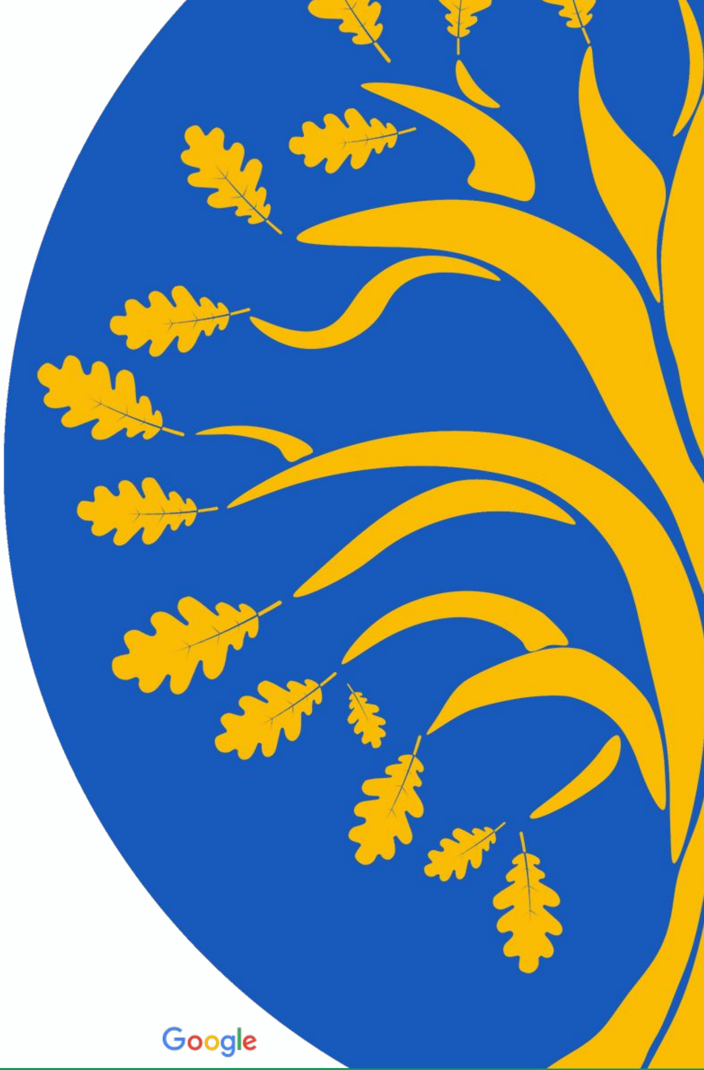
Exploring Heterogeneous Characteristics of Layers in ASR Models for More Efficient Training

Authors:

Lillian Zhou*, Dhruv Guliani*, Andreas Kabel,
Giovanni Motta, Françoise Beaufays

Presented at:

IEEE ICASSP 2022



* Equal contribution

Google

Motivation

- Federated Learning (FL): train models on device, aggregate to central model [17].
- Automatic Speech Recognition (ASR) models are costly to train.
- Must reduce transport and memory costs.

