MAFD: A Federated Distillation Approach with Multi-head Attention for Recommendation Task

Aming Wu, Young-Woo Kwon
Computer Science and Engineering, Kyungpook National University
wuaming@knu.ac.kr, ywkwon@knu.ac.kr

Abstract

Data isolation and privacy protection are becoming new challenges for deep learning. Federated learning (FL) has been shown to be a key example of privacy-preserving global training models derived from decentralized data in order to address such issues. This research proposes a new personalized federated knowledge distillation model based on a multi-head attention mechanism for a recommendation system. Compared to numerous benchmark models, our study demonstrates that our technique yields promising results.

Ⅰ. Introduction

Federated learning (FL) is a distributed machine learning framework first proposed by McMahan et al. [1]. It can conduct collaborative training without sharing private data and has achieved unprecedented success in data privacy.

With the higher complexity of the model of the recommendation system, more weight coefficients need to be exchanged by the federated learning, which brings severe challenges to the communication overhead of the edge devices. Han et al. [2] Proposed a general federated sequence recommendation model (DeepRec), which effectively avoids the risk of privacy disclosure. However, DeepRec did not consider different preferences and hidden microscopic behaviors. Xuan et al. [3] introduced knowledge distillation into the federated learning scenario and proposed a personalized FedCodl model, effectively solving the communication problem between model parameters and drinking devices.

However, in the recommendation system, since the highly heterogeneous data among users, the convergence speed and the accuracy of the global model are low in practical applications.

To address such limitations, this research introduces federated distillation based on a multi-head attention mechanism for a recommendation system, namely MAFD. Compared with the traditional federated learning model, we add the Wasserstein Distance (WD) and regular terms to the joint objective function to reduce the impact on the global model caused by the difference between the teacher and student network. Also, the improved multi-head attention mechanism is presented in the end process of federated distillation devices to enrich the feature encoding information.

Ⅱ. Proposed MAFD model

This section presents the detailed workflow that we proposed the personalized federated distillation strategy based on a multi-head attention mechanism.

(i) Assuming that the entire recommendation system has d devices, a deep learning recommendation model (student model) is initialized for each device D, and the model is trained using local data. It is worth noting that the local model uses the attention mechanism to encode the local user and item features, fuse the feature cross information to obtain the feature embedding expression, and train these expressions as the input of the local model. The attention mechanism can capture more implicit features, and the coding can reduce the risk of data leakage from local users.

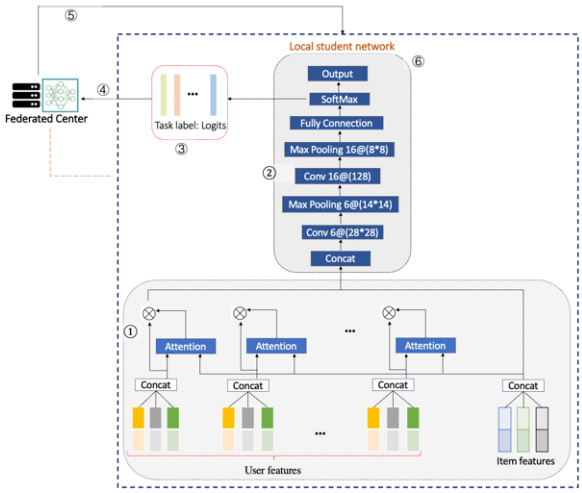

Figure 1: The MAFD model framework

(ii) The local devices upload the trained model parameters to the Federated Center. Here the parameters uploaded by the federated distillation method are

Logits vectors calculated by the last Softmax layer of the local student model instead of the model weight matrix, which can significantly reduce the number of uploaded parameters and relieve the communication pressure of the central equipment.

(iii) The Federated Center integrates the Logits vectors of all received devices into a new global Logits vector. For each device, the federated learning is used to build the teacher model of the device and distribute it to each device for updating.

