Promoting Ensemble Diversity with Interactive Bayesian Distributional Robustness for Fine-tuning Foundation Models

Ngoc-Quan Pham*, Tuan Truong*, Quyen Tran, Tan Nguyen, Dinh Phung, Trung Le

Qualcomm

NUS National University of Singapore

MONASH University

How do we utilize large pre-trained models?

- Foundational models are increasingly demonstrating remarkable capabilities over a wide array of tasks.
- **Goal**: Effectively utilizing these pre-trained models for **downstream tasks**. However, adapting these models via **full fine-tuning** presents significant limitations: **high computational cost**, **overfitting**, **storage overhead**…

Parameter-efficient Fine-tuning (PEFT)

- Techniques to adapt large pre-trained models with minimal parameter updates.
- **Pros:** Reduce computational cost and memory usage while maintaining performance, and preserving pre-trained knowledge.
- **Cons:** PEFT methods can lead to overconfident predictions, especially when fine-tuned on small datasets.

Motivations

Our method relies on:

- **Bayesian Inference**: enhances robustness, tackling uncertainty
- **Flat minimizers**: improve neural network generalization by helping models find broader local minima, making them more robust
- **Distributional Robustness**: a framework for learning under distributional uncertainty, which seeks the worst-case among a ball of "local distributions".

Motivations

- We also want to promote ensemble diversity
- Define the **approximate posterior distribution** Q^K, where samples are concatenated models $\theta_{1:K}$
- We learn Q so that the sampled models reside in *low-loss, low-sharpness* regions while *maintaining ensemble diversity.*
- Leverage DRO to alleviate training instabilities

Interactive Bayesian Distributional Robustness

- We propose a *distributional population loss* where l_{div} encourages the diversity among model particles

$$\mathcal{L}_{\mathcal{D}}\left(Q^K\right) = \mathbb{E}_{\boldsymbol{\theta} \sim Q^K}\left[\frac{1}{K}\sum_{i=1}^{K} \mathcal{L}_{\mathcal{D}}(\theta_i) + \alpha \mathbb{E}_{\mathcal{D}}\left[l_{div}(\theta_{1:K}; x, y)\right]\right].$$

Theorem 4.1. *With the probability at least* $1-\delta$ *over the choice of* $\mathcal{S} \sim \mathcal{D}^N$*, we have*

$$\mathcal{L}_{\mathcal{D}}\left(Q^K\right) \leq \min_{\lambda \geq 0}\left\{\lambda\rho + \mathbb{E}_{\boldsymbol{\theta} \sim Q^K}\left[\max_{\boldsymbol{\theta}'}\left\{\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}') - \lambda c^K(\boldsymbol{\theta}, \boldsymbol{\theta}')\right\}\right]\right\} + L\sqrt{\frac{KD_{KL}\left(Q, P\right) + \log\frac{1}{\delta}}{2N}}$$

Remark: This framework operates on the joint distribution Q^K and incorporates the divergence loss l_{div}, enabling us to model interactions between the particle models $\theta_{1:K}$

Divergence Loss

- Let f^i_{-y} be the non-maximal prediction probabilities by eliminating the prediction probability of the ground-truth label y

- We encourage the non-maximal predictions to diverge, while maximizing prediction probability of the ground-truth label.

- Motivated by the theory of Determinantal Point Processes, we define the ensemble diversity:

$$l_{div}\left(\theta_{1:K}; x, y\right) = \mathrm{Vol}^2\left(\left[\tilde{f}^i_{-y}\right]_{i\in[C]}\right)$$

where $\tilde{f}^i_{-y} = \frac{f^i_{-y}}{\|f^i_{-y}\|}, \left[\tilde{f}^i_{-y}\right]_{i\in[C]} \in \mathbb{R}^{(C-1)\times K}, \left[C\right] = \left\{1, ..., C\right\}.$

Practical Method

- Define $Q = \frac{1}{K}\sum_{i=1}^{K}\mathcal{N}\Big(\mu_i, \sigma^2\mathbb{I}\Big)$ and $P = \mathcal{N}\Big(\mathbf{0}, \mathbb{I}\Big)$. With a few relaxations, the problem becomes

$$\min_{\mu_{1:K},\sigma}\min_{\lambda\geq 0}\Big\{\lambda\rho + \mathbb{E}_{\theta_{1:K}\sim Q}\Big[\frac{1}{K}\sum_{i=1}^{K}\max_{\theta_i'}\tilde{\ell}(\theta_i',\theta_i;x,y)\Big]\Big\} + \frac{\beta}{K}\Big[\sum_{i=1}^{K}\|\mu_i\|^2 + d(\sigma - \log\sigma)\Big]$$

where $\tilde{\ell}(\theta_i',\theta_i;x,y) = l(\theta_i';x,y) + \alpha l_{div}(\theta_i',\theta_{-i};x,y) - \lambda c(\theta_i,\theta_i')$

