MODEL COMPRESSION BEYOND SIZE REDUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Current Deep Neural Network models are large by design. Model compression methods aim to reduce the size of models maintaining their performance. However, these techniques alter the behavior of the network beyond reducing its size. Furthermore, they can be re-purposed for an entirely different problem than efficiency. This paper aims to draw attention to the matter by highlighting present works around Explaniability, Fairness, Neural Architecture Search, Security, Out-of-distribution generalization, and Life-long learning. It also puts forward suggestions for future research directions.

1 INTRODUCTION

Breakthrough advances in Artificial Intelligence algorithms are consequences of Neural Networks which are the foundations for Deep Learning algorithms, a family of Machine Learning algorithms behind successful Artificial Intelligence tasks like Voice Recognition, Image Classification Krizhevsky et al. (2012), Human Language understanding Devlin et al. (2019), etc. Deep Learning models guarantee to represent, with the help of their underlying neural networks, any continuous function in the real world but the exact size and arrangement of Neural Network parameters for a given problem is still a hyper-parameter, the decision of the designer. Thus, practitioners usually start with over-parameterized deep learning models to ensure representation capacity, thus performance, and aim for efficiency via model compression techniques which allow reducing the size of large models without loss of performance.

Model Compression methods in the literature, for convenience, can be classified into five major parts: Pruning Han et al. (2015), removing an unwanted structure from a trained network, Knowledge Distillation Hinton et al. (2015), a mechanism to pass along the knowledge of a bigger model to a smaller model, Quantization Jacob et al. (2018), reducing the number of bits of model parts required to be represented, Low-rank tensor decomposition Rigamonti et al. (2013), a way to represent weight tensors with their most representative dimensions in a compact form, and other methods such as Wiedemann et al. (2020).

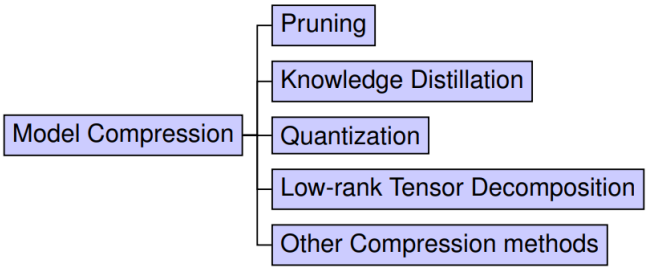

Figure 1: A simplified taxonomy of model compression methods

Either by transforming a large model or by training a new alternative model from scratch, model compression techniques give us an efficient model that has a comparable performance with the original model. But this efficiency can alter the behavior of the network that we care about including the explainability or interpretability, security, bias, and out-of-distribution generalization. On the other hand, beyond their behavior alteration, they can also be effectively repurposed for a task entirely different from efficiency such as Neural Architecture search and Life-long learning.

2 MODEL COMPRESSION BEYOND SIZE REDUCTION

The purpose of Model Compression is to reduce the size of networks, but the size reduction can alter the behavior of the network and this change can benefit or harm performance. Where it benefits, it will be an extra advantage. In fact, in some cases, it can be done for the extra advantage LeCun et al. (1989). All decisions that were made about the network before compressing it can be questioned after the compression, but mentioned here are a few of them for there is a lack of enough work in the area. Furthermore, works that apply compression for other purposes are included.

2.1 REDUCING OVERFITTING

Ideally, a model is expected to generalize and not overfit. Overfitting is when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance. Extreme compression with any kind of compression damage generalization Han et al. (2015), Romero et al. (2015). But the introduction of noise through compression as a means to regularize a neural network is the earliest and most famous practice to reduce overfitting.

In fact, in the early days of neural networks, the purpose of pruning was to reduce overfitting LeCun et al. (1989),Mozer & Smolensky (1988). Recently, dropouts, randomly dropping out weights from the network with a certain probability, Srivastava et al. (2014) enabled models to go deeper than usual and reignited the consideration of compression as a cure for overfitting. In Knowledge Distillation the student can outperform the teacher network Hinton et al. (2015). Similarly, Quantization helps introduce noise Boo et al. (2021) that help in performance.

