Improved Regularization and Robustness for Fine-tuning in Neural Networks

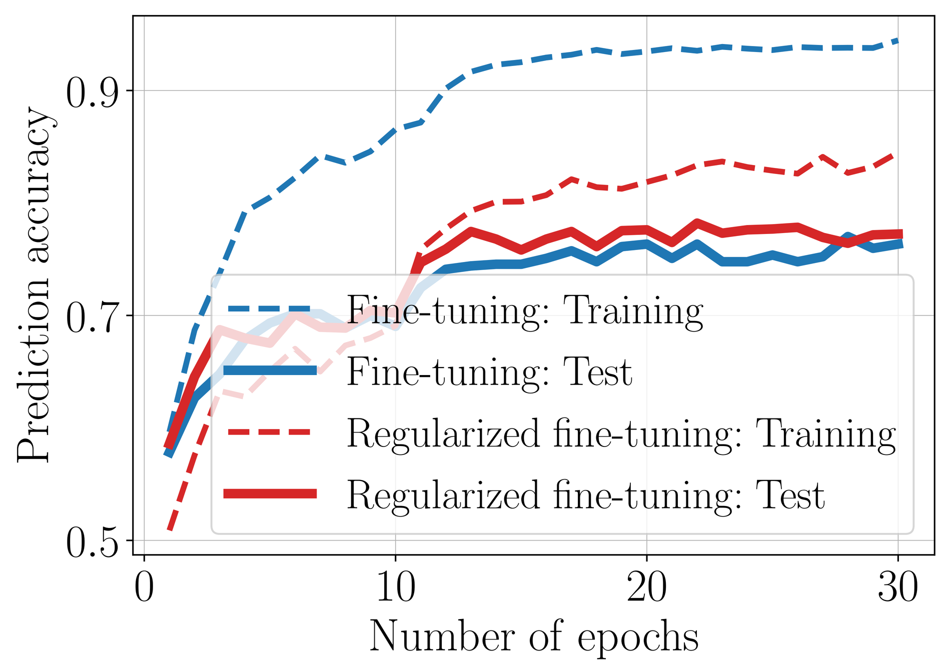
Dongyue Li and Hongyang R. Zhang

Northeastern University

Motivation

Fine-tuning a pre-trained model on a target data set with limited labels has been successful.

However, fine-tuning is prone to overfitting. Regularization helps alleviate this issue.



Mystery around fine-tuning:

- Regularization may or may not help fine-tuning; not well-understood (Li et al., 2020).
- Applying adversarial training during pre-training leads to models with better transfer to downstream tasks (Salman et al., 2020).

Practical concern in fine-tuning:

- Label noise is common in transfer learning, for example, if labels are created via weak supervision (Ratner et al., 2016).

Problem Setup

Data. $(x_1^{(t)}, y_1^{(t)}), \dots, (x_{n^{(t)}}^{(t)}, y_{n^{(t)}}^{(t)}) \in P^{(t)}$.

Model. L-layer feed-forward neural networks

$$f_W(x) = \phi_L \circ \phi_{L-1} \circ \cdots \phi_1(x)$$

where $\phi_i(z) = \psi_i(W_i z)$ and $W = [W_1, \dots, W_L]$.

Prediction error.

$$\mathcal{L}^{(t)}(f_W) = \mathbb{E}_{(x,y)\sim\mathcal{P}^{(t)}}\left[\ell(f_W(x), y)\right].$$

where $\ell(\cdot)$ is both convex and 1-Lipschitz.

Regularized fine-tuning problem (Gouk et al., 2021).

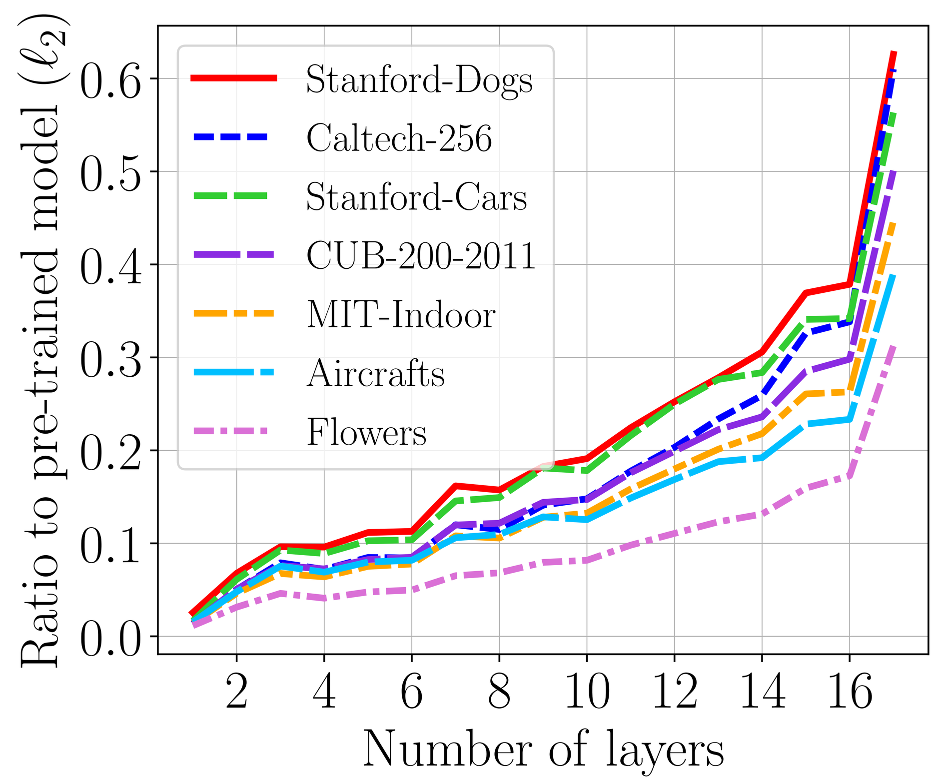
$$\hat{W} \leftarrow \arg\min \hat{\mathcal{L}}^{(t)}(f_W)$$

