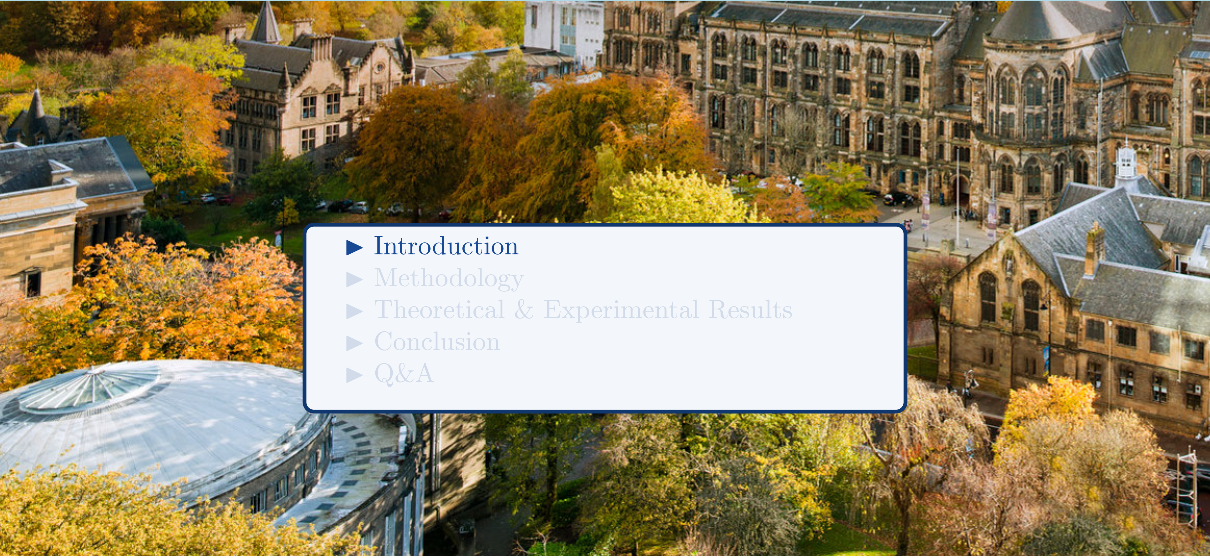
FedDIP: Federated Learning with Extreme Dynamic Pruning and Incremental Regularization

Qianyu Long[1] Christos Anagnostopoulos[1] Shameem Puthiya Parambath[1] Daning Bi[2]

[1]University of Glasgow, UK [2]Hunan University, CN

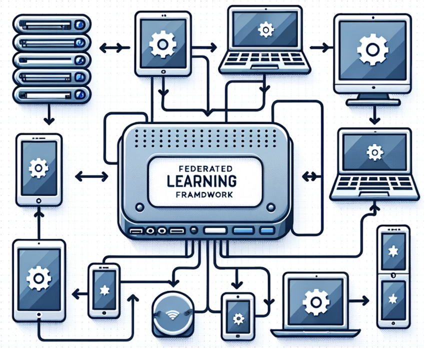
Contents



Knowledge & Data Engineering Systems

Applications

- **Healthcare**
- **Mobile Devices**
- **Autonomous Vehicles**



Challenges

- Computation
- Communication
- Model Inference

Efficient Federated Learning with **Compression**!

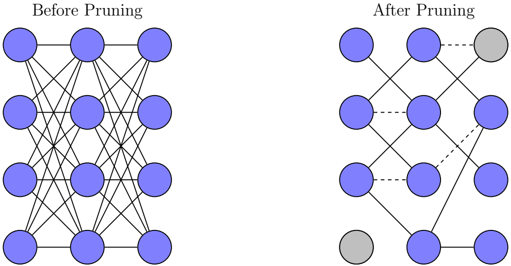
(SOTAs) Categorized by model training

- *Dense* to *Dense*: Only *gradients*
- *Dense* to *Sparse*: Excessive memory
- *Sparse* to *Sparse*: Insufficient sparsity level

