TSUBAME 共同利用　令和 6 年度　学術利用　成果報告書

利用課題名　計算資源が限られた状況下における Low-Rank Adaptation Fine-Tuning に関する研究
英文：A Study on Low-Rank Adaptation Fine-Tuning under Limited Computational Resources

利用課題責任者
波多野 賢治
Kenji Hatano

同志社大学文化情報学部
Doshisha University, Faculty of Culture and Information Science
https://www-mil.cis.doshisha.ac.jp

本課題では，学習済みの LoRA アダプタが，重みのビット幅が異なる基盤モデル間で転移可能であるかどうかを検証した．具体的には，複数のタスクとモデルを用いて，低ビット幅へ量子化された基盤モデル上で学習された QLoRA アダプタが，より大きなビット幅の基盤モデル上でも有効に動作する事を明らかにした．特に，重みを 4 ビット幅や 3 ビット幅に量子化した基盤モデル上で学習したアダプタであっても，推論時に重みを 16 ビット幅に復元することで，基盤モデルを量子化しない通常の LoRA とほぼ同等の推論精度を達成できることを示した．本手法は，学習済み QLoRA モデルの性能向上や，量子化によって得られた余剰メモリを活用したバッチサイズの拡大による LoRA の学習速度の向上などに活用可能である．

In this study, we investigated whether trained LoRA adapters can be transferred across foundation models with different weight bit-widths. Specifically, we showed that LoRA adapters trained on foundation models quantized to low bit-widths can still function effectively when applied to foundation models with higher bit-widths, using multiple tasks and models. We demonstrated that even adapters trained on 4-bit or 3-bit quantized foundation models can achieve inference accuracy comparable to standard LoRA trained at 16-bit precision, by restoring the foundation model to 16-bit at inference time. This approach can be applied to enhance the performance of pre-trained QLoRA models and to accelerate LoRA training by expanding the batch size using memory savings gained through quantization.

Keywords: PEFT, LoRA, QLoRA, Quantization, Transferability

背景と目的

大規模言語モデルをはじめとする深層学習モデルは，重みの勾配や最適化状態を保持するため，推論時よりも訓練時に要求される GPU メモリ量が多いことが知られている．したがって，例えば同一の計算資源上で基盤モデルのファインチューニングと推論を行う際には，基盤モデルのサイズは訓練時の資源制約から決定されるため，推論時にはメモリの余剰が発生する．

その一方で，既存手法はこの余剰メモリを考慮しておらず活用できていない[1,2,3]．そこで本課題では，この推論時のメモリの余剰を活用して LoRA モデルの性能を向上させる新たな 量子化-LoRA フレームワークとして，Post LoRA Restoration（PLR）を提案する．PLR では，訓練時の資源制約から生まれる推論時のメモリの余剰を活用し，量子化された基盤モデルの重みのビット幅の復元を行う．本手法はあるビット幅の基盤モデル上で学習したアダプタが他のビット幅の基盤モデルの上でも効果的に機能できるという転移性を LoRA アダプタが持つという仮説に基づいている．

TSUBAME4.0 の計算資源を用いて評価実験を行った結果，PLR の適用による精度の向上を確認でき，アダプタの転移性および提案手法の有効性を確認できた．

概要

提案手法である PLR の概念図を図1に示す．モデルの重みを k ビットへ量子化した後に LoRA を実施し推論を行う QLoRA [2] に対し，PLR では推論時のメ

