BILINGUAL PUBLISHING GROUP

Artificial Intelligence Advances

https://journals.bilpubgroup.com/index.php/aia

RESEARCH ARTICLE

Double-Compressed Artificial Neural Network for Efficient Model Storage in Customer Churn Prediction

***Lisang Zhou*[1*], *Huitao Zhang*[2], *Ning Zhou*[3]**

[1] *Bazaarvoice Inc., Austin, 78759, TX, United States*

[2] *Northern Arizona University, Flagstaff, 86011, AZ, United States*

[3] *Zhejiang Future Technology LLC, Hangzhou, 311200, Zhejiang, China*

ABSTRACT

In the rapidly evolving field of Artificial Intelligence (AI), efficiently storing and managing AI models is crucial, particularly as their complexity and size increase. This paper explores the strategic importance of AI model storage, focusing on performance, cost-efficiency, and scalability within the realm of customer churn prediction, utilizing model compression technologies. Deep learning networks, integral to AI models, have become increasingly large, necessitating millions of parameters. These parameters make the models computationally expensive and voluminous in storage requirements. Addressing these issues, the paper discusses the application of model compression techniques—specifically pruning and quantization—to mitigate the storage and computational challenges. The experimental results demonstrated the effectiveness of the proposed method. These techniques reduce the physical footprint of AI models and enhance their processing efficiency, making them suitable for deployment on resource-constrained devices. Using these models in customer churn prediction in telecommunications illustrates their potential to improve service delivery and decision-making processes. By compressing models, telecom companies can better manage and analyze large datasets, enabling more effective customer retention strategies and maintaining a competitive edge in a dynamic market.

Keywords: Component; Model storage; Model compression; Machine learning; Customer churn prediction

*CORRESPONDING AUTHOR:

Lisang Zhou, Bazaarvoice Inc., Austin, 78759, TX, United States; Email: lzhou@berkeley.edu

ARTICLE INFO

Received: 2 April 2024 | Accepted: 23 April 2024 | Published Online: 29 April 2024

DOI: https://doi.org/10.30564/aia.v6i1.6377

CITATION

Zhou, L., Zhang, H., Zhou, N., 2024. Double-Compressed Artificial Neural Network for Efficient Model Storage in Customer Churn Prediction.. Artificial Intelligence Advances. 6(1): 1–12. DOI: https://doi.org/10.30564/aia.v6i1.6377

1. Introduction

In the rapidly evolving field of Artificial Intelligence (AI), the development and deployment of AI models have become central to transforming industries, enhancing decision-making processes, and improving service delivery across various sectors [1–5]. For instance, Luo et al. developed an utterance-based parallel neural network for effective audio sentiment analysis and demonstrated its effectiveness [1]. Qiu et al. first proposed a Siamese network for pose-guided matching in rehabilitation training and deployed the algorithm in the software [2]. Li et al. introduce DDN-SLAM, the pioneering real-time dense dynamic neural implicit SLAM system that integrates semantic features, achieving excellent performance [3]. However, as these models increase in complexity and size, the challenge of efficiently storing and managing these models becomes critical. This paper explores the strategic importance of AI model storage, particularly focusing on its impact on performance, cost-efficiency, and scalability in the realm of customer churn prediction using model compression technologies.

AI models, particularly deep learning networks, have grown not only in sophistication but also in the size of their architectures and the datasets they require [6, 7]. It is inspired by biology domain which has achieved significant progress in the last decades [8–14]. These models are typically composed of millions, if not billions, of parameters, making them both computationally expensive and large in terms of storage requirements. Efficient model storage is not merely a technical requirement but a strategic one, impacting everything from the speed of model deployment to the cost of operations and the feasibility of real-time analytics.

The significance of AI model storage can be discerned through several critical lenses: performance optimization, cost reduction, and regulatory compliance. Each of these aspects is essential for businesses that rely on AI to drive customer insights and operational efficiencies. For instance, in the telecommunications industry, predicting customer churn allows companies to proactively engage at-risk customers with retention strategies.

