Enhanced Deep Neural Network Model Pruning with Causal Inference

Anishka Ratnawat
CSA, IISc
SR no. 22777
anishkar@iisc.ac.in

Boul Chandra Garai
CSA, IISc
SR no. 23318
chandraboul@iisc.ac.in

Snigdha Shekhar
CSA, IISc
SR no. 22453
snigdhas@iisc.ac.in

Vaisakh P S
CSA, IISc
SR no. 23843
vaisakhp@iisc.ac.in

***Abstract*—Deep Neural Networks (DNNs) have demonstrated remarkable performance across various tasks, but their large size and computational complexity pose significant challenges for deployment across resource-constrained devices. This project will study and evaluate model pruning methods. Furthermore, leverage causal inference techniques for pruning DNNs to improve explainability and pruning operation enhancement.**

I. Introduction

Deep Neural Networks (DNNs) have revolutionized computer vision tasks, achieving state-of-the-art accuracy in various application areas such as autonomous vehicles, natural language processing, e-commerce, and computer vision. However, their computational and memory demands remain substantial, hindering their deployment on resource-constrained platforms such as edge devices and mobile phones. Weight pruning is a powerful technique to mitigate these challenges by reducing the model size while maintaining performance.

II. Model Pruning

DNNs are characterized by their many parameters, which contribute significantly to their memory footprint and computational cost. Pruning aims to **trim down** these networks by selectively removing weights or entire channels, leading to a more compact representation. The motivations for pruning include Resource Efficiency, Energy Savings and Generalization Improvements.

Pruned models with **Resource Efficiency** occupy less memory and require fewer computations during inference, making them suitable for deployment on edge devices and mobile platforms. Reduced model size translates to **Energy Savings** with lower power consumption, crucial for battery-powered devices. Pruning can act as a form of regularization, preventing over-fitting and enhancing the model's ability to generalize to unseen data, hence achieving **Generalization Improvement**.

As an outcome, Model Pruning offers several benefits, such as **Model Compression** by eliminating redundant parameters, pruned models achieve a smaller memory footprint, enabling efficient storage and faster inference. **Speedup** due to reduced computation due to sparsity leading to faster predictions, crucial for real-time applications. **Fine-Grained Control** to tailor the model size according to specific hardware constraints or deployment scenarios. Pruned models serve as excellent starting points for **Transfer Learning**, as they retain essential features while being more lightweight.

A. Existing work on Model Pruning

Pruning-at-initialization (PAI) [1] methods prune unimportant weights right after initialization, avoiding expensive fine-tuning, and their method improves accuracy. SkimCaffe [2] presented a technique to realize size economy and speed improvement while pruning CNNs simultaneously. An Iterative Pruning and Fine-tuning [3] based pruning technique for deploying CNNs on resource-constrained mobile devices surpasses state-of-the-art pruning algorithms and even neural architecture search-based algorithms.

B. Challenges in existing pruning methods

Computational Cost: Traditional weight pruning methods involve iterative removal of ineffective weights or activations, followed by fine-tuning. This process can be computationally expensive. [1] proposed Pruning-at-initialization (PAI) methods prune individual weights right after initialization, avoiding expensive fine-tuning or retraining of the pruned model. However, the pruned model still requires unstructured sparse matrix computation, making it challenging to achieve wall-clock speedups.

Accuracy Degradation: While pruning reduces model size, it often leads to accuracy degradation. Striking a balance between model compactness and performance remains challenging. [4] proposed a method for faster CNNs with direct sparse convolutions and guided pruning, but maintaining accuracy in the compact network is unstable. Recent work [5] investigates mutual information as a criterion for identifying unimportant filters to prune.

Hyper-parameter Sensitivity: Training-based pruning methods introduce additional hyperparameters, such as learning rates for fine-tuning and the number of epochs before rewinding. These hyperparameters can complicate the pruning process and affect reproducibility. J Chang et al. [6] discussed pruning techniques for deploying CNNs on resource-constrained mobile devices, emphasizing the trade-off between accuracy and pruning cost.

Structural Integrity Neglect: Existing pruning strategies often oversimplify the process and neglect the structural integrity and global correlation between layers in the CNN

model. Some pruning methods fail to consider the global context and inter-layer dependencies [5].

Explanation of Effectiveness: Pruning methods need to provide clear explanations for the effectiveness of the proposed network. Understanding why certain filters or channels are pruned is essential. Challenges exist in defining filters' importance and determining the number of pruned filters [7].

III. Problem Description and Strategy

Pruning using Deep Reinforcement Learning [8] and Filter Similarity Analysis [9] explores the problem statement of pruning DNNs. Gradient-based methods effectively visualize CNNs but cannot produce a causal explanation for feature classification. Causal explanation methods give causal interpretation for CNN decisions or abstract CNN models and transform them into causal models.

In response, a class of causal explanation methods has risen to prominence to explain CNN decisions and give a causal interpretation of CNNs [10]. A causal model [11] was built over human-understandable image-based abstractions of the model, which helps in asking counterfactual questions. Based on these ideas, a new structural causal model [12] was introduced characterizing filters as causes and assigning their counterfactual influence at each convolutional layer to understand which filters affect the outputs.

A new approach CexCNN was developed in [13], which learns the most important features by estimating the responsibility, blame, and causality abstraction of each causal filter in the last convolution layer and employs counterfactual reasoning, which comes at the top of the causal hierarchy. In other words, CexCNN identifies salient regions in input images based on the most responsible filters of the last convolution layer.

Research into Causal insights to methods such as Reinforcement Learning [14] [15] [16] has also revealed improvements.