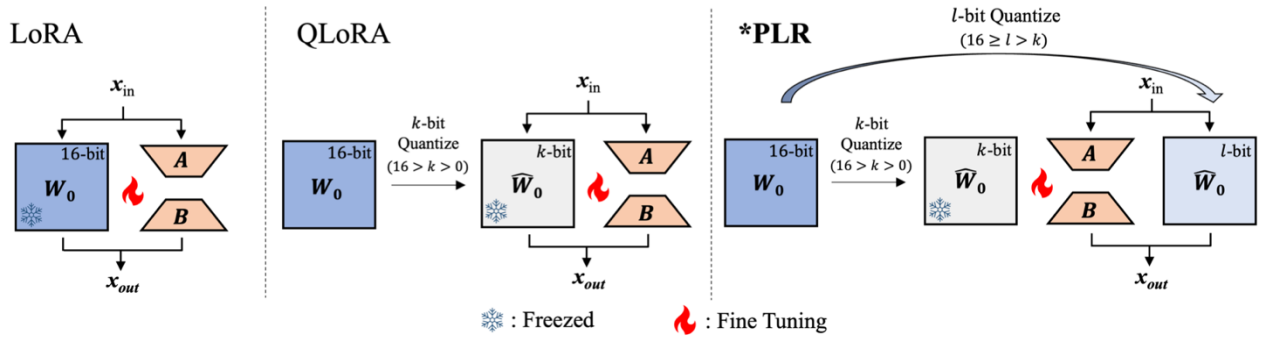

図1：LoRA, QLoRA, PLR の概念図

モリ余剰に着目して，量子化-LoRA モデルを学習したのちに 量子化-LoRA の基盤モデルの重みの精度をより高いビット幅へ復元する．

PLR の具体的な流れは，以下の通りである．基盤モデルの重みを $\boldsymbol{W}_0^{16\text{-bit}}$，LoRA アダプタの重みを $\Delta\boldsymbol{W}_{\text{LoRA}}$ とすると，LoRA 適用後の各層の重み $\boldsymbol{W}$ は，$\boldsymbol{W} = \boldsymbol{W}_0^{16\text{-bit}} + \Delta\boldsymbol{W}_{\text{LoRA}}$ で表される．QLoRA などの量子化-LoRA 手法では，$\boldsymbol{W}_0^{16\text{-bit}}$ を k ビット幅へ量子化し $\boldsymbol{W}_0^{k\text{-bit}}$ に置き換えた，$\boldsymbol{W} = \boldsymbol{W}_0^{k\text{-bit}} + \Delta\boldsymbol{W}_{\text{LoRA}}$ について，$\Delta\boldsymbol{W}_{\text{LoRA}}$ の訓練を行う．ここで，k は $16 > k > 0$ である．PLR では，QLoRA の訓練完了後，$\boldsymbol{W}_0^{k\text{-bit}}$ を $\boldsymbol{W}_0^{l\text{-bit}}$ の重みへ置き換えた，$\boldsymbol{W} = \boldsymbol{W}_0^{l\text{-bit}} + \Delta\boldsymbol{W}_{\text{LoRA}}$ によって推論が行われる．ここで，l は $l > k$ を満たし，推論を行う計算資源の余剰に合わせて選択される．

一般的に，LLM の量子化は低いビット幅であるほど近似値に丸め込まれた際に生じる量子化誤差が大きく，性能が低下することが知られている．PLR を適用することでよりビット幅の大きく劣化の少ない基盤モデルを用いることが可能となるため，量子化誤差から生じる基盤モデルの性能低下を回避し QLoRA モデルの性能を復元できることが期待される．

PLR の有効性を確かめるために，QLoRA モデルに対し PLR を適用した際に下流タスクの正解率が向上するかを検証する．使用するモデルは，Llama3 ファミリー[4]のうち，Llama 3.2-1B, Llama 3.2-3B, Llama 3.1-8B の三種類のモデルサイズを選定した．また，評価対象とするタスクには，Grade School Math 8K (GSM8K) [4] および SQL Create Context (SCC) を使用する．

次に，学習の設定について述べる．LoRA および QLoRA では，Adapter の Rank と は，すべてのモデルで 16 に設定し，LoRA を適用する LLM の層は Attention 層のみとした．学習率は 2e-4，バッチサイズは 32 とし，10 エポック学習した内，検証データに対する Loss が最も低いエポックの重みを評価に使用する，QLoRA および PLR で用いる基盤モデルの量子化手法には，GPTQ [7] を使用し，8,4,3,2 ビットの4パターンのビット幅で PLR の有効性を検証する．GPTQ は，重みの量子化後にキャリブレーションデータを用いて補正を行い，量子化誤差を最小限に抑える手法である．キャリブレーションデータには，訓練データから無作為抽出した 500 件を使用する．これらの実装は，Python 3.9.18，Cuda 12.8，PyTorch 2.5.0，Transformers 4.47.0，Peft 0.14.0 のソフトウェアおよびモジュールを用いて，TSUBAME4.0 の計算機上で行われた．

結果および考察

上記の設定で学習した LoRA，QLoRA，QLoRA+PLR モデルの各タスクに対する正解率を，表1に示す．なお，表中の PLR16 や PLR8 は，QLoRA の学習後にそれぞれ 16-bit，8-bit へ基盤モデルの重みを復元することを指す．

モデルとデータセット全体を見ると，ほぼすべてのケースにおいて PLR の適用によって正解率が向上しており，PLR の有効性が確認された。一方で，8-bit においては唯一 PLR の性能が低下している．ここで，16-bit LoRA に着目すると，16-bit LoRA よりも 8-bit QLoRA および 8-bit QLoRA + PLR16 の正解率が高い．このことから，基盤モデルとしての性