- We alternatively update $\mu_{1:K}$ and λ with gradient descent, and update θ_i' with a gradient ascent

Interactive Bayesian Distributional Robustness

Algorithm 1 **I**nteractive **B**ayesian **D**istributional **R**obustness (IBDR)

Input: Initial particle means $\mu_{1:K}$; ascend step size α_1; learning rates $\alpha_\lambda, \alpha_\mu$
Output: Optimal particle means $\mu_{1:K}$
while not converged **do**
  Sample batch $\mathcal{B} = \{(x_1, y_1), \ldots, (x_b, y_b)\}$
  Sample $\epsilon_i \sim \mathbb{N}(0, \mathbb{I})$ and $\theta_i \leftarrow \mu_i + \sigma\epsilon_i$
  Compute $\theta_i' \leftarrow \theta_i + \alpha_1 \nabla_{\theta_i} \tilde{\ell}(\theta_i', \theta_i; x, y)$
  Compute $\lambda \leftarrow \lambda - \alpha_\lambda \nabla_\lambda \overline{\mathcal{L}}(\lambda, \theta_i', \theta_i; x, y)$
  Compute $\mu_i \leftarrow \lambda - \alpha_\mu \nabla_{\mu_i} \overline{\mathcal{L}}(\lambda, \theta_i', \theta_i; x, y)$
end while
return $\mu_{1:K}$

Experiments

Image Classification

Table 1. Top-1 Accuracy on VTAB-1K. The accuracies are reported with ViT-B/16 pre-trained on ImageNet-21K

	Natural							Specialized				Structured								
Method	CIFAR100	Caltech101	DTD	Flowers102	Pets	SVHN	Sun397	Camelyon	EuroSAT	Resisc45	Retinopathy	Clevr-Count	Clevr-Dist	DMLab	KITTI	dSpr-Loc	dSpr-Ori	sNORB-Azim	sNORB-Ele	AVG
FFT	68.9	87.7	64.3	97.2	86.9	87.4	38.8	79.7	**95.7**	84.2	73.9	56.3	58.6	41.7	65.5	57.5	46.7	25.7	29.1	62.3
LoRA	67.1	90.7	68.9	98.1	90.1	84.5	54.2	84.1	94.9	84.4	73.6	**82.9**	69.2	49.8	78.5	75.7	47.1	**31.0**	**44.0**	68.4
SAM	72.7	90.3	71.4	99.0	90.2	84.4	52.4	82.0	92.6	84.1	74.0	76.7	68.3	47.9	74.3	71.6	43.4	26.9	39.1	70.5
SA-BNN	65.1	91.5	71.0	98.9	89.4	89.3	55.2	**86.2**	94.5	86.4	75.2	61.4	63.2	40.0	71.3	64.5	34.5	27.2	31.2	68.2
SGLD	68.7	91.0	67.0	98.6	89.3	83.0	51.6	81.2	93.7	83.2	76.4	80.0	70.1	48.2	76.2	71.1	39.3	31.2	38.4	68.4
DeepEns	68.6	88.9	67.7	98.9	90.7	85.1	54.5	82.6	94.8	82.7	75.3	46.6	47.1	47.4	68.2	71.1	36.6	30.1	35.6	67.0
BayesTune	68.2	91.7	69.5	99.0	90.7	86.4	51.2	84.9	95.3	84.1	75.1	82.8	68.9	49.7	79.3	74.3	46.6	30.3	42.8	68.5
SVGD	71.3	90.2	71.0	98.7	90.2	84.3	52.7	83.4	93.2	86.7	75.1	75.8	70.7	49.6	79.9	69.1	41.2	30.6	33.1	70.9
IBDR	**73.0**	**92.1**	**71.7**	**99.3**	**91.4**	**91.3**	**56.7**	85.1	95.0	**87.3**	**76.5**	78.1	**75.1**	**53.6**	**80.4**	**77.1**	**49.3**	28.9	40.1	**73.6**
	(.11)	(.31)	(.12)	(0.15)	(0.16)	(.36)	(.18)	(.24)	(.44)	(.14)	(.12)	(.11)	(.24)	(.42)	(.26)	(.29)	(.19)	(.13)	(.37)	

Experiments

Image Classification

Table 2. Expected Calibration Errors (ECE) on `VTAB-1K`. The results are reported with `ViT-B/16` pre-trained on `ImageNet-21K`

