Explaining why *Lottery Ticket Hypothesis* works or fails

Shantanu Ghosh
shantan2@andrew.com

Aishwarya Ravi
aravi2@andrew.com

1 Motivation

Discovering a tiny, sparse network instead of a massive neural network can be advantageous for deploying them on devices with limited storage space, such as mobile phones. Additionally, model explainability is essential to engendering trust in AI. The Lottery Ticket Hypothesis (LTH) aims to find a sub-network within a deep network that achieves similar or better performance than the initial deep network. Initial LTH algorithm was evaluated against relatively easier datasets like MNIST, CIFAR *etc.*. In addition, the researchers apply a variety of pruning strategies to enhance the performance of the pruned networks without paying much attention to integrating interpretability and explainability to the pruned subnetworks. In this work, we want to examine why the performance of the pruned networks gradually increases or decreases. To investigate this phenomenon, we utilize the tools from X-AI research and hypothesize that the explanations from the pruned networks are either consistent or inconsistent with the original network. Specifically, we seek to evaluate LTH against a complex dataset, CUB-200. Also, we focus on studying the explainability of the pruned networks in terms of both pixels and high-level concepts using GRAD-CAM and TCAV respectively.

2 Proposal

LTH [1] aims to find a subnetwork within a randomly initialized massive neural network without compromising accuracy. Specifically, (1) we randomly initialize a neural network, (2) train the network for an arbitrary number of iterations, (3) prune a certain number of parameters, (4) reset the remaining parameters to the initial weights in step 1, and (5) continue this process. As a result, we obtain subnetworks, which are also deep networks with fewer parameters than the original network and are referred to as "winning tickets" in [1]. In this paper, we aim to answer the question, *"are these subnetworks utilizing the same features as the big network to perform the downstream classification?"* The answer to this question explains why the subnetworks do not compromise accuracy. To answer this question, we borrow tools from X-AI research and seek to explain the predictions of the subnetworks.

To evaluate the explanations from the subnetworks in terms of pixels, we utilize the famous saliency map [7] based method. This post-hoc-based explanation method aims to identify the important features in the input that contribute the most to the network's output. In this work, we use the gradient-based saliency map method Grad-CAM [6], which integrates the gradient of the predicted output w.r.t the final activation for the computation of saliency maps. However, saliency maps suffer from undesirable drawbacks such as lack of fidelity [3],being inconsistent [4] or providing mechanistic explanation [5].

To mitigate such pitfalls of saliency maps, the researchers focus on explaining a network's prediction in terms of high-level interpretable objects, termed as *concepts* [2]. For example, to explain a network's prediction for "zebra", saliency maps highlight the important region in the image of a "zebra". However, in [2], the researchers explain a "zebra" using a concept *stripness*, important to predict an animal as "zebra". They denote a trained neural network as a composition $f(.) = h(g(.))$, where h is a bottleneck. Using the embeddings from this bottleneck, they learn a binary classifier that distinguishes examples without a concept from examples with a concept. Then, they define the concept activation vector (CAV) as the vector orthogonal to the classifier's decision boundary. Finally, they quantify the importance of a concept as the gradient of the final prediction of the network f w.r.t the CAV and term this metric as the TCAV score of the concept. Along with the saliency maps, we will also compute the TCAV scores for the different concepts for each subnetwork to estimate the importance of the corresponding concept for the final prediction.

Future implications If the subnetworks use the same concepts or features for prediction as the original network, this will explain why their accuracy is not compromised. Also for a complicated dataset like CUB-200, if the performance of the subnetworks decreases gradually, the saliency maps and the TCAV of different concepts may be inconsistent for the subnetworks and the main network. As a future work, this observation play an instrumental role to devise new pruning algorithms, incorporating the tools from X-AI to increase the performance of the subnetworks.

3 Method

(1) Implement unstructured and iterative LTH using [1] (2) Compare the Grad-CAM output [6] on all the pruned subnetworks (3) Get the CAVs for the concepts using [2] (4) Compare the TCAV scores of the important concepts for the all the subnetwork.

4 Dataset

CUB-200. Also if time permits, we will use MNIST.

References

[1] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

[2] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav).(2017). *arXiv preprint arXiv:1711.11279*, 2017.

[3] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 267–280. Springer, 2019.

[4] Vipin Pillai, Soroush Abbasi Koohpayegani, Ashley Ouligian, Dennis Fong, and Hamed Pirsiavash. Consistent explanations by contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10213–10222, 2022.

[5] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[6] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[7] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.