Enabling On-Device Training of Speech Recognition Models with Federated Dropout

Paper#: 2057

Dhruv Guliani (dguliani@google.com), Lillian Zhou, Changwan Ryu, Tien-Ju Yang, Harry Zhang, Yonghui Xiao, Françoise Beaufays, Giovanni Motta

Google Research

icassp 2022 Singapore

Summary of Contributions

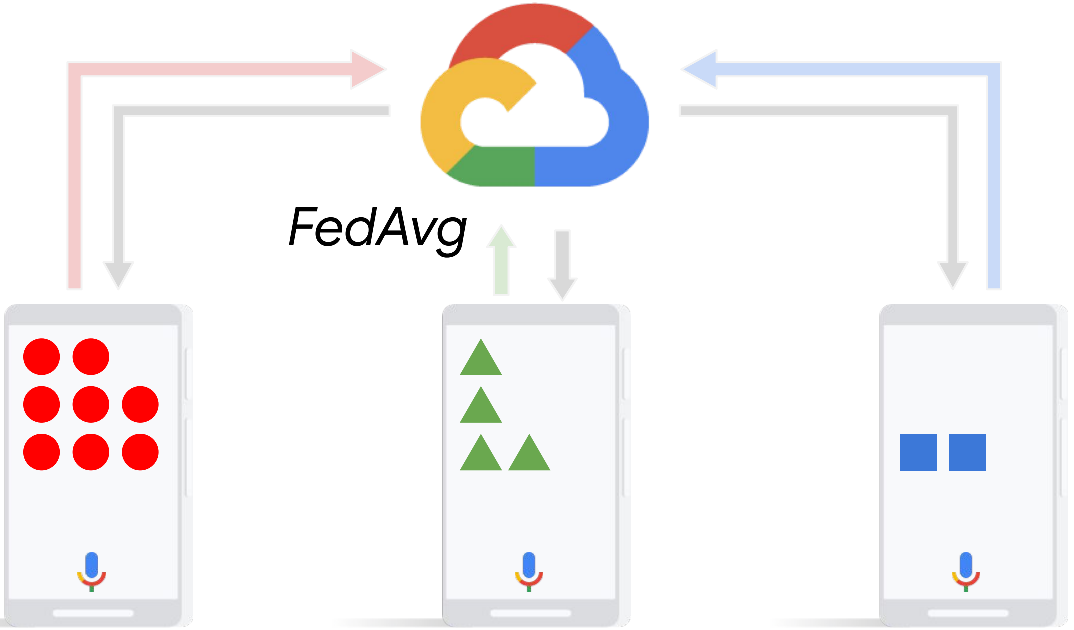
- Propose using **federated dropout (FD)** to reduce the size of client models while training a full-size model server side
- Show that FD can be successfully applied to ASR to provides a quality/cost trade-off
- Extend the technique to Google-scale workloads and show that the trade-off still applies
- Use per-layer varying FD rates to improve quality while keeping cost constant
- Show that FD effectively trains high quality **sub-models** within the full-size model, enabling the size to be reduced for on-device inference