Prior research has demonstrated that not all parameters in neural networks (NNs) are crucial [15], leading to over-parameterization [16]. A major challenge in deep neural networks (DNNs) is minimizing computational costs and storage needs to facilitate deployment on devices with limited resources [17]. While cloud deployment of deep learning (DL) models offers substantial computation and storage, it suffers from low throughput and extended response times. Consequently, there is a growing trend to shift inference processes from the cloud to edge devices for real-time tasks like video object detection and segmentation. Current edge device capabilities, however, limit their use for real-time DL model inference and the application in other domains such as biology and finance [18–21]. Moreover, transferring data over networks consumes more energy than local processing due to the high cost of network transmission. Mobile computing, prevalent across various domains, benefits from local data and image processing on the device itself, avoiding the latency involved in server-based processing. Pre-deployment, DL models require training on large datasets, a process that is both time-intensive and dependent on GPUs for speed. Training variants of the popular VGG DL model on the ImageNet dataset, for instance, can take 2-3 weeks based on the network architecture [22].

The telecommunications industry, characterized by its intensive data generation, sees customers continually producing vast amounts of information. One vital application of data mining within this sector is customer segmentation, which organizes customers into different groups based on their attributes, ensuring that those within the same group are highly similar, while those in different groups are distinctly different. However, the predominant focus of research has been on analyzing the data from current customers to devise strategies for attracting new ones. In reality, the expense of acquiring potential customers far exceeds the cost of retaining those at risk of leaving. Hence, there is substantial potential for companies to enhance revenues at a lower cost by examining the data of customers who are either

lost or in the process of leaving to inform marketing decisions. Incorporating model compression techniques can be pivotal in this context, which this article aims to explore further. Model compression not only reduces the computational demands and storage requirements of deploying deep learning models on resource-constrained devices, but it also enhances the efficiency of processing large datasets typical of the telecom industry. By applying these techniques, telecom companies can leverage advanced data analytics for customer segmentation and retention strategies more effectively, enabling faster and more cost-efficient decision-making processes that are crucial for maintaining competitive advantage in a rapidly evolving market.

This paper is structured as follows: section 2 details the related works of customer churn prediction and model compression technologies. Subsequently, section 3 provides the workflow of the proposed method. The experimental results and corresponding discussion are provided in section 4. Finally, section 5 provides a comprehensive conclusion of this paper.

2. Literature Review

2.1 The progresses of customer churn prediction

In the realm of customer churn prediction within the difference industries, a variety of advanced data mining techniques have been explored to improve accuracy and comprehensibility of predictive models. These methodologies underscore the integration of hybrid modeling approaches and advanced algorithmic strategies to address challenges such as large, unbalanced datasets and the need for actionable insights.

M.A.H. Farquad proposed a three-phase hybrid Support Vector Machine (SVM) model aimed at overcoming the opacity of traditional SVMs [23]. This method begins with SVM-Recursive Feature Elimination to minimize the feature set, followed by the generation of an SVM model from which support vectors are extracted. In the final phase, a Naive Bayes Tree—a blend of decision tree and naive Bayesian classifier—is used to formulate rules. Despite its innovative approach, the model struggled with scalability on large datasets.

Qiu et al. introduces a K-means++ approach for segmenting silent customers [24]. Initially, essential variables for the segmentation model were selected, followed by preprocessing of the original data. Subsequently, silent customers were grouped using this method, and the Calinski-Harabasz index was employed to confirm the optimal clustering outcome at k=6. Finally, through radar chart analysis, recommendations were provided to enhance operational and maintenance management and support decision-making in precision marketing.

Chih-Fong Tsai explored hybrid neural network techniques, particularly using back-propagation Artificial Neural Network (ANNs) and self-organizing maps (SOMs) [25], to predict customer churn in datasets provided by American telecom companies. This approach involved using data reduction to enhance prediction accuracy, with the ANN+ANN hybrid model demonstrating superior performance over the SOM+ANN model in various test scenarios.

Wouter Verbeke introduced the application of Ant-Miner+ and ALBA algorithms for churn prediction [26], which utilize Ant Colony Optimization and rule extraction to enhance model comprehensibility and incorporate domain knowledge through monotonicity constraints. This method emphasizes the creation of intuitive and accurate classification rules.

Ning Lu advocated for the use of boosting algorithms to improve churn prediction models [27]. By segmenting customers into clusters based on their risk profiles, and applying logistic regression within these clusters, the boosting method effectively differentiated between customer segments, enhancing the predictive performance compared to single-model approaches.