$$\text{s.t. } \|W_i - \hat{W}_i^{(s)}\|_F \le D_i,\ \forall\, i = 1, \dots, L.$$

Regularization methods

Through a PAC-Bayesian analysis, we identify **two key components** to determine the fine-tuning generalization performance.

- **Layer-wise Distances $\{\mathbf{D_i}\}_{\mathbf{i=1,\dots,L}}$.** Fine-tuned distances are relatively small compared to the pre-trained network and grow with layers.



Implication. *If we set the same value for D_i, only regularize top layers. Instead, set D_i proportional to fine-tuned distance.*

- **Perturbed Loss** $\mathbb{E}_{\mathbf{U}}[\ell(\mathbf{f}_{\hat{\mathbf{W}}+\mathbf{U}}(\mathbf{x}), \mathbf{y})]$. Models fine-tuned from a pre-trained initialization (Fine-tune) are *more stable than models trained from random initialization (Random).*

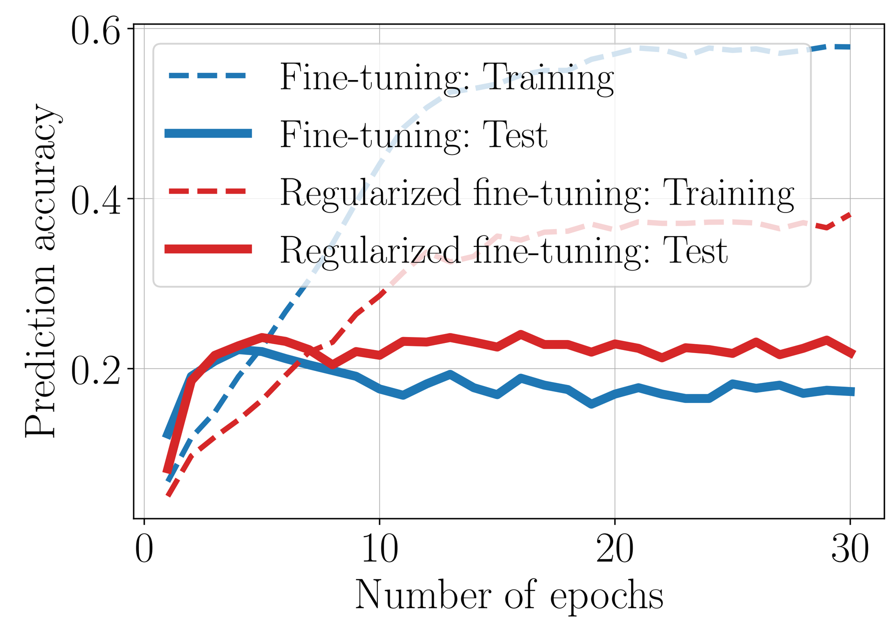
Adversarial robustness. *Fine-tuning from adversarial pretrained models incurs lower perturbed losses than fine-tuning from (standard) pre-trained initializations.*

		CUB-200-2011	
σ	Random	Pre-trained	Adversarial
10^{-2}	3.77±0.42	**1.45±0.13**	1.76±0.09
10^{-3}	0.82±0.07	0.62±0.03	**0.54±0.03**
10^{-4}	0.81±0.04	0.61±0.03	**0.61±0.01**
		Indoor	
σ	Random	Pre-trained	Adversarial
10^{-2}	2.51±0.34	1.11±0.09	**0.97±0.07**
10^{-3}	0.49±0.09	0.36±0.05	**0.32±0.04**
10^{-4}	0.44±0.03	0.33±0.02	**0.30±0.04**