Intro

Federated Learning and Federated Dropout

Google Research

Federated Learning



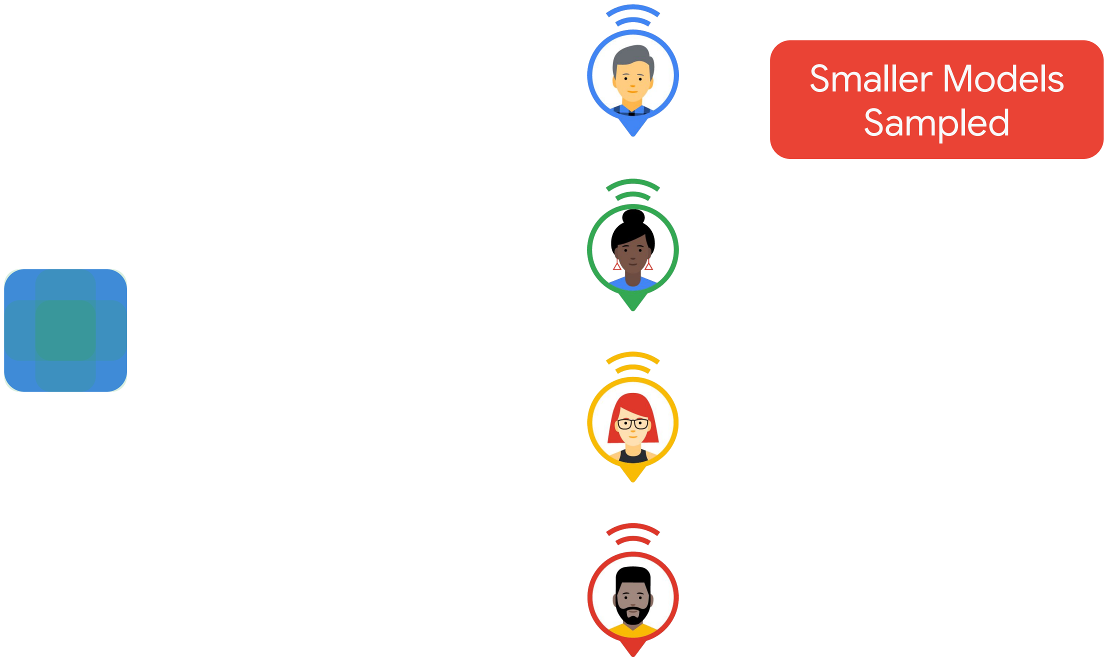
Repeat

Incorporates **Privacy** into model training through data separation, differential privacy, secure aggregation, etc.

