Workshop at the 6th Symposium on Advances in Approximate Bayesian Inference (non-archival), 2024 1–15

Gaussian Stochastic Weight Averaging for Bayesian Low-rank Adaptation of Large Language Models

Anonymous Authors
Anonymous Institution

Abstract

Fine-tuned Large Language Models (LLMs) often suffer from overconfidence and poor calibration, particularly when fine-tuned on small datasets. To address these challenges, we propose a simple merging of Low-Rank Adaptation (LoRA) with Gaussian Stochastic Weight Averaging (SWAG), facilitating approximate Bayesian inference in LLMs. Through extensive testing across several Natural Language Processing (NLP) benchmarks, we demonstrate that our straightforward and computationally efficient approach improves model generalization and calibration. We further show that our method exhibits greater robustness against distribution shift, as reflected in its improved performance on out-of-distribution tasks.

1. Introduction

In recent years, LLMs have demonstrated exceptionally strong performance across a wide range of natural language processing tasks (Radford et al., 2019; Touvron et al., 2023a; Brown et al., 2020). Due to extremely large numbers of parameters in state-of-the-art foundation models like RoBERTa (Liu et al., 2019) or GPT (OpenAI, 2023), approximate Bayesian inference has been difficult. Moreover, fine-tuning these models on the full number of weights is inefficient and prohibitively expensive for practitioners without an abundance of computational resources. In response to this problem, recent work has explored adapters for parameter-efficient fine-tuning (PEFT; Ding et al., 2023) of LLMs on downstream tasks. Our research focuses on utilizing these newly developed PEFT techniques for Bayesian subspace inference in these tuning parameters. One of the currently most popular PEFT methods is low-rank adaptation (LoRA; Hu et al., 2022), which fine-tunes a model indirectly by freezing the pre-trained weights and introducing a set of low-rank matrices which are injected into several layers throughout the model. While LoRA can help resolve the issue of inefficiency in fine-tuning, the resulting LLMs still suffer from a significant limitation: they have been shown to be overly confident in their predictions and exhibit poor calibration (Jiang et al., 2021; Xiao et al., 2022; He et al., 2023). The predictive probabilities produced by neural networks in classification settings are often incorrectly interpreted as the confidence of the model. However, models may still be uncertain in their predictions, despite yielding a high softmax output (Gal and Ghahramani, 2016). This phenomenon can be dangerous, especially in safety-critical applications such as medical diagnoses (Singhal et al., 2023), and has prompted the call for better Bayesian treatments of LLMs (Papamarkou et al., 2024).

Many previous works have already proposed various approximate Bayesian inference methods that can be applied in deep neural networks to address this issue (see Appendix

A for related works). In this paper, we propose a simple alternative to these methods which avoids any significant numerical and implementation challenges while also consistently improving model generalization and calibration despite its simplicity. We integrate Gaussian Stochastic Weight Averaging (SWAG; Maddox et al., 2019) with Low-Rank Adaptation (LoRA) to inexpensively enable approximate Bayesian inference with LLMs.

We evaluate our method's performance against those of standard methods such as standard (non-Bayesian) LoRA fine-tuning as well as Monte Carlo (MC) dropout (Gal and Ghahramani, 2016) and LoRA ensembles (Wang et al., 2023) on a number of benchmark commonsense reasoning NLU tasks. We find competitive performance both in terms of accuracy and calibration.

We provide a description of our proposed method SWAG-LoRA in Section 2 and follow this up with a discussion of our empirical results in Section 3. A detailed treatment of the related work is deferred to Appendix A.

2. Methods

2.1. Low-Rank Adaptation

Low-Rank Adaptation (LoRA) is PEFT method which significantly reduces the number of parameters needed for fine-tuning without significantly compromising performance. LoRA adds low-rank adaptation matrices to certain layers (most commonly the query and value projection matrices in attention layers and the classification head (Hu et al., 2022)). LoRA enables efficient and effective fine-tuning with minimal computational resources. Equivalently, it enables the fine-tuning of very large models using limited computational resources. In our context, we leverage LoRA to greatly reduce the number of trainable parameters so that we can feasibly use approximate Bayesian inference methods with LLMs with billions of parameters. LoRA enables us to evaluate our method with LLama-2 with 7 billion parameters.

