Semi-Supervised Learning with Meta-Gradient

Xin-Yu Zhang[1], Taihong Xiao[2], Haolin Jia[3], Ming-Ming Cheng[1], Ming-Hsuan Yang[2]

[1]Nankai University, [2]University of California, Merced, [3]Tongji University

arXiv: https://arxiv.org/abs/2007.03966
Github: https://github.com/Sakura03/SemiMeta
Email: xinyuzhang@mail.nankai.edu.cn

Motivation

- Semi-supervised learning (SSL) aims at utilizing unlabeled data together with a small amount of labeled data to improve the generalization ability.
- Consistency-based SSL methods assume the predictions should be consistent against small perturbations of training data or parameters.
- Existing consistency-based algorithms **do not fully exploit the label information** when computing the consistency regularization.
- In this work, we propose a meta-learning algorithm in which the consistency loss is designed specifically for the underlying task.

Challenge

- The main challenge is that the consistency loss, which is calculated from the unlabeled data, seems to have no relationship with the labeled data.
- We borrow the idea of meta-learning, and build the relationship by unfolding and differentiating one SGD step, as shown in Fig. 1.

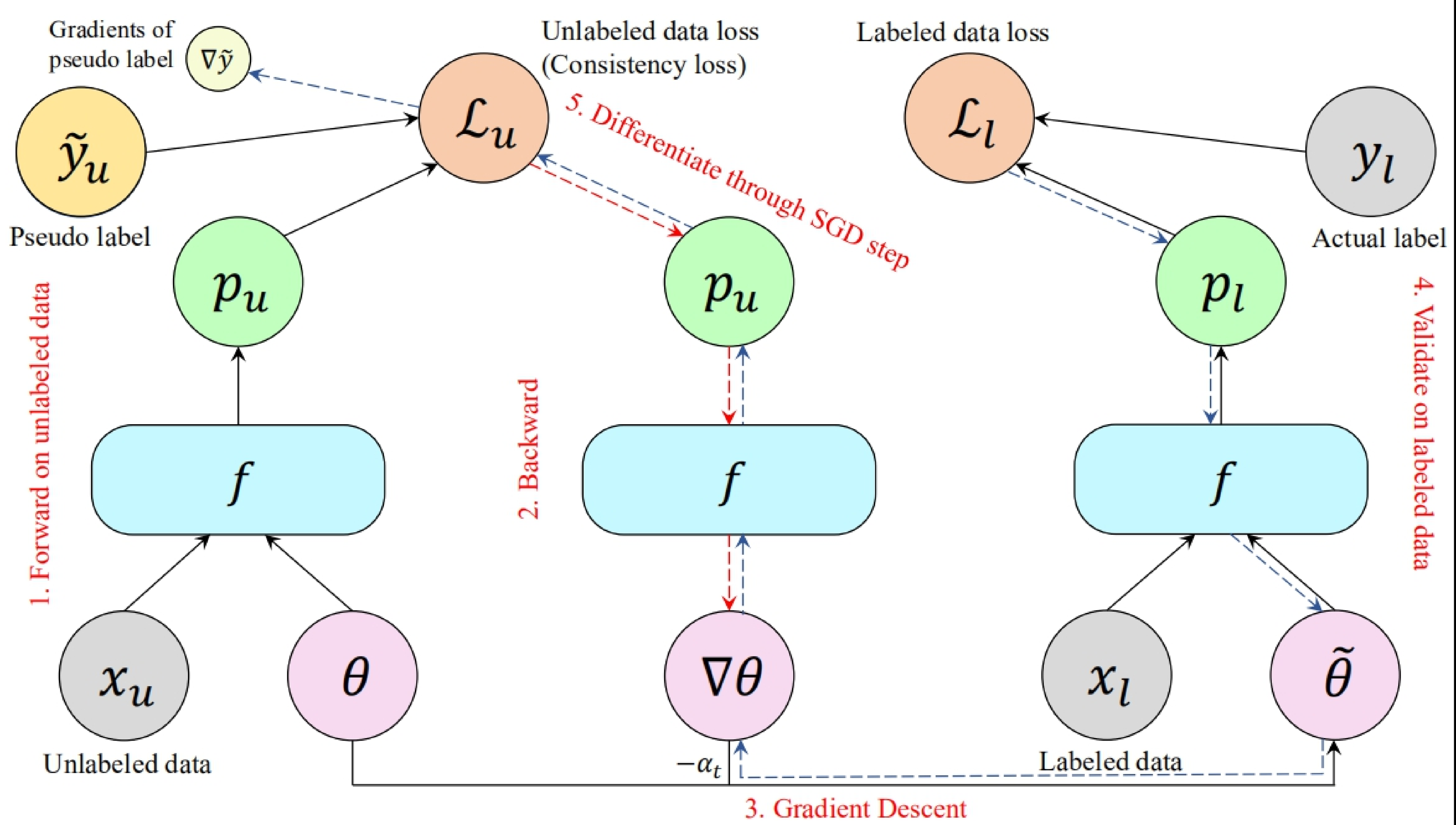


Fig 1. The magic of meta learning.

Algorithm

- Our formulation is as follows:

$$\min_{\mathcal{Y}} \sum_{k=1}^{N^l} \mathcal{L}(x_k^l,\ y_k; \theta^*(\mathcal{Y}))$$

$$s.t.\quad \theta^*(\mathcal{Y}) = \operatorname*{argmin}_{\theta} \sum_{i=1}^{N^u} \mathcal{L}(x_i^u,\ \hat{y}_i; \theta).$$

- Directly solving the above bi-level optimization problem is computationally prohibitive. We adopt an online approximation approach.

1. Initialize the pseudo labels of the unlabeled data:
$$\tilde{y}_i = f(x_i^u; \theta_t);$$
2. Compute the unlabeled data loss:
$$\mathcal{L}(x_i^u, \tilde{y}_i; \theta_t) = \Phi(f(x_i^u; \theta_t),\ \tilde{y}_i);$$
3. Back-propagate the unlabeled loss *w.r.t.* the model parameters:
$$\nabla\theta_t = \frac{1}{B^u}\sum_{i=1}^{B^u} \nabla_\theta \mathcal{L}(x_i^u, \tilde{y}_i; \theta_t);$$
4. Apply one SGD step on the model parameters:
$$\tilde{\theta}_{t+1} = \theta_t - \alpha_t \nabla\theta_t;$$
5. Evaluate the updated parameters on the labeled data and compute the labeled loss:
$$\mathcal{L}(x_k^l, y_k; \tilde{\theta}_{t+1}) = \Phi\big(f(x_k^l; \tilde{\theta}_{t+1}), y_k\big);$$
6. Back-propagate the labeled loss *w.r.t.* the pseudo labels:
$$\nabla\tilde{y}_i = \frac{1}{B^l}\sum_{k=1}^{B^l} \nabla_{\tilde{y}_i} \mathcal{L}(x_k^l, y_k; \tilde{\theta}_{t+1});$$
7. Perform one SGD step on the pseudo labels:
$$\hat{y}_i = \tilde{y}_i - \beta_t \nabla\tilde{y}_i;$$
8. Compute the consistency regularization with the updated pseudo labels:
$$\mathcal{L}_{cons} = \frac{1}{B^u}\sum_{i=1}^{B^u} \mathcal{L}(x_i^u, \hat{y}_i; \theta_t)$$

Remarks

- Since the initial pseudo labels are exactly the prediction of the current model, the unlabeled loss in Step 2 and the gradients in Step 3 are both zero. However, the update in Step 4 is still meaningful since the Jacobian of $\nabla\theta_t$ *w.r.t.* $\tilde{y}_i$ is non-zero, which lays in the essence of the relationship between consistency loss and labeled data;
- The equation in Step 6 actually involves the second-order derivative. We avoid this by the first-order approximation (see our paper);
- In Step 8, $\hat{y}_i$ is "detached" from θ_t, namely, the gradients of $\hat{y}_i$ do not propagate to θ_t when differentiating $\mathcal{L}_{cons}$.