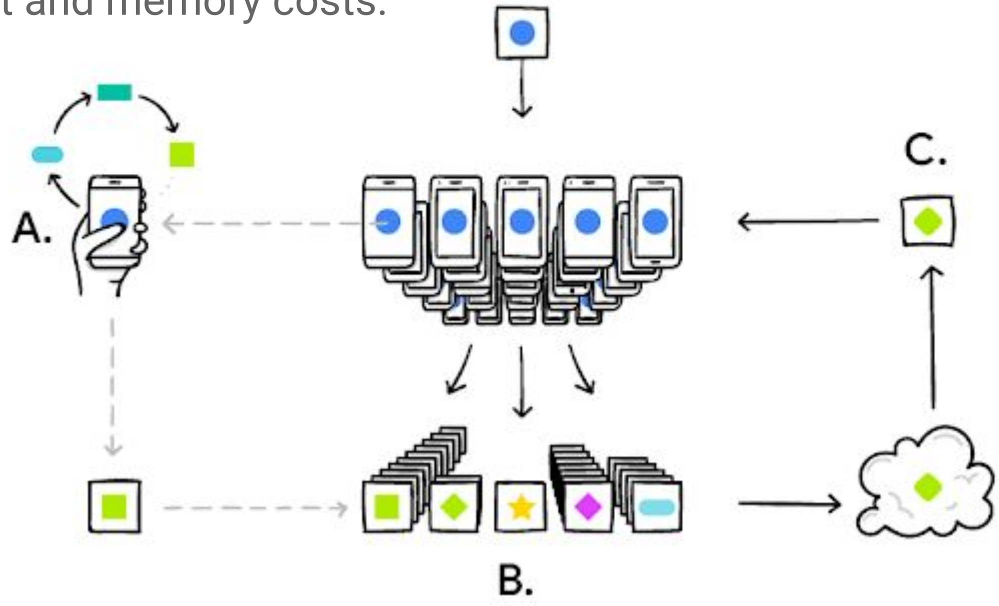


Image source: https://ai.googleblog.com/2017/04/federated-learning-collaborative.html.

Motivation

- In a neural model, certain layers are more important than others [11].
    - Does this apply to Conformer models, the SOTA for ASR?
    - Can we determine which layers are more important (*critical*), and which are less important (*ambient*)?
- If so, we could:
    - Only train most important layers.
    - Target compression techniques to least important layers.
        - Federated Dropout (FD)→ drop more in unimportant layers.
        - Transport compression → allocate bit budget by each layer's importance.

Do layers in ASR models also vary in importance?
Can we reliably rank layers by importance?

Experiment 1

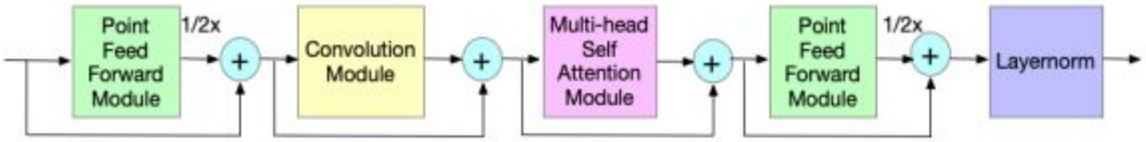
- Tested stability of these properties and variance across model sizes.
- Data and model details:
  - Librispeech corpus [20].
  - Three different model sizes of non-streaming Conformer [3].

Model	Conf Params	Conf Layers	Total Params
ConformerS	8.1M	16×0.5M	10.3M
ConformerM	25.4M	16×1.6M	30.7M
ConformerL	107.5M	17×6.3M	118.6M

Experiment 2

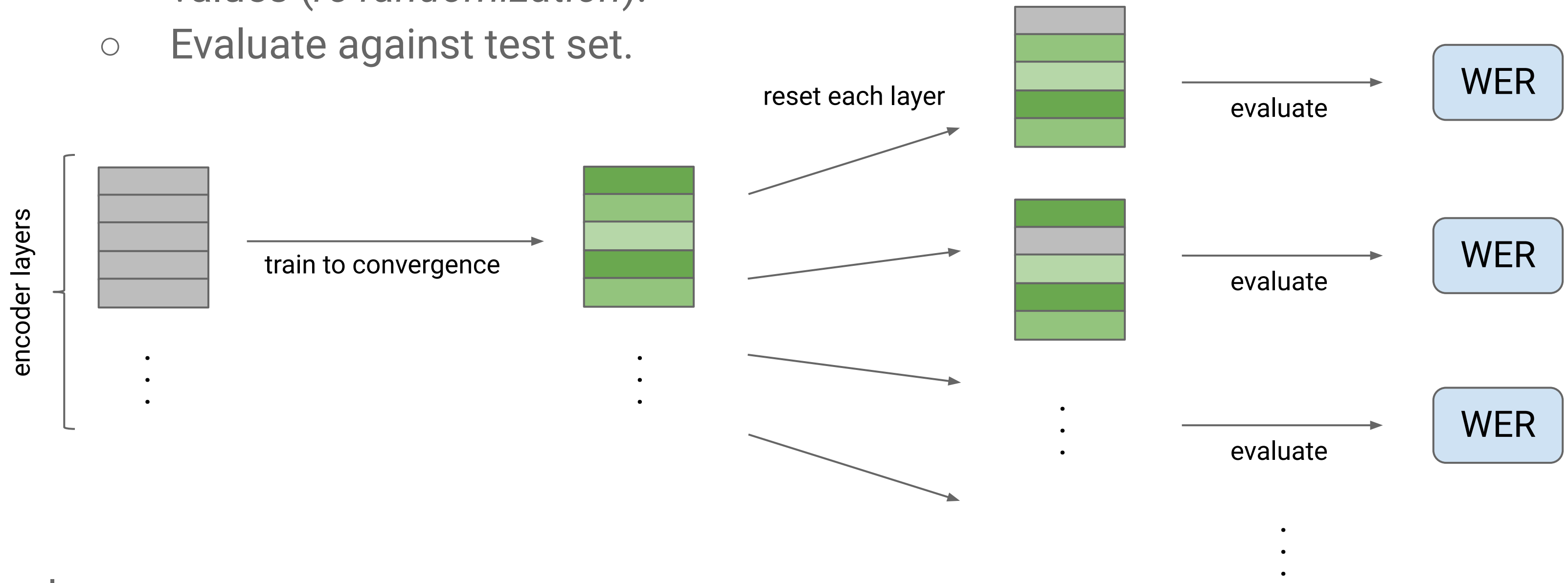
- Tested application in practical dataset on state-of-the-art model.
- Data and model details:
  - Multi-domain dataset (MD), with and without Short-form domain (SF) held out. [21]
  - Streaming Conformer model. [4]