(iv) The local equipment receives the teacher model and optimizes the joint loss function by combining the adaptive learning rate strategy to guide the training of the student model.

Experiments: The effectiveness of the MAFD model on the MovieLens 1M [4] is verified. The MovieLens 1M dataset contains 1500 users and user features, 200movies, as well as the label attribute information of movies. In the experiment, movies with more than 20 movie reviews and users with more than or equal to 10 movie reviews were selected as training samples. To compare the performance of different federated algorithms, we compared the basic MAFD combined with the convolutional neural network (CNN) proposed in this paper with two other federated models: (i) FD + CNN: Federated knowledge distillation algorithm combined with CNN [5]; (II) FDIN: Federated learning (FedAvg) combined with deep interest network (DIN) [6].

Evaluation: To measure the recommendation performance of the model, we first evaluated the model's accuracy, and then used mean absolute error (MAE) to analyze the model's performance.

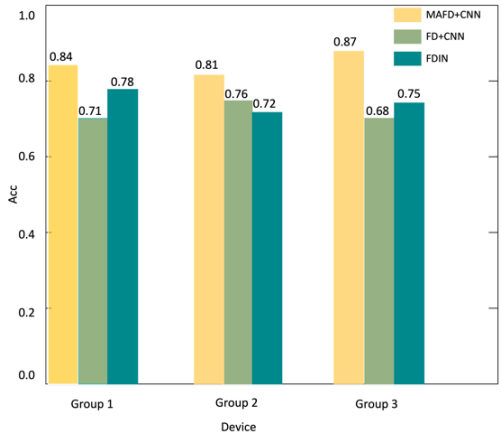

Figure 2: Different model's accuracy on MovieLens

The results of three different model frameworks on the same data set are presented in Figure 2, which clearly suggests that the accuracy of the MAFD model proposed is higher than that of other models in each group, the average accuracy is above 80%.

Except the RMSE indicator, in this research, we introduce the MAE. Table1 is observed that among the three models, the MAE value of the MAFD model is the smallest and below 0.2 in each group devices, it reveals that the proposed MAFD is promising performance.

Table1: Different models' MAE in three groups

Index and model		Group1	Group2	Group3	Global
MAE	MAFD+ CNN	**0.15**	**0.21**	**0.19**	**0.18**
	FD+ CNN	0.18	0.28	0.26	0.24
	FDIN	0.16	0.25	0.23	0.21

Ⅲ. Conclusion

In this paper, we presented a federated distillation method based on a multi-head attention mechanism to improve the model's performance. Through multiple rounds of experiments on the MovieLens 1M dataset, compared with other models, the MAFD model achieves an accuracy of more than 80% and significantly improves the convergence speed and running time. In future research, we will further test the robustness of the model in different data sets and try to adopt the combination of joint distillation and multi-task learning to formulate different strategies for different devices while compatible with multiple training tasks, to reduce the flow and model running time significantly.

Reference

[1] Koneč ný, J., McMahan, B., & Ramage, D. (2015). Federated optimization: Distributed optimization beyond thedatacenter. arXiv preprint arXiv:1511.03575.

[2] Han, J., Ma, Y., Mei, Q., & Liu, X. (2021, April). Deeprec: On-device deep learning for privacy-preserving sequential recommendation in mobile commerce. In Proceedings of the Web Conference 2021 (pp. 900-911).

[3] Ni, X., Shen, X., & Zhao, H. (2022). Federated optimization via knowledge codistillation. Expert Systems with Applications, 191, 116310.

[4] "MovieLens 1M Dataset" https://grouplens.org/datasets/movielens/1m.

[5] Seo, H., Park, J., Oh, S., Bennis, M., & Kim, S. L. (2020). Federated knowledge distillation. arXiv preprint arXiv:2011.02367.

[6] Zhou, G., Mou, N., Fan, Y., Pi, Q., Bian, W., Zhou, C., ... & Gai, K. (2019, July). Deep interest evolution network for click-through rate prediction. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 5941-5948).