This project will delve into pruning decision impacts and causal reasoning and attempt to improve methods with the obtained causal insights. Focus on evaluating the effect of causal inference and causal modelling on various pruning methods discussed earlier [8] [9] and evaluating its efficacy across different device environments such as laptops, PCs and edge devices.

IV. Project Risks and Consideration

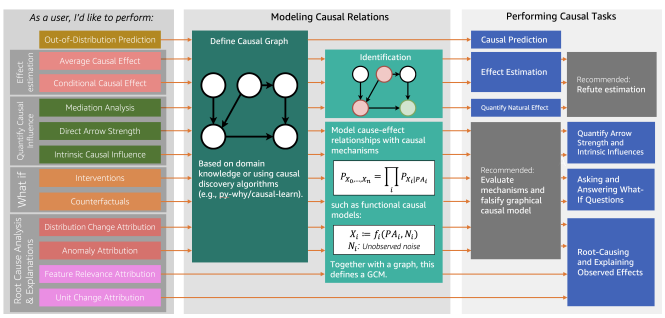

Fig. 1. Causal analysis overview with doWhy [17]

Causal inferences require the generation of causal graphs in estimating causal effects. The availability of software packages such as doWhy [17] [18] does help in handling this complexity, the workflow of which is shown in Figure 1. Yet, the efficiency of Causal Inference through these packages and methods is to be experimented with. Its shortcoming may require further exploration into Causal Discovery [19] methods, which could impact project completion. Furthermore, hardware platforms for evaluating pruned models must be finalized and secured.

References

[1] Y. Cai, W. Hua, H. Chen, G. E. Suh, C. D. Sa, and Z. Zhang, "Structured pruning is all you need for pruning cnns at initialization," 2022.

[2] J. Park, S. Li, W. Wen, P. T. P. Tang, H. Li, Y. Chen, and P. Dubey, "Faster cnns with direct sparse convolutions and guided pruning," *arXiv preprint arXiv:1608.01409*, 2016.

[3] W. Wang, M. Chen, S. Zhao, L. Chen, J. Hu, H. Liu, D. Cai, X. He, and W. Liu, "Accelerate cnns from three dimensions: A comprehensive pruning framework," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 717–10 726.

[4] Y. Cai, W. Hua, H. Chen, F. Li, G. E. Suh, C. D. Sa, and Z. Zhang, "Structured pruning of CNNs at initialization," 2023. [Online]. Available: https://openreview.net/forum?id=iA8XoWjDeGK

[5] Y. Xue, W. Yao, S. Peng, and S. Yao, "Automatic filter pruning algorithm for image classification," *Applied Intelligence*, vol. 54, no. 1, pp. 216–230, 2024.

[6] J. Chang, Y. Lu, P. Xue, Y. Xu, and Z. Wei, "Automatic channel pruning via clustering and swarm intelligence optimization for cnn," *Applied Intelligence*, vol. 52, no. 15, pp. 17 751–17 771, 2022.

[7] Z. Wu, F. Li, Y. Zhu, K. Lu, M. Wu, and C. Zhang, "A filter pruning method of cnn models based on feature maps clustering," *Applied Sciences*, vol. 12, no. 9, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/9/4541

[8] A. M. Hadi, Y. Jang, and K. Won, "Deep reinforcement learning agent for dynamic pruning of convolutional layers," in *Proceedings of the 2023 International Conference on Research in Adaptive and Convergent Systems*, ser. RACS '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3599957.3606236

[9] L. Geng and B. Niu, "Pruning convolutional neural networks via filter similarity analysis," *Mach. Learn.*, vol. 111, no. 9, p. 3161–3180, sep 2022. [Online]. Available: https://doi.org/10.1007/s10994-022-06193-w

[10] P. Schwab and W. Karlen, "Cxplain: Causal explanations for model interpretation under uncertainty," 2019.

[11] M. Harradon, J. Druce, and B. Ruttenberg, "Causal learning and explanation of deep neural networks via autoencoded activations," 2018.

[12] T. Narendra, A. Sankaran, D. Vijaykeerthy, and S. Mani, "Explaining deep learning models using causal inference," *arXiv preprint arXiv:1811.04376*, 2018.

[13] H. Debbi, "Causal explanation of convolutional neural networks," *Oliver, N., Pérez-Cruz, F., Kramer, S., Read, J., Lozano, J.A. (eds) Machine Learning and Knowledge Discovery in Databases. Research Track. ECML PKDD 2021. Lecture Notes in Computer Science(), vol 12976. Springer, Cham.*, 2021.

[14] Z. Deng, J. Jiang, G. Long, and C. Zhang, "Causal reinforcement learning: A survey," 2023.

[15] T. He, J. Gajcin, and I. Dusparic, "Causal counterfactuals for improving the robustness of reinforcement learning," 2023.

[16] M. Seitzer, B. Schölkopf, and G. Martius, "Causal influence detection for improving efficiency in reinforcement learning," 2021.

[17] A. Sharma and E. Kiciman, "Dowhy: An end-to-end library for causal inference," *arXiv preprint arXiv:2011.04216*, 2020.

[18] P. Blöbaum, P. Götz, K. Budhathoki, A. A. Mastakouri, and D. Janzing, "Dowhy-gcm: An extension of dowhy for causal inference in graphical causal models," *arXiv preprint arXiv:2206.06821*, 2022.

[19] Y. Zheng, B. Huang, W. Chen, J. Ramsey, M. Gong, R. Cai, S. Shimizu, P. Spirtes, and K. Zhang, "Causal-learn: Causal discovery in python," *arXiv preprint arXiv:2307.16405*, 2023.