Dataset	Hours
Multi-domain (MD)	400k
Short-form domain (SF)	27k
Short-form held out (MD-SF)	373k



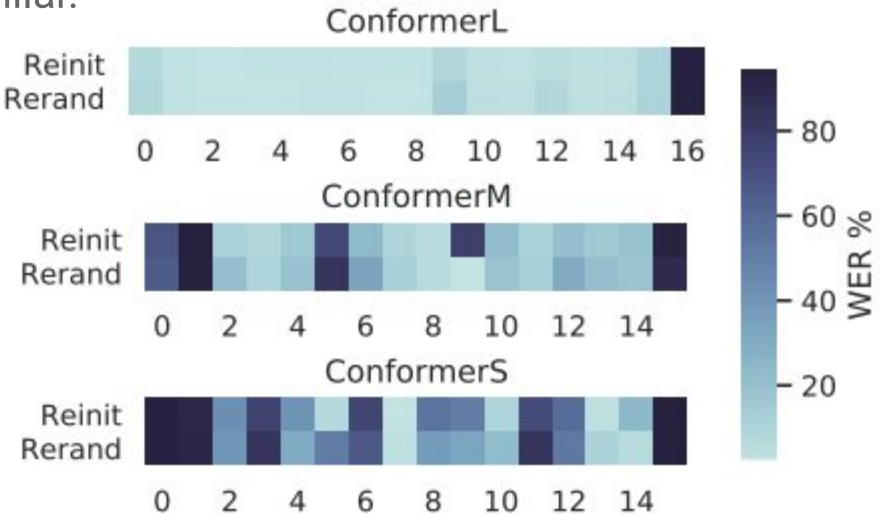
Methodology

- Train models to convergence.
- For each layer in the encoder:
  - Reset the layer weights to initial values (*re-initialization*) or random values (*re-randomization*).
  - Evaluate against test set.



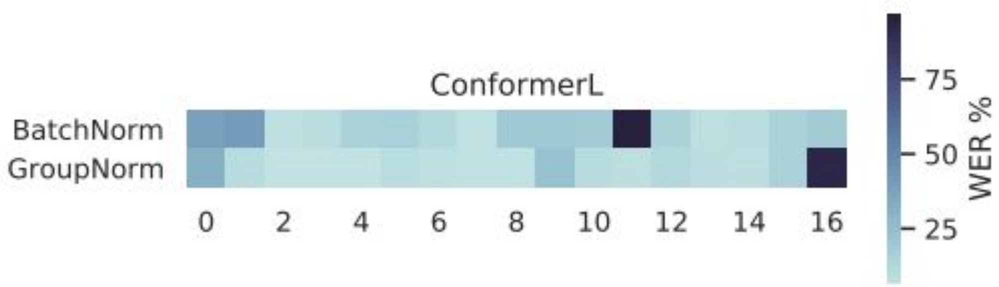
Model Size

- Columns show WER when each layer is reset.
  - Certain layers can be reset without penalty: "ambient layers".
  - Others have catastrophic impact: "critical layers".
- The larger the model, the more robust to having entire layers reset.
- Re-initialization vs re-randomization similar.