2.2 EXPLAINABILITY

Explainability or interpretability is an effort to try to understand the decisions of neural networks. It is an area that is getting more and more attention due to high stake applications of neural networks. Under conventional settings, model compression methods impact attribution interpretability Joseph et al. (2020), Park et al. (2020). But they can also help solve it via other means, especially Knowledge Distillation. In Frosst & Hinton (2017), the researchers distilled the knowledge of a high-performing neural network into a decision tree model which is inherently explainable. A later work Liu et al. (2018) generalized the application of Knowledge Distillation for interpretability by formulating the problem as a multi-output regression problem. The result is a Decision tree that performs better than one trained on the data directly but not better than the original neural network. Thus, trading a little bit of accuracy for interpretability. With pruning, it is also possible to remove non-informative features in Convolutional Neural Networks making interpretation easier Hamblin et al. (2022).

The relationship between explainability goes two ways. There is a substantial amount of work demonstrating the use of explainability for Pruning and Quantization Sabih et al. (2020), Yu et al. (2018), Yao et al. (2021). There is still a lack of detailed work between compression and Explainability. For example, what model compression does to Mechanstic Interpretability, an effort trying to understand what happens inside Neural Networks, remained a mystery Olah (2022).

2.3 NEURAL ARCHITECTURE SEARCH

Neural Architecture Search, a relatively recent approach in the AI community, is an attempt to find optimal network architecture in an educated way. This is because existing novel architectures are almost human choices and could have been different. The relationship of Neural Architecture Search with model compression is also intuitive and well recognized in literature Cheng et al. (2017). The task of optimal compression can be seen as a search in the space of sub-architectures. For example, the task of Pruning can be taken as a search in the space of architectures that are sub-networks of the original network Yang et al. (2020), Ashok et al. (2018). Quantization also has a similar relationship with Neural Architecture Search as pruning Wu et al. (2018) but in precision space.

2.4 ALGORITHMIC FAIRNESS

Algorithmic fairness has become increasingly important due to the increasing impact of AI on our society. Particularly, as they are being adopted in various fields of high social importance or automated decision-making, the question of how fair the algorithm is is critical now more than ever. Since Model Compression is becoming a default component of Machine Learning deployment, its impact on fairness has to be examined.

Pruning, and weight decomposition, seem to exacerbate the bias in a networkStoychev & Gunes (2022). Although, both Pruning and Quantization can damage prediction capability on certain parts of a computer vision dataset Hooker et al. (2020) which must be audited, Pruned networks suffer more from this Hooker et al. (2019). But in Natural Language Processing Knowledge Distillation can potentially improve the fairness of the model Xu & Hu (2022). Furthermore, Knowledge Distillation can be used to mitigate the bias introduced by pruning Blakeney et al. (2021).

2.5 SECURITY

Security issues regarding Neural Networks models include Gradient Leakage Attacks, where an attacker can gain access to private training data from a model's gradient, and Adversarial Samples, where an attacker misleads a trained classifier with carefully designed inputs Goodfellow et al. (2014), and Membership Inference Attacks, where an attacker learns about the training data by making repeated inferences Wang et al. (2020a).

In general, both Quantization and Pruning do not help mitigate Adversarial attacks but both of them can be made helpful at the cost of accuracy Zhao et al. (2018). With careful design, Quantization can achieve superior robustness than the original to attain both efficiency and robustness at the same time Lin et al. (2019). Knowledge Distillation is extensively used as a defense mechanism for Adversarial attacks Papernot et al. (2015). Pruning can also be used to prevent Membership Inference Attacks Wang et al. (2020a).

2.6 OUT OF DISTRIBUTION GENERALIZATION

In general, models are trained on data collected at one time. There can be a distribution shift in the data the models are trained with. This can happen due to changes in the underlying data source, environmental noise, recording mechanism, etc. The data is termed out-of-distribution data and the ability of a model to perform well on the new data is an out-of-distribution generalization. This is critical quality models need to have as real-world data keeps changing.

Unfortunately, out-of-distribution performance is correlated positively with larger model size and data Hendrycks et al. (2020). Both pruning and Knowledge Distillation negatively impact the out-of-distribution generalization capability of the model significantly in language tasks Du et al. (2021). But the performance can be preserved between pruned and the original network with a controlled pruning ratio in vision architectures Liebenwein et al. (2021). Fortunately, it is not impossible to have both efficiency and out-of-distribution generalization capability at the same time Diffenderfer et al. (2021).

