MLSP-20.1
Paper #4685
icassp 2022
Singapore

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

Lillian Zhou*, Dhruv Guliani*, Andreas Kabel, Giovanni Motta, Françoise Beaufays {lqz, dguliani, aka}@google.com
* Equal Contributions.

Introduction

Motivation

As advancements are made in Automatic Speech Recognition (ASR), model sizes and associated training costs also grow. This becomes especially prohibitive for training on edge devices with limited computational resources.

One application of training ASR models on-device is for Federated Learning (FL) [17]: an exciting, privacy-preserving technique that allows training models on individual users' devices without their data ever being sent to a central server (Fig. 1).

In order to enable ASR training in this and other on-device scenarios, it is critical to find optimizations that can reduce the associated memory and transport costs.

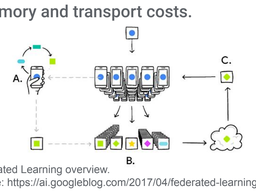

Fig. 1: Federated Learning overview.
Image source: https://ai.googleblog.com/2017/04/federated-learning-collaborative.html.

Layer Importance

It has been shown that different layers of Transformer models have differing importance to the model's function [11]. Layers can be considered *ambient* if they are less important to the model's function, and *critical* if they're more important.

If state-of-the-art ASR models' Conformer architecture [3] also displays this variation in importance, and if we can reliably rank the importance of different layers, this could point to an array of potential improvements in training efficiency. For example, we could only train the most important (critical) layers, or target compression techniques to the least important (ambient) layers.

References

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.

Ablation Studies

Data and Models

To find layers of differing importance, we ran two sets of experiments. First, to test the stability of these properties, and their variance across model sizes, we experimented with three different sizes of non-streaming Conformer [3] (Fig. 2), all trained on the Librispeech corpus [20].

Next, we applied our findings to a state-of-the-art streaming Conformer model [4], using a practical Multi-domain dataset (MD), both with and without a particular Short-form domain (SF) held out [21] (Fig. 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

Fig. 2: Sizes of non-streaming Conformer used.

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

Fig. 3: Datasets used with streaming Conformer.

Methodology

First, we trained the models to convergence. Then, for each encoder layer, we reset its weights to either the initial values (*re-initialization*) or random values (*re-randomization*), and evaluated the resulting ablated model, as illustrated in Fig. 4. To test stability, we repeated this process five times per model.

The results are presented in Figs. 5-7 below, where each column shows one conformer layer, and the color or vertical position shows the Word Error Rate (WER) result of evaluation.

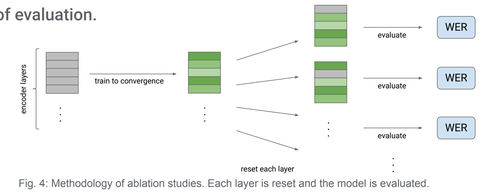

Fig. 4: Methodology of ablation studies. Each layer is reset and the model is evaluated.

Usefulness of Group Normalization

In prior work [11], it was found that Batch Normalization interferes with the formation of ambient layers. Our experiments showed that using Group Normalization [18] can still yield ambient layers (Fig. 5). This is especially beneficial for the FL setting, where Group Normalization is preferred [23].

Fig. 5: WER of layer reset after Batch Normalization vs Group Normalization.

Stability Across Model Sizes

Across our experiments, we found that the larger the model, the more ambient layers it had (Fig. 6). One hypothesis is that this could be due to model overparameterization, which has been shown to benefit neural networks [8,9]. Additionally, we found that larger models also displayed more stability across layers (Fig. 7), meaning that the same layers were found to be ambient across multiple runs.

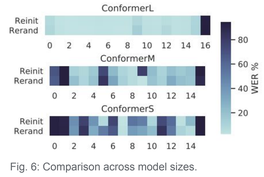

Fig. 6: Comparison across model sizes.

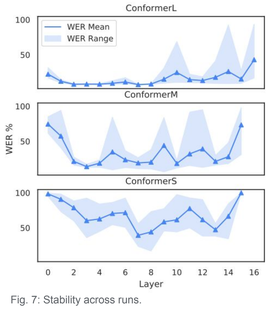

Fig. 7: Stability across runs.

