Model Fusion Through Bayesian Optimization in Language Model Fine-Tuning

Chaeyun Jang[1], Jungtaek Kim[2], Hyungi Lee[1], Juho Lee[1]
Korea Advanced Institute of Science & Technology[1], University of Pittsburgh[2]

The 3rd NeurIPS Workshop On Efficient Natural Language and Speech Processing (ENLSP)

Overview

1. Seeking to alleviate the constraints of conventional Weight Averaging (WA) in the fine-tuning of Pretrained Language Models (PLMs).
2. Intending to propose an efficient model fusion methodology that incorporates Hyperparameter Optimization (HP).

Contributions

1. Multi Objective Bayesian Optimization SWA (MOBO-SWA):
   - Employing MOBO optimizes model averaging coefficients by considering both loss and metric.

2. Bilevel Model Fusion (Bilevel-BO-SWA):
   - The process is structured as a bilevel procedure, with an outer BO optimizing hyperparameters for PLMs fine-tuning, and an MOBO-SWA for model fusion.

Backgrounds

Model Fusion for Pre-trained Language Models

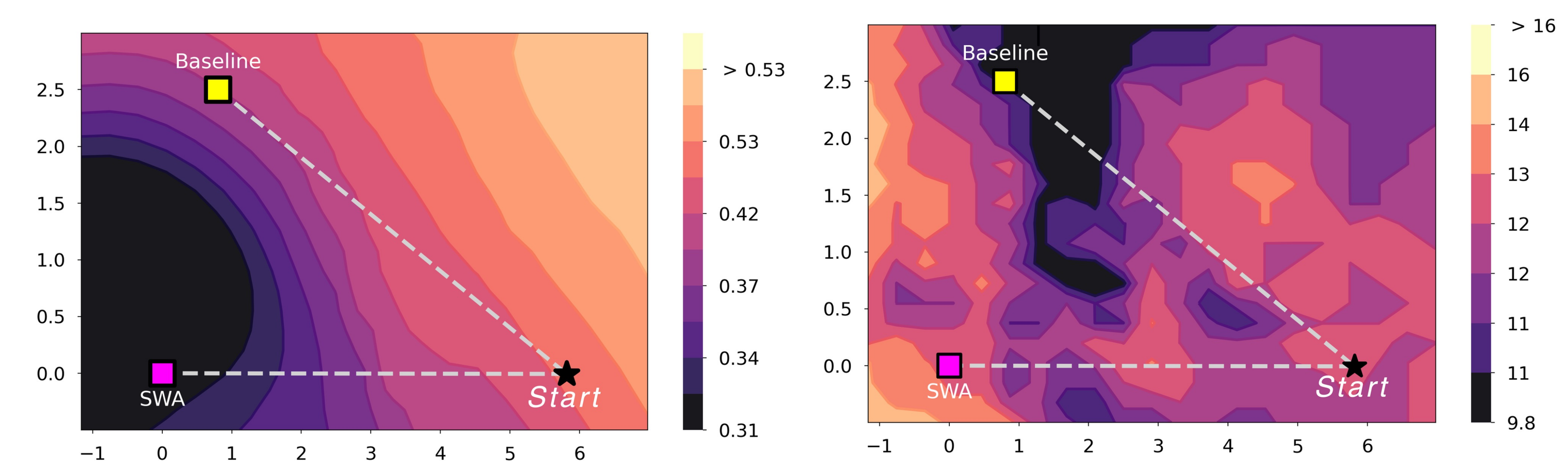
- The cost of fine-tuning PLMs is **high**, making deep ensemble methods **inefficient**.
- WA methods like Stochastic Weight Averaging (SWA) [1] and Model Soups [2] are more feasible for model fusion, although not always adequate for PLMs [3].
- We discovered that the loss surface of PLMs has a best combination that cannot be found by simple averaging.

Bayesian Optimization (BO)

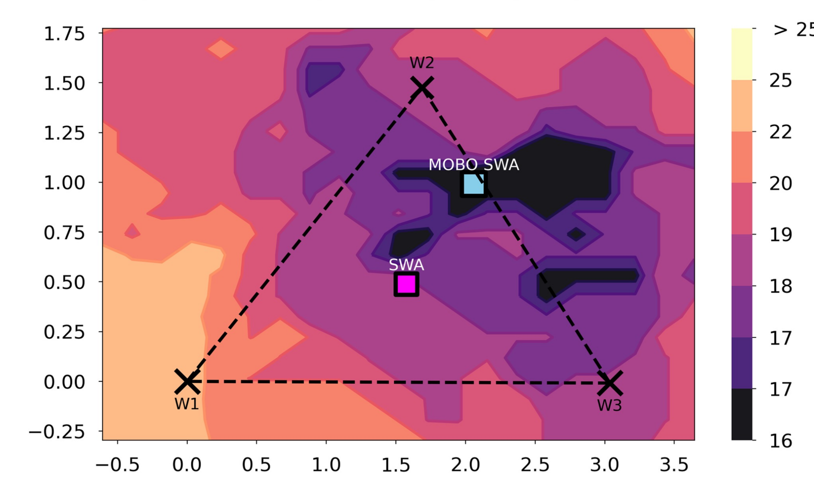
- BO is a strategy for optimizing black-box functions that are **costly to evaluate** using the surrogate model.
- Applications range across various domains, and while Gaussian Processes are commonly used as the surrogate model, other models like Bayesian neural networks and tree-based models are also options [4].

