JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

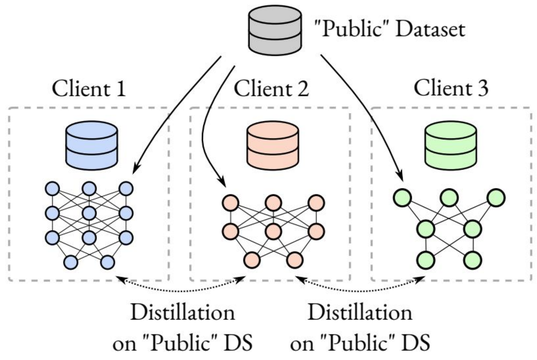
Decentralized Learning with Multi-Headed Distillation

Andrey Zhmoginov, Mark Sandler, Nolan Miller, Gus Kristiansen, Max Vladymyrov – Google Research

Google Research

Abstract

- A novel distillation-based fully decentralized learning technique that allows multiple agents with private data to learn from each other, without having to share their data, weights or weight updates.
- Our approach is communication efficient, utilizes an unlabeled public dataset and uses multiple auxiliary heads for each client, greatly improving training efficiency in the case of heterogeneous data.
- Individual models can preserve and enhance performance on their private tasks while also dramatically improving their performance on the global data distribution.
- We study the effects of data and model architecture heterogeneity as well as communication graph topology.
- We show that our agents can significantly improve their performance compared to learning in isolation and that in heterogeneous ensembles, small models benefit from distilling from large models and large models can aggregate more information and get higher accuracy than that achievable by small models.



Method

- Each agent is equipped with a main head learning a private task (personal task of a client) and multiple auxiliary heads distilling the knowledge from other clients.
- The *main head* is only trained on the client's private task.
- *Auxiliary heads* use the shared public dataset to distill the knowledge from other clients.
- Having multiple distillation targets, each auxiliary head distills embeddings from others and also distills predictions choosing the client with the most "confident" prediction (distilling to "soft" predictions).

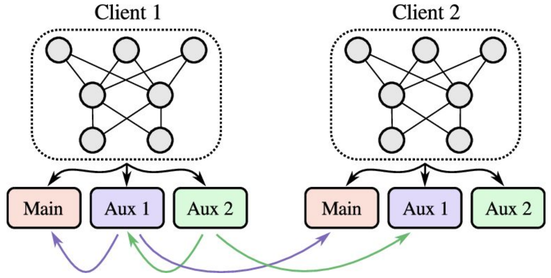
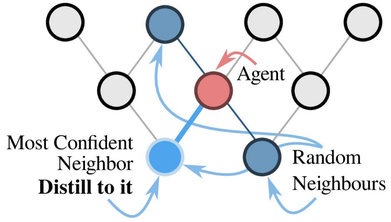


Figure. Distillation connections between the main and auxiliary heads of the clients. *Aux1* heads distill information from the *Main* model heads, *Aux2* heads distill information from *Aux1*, etc.

- We use the maximum logit value as the measure of *confidence*.

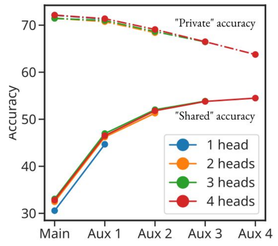


Experimental Setup

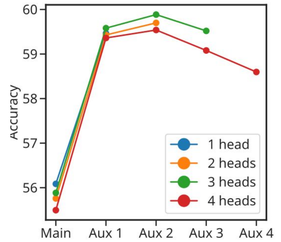
- ImageNet with ResNet34 and ResNet18 models
- 90% of data for *private tasks* and 10% as a *public dataset*
- Ensemble of 8 clients / IID and non-IID (each client uses samples from 250 labels out of 1000) private datasets.

Experimental Results

- IID: Aux heads *have the highest accuracy*, but adding too many hurts performance.
- non-IID: Adding aux heads *improves "private" main head accuracy* while also *increasing shared accuracy* on aux heads.



non-IID Dataset
(shared task != private)

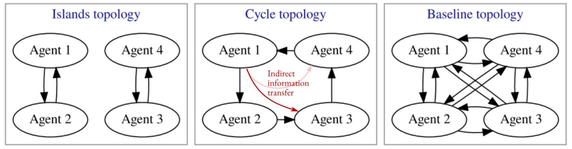


IID Dataset
(shared task=private)

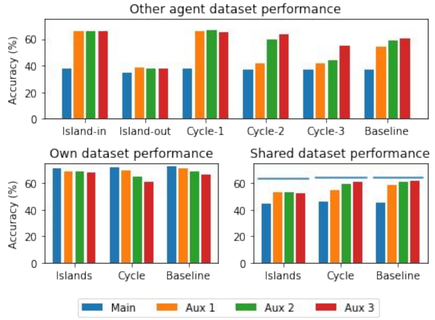
- With increased amount of public data: IID setup approaches the supervised model accuracy. In the non-IID setup, it is comparable to Federated Averaging (FA) with infrequent updates (gradients aggregated every u steps):

IID	*Accuracy*	non-IID	*Accuracy*
Separate	46.3%	Separate	25.1%
MHD (Ours)	68.6%	MHD (Ours)	63.4%
FA, $u = 200$	70.5%	FA, $u = 200$	68.0%
FA, $u = 1000$	69.1%	FA, $u = 1000$	65.7%
Supervised	68.9%	–	–

Communication Topology Effects



- Auxiliary heads enable transitive distillation (information transfer across indirectly connected clients) and additional heads make learning across additional degrees of separation more efficient.



Learning in Heterogeneous Systems

- The presence of a larger ResNet34 model improved the accuracy of smaller ResNet18 clients (66.2% to 66.7%).
- Step towards *training foundation models from weak learners* – witness that larger model trained with an ensemble of small ones can accumulate knowledge and reach higher accuracy than that achievable with small models only (68.6% vs 67.7%).

arXiv: https://arxiv.org/abs/2211.15774

Contact: azhmogin@google.com