連合学習における LoRA の統合数と精度の関係の検証

尹子旗　村田栄樹　河原大輔
早稲田大学理工学術院
{yinziqi2001@toki., eiki.1650-2951@toki., dkw@}waseda.jp

概要

本研究では、大規模言語モデル（LLM）の連合学習における LoRA の統合数がモデル性能に与える影響について検証する。LLM に基づく連合学習の多くは、複数のクライアントがローカルデータを用いて LoRA を学習し、その重みを統合することでグローバルモデルを構築する。しかし、LoRA の統合数が増加するにつれ、モデル性能の低下が予想されるため、その定量的評価と原因の解析が必要である。本研究では、データ分布が同一であることを仮定し、テキスト要約タスクおよび数学的推論タスクを対象に実験する。その結果、LoRA の統合数が増加するにつれてモデル性能が低下する傾向が観察された。特に数学的推論タスクでは、統合数の増加によって正解率が顕著に低下した。これらの結果は、従来の統合手法である FedAvg が LoRA 重みの統合において性能劣化を招くことを示しており、統合手法の改善が求められることを示唆する。

1　はじめに

連合学習（Federated Learning）は、データ共有の制約が厳しい環境で、複数のクライアントが協力してモデルを学習するための手法として注目されている [1]。例えば、複数の医療機関が患者データを共有せずに診断モデルを学習する場合や、複数の企業が競争力を保ちながらデータ分析モデルを構築する場合などが想定される。このような場面では、各クライアントが独自のデータを持ちながらも、単一のクライアントのデータだけでは十分なモデル性能が得られない場合が多い。一方で、データ共有の制約があるため、クライアント間で共有されるのは学習したモデルの重みや勾配情報に限られる。連合学習により、各クライアントはローカルデータを共有せずにプライバシを保護しつつ、学習した情報を集約することによって単独のクライアントでは得られない高性能なモデルを構築することが可能となる。

LoRA（Low-Rank Adaptation）[2] は、LLM の効率的な学習手法として提案され、連合学習環境においても通信コストを削減しつつ高いモデル性能を維持する技術として注目されている [3, 4]。LoRA を活用した連合学習では、複数のクライアントがそれぞれのデータで学習した LoRA の重みをサーバに送信し、それらを統合してグローバルモデルを更新する。この統合には、通常、FedAvg（Federated Averaging）[5] が使用される。FedAvg は統合時に重みの算術平均を計算する手法で、広く採用されている。しかし、クライアント数が増加するにつれ、性能がどのように変化するかの検証が行われていなかった。

本研究は、連合学習における LoRA の統合数の増加がモデル性能に与える影響を定量的に評価することを目的とする。連合学習が適用される多くの場面において、クライアント間のデータ分布が同一であるという仮定が成り立つことを考慮する。この仮定の下、一つのデータセットをランダムに分割して対応する LoRA の学習に用い、FedAvg によって統合する形で実際の応用場面を再現し、その性能を検証する。

検証の結果、FedAvg を適用する際、LoRA の統合数の増加につれモデルの性能が低下する傾向が見られた。この結果は、LoRA の統合における新たな手法の必要性を示唆しており、今後の効率的な連合学習モデルの設計に向けた新たな指針を提供するものである。

2　関連研究

2.1　連合学習

連合学習は、分散学習の一種であり、複数のクライアントが自身のデータを共有せずに協調して機械学習モデルを訓練する手法である。各クライアントはローカルにあるデータを用いてモデルを学習し、その更新情報である重みや勾配情報のみをサーバに

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

送信する。サーバはこれらの更新情報を集約してグローバルモデルを更新し、再度クライアントに配布する形で学習を進める。このデータを共有しない手法により、プライバシ情報を保護しつつモデルの学習ができる。

LLM における連合学習は、同じ仕組みで学習を行う [6]。この手法はデータのプライバシ保護を可能にする一方で、LLM の規模により通信コストの課題が存在する。この課題を解決するために、モデル全体の重みの代わりに LoRA の重みのみをサーバに送信する手法が提案されている [7]。