2.2. SWA & SWAG

Stochastic weight averaging (SWA) consists of a simple averaging of model weights from the trajectory of Stochastic Gradient Descent (SGD). SGD iterates to average are obtained using a cyclical or large constant learning rate, which explores the region around the MAP estimate in the loss landscape. SWA has been shown to find flatter solutions than SGD, improving robustness and generalization performance (Izmailov et al., 2018). SWA-Gaussian is a simple extension of SWA, which treats SGD iterates around the MAP as samples from a Gaussian distribution. SWAG estimates this Gaussian distribution using the results of SWA as its first moment $\mathbf{w}_{\text{SWA}}$ and a low-rank plus diagonal covariance Σ, also derived from the SGD iterates, as its second moment, thus creating an approximate posterior distribution over the weights of the network:

$$p(\mathbf{w}|D) \approx \mathcal{N}\left(\mathbf{w}_{\text{SWA}}\ , \frac{1}{2} \cdot (\Sigma_{\text{diag}}\ + \Sigma_{\text{low-rank}}\ )\right).$$

The Bayesian Model Average (BMA) can be approximated by sampling from this distribution. This approach is particularly attractive for obtaining uncertainties for deep learning

models, as it often improves generalization, calibration, and uncertainty quantification with virtually no computational overhead (Maddox et al., 2019). MultiSWAG is a natural extension of SWAG that leverages an ensemble of SWAG models, enabling an effective mechanism for Bayesian marginalization within multiple basins in the loss landscape to further improve model generalization and calibration (Wilson and Izmailov, 2020).

Our implementations of SWAG and MultiSWAG estimate the posterior distribution only over the parameters in the LoRA modules in our network, thereby avoiding the significantly more computationally expensive task of estimating a posterior distribution over all network parameters.

3. Experiments

3.1. Experimental Setup

Implementation Details All of the experiments reported in this paper are conducted using Llama-2-7B (Touvron et al., 2023b) (finetuned in 16-bit and evaluated in 32-bit). We leverage the PEFT library (Mangrulkar et al., 2022) implementation of LoRA to adapt the queries, values, and the causal language modeling head of Llama-7B, using the LoRA rank $r = 8$, alpha $\alpha = 16$ and a LoRA dropout probability of 0.1.

We adapted the official implementation of SWAG (Zellers et al., 2018) to work seamlessly with PEFT's LoRA adapters. We experimented with many schedulers for SWAG and ultimately used a constant learning rate schedule for which we determined the constant learning rate by dividing the fine-tuning optimizer's maximum learning rate by 2. We also implement MultiSWAG, whose performance we evaluate against ensembles of standard LoRA-finetuned Llama-7B models for a fairer comparison. All ensemble or MultiSWAG models consist of five individual members in the ensemble. When evaluating (Multi)SWAG and MC Dropout, we always sample 15 models from our approximate posterior distributions. Furthermore, for SWAG sampling we use a sample scale of 1.0.

Datasets We evaluate our models on several well-known multiple-choice question answering (MCQA) benchmarks. We evaluate our methods on Open Book Question Answering (OBQA; Mihaylov et al., 2018), Commonsense QA (CQA; Talmor et al., 2019), AI2 Reasoning Challenge Easy and Challenge (ARC-E & ARC-C; Clark et al., 2018), and Measuring Massive Multitask Language Understanding (MMLU; Hendrycks et al., 2021).

The CQA dataset does not have public test labels. We therefore randomly split the training data into new train and validation sets and use the original validation set to report test set performance. For the remainder of the paper, when speaking of results on the CQA test set, this is the split we will be referring to.

Calibration and uncertainty quantification. Another challenge that comes along with integrating uncertainty estimation into neural network models is the measurement of the quality of these uncertainties. One of the most common metrics in this regard is the expected calibration error (ECE), which is used to measure the discrepancy between a model's predicted probabilities and the true frequencies of the outcomes. However, Gruber and Buettner (2022) show that most commonly used calibration errors are actually lower bounds on the true error and provide a taxonomy of a number of these metrics used in practice. They find that the ECE lower bound (which suffers from shortcomings such as

sensitivity to the size of datasets) is the least tight of all the considered metrics, while the Brier score is the tightest. For this reason, in addition to the more common evaluation metrics negative log-likelihood (NLL) and ECE, we also include the Brier score, which gives a better estimate of the true calibration error.

