Extending Audio Lottery: Generalization by Sparsity for Multilingual Speech Tasks

Andrew Kondrich
Stanford University
andrewk1@stanford.edu

Abstract

As models become larger and deployed into real-time applications, it becomes crucial to optimize their inference latency, memory footprint, and robustness under distribution shift. Modern neural network pruning techniques leverage the "Lottery Ticket Hypothesis" (Frankle and Carbin, 2018) (LTH) to find subnetworks inside larger models that can increase performance across these desiderata. We explore how LTH impacts performance on utterance intent classification tasks in multilingual settings. While LTH pruned models outperformed the original models at less than 20% of the weights for intent classification in English, they performed worse as the total number of languages in the training data increased. We posit that this behavior supports a "probing" (Cao et al., 2021) explanation to LTH, where LTH pruning discovers the particular subnetwork in the pretrained model that encodes the intent classification task.

1 Introduction

Introduction to Lottery Ticket Hypothesis Deep Learning model compression is an exciting and crucial area of research exploring how to reduce the memory requirements of neural network architectures while maintaining performance. Modern approaches to parameter pruning leverage the Lottery Ticket Hypothesis (Frankle and Carbin, 2018), the belief that a "winning ticket" subnetwork exists in every model that, when trained in isolation, can fully recover the performance of the full-sized model. Deviating from standard after-training pruning approaches, LTH algorithms are iteratively applied during training time to find subnetworks, called "winning tickets", that can be trained effectively in isolation. These winning tickets were shown to not only have significantly less parameters but also improve model accuracy and robustness, inspiring a surge of works across data domains identifying further "unintended" advantages to LTH pruned models ().

LTH in Audio Tasks In a seminal work, (Ding et al., 2022) demonstrated LTH can improve the word error rate for ASR transcription models. We seek to further identify LTH impact on audio tasks, particularly studying multilingual intent classification using popular self-supervised learning (SSL) pretrained models like wav2vec 2.0 (Baevski et al., 2020). Through a series of ablations, we identify how much pruned pretrained models can improve performance when fine-tuned on intent classification tasks and how these advantages hold when increasing the number of languages in the dataset. We seek to demonstrate both the utility of LTH to practitioners working with pretrained audio models and to shed light on what LTH pruning may be doing under the hood.

2 Related Works

LTH and Robustness (Frankle and Carbin, 2018) first introduced the Lottery Ticket Hypothesis, a framework for finding subnetworks in deep neural networks that can be trained from scratch to math or outperform the original network. In incredibly overparameterized modern neural nets, LTh posits that there exists a subnetwork, called a "winning ticket" that can match or exceed the performance of the original model while training for a similar number of iterations. To find winning ticket subnetworks, the work introduces the Iterative Weight Magnitude Pruning (IMP) algorithm, which iteratively removes the smallest $s\%$ of weights across multiple retrains of the model. Once model performance starts to decrease significantly (due to the sparsity), we may stop pruning and choose one of the previous subnetworks per our desired tradeoff between sparsity and performance. In this work, we consider an improved variant of

IMP that resets weights to a kth checkpoint instead of a random initialization (Frankle et al., 2019).

LTH-motivated pruning is a form of learning in itself and provides performance and regularization benefits: (Zhou et al., 2019) showed that a primary effect of IMP is to set weights to 0 that would have converged to 0 over time anyways, and that IMP encodes an inductive bias in the architecture such that even a randomly initiated subnetwork significantly outperforms a random initiation of the full network. (Morcos et al., 2019) showed winning tickets can demonstrate robustness under data distribution shift by "pretraining" the winning ticket on a larger dataset than using the architecture for a smaller one.

(Ding et al., 2022) recently demonstrated for the first time that the Lottery Ticket Hypothesis applies to speech recognition models, demonstrating that 80-95% of weights can be pruned without any impact on the Word Error Rate using IMP. The paper also demonstrates that these pruned models can significantly outperform their full-weight counterparts in environments with background noise, suggesting that LTH pruning approaches can also improve ASR model robustness. This result makes sense in the context of the observation of distribution shift robustness by (Morcos et al., 2019).

Pruning Deep Audio Models Model Pruning techniques have been explored in speech models. For RNN-based speech models, modern pruning techniques like LTH tend to be outperformed by simple random pruning schemes, and approaches that demonstrate success rely on manual observations of weight behavior patterns in the architectures thesmelves to find prunable weights (Zhang and Stadie, 2019). Conflicting evidence exists on the efficacy of LTH in speech tasks: (Lai et al., 2021) observed that LTH approaches do not demonstrate notable gains on WER in finetuning SSL audio models and proposed a better performing non-LTH approach, while (Ding et al., 2022) showed significant WER decrease using off-the-shelf LTH approaches when training speech recognition models end-to-end. To the best of our knowledge, no existing work explores LTH or weight pruning for tasks apart from WER in speech models.