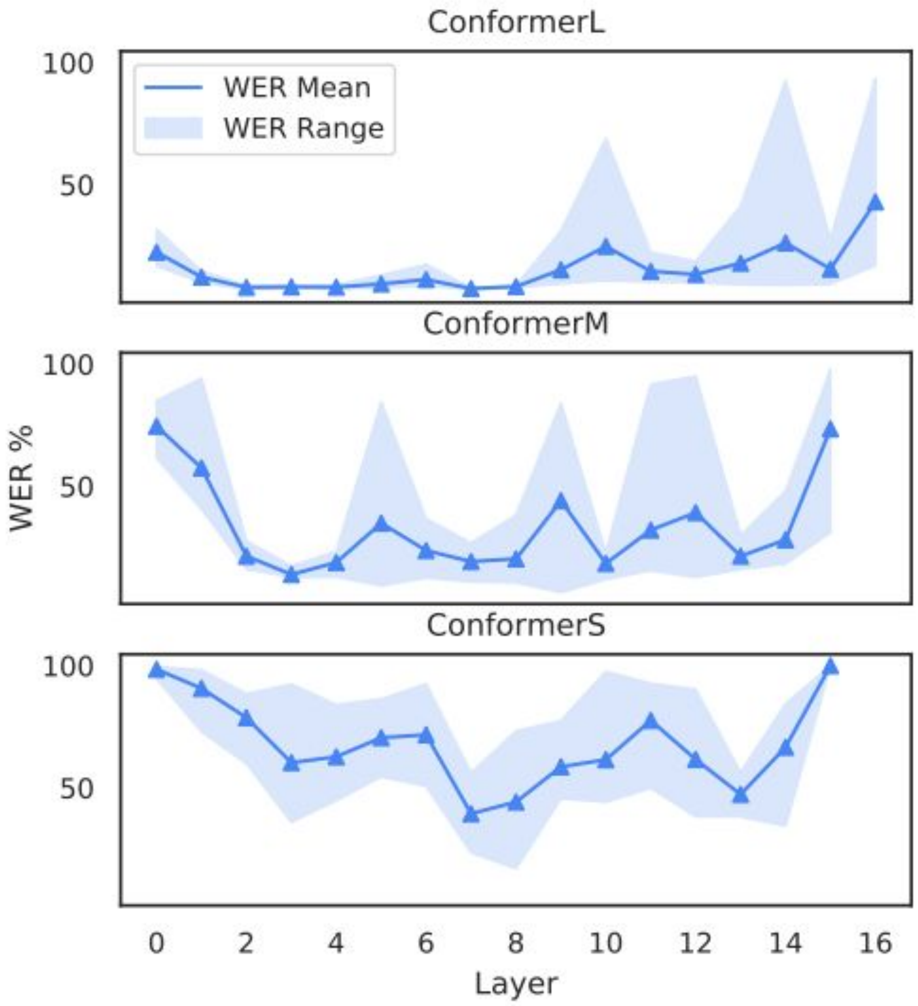
Batch Normalization vs Group Normalization

- Batch Normalization interferes with formation of ambient layers, so was left out of experiments in prior work [11].
- We find that Group Normalization [18] yields ambient layers.
  - Also ideal for Federated Learning [23].



Stability

- Reran the same experiment 5 times, including training from scratch.
- Larger models are more stable across runs.



Google

Takeaways

- We can rank layers of ASR model by importance.
- Group Normalization can be used.
- "Ambient" layers can be reset after training without much consequence.
  - Position of ambient layers is somewhat stable.
  - Larger models yield more ambient layers that are more stable.