We also evaluate our methods' calibration across both in-distribution (ID) and out-of-distribution (OOD) tasks. In addition to OOD evaluation, we test our methods' uncertainty quantification performance by evaluating whether our uncertainty metrics can reliably detect OOD samples. We investigated several different uncertainty metrics across our experiments and ultimately found that those most effective at detecting OOD samples were the standard entropy calculated over class probabilities (here referred to as *entropy* or *cross-prediction entropy*), as well as (only for our ensemble/SWAG/MultiSWAG/MC dropout methods) an entropy-adjacent metric we refer to as *cross-model entropy* which calculates the average per-class uncertainties between the models of the ensemble.

While entropy is the more common method of uncertainty quantification, it has been argued that it is not maximally informative, as it does not actually capture the disentanglement of epistemic and aleatoric uncertainty (Malinin et al., 2020). Entropy uses the average prediction in the BMA, meaning it cannot distinguish between a scenario in which all hypotheses confidently disagree (epistemic uncertainty) and a scenario in which all hypotheses agree on being highly uncertain (aleatoric uncertainty). Cross-model disagreement on the other hand considers the variance in hypotheses ouput by each ensemble member, thus measuring epistemic uncertainty alone.

In 3.2.2 we present OOD evalutaion and detection using the aforementioned metrics in combination with the area under the ROC curve (AUROC; Liang et al., 2018).

3.2. Results

3.2.1. Baseline results

In Table 1, we present the test set accuracy (ACC), negative log-likelihood (NLL), expected calibration error (ECE), and Brier score for Llama-7B fine-tuned on each of our benchmark datasets. We compare (Multi)SWAG against the MAP, MC Dropout, Ensemble and SWA.

We find several interesting results from our experiments. MC Dropout and Ensembles do not appear to have a significant effect on either the performance or the calibration of the MAP model. This is surprising as other works have focused demonstrated such improvements from ensembling deep networks (Wang et al., 2023; D'Angelo and Fortuin, 2021).

We further notice that SWA significantly improves the generalization performance over the MAP, albeit at the cost of calibration. SWAG offers a middle ground between SWA and the MAP, by improving both generalization and calibration over the MAP model. Moreover, MultiSWAG yields the same improvements at a slightly larger scale (albeit an additional computational cost). It's important to note that the benefit from MultiSWAG appears to arise from both the ensembling as well as the exploration of basins of attraction in the loss landscape. MultiSWAG consistently outperforms all other methods in both accuracy and calibration.

3.2.2. OOD results

Our results for OOD evaluation and detection are shown in Table 2. We find that while SWAG and MultiSWAG have lower AUROC scores than non-SWAG methods, the OOD evaluation accuracy is systematically higher for the SWAG methods. This is consistent with the findings of (Wang et al., 2023), who posit that *"we suspect the models would fail to detect OOD samples if they can generalize to them, as the accuracy increases throughout fine-tuning"*.

Interestingly, when comparing AUROC results on the two selected metrics, cross-prediction entropy and cross-model entropy, we find that while the choice of entropy does not affect the AUROC scores for non-SWAG methods, cross-model entropy consistently yields better results for SWAG by up to 4 percentage points, allowing us to almost match AUROC performance of the non-SWAG counterparts.

Table 1: Baseline evaluation of accuracy and calibration of SWAG-LoRA and other uncertainty-aware LoRA methods with Llama-2-7B across benchmark datasets. MultiSWAG consistently outperforms the other methods, improving test set accuracy and calibration across benchmark datasets. SWAG demonstrates similar improvements to a lesser degree. Results presented are averaged over three separate training runs, with standard errors shown as subscripts [2]. Each best result is bolded and the second best is underlined.