Probing Neural Networks In our experiments, we find parallels between the results of LTH pruning and the probing literature. Neural net probing identifies where certain properties or tasks are exhibited inside of a deep neural network (Adi et al., 2017; Conneau et al., 2018). The primary mechanism to identifying properties is to train a shallow neural net on top of a pretrained embeddings, but (Cao et al., 2021) notably demonstrates that a pruning-based approach can also extract subnetworks directly that perform certain linguistic tasks. While (Cao et al., 2021) does not use LTH, we suspect that similar behaviors may be exhibited while applying IMP during intent classification finetuning on SSL pretrained networks.

3 Approach

We design an ablation study to understand how IMP pruned SSL pretrained models perform on intent classification finetuning in multilingual settings. We seek to understand the difference between the base model performance and the pruned models, as well as how many of the weights can be pruned before performance suffers.

We implement Iterative Weight Magnitude Pruning with rewinding from (Frankle et al., 2019). The algorithm is implemented in 4 main steps:

1. Train randomly initialized network for k steps.
2. Store weights at k, train for n steps.
3. Zero bottom $p\%$ of all weights by L1 magnitude, reset remaining weights to iteration k checkpoint.
4. Repeat steps 2-3 until reaching desired sparsity.

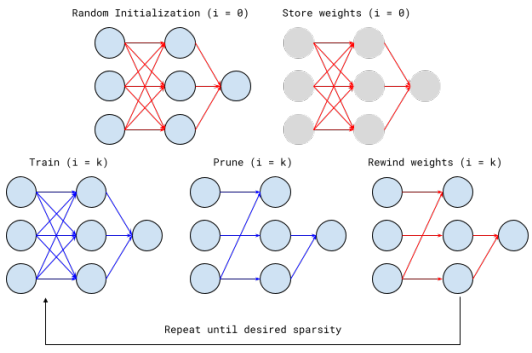

Figure 1: IMP algorithm from (Frankle et al., 2019) visualized for $k = 0$. n, p, k are hyperparameters set via observed validation performance.

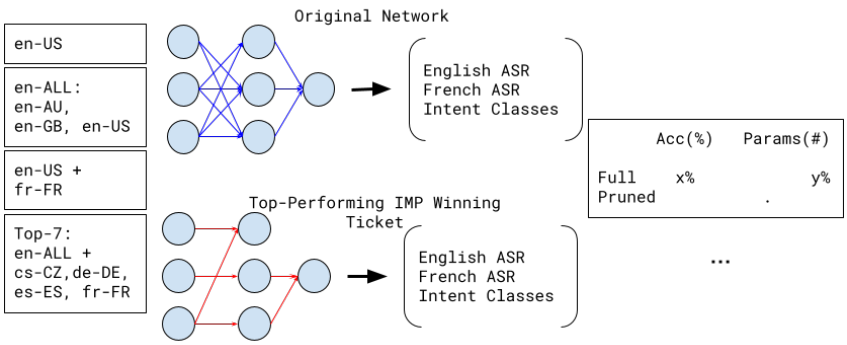

Figure 2: Experimental design across our 4 datasets, in increasing language diversity. For each dataset, we compare performance by training the original model for 50 epochs, training with IMP to select a winning ticket subnetwork, then retraining the subnetwork for 50 epochs, and comparing performance. We choose the winning ticket by picking the IMP checkpoint with the highest validation performance. For our choices of n, p, k as described in Fig 1. for IMP, we use $k = 2$ epochs, $n = 5$ epochs, and $p = 0.2$.

4 Experiments

Dataset We use the Minds14 Dataset from (Gerz et al., 2021a) as the basis for our experiments, consisting of 500 samples per each of 14 languages evenly split across 14 intent classes (Intents and languages listed in Fig.7 and Fig.8). Collected from e-bank call center interactions, the data includes customer utterances indicative of a certain action they would look to complete, for example ”I want to change my address” would fall under the ”Address” intent.

To understand how language diversity impacts LTH performance, we study 4 datasets with increasing number of languages included, starting with two datasets consisting of English with varying number of dialects, then English and French jointly and lastly a 7 language subset called ”Top-7”. The details are shown in Fig. 2.