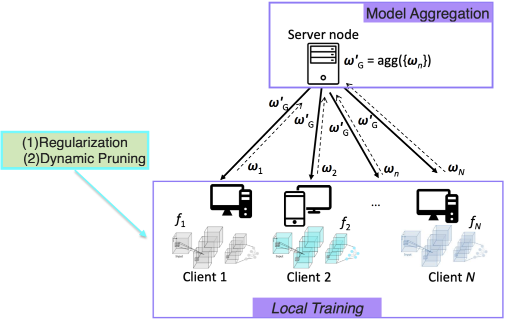
Highlights

- Dynamic pruning with error feedback and regularization in FL
- *Sparse* to Extreme *Sparse* ($s_p > 0.8$ for i.i.d and non-i.i.d)
- Convergence analysis with theoretical support

Model Pruning



Contents



Dynamic Pruning with error Feedback (**DPF**[a]):

$$\boldsymbol{\omega}_{t+1} = \boldsymbol{\omega}_t - \eta_t \nabla f(\boldsymbol{\omega}_t \odot \mathbf{m}_t) \quad (1)$$

$$= \boldsymbol{\omega}_t - \eta_t \nabla f(\boldsymbol{\omega}_t + \mathbf{e}_t) \quad (2)$$

Inspired by **GReg**[b], our incremental regularization:

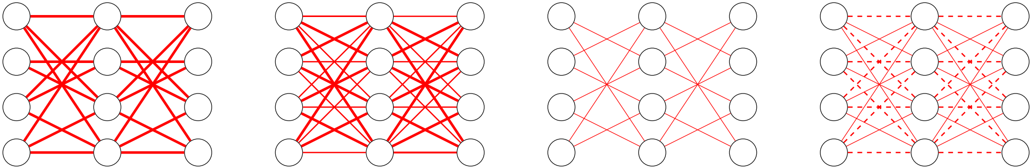
$$\lambda_t = \begin{cases} 0 & \text{if } 0 \le t < \frac{T}{Q} \\ \vdots & \vdots \\ \frac{\lambda_{\max}(Q-1)}{Q} & \text{if } \frac{(Q-1)T}{Q} \le t \le T \end{cases} \quad (3)$$

[a] Tao Lin et al. (2019). "Dynamic Model Pruning with Feedback". In: *International Conference on Learning Representations (ICLR)*.

[b] Huan Wang et al. (2021). "Neural Pruning via Growing Regularization". In: *International Conference on Learning Representations (ICLR)*.

FedDIP Diagram

Methodology 2



Sparse Model | After Regularization | After Pruning | After Error Feedback

Benefits:

- DPF in FL:
  - Reduce downloading cost
  - Avoid long time post-pruning fine-tune
  - Suitable for high sparsity pruning
- IR with DPF:
  - Consistent weight importance scoring
  - Hessian Information for accurate pruning

Contents

Theorem (FedDIP Convergence)

Consider the general assumption, and let $\eta_t = \frac{1}{tL}$, $L > 0$. *Then, the convergence rate of the FedDIP process is bounded by:*

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\bar{\boldsymbol{\omega}}'^{(t)})\|^2 \leq 2L\mathbb{E}(f(\boldsymbol{\omega}_1) - f^*) + \boxed{2L\sum_{t=1}^{T}[\mu\mathbb{E}[\sqrt{\delta_{t+1}}\|\bar{\boldsymbol{\omega}}^{t+1}\|]]} + \frac{\pi^2}{3L^2}\chi, \quad (4)$$

where $f(\boldsymbol{\omega}_1)$ *and* f^* *stand for the initial loss and the final convergent stable loss, with* $\chi = \frac{(\gamma-1)L^2+L}{2K}\sum_{n=1}^{N}\rho_n\sigma_n^2 + \frac{(\gamma-1)\gamma E_l^2 L^2 G^2}{2}$, *and* γ *is the data dispersion degree.*

Datasets & Models: Evaluate *LeNet-5* ($s_p = 0.9$) on Fashion-MNIST, *AlexNet* ($s_p = 0.9$) on CIFAR10, and *ResNet-18* ($s_p = 0.8$) on CIFAR100.

