MOON: Model-Contrastive Federated Learning

Stefanos A Frilingos ELEC70082: Distributed Optimisation and Learning

I. Introduction

Federated learning is an emerging paradigm for training machine learning models across distributed clients, coordinated by a central server, without sharing raw data, thereby preserving privacy. A major challenge in this setting is the presence of non-IID (non-independent and identically distributed) data across clients, which often leads to degraded model performance. This report reviews the **Mo**del-**Con**trastive Federated Learning (MOON) algorithm [1], which addresses this issue through a novel model-level contrastive approach. By aligning local model representations with the global model during training, MOON effectively mitigates representation drift caused by heterogeneous data distributions. The core contributions and methodology of the paper are discussed, followed by a reproduction of a key experimental result to assess the algorithm's effectiveness.

II. Background

Several algorithms have been proposed to address the optimisation challenges introduced by non-IID data in federated learning. Classical approaches such as Federated Averaging (FedAvg) [2] aggregate local updates without accounting for divergence between client and server objectives, often resulting in suboptimal convergence. To mitigate this, FedProx [3] introduces a proximal term to constrain local updates, while SCAFFOLD [4] employs control variates to reduce the impact of client drift. Despite these improvements, such methods may struggle to scale effectively in deep learning applications with highly heterogeneous data, especially complex images.

MOON introduces a novel approach by incorporating ideas from contrastive learning. While traditional contrastive methods such as SimCLR [5] operate on input-level augmentations, MOON applies contrastive learning at the model level. Each client aligns its current local representation with the global model's representation and contrasts it with its own previous version, thereby reducing representation drift caused by non-IID data.

The MOON framework directly relates to several key themes in this module. It builds upon distributed stochastic optimisation, particularly through the use of local SGD and global aggregation. MOON is grounded in federated learning, which involves a client-server architecture where learning is coordinated without direct data sharing. Its contrastive objective introduces an implicit regularisation, akin to dual variables in primal-dual optimisation, to promote alignment between distributed model updates. Furthermore, MOON contributes to the development of new distributed learning algorithms that are robust to data heterogeneity and system-level constraints. As such, it exemplifies how principled algorithm design can address performance trade-offs in real-world distributed optimisation problems.

III. MOON Algorithm

MOON introduces a model-level contrastive learning objective to mitigate the effects of data heterogeneity in federated learning. Unlike conventional methods such as FedAvg, which aggregate local updates without correcting for representation drift, MOON explicitly aligns local and global model representations during training.

A. Contrastive Learning in Model Space

Contrastive learning has proven effective in self-supervised learning by enforcing similarity between augmented views of the same input. MOON extends this principle to the model space: it encourages the current local model to remain close to the global model while distancing itself from its own previous state.

For a given input x, define:

- $z = R_{w_i^t}(x)$: representation from the current local model,
- $z_{\text{glob}} = R_{w^t}(x)$: representation from the global model,
- $z_{\text{prev}} = R_{w_i^{t-1}}(x)$: representation from the previous local model.

The model-contrastive loss is defined as:

$$\mathcal{L}_{\text{con}} = -\log \frac{\exp(\text{sim}(z, z_{\text{glob}})/\tau)}{\exp(\text{sim}(z, z_{\text{glob}})/\tau) + \exp(\text{sim}(z, z_{\text{prev}})/\tau)}, \tag{1}$$

where the cosine similarity is given by:

$$\text{sim}(a, b) = \frac{a \cdot b}{\|a\|\|b\|}, \tag{2}$$

and τ is a temperature parameter controlling the sharpness of the similarity.

B. Federated Optimisation with MOON

The local training objective incorporates both supervised and contrastive terms:

$$\mathcal{L} = \mathcal{L}_{\text{sup}}(w_i^t; (x, y)) + \mu \mathcal{L}_{\text{con}}(w_i^t, w_i^{t-1}, w^t; x), \tag{3}$$

where μ balances the influence of the contrastive loss.

Each client updates its model using stochastic gradient descent:

$$w_i^{t+1} = w_i^t - \eta \nabla \mathcal{L}, \tag{4}$$

and sends the result to the server, which aggregates the updates:

$$w^{t+1} = \sum_{i=1}^{N} \frac{|D_i|}{|D|} w_i^{t+1}. \tag{5}$$

Through this contrastive alignment of model representations, MOON reduces local update drift and achieves improved convergence in federated settings with non-IID data.

IV. Results

This section evaluates the performance of MOON and baseline algorithms using two primary metrics: top-1 test accuracy and training loss over communication rounds. While the original MOON paper [1] reports results over 100 rounds on CIFAR datasets, this reproduction experiment was conducted over 40 rounds due to practical constraints. The training was carried out on Google Colab with an NVIDIA L4 GPU, which enabled full participation of 20 clients using the entire CIFAR-10 dataset [6].

Data heterogeneity was introduced using a Dirichlet distribution with concentration parameter $\beta = 0.5$, resulting in non-IID data across clients. All algorithms—FedAvg [2], FedProx [3], SCAFFOLD [4], and MOON—were evaluated under the same conditions. MOON was run with contrastive weight $\mu = 0.1$, while other hyperparameters were kept consistent across methods.

Figures 1 and 2 display the top-1 test accuracy and training loss over 40 communication rounds. All methods show steady improvement in accuracy and a general downward trend in training loss. MOON performs comparably to FedAvg and FedProx, while SCAFFOLD exhibits faster early convergence and reaches the highest accuracy within the evaluation window.

