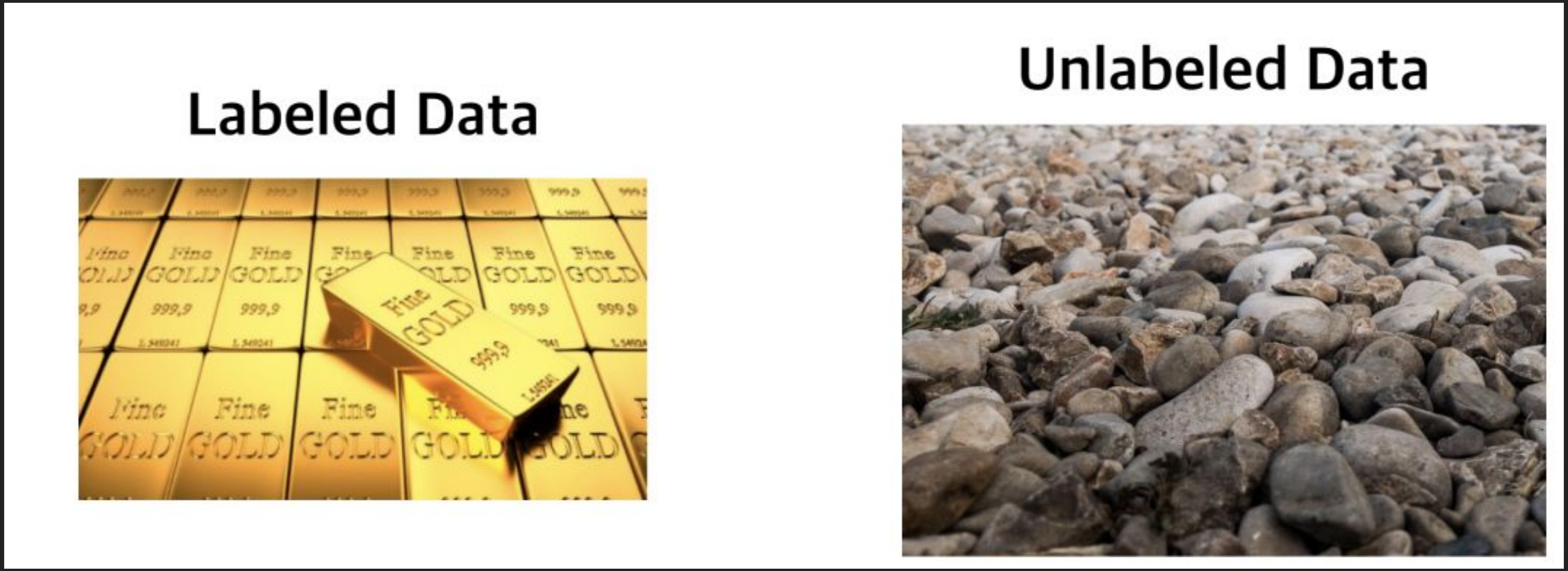
Label Propagation for Deep Semi-supervised Learning

Presented by Minoo

Labeling is expensive



What is semi-supervised learning?

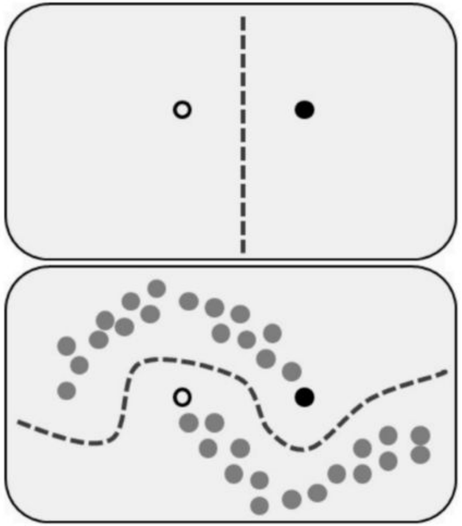
Supervised Learning

$$(x, y) \sim p(x, y)$$

$$\max \quad \mathbb{E}_{x,y\sim p(x,y)}\left[\log p(y|x)\right]$$

Semi-Supervised Learning

$$D_U : x \sim p(x), D_S : (x, y) \sim p(x, y)$$



Basic assumptions of semi-supervised learning

- Continuity assumption
- Cluster assumption
- Manifold assumption

In deep semi supervised learning all three assumptions are connected

Label Propagation for Deep Semi-supervised Learning

Ahmet Iscen[1] Giorgos Tolias[1] Yannis Avrithis[2] Ondřej Chum[1]
[1]VRG, FEE, Czech Technical University in Prague [2]Univ Rennes, Inria, CNRS, IRISA