Experimental Design Details of the training schedules performed are described in Fig. 2. We finetune wav2vec 2.0 (Baevski et al., 2020) pretrained models on Minds14 intent classifications. We attempt various layer freezing strategies. First, training the model end-to-end was unreasonably slow and not able to outperform random guessing on the validation set within 50 epochs. We observed optimal performance when freezing the convolutional feature encoder and finetuning through the contextual sequence representations learned by the Transformer (Vaswani et al., 2017) layer. A task head consisting of a small MLP is trained via Additive Margin Softmax Loss (Wang et al., 2018) to classify intents from the internal sequence representations of the Transformer layer. Model checkpoints and datasets are downloaded via the Huggingface hosted repositories (Lhoest et al., 2021).

Results - English models We include table and graphs illustrating the results on en-US and en-ALL datasets. Evidence suggests that LTH holds well for intent classification on English datasets: We are able to prune over 80% of weights in both experiments while achieving significant performance gains.

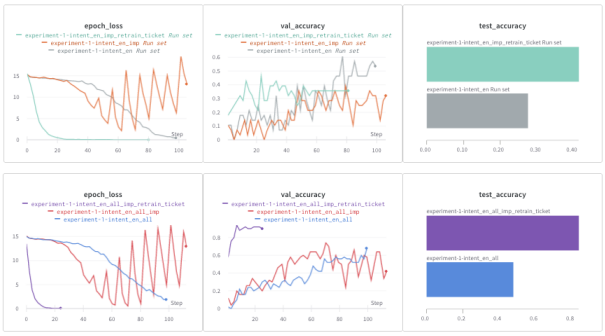

Figure 3: en-US (top) and en-ALL (bottom) charts for loss, validation and test accuracies respectively. Original Model (grey, blue) curves take longer to reach low loss and accuracy performance, while pruned models (green, purple) quickly optimize the objective and reach higher accuracies. Jagged orange and red curves are IMP loss curves, demonstrating how optimization becomes harder with increased weight sparsity.

Training Multilingual models For multilingual intent classification, we initially explored using cross-lingual self-supervised embeddings from XLSR-Wav2Vec2 (Conneau et al., 2020). However, finetuning these models was not performant

	en-US	en-ALL
Achieved Sparsity %	86.7	83.5
Intent Acc - Pruned %	**41.4**	**84**
Intent Acc - Orig %	27.6	48

Table 1: Performance and achieved sparsity - EN datasets

	fr-Fr / en-EN	Top-7
Achieved Sparsity %	67.2	79.1
Intent Acc - Pruned %	**28.6**	16
Intent Acc - Orig %	**28.6**	**22**

Table 2: Performance and achieved sparsity - en/fr and top-7 subdatasets

and were unable to outperform random baselines on the intent classifications. For the remainder of the experiments, we tested multilingual finetuning off of the English-only Wav2Vec pretrained models, with the hypothesis that the frozen encoder is sufficiently general purpose to parse utterances that we could successfully finetune on cross-lingual downstream tasks.

Figure 4: Finetuning off of multilingual embeddings (green) significantly underperformed finetuning english-only pretrained models (purple).

We next show tables and graphs with results on en-EN/fr-FR and Top-7 datasets. In general, we saw that the best performing LTH checkpoints were less sparse than en-only counterparts and either rivaled or underperformed the original model checkpoints.

Figure 5: En-fr (top) and Top-7 (bottom) charts. Green and red lines indicate winning ticket retrain.

5 Probing Hypothesis

We introduce a hypothesis to describe the observed behavior across the datasets and offer some followup questions to be considered in future work.

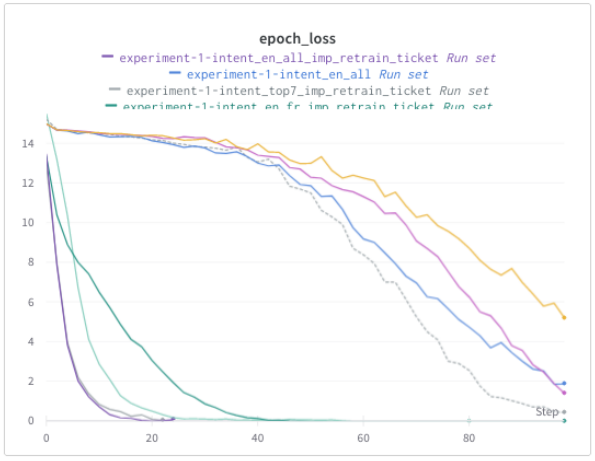

Figure 6: In all cases, we see significant training efficiency gains when training the winning ticket subnetwork from scratch as opposed to the original models, potentially indicating the winning ticket is simply the subnetwork in the pretrained model that implicitly learned features relevant to intent classification via self-supervision.