2.2 LoRA

LoRA（Low-Rank Adaptation）は、LLM の効率的なファインチューニングを可能にする手法の一つである [2]。この手法では、事前学習済みモデルの重みを固定し、各層に訓練可能な低ランク行列を追加することで、訓練の効率化を実現する。LoRA の最大の利点は、通常のファインチューニングに比べて学習パラメータの数を大幅に削減できる点にある。これにより、LLM のファインチューニングに必要な計算資源を大幅に抑えつつ、高い性能を維持することが可能になる。

LoRA の構造は、モデルの重み行列 $W \in \mathbb{R}^{d\times k}$ に対して、低ランク行列 $A \in \mathbb{R}^{d\times r}$ および $B \in \mathbb{R}^{r\times k}$ を導入し、以下のように表現される：

$$W' = W + BA. \tag{1}$$

ここで、W は元のモデルの重み行列、W' は更新後のモデルの重み行列、d は行列の行数（特徴量の次元数）、k は行列の列数（出力の次元数）、A および B はそれぞれ低ランク行列である。r はランクを表し、$r \ll \min(d,k)$ かつ 導入されるパラメータの数は $r(d+k)$ であるため、この数は抑えられる。

連合学習においては、クライアントが LoRA の重み (A, B) をサーバに送信することで、通信コストの削減が実現できる。通常、LLM 全体の重みを共有する場合、数百 MB から数 GB 単位の通信が必要となる。しかし、LoRA のアプローチでは、共有されるのは低ランク行列の情報のみであり、これにより通信量がモデル全体の重みに比べて大幅に削減される。

2.3 連合学習におけるモデル統合

LoRA を用いた連合学習における更新情報を集約してグローバルモデルを更新する際は、複数の LoRA 重みを統合する手法を適用することができる。よく用いられる手法として、FedAvg [5] という重みの算術平均をとる手法がある。FedAvg は単純かつ計算効率が高い一方で、LoRA のように学習した重みが特定のタスクや局所的な解に依存する場合、統合後のモデル性能が劣化する問題がある。この問題を克服するために、重みに対する算術操作を避ける Mixture-of-Experts (MoE) の手法を LoRA に応用した MoLE という手法が提案されている [8]。MoLE は、FedAvg という勾配に対する手法をこのまま LoRA の重みに適用することによるモデルの悪化を回避できる特徴を持つ。しかし、エキスパートの数が過大に増加するにつれ、性能の低下が見られるため [8]、連合学習のような多数のクライアントから送信された LoRA エキスパートの統合に不適であることがわかる。そのため、FedAvg をはじめとする算術平均をとる手法が依然として主流である。一方、それらの研究において、クライアントの数は非常に限定され、数十 LoRA の統合しか実験されておらず、大規模な実験は実施されていなかったほか、統合するときにノイズが加えられることも懸念される [9]。

3 問題設定

本研究では、連合学習における LoRA の統合数の増加がモデル性能に及ぼす影響を分析し、この関係性を解明することを目指す。連合学習の応用場面の多くでは、すべてのクライアントにおけるデータ分布は同じと考えられる。このため、本研究はこのような場面を想定し、一つのデータセットをランダムに分割し、それぞれに対応するクライアントで LoRA を学習する。

具体的には、データセット $\mathcal{D}$ を N 個のクライアントに分割し、各クライアント k が保有するデータ $\mathcal{D}_k$ を次のように定義する：

$$\mathcal{D} = \bigcup_{k=1}^{N} \mathcal{D}_k, \quad \mathcal{D}_i \cap \mathcal{D}_j = \emptyset \ (\forall i \neq j), \quad |\mathcal{D}_k| = \frac{|\mathcal{D}|}{N}. \tag{2}$$

ここで、$|\mathcal{D}_k|$ は各クライアントのデータサイズであり、すべてのクライアントにおいて等しい。また、データはランダムに割り当てられるため、データの分布はすべての $\mathcal{D}_k$ に対して均等である。

各クライアント LoRA に対して $LoRA_k$ はローカルデータ $\mathcal{D}_k$ を用いて LoRA の重み A_k と B_k を学習する。そして FedAvg の手法でこれらの LoRA を統