Abstract

Semi-supervised learning is becoming increasingly important because it can combine data carefully labeled by humans with abundant unlabeled data to train deep neural networks. Classic methods on semi-supervised learning that have focused on transductive learning have not been fully exploited in the inductive framework followed by modern deep learning. The same holds for the manifold assumption—that similar examples should get the same prediction. In this work, we employ a transductive label propagation method that is based on the manifold assumption to make predictions on the entire dataset and use these predictions to generate pseudo-labels for the unlabeled data and train a deep neural network. At the core of the transductive method lies a nearest neighbor graph of the dataset that we create based on the embeddings of the same network. Therefore our learning process iterates between these two steps. We improve performance on several datasets especially in the few labels regime and show that our work is complementary to current state of the art.

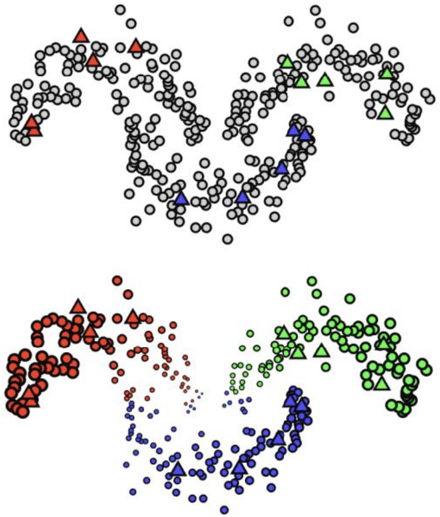
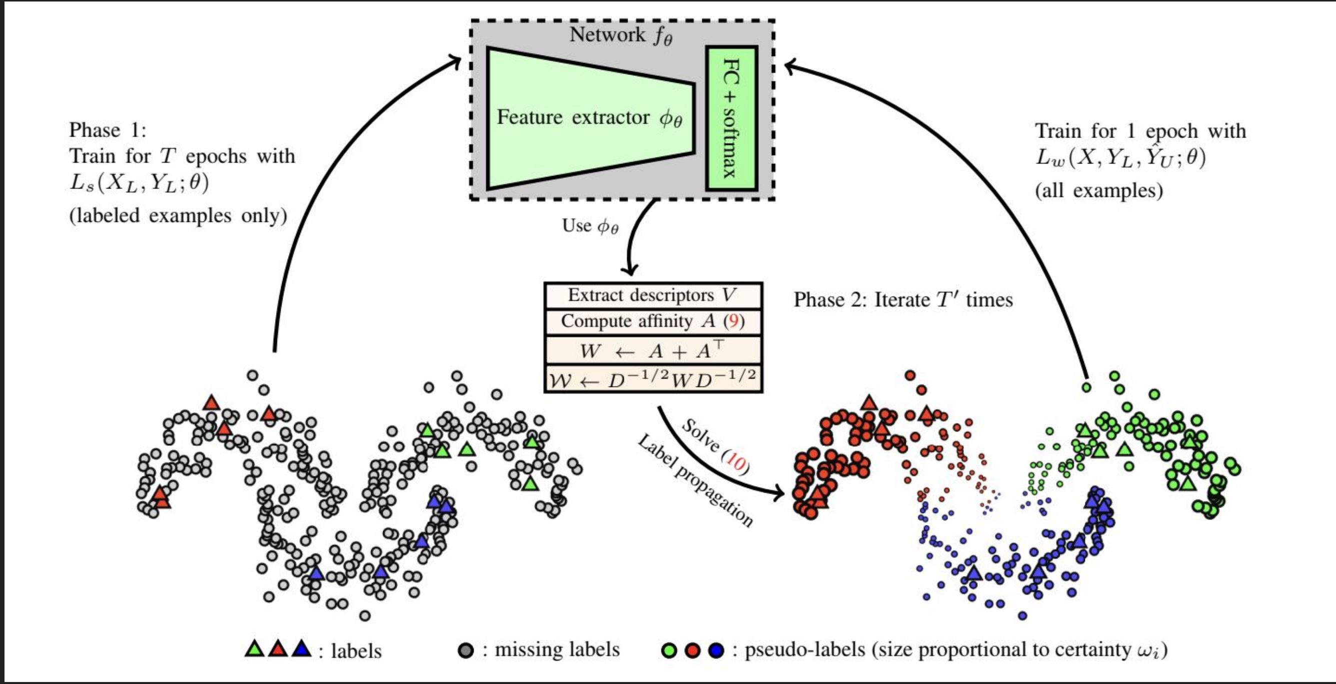