Numerical Signatures

Measure of Change

To investigate the numerical basis of ambient layer formation, we examined the model weights before and after training. Assuming that resetting the weights with least change during training would also be the least damaging, we hypothesized that the weights that changed least would hold some correlation to the layers we found to be ambient.

Using the Frobenius norm to measure the change of model weights between initial time, 0, and a fixed time, *t*, we compared the change for each module, *m*, across layers, *l*:

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

Figs. 8-9 plot this value across layers for each module.

Findings

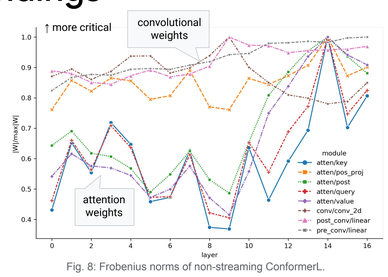

Fig. 8: Frobenius norms of non-streaming ConformerL.

For the non-streaming conformer (Fig. 8), we found that the plotted Frobenius norm of the attention-related modules formed an interesting geometric shape. Most of the change was in the upper layers, while the lower layers were less changed, showing some similarity to the ambient layers that we found through ablation studies.

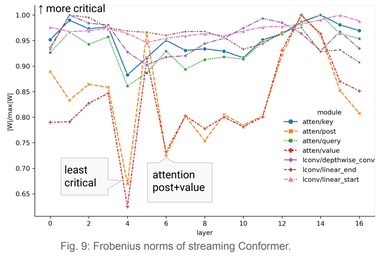

Fig. 9: Frobenius norms of streaming Conformer.

For the streaming Conformer (Fig. 9), we similarly saw a strong variation in two attention modules, post and value, that also bears some resemblance to our empirically-determined ambient layers. In particular, we saw a strong dip at stacking layer 4, which was always most ambient in our experiments.

Also similarly to our empirical results, we observed these properties to emerge during training, to be roughly stable under different initial weights, and to be less pronounced for smaller models. These findings suggest future per-module ablation studies, on top of the per-layer ones we have already shown.

Application

Federated Dropout

Unlike regular Dropout, a regularization technique, Federated Dropout (FD) [19] aims reduce model training costs. As illustrated in Fig. 10, one method is to drop entire rows and columns of weights, reducing the size of the final model to be shipped to device, trained, and shipped back to the server. To make best use of this technique, it would be ideal to find the optimal rows and columns to drop.

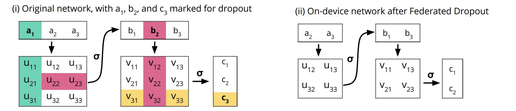

Fig 10: Illustration of Federated Dropout.
Image source: [19] "Expanding the Reach of Federated Learning by Reducing Client Resource Requirements", Caldas et. al.

Targeting Ambient Layers

In these experiments, we trained a model to convergence, ranked the layers by importance, and then performed fine-tuning on data from a held-out domain. For the fine-tuning step, we applied 50% FD to the *n* most critical or ambient layers.

Because stacking layer 4, the most ambient layer, is also twice the size of the other layers, we compared results in Fig. 11 based on the number of parameters dropped, rather than the number of layers. Thus, dropping the 2 most ambient layers was comparable to dropping the 3 most critical layers in terms parameters saved, but the WER was 7% lower. Dropping the 3 most ambient layers was as efficient as dropping the 4 most critical layers, but gave 22% WER improvement. Finally, compared to a flat 20% dropout across the model, we were able to achieve the same WER with fewer params when we targeted dropout to the most ambient layers instead.

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

Fig. 11: Results of targeting FD to ambient layers.

Conclusion

Conclusion

When training the ASR model on-device, memory and transport efficiency are precious, and it is crucial to know which parts of the model may be compressed with least impact to its quality.

Our ablation experiments showed that SOTA ASR model layers varied in importance, and explored how that variation was impacted by model size and normalization technique.

We further examined the model weights and showed interesting geometric signatures of the model's attention modules, suggesting a future direction of research in per-module ablation studies.

Finally, we demonstrated an application of these properties in targeting layers for Federated Dropout, affording computational savings without sacrificing WER.