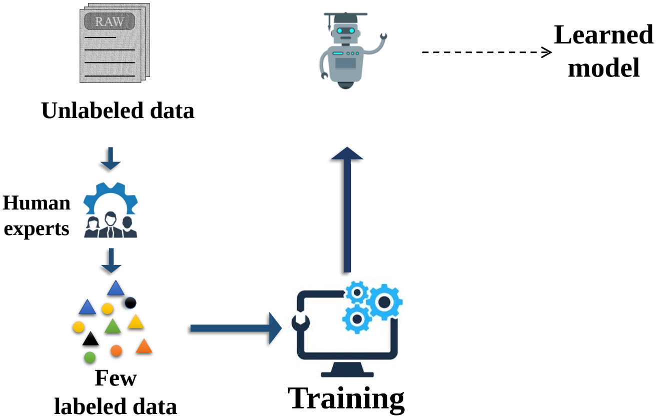
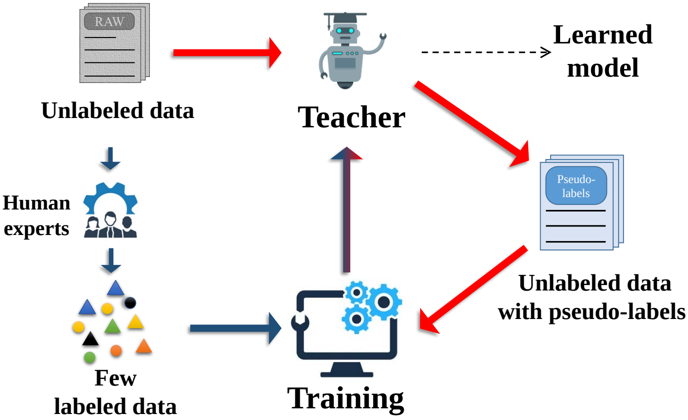
Self-training



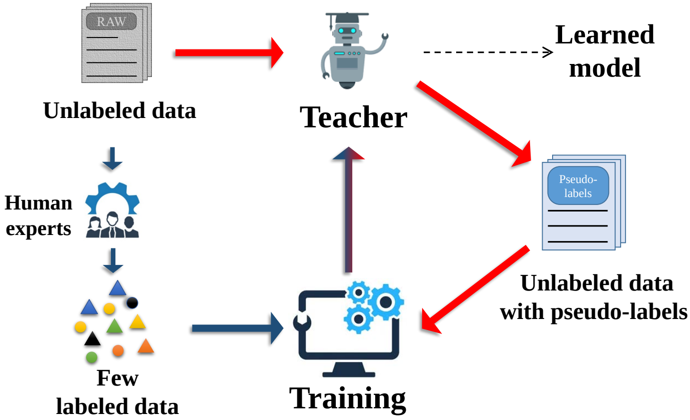
Self-training



Self-training

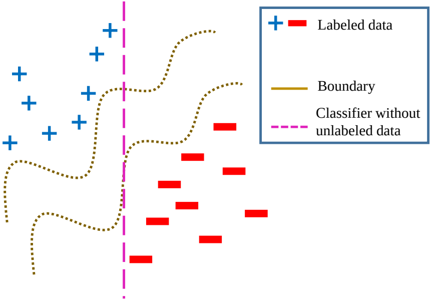
Why self-training?

- Labeled data is hard to get while unlabeled data is cheap.
- Unlabeled data can improve the performance.



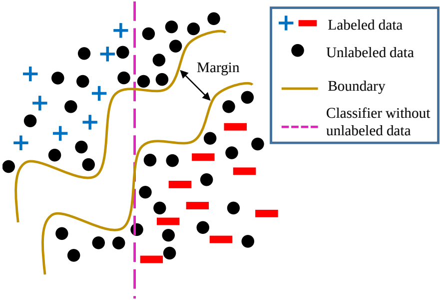
Related Works: Benefits of unlabeled data and Limitations of Self-training

- Unlabeled data benefit boundary identification.



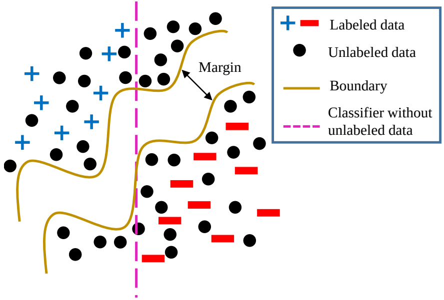
Related Works: Benefits of unlabeled data and Limitations of Self-training

- Unlabeled data benefit boundary identification.



