Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 11
TPCEE 2025

Review on the Principles and Cutting-Edge Methods of Deep Learning Dynamic Pruning and Model Compression Techniques

Yongxi Zhao

China School of Beijing Institute of Technology Zhuhai, Zhuhai, China

Abstract. Deep learning models have achieved remarkable success in fields such as computer vision, natural language processing, and speech recognition. However, their massive parameter counts and high computational complexity severely restrict their deployment in resource-constrained scenarios like edge devices and embedded systems. As a core model compression technology, neural network pruning achieves model lightweight while maintaining performance by removing redundant connections, neurons, or filters. Based on research related to neural network pruning and more cutting-edge achievements, this paper systematically sorts out the core principles, method classifications, application scenarios, and future challenges of pruning technology, providing a comprehensive reference for the research and application of pruning technology.

Keywords: Pruning; Neural Network; Deep Learning; Computer Visio.

1. Core Principles and Classification of Pruning Technology

1.1. Pruning Based on Neuron Activation Characteristics

This type of method identifies redundant units by evaluating the activation frequency or contribution of neurons during training, with the core idea that "less activation means less importance". Activation-Based Pruning [1] proposes to record the number of activations of neurons during the forward propagation of training (activation counter) and prune neurons with low activation frequencies. This method is divided into global pruning (cross-layer evaluation) and local pruning (layer-wise evaluation). Through iterative pruning and retraining, it achieves sparse low-rank matrix approximation of the hidden layer, which is theoretically equivalent to introducing weighted nuclear norm regularization into the objective function. On the Fashion MNIST dataset, after pruning 80% of the parameters, the accuracy loss is ≤5%, the low-rank approximation effect is better than that of Principal Component Analysis (PCA), and the computational complexity (FLOPs) is reduced by more than 80%. This method can induce a low-rank structure without additional regularization and has a significant effect on fully connected networks, but it needs to be combined with structured pruning for convolutional layers.

Literatures [2-3] use first-order Taylor expansion to approximate and evaluate the contribution of neurons to the loss function, and quantify the importance through the product of gradients and weights. This method belongs to importance-driven unstructured pruning and can be combined with various optimization algorithms. In the metal corrosion image segmentation task, the U-Net model based on Taylor pruning has an Intersection over Union (IoU) loss ≤10% when 90% of the model is pruned, and local pruning (such as layer-wise progressive pruning) is more efficient than global pruning.

1.2. Structured Pruning and Network Architecture Optimization

Structured pruning is based on the premise of retaining the network layer structure and achieves compression by removing entire groups of channels, filters, or layers, resulting in better hardware compatibility. Variable Scale Pruning [4] is proposed for the Transformer architecture. It is found that the weights of the feed-forward layer increase with the network depth, and the weights of the deeper layers are more important. Therefore, a decreasing pruning rate is adopted (e.g., the pruning rate of the encoder layer decreases from 30% to 19%). This method combines Recursive Least Squares (RLS) optimization to dynamically adjust the pruning ratio of each layer. On the Libri-trans

speech recognition dataset, when 59.5% of the model is pruned, the Word Error Rate (WER) loss is <10%, the computational complexity is reduced by 30% compared with global pruning, and it is suitable for the long-sequence modeling requirements of the Transformer.

Channel/Filter Pruning [5] compares weight pruning (unstructured) and channel pruning (structured), and points out that channel pruning retains the network layer structure by removing entire groups of convolutional filters, making it easier to be accelerated by hardware (such as GPU convolution operations). Literature [6] further proposes vectorized kernel pruning, which decomposes convolutional kernels into similar sub-kernel clusters to achieve structured compression. On the AlexNet model, after pruning 70% of the parameters through channel pruning, the inference speed is doubled, and the accuracy loss is ≤3%, while the accuracy loss of weight pruning at the same compression rate exceeds 5%. Dynamic Channel Pruning reduces computational costs by dynamically adjusting the inference path, but the existing methods are prone to a sharp decline in classification performance as the pruning rate increases. To this end, the Dynamic Channel Pruning via Activation Gates (DCPAG) method is proposed: a Channel Pruning Auxiliary (CPA) is used to generate a pruning strategy that balances representational ability and computational cost, which is embedded into the Dynamic Rectified Linear Unit (DyReLU) to form an Embedded Dynamic Rectified Linear Unit (EB-DyReLU), realizing the balance between dynamic pruning and representational ability; at the same time, samples are self-classified according to the difficulty of recognition, and additional training is conducted on hard samples to improve performance. On the CIFAR-10 and ImageNet datasets, under the same pruning rate, the classification accuracy is improved by 0.5%-1.5%, the computational cost is reduced by 5%-20%, and the performance is better than similar channel-based methods.