Collectively, these studies highlight a trend towards combining multiple data mining techniques to tackle the complexities of churn prediction [28]. Hybrid models and sophisticated algorithmic enhancements are central to advancing the field, offering better accuracy and model transparency which are crucial for practical applications in data-rich indus-

tries like telecom and banking.

2.2 The advancements of model compression techniques

Model compression can be broadly categorized into four main techniques: pruning, quantization, knowledge distillation, and compact network design.

Pruning. It involves removing redundant or non-significant parameters from a model [29]. This can be achieved through various methods, such as weight pruning, where small weight values are zeroed out to reduce the model size, and neuron pruning, which removes entire neurons or layers deemed less important. The challenge with pruning lies in determining which parameters to remove without adversely impacting the model's accuracy.

Quantization. It reduces the precision of the numerical values used in the model [30]. By converting floating-point representations to lower-bit quantized versions, quantization can significantly decrease the model's memory requirements and speed up inference, often with minimal loss in accuracy. Techniques range from simple uniform quantization to more complex mixed-precision and adaptive methods.

Knowledge distillation. It is a technique where a smaller "student" model is trained to mimic the behavior of a larger "teacher" model [30]. The student learns from the soft output distributions of the teacher, which carry richer information than hard labels alone. This approach not only reduces the size of the model but often retains much of the teacher's predictive power.

Compact network designs. It involves creating new architectures that are inherently smaller and more efficient yet maintain high performance. Examples include MobileNets, EfficientNets, and SqueezeNet, which use depthwise separable convolutions, network scaling, and bottleneck layers, respectively, to achieve compactness.

Recent studies have shown that combining these techniques can lead to even more efficient models. Model compression not only helps in deploying deep learning models in memory and power-constrained environments but also reduces the carbon footprint associated with training and running AI systems. As the demand for AI on edge devices continues to grow, model compression will play a crucial role in making AI ubiquitous and sustainable.

3. Method

3.1 Dataset Preparation

Churn customers in the communication industry are defined as those who were previously active with frequent communication sessions but have become less active over time and rarely communicate now. These are customers who have been lost or are in the process of being lost. By utilizing the data min-

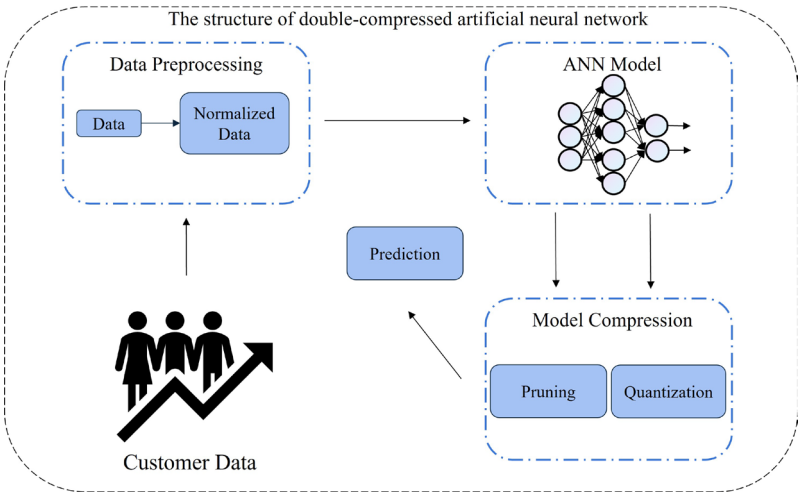

Figure 1. The general workflow of customer churn prediction based on double-compressed artificial neural network.

ing technique of clustering analysis, it is possible to group churn customers, characterize each group, and analyze their properties to devise targeted strategies for customer retention and minimize customer loss.

The study collected data on 125,296 churn customers from a large communication company in one Province, representing all churn customers in a specific city over certain months. The data for churn customers from October and November of the current year was used to ensure the timeliness of the analysis results. The original dataset contains 26 features, which are categorized into customer identity information, customer package fee, package usage, and customer communication frequency.