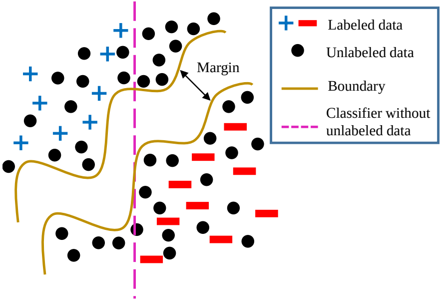
Related Works: Benefits of unlabeled data and Limitations of Self-training

- Unlabeled data benefit boundary identification.
- Limitations of existing theoretical works:
  - ☐ *Linear models* [Chen et al.20a, Raghunathan et al.20, Oymak and Gulcu.20].
  - ☐ Unlabeled data in non-linear model can sometime hurt the performance [Wei et al.20].
  - ☐ Infinite number of unlabeled data.



Related Works: Benefits of unlabeled data and Limitations of Self-training

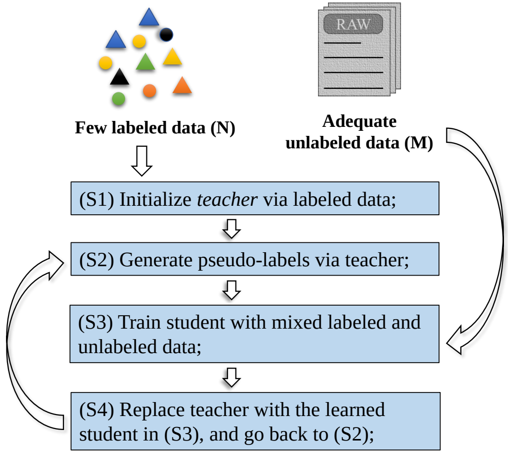
- Unlabeled data benefit boundary identification.
- Limitations of existing theoretical works:
  - ☐ *Linear models* [Chen et al.20a, Raghunathan et al.20, Oymak and Gulcu.20].
  - ☐ Unlabeled data in non-linear model can sometime hurt the performance [Wei et al.20].
  - ☐ Infinite number of unlabeled data.



Question?

1. *How to set hyperparameters* that ensure enhanced accuracy?
2. *How much unlabeled data* is required to obtain a specific improvement in test accuracy?

Iterative Self-training Algorithm



Input: Labeled data $\mathcal{D} = \{\boldsymbol{x}_n, y_n\}_{n=1}^{N}$, unlabeled data $\widetilde{\mathcal{D}} = \{\tilde{\boldsymbol{x}}_m\}_{m=1}^{M}$, and loss parameter λ.

1. Obtain teacher $\boldsymbol{W}^{(0)}$ by minimizing $f_{\mathcal{D}}(W)$ with respect to labeled data.

For $\ell = 0, 1, 2, \cdots, L$ do

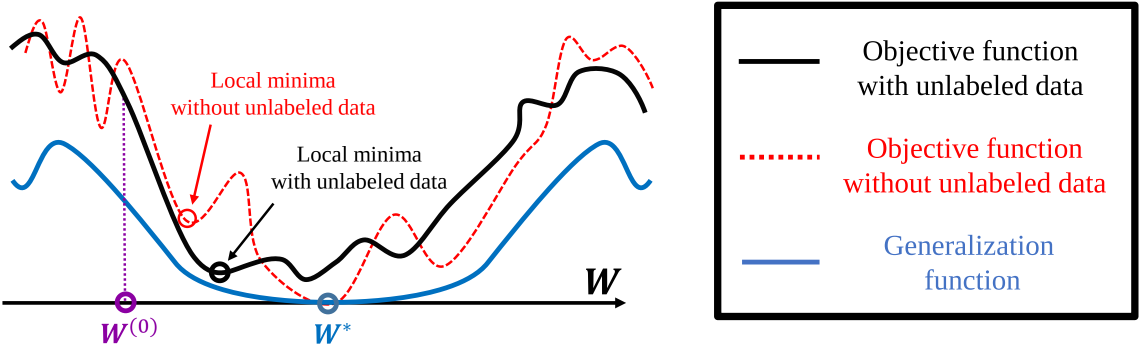
2. Generate pseudo-labels $\tilde{y}_m^{(\ell)}$ for the unlabeled data in $\widetilde{\mathcal{D}}$ using teacher $\boldsymbol{W}^{(\ell)}$, i.e., $\tilde{y}_m^{(\ell)} = g(\boldsymbol{W}^{(\ell)}; \tilde{\boldsymbol{x}}_m)$.
3. Train a student $\widehat{\boldsymbol{W}}$ by minimizing:

$$f(\boldsymbol{W}) = \lambda \cdot f_{\mathcal{D}} + (1 - \lambda) \cdot f_{\widetilde{\mathcal{D}}}^{(\ell)}.$$