Metrics	Methods	OBQA	CQA	ARC-E	ARC-C
ACC ↑	MAP	$77.9_{0.2}$	$76.2_{0.3}$	$78.3_{0.5}$	$57.8_{0.5}$
	MC Dropout	$77.7_{0.1}$	$75.8_{0.3}$	$77.8_{0.7}$	$55.4_{2.0}$
	Ensemble	77.6	77.1	77.7	57.6
	SWA	$81.2_{0.4}$	$\underline{77.0}_{0.4}$	$\underline{82.7}_{0.2}$	$\underline{63.5}_{0.5}$
	SWAG	$\underline{81.9}_{0.3}$	$\underline{77.0}_{0.5}$	$80.9_{1.1}$	$62.3_{0.4}$
	MultiSWAG	**83.2**	**78.9**	**83.8**	**66.4**
NLL ↓	MAP	$0.68_{0.00}$	$0.69_{0.01}$	$0.68_{0.00}$	$1.14_{0.01}$
	MC Dropout	$0.68_{0.01}$	$0.69_{0.01}$	0.68	$1.20_{0.07}$
	Ensemble	0.68	$\underline{0.65}$	$0.67_{0.02}$	$1.07_{0.03}$
	SWA	$1.00_{0.03}$	$0.89_{0.01}$	$1.06_{0.02}$	$1.87_{0.03}$
	SWAG	$\underline{0.60}_{0.02}$	$0.79_{0.05}$	$\underline{0.62}_{0.02}$	$\underline{1.12}_{0.04}$
	MultiSWAG	**0.50**	**0.58**	**0.53**	**0.91**
ECE ↓	MAP	$9.1_{0.2}$	$6.4_{0.4}$	$8.3_{0.3}$	$30.3_{19.4}$
	MC Dropout	$8.7_{0.2}$	$6.6_{0.1}$	$7.5_{0.9}$	$7.8_{1.4}$
	Ensemble	9.0	$\underline{4.6}$	$\underline{6.5}$	**4.0**
	SWA	$14.9_{0.5}$	$14.2_{0.4}$	$014.2_{0.3}$	$27.9_{0.6}$
	SWAG	$\mathbf{4.8}_{0.4}$	$11.0_{3.2}$	$\underline{6.5}_{0.7}$	$10.9_{1.5}$
	MultiSWAG	$\underline{5.5}$	**4.3**	**6.3**	$\underline{4.3}$
Brier ↓	MAP	$0.33_{0.00}$	$0.35_{0.00}$	$0.32_{0.00}$	$0.58_{0.00}$
	MC Dropout	$0.33_{0.00}$	$0.35_{0.01}$	$0.33_{0.01}$	$0.62_{0.00}$
	Ensemble	0.33	$\underline{0.33}$	0.32	0.56
	SWA	$0.34_{0.01}$	$0.37_{0.01}$	$0.31_{0.01}$	$0.63_{0.01}$
	SWAG	$\underline{0.28}_{0.01}$	$0.36_{0.01}$	$\underline{0.29}_{0.01}$	$\underline{0.53}_{0.01}$
	MultiSWAG	**0.25**	**0.30**	**0.25**	**0.47**

Table 2: Comparison of OOD evaluation and detection performance of SWAG-LoRA with other uncertainty-aware LoRA methods. SWA and SWAG systematically outperform the other methods in OOD accuracy, while suffering from worse calibration. All models were trained on OBQA as the ID dataset. Results presented are averaged over three separate training runs, with standard errors shown as subscripts.