To mitigate the impact of input variables with differing unit dimensions on distance calculation in the clustering model, it is essential to standardize the key variables using the equations (1) (2) mentioned below. Following the removal of some anomalous data, zero-mean normalization was employed. Thereinto, 80% of data was chosen as training data for model training while the remaining dataset was used for evaluating the model performance.

$$x_{mean}=\frac{1}{n}\Sigma_{i=1}^{n}x_i \tag{1}$$

$$s=\sqrt{\frac{1}{n-1}\Sigma_{i=1}^{n}x_i\ \ (x_i-x_{mean})^2} \tag{2}$$

n is the number of data, x_{mean} and s respectively represent the mean and variance of selected variables.

3.2 Double-compressed ANN model

Figure 1 presents the general workflow of customer churn prediction based on double-compressed artificial neural network.

ANN baseline model

Artificial Neural Networks are a cornerstone of modern machine learning, simulating the way the human brain analyzes and processes information. ANNs are the foundations of deep learning architectures, which have substantially improved the performance of systems in various domains such as image recognition, natural language processing, and predictive analytics. The basic building block of an ANN is the neuron, or node, which receives input from external sources or other neurons and computes an output. Each neuron is connected by weights that are adjusted during the training process to minimize prediction error. The network's ability to learn complex patterns depends on the arrangement of layers and nodes within each layer.

The typical architecture of an ANN shown in **Figure 2** includes an input layer, one or more hidden layers, and an output layer. The input layer receives raw data aligned with the features of the problem domain. The hidden layers perform the majority of

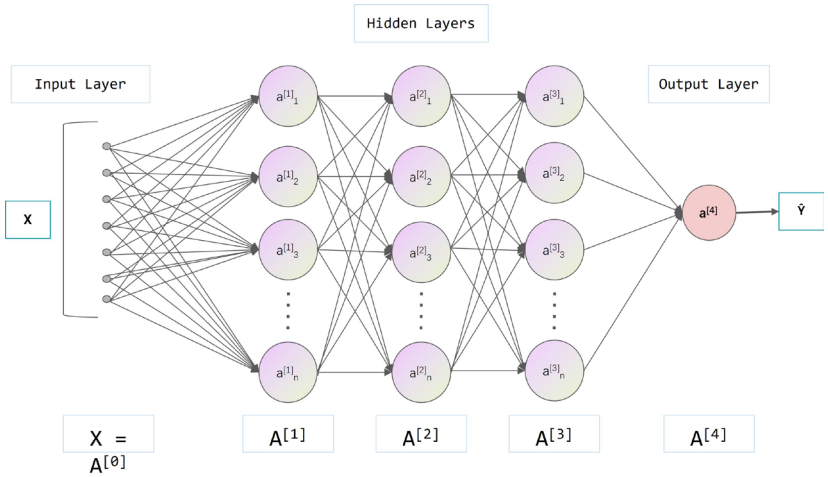

Figure 2. The architecture of the artificial neural network.

computations through their neurons, which are designed to extract progressively higher-level features from the input data. The output layer produces the final prediction or classification based on the learned features and relationships in the data.

This paper introduces a tailored ANN designed specifically for categorizing churn customers into two distinct classes, reflecting the silent customer categories identified in our research. The proposed neural network architecture comprises four hidden layers, which enhances its ability to capture complex, non-linear relationships in the data.

The four hidden layers of the ANN are structured with varying numbers of neurons: 32, 16, 8, and 4, respectively. This configuration allows the network to refine and abstract customer data through each successive layer, providing a robust mechanism for understanding the nuanced behaviors that characterize silent and potentially churning customers. Each layer serves to transform the input data with increasing granularity, ensuring that by the time information reaches the output layer, the network has a comprehensive understanding of the customer profiles.

The final layer of the network is the output layer, which consists of 2 neurons. Each neuron corresponds to one of the two categories of churn customers identified: those at immediate risk of termination and those showing signs of potential future disengagement. This bifurcation enables precise interventions tailored to the specific needs and behaviors of different customer segments.

By implementing a four-layer hidden structure, the ANN can effectively learn from a large dataset of churn customers, identifying complex patterns that are not immediately apparent through traditional analytic techniques. This deep learning model offers a sophisticated tool for telecommunications companies to preemptively address customer churn and enhance retention strategies, ultimately leading to improved customer satisfaction and loyalty.

