Robust Fine-Tuning of Deep Neural Networks with Hessian-based Generalization Guarantees

Dongyue Li

Northeastern University, Boston, MA

Joint work w/ Haotian Ju and Hongyang Zhang

Fine-tuning in Transfer Learning

- A modern approach for transfer learning: Download a pretrained network; Fine-tune it on a target task.
  - **Pretrained models** (e.g., ResNet, BERT, GPT-3, CLIP) are accessible online.
  - **Fine-tuning:** Train the whole network initialized from pretrained weights [HGD'19, RSR+'20].

BERT

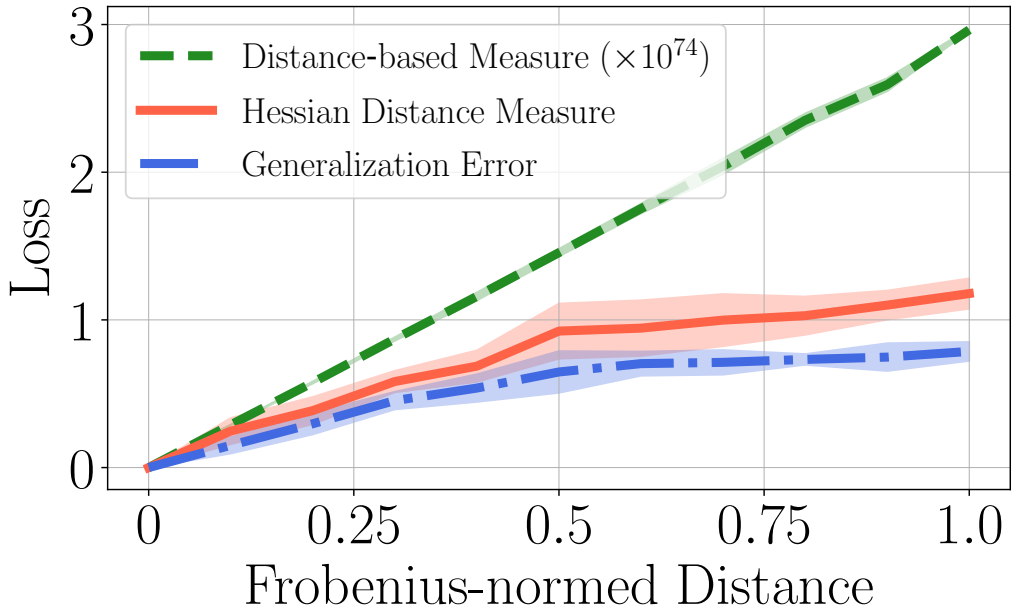
GPT-3

T-5

- Fine-tuning is also an efficient approach in multi-task learning [Rud'17], meta-learning [FAL'17], and zero-shot learning [WIL+'22].

Generalization of Fine-tuned Models

- What is being transferred during fine-tuning? [NSZ'20]
  - Previous works highlight the role of **distance from initialization** for studying the generalization properties of fine-tuning [GHP'21; LZ'21].
  - Motivating observation: In addition to distance from initialization, **Hessians affect generalization** (through noise stability of a model against noise injection).



PAC-Bayesian Bounds

- **Generalization error of fine-tuned models:**
  - f_W: a multi-layer feedforward neural network with weight W.
  - $L(f_W)$: expected loss.
  - $\hat{L}(f_W)$: empirical loss.
  - $L(f_W) - \hat{L}(f_W)$: generalization error.

- **Technique:**
  - **PAC-Bayesian bounds** [McA'99; McA'01]: Crucial in studying generalization errors of deep neural nets [NBS'18; JNM+'20].
  - **Noise stability:** A key notion to quantify the loss stability of a model under noise perturbations. [AGN+'18].
    - $I(f_W) = \mathbb{E}[L(f_{W+U})] - L(f_W)$: loss stability under noise perturbation U.

Noise Stability and Hessians

- **Claim:** $I(f_W) = \sum \langle \Sigma_i, H_i[L(f_W)] \rangle + \xi$.
  - Σ_i: covariance matrix of U_i at layer i.
  - H_i: loss Hessian matrix at layer i.
- Comparing Hessian approximation to noise stability on CIFAR-100.
  - The Hessian approximation is accurate in practice.