Aiming at the difficulty of hardware adaptation for sparse matrices in unstructured pruning [7] and the loss of flexibility in structured coarse-grained pruning, Dynamic Probabilistic Pruning (DPP) proposes a multi-granularity (weight, kernel, feature map) pruning mechanism. It realizes differentiable "select 1 out of every k" sampling through Gumbel-softmax relaxation and supports end-to-end optimization (e.g., pruning 1 out of every k weight for each output neuron, or pruning 1 out of every k kernel for each feature map). In image classification tasks, it achieves competitive compression rates and classification accuracy when pruning mainstream deep learning models; its dynamic mask supports the joint optimization of pruning and weight quantization, which can further compress the network; information theory indicators show that its pruning mask has high confidence and diversity.

1.3. Pruning Strategies Driven by Optimization Algorithms

The pruning problem is transformed into an optimization problem, and adaptive pruning is achieved by dynamically adjusting parameters. RLS-Based Pruning [8] proposes to combine RLS optimization with structured pruning, uses the inverse input autocorrelation matrix and weight matrix to evaluate channel importance, and executes pruning when the test loss drops to the level of the unpruned model as the trigger condition. The convergence speed of RLS is 50% faster than that of Stochastic Gradient Descent (SGD), and iterative pruning can be completed in fewer training epochs. On the CIFAR-10 dataset, the accuracy of the VGG-16 model decreases by only 0.12% when 88.74% of the parameters are pruned, and the accuracy of ResNet-50 decreases by 1.02% when 64.27% of the parameters are pruned, which is better than the traditional Taylor pruning algorithm.

Literature [9] proposes to model the efficiency of network elements based on Multi-Output Gaussian Process (MOGP), judge whether to prune by predicting the confidence level, and introduce Lagrange multipliers to balance training costs and performance. This method belongs to probabilistic pruning and can dynamically adjust the pruning threshold. In the ResNet-50 image classification task, when 97.7% of the model is pruned, the accuracy is 20% higher than that of random pruning, and combining Initialization Pruning (BEP-LITE) can reduce the training time by 15%.

To meet the demand for efficiency improvement of ranking-based pruning algorithms [10], Combined Dynamic Programming Ensemble Pruning (ComDPEP) integrates dynamic programming into the classic Complementarity-based Ensemble Pruning (ComEP) algorithm, and constructs an efficient dynamic form (ComDPEP) with the help of two auxiliary tables. On four benchmark classification datasets, the efficiency of ComDPEP is significantly higher than that of ComEP, and it is also better than two advanced algorithms based on uncertainty-weighted accuracy and error reduction pruning, while maintaining the same effectiveness as ComEP.

1.4. Hybrid Pruning and Special Scenario Optimization

Pruning strategies designed by combining multiple technologies or for specific tasks. Literature [11] introduces a logistic growth differential equation to simulate the dynamic change of filter importance, and avoids the sharp accuracy decline of hard pruning by smoothly adjusting the pruning threshold, which belongs to the category of soft pruning. In the large-scale image classification task on ImageNet, when 60% of the filters are pruned, the accuracy loss is 4 percentage points lower than that of traditional methods, and the convergence speed is increased by 30%.

Edge Device Pruning proposes pruning based on filter similarity. By clustering redundant filters and retaining representative features, combined with layer-wise optimization thresholds, low-power deployment is achieved [12-13]. This method significantly reduces the number of memory accesses while maintaining model accuracy (e.g., the cache hit rate is increased by 40%). In the embedded vision system, the inference speed of the pruned CNN model on Raspberry Pi 4 is increased by 1.8 times, and the power consumption is reduced by 25%.

