ZeroFL: Efficient On-Device Training for Federated Learning with Local Sparsity

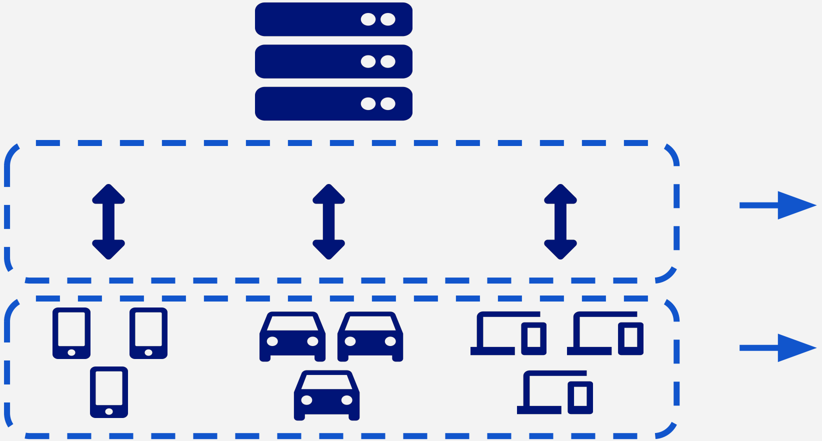





Xinchi Qiu, Javier Fernandez-Marques
Pedro PB Gusmão, Yan Gao, Titouan Parcollet, Nicholas D. Lane

Federated Learning

- FL is a form of distributed ML
- FL *clients* (i.e. compute nodes) are embedded devices
- FL clients collaboratively learn a single *global model*
- Data stays in the client



Model exchange every round, large communication cost

Training on resource constrained device

Reducing on-device training costs

- FL training is costly in terms of **compute** and **communication**
- The energy footprint of FL can be higher than centralised training (Qiu et al. 2021)*
- Multiple ways to address challenges: quantization, pruning, distillation, ...
- ZeroFL:
  - reduces on-device compute costs thanks to highly sparse OPs
  - reduces uplink communication with client-specific masking

**A First Look into the Carbon footprint of Federated Learning:* https://arxiv.org/abs/2102.07627

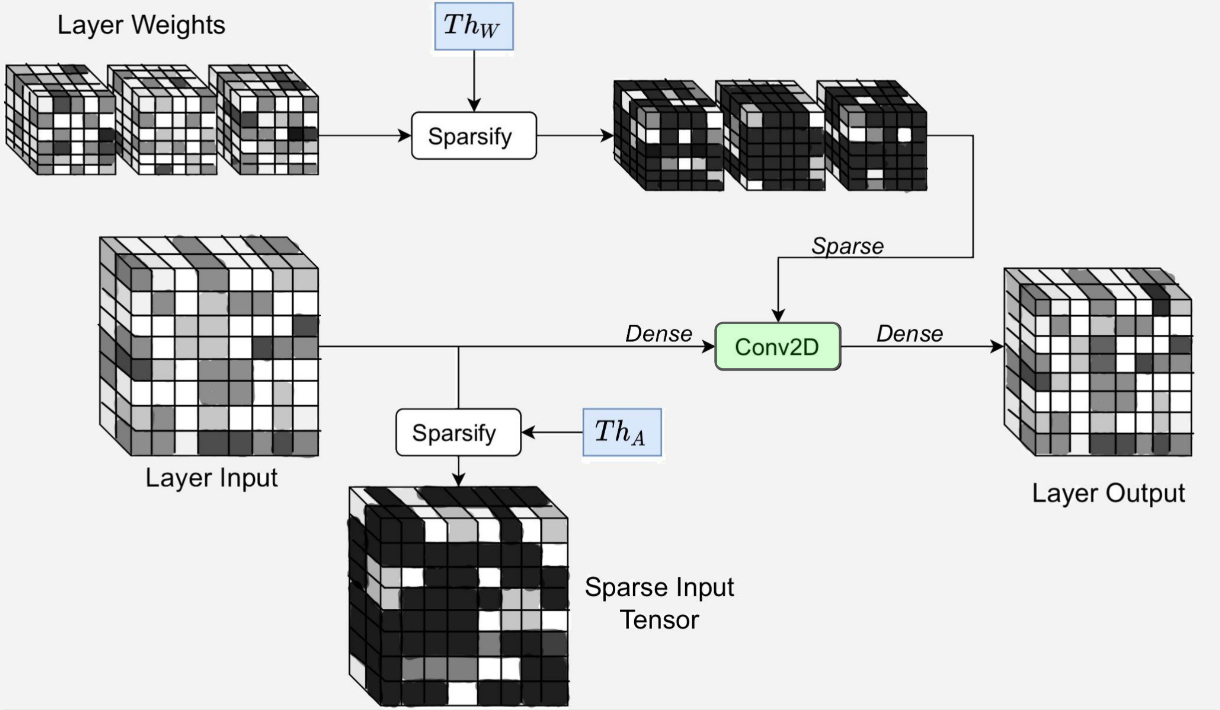
Reducing on-device training costs

- FL training is costly in terms of **compute** and **communication**
- The energy footprint of FL can be higher than centralised training (Qiu et al. 2021)*
- Multiple ways to address challenges:
  - Having smaller models – limits learning
  - Compressing model updates (Konečný et al. 2017)
  - Learning by distilling (FedGKT - He et al. 2020)
  - Pruning model based on compute capabilities of client (FederatedDropout - Caldas et al. 2018 , FjORD - Horvath et al. 2021)
- ZeroFL:
  - reduces on-device compute costs thanks to highly sparse OPs
  - reduces uplink communication with client-specific masking