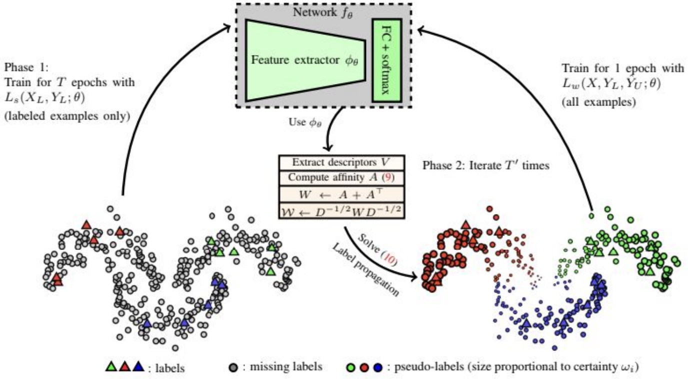
Figure 1. Label propagation on manifolds toy example. Triangles denote labeled, and circles un-labeled training data, respectively. Top: color-coded ground truth for labeled points, and gray color

Methodology:



Algorithm 1 Label propagation for deep SSL

procedure LPDSSL(Training examples X, labels Y_L)
  $\theta \leftarrow$ initialize randomly
  for epoch $\in [1, \dots, T]$ **do**
    $\theta \leftarrow \textsc{Optimize}(L_s(X_L, Y_L; \theta))$ ▷ mini-batch optimization
  end for
  for epoch $\in [1, \dots, T']$ **do**
    for $i \in \{1, \dots, n\}$ **do** $\mathbf{v}_i \leftarrow \phi_\theta(x_i)$ ▷ extract descriptors
    for $(i, j) \in \{1, \dots, n\}^2$ **do** $a_{ij} \leftarrow$ affinity values (9)
    $W \leftarrow A + A^\top$ ▷ symmetric affinity
    $\mathcal{W} \leftarrow D^{-1/2} W D^{-1/2}$ ▷ symmetrically normalized affinity
    $Z \leftarrow$ solve (10) with CG ▷ diffusion
    for $(i, j) \in U \times C$ **do** $\hat{z}_{ij} \leftarrow z_{ij} / \sum_k z_{ik}$ ▷ normalize Z
    for $i \in U$ **do** $\hat{y}_i \leftarrow \arg\max_j \hat{z}_{ij}$ ▷ pseudo-label
    for $i \in U$ **do** $\omega_i \leftarrow$ certainty of $\hat{y}_i$ (11) ▷ pseudo-label weight
    for $j \in C$ **do** $\zeta_j \leftarrow (|L_j| + |U_j|)^{-1}$ ▷ class weight/balancing
    $\theta \leftarrow \textsc{Optimize}(L_w(X, Y_L, \hat{Y}_U; \theta))$ ▷ mini-batch optimization
  end for
end procedure



$$a_{ij} := \begin{cases} [\mathbf{v}_i^\top \mathbf{v}_j]_+^\gamma, & \text{if } i \neq j \wedge \mathbf{v}_i \in \mathrm{NN}_k(\mathbf{v}_j) \\ 0, & \text{otherwise} \end{cases}$$

$$W := A + A^\top$$

$$\downarrow$$

$$(I - \alpha \mathcal{W})Z = Y,$$

conjugate gradient $\downarrow$

$$Z$$

$$L_w(X, Y_L, \hat{Y}_U; \theta) := \sum_{i=1}^{l} \zeta_{y_i} \ell_s \left(f_\theta(x_i), y_i\right) + \sum_{i=l+1}^{n} \omega_i \zeta_{\hat{y}_i} \ell_s \left(f_\theta(x_i), \hat{y}_i\right),$$

sample-wise confidence weight

class balancing weight

$$\zeta_j := \left(|L_j| + |U_j|\right)^{-1}$$

normalized over all classes

normalized over all samples

$$\omega_i := 1 - \frac{H(\hat{\mathbf{z}}_i)}{\log(c)},$$

Experiment:

- Dataset: CIFAR-10, CIFAR-100, Mini-ImageNet
- Augmentations: translation, (rotation), horizontal flip
- Classifiers: 13-layers CNN (CIFAR-10, CIFAR-100), ResNet-13 (Mini-ImageNet)
- L2 normalization before the FC layer (for graph construction)
- Training: After supervised learning with only labeled data, 180 epochs training.
- Cosine learning rate schedule (LR zero at 210 epochs)
- Initial LR = 0.05 (CIFAR-10), 0.2 (the others)

Result:

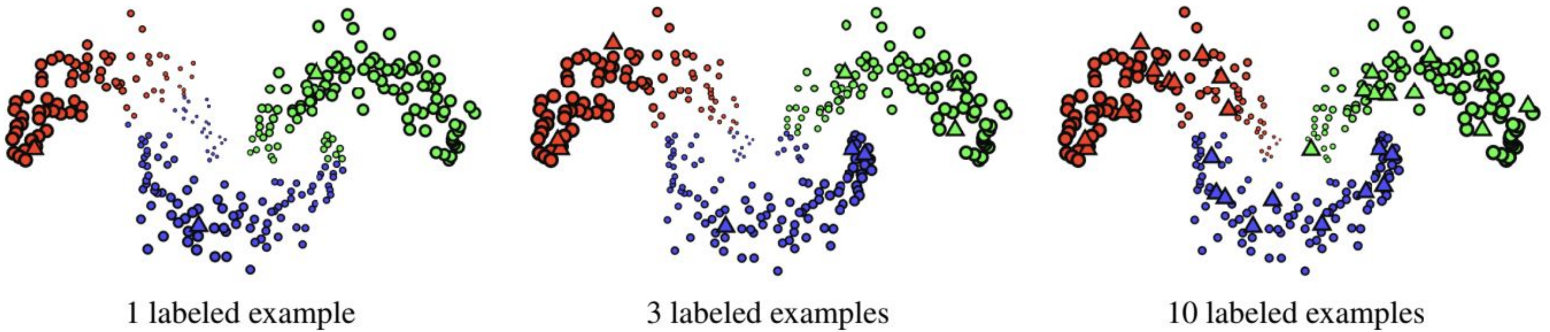


Figure 3. Toy example with 300 examples demonstrating label propagation for different number of labeled examples. Triangle markers correspond the labeled examples and circles to the unlabeled ones which are finally pseudo-labeled by label propagation. The class is color-coded and the size of the circles corresponds to weight ω_i. The true labels are the same as the example of Figure 1 (top).

Dataset	CIFAR-10			
Nb. labeled images	500	1000	2000	4000
Fully supervised	49.08 ± 0.83	40.03 ± 1.11	29.58 ± 0.93	21.63 ± 0.38
TDCNN [36]†	-	32.67 ± 1.93	22.99 ± 0.79	16.17 ± 0.37
Network prediction (1) + weights	35.17 ± 2.46	23.79 ± 1.31	16.64 ± 0.48	13.21 ± 0.61
Ours: Diffusion prediction (7) + weights	32.40 ± 1.80	22.02 ± 0.88	15.66 ± 0.35	12.69 ± 0.29
VAT [26]†	-	-	-	11.36
Π model [23]†	-	-	-	12.36 ± 0.31
Temporal Ensemble [23]†	-	-	-	12.16 ± 0.24
MT [38]†	-	27.36 ± 1.30	15.73 ± 0.31	12.31 ± 0.28
MT [38]	27.45 ± 2.64	19.04 ± 0.51	14.35 ± 0.31	11.41 ± 0.25
MT + Ours	$\mathbf{24.02 \pm 2.44}$	$\mathbf{16.93 \pm 0.70}$	$\mathbf{13.22 \pm 0.29}$	$\mathbf{10.61 \pm 0.28}$

Table 2. Comparison with the state of the art on CIFAR-10. Error rate is reported. "13-layer" network is used. The top part of the table corresponds to training with pseudo-labels, while the bottom part of the table includes methods that are complementary to ours, as shown by the combination of our method with MT. † denotes scores reported in prior work.

Dataset	CIFAR-100		Mini-ImageNet-*top1*		Mini-ImageNet-*top5*	
Nb. labeled images	4000	10000	4000	10000	4000	10000
Fully supervised	55.43 ± 0.11	40.67 ± 0.49	74.78 ± 0.33	60.25 ± 0.29	53.07 ± 0.68	38.28 ± 0.38
Ours	46.20 ± 0.76	38.43 ± 1.88	$\mathbf{70.29 \pm 0.81}$	57.58 ± 1.47	$\mathbf{47.58 \pm 0.94}$	36.14 ± 2.19
MT [38]	45.36 ± 0.49	36.08 ± 0.51	72.51 ± 0.22	57.55 ± 1.11	49.35 ± 0.22	32.51 ± 1.31
MT + Ours	$\mathbf{43.73 \pm 0.20}$	$\mathbf{35.92 \pm 0.47}$	72.78 ± 0.15	$\mathbf{57.35 \pm 1.66}$	50.52 ± 0.39	$\mathbf{31.99 \pm 0.55}$

Table 3. Performance comparison on CIFAR-100 and Mini-ImageNet with 4k and 10k labeled images. Error rate is reported. "13-layer" network is used for CIFAR-100 and Resnet-18 is used for Mini-ImageNet. All methods are reproduced by us.

Thank you

References

- Iscen, Ahmet, et al. "Label propagation for deep semi-supervised learning." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.
- https://drive.google.com/file/d/1T8VxIoxG2SaN4zo8bCzsao0_R5jUAQq2/view