To address the challenge of deploying Deep Neural Networks (DNNs) on resource-constrained devices, Input Difficulty-Adaptive Dynamic Pruning proposes a dynamic pruning method that considers the difficulty of input images, and adjusts the pruning strategy according to the complexity of the input during inference. Experiments on various advanced DNN models on the ImageNet dataset show that this method can reduce the model size and computational complexity without retraining or fine-tuning, providing a direction for the design of lightweight DNN frameworks [14].

Dynamic pruning in information retrieval systems improves query efficiency without reducing effectiveness by setting an upper bound to omit the scoring of documents that are unlikely to enter the final retrieval set [15]. Traditional methods pre-calculate the upper bound during index construction, which limits the dynamic adjustment of the weighted model; the improved method approximately calculates the upper bound based on term statistical information during querying, supporting real-time adjustment of retrieval strategies. A unified notation is used to elaborate on the problem of term upper bound determination, the limitations of existing approximation methods are analyzed, and an upper bound approximation method based on the constrained nonlinear maximization problem is proposed. It is proved that this method does not affect the retrieval effectiveness of modern weighted models and is suitable for Markov Random Field (MRF) neighborhood models. On large-scale web test sets, empirical evidence shows that the accuracy of upper bound approximation can affect the number of scored postings and efficiency.

Dynamic Convolution and Filter Pruning (for surveillance video analysis) aim at the computational resource requirements of large-scale surveillance video analysis [16]. Taking advantage of the high scene similarity, a dynamic convolution architecture is proposed to reuse previous feature maps to reduce computational complexity, and combined with filter pruning to further optimize performance. On 45 surveillance videos of different scenes, the hybrid pruning architecture maintains a Mean Average Precision (mAP) loss $\leq$1.34%, reduces FLOPs by up to 80.4%, and the processing speed is 2.8 times higher than that of traditional single-stage multi-box detectors.

To solve the problems of high computational complexity and low generality of Neural Architecture Search (NAS), Dynamic Distribution Pruning-based Neural Architecture Search (DDPNAS) proposes an efficient unified framework DDPNAS [17]. It provides theoretical bounds for accuracy and efficiency through dynamic distribution pruning, samples architectures from the joint

classification distribution, dynamically prunes the search space and updates the distribution every several training cycles, and directly obtains the optimal architecture under given constraints through an efficient network generation method. On CIFAR-10 and ImageNet (mobile device settings), the searched architectures achieve top-1 accuracies of 97.56% and 77.2% respectively, with the fastest search speed (only 1.8 GPU hours on Tesla V100), and are suitable for different search spaces and device constraints.

2. Application of Pruning Technology in Key Fields

In image classification tasks, lightweight models (such as MobileNet) often adopt channel pruning [7, 9], while high-performance models (such as ResNet) are more suitable for combining activation-based pruning with RLS optimization [1, 4]. Dynamic Probabilistic Pruning (DPP) and input difficulty-adaptive pruning also show excellent performance in image classification.

Literature [18] compares various pruning strategies on CIFAR-100 and finds that progressive pruning (e.g., pruning 5% every 10 epochs) has an accuracy loss of 3-5 percentage points lower than one-time pruning, and combining knowledge distillation can further improve performance. The architecture searched by DDPNAS on CIFAR-10 achieves a top-1 accuracy of 97.56%, verifying the potential of combining pruning with NAS.

In medical image analysis tasks, tasks such as tumor classification have extremely high accuracy requirements, so it is necessary to balance the compression rate and pixel-level segmentation accuracy. Literature [19] proposes combining attention mechanism with pruning. By strengthening the retention of feature channels in tumor regions, the segmentation accuracy loss is <2% when 50% of the model is pruned on the breast tumor dataset.