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

合する。統合されたモデルは

$$W' = W + \frac{1}{N}\sum_{i=1}^{N} B_i \times \frac{1}{N}\sum_{i=1}^{N} A_i \tag{3}$$

となる。

4 実験

4.1 実験設定

連合学習における LoRA の統合数の増加がモデル性能に与える影響を評価するため、複数のタスクを対象に実証実験を行う。具体的には、事前学習済みの Llama3.2-1B[1] をベースモデルとし、テキスト要約タスクと数学的推論タスクの 2 つのタスクを選定する。それぞれのタスクにおいて、異なる LoRA の統合数の設定でモデル性能を比較する。

4.1.1 テキスト要約タスク

テキスト要約タスクには、広く利用されている CNN/Daily Mail データセット [10] を用いる。このタスクにおいては、LoRA の統合数 N を 1, 8, 64, 256, 1024 に設定する。公平性を担保するために、すべての LoRA の学習時に計算量の合計を一定にする方法として、epoch 数を 1 にする。ただし LoRA の統合数が 1024 の場合は、LoRA の収束を確保するために epoch 数は 2 にした。要約性能の評価には、一般的な指標である ROUGE-1 と ROUGE-L [11]、および文脈的意味を考慮する BERTScore [12] を用いることで、要約の品質を多角的に評価する。

4.1.2 数学的推論タスク

数学的推論タスクには、算術推論のベンチマークデータセットである GSM8K（Grade School Math 8K）[13] を使用する。GSM8K は小学生レベルの数学問題を対象としたデータセットであり、正確な推論能力を評価するためのタスクとして適している。このタスクにおいては LoRA の統合数を 1, 4, 16, 64, 256 に設定し、モデル性能を比較する。同様に公平性を担保するために、epoch 数を 20 にする。ただし LoRA の統合数が 256 の場合は、収束を確保するために epoch 数は 40 にした。数学的推論の評価指標としては、出力が正解と完全一致する割合を示す Exact Match を採用する。Exact Match の基準として、GSM8K の評価方法と同様に、最終回答のフォーマット（付録 A 参照）の前の回答過程を無視し、最終回答のフォーマットの後ろにある内容のみを評価の対象とする。

1） https://huggingface.co/meta-llama/Llama-3.2-1B

表 1 CNN/Daily Mail における LoRA 数と統合したモデルの性能．BERTScore の P は Precision，R は Recall を示す．

LoRA 数	ROUGE		BERTScore		
	1	L	P	R	F1
1	37.11	24.65	88.62	86.25	87.40
8	35.24	23.80	88.26	85.93	87.06
64	35.12	23.75	88.34	85.85	87.06
256	34.73	23.54	88.22	85.82	86.99
1024	34.91	23.62	88.11	85.94	86.99

表 2 CNN/Dailymail における LoRA 数と統合したモデルの性能．BERTScore の P は Precision，R は Recall を示す．

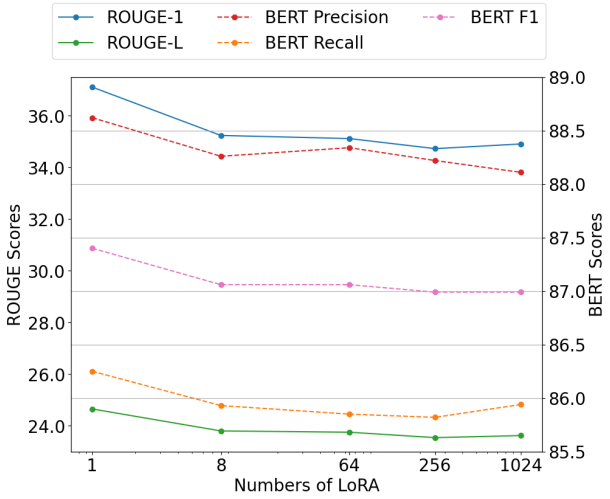

図 1 CNN/Dailymail における LoRA 数と統合したモデルの性能

4.2 実験結果

4.2.1 テキスト要約タスク

テキスト要約タスクの実験結果を表 2 に示す。LoRA の統合数が増加するにつれ、ROUGE スコアおよび BERTScore が減少する傾向が観察された。例えば、統合数が 1 から 256 に増加した場合、ROUGE-1 は 37.11 から 34.73 へ、ROUGE-L は 24.65 から 23.54 へと減少した。BERTScore においても全ての指標が低下した。これらの結果は、LoRA 統合数の増加がモデルの要約能力に一定の悪影響を及ぼすこと示している。一方で、統合数が増えても性能低下は緩やかであり、LoRA 統合のノイズに対する耐性を持つ可能性があると考えられる。