**A First Look into the Carbon footprint of Federated Learning:* https://arxiv.org/abs/2102.07627

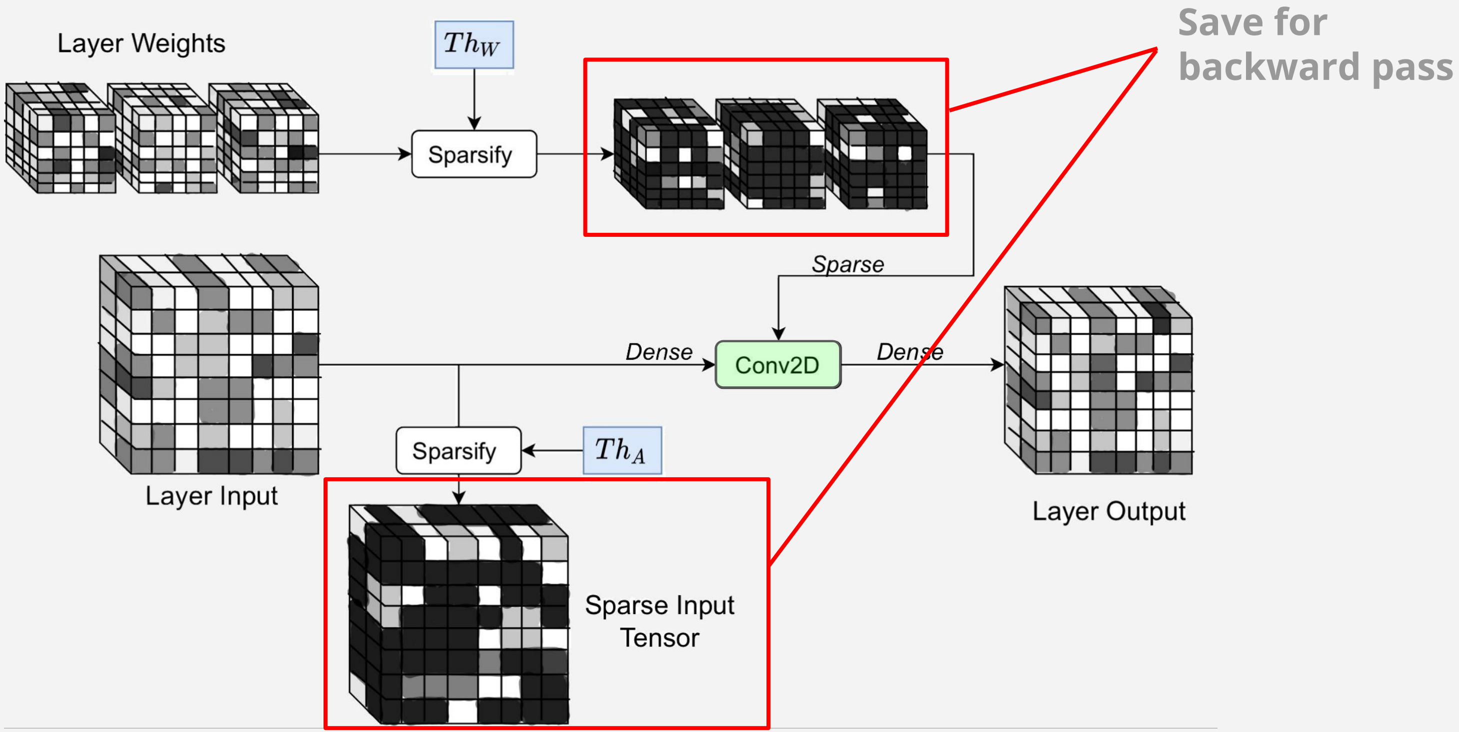
Sparse on-device training for FL

- We borrow inspiration from SWAT (Raihan & Aamodt, 2020)