σ	ResNet-18		ResNet-50	
	Noise Stability	Hessian Approx.	Noise Stability	Hessian Approx.
0.01	0.86 ± 0.19	1.07	0.39 ± 0.10	1.24
0.011	1.13 ± 0.24	1.29	1.12 ± 0.20	1.50
0.012	1.45 ± 0.29	1.54	1.56 ± 0.26	1.79
0.013	1.82 ± 0.35	1.80	2.10 ± 0.33	2.10
0.014	2.23 ± 0.40	2.09	2.71 ± 0.38	2.43
0.015	2.65 ± 0.43	2.34	3.30 ± 0.40	2.79
0.016	3.07 ± 0.45	2.73	3.80 ± 0.40	3.18
0.017	3.47 ± 0.47	3.08	4.17 ± 0.40	3.59
0.018	3.84 ± 0.49	3.46	4.60 ± 0.44	4.02
0.019	4.15 ± 0.51	3.85	4.77 ± 0.44	4.48
0.020	4.43 ± 0.55	4.27	5.03 ± 0.82	4.97
Relative RSS	**0.75%**		**2.98%**	

Result: Hessian-based Generalization Bounds

- **Theorem** (informal):
  - $l(\cdot,\cdot)$: a loss function bounded between 0 and C.
  - v_i: the weight vector from initialization of $W_i - W_i^{(s)}$ at layer i.
  - $H_i^+[l(f_W(x),\ y)]$: the loss Hessian matrix at given sample (x, y).
  - n: the size of training dataset.
  - The generalization error of f_W scales as

$$O\left(\frac{\sum_{i=1}^{L}\sqrt{C\max_{x,y} v_i^T H_i^+[l(f_W(x),y)]v_i}}{\sqrt{n}}\right)$$

Result: Applies to Various Fine-tuning Methods

- Our result applies to a wide range of fine-tuning methods.
  - Distance-based regularization: $\left|\left|W_i - W_i^{(s)}\right|\right|_F \leq \alpha_i$, for $i = 1, \cdots, L$.

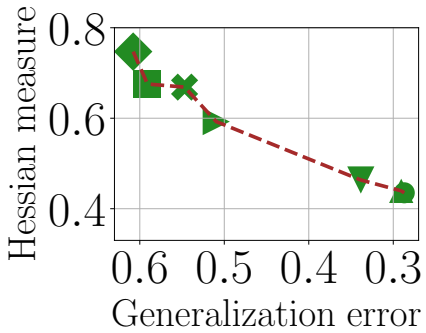
$$O\left(\frac{\sum_{i=1}^{L}\sqrt{\mathrm{Tr}\left[H_i^+\right]\cdot\left|\left|v_i\right|\right|^2}}{\sqrt{n}}\right)$$

  - Fine-tuning with consistent losses.
    - Labels are independently flipped: Reweight loss with the inverse of label confusion matrix F [NDR+'13; PRK+'17].

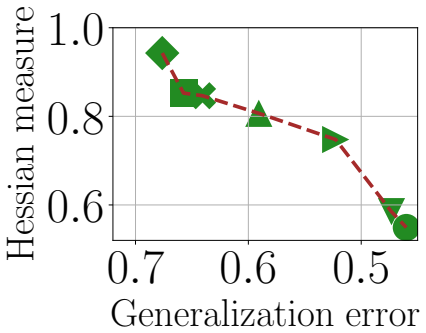
$$O\left(\frac{\sum_{i=1}^{L}\sqrt{||(\mathrm{F}^{-1})^T||_{1,\infty}\mathrm{Tr}[H_i]\cdot\left|\left|v_i\right|\right|^2}}{\sqrt{n}}\right)$$

Measuring Generalization with Hessians

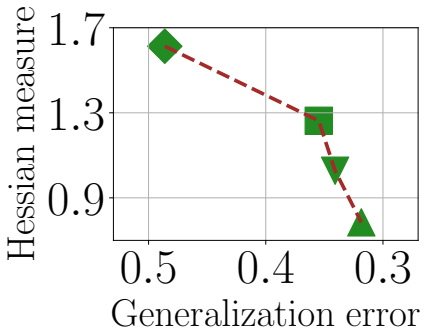
- Applied to various datasets and regularization methods, the Hessian-based measures correlate with empirical generalization errors accurately.