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 3　GSM8K における LoRA 数と統合したモデルの性能

LoRA 数	Exact Match
1	13.04
4	10.54
16	8.34
64	5.38
256	3.11

4.2.2　数学的推論タスク

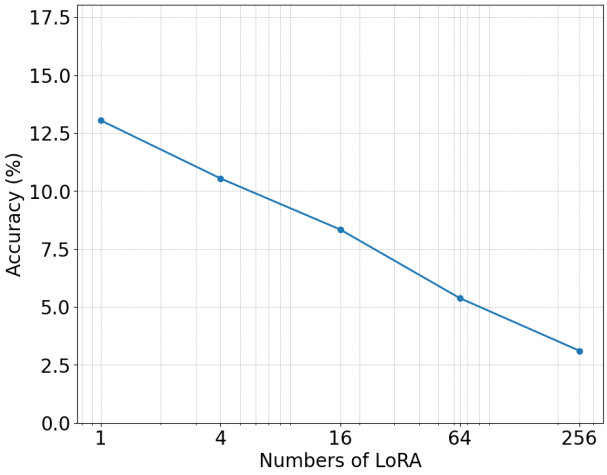

図 2　GSM8K における LoRA の統合数と統合したモデルの性能

数学的推論タスクの実験結果を表 3 に示す。表より、LoRA 統合数の増加がモデル性能に対して顕著な悪影響を与えることが分かる。具体的には、統合数が 1 から 256 に増加するに従い、完全一致率は 13.04 から 3.11 へと大幅に低下した。特に統合数が 64 を超えた段階で、正解率の急激な悪化が観察された。この結果は、数学的推論タスクが高い精度と一貫性を必要とする性質を持つため、FedAvg による LoRA 統合の過程で発生するノイズや局所最適解のばらつき [14] が、グローバルモデルの性能に致命的な影響を与える可能性を示している。

4.2.3　分析

テキスト要約タスクと数学的推論タスクの異なるクライアントの学習ロスを図 3 と図 4 に示す。テキスト要約タスクでは、モデルが学習するべき解が柔軟で、多様性が許容されやすい [15]。このため、ロス地形の谷が広くなだらかであり、各 LoRA が似通った谷底に収束しやすく、統合時にこれらの類似した局所最適解を平均化しても、グローバルモデルの性能に大きな悪影響を及ぼさない結果となったと考えられる。

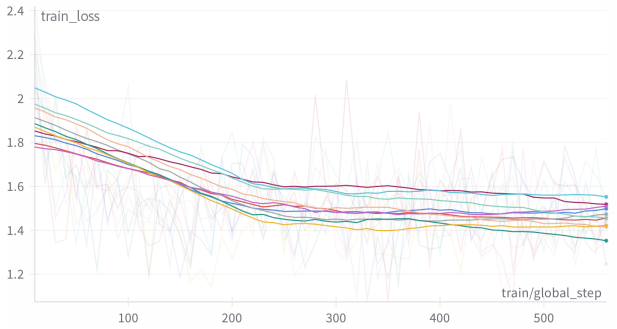

図 3　CNN/Daily Mail における学習ロス (N=256)

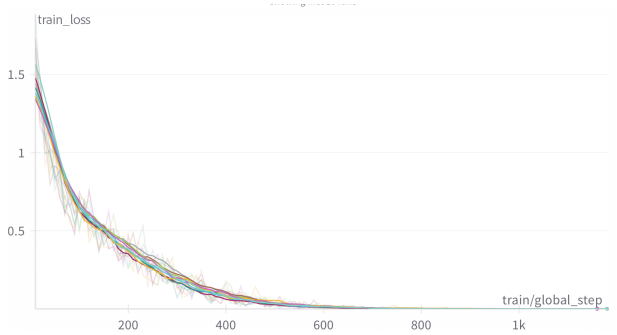

図 4　GSM8K における学習ロス (N=64)