Metrics	Methods	ARC-E	ARC-C	MMLU law	MMLU cs
ACC ↑	MAP	$60.2_{0.2}$	$50.7_{0.3}$	$34.4_{0.5}$	$42.8_{0.7}$
	MC Dropout	$60.1_{0.2}$	$50.8_{0.3}$	$34.4_{0.5}$	$42.6_{0.7}$
	SWA	$\mathbf{66.5_{0.11}}$	$\mathbf{56.0_{0.13}}$	$\mathbf{40.4_{0.10}}$	$\mathbf{46.2_{2.1}}$
	SWAG	$\underline{64.6}_{0.11}$	$\underline{53.5}_{0.1}$	$\underline{40.1}_{0.23}$	$\underline{44.1}_{0.9}$
NLL ↓	MAP	$\mathbf{0.96_{0.00}}$	$\mathbf{1.15_{0.01}}$	$\mathbf{1.40_{0.01}}$	$\mathbf{1.29_{0.00}}$
	MC Dropout	$\mathbf{0.96_{0.00}}$	$\mathbf{1.15_{0.01}}$	$\mathbf{1.40_{0.01}}$	$\mathbf{1.29_{0.00}}$
	SWA	$1.43_{0.04}$	$1.84_{0.05}$	$2.39_{0.24}$	$2.56_{0.30}$
	SWAG	$\underline{1.02}_{0.03}$	$\underline{1.28}_{0.02}$	$\underline{1.64}_{0.10}$	$\underline{1.60}_{0.07}$
ECE ↓	MAP	$\mathbf{4.9_{0.3}}$	$\underline{10.9}_{0.5}$	$\underline{15.2}_{0.9}$	$\mathbf{11.4_{0.2}}$
	MC Dropout	$\underline{5.0}_{0.2}$	$\mathbf{10.8_{0.5}}$	$\mathbf{15.1_{1.0}}$	$\underline{11.6}_{0.5}$
	SWA	$19.6_{1.1}$	$27.8_{1.5}$	$33.1_{4.1}$	$35.3_{3.7}$
	SWAG	$7.4_{1.3}$	$15.5_{0.4}$	$20.1_{3.6}$	$20.7_{0.4}$
Brier ↓	MAP	$\underline{0.51}_{0.00}$	$\mathbf{0.62_{0.00}}$	$\mathbf{0.75_{0.01}}$	$\mathbf{0.69_{0.00}}$
	MC Dropout	$\underline{0.51}_{0.00}$	$\mathbf{0.62_{0.00}}$	$\mathbf{0.75_{0.01}}$	$\mathbf{0.69_{0.00}}$
	SWA	$0.53_{0.02}$	$0.69_{0.02}$	$0.89_{0.04}$	$0.88_{0.03}$
	SWAG	$\mathbf{0.48_{0.01}}$	$\underline{0.62}_{0.01}$	$\underline{0.79}_{0.03}$	$\underline{0.75}_{0.01}$
AUROC(H_{Pred}) ↑	MAP	$\mathbf{0.78_{0.00}}$	$\mathbf{0.81_{0.00}}$	$\mathbf{0.92_{0.00}}$	$\mathbf{0.89_{0.00}}$
	MC Dropout	$\mathbf{0.78_{0.00}}$	$\mathbf{0.81_{0.00}}$	$\mathbf{0.92_{0.00}}$	$\mathbf{0.89_{0.00}}$
	SWA	$\underline{0.73}_{0.02}$	$\underline{0.76}_{0.02}$	$0.84_{0.05}$	$0.80_{0.08}$
	SWAG	$0.72_{0.02}$	$0.75_{0.01}$	$\underline{0.84}_{0.02}$	$\underline{0.81}_{0.01}$
AUROC(H_{Model}) ↑	MC Dropout	$\mathbf{0.78_{0.00}}$	$\mathbf{0.81_{0.00}}$	$\mathbf{0.92_{0.00}}$	$\mathbf{0.87_{0.00}}$
	SWAG	$0.75_{0.00}$	$0.78_{0.01}$	$0.88_{0.02}$	$0.85_{0.00}$

4. Conclusion

In this work, we demonstrated that the popular and simple SWAG method can be applied with low-rank adaptation to achieve a competitive performance with more sophisticated and computationally expensive methods that have recently been proposed for Bayesian LoRA. We show our approach's ability to both improve model generalization and calibration on a variety of widely used natural language understanding benchmarks. We show that our approach is also robust to domain shift and can generalize to OOD data better than a conventionally fine-tuned LLM. We hope that this exploration can serve to further the endeavor to enable uncertainty awareness in foundation models without introducing prohibitive computational overhead.

2. Due to computational and time constraints, we were unable to finalize the repeat runs for Ensemble and MultiSWAG, and as such only the single run metrics are reported.

References

Jakob Smedegaard Andersen and Walid Maalej. Efficient, uncertainty-based moderation of neural networks text classifiers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1536–1546. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.FINDINGS-ACL.121. URL https://doi.org/10.18653/v1/2022.findings-acl.121.

Oleksandr Balabanov and Hampus Linander. Uncertainty quantification in fine-tuned llms using lora ensembles, 2024.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1–2):151–175, may 2010. ISSN 0885-6125. doi: 10.1007/s10994-009-5152-4. URL https://doi.org/10.1007/s10994-009-5152-4.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL http://dx.doi.org/10.1023/A%3A1010933404324.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Arnav Chavan, Zhuang Liu, Deepak Gupta, Eric Xing, and Zhiqiang Shen. One-for-all: Generalized lora for parameter-efficient fine-tuning. *arXiv preprint arXiv:2306.07967*, 2023.