Loss Surface of the Language Model

Valid loss surface vs Valid metric surface



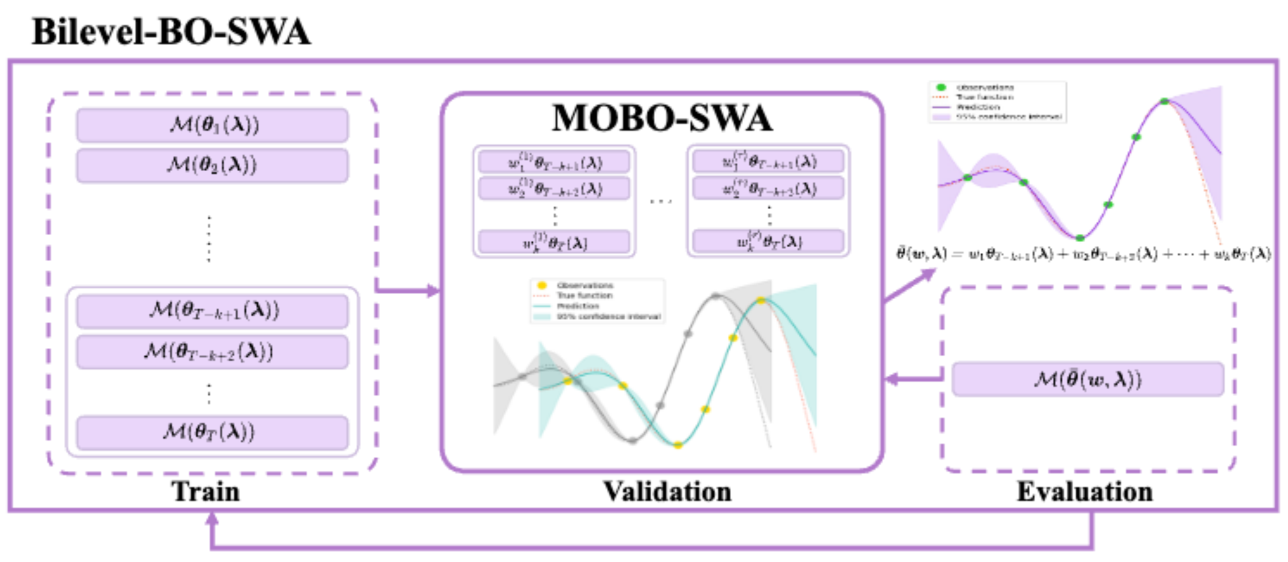
SWA vs MOBO-SWA



- ✓ While SWA tends to lead to a **flat loss minima**, this is not necessarily related to better performance according to **evaluation metrics**.
- ✓ There exists the **best combination** for combining model weights that is not achievable through SWA.

* Following Garipov et al.'s [5] method, we collect SWA models w_1, w_2, w_3 and establish an ***orthonormal basis*** $\boldsymbol{u}_1$, $\boldsymbol{u}_2$ to represent their parameter space on the x-axis and y-axis.

Model Fusion Through Bayesian Optimization



- Our target language model as $\mathcal{M}(\boldsymbol{\theta}(\boldsymbol{\lambda}))$ where $\boldsymbol{\lambda}$ is a hyperparameter vector and $\boldsymbol{\theta}(\boldsymbol{\lambda})$ is model parameter trained with $\boldsymbol{\lambda}$

$$\bar{\boldsymbol{\theta}}(\boldsymbol{w}, \boldsymbol{\lambda}) = w_1 \boldsymbol{\theta}_{T-k+1}(\boldsymbol{\lambda}) + w_2 \boldsymbol{\theta}_{T-k+2}(\boldsymbol{\lambda}) + \cdots + w_k \boldsymbol{\theta}_T(\boldsymbol{\lambda}), \quad (1)$$

  Where $w_1, w_2, \ldots w_k \in [0,1]$ subject to $\sum_{i=1}^{k} w_k = 1$
- Our objective is to aggregate the final k epoch models into a proficient single model $\mathcal{M}(\bar{\boldsymbol{\theta}}(\boldsymbol{\lambda}))$, using a weighted combination shown in equation (1).
- Multi-Objective Bayesian Optimization for Model Fusion

$$\mathcal{P} = \left\{ \boldsymbol{w}^\dagger \mid \boldsymbol{w}^\dagger = \arg\min_{\boldsymbol{w}} \left(f_{\text{loss}}(\mathcal{M}(\bar{\boldsymbol{\theta}}(\boldsymbol{w}, \boldsymbol{\lambda}))), f_{\text{metric}}(\mathcal{M}(\bar{\boldsymbol{\theta}}(\boldsymbol{w}, \boldsymbol{\lambda}))) \right) \right\}. \quad (2)$$

  We optimize model fusion coefficients with MOBO-SWA (2) considering both loss and metric, employing **qNEHVI** [6] for the hypervolume improvement objective optimization.
- Bilevel Bayesian Optimization for Model Fusion

  We further refine hyperparameters through Bilevel-BO-SWA with the objective $f_{\text{metric}}(\mathcal{M}(\bar{\boldsymbol{\theta}}(\tilde{\boldsymbol{w}}^\dagger, \boldsymbol{\lambda})))$

Experiment Results

Main result

Table 1: **Results on the GLUE dataset using the RoBERTa-base.** Numerical results in boldface and with an underscore indicate the best and the second-best results in the respective datasets, respectively.