In end-to-end speech recognition tasks [2], aiming at the encoder-decoder structure of the Transformer architecture, feed-forward layer hierarchical pruning is proposed, and more neurons are retained in the deeper layers (e.g., the pruning rate of the last three layers is ≤20%). On the VoxforgeIT dataset, the model size is compressed by 40%, and the speech recognition accuracy is maintained above 92%.

In natural language processing, language models have a large number of parameters (e.g., BERT-base contains 110 million parameters), and unstructured pruning is difficult to adapt to hardware. Literature [20] proposes the joint optimization of pruning and quantization, and performs structured pruning on the Transformer attention heads (e.g., deleting redundant heads). On the GLUE benchmark task, the performance loss is <1% when 30% of the parameters are pruned.

In the metal corrosion detection task, real-time performance and robustness are required. Literature [3] combines local activation-based pruning with Taylor pruning. In the corrosion images collected by industrial cameras, the pruned Feature Pyramid Network (FPN) model achieves an inference speed of 25 frames per second on the embedded GPU, meeting the needs of online monitoring.

In edge device deployment, Literature [21] proposes heuristic multi-technology fusion: first, reduce the number of filters through channel pruning, then reduce the precision through weight quantization, and finally perform fine-tuning through knowledge distillation. In the smart home vision sensor, the model size is compressed by 75%, and the inference delay is <50ms. The dynamic convolution and filter pruning in surveillance video analysis also provide an efficient solution for edge device deployment, increasing the processing speed by 2.8 times while maintaining accuracy.

The dynamic pruning strategy of information retrieval systems realizes the improvement of retrieval efficiency by approximately calculating the term upper bound during querying, without affecting the effectiveness of various modern weighted models and Markov Random Field (MRF) neighborhood models, and is suitable for efficient retrieval of large-scale web test sets.

3. Current Challenges and Future Directions

3.1. Technical Challenges

The sparse matrices generated by unstructured pruning (such as random weight zeroing) are difficult to be efficiently accelerated by existing GPUs/TPUs, so it is necessary to develop dedicated sparse computing architectures. Most existing methods use a fixed pruning rate, which is difficult to cope with different task complexities. Cross-modal models (such as image-text joint understanding) have complex structures, and it is necessary to balance the feature importance of different modalities during pruning. The existing single-modal pruning technology has poor transferability.

3.2. Future Research Directions

Combine Neural Architecture Search (NAS) to realize the automatic optimization of pruning strategies. For example, DDPNAS demonstrates the potential of combining dynamic distribution pruning with NAS. The layer-wise threshold optimization proposed in Literature [14] can be extended to end-to-end search to reduce the cost of manual parameter tuning. Learn from the synaptic pruning mechanism of biological neural networks to develop pruning algorithms based on plasticity. For example, the logistic growth model in Literature [8] can be further integrated with Hebbian learning rules to improve the rationality of pruning. In the federated learning scenario, study collaborative pruning in a distributed environment to realize the linkage of model compression strategies between edge devices and central servers. For example, the Bayesian modeling in Literature [5] can be extended to distributed inference. Promote the multi-granularity pruning concept of DDP, combine strategies such as input difficulty adaptation and dynamic convolution, and develop a more flexible dynamic pruning framework to adapt to real-time requirements in complex scenarios.

4. Conclusion

Neural network pruning technology has developed from early empirical pruning to an interdisciplinary field integrating optimization theory, hardware characteristics, and task requirements. The combination of structured pruning and hardware acceleration, optimization algorithm-driven adaptive pruning, joint compression of cross-modal tasks, and the integration of pruning with technologies such as NAS/quantization will become the focus of future research. With the surging demand for lightweight models in scenarios such as edge computing, autonomous driving, and smart cities, pruning technology needs to further balance compression efficiency, inference speed, and model accuracy to promote the application of deep learning in a wider range of fields.

Acknowledgements

This paper did not receive any financial support from funding agencies.

References

[1] Ganguli, T., & Chong, E. K. P. (2024). Activation-Based Pruning of Neural Networks. Algorithms, 17 (1), 48. https://doi.org/10.3390/a17010048.