2.7 LIFE-LONG LEARNING

Once a network has been trained on certain data, its performance can degrade over time as new data classes come in. This phenomenon is termed Catastrophic Forgetting. This problem can potentially be exacerbated by the fact that the original training data might not be available to work with later fine-tuning. In this area, Pruned networks have been favored as the pruned structures make space for learning new ones Geng et al. (2021), Liew et al. (2019), Wang et al. (2020b) but Knowledge Distillation solves it in an entirely different way by remembering the old information via data free approach Binici et al. (2021), Lee et al. (2019), Shmelkov et al. (2017). Therefore, the current model needs to take into consideration this trade-off: to prune and leave enough real estate for future parameters, or distill and try to remember the data for when it is needed. Which approach can solve it better is still an open question to the best of our knowledge.

Table 1: Summary of Model Compression Beyond Size Reduction

NETWORK BEHAVIOUR	DESCRIPTION
Overfitting	- Any dramatic compression damages it slight compression can improve by introducing noise that helps reduce it LeCun et al. (1989) Boo et al. (2021), Hinton et al. (2015)
Explainability	- All of them damage attribution Joseph et al. (2020), but KD can solve it Frosst & Hinton (2017), Liu et al. (2018)
Neural Architecture Search	- Has a positive relationship with all Cheng et al. (2017),Ashok et al. (2018).
Bias	- All exacerbate existing bias in computer vision Stoychev & Gunes (2022) but KD can potentially improve it in language Xu & Hu (2022)
Security(Adversarial attacks)	- Pruning and Quantization harm security Zhao et al. (2018) while KD is used as as solution Zhao et al. (2018)
Security(Membership Inference Attack)	- Pruning can help Wang et al. (2020a).
Out of Distribution Generalization	- It is impacted by all compression types Liebenwein et al. (2021), Du et al. (2021), but its possible to design Diffenderfer et al. (2021).
Life-long learning	- Pruning is used as a solution Geng et al. (2021), Wang et al. (2020b). distillation can assist remember old data in a model Binici et al. (2021) , Lee et al. (2019)

3 DISCUSSION

Model Compression methods beyond size reduction can be seen in perspectives: effect on the network after and re-purposed. Re-purposing compression methods means using them for an entirely different purpose than size reduction. This is intuitive as most of them are adapted from already techniques such as information theory.

It can also be observed that most Model Compression methods, Pruning, Quantization, and Knowledge Distillation, are somehow connected to some natural phenomena outside of their discipline: Pruning is related to how we humans take exploit existing neural connections made earlier Frankle & Carbin (2019), Quantization is similar to how the human brain encodes and stores information Gholami et al. (2022), and Knowledge Distillation as well is related with the concept of how insects have a form that helps them to learn and grow at an early stage but then have an entirely different form once fully grown Hinton et al. (2015). This raises the question if we can find other similar ideas that can be of help. Low-rank decomposition was originally used to factor out representative features from a psychological dataset.

In most cases where they are applied sequentially in combination, more than one of them, the first serves as compression, then the second compression technique serves as a recovery tool rather than a reduction tool.

Whenever there exists a relationship between any network behavior such as explainability, the relationship goes two ways: the compression method affects it and is also affected by it. For example, in Park et al. (2020) explainability is impacted by compression whereas in Sabih et al. (2020) compression is assisted by explainability.

4 OPEN RESEARCH QUESTIONS

Despite efforts to train smaller networks from scratch Frankle & Carbin (2019), most Model Compression methods are applied after a big model is trained. Thus, the compression ought to affect other aspects of the model. For example, as pointed out earlier and also mentioned in Cheng et al. (2017), there are remaining works that can further bridge the concept of a model's size (compression) and its explainability. Some works have been mentioned, but formalized future work at the intersection of model characteristics and compression can be fruitful. Furthermore, as indicated in the discussion section, whenever there is a relationship, it is likely a two-way relationship which makes research in the indicated directions more valuable.

Model Compression is going to become even more important with the advent of Large Language Models and to make them useful practically, a study of their consequences will be an important research direction.