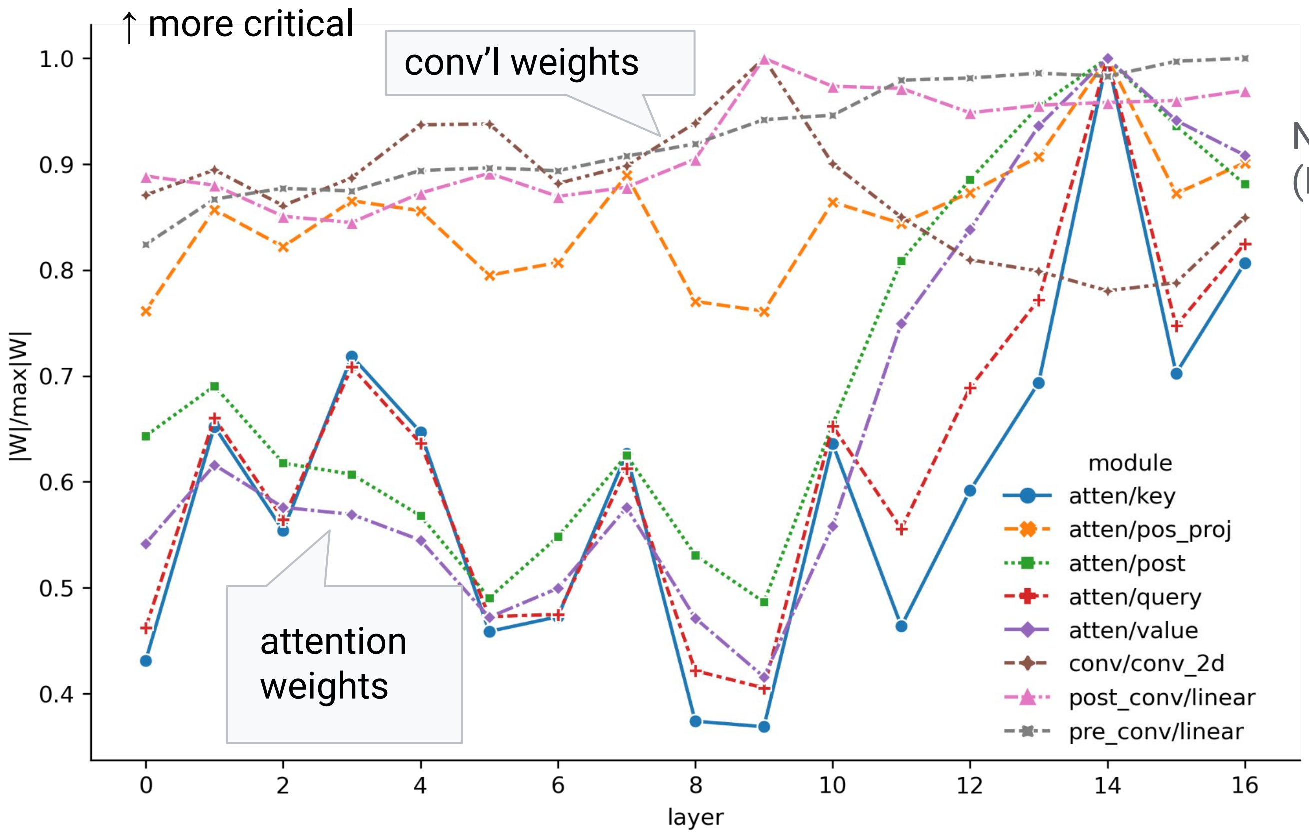
Can we rank layers from in-training metrics (without offline ablation studies)?

Numerical Signatures

- Hypothesis:
  - Weights that change least during training may be least damaging to reset.
  - These may correspond to ambient layers.
- Methodology:
  - Use Frobenius norm to measure change away from initial value at a fixed time, *t*.
  - Compare normalized change for each module, *m*, across layers, *l*.
  - Plot onto [0,1] to show relative importance of a layer wrt a module.

$$\frac{\left|\mathbf{W}^{m,l}(t) - \mathbf{W}^{m,l}(0)\right|_F}{\max_{l'} \left|\mathbf{W}^{m,l'}(t) - \mathbf{W}^{m,l'}(0)\right|_F}$$