[2] Yu, V. F., Santiyuda, G., Lin, S. -W., Pasaribu, U. S., & Afrianti, Y. S. (2025). Neural Network Pruning for Lightweight Metal Corrosion Image Segmentation Models. IEEE Access, 13, 71673 - 71687. https://doi.org/10.1109/ACCESS.2025.3562435.

[3] Yu, T., Zhang, C., Ma, M., & Wang, Y. (2023). Recursive least squares method for training and pruning convolutional neural networks. Applied Intelligence, 53 (24), 24603 - 24618. https://doi.org/10.1007/s10489 - 023 - 04740 - z.

[4] Ben Letaifa, L., & Rouas, J. -L. (2023). Variable Scale Pruning for Transformer Model Compression in End-to-End Speech Recognition. Algorithms, 16 (9), 398. https://doi.org/10.3390/a16090398.

[5] Rajpal, M., Zhang, Y., & Low, B. K. H. (2023). Pruning during training by network efficacy modeling. Machine Learning, 112 (14), 2653 - 2684. https://doi.org/10.1007/s10994 - 023 - 06304 - 1.

[6] Sivakumar M, Padmapriya S T. Improving Efficiency of Brain Tumor Classification Models Using Pruning Techniques [J]. Current Medical Imaging, 2024, 20 (1): e15734056303076.

[7] Malihi L, Heidemann G. Matching the ideal pruning method with knowledge distillation for optimal compression [J]. Applied System Innovation, 2024, 7 (4): 56.

[8] Hu C, Zhang S, Tao K, et al. SFPBL: Soft Filter Pruning Based on Logistic Growth Differential Equation for Neural Network [J]. Computers, Materials & Continua, 2025, 82 (3).

[9] Koo K, Kim H. V-skp: Vectorized kernel-based structured kernel pruning for accelerating deep convolutional neural networks [J]. IEEE Access, 2023, 11: 118547 - 118557.

[10] Bibi U, Mazhar M, Sabir D, et al. Advances in pruning and quantization for natural language processing [J]. IEEE Access, 2024.

[11] Tian D, Yamagiwa S, Wada K. Heuristic method for minimizing model size of CNN by combining multiple pruning techniques [J]. Sensors, 2022, 22 (15): 5874.

[12] Wu T, Song C, Zeng P. Model pruning based on filter similarity for edge device deployment [J]. Frontiers in Neurorobotics, 2023, 17: 1132679.

[13] Pachon C G, Pinzon-Arenas J O, Ballesteros D. Pruning Policy for Image Classification Problems Based on Deep Learning [C]//Informatics. MDPI, 2024, 11 (3): 67.

[14] Ding Y, Chen D R. Optimization based layer-wise pruning threshold method for accelerating convolutional neural networks [J]. Mathematics, 2023, 11 (15): 3311.

[15] Macdonald C, Ounis I, Tonellotto N. Upper-bound approximations for dynamic pruning[J]. ACM Transactions on Information Systems (TOIS), 2011, 29 (4): 1 - 28.

[16] Liu S Q, Yang Y X, Gao X J, et al. Dynamic channel pruning via activation gates [J]. Applied Intelligence, 2022, 52 (14): 16818 - 16831.

[17] Dai Q, Han X. An efficient ordering-based ensemble pruning algorithm via dynamic programming [J]. Applied Intelligence, 2016, 44 (4): 816 - 830.

[18] Gonzalez-Carabarin L, Huijben I Am M, Veeling B, et al. Dynamic probabilistic pruning: A general framework for hardware-constrained pruning at different granularities [J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 35 (1): 733 - 744.

[19] Pentsos V, Spantidi O, Anagnostopoulos I. Dynamic image difficulty-aware DNN pruning [J]. Micromachines, 2023, 14 (5): 908.

[20] Zheng X, Yang C, Zhang S, et al. Ddpnas: Efficient neural architecture search via dynamic distribution pruning [J]. International Journal of Computer Vision, 2023, 131 (5): 1234 - 1249.

[21] Tsai C Y, Gao D Q, Ruan S J. An effective hybrid pruning architecture of dynamic convolution for surveillance videos [J]. Journal of Visual Communication and Image Representation, 2020, 70: 102798.