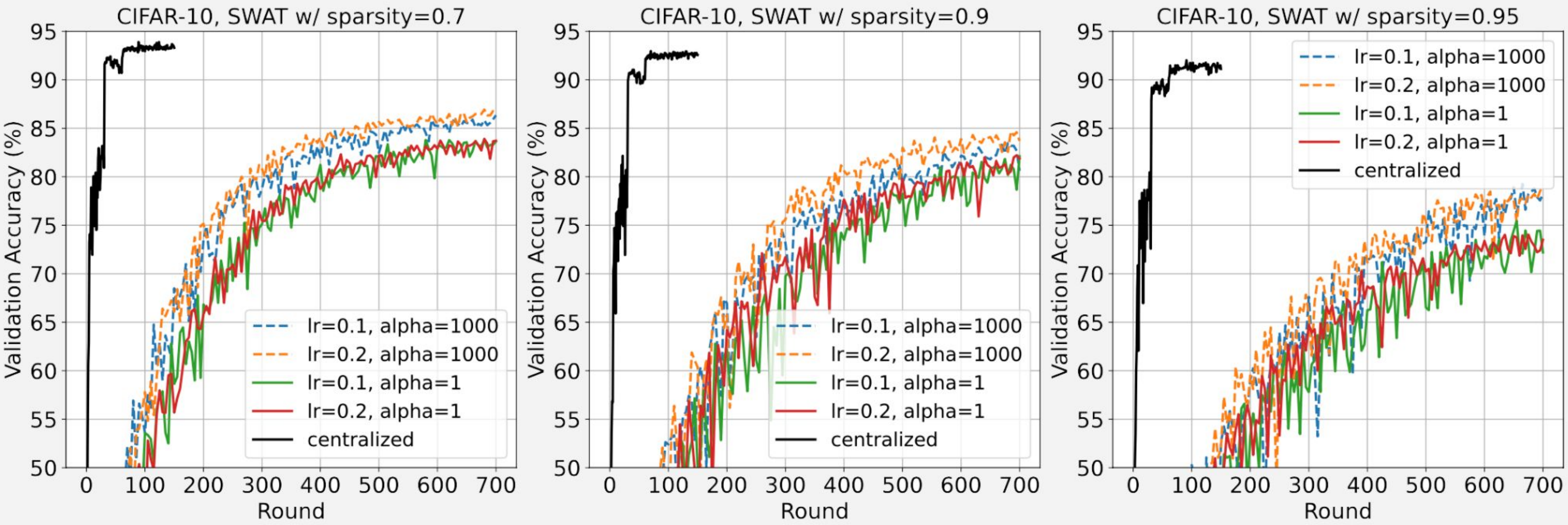
Sparse on-device training for FL

- We borrow inspiration from SWAT (Raihan & Aamodt, 2020)



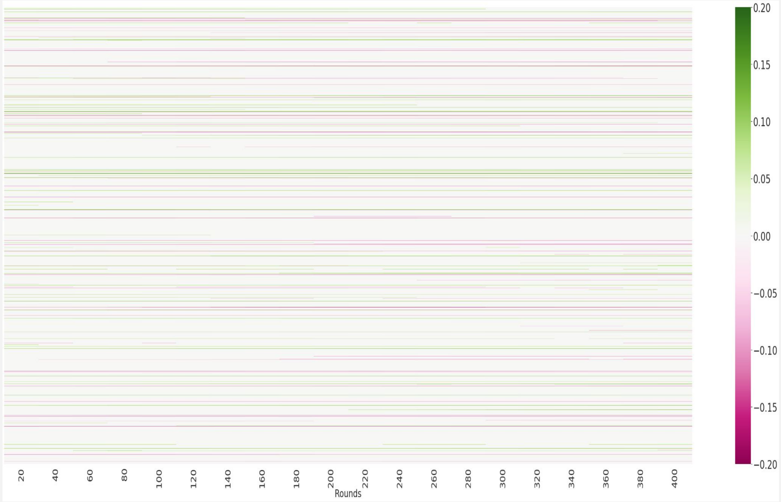
Sparse on-device training for FL

- Adapt SWAT to FL by to treat each local training as 1 centralised training
- Unlike centralised training, FL with sparse on-device training degrades rapidly



Improving sparse FL on-device training

- What needs to be investigated to make sparse training work better in FL?
    - Only top-k weights are used in forward propagation in evaluation
    - Non-zero weights remain at constant locations throughout the training process; sparsified weights tend to be the same
    - We not only save in compute but also communication



Only communicate top-k weights for aggregation; k=(1-sp)+ mask ratio

Results

- Datasets we use: CIFAR10, Speech Commands, FEMNIST
- Summary of results:
  - Generally mask ratio 0.1 or 0.2 perform better than 0
  - Trade-off between communication and performance
- Potential expansion directions
  - Structure sparsity: block masking etc.
  - Different masking method

	Sparsity Level	SWAT Full Model	ZeroFL (m=0.2)	File Size (MB)	Comms Save
CIFAR-10	90%	80.62%	81.04%	27.3	1.6x
	95%	74.00%	75.54%	23.0	1.9x
Speech Commands	90%	82.81%	84.90%	27.3	1.6x
	95%	81.12%	82.02%	23.0	1.9x
FEMNIST	95%	83.34%	83.78%	4.4	5.2x

ICLR

UNIVERSITY OF OXFORD

AVIGNON UNIVERSITÉ

Thanks!






Xinchi Qiu, Javier Fernandez-Marques
Pedro PB Gusmão, Yan Gao, Titouan Parcollet, Nicholas D. Lane