There are again not many works on how the Low-rank weight decomposition Rigamonti et al. (2013) of weights alters the behavior of the network. The fact that it is done layers-wise creates more questions to be researched because the concepts can be explored layer-wise, across tasks, and against the different network characteristics mentioned earlier.

One can raise different behaviors of a network and ask how compression impacts it but also how these methods can be re-purposed as a solution for other problems. Presented here are only a limited number of them because there is still much work to be done in the area. Existing works mainly focus on a specific field such as language or vision, therefore research across vision, language, and multimodal models can also be explored. A comprehensive survey on the effect of Model Compression beyond size reduction and the different ways they can be re-purposed can be extremely helpful to build Model Compression without ramifications and beyond.

REFERENCES

Anubhav Ashok, Nicholas Rhinehart, Fares N. Beainy, and Kris M. Kitani. N2n learning: Network to network compression via policy gradient reinforcement learning. *ArXiv*, abs/1709.06030, 2018.

Kuluhan Binici, Nam Trung Pham, Tulika Mitra, and Karianto Leman. Preventing catastrophic forgetting and distribution mismatch in knowledge distillation via synthetic data. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3625–3633, 2021.

Cody Blakeney, Nathaniel Huish, Yan Yan, and Ziliang Zong. Simon says: Evaluating and mitigating bias in pruned neural networks with knowledge distillation. *ArXiv*, abs/2106.07849, 2021.

Yoonho Boo, Sungho Shin, Jungwook Choi, and Wonyong Sung. Stochastic precision ensemble: self-knowledge distillation for quantized deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 6794–6802, 2021.

Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. A survey of model compression and acceleration for deep neural networks. *ArXiv*, abs/1710.09282, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.

James Diffenderfer, Brian Bartoldson, Shreya Chaganti, Jize Zhang, and Bhavya Kailkhura. A winning hand: Compressing deep networks can improve out-of-distribution robustness. *Advances in Neural Information Processing Systems*, 34:664–676, 2021.

Mengnan Du, Subhabrata Mukherjee, Yu Cheng, Milad Shokouhi, Xia Hu, and Ahmed Hassan Awadallah. What do compressed large language models forget? robustness challenges in model compression. *ArXiv*, abs/2110.08419, 2021.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv: Learning*, 2019.

Nicholas Frosst and Geoffrey E. Hinton. Distilling a neural network into a soft decision tree. *ArXiv*, abs/1711.09784, 2017.

Binzong Geng, Min Yang, Fajie Yuan, Shupeng Wang, Xiang Ao, and Ruifeng Xu. Iterative network pruning with uncertainty regularization for lifelong sentiment classification. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pp. 1229–1238, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3462902. URL https://doi.org/10.1145/3404835.3462902.

Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. *ArXiv*, abs/2103.13630, 2022.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2014.

Christopher Hamblin, Talia Konkle, and George A. Alvarez. Pruning for interpretable, feature-preserving circuits in cnns. *ArXiv*, abs/2206.01627, 2022.

Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Lixuan Zhu, Samyak Parajuli, Mike Guo, Dawn Xiaodong Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8320–8329, 2020.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015.

Sara Hooker, Aaron C. Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget. *arXiv: Learning*, 2019.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily L. Denton. Characterising bias in compressed models. *ArXiv*, abs/2010.03058, 2020.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew G. Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2704–2713, 2018.

Vinu Joseph, Shoaib Ahmed Siddiqui, Aditya Bhaskara, Ganesh Gopalakrishnan, Saurav Muralidharan, Michael Garland, Sheraz Ahmed, and Andreas R. Dengel. Going beyond classification accuracy metrics in model compression. 2020.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 – 90, 2012.

Yann LeCun, John S. Denker, and Sara A. Solla. Optimal brain damage. In *NIPS*, 1989.

Kibok Lee, Kimin Lee, Jinwoo Shin, and Honglak Lee. Overcoming catastrophic forgetting with unlabeled data in the wild. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 312–321, 2019.

Lucas Liebenwein, Cenk Baykal, Brandon Carter, David Gifford, and Daniela Rus. Lost in pruning: The effects of pruning neural networks beyond test accuracy. *ArXiv*, abs/2103.03014, 2021.