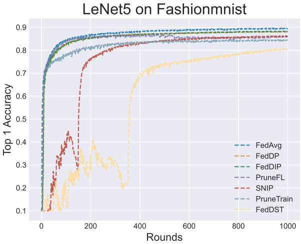


Figure: LeNet-5: Accuracy vs Training Rounds

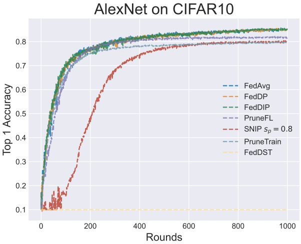


Figure: AlexNet: Accuracy vs Training Rounds

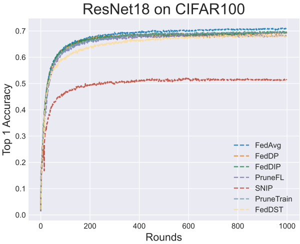


Figure: ResNet-18: Accuracy vs Training Rounds

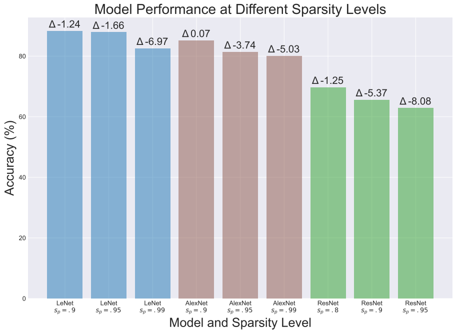


Figure: Performance under Extreme Sparsities

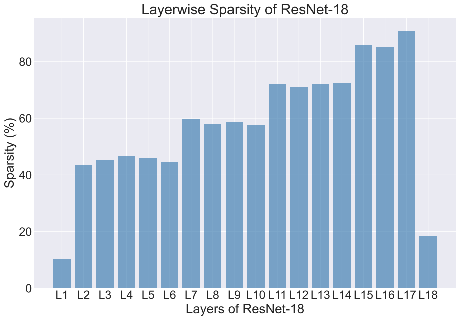
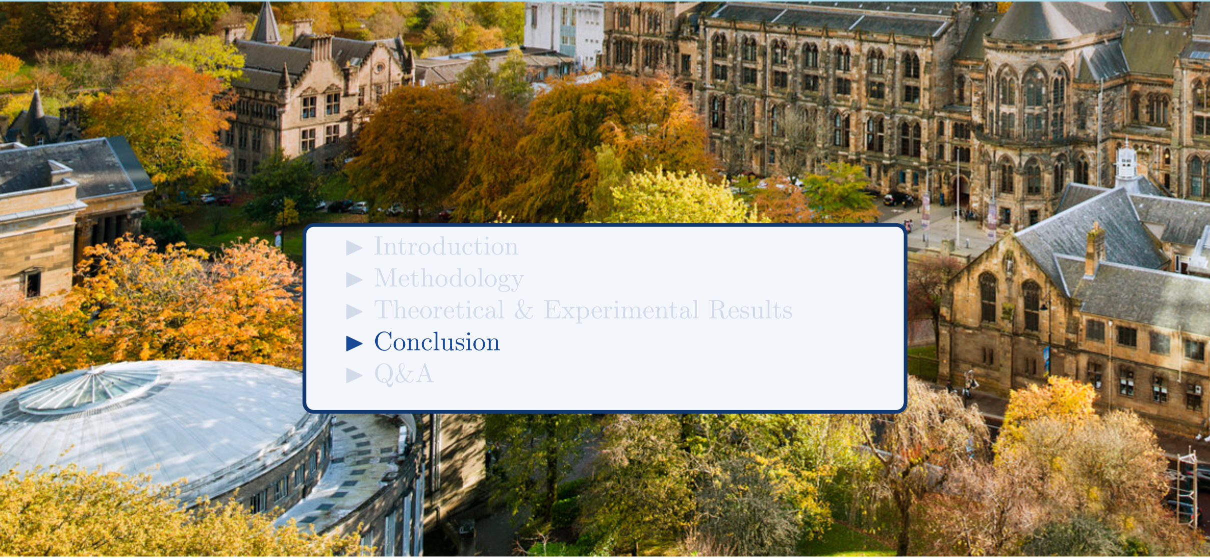


Figure: Sparsity Distribution of ResNet18

Contents



Conclusion and Future Work

Impact: Because of extreme pruning in FL, FedDIP contributes as follows.

- Faster Training
- Less Memory Required
- Less Downloading Cost

Future Work

- Mask Personalization
- Decentralized FL

Contents

Q&A

Thank you!