Eliminates need for data collection on central servers

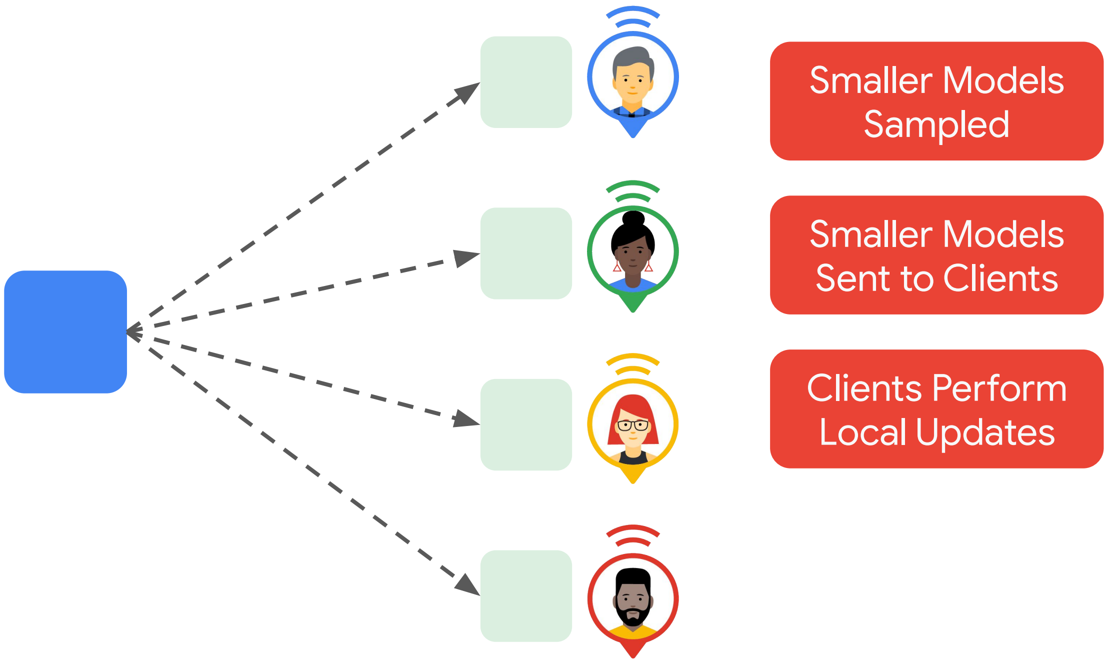
Google Research

Federated Dropout



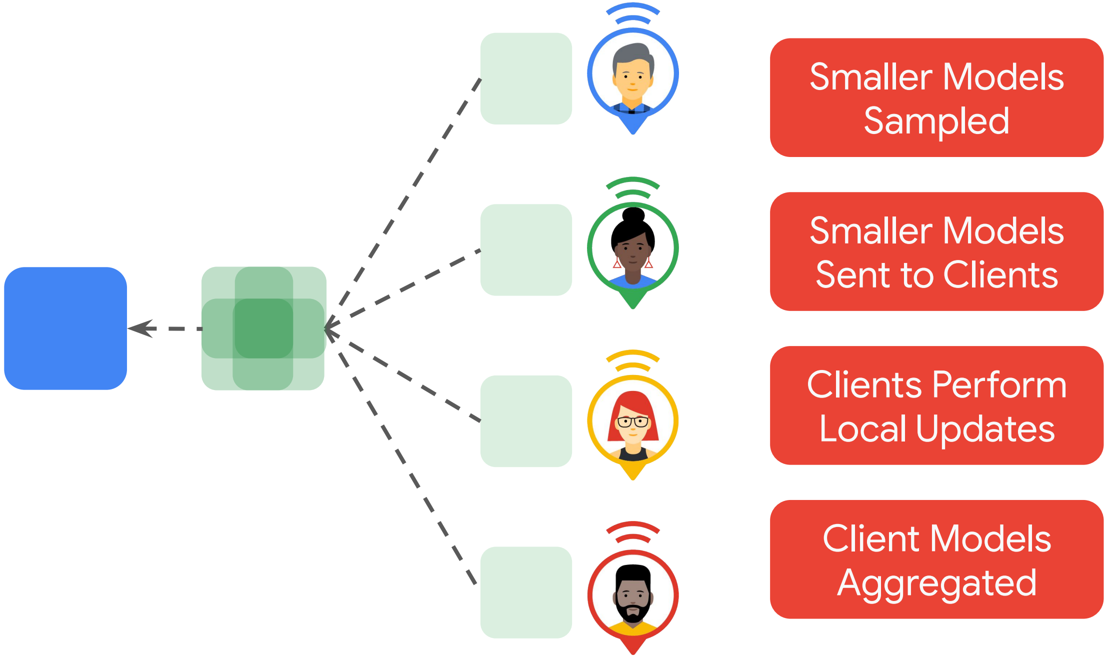
[12]

Federated Dropout



[12]