Pruning

The schematic of general pruning process is shown in **Figure 3**. The "LayerDrop" technique represents a groundbreaking method for managing over-parameterized neural networks, eliminating the need for traditional post hoc pruning while still allowing for the extraction of efficient sub-networks without loss of performance. This approach involves strategically reducing model weights during the training process to create smaller, sampleable sub-networks. These sub-networks are robust and can later be pruned more effectively, extending the utility and flexibility of the network.

Distinct from previous methods, which primarily relied on dropping layers during the training phase to enable the extraction of shallower sub-networks, LayerDrop introduces a novel capability. It allows for the dynamic selection of sub-networks at varying depths during the inference stage. This flexibility ensures that one can tailor the network depth according to specific needs or constraints without compromising the overall network integrity.

This adaptive feature of LayerDrop facilitates the stabilization of training processes for significantly deeper networks than typically feasible, thereby enhancing the model's ability to generalize from training data to real-world applications. The efficacy of this technique is demonstrated in this study by applying it to the DistilBERT model, a streamlined version of the larger BERT architecture known for its efficiency and performance. For this application, we set the LayerDrop rate at $p = 0.2$, carefully chosen to balance network depth with performance during inference.

Quantization

Quantization is a key technique in model compression, aimed at reducing the numerical precision of a model's weights and activations to streamline its computational demands. This approach is particularly beneficial for deploying deep learning models on resource-constrained platforms, where memory bandwidth and storage capacity are limited. In our study, we implemented both static and dynamic quantization methods on various components of the network to enhance its performance efficiency.

The process commenced with the extraction of original floating-point (float32) weights from a previously pruned model. These weights underwent a

specialized quantization procedure designed specifically for converting float weights without affecting the model's precision requirements—this technique is known as weight-only float quantization. As part of this step, the float32 precision embeddings were transformed into quantized embeddings, significantly reducing their size and computational complexity.

Further, we leveraged dynamic quantization, particularly focusing on the model's linear layers, where the majority of computations take place. Using the torch.quantization.quantize_dynamic function, we converted the weights in these linear layers from float32 values to compact 8-bit integer (int8) values. This method dynamically adjusts the quantization parameters based on the distribution of the weights, which is crucial for maintaining the performance of the model while significantly reducing its computational overhead.

By integrating static and dynamic quantization, our approach effectively minimizes the resource demands of the neural network without compromising its accuracy or latency. This dual-quantization strategy exemplifies a practical balance between performance and efficiency, making it a viable solution for real-world applications where speed and size are critical constraints.

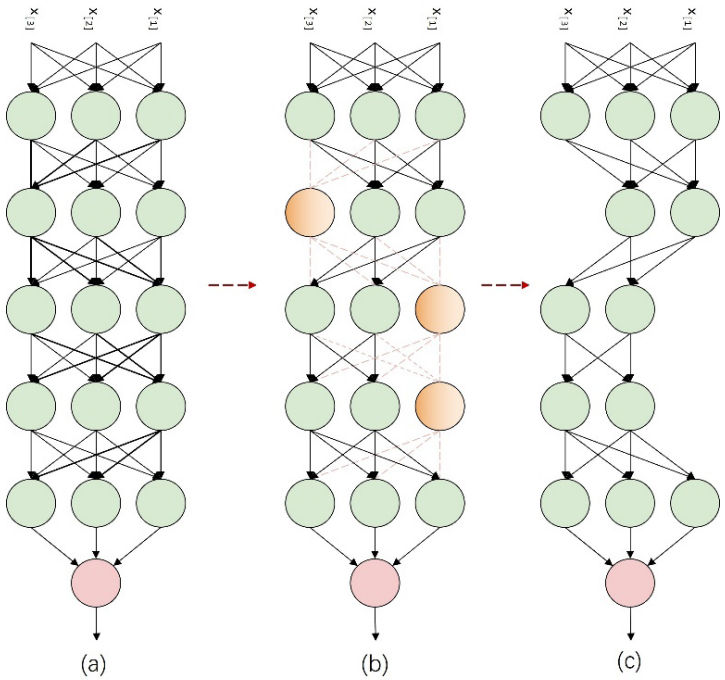

Figure 3. The schematic of pruning: (**a**) baseline ANN model (**b**) the neurons required to be pruned (**c**) the ANN model after pruning.