Wenlong Chen and Yingzhen Li. Calibrating transformers via sparse gaussian processes. *arXiv preprint arXiv:2303.02444*, 2023.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*, 2023.

Tristan Cinquin, Alexander Immer, Max Horn, and Vincent Fortuin. Pathologies in priors and inference for Bayesian transformers. In *Fourth Symposium on Advances in Approximate Bayesian Inference*, 2022.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*, 2018.

Francesco D'Angelo and Vincent Fortuin. Repulsive deep ensembles are Bayesian. *Advances in Neural Information Processing Systems*, 34:3451–3465, 2021.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models. *arXiv preprint arXiv:2203.06904*, 2022.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, 2023.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International conference on machine learning*, pages 1050–1059. PMLR, 2016.

Adam Gleave and Geoffrey Irving. Uncertainty estimation for language reward models. *arXiv preprint arXiv:2203.07472*, 2022.

Sebastian Gruber and Florian Buettner. Better uncertainty calibration via proper scores for classification and beyond. *Advances in Neural Information Processing Systems*, 35: 8618–8632, 2022.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Guande He, Jianfei Chen, and Jun Zhu. Preserving pre-trained features helps calibrate fine-tuned language models. In *International Conference on Learning Representations*, 2023.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*, 2017.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.

John Hewitt, Xiang Lisa Li, Sang Michael Xie, Benjamin Newman, and Percy Liang. Ensembles and cocktails: Robust finetuning for natural language generation. In *Proceedings of the Workshop on Distribution Shifts, 35th Conference on Neural Information Processing Systems*, 2021.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Pavel Izmailov, Wesley J Maddox, Polina Kirichenko, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Subspace inference for bayesian deep learning. In *Uncertainty in Artificial Intelligence*, pages 1169–1179. PMLR, 2020.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Ananya Kumar, Tengyu Ma, Percy Liang, and Aditi Raghunathan. Calibrated ensembles can mitigate accuracy tradeoffs under distribution shift. In James Cussens and Kun Zhang, editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 1041–1051. PMLR, 01–05 Aug 2022. URL https://proceedings.mlr.press/v180/kumar22a.html.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1VGkIxRZ.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. *Advances in neural information processing systems*, 32, 2019.

Andrey Malinin, Bruno Mlodozeniec, and Mark J. F. Gales. Ensemble distribution distillation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=BygSP6Vtvr.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260.

R OpenAI. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2:13, 2023.

Kazuki Osawa, Siddharth Swaroop, Mohammad Emtiyaz E Khan, Anirudh Jain, Runa Eschenhagen, Richard E Turner, and Rio Yokota. Practical deep learning with bayesian principles. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/b53477c2821c1bf0da5d40e57b870d35-Paper.pdf.

Theodore Papamarkou, Maria Skoularidou, Konstantina Palla, Laurence Aitchison, Julyan Arbel, David Dunson, Maurizio Filippone, Vincent Fortuin, Philipp Hennig, Aliaksandr Hubin, Alexander Immer, Theofanis Karaletsos, Mohammad Emtiyaz Khan, Agustinus Kristiadi, Yingzhen Li, Stephan Mandt, Christopher Nemeth, Michael A Osborne, Tim GJ Rudner, David Rügamer, Yee Whye Teh, Max Welling, Andrew Gordon Wilson, and Ruqi Zhang. Position paper: Bayesian deep learning in the age of large-scale AI. *arXiv preprint arXiv:2402.00809*, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Florian Seligmann, Philipp Becker, Michael Volpp, and Gerhard Neumann. Beyond deep ensembles: A large-scale evaluation of bayesian deep learning under distribution shift. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=jX49iKr6vb.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Meiqi Sun, Wilson Yan, Pieter Abbeel, and Igor Mordatch. Quantifying uncertainty in foundation models via ensembles. In *NeurIPS 2022 Workshop on Robustness in Sequence Modeling*, 2022. URL https://openreview.net/forum?id=LpBlkATV24M.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Dustin Tran, Mike Dusenberry, Mark Van Der Wilk, and Danijar Hafner. Bayesian layers: A module for neural network uncertainty. *Advances in neural information processing systems*, 32, 2019.

