The Early-Bird Gets the WORM: An Optimization-Based Method to Search for Ideal Early-Bird Tickets (Proposal)

Adithya Vasudev

Abstract

Using the early-bird strategy has been shown to be an energy-efficient method of searching through and identifying winning tickets when pruning neural networks. However, the main bottleneck when identifying early-bird tickets is searching for the ideal early-bird stopping point. The current approach to identifying early-bird tickets involves determining when the early-bird mask difference approaches a threshold set as a hyperparameter. I propose an investigation into finding these early-bird tickets more algorithmically by phrasing the search as an optimization problem. I then investigate two techniques for solving this optimization problem: WORM (Weighted Optimization for Reducing Models), which is a gradient-based approach, and WORM-RL (Weighted Optimization for Reducing Models with Reinforcement Learning), which is a reinforcement learning-based approach.

1. Target Problem

Pruning neural networks has been a popular technique for scaling model sizes and reducing the computational cost of inference on edge devices. However, identifying the optimal pruned network is a combinatorially explosive problem, as the number of possible networks to search through grown exponentially, and thus result in long search and train times to successfully downsample a model. Improvements in finding the optimal pruned network have been made through the identification of winning tickets [3], which are subnetworks that can be trained in isolation to achieve similar performance to the original network, and early-bird tickets [9], which is a method of finding these winning tickets by pruning midway through training. However, there is no fixed mathematical method to determine these early-bird tickets; current approaches use a threshold on the mask distance that's determined by the user. [9].

2. Existing Approaches

Optimization-based approaches to pruning have been proposed. Algorithms such as GraSP [8] and PDP [1] have formulated the generalized lottery-ticket hypothesis into optimization problems, using various regularization-style techniques, and then solving them via gradient based approaches. [8] [1] However, these approaches do not take into account the early-bird strategy [1] [8], which tends to be a more efficient method of finding winning tickets [9]. As a result, while these methods are efficient, they are not optimized by means of the early-bird strategy. Early-bird ticket bounds are typically set by the model designer [9], and techniques for finding the optimal bound for this threshold is either exhaustive or inefficient.

3. Proposed Solution

I propose an investigation into two alternative approaches for identifying early-bird tickets. Both methods involve phrasing the early-bird ticket finding problem as an optimization problem, and then solving the optimization problem via two techniques. The first is a gradient-based approach. The crux of this approach is to modify the approach of using mask distance by turning it into a component of the loss function. I then combine this penalty term with a second penalty term penalizing the duration of the search, and optimize this composite loss function directly via gradient descent.

The second is a reinforcement learning-based approach to solve the optimization problem. The idea behind the approach is to maximize a reward function that takes into account how well a pruned model performs on a validation set, and how early the model was pruned. This approach does away with the mask bound entirely, using Deep Q-Learning to find an optimal model directly.

I plan to test these methods on various models, such as ResNet-20 [4], BERT [2], and Gemma-2B [6], and datasets such as CIFAR-10 [5] and GLUE [7], among others. These models and datasets were chosen to represent a wide variety of model architectures in a variety of different applications and granularities. This allows us to both compare the performance of our method against the performance on ResNet-20 from the original paper [9], while investigating its generalizablity to more modern architectures.

Performance gains will be measured by comparing top-k performance (for the CNN), and n-shot performance (for the Large Language Models), compared to both SOTA for non-pruned, differentiably-pruned models, and early-bird models. Efficiency will be measured by runtime, operation count, and energy consumption.

4. Timeline

February 26th: Submit Proposal
March 11th: Basic Mathematical Formulation of WORM
March 25th: Complete Implementation on CNN
April 1st: Complete Implementation on Transformers
April 8th: Gather Results
April 22nd: Presentation and Paper Completed

5. Expected Outcomes

I expect to find similar performance to the original early-bird strategy, but with a more efficient search process, thereby resulting in theoretical efficiency and energy gains. I also expect similar performance from both the gradient-based and reinforcement learning-based approaches, but with the reinforcement learning-based approach being more generalizable to other optimization problems, due to its less restrictive nature.

References

[1] Minsik Cho, Saurabh Adya, and Devang Naik. Pdp: Parameter-free differentiable pruning is all you need. *ICML*, 2023. 1

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ACL*, 2019. 1

[3] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Training pruned neural networks. *CoRR*, abs/1803.03635, 2018. 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2015. 1

[5] Alex Krizhevsky. Learning multiple layers of features from tiny images. *Technical Report*, 2009. 1

[6] Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, and et al. Gemma. *Technical Report*, 2024. 1

[7] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *ICLR*, 2019. 1

[8] Chaoqi Wang, Guodong Zhang, and Roger Grosse. Picking winning tickets before training by preserving gradient flow. *ACM*, 2020. 1

[9] Haoran You, Chaojian Li, Pengfei Xu, Yonggan Fu, Yue Wang, Xiaohan Chen, Richard G. Baraniuk, Zhangyang Wang, and Yingyan Lin. Drawing early-bird tickets: Towards more efficient training of deep networks. *ICLR*, 2022. 1