Robustness w.r.t. label noise

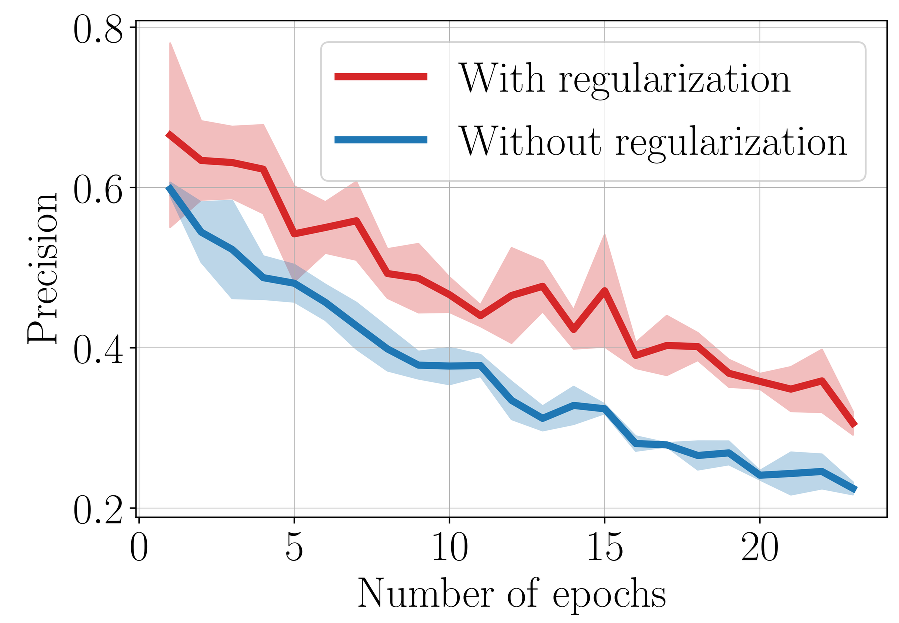
Memorizing during fine-tuning.

- Test accuracy increases at first and ends up overfitting to noisy labels.
- Though regularization can improve performance, the generalization error is still large.
- The model picks up some discriminative power in the starting phase.

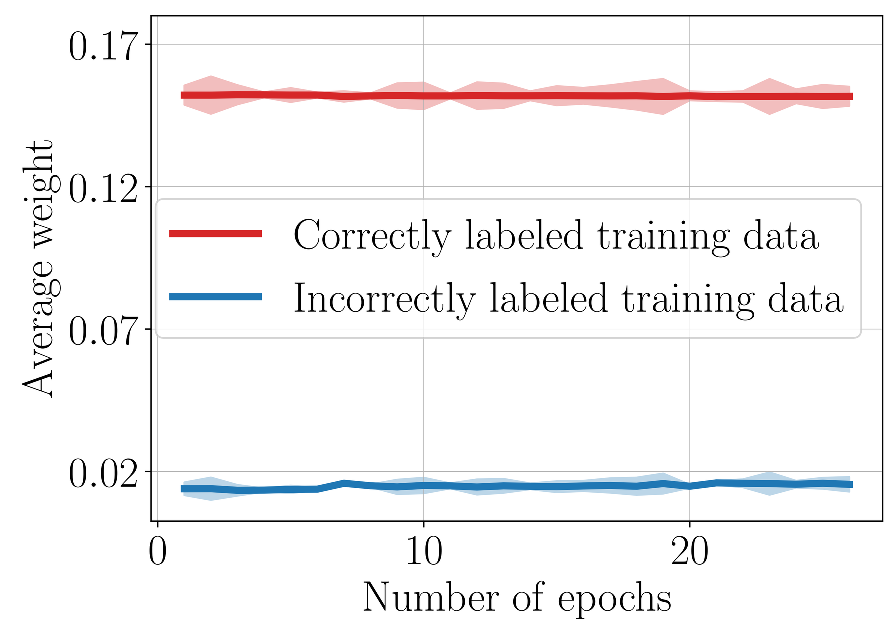


Implication. Leverage the model predictions to relabel the data.

- **Self label-correction** relabels the data point if the model is confident.



- **Self label-removal** down-weights the loss of large-loss data.



Experimental Results

Regularized self-labeling (RegSL) combines both layer-wise regularization and self-labeling.

Comparing regularization methods.

Fine-tuning ResNet-101 on seven image classification data sets.

- Average **1.76%** improvement compared to the constant regularization (Gouk et al., 2021).

Fine-tuning ResNet-18 on ChestX-ray14 data set (Wang et al., 2017; Rajpurkar et al., 2017).

Comparing robustness w.r.t. label noise.

Fine-tuning ResNet-18 on MIT-Indoor data set with both independent and correlated noises.

- Average **3.56%** improvement over regularization methods and previous supervised training methods.

Fine-tuning Vision Transformer (Dosovitskiy et al., 2020) on noisy labels.

Key insight.

We identify a pipeline of applying regularization.

1. Run fine-tuning on the pre-trained network and plot the "fine-tuned" layer-wise distances.
2. Encode the layer-wise distance patterns using explicit regularization constraints.

Regularization and self-labeling complement each other during fine-tuning.

Methods	independent noise 20%	correlated noise 25.18%
RegSL (ours)	**72.51±0.46**	**70.12±0.83**
w/o regularization	71.94±0.43	69.43±0.36
w/o self-labeling	70.23±0.25	69.05±0.09

Extension to other settings.

Few-show image classification tasks (fine-tuning ResNet-12 over 600 meta-test splits of miniImageNet.);

Transfer learning in sentence classification tasks (training three-layer MLPs on SST , MR , CR , MPQA , SUBJ , and TREC).