Xi Wang, Laurence Aitchison, and Maja Rudolph. Lora ensembles for large language model fine-tuning. *arXiv preprint arXiv:2310.00035*, 2023.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gEZrGCozdqR.

Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *CoRR*, abs/2002.08791, 2020. URL https://arxiv.org/abs/2002.08791.

Yuxin Xiao, Paul Pu Liang, Umang Bhatt, Willie Neiswanger, Ruslan Salakhutdinov, and Louis-Philippe Morency. Uncertainty quantification with pre-trained language models: A large-scale empirical analysis. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7273–7284, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.538. URL https://aclanthology.org/2022.findings-emnlp.538.

Boyang Xue, Jianwei Yu, Junhao Xu, Shansong Liu, Shoukang Hu, Zi Ye, Mengzhe Geng, Xunying Liu, and Helen Meng. Bayesian transformer language models for speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7378–7382. IEEE, 2021.

Adam Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. Bayesian low-rank adaptation for large language models. In *Socially Responsible Language Modelling Research*, 2023. URL https://openreview.net/forum?id=9O6hbiDLU7.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F. Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. Benchmarking llms via uncertainty quantification, 2024.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1009. URL https://aclanthology.org/D18-1009.

Ruqi Zhang, Chunyuan Li, Jianyi Zhang, Changyou Chen, and Andrew Gordon Wilson. Cyclical stochastic gradient mcmc for bayesian deep learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkeS1RVtPS.

Appendix A. Related Work

LLM Fine-tuning. In recent years, a number of methods for making more efficient use of fine-tuning have emerged. Two notable mentions are transfer learning (Houlsby et al., 2019), where a pre-trained LLM is adapted to new tasks or domains, enabling models to leverage vast, pre-learned knowledge bases for a wide range of applications, and zero-shot learning (Wei et al., 2022), where models infer correct responses without prior exposure to specific task examples, which showcases the impressive generalization capabilities of LLMs. While the aforementioned methods use traditional fine-tuning to efficiently generalize and transfer knowledge, Parameter Efficient Fine-Tuning (PEFT; Ding et al., 2022, 2023; He et al., 2023) targets the computational efficiency of the underlying fine-tuning methods.

One particularly noteworthy PEFT method (and the one that we focus on in our experiments) is low-rank adaptation (LoRA; Hu et al., 2022). LoRA introduces a set of low-rank matrices whose outputs are concatenated with several layers throughout the model and optimizes only this comparatively small number of fine-tuning parameters. Consequently, LoRA reduces the number of trainable parameters by three to four orders of magnitude from full model fine-tuning, despite achieving comparable or even superior performance. A key advantage of LoRA is that the low-rank matrices can simply be added to the pre-trained weights, thereby not inducing any increased inference time. As the number of trainable parameters is drastically reduced compared to traditional fine-tuning, the memory requirements of the optimizer states experience the same reduction, which allows for the fine-tuning of much larger models given identical hardware constraints. For instance, a recent study (Chen et al., 2023) has illustrated that fine-tuning LLMs with LoRA can significantly enhance their capacity to handle larger context windows with only a minimal increase in computational costs. Furthermore, LoRA modules trained on different tasks can be stored and used to efficiently switch between models optimized for different downstream tasks, significantly reducing the storage required for the usage of several different fine-tuned LLMs. Recently, a number of other works around LoRA have been published, attempting to further improve its efficiency and flexibility, among others, QLoRA (Dettmers et al., 2023) and GLoRA (Chavan et al., 2023).

Uncertainty estimation for LLMs. The bulk of previous work on Bayesian inference for LLMs has focused on pre-training (Tran et al., 2019; Xue et al., 2021; Cinquin et al., 2022; Zhang et al., 2020; Chen and Li, 2023), which is quite computationally expensive and additionally does not improve these models much, as large pre-trained models typically already have reasonably good calibration (Kadavath et al., 2022). Additionally, it has been shown that approximate Bayesian inference on posteriors over subspaces of the full parameter space actually produces accurate predictions as well as well-calibrated predictive uncertainty in both regression and classification settings (Izmailov et al., 2020). Therefore, subspace inference is a particularly interesting approach to making LLMs Bayesian.