Method	RTE	MRPC	CoLA	STS-B	SST-2	Avg.
Fine-tune	74.94 ± 2.28	91.65 ± 0.65	56.34 ± 2.90	89.86 ± 0.16	94.49 ± 0.04	81.46
SWA [16]	77.19 ± 1.04	91.31 ± 1.82	55.05 ± 3.09	**89.89** ± 0.20	94.49 ± 0.08	81.59
Greedy SWA [37]	76.52 ± 1.31	91.84 ± 0.14	56.47 ± 3.20	<u>89.87</u> ± 0.18	94.36 ± 0.05	81.81
Learned SWA [37]	77.82 ± 3.58	90.62 ± 1.87	<u>59.02</u> ± 2.60	89.65 ± 0.09	94.19 ± 0.00	82.26
MOBO-SWA (ours)	<u>77.86</u> ± 0.28	<u>92.05</u> ± 1.05	58.20 ± 1.72	89.58 ± 0.12	<u>94.55</u> ± 0.12	<u>82.45</u>
Bilevel-BO-SWA (ours)	**78.43** ± 0.26	**92.38** ± 0.68	**59.21** ± 3.53	89.86 ± 0.01	**94.97** ± 0.08	**82.97**
Best subset (oracle)	80.66 ± 0.52	92.90 ± 0.22	60.01 ± 1.88	89.93 ± 0.20	95.08 ± 0.06	83.72

Loss vs Metric discrepancy

Table 2: **Results on GLUE benchmark for RoBERTa-base.** Evaluation results of SWA and naive fine-tuned model on the RTE, MRPC, SST-2. We used custom validation sets for the evaluation. Here *NLL* is the loss function and *Error rate* is the 1 - *Accuracy* for the RTE and SST-2, and the *F1 score* for the MRPC. The lower value is the better for all the evaluation functions. Please refer to Appendix B to see how we split the custom validation sets.

		Task		
		RTE	MRPC	SST-2
NLL (↓)	Fine-tune	0.97 ± 0.01	0.54 ± 0.02	0.28 ± 0.00
	SWA	**0.87** ± 0.03	**0.53** ± 0.00	**0.22** ± 0.00
Error rate (↓)	Fine-tune	**21.21** ± 0.69	**7.82** ± 0.01	**4.94** ± 0.26
	SWA	21.71 ± 1.47	7.90 ± 0.01	5.16 ± 0.24

Ablation on MOBO and BO

Table 3: **Using RoBERTa-base, Performance Analysis of Basic BO on the GLUE Dataset.** When employing BO that focuses solely on a single objective, specifically the metric, it was observed that MOBO-SWA exhibited commendable effectiveness in comparison to BO-SWA, which takes into account both the loss and metric.

Method	RTE	MRPC	CoLA	STS-B	SST-2	Avg.
BO-SWA	77.20 ± 1.97	91.92 ± 0.66	57.56 ± 0.30	**89.63** ± 0.05	94.47 ± 0.15	82.16
MOBO-SWA	**77.86** ± 0.28	**92.05** ± 1.05	**58.20** ± 1.72	89.58 ± 0.12	**94.55** ± 0.12	**82.45**

Ablation on the Effectiveness of the Outer BO

Table 4: **Comparative Performance Analysis Applying Outer BO on Various Baselines.** The table shows that Bilevel-BO-SWA outperforms other strategies on RTE and MRPC datasets, according to key performance metrics.

	Baseline	SWA	Greedy SWA	Learned SWA	Bilevel-BO-SWA
RTE	77.20	76.68	76.68	78.22	79.60
MRPC	91.41	90.57	90.57	90.03	93.39

References

[1] Izmailov, Pavel, et al. "Averaging Weights Leads to Wider Optima and Better Generalization."
[2] Wortsman, Mitchell, et al. "Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time." International Conference on Machine Learning. PMLR, 2022.
[3] Kaddour, Jean, et al. "When do flat minima optimizers work?." Advances in Neural Information Processing Systems 35 (2022): 16577-16595.
[4] Frazier, Peter I. "A tutorial on Bayesian optimization." arXiv preprint arXiv:1807.02811 (2018).
[5] Garipov, Timur, et al. "Loss surfaces, mode connectivity, and fast ensembling of dnns." Advances in neural information processing systems 31 (2018).
[6] Daulton, Samuel, Maximilian Balandat, and Eytan Bakshy. "Parallel bayesian optimization of multiple noisy objectives with expected hypervolume improvement." Advances in Neural Information Processing Systems 34 (2021): 2187-2200.