4. Results and Discussion

4.1 The performance of the baseline model

Figure 4, **Figure 5** and **Figure 6** provide a comprehensive overview of the performance of different ANN models employing various compression techniques, focusing on accuracy, inference time, and model size. Across these metrics, distinct trends emerge, shedding light on the efficacy of pruning and quantization methods in optimizing ANN performance. The first bar chart displays the Accuracy of the various ANN models, providing a clear comparison of their predictive performance after different compression techniques are applied. The second chart illustrates the Inference Time in seconds for each model, where shorter bars represent faster processing, which is advantageous for real-time applications. The third chart depicts the Size in megabytes (MB) of each model, showcasing the reduction in model footprint post-compression techniques. This is particularly relevant for deployment in environments with limited storage capacity.

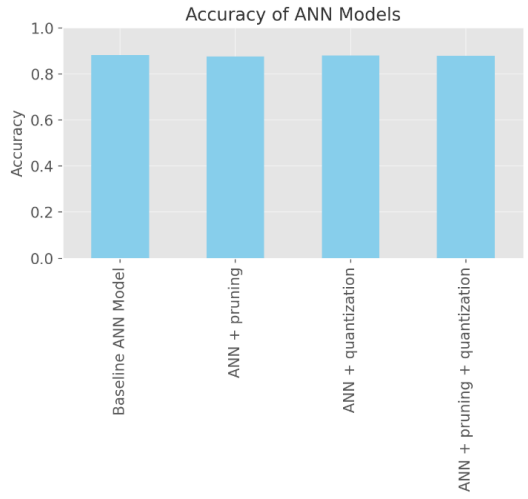

Figure 4. The accuracy of different models.

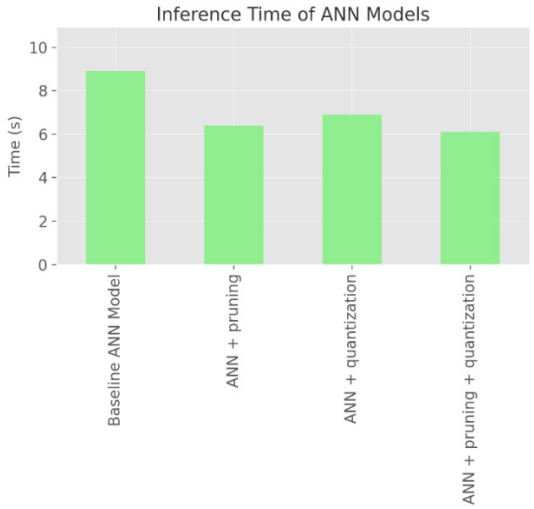

Figure 5. The inference time of different models.

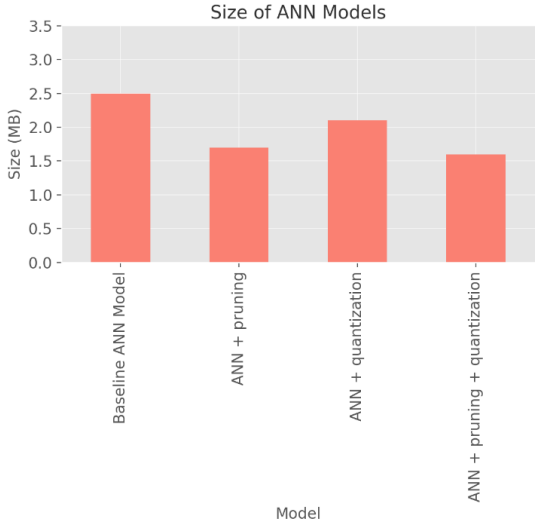

Figure 6. The size of different models.

Table 1. Performance metrics for different models.

Model	Metrics		
	Accuracy	***Inference Time (s)***	***Size (MB)***
Baseline ANN Model	0.883	8.9	2.5M
ANN + pruning	0.876	6.4	1.7M
ANN+ quantization	0.881	6.9	2.1M
ANN + pruning+ quantization	0.879	6.1	1.6M

Beginning with accuracy, it's evident that both the baseline ANN model and the ANN model incorporating both pruning and quantization techniques exhibit relatively higher accuracy compared to those employing only pruning or quantization. This suggests that even with model compression, these techniques can maintain model accuracy, especially when used in conjunction. Regarding inference time, the ANN model utilizing both pruning and quantization techniques demonstrates the fastest inference time, followed by the model employing quantization alone. This indicates that these compression techniques significantly accelerate the inference process, enhancing the model's suitability for real-time applications. Lastly, in terms of model size, the ANN model employing both pruning and quantization techniques exhibits the smallest model size, followed by the model employing quantization alone. This illustrates the effectiveness of these compression techniques in reducing the model's storage requirements, consequently lowering deployment and transmission costs.