Recent works that combined language model fine-tuning with ensembles consider either full fine-tuning (Gleave and Irving, 2022; Sun et al., 2022) or introduce ensembles consisting of two members: one trained with full fine-tuning, and the other fine-tuned with LoRA adapters (Hewitt et al., 2021). The first of these methods requires the storage of M sets of model parameters, one for each of the M ensemble members, which can become impractical

for the larger LLMs. While the latter method requires fewer parameters to be stored, it is limited by the small ensemble size.

More recently, methods for combining uncertainties with LoRA were proposed, which are Laplace-LoRA (Yang et al., 2023), LoRA Ensembles (Balabanov and Linander, 2024; Wang et al., 2023), MC-Dropout, and Bayes by Backprop (Andersen and Maalej, 2022). Moreover, non-Bayesian methods of quantifying uncertainties such as conformal prediction (Ye et al., 2024) have also been applied in this context. Laplace-LoRA imposes a Laplace approximation on the posterior over the LoRA parameters, while LoRA ensembles construct an ensemble over the LoRA adapters.

In contrast to these existing works, we show that it is possible to achieve similar results with the use of a simple and inexpensive stochastic weight averaging approach.

Out-of-distribution (OOD) detection. Detecting out-of-distribution (OOD) data is crucial for ensuring the reliability and safety of machine learning systems. In open-world scenarios, many models may encounter test samples that are out-of-distribution (OOD), requiring careful handling. Such distributional shifts may arise from semantic changes (Hendrycks and Gimpel, 2017), where OOD samples belong to different classes, or from covariate shifts (Ben-David et al., 2010), where OOD samples come from a different domain.

The integration of Bayesian components into Large Language Models (LLMs) and the resulting uncertainty estimates naturally equip Bayesian LLMs with the capability to detect out-of-distribution (OOD) samples. Given the diverse landscape of Bayesian Deep Learning (BDL) methods and settings, direct comparisons are challenging. However, a recent review by Seligmann et al. (2023), covering a range of BDL approaches in realistic OOD scenarios, found that fine-tuning just the final layers of pre-trained models with BDL algorithms significantly improves both generalization accuracy and calibration on data with realistic distribution shifts, while only slightly increasing runtime overhead. Moreover, these models are often comparable to or even exceed the performance of specialized OOD generalization methods. This finding is supported by the notable enhancements reported by recent Bayesian LoRA-based methodologies (Yang et al., 2023; Balabanov and Linander, 2024; Wang et al., 2023).

Particularly, the post-hoc implementation of Laplace approximation and LoRA fine-tuning, as demonstrated in Yang et al. (2023), has improved calibration and uncertainty estimation. Our work shares the same intuition, drawing on the premise that Bayesian approaches, as they have employed, are instrumental in improving model calibration. This is particularly relevant for LLMs in fine-tuning contexts, where available data is scarce compared to the pre-training phase, highlighting the efficacy of Bayesian methods in navigating uncertainty with limited datasets.

Drawing on the empirical results and insights of the work by Wang et al. (2023), our work is motivated by the findings that LoRA ensembles significantly enhance both the accuracy and calibration of these models, surpassing the outcomes of basic LoRA fine-tuning and other methods like last-layer fine-tuning or Monte Carlo dropout. This evidence lends strong support to the efficacy of ensembling techniques, to which SWAG presented in this work belongs, in not just boosting model performance but also in fine-tuning calibration. Wang et al. (2023) also explore the realm of regularized LoRA ensembles, echoing the classical notion (Breiman, 2001) that diversity among ensemble components is pivotal for

generalization. Recent works (Lakshminarayanan et al., 2017; Kumar et al., 2022; Fort et al., 2019) suggest that these ensembling principles hold true for deep learning architectures as well.

To evaluate how well our models are calibrated and their effectiveness in OOD detection, we utilize established metrics such as negative log-likelihood (NLL), expected calibration error (ECE), and the Brier score (Guo et al., 2017; Osawa et al., 2019). We also report on entropy and cross-model uncertainty metrics, following the argument by Malinin et al. (2020) that the entropy score alone cannot distinguish between epistemic (related to the methodology or model) and aleatoric (inherent to the data) uncertainties, while cross-model uncertainty like model disagreement offers an estimate of epistemic uncertainty only (i.e., model-based). Further discussions are presented in Subsection 3.1.