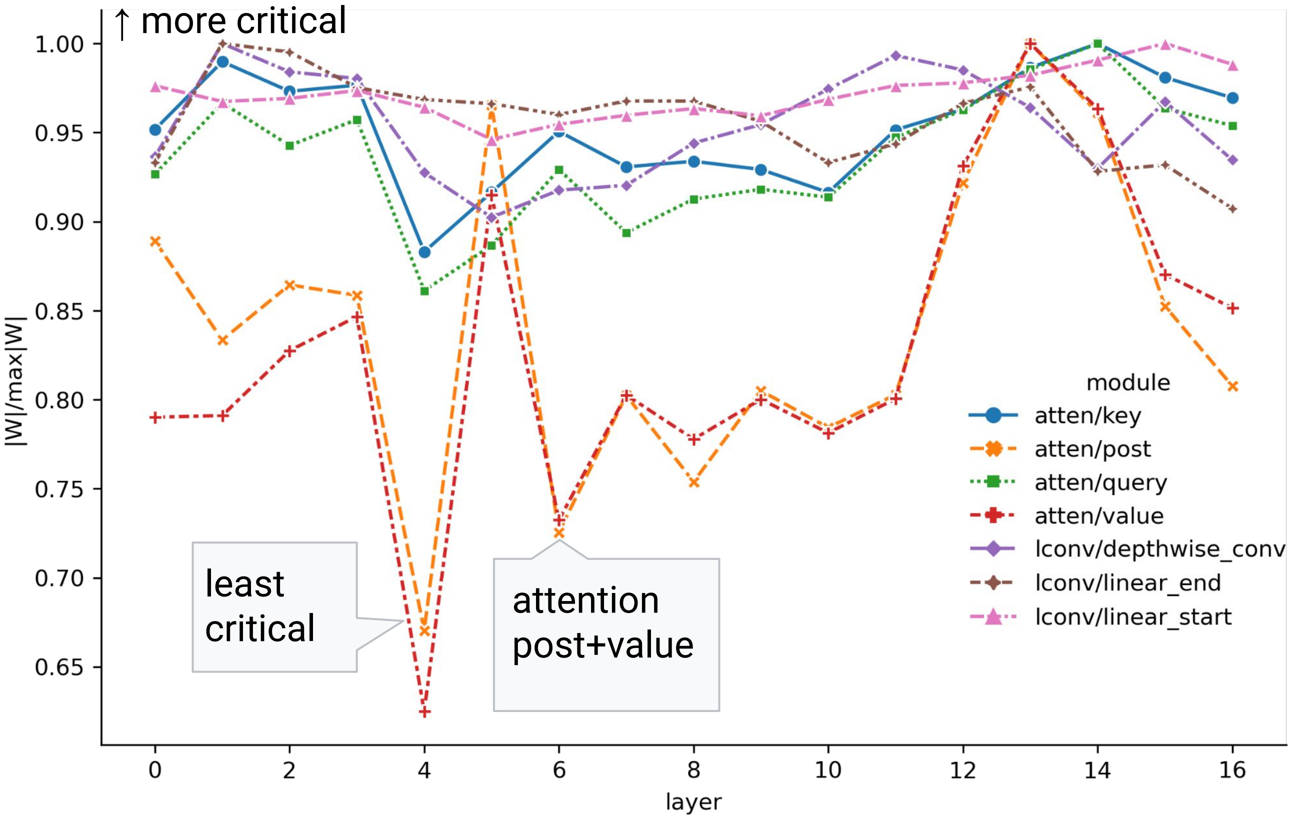
Numerical Signatures



Non-streaming Conformer (LibriSpeech, *t*=70k)

- upper layers experience more updates for *all* attention-related weights
- convolutional weights roughly equidistributed over layers

Numerical Signatures



Streaming Conformer
(MD – SF, t=300k)

- strong variation in two attention weights
- dip at layer 4, the least critical layer

