AUDIO LOTTERY: SPEECH RECOGNITION MADE ULTRA-LIGHTWEIGHT, TRANSFERABLE, AND NOISEROBUST

Review by team **AudioHeads**
[20d070009,20d070067,20d070047]

16th March 2023

Acknowledgements:

This paper was authored by Shaojin Ding from the Texas A&M University and Tianlong Chen, Zhangyang Wang from the University of Texas at Austin.

Summary of the paper:

Due to limited computational resources, large pre-trained Automatic Speech Recognition (ASR) models are compressed to develop on-device models. However, a trade-off exists between computational efficiency and model performance and hence it is hard to develop efficient on-device models in this manner. A recent lottery ticket hypothesis (LTH) revealed the existence of highly sparse subnetworks or winning tickets, that can be trained in isolation without sacrificing the full model performance. This paper investigates the use of the LTH for the development of on-device ASR models.

The authors conducted extensive experiments to analyze the existence of winning tickets in three ASR backbones. Firstly, their results verified the existence of lightweight ASR models or winning tickets in CNN-LSTM, RNN-Transducer, and Conformer models, even at high sparsity. Winning tickets having less than 20% of full model weights were obtained on all backbones. Infact, the most lightweight winning ticket kept only 4.4% weights. The data used in these experiments was from the TED-LIUM (Rousseau et al., 2012), the Common Voice (Ardila et al., 2020), and the LibriSpeech (Panayotov et al., 2015) datasets. Secondly, they compared the iterative weight magnitude pruning (IMP) method with other network compression methods and found that matching subnetworks extracted by IMP significantly exceed those extracted by random pruning or random tickets. Their approach achieved state-of-the-art performance on ASR model compression and showed that both binary pruning masks and weight initialization are indispensable in finding winning tickets. Lastly, they found that the use of weights from pre-trained models to initialize IMP allows it to identify more effective winning tickets, which have higher parameter efficiency, as compared to random initialization.

Throughout their experiments, the authors observed that winning tickets can generalize to structured sparsity without performance degradation. Additionally, it was found that winning tickets identified from large source datasets can achieve matching performance on various target datasets, thus verifying transferability. Lastly, when the training utterances have high background noises, the winning tickets (identified from either target dataset or source datasets) even substantially outperform the full models, showing the extra bonus of noise robustness by inducing sparsity. Thus, we can see the benefits that LTH can bring to server-side and on-device ASR systems.

Strengths:

1. These results jointly demonstrate the benefits of winning tickets in on-device ASR applications.
2. Previous studies have explained the theoretical correctness of the LTH theory. However, this study is the first attempt to apply LTH to real-world use cases by investigating three unique properties that were rarely studied in previous LTH research but are key to user-interactive ASR devices.
3. The study reveals the existence of winning tickets in the context of ASR for the first time. The most lightweight winning tickets from CNN-LSTM, RNN-Transducer, and Conformer backbones only possess 21.0%, 10.7%, and 8.6% remaining nonzero weights, respectively. It also shows that the IMP technique gives winning tickets that outperform tickets that use other network pruning methods.
4. The authors are the first to study the use of structured sparsity in LTH and have successfully found highly sparse winning tickets without performance degradation as compared to using unstructured sparsity.
5. It is observed that winning tickets, especially those from large source datasets, have exceptional transferability across different datasets, and are notably better than full models.
6. For the first time in LTH studies, the authors achieve noise robustness by inducing sparsity. In the presence of background noise, the winning tickets achieve significantly better WERs than full models.

Weaknesses:

1. The paper evaluates the models on TED-LIUM, CommonVoice, and LibriSpeech datasets which are not diverse enough for representing all the different accents and dialects, hence the robustness in terms of accents cannot be confirmed for the winning tickets.
2. There have not been any attempts in the paper to test the model on real world data, and all the inferences and results shown in the paper are based on pre-constructed datasets, thus the claim of noise-robustness is doubtful.
3. The pruning and lottery ticket selection have been done only for some baseline models, and not upon the state-of-the-art models that are currently used for ASR applications, thus the paper does not provide any insight into how well this technique would work in improving the current models
4. There has been absolutely no discussion on the effect of hyperparameters in the entire paper, the paper could have included factors like learning rate, batch size, or the threshold used and their effects on network pruning.
5. While there has been a lot of discussion on the winning tickets and the sub-networks, the paper does not provide much insight into the properties of these tickets and is not helpful for anyone trying to understand the architecture of the sub-networks.
6. The paper could have included a comparison between the result parameters and the extent of compression of the model for a better understanding of the tradeoff involved in using repeated compression with ticket selection.

Questions/Comments:

1. The authors have compared the WER of the IMP algorithm with pre-existing state-of-the-art pruning methods, have they also tried to investigate the transferability and noise robustness of those methods?
2. Have the authors attempted to use this model of theirs for multilingual ASR? Do they expect their model to have similar transferability and noise-robustness for languages other than English.
3. The paper cites a lot of references, and they are not even numbered, the presence of so many unnumbered references sprinkled throughout the paper hampers the overall reading experience