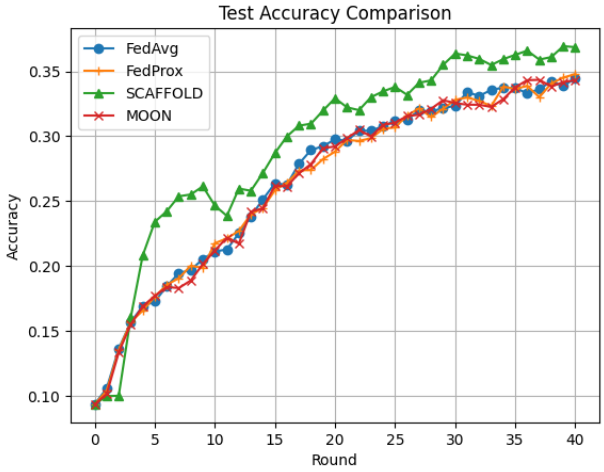

Fig. 1: Top-1 test accuracy over 40 communication rounds.

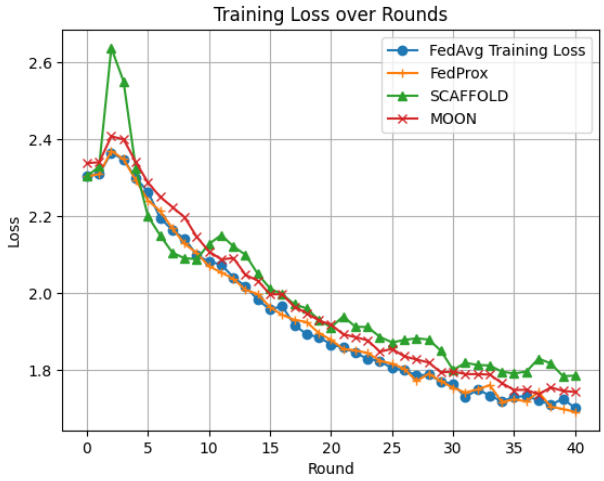

Fig. 2: Training loss over 40 communication rounds.

V. Discussion

While the original MOON paper reported clear gains in accuracy and convergence speed, our reproduction shows that MOON performs competitively but does not consistently outperform the baselines within the 40-round training window. SCAFFOLD achieves the fastest convergence and highest test accuracy, particularly in early rounds, likely due to its control variate mechanism that actively corrects for client drift. MOON, in contrast, shows steady progress but remains closely aligned with FedAvg and FedProx throughout training.

These trends suggest that MOON's contrastive loss may require a longer training horizon to fully demonstrate its benefit, as its alignment effect accumulates over successive rounds. Additionally, MOON's performance is known to be sensitive to the hyperparameter μ, which controls the weight of the contrastive regularisation term. In our experiments, the originally recommended value of $\mu = 5$ yielded worse performance. Further tests of $\mu = [0.01, 0.1, 1]$ showed that the most stable results came from $\mu = 0.1$. This highlights the importance of careful tuning in practice, and reinforces the idea that MOON's effectiveness may depend heavily on task-specific calibration and sufficient training rounds.

Despite these empirical limitations, MOON presents a novel contribution to federated learning by introducing a model-level contrastive regularisation strategy. Unlike methods that modify aggregation rules or require global coordination, MOON integrates its alignment mechanism directly into local training. This modular design allows it to be paired with various existing optimisation frameworks without altering the communication protocol.

In the context of distributed optimisation, MOON can be interpreted as implicitly enforcing a coherence constraint between local and global updates—similar to regularisation terms in a primal-dual setting. It offers a new way to mitigate heterogeneity-driven divergence without sacrificing privacy or adding communication overhead.

VI. Conclusion

This report examined the MOON algorithm, which introduces a model-level contrastive objective to mitigate the effects of data heterogeneity in federated learning. By aligning local, global, and previous model states, MOON provides a structured mechanism to reduce representation drift without altering the aggregation process.

In our reproduction study, MOON performed competitively with standard baselines but did not exhibit a consistent advantage within the 40-round training window. Its primary strengths lie in its modularity, ease of integration into local training, and its robustness to non-IID data distributions—particularly in image classification tasks, where traditional methods like FedAvg may struggle. However, its effectiveness is sensitive to hyperparameter selection and appears to require longer training to fully realise its benefits.

Looking ahead, MOON's contrastive learning principle could be adapted to fully decentralised, peer-to-peer learning settings, where no central server exists to coordinate updates. By enforcing local alignment between neighbouring agents' models, MOON-like mechanisms could enable coherent learning over graph-based networks. This opens promising directions for distributed optimisation in collaborative applications such as edge intelligence, autonomous vehicle fleets, or sensor networks operating under heterogeneous data conditions and sparse connectivity.

References

[1] Q. Li, B. He, and D. Song, "Model-contrastive federated learning," *arXiv preprint arXiv:2103.16257*, 2021.

[2] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," *arXiv preprint arXiv:1602.05629*, 2016.

[3] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proceedings of the 3rd Conference on Machine Learning and Systems (MLSys)*, 2020.

[4] S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for on-device federated learning," in *International Conference on Machine Learning (ICML)*, 2020.

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning (ICML)*, 2020.

[6] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.