4. Set the student $\widehat{\boldsymbol{W}}$ as the new teacher $\boldsymbol{W}^{(\ell+1)}$ and $\ell \longleftarrow \ell + 1$.

Intuition From the Landscape Analysis

Adding unlabeled data can shift the convergent point towards the desired model $\boldsymbol{W}^\star$.



[Zhang et al. ICLR'22]

Main Theoretical Findings

Takeaway: iteration $\{\boldsymbol{W}^{(\ell)}\}_{\ell=1}^{L}$ converge linearly to ground truth $\boldsymbol{W}^*$ up to bounded error term depending on λ and unlabeled data amount M.

- $\boldsymbol{W}^*$: the desired model.
- N^*: the *required* labeled data for finding W^*. (Desired number of labeled data we want)

[Zhang et al. ICLR'22]

Main Theoretical Findings

Takeaway: iteration $\{\boldsymbol{W}^{(\ell)}\}_{\ell=1}^{L}$ converge linearly to ground truth $\boldsymbol{W}^*$ up to bounded error term depending on λ and unlabeled data amount M.

- W^*: the desired model.
- N^*: the *required* labeled data for finding W^*. (Desired number of labeled data we want)
- N: the number of labeled data we have, $N < N^*$. (Actual number of labeled data we have)
- M: the number of unlabeled data.
- $W^{(0)}$: the initial weights learned from N labeled data.

[Zhang et al. ICLR'22]

Main Theoretical Findings

Takeaway: iteration $\{\boldsymbol{W}^{(\ell)}\}_{\ell=1}^{L}$ converge linearly to ground truth $\boldsymbol{W}^*$ up to bounded error term depending on λ and unlabeled data amount M.

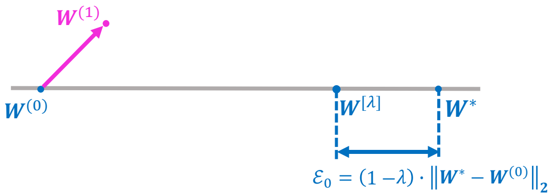
- W^*: the desired model.
- N^*: the *required* labeled data for finding W^*. (Desired number of labeled data we want)
- N: the number of labeled data we have, $N < N^*$. (Actual number of labeled data we have)
- M: the number of unlabeled data.
- $W^{(0)}$: the initial weights learned from N labeled data.
- $W^{[\lambda]}$: $\boldsymbol{W}^{[\lambda]} = (1-\lambda)\boldsymbol{W}^{(0)} + \lambda \boldsymbol{W}^*$.

  $\lambda \in \left[\frac{1}{2}, \sqrt{\frac{N}{N^*}}\right]$.

[Zhang et al. ICLR'22]

Main Theoretical Findings

Takeaway: iteration $\{\boldsymbol{W}^{(\ell)}\}_{\ell=1}^{L}$ converge linearly to ground truth $\boldsymbol{W}^*$ up to bounded error term depending on λ and unlabeled data amount M.



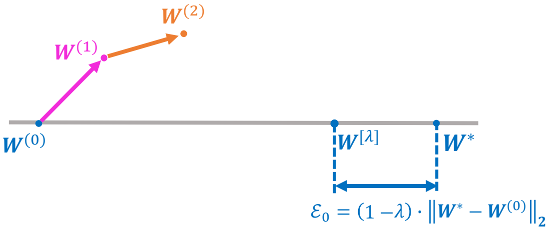
- $\boldsymbol{W}^*$: the desired model.
- N^*: the *required* labeled data for finding W^*. (Desired number of labeled data we want)
- N: the number of labeled data we have, $N < N^*$. (Actual number of labeled data we have)
- M: the number of unlabeled data.
- $\boldsymbol{W}^{(0)}$: the initial weights learned from N labeled data.
- $\boldsymbol{W}^{[\lambda]}$: $\boldsymbol{W}^{[\lambda]} = (1-\lambda)\boldsymbol{W}^{(0)} + \lambda \boldsymbol{W}^*$.