表1：LoRA，QLoRA および QLoRA+PLR の各モデルのそれぞれのタスクに対する正解率

Datasets	Models	16-bit LoRA	8-bit QLoRA		4-bit QLoRA			3-bit QLoRA				2-bit QLoRA				
			QLoRA	PLR16	QLoRA	PLR8	PLR16	QLoRA	PLR4	PLR8	PLR16	QLoRA	PLR3	PLR4	PLR8	PLR16
GSM8k	Llama 3.2-1B	21.60	**22.13**	21.83	18.19	**18.95**	**19.41**	11.22	**15.31**	**16.91**	**16.53**	**3.18**	3.11	3.03	2.81	2.43
	Llama 3.2-3B	42.91	**43.06**	42.07	37.75	**41.32**	**42.38**	29.56	**34.50**	**37.76**	**37.30**	4.17	**6.60**	**15.61**	**15.39**	**15.85**
	Llama 3.1-8B	59.66	58.75	**58.98**	56.56	**58.15**	**58.00**	50.34	**57.31**	**58.98**	**59.59**	5.16	**18.27**	**33.35**	**33.73**	**34.34**
	Avg.	41.39	**41.31**	40.96	37.50	**39.47**	**39.93**	30.37	**35.71**	**37.88**	**37.81**	4.17	**9.33**	**17.33**	**17.31**	**17.54**
SCC	Llama 3.2-1B	69.21	**67.29**	63.33	25.84	**30.11**	**31.60**	23.90	22.19	**37.69**	**37.63**	0.00	0.00	**0.22**	**0.22**	**0.19**
	Llama 3.2-3B	72.30	**83.02**	82.79	75.22	**80.49**	**80.41**	68.65	68.15	**75.54**	**75.44**	2.97	**32.93**	**39.51**	**35.83**	**35.87**
	Llama 3.1-8B	84.96	**84.30**	84.21	84.38	**84.47**	**84.46**	77.35	**79.46**	**80.04**	**79.94**	4.81	**36.73**	**32.20**	**36.13**	**36.20**
	Avg.	75.49	**78.20**	76.78	61.81	**65.02**	**65.49**	56.63	56.60	**64.42**	**64.34**	2.59	**23.22**	**23.98**	**24.06**	**24.09**
Avg.	Llama 3.2-1B	45.40	**44.71**	42.58	22.02	**24.53**	**25.51**	17.56	**18.75**	**27.30**	**27.08**	1.59	1.55	**1.63**	1.52	1.31
	Llama 3.2-3B	57.61	**63.04**	62.43	56.49	**60.91**	**61.40**	49.11	**51.33**	**56.65**	**56.37**	3.57	**19.77**	**27.56**	**25.61**	**25.86**
	Llama 3.1-8B	72.31	71.53	**71.60**	70.47	**71.31**	**71.23**	63.85	**68.39**	**69.51**	**69.77**	4.99	**27.50**	**32.78**	**34.93**	**35.27**
	Avg.	58.44	**59.76**	58.87	49.66	**52.25**	**52.71**	43.50	**46.15**	**51.15**	**51.07**	3.38	**16.27**	**20.65**	**20.69**	**20.81**

能が 16-bit のものよりも 8-bit の方が優れていることが推測でき，これが基盤モデルを 16-bit へ復元する PLR16 が 8-bit QLoRA に敗北した要因であると解釈できる．この 8-bit が 16-bit に打ち勝つ現象は，量子化による正則化効果により 16-bit よりも 8-bit の方がアダプタの学習が効率的に進んだためであると考えられる．

次に，モデルサイズごとに結果を比較すると，モデルサイズが大きく，性能が高いほど，PLR の有効性が高まっていることが読み取れる．特に，GSM8k の 3-bit QLoRA+PLR16 や SCC の 4-bit QLORA+PLR16 の精度は，いずれも 16-bit LoRA と同等の精度に達している，これは，4-bit や 3-bit 幅の量子化基盤モデル上で，通常の LoRA と同程度にアダプタがタスクの解法を学習できていることを示唆している．この結果から，LoRA と比較した際の QLoRA による性能低下は量子化誤差による基盤モデルの性能劣化の影響を受けたものであり，一定のビット幅まではアダプタ自体の学習は順調に進んでいることが明らかになった．

まとめと今後の課題

本課題では，低ビット幅の基盤モデル上で学習された LoRA アダプタが，より高いビット幅の基盤モデル上でも有効に動作するかを検証した．実験の結果，3 ビットまでの QLoRA アダプタは，通常の LoRA アダプタと同程度の性能を有することが確認できた．本手法を適用することにより，学習済み QLoRA モデルの性能を追加の学習なしに向上させることが可能である．

今後の課題としては，具体的にどれほどの余剰メモリがある場合に PLR が適用可能かについての調査を行う予定である．また，仮に余剰メモリがないケースでも，処理の一部を CPU に分散させるオフローディング技術などの推論時の最適化手法を用いることで，余剰メモリを創り出せる可能性がある．そうした最適化手法の併用も追加実験として行う必要がある．

参考文献

[1] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In International Conference on Learning Representations, 2021.

[2] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: efficient finetuning of quantized llms. In Proceedings of the 37th International Conference on Neural Information Processing Systems, pp. 10088–10115, 2023.

[3] Yuhui Xu, Lingxi Xie, Xiaotao Gu, Xin Chen, Heng Chang, Hengheng Zhang, Zhengsu Chen, XIAOPENG ZHANG, and Qi Tian. Qa-lora: Quantization-aware low-rank adaptation of large language models. In The Twelfth International Conference on Learning Representations, 2024.

[4] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

[5] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. arXiv preprint

arXiv:2110.14168, 2021.

[6] b mc2. sql-create-context dataset, 2023. URL https://huggingface.co/datasets/b-mc2/sql-create-context.

[7] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. In The Eleventh International Conference on Learning Representations, 2023.