	Natural							Specialized				Structured								
Method	CIFAR100	Caltech101	DTD	Flowers102	Pets	SVHN	Sun397	Camelyon	EuroSAT	Resisc45	Retinopathy	Clevr-Count	Clevr-Dist	DMLab	KITTI	dSpr-Loc	dSpr-Ori	sNORB-Azim	sNORB-Ele	AVG
FFT	0.29	0.23	0.20	0.13	0.27	0.19	0.45	0.21	0.13	0.18	0.17	0.41	0.44	0.42	0.22	0.14	0.23	0.24	0.40	0.26
LoRA	0.38	0.19	0.18	0.05	0.09	0.10	0.14	**0.11**	0.09	0.12	0.11	0.12	0.19	0.34	0.18	0.14	0.21	0.18	0.31	0.17
SAM	0.21	0.25	0.20	0.11	0.12	0.15	0.14	0.17	0.16	0.14	**0.09**	0.12	0.17	0.24	0.16	0.21	**0.19**	0.13	0.16	0.16
SA-BNN	0.22	0.08	0.19	0.15	0.12	0.12	0.24	0.13	**0.06**	0.12	0.18	0.14	**0.21**	0.22	0.24	0.25	0.41	0.46	0.34	0.20
SGLD	0.26	0.20	**0.17**	0.05	0.18	0.14	0.23	0.18	0.09	0.12	0.32	0.26	0.29	**0.21**	0.26	0.42	0.39	**0.11**	0.24	0.22
DeepEns	0.24	0.12	0.22	0.04	0.10	0.13	0.23	0.16	0.07	0.15	0.21	0.31	0.32	0.36	0.13	0.32	0.31	0.16	0.29	0.20
BayesTune	0.32	0.93	0.20	0.03	0.85	0.12	0.22	0.13	0.07	0.13	0.22	**0.12**	0.23	0.30	0.24	0.28	0.28	0.31	0.26	0.23
SVGD	0.20	0.13	0.19	0.04	0.16	0.09	0.20	0.15	0.11	0.13	0.12	0.17	0.21	0.30	0.18	0.21	0.25	0.14	0.26	0.18
IBDR	**0.16**	**0.08**	0.19	**0.02**	**0.07**	**0.07**	**0.13**	0.12	0.06	**0.11**	0.11	0.13	0.24	0.30	**0.12**	**0.11**	0.30	0.30	**0.16**	**0.14**
	(.03)	(.02)	(.02)	(.01)	(.01)	(.01)	(.02)	(.03)	(.02)	(.02)	(.01)	(.01)	(.02)	(.03)	(.01)	(.01)	(.05)	(.04)	(.02)	

Experiments

Commonsense Reasoning

Table 3. Accuracy/ECE on six common-sense reasoning datasets

Metric		Datasets						
Type	Method	WG-S	ARC-C	ARC-E	WG-M	OBQA	BoolQ	AVG
ACC (↑)	MLE	68.99	69.10	85.65	74.53	81.52	86.53	77.72
	MAP	68.62	67.59	86.55	75.61	81.38	86.50	77.71
	MCD	69.26	68.43	86.07	76.18	81.49	87.15	78.10
	ENS	69.57	66.20	84.40	75.32	81.38	87.09	77.33
	BBB	67.54	68.11	85.63	73.41	81.72	**87.19**	77.27
	LAP	69.20	66.78	80.05	75.55	82.12	86.95	76.78
	BLoB	70.89	**70.83**	86.68	74.55	82.73	86.80	78.75
	IBDR	**72.51**	70.56	**86.95**	**76.46**	**84.60**	86.89	**79.66**
ECE (↓)	MLE	29.83	29.00	13.12	20.62	12.55	3.18	18.05
	MAP	29.76	29.42	12.07	23.07	13.26	3.16	18.46
	MCD	28.06	27.73	12.31	18.27	15.12	3.49	17.50
	ENS	28.52	29.16	12.57	20.86	15.34	9.61	19.34
	BBB	21.93	25.84	12.42	15.89	11.23	3.76	15.18
	LAP	**4.15**	**16.25**	33.29	**7.40**	8.70	**1.30**	**11.85**
	BLoB	20.62	20.61	**9.43**	11.23	8.36	2.46	12.12
	IBDR	24.17	21.20	9.71	11.19	**5.82**	1.54	12.27

Conclusion

- We introduce a novel Bayesian framework that explicitly models the interaction between particles
- We propose Interactive Bayesian Distributional Robustness, which simultaneously enhances ensemble diversity, generalization ability, and distributional robustness

Thank you