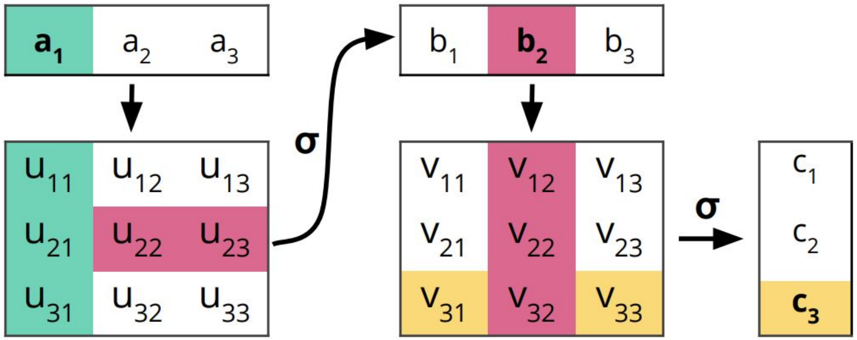
Takeaways

- Attention-layer weight matrices show strong per-layer signature:
  - emerges during training.
  - stable under different seeds.
  - less pronounced for smaller models.
  - shares some features with WER in re-init and re-rand experiments.
- Suggests per-module ablation studies.

Can we use these findings to reduce model training costs in FL?

Applications: Federated Dropout

(i) Original network, with a_1, b_2, and c_3 marked for dropout



(ii) On-device network after Federated Dropout

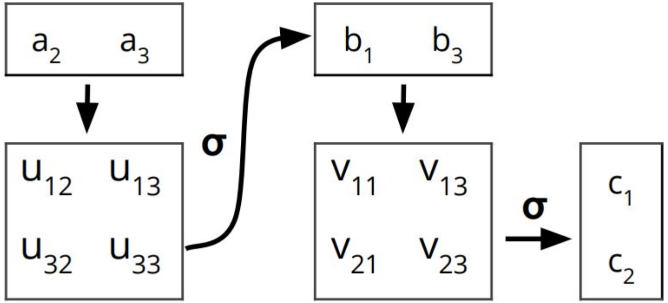


Image source: [19] "Expanding the Reach of Federated Learning by Reducing Client Resource Requirements", Caldas et. al.

Applications: Federated Dropout

- Setup:
  - Fine-tuning on a held-out domain (SF).
  - Apply 50% Federated Dropout (FD) to *n* most critical or ambient layers.
- Comparing settings with same number of parameters trained:
  - Amb-2 vs Crit-3: 7% WER difference.
  - Amb-3 vs Crit-4: 22% WER difference
- Comparing dropping ambient layers to flat dropout across the model:
  - More dropout, same WER.

Dropout	Params Dropped	WER
Crit-2 50%	8%	6.9
Amb-2 50%	9%	6.3
Crit-3 50%	9%	7.0
Amb-3 50%	10%	6.5
Crit-4 50%	10%	7.3
Flat 20%	11%	6.6
Amb-4 50%	12%	6.6

Conclusion

- Ambient properties exists in ASR Conformer.
- Larger models show a higher number of stable ambient layers.
- Attention modules have interesting geometric signature and show some of the per-layer signatures of ambient-ness.
- Up to 22% relative WER improvement when targeting ambient layers for FD, with same number of parameters trained.

Key References

[3] A. Gulati, C.-C. Chiu, J. Qin et al., Eds., "Conformer: Convolution-augmented Transformer for Speech Recognition", 2020.

[4] B. Li, A. Gulati, J. Yu et al., "A better and faster end-to-end model for streaming asr," 2021.

[11] C. Zhang, S. Bengio, and Y. Singer, "Are all layers created equal?" CoRR, vol. abs/1902.01996, 2019.

[17] H. B. McMahan, E. Moore et al., "Communication-Efficient Learning of Deep Networks from Decentralized Data", 2017

[18] Y. Wu and K. He, "Group normalization," CoRR, vol. abs/1803.08494, 2018.

[19] S. Caldas, J. Konečný, B. McMahan *et al.*, "Expanding the Reach of Federated Learning by Reducing Client Resource Requirements," 2018.

[20] V. Panayotov, G. Chen, D. Povey et al., "Librispeech: An ASR Corpus Based on Public Domain Audio Books", 2015

[21] A. Misra, D. Hwang, Z. Huo et al., "A Comparison of Supervised and Unsupervised Pre-Training of End-to-End Models", 2021.

[23] K. Hsieh, A. Phanishayee, O. Mutlu et al., "The non-iid data quagmire of decentralized machine learning", 2019.