CIFAR-10
ResNet-50



CIFAR-100
ResNet-50



MRPC
BERT-Base

Comparison to Previous Results

- Numerically, our generalization bound is **orders of magnitude smaller** than previous results.
  - Our results are also applicable to BERT models.

Generalization bound	CIFAR-10	CIFAR-100	MRPC	SST2
Pitas et al. [PDV17]	5.51E+10	3.13E+12	/	/
Arora et al. [AGN+18]	1.62E+06	9.66E+07	/	/
Long and Sedghi [LS20]	6.30E+13	8.32E+13	/	/
Gouk et al. [GHP21]	2.04E+69	2.72E+69	/	/
Li and Zhang [LZ21]	1.63E+27	1.31E+29	/	/
Neyshabur et al. [NBS18]	3.13E+11	1.62E+13	/	/
Our result	2.26	7.23	3.83	9.71

Algorithm

- Based on the analysis, we design an algorithm that incorporates **consistent losses** and **distance-based regularization** for fine-tuning.
  - Consistent loss reweighting:

$$\bar{L}(f_W(x)) = F^{-1} \cdot [l(f_W(x), 1), \cdots, l(f_W(x), k)]^T$$

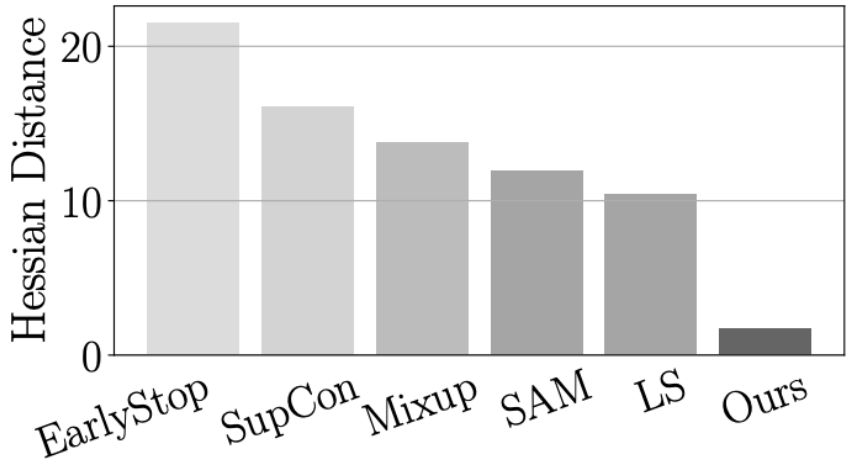
  - Distance-based regularization:

$$\left|\left|W_i - W_i^{(s)}\right|\right|_F \leq \alpha_i, \text{ for } i = 1, \cdots, L$$

  - Our result provides **a provable generalization guarantee** for this algorithm under class conditional label noise.

Empirical Results

- Our algorithm is competitive with or even outperforms baseline methods under various noisy environments and architectures.
  - Average **6.43%** and **3.26%** accuracy improvement in fine-tuning ResNet on **synthetic random noise** and **correlated label noise**.
  - Our results extend to **RoBERTa** and **Vision Transformer** models.
  - Reduces the Hessian distance measure **six times more** than previous fine-tuning algorithms.



Conclusion

Takeaways:

- We develop Hessian-based generalization bounds for fine-tuned models.
- Hessian-based measures accurately correlate with generalization errors.
- An algorithm for fine-tuning with noisy labels.

Further info:

- Paper: https://arxiv.org/pdf/2206.02659.pdf
- Code: https://github.com/NEU-StatsML-Research/Robust-Fine-Tuning
- Email: {h.ju, li.dongyu, ho.zhang}@northeastern.edu

Thanks for watching!