一方、数学的推論タスクでは、統合数の増加に伴い、モデル性能が急激に低下する傾向が観察された。この原因として、数学的推論タスクが高精度かつ一貫した推論を必要とする点が挙げられる。このようなタスクでは、ロス地形が複雑で、谷が細かく分断されることが多い [14] と考えられる。各 LoRA が異なるデータセットで学習した場合、それぞれが異なる局所的最適解に到達する可能性が高く、統合時にはこれらの異なる最適解がノイズとなって統一された解から外れることがある。

5　おわりに

本研究では、連合学習における LoRA の統合数がモデル性能に与える影響を分析し、統合数の増加に伴い性能が低下する傾向を明らかにした。特に、数学的推論タスクでは大幅な性能低下が見られ、これは FedAvg が LoRA 重みの統合において不適切である可能性を示唆する。一方、テキスト要約タスクでは性能低下が緩やかであり、タスク特性に応じた統合手法の選択が重要であることが確認された。本研究は、LoRA 統合における課題を浮き彫りにするとともに、タスク特性を考慮した新たな統合手法や非独立同分布環境での検証、通信効率向上のための技術開発など、LoRA を活用した連合学習の実用化に向けた今後の研究の方向性を示唆する。

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

謝辞

本研究は「戦略的イノベーション創造プログラム (SIP)」「統合型ヘルスケアシステムの構築」JPJ012425 の補助を受けて実施した。

参考文献

[1] 米谷竜. 連合学習入門. 精密工学会誌, Vol. 87, No. 8, pp. 662–665, 2021.

[2] 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, 2021.

[3] Liping Yi, Han Yu, Gang Wang, Xiaoguang Liu, and Xiaoxiao Li. pfedlora: Model-heterogeneous personalized federated learning with lora tuning, 2024.

[4] Jiaxing QI, Zhongzhi Luan, Shaohan Huang, Carol Fung, Hailong Yang, and Depei Qian. Fdlora: Personalized federated learning of large language model via dual lora tuning, 2024.

[5] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data, 2023.

[6] Chaochao Chen, Xiaohua Feng, Yuyuan Li, Lingjuan Lyu, Jun Zhou, Xiaolin Zheng, and Jianwei Yin. Integration of large language models and federated learning, 2024.

[7] Rui Ye, Rui Ge, Xinyu Zhu, Jingyi Chai, Yaxin Du, Yang Liu, Yanfeng Wang, and Siheng Chen. Fedllm-bench: Realistic benchmarks for federated learning of large language models, 2024.

[8] Xun Wu, Shaohan Huang, and Furu Wei. Mixture of lora experts, 2024.

[9] Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. Improving lora in privacy-preserving federated learning, 2024.

[10] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks, 2017.

[11] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[12] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In **International Conference on Learning Representations**, 2020.

[13] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. **arXiv preprint arXiv:2110.14168**, 2021.

[14] Hao Liao, Wei Zhang, Zhanyi Huang, Zexiao Long, Mingyang Zhou, Xiaoqun Wu, Rui Mao, and Chi Ho Yeung. Exploring loss landscapes through the lens of spin glass theory, 2024.

[15] Tanya Goyal, Nazneen Fatema Rajani, Wenhao Liu, and Wojciech Kryściński. Hydrasum: Disentangling stylistic features in text summarization using multi-decoder models, 2022.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

A　GSM8K の評価方法

以下は数学推論タスク (4.2.2 節) に使用される GSM8K データセットの一例である。

質問文	参照回答
Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?	Natalia sold 48/2 = <<48/2=24>>24 clips in May. Natalia sold 48+24 = <<48+24=72>>72 clips altogether in April and May. #### 72

GSM8K [13] の実験において、評価は終了トークンの####の後続の内容を抽出し、正誤判定を行う方法を使用している。本研究においても同様の方法を採用する。

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).