It is evident that they excel in reducing model size and inference time while maintaining high accuracy. Pruning reduces model complexity and storage demands by eliminating unnecessary connections and parameters. Quantization further reduces model size by reducing parameter bit-width and accelerates inference speed. When combined, these techniques complement each other, further enhancing model performance while preserving accuracy.

4.2 Discussion

Analyzing the impact of compression techniques

on ANNs provides crucial insights into their efficiency and applicability in various technological environments. The experimental results outlined in the provided charts examine the effectiveness of pruning and quantization on three key metrics: accuracy, inference time, and model size. Each metric is critical for determining the viability of deploying compressed ANNs in real-world applications, particularly where performance, speed, and storage are constrained.

Accuracy. The data shows that ANNs employing both pruning and quantization maintain high accuracy, suggesting that these methods, when combined, effectively compensate for any potential degradation caused by model compression alone. This is particularly noteworthy as maintaining high accuracy is paramount in applications such as medical imaging or autonomous driving where decision-making is based on model predictions. The synergy between pruning and quantization could be attributed to their complementary nature; while pruning eliminates redundant connections, potentially making the model more generalizable, quantization simplifies the computational demand without significantly affecting the predictive power.

Inference Time. In the realm of real-time applications, the speed of inference is as crucial as accuracy. The experimental results highlight that models utilizing both pruning and quantization excel in minimizing inference time. This improvement in speed can be transformative for applications requiring real-time data processing, such as video surveillance and real-time transaction monitoring. By reducing the inference time, these compression techniques enhance the responsiveness of ANNs, thereby supporting more agile and efficient operational capabilities.

Model Size. The reduction in model size is another significant advantage of employing compression techniques. Smaller models are not only easier to store and transmit but also less costly in terms of computational resources. This makes compressed models particularly advantageous for deployment in mobile and embedded systems where memory and processing power are limited. The reduction in model size can also facilitate faster updates and scalability across networks, which is essential for cloud-based AI services and IoT devices.

When pruning and quantization are applied together, they can compound each other's limitations. For example, a model that has been both pruned and quantized might exhibit exacerbated information loss or further reduced robustness to variations in input data. The combined effects of these techniques need thorough testing across various scenarios to ensure that the resultant models can still meet the required standards of reliability and accuracy.

Despite these advantages, the deployment of compressed models is not without challenges. The complexity of choosing the right balance between compression level and performance metrics requires careful tuning and validation to ensure optimal operation. Additionally, certain applications may exhibit sensitivity to reduced precision, leading to potential biases or errors under specific conditions.

In future studies, the potential for further innovations in model compression techniques holds promise for even greater enhancements in ANN performance. Advances might include more sophisticated algorithms for dynamic pruning and quantization that adapt to real-time data inputs [31–38], potentially increasing both accuracy and efficiency. In addition, the hardware should be also improved to be more compatible with these algorithms [39–45]. Furthermore, as edge computing grows, the demand for lightweight models that can operate independently of large data centers will likely increase, amplifying the importance of effective compression techniques.

5. Conclusion

This study has highlighted the pivotal role of model compression techniques in enhancing the efficiency and applicability of AI models, particularly in the context of predicting customer churn in the telecommunications industry. By implementing pruning and quantization, the study demonstrates that it is possible to maintain high accuracy, reduce inference times, and decrease model sizes effectively. These changes are crucial for deploying models in envi-

ronments with limited computational and storage resources, and they facilitate real-time processing and decision-making capabilities essential for customer retention and satisfaction. Pruning and quantization individually and in combination help in simplifying the models while ensuring they remain effective and agile for real-world applications. However, the study also recognizes the inherent limitations associated with these techniques, such as potential loss of model robustness and the challenges in balancing compression with performance. Moving forward, the development of more sophisticated, adaptive compression algorithms could further enhance model performance and efficiency.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Luo, Z., Xu, H., Chen, F.,2019. Audio Sentiment Analysis by Heterogeneous Signal Features Learned from Utterance-Based Parallel Neural Network. In AffCon@ AAAI. pp. 80-87.