  $\lambda \in \left[\frac{1}{2}, \sqrt{\frac{N}{N^*}}\right]$.

[Zhang et al. ICLR'22]

Main Theoretical Findings

Takeaway: iteration $\{\boldsymbol{W}^{(\ell)}\}_{\ell=1}^{L}$ converge linearly to ground truth $\boldsymbol{W}^*$ up to bounded error term depending on λ and unlabeled data amount M.



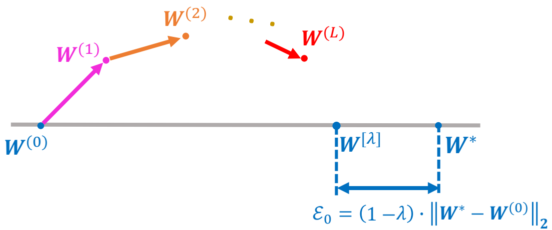
- $\boldsymbol{W}^*$: the desired model.
- N^*: the *required* labeled data for finding W^*. (Desired number of labeled data we want)
- N: the number of labeled data we have, $N < N^*$. (Actual number of labeled data we have)
- M: the number of unlabeled data.
- $\boldsymbol{W}^{(0)}$: the initial weights learned from N labeled data.
- $\boldsymbol{W}^{[\lambda]}$: $\boldsymbol{W}^{[\lambda]} = (1-\lambda)\boldsymbol{W}^{(0)} + \lambda \boldsymbol{W}^*$.

  $\lambda \in \left[\frac{1}{2}, \sqrt{\frac{N}{N^*}}\right]$.

[Zhang et al. ICLR'22]

Main Theoretical Findings

Takeaway: iteration $\{\boldsymbol{W}^{(\ell)}\}_{\ell=1}^{L}$ converge linearly to ground truth $\boldsymbol{W}^*$ up to bounded error term depending on λ and unlabeled data amount M.



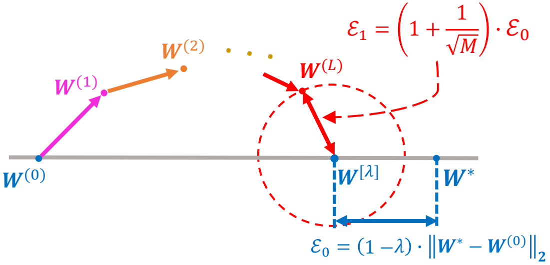
- W^*: the desired model.
- N^*: the *required* labeled data for finding W^*. (Desired number of labeled data we want)
- N: the number of labeled data we have, $N < N^*$. (Actual number of labeled data we have)
- M: the number of unlabeled data.
- $W^{(0)}$: the initial weights learned from N labeled data.
- $W^{[\lambda]}$: $\boldsymbol{W}^{[\lambda]} = (1-\lambda)\boldsymbol{W}^{(0)} + \lambda \boldsymbol{W}^*$.

  $\lambda \in \left[\frac{1}{2}, \sqrt{\frac{N}{N^*}}\right]$.
- $W^{(L)}$: the convergent point.

[Zhang et al. ICLR'22]

Main Theoretical Findings

Takeaway: iteration $\{\boldsymbol{W}^{(\ell)}\}_{\ell=1}^{L}$ converge linearly to ground truth $\boldsymbol{W}^*$ up to bounded error term depending on λ and unlabeled data amount M.



[Zhang et al. ICLR'22]

- $\boldsymbol{W}^*$: the desired model.
- N^*: the *required* labeled data for finding W^*. (Desired number of labeled data we want)
- N: the number of labeled data we have, $N < N^*$. (Actual number of labeled data we have)
- M: the number of unlabeled data.
- $\boldsymbol{W}^{(0)}$: the initial weights learned from N labeled data.
- $\boldsymbol{W}^{[\lambda]}$: $\boldsymbol{W}^{[\lambda]} = (1-\lambda)\boldsymbol{W}^{(0)} + \lambda \boldsymbol{W}^*$.