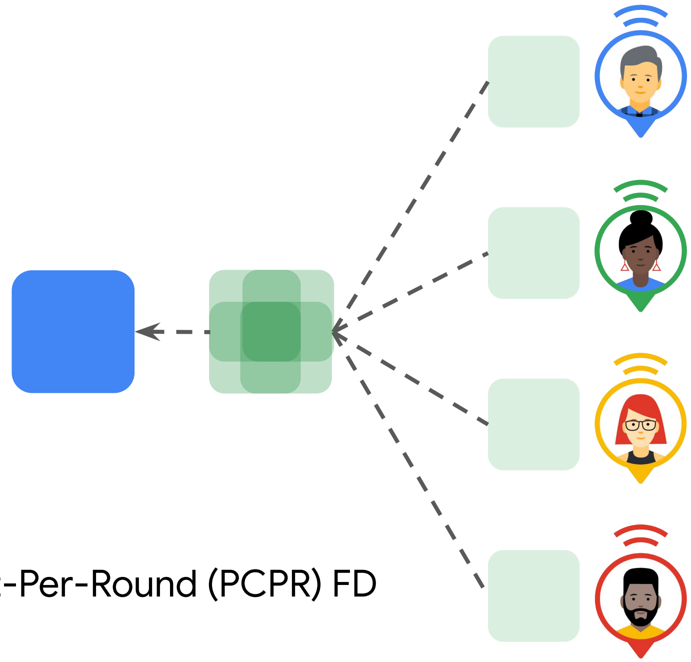
Federated Dropout



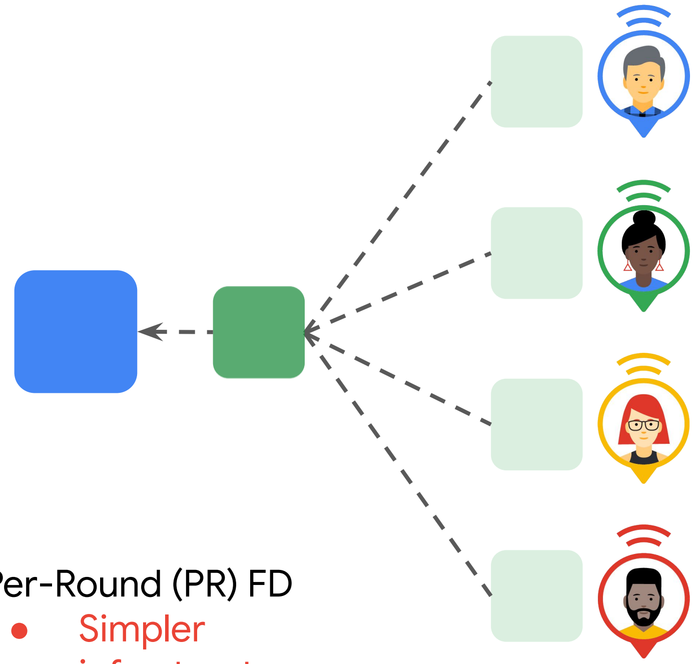
Reduces both **Communication** and **On-Device Computation** Cost.

[12]

Federated Dropout Flavours



Per-Client-Per-Round (PCPR) FD



Per-Round (PR) FD

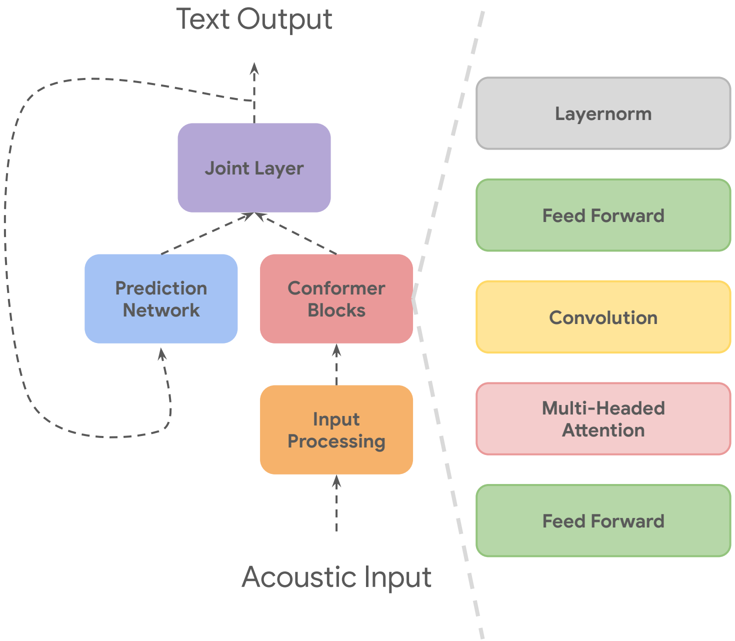
- Simpler infrastructure requirements

Google Research

Methodology

Model and Data

Models and Datasets



Non Streaming Conformer [2]

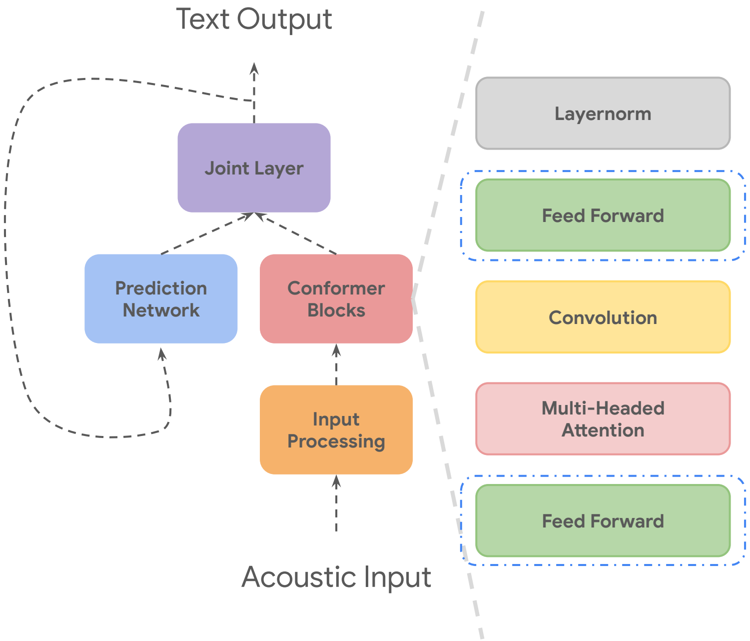
- 119M Parameters
- Trained on speaker-split Librispeech from scratch