(Cao et al., 2021) showed that, by learning a weight mask via gradient descent, subnetworks can be identified in SSL pretrained models that solve key linguistic tasks. We posit if we finetuned the particular subnetworks key to intent classification present in the pretrained wav2vec2 model, we would observe comparable training loss efficiency and test accuracy gains as seen with the LTH experiments (fig. 6). Furthermore, a reasonable hypothesis for why the multilingual models were not performing well in our experiments may be because there existed no ”subnetwork” in the pretrained model that was capable of solving multilingual linguistic tasks, since the model is a pretrained on english only. This suggests a number of following questions that can validate this hypothesis. For example, if we freeze the entire pretrained model but still attempt to find a subnetwork for the classification task via pruning, could we see comparable performance to our current finetuning scheme in English and even worse performance in multilingual? Across

multiple repeat IMP runs (with different seeds), do we see the same subnetwork being identified in the fixed pretrained model that is ended up as the winnig ticket for intent classification? We pose this problem as an interesting extension to the current line of work and a potential avenue to explore model interpretability and probing further via LTH subnetworks.

6 Conclusion

We demonstrate that LTH hypothesis can hold for intent classification on English tasks using English pretrained backbones with significant accuracy and parameter efficiency gains. Winning ticket subnetworks are capable of much faster training (Fig. 6) and higher test accuracy than their full-weight counterparts. In the multilingual setting, we did not see an improvement in pruning models finetuned off of English base models, and provide further line of experiments that may explain the lack of performance via a subnetwork probe intepretation of LTH. LTH seems to be a simple yet powerful way to both improve model performance and to increase it's efficiency. Further study is warranted to understand how LTH applies to finetuning pretrained audio models particularly under significant task and data distribution shifts.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477.

Steven Cao, Victor Sanh, and Alexander M. Rush. 2021. Low-complexity probing via finding subnetworks.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *CoRR*, abs/2006.13979.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Shaojin Ding, Tianlong Chen, and Zhangyang Wang. 2022. Audio lottery: Speech recognition made ultra-lightweight, noise-robust, and transferable. In *International Conference on Learning Representations*.

Jonathan Frankle and Michael Carbin. 2018. The lottery ticket hypothesis: Training pruned neural networks. *CoRR*, abs/1803.03635.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. 2019. The lottery ticket hypothesis at scale. *CoRR*, abs/1903.01611.

Daniela Gerz, Pei-Hao Su, Razvan Kusztos, Avishek Mondal, Michal Lis, Eshan Singhal, Nikola Mrksic, Tsung-Hsien Wen, and Ivan Vulic. 2021a. Multilingual and cross-lingual intent detection from spoken data. *CoRR*, abs/2104.08524.

Daniela Gerz, Pei-Hao Su, Razvan Kusztos, Avishek Mondal, Michal Lis, Eshan Singhal, Nikola Mrksic, Tsung-Hsien Wen, and Ivan Vulic. 2021b. Multilingual and cross-lingual intent detection from spoken data. *CoRR*, abs/2104.08524.

Cheng-I Jeff Lai, Yang Zhang, Alexander H. Liu, Shiyu Chang, Yi-Lun Liao, Yung-Sung Chuang, Kaizhi Qian, Sameer Khurana, David D. Cox, and James R. Glass. 2021. PARP: prune, adjust and re-prune for self-supervised speech recognition. *CoRR*, abs/2106.05933.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Sasko, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clement Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander M. Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. *CoRR*, abs/2109.02846.

Ari S. Morcos, Haonan Yu, Michela Paganini, and Yuandong Tian. 2019. One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. 2018. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930.

Matthew Shunshi Zhang and Bradly C. Stadie. 2019. One-shot pruning of recurrent neural networks by jacobian spectrum evaluation. *CoRR*, abs/1912.00120.

Hattie Zhou, Janice Lan, Rosanne Liu, and Jason Yosinski. 2019. Deconstructing lottery tickets: Zeros, signs, and the supermask. *CoRR*, abs/1905.01067.

A Appendices

```
BUSINESS LOAN
FREEZE
ABROAD
APP ERROR
DIRECT DEBIT
CARD ISSUES
JOINT ACCOUNT
BALANCE
HIGH VALUE PAYMENT
ATM LIMIT
ADDRESS
PAY BILL
CASH DEPOSIT
LATEST TRANSACTIONS
```

Figure 7: Minds14 Dataset Intent Classes (Gerz et al., 2021b)

```
CS 574
DE 611
EN-AU 654
EN-GB 592
EN-US 563
ES 486
FR 539
IT 696
KO 592
NL 654
PL 562
PT 604
RU 539
ZH 502
```

Figure 8: Minds14 Dataset Languages and Dataset Sizes (Gerz et al., 2021b)