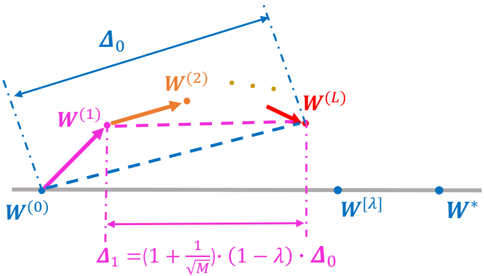
  $\lambda \in \left[\frac{1}{2}, \sqrt{\frac{N}{N^*}}\right]$.
- $\boldsymbol{W}^{(L)}$: the convergent point.

  Generalization error:

$$\|\boldsymbol{W}^{(L)} - \boldsymbol{W}^*\|_2 \leq \varepsilon_0 + \varepsilon_1.$$

Main Theoretical Findings

Takeaway: iteration $\{\boldsymbol{W}^{(\ell)}\}_{\ell=1}^{L}$ converge linearly to ground truth $\boldsymbol{W}^*$ up to bounded error term depending on λ and unlabeled data amount M.

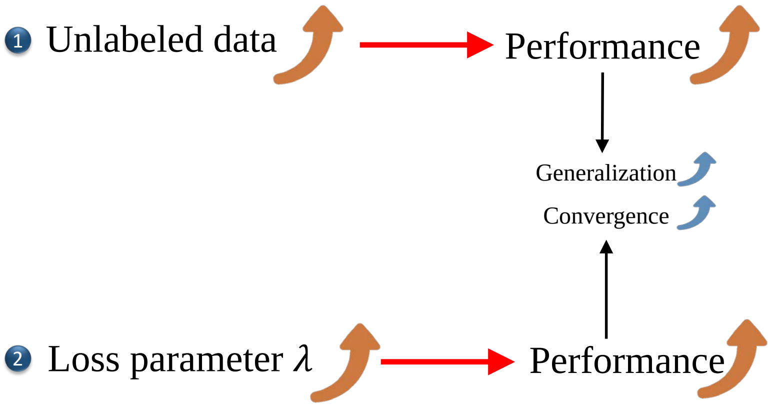


[Zhang et al. ICLR'22]

- W^*: the desired model.
- N^*: the *required* labeled data for finding W^*. (Desired number of labeled data we want)
- N: the number of labeled data we have, $N < N^*$. (Actual number of labeled data we have)
- M: the number of unlabeled data.
- $W^{(0)}$: the initial weights learned from N labeled data.
- $W^{[\lambda]}$: $\boldsymbol{W}^{[\lambda]} = (1-\lambda)\boldsymbol{W}^{(0)} + \lambda \boldsymbol{W}^*$. $\lambda \in \left[\frac{1}{2}, \sqrt{\frac{N}{N^*}}\right]$.
- $W^{(L)}$: the convergent point. Convergence rate:

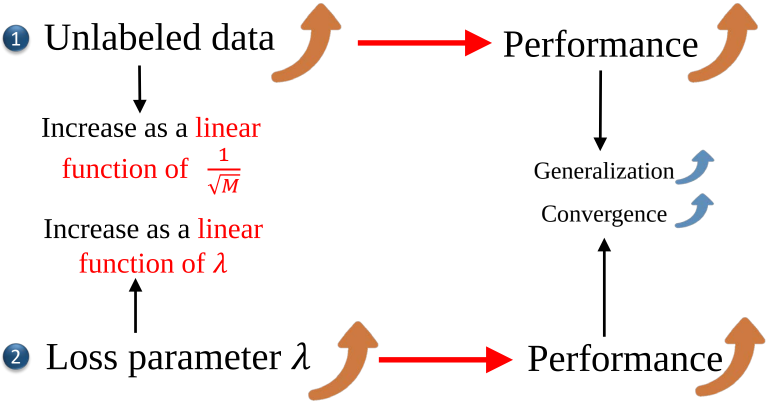
$$\frac{\Delta_1}{\Delta_0} \leq (1+\frac{1}{\sqrt{M}}) \cdot (1-\lambda).$$