Streaming Conformer [3]

- 137M Parameters
- Trained on Google scale multi-domain (MD:374k hours) data centrally and then trained using FL on of 26k hours of medium-form (MF) data
- Domain Adaptation task

Google Research

Models and Datasets



Non Streaming Conformer [2]

- Feedforward layers contain 60% of all model parameters

Streaming Conformer [3]

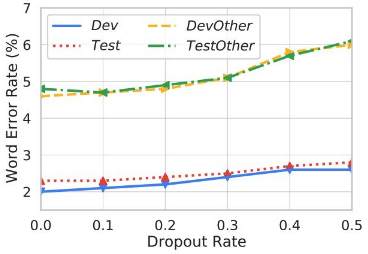
- Feedforward layers contain 55% of all model parameters

Google Research

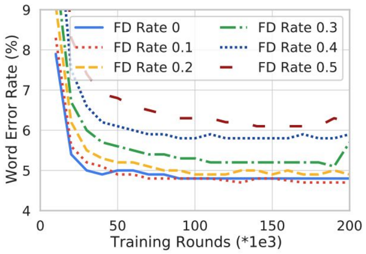
Experiments

Results

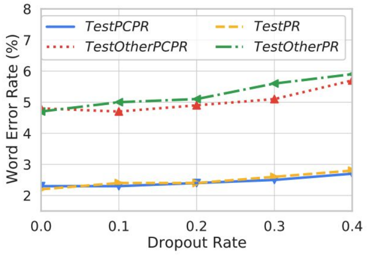
Non-Streaming Conformer Results



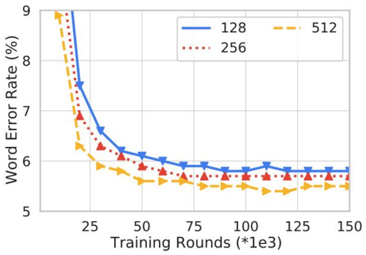
(a) Dropout Rate vs. WER



(b) Convergence Time



(c) PCPR vs. PR



(d) Clients Per Round

Takeaways:

a) FD provides a quality/cost(model size) trade-off

b) Higher FD rates usually converge slower

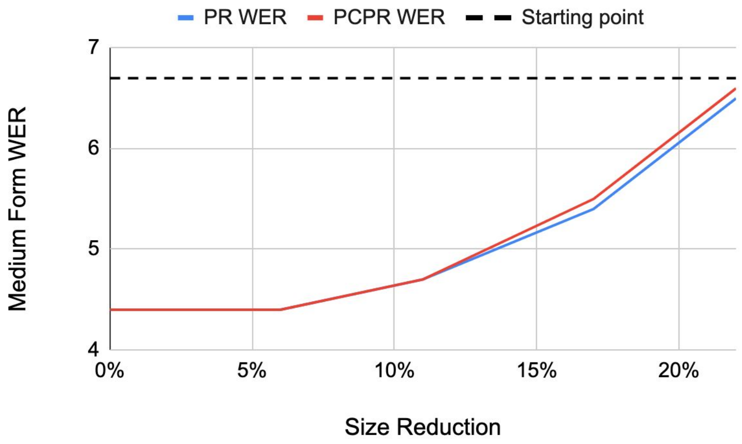
c) PR is slightly worse than PCPR, but usable if eng. resources limited