Convergence Analysis

Theorem 1 (Convergence guarantee). Let

$$G(\theta; \mathcal{D}^l) = \frac{1}{N^l}\sum_{k=1}^{N^l} \nabla_\theta \mathcal{L}(x_k^l, y_k; \theta)$$

be the loss of the labeled data. Under mild conditions (see our paper), as long as the regular learning rate α_t and meta learning rate β_t are sufficiently small, each SGD step will decrease the validation loss $G(\theta)$, regardless of the selected unlabeled batch, *i.e.*,

$$G(\theta_{t+1}) \leq G(\theta_t),\ \ for\ each\ \ t.$$

Theorem 2 (Convergence rate). Under the same condition as above, as long as the learning rates α_t, β_t are moderate (not too small or too large), then the meta-learning algorithm achieves $\mathbb{E}[\|\nabla_\theta G(\theta_t)\|^2] \leq \varepsilon$ in $O(1/\varepsilon^2)$ steps, *i.e.*,

$$\min_{1\leq t\leq T} \mathbb{E}[\|\nabla_\theta G(\theta_t)\|^2] \leq \frac{C}{\sqrt{T}},$$

where C is a constant independent of the training process.

Future Perspective

- Incorporation with the concurrent self-supervised SSL methods, *e.g.*, FixMatch.
- Extension beyond semi-supervised classification, *e.g.*, weakly-supervised segmentation.
- (May be too aggressive) Generalize the current bi-level optimization towards multi-level optimization.

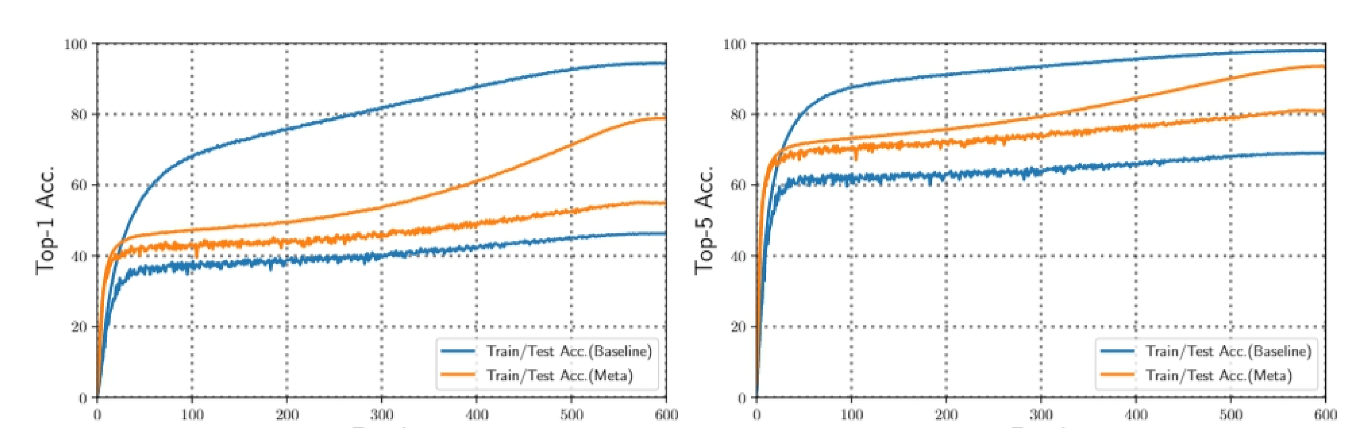
Experiments

- Comparison with SOTA methods on SVHN, CIFAR datasets.

Method	SVHN	CIFAR-10	CIFAR-100
Π-Model (Laine and Aila, 2017)	4.82%	12.36%	39.19%
TE (Laine and Aila, 2017)	4.42%	12.16%	38.65%
MA-DNN (Chen et al., 2018)	4.21%	11.91%	34.51%
Co-training (Qiao et al., 2018)	3.29%	8.35%	34.63%
MT+fastSWA (Athiwaratkun et al., 2019)	-	9.05%	33.62%
TNAR-VAE (Yu et al., 2019)	3.74%	8.85%	-
ADA-Net (Wang et al., 2019)	4.62%	10.30%	-
Ours	**3.15%**	**7.78%**	**30.74%**
Fully-Supervised	2.67%	4.88%	22.10%

- Performance on ImageNet dataset and the training/testing accuracy curves.

Method	Top-1	Top-5
Labeled-Only	53.65%	31.01%
MT	49.07%	23.59%
Co-training	46.50%	22.73%
ADA-Net	44.91%	21.18%
Ours	**44.87%**	**18.88%**
Fully-Supervised	29.15%	10.12%



- Feature visualization of the baseline method and ours.