Insights of the Theoretical Results



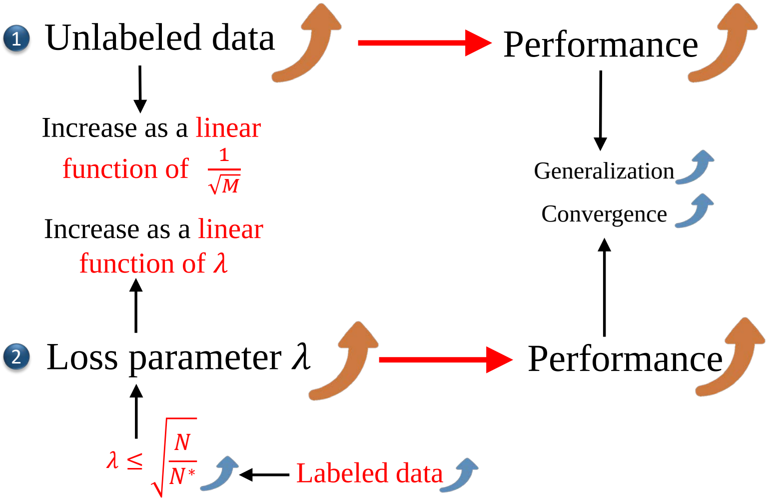
[Zhang et al. ICLR'22]

Insights of the Theoretical Results



[Zhang et al. ICLR'22]

Insights of the Theoretical Results



[Zhang et al. ICLR'22]

Empirical Results: ResNet-28 on CIFAR-10

- Ten-class image classification: Labeled data from CIFAR-10, unlabeled data from Tiny Images, 28-layer ResNet.
- CIFAR-10 dataset contains 60,000 32x32 color images in 10 different classes.
- ResNet: network with Residual blocks via skip connections.

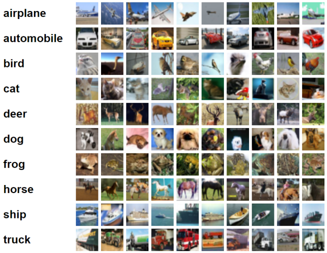


Figure 8: Illustration of the CIFAR-10 dataset (labeled subsets of Tiny Images)

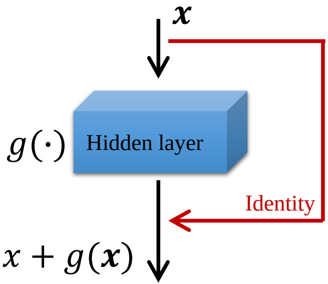


Figure 9: Illustration of Residual modular

Empirical Results: ResNet-28 on CIFAR-10

- From the line with rectangular mark ($N = 50K$), the test accuracy is improved by 7% by using unlabeled data (82.79% to 89.61% as the unlabeled data from 0 to 500K).
- The improved test accuracy and convergence rate are in the order of $1/\sqrt{M}$, matching our theoretical findings.

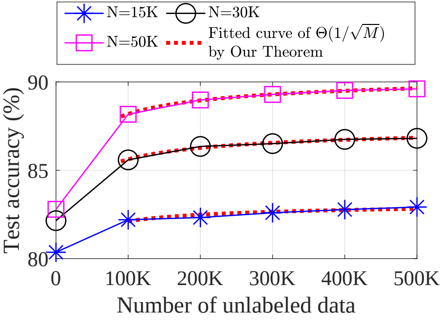


Figure 9: The test accuracy against the number of unlabeled data M

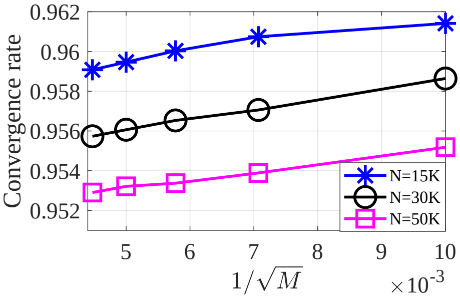


Figure 10: The convergence rate against the number of unlabeled data M

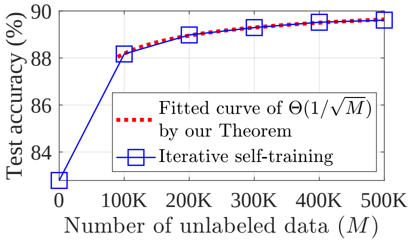
Self-training for Sample Efficient Deep Learning

Self-training algorithms augment limited labeled data with a large size of unlabeled data.

- Unlabeled data is widely available while labeled data is expensive.
- Unlabeled data can improve the performance

Our contributions:

- Theoretical guidance for the hyperparameter selection with guaranteed improved performance.
- Quantitative characterization of unlabeled data amount improves performance with theoretical guarantees.



[Zhang et al. ICLR'22]