Wei Shiung Liew, Chu Kiong Loo, Vadym Gryshchuk, Cornelius Weber, and Stefan Wermter. Effect of pruning on catastrophic forgetting in growing dual memory networks. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2019. doi: 10.1109/IJCNN.2019.8851865.

Ji Lin, Chuang Gan, and Song Han. Defensive quantization: When efficiency meets robustness. *arXiv preprint arXiv:1904.08444*, 2019.

Xuan Liu, Xiaoguang Wang, and Stan Matwin. Improving the interpretability of deep neural networks with knowledge distillation. *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 905–912, 2018.

Michael C. Mozer and Paul Smolensky. Skeletonization: A technique for trimming the fat from a network via relevance assessment. In *NIPS*, 1988.

Chris Olah. Mechanistic Interpretability, Variables, and the Importance of Interpretable Bases. https://transformer-circuits.pub/2022/mech-interp-essay/index.html, jun 22 2022. [Online; accessed 2022-12-14].

Nicolas Papernot, Patrick Mcdaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. *2016 IEEE Symposium on Security and Privacy (SP)*, pp. 582–597, 2015.

Geondo Park, June Yong Yang, Sung Ju Hwang, and Eunho Yang. Attribution preservation in network compression for reliable network interpretation. *Advances in Neural Information Processing Systems*, 33:5093–5104, 2020.

Roberto Rigamonti, Amos Sironi, Vincent Lepetit, and Pascal V. Fua. Learning separable filters. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2754–2761, 2013.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *CoRR*, abs/1412.6550, 2015.

Muhammad Sabih, Frank Hannig, and Juergen Teich. Utilizing explainable ai for quantization and pruning of deep neural networks. *arXiv preprint arXiv:2008.09072*, 2020.

Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. Incremental learning of object detectors without catastrophic forgetting. In *Proceedings of the IEEE international conference on computer vision*, pp. 3400–3409, 2017.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15: 1929–1958, 2014.

Samuil Stoychev and Hatice Gunes. The effect of model compression on fairness in facial expression recognition. *ArXiv*, abs/2201.01709, 2022.

Yijue Wang, Chenghong Wang, Zigeng Wang, Shangli Zhou, Hang Liu, Jinbo Bi, Caiwen Ding, and Sanguthevar Rajasekaran. Against membership inference attack: Pruning is all you need. In *International Joint Conference on Artificial Intelligence*, 2020a.

Zifeng Wang, Tong Jian, Kaushik Chowdhury, Yanzhi Wang, Jennifer Dy, and Stratis Ioannidis. Learn-prune-share for lifelong learning. In *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 641–650. IEEE, 2020b.

Simon Wiedemann, Heiner Kirchhoffer, Stefan Matlage, Paul Haase, Arturo Marbán, Talmaj Marinc, David Neumann, Tung Nguyen, Heiko Schwarz, Thomas Wiegand, Detlev Marpe, and Wojciech Samek. Deepcabac: A universal compression algorithm for deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 14:700–714, 2020.

Bichen Wu, Yanghan Wang, Peizhao Zhang, Yuandong Tian, Peter Vajda, and Kurt Keutzer. Mixed precision quantization of convnets via differentiable neural architecture search. *arXiv preprint arXiv:1812.00090*, 2018.

Guangxuan Xu and Qingyuan Hu. Can model compression improve nlp fairness. *ArXiv*, abs/2201.08542, 2022.

Zhaohui Yang, Yunhe Wang, Xinghao Chen, Boxin Shi, Chao Xu, Chunjing Xu, Qi Tian, and Chang Xu. Cars: Continuous evolution for efficient neural architecture search. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1826–1835, 2020.

Kai Yao, Feilong Cao, Y. W. Leung, and Jiye Liang. Deep neural network compression through interpretability-based filter pruning. *Pattern Recognit.*, 119:108056, 2021.

Ruichi Yu, Ang Li, Chun-Fu Chen, Jui-Hsin Lai, Vlad I. Morariu, Xintong Han, Mingfei Gao, Ching-Yung Lin, and Larry S. Davis. Nisp: Pruning networks using neuron importance score propagation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9194–9203, 2018.

Yiren Zhao, Ilia Shumailov, Robert D. Mullins, and Ross Anderson. To compress or not to compress: Understanding the interactions between adversarial attacks and neural network compression. *ArXiv*, abs/1810.00208, 2018.