d) Higher report goals could improve convergence speed and quality

Google Research

Detailed analysis in paper

Streaming Conformer Results

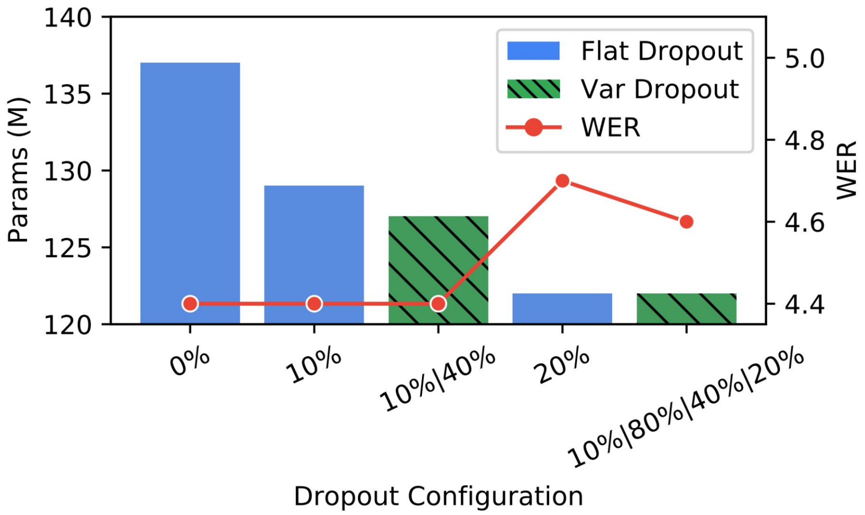


Takeaways:

1. FD scales to larger datasets and the domain adaptation task
2. PR remains only slightly worse than PCPR at higher FD rates

Detailed analysis in paper

Per-Layer Dropout Results

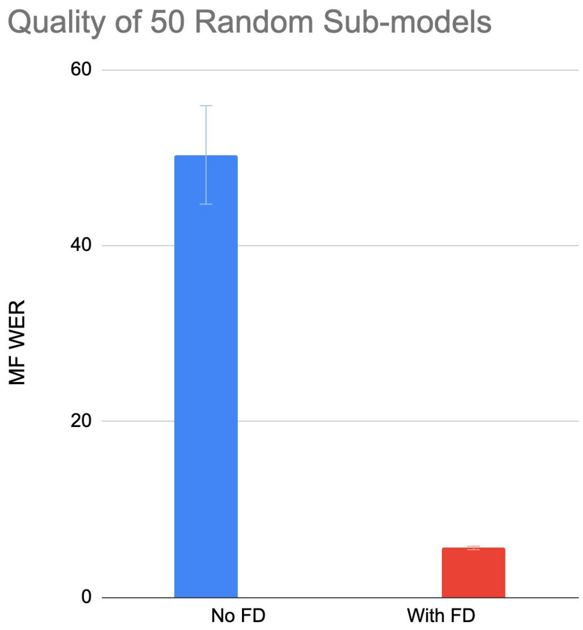


Experimented with varying FD rates per layer (chosen according to estimates of layer importance [22])

Takeaways:

1. Quality/cost trade-off can be improved.
2. New search space.

High-Quality Sub-models



Sampled 50 submodels with the same method as 50% FD from 2 experiments: one trained under FL with 50% FD and one without.

Takeaways:

1. FD improves the quality of sub-models within the larger model
2. Can deploy the same model to devices with various compute capabilities

Google Research

Conclusions

- Federated dropout is a promising technique to reduce the cost of training ASR models under FL and provides a tangible cost/quality trade-off
- Federated dropout scales to large, real-world workloads
- Varying per-layer dropout can yield more performant or lower cost configurations of FD
- FD causes capable sub-models to form within larger models, opening up possibilities to downsample models for inference

Select References

[1] Conformer: Convolution-augmented Transformer for Speech Recognition

[2] A Better and Faster End-to-End Model For Streaming ASR

[4] Training Speech Recognition Models with Federated Learning: A Cost/Quality Framework

[12] Expanding the Reach of Federated Learning by Reducing Client Resource Requirements

[22] Are All Layers Created Equal?

Thank You!

Team