[2] Chen, F., Luo, Z., Xu, Y., et al., 2019. Complementary fusion of multi-features and multi-modalities in sentiment analysis. arXiv preprint arXiv:1904.08138.

[3] Luo, Z., Zeng, X., Bao, Z., et al., 2019, July. Deep learning-based strategy for macromolecules classification with imbalanced data from cellular electron cryotomography. In 2019 International Joint Conference on Neural Networks (IJCNN). pp. 1-8.

[4] Qiu, Y., Wang, J., Jin, Z., et al., 2022. Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. Biomedical Signal Processing and Control, 72, 103323.

[5] Qiu, Y., Yang, Y., Lin, Z., et al., 2020. Improved denoising autoencoder for maritime image denoising and semantic segmentation of USV. China Communications. 17(3), 46-57.

[6] Sun, G., Zhan, T., Owusu, B.G., et al., 2020. Revised reinforcement learning based on anchor graph hashing for autonomous cell activation in cloud-RANs. Future Generation Computer Systems. 104, 60-73.

[7] Qiu, Y., Chang, C.S., Yan, J.L., et al., 2019. Semantic segmentation of intracranial hemorrhages in head CT scans. In 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS). pp. 112-115.

[8] Shen, Y., Gu, H.M., Qin, S., et al., 2022. Surf4, cargo trafficking, lipid metabolism, and therapeutic implications. Journal of Molecular Cell Biology. 14(9), 63.

[9] Wang, M., Alabi, A., Gu, H.M., et al., 2022. Identification of amino acid residues in the MT-loop of MT1-MMP critical for its ability to cleave low-density lipoprotein receptor. Frontiers in Cardiovascular Medicine. 9, 917238.

[10] Shen, Y., Gu, H.M., Zhai, L., et al., 2022. The role of hepatic Surf4 in lipoprotein metabolism and the development of atherosclerosis in apoE−/− mice. Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids. 1867(10), 159196.

[11] Wang, B., Shen, Y., Zhai, L., et al., 2021. Atherosclerosis-associated hepatic secretion of VLDL but not PCSK9 is dependent on cargo receptor protein Surf4. Journal of Lipid Research. 62.

[12] Deng, S.J., Shen, Y., Gu, H.M., et al., 2020. The role of the C-terminal domain of PCSK9 and SEC24 isoforms in PCSK9 secretion. Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids. 1865(6), 158660.

[13] Shen, Y., Wang, B., Deng, S., et al., 2020. Surf4 regulates expression of proprotein convertase subtilisin/kexin type 9 (PCSK9) but is not required for PCSK9 secretion in cultured human hepatocytes. Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids. 1865(2), 158555.

[14] Xia, D., Alexander, A.K., Isbell, A., et al., 2017. Establishing A Co-Culture System For

Clostridium Cellulovorans And Clostridium Aceticum For High Efficiency Biomass Transformation. J. Sci. Heal. Univ. 14, 8-13.

[15] Denker, J., LeCun, Y., 1990. Transforming neural-net output levels to probability distributions. Advances in neural information processing systems, 3.

[16] Denil, M., Shakibi, B., Dinh, L., et al., 2013. Predicting parameters in deep learning. Advances in neural information processing systems. 26.

[17] Zhu, M., Gupta, S., 2017. To prune, or not to prune: exploring the efficacy of pruning for model compression. arXiv preprint arXiv:1710.01878.

[18] Horne, J., Beddingfield, E., Knapp, M., et al., 2020. Caffeine And Theophylline Inhibit B-Galactosidase Activity And Reduce Expression In Escherichia Coli. Acs Omega. 5(50), 32250-32255.

[19] Mock, M.B., Zhang, S., Pniak, B., et al., 2021. Substrate Promiscuity of The Ndmcde N7-Demethylase Enzyme Complex. Biotechnology Notes, 2, pp.18-25.

[20] Tarca, A.L., Carey, V.J., Chen, X.W., et al., 2007. Machine learning and its applications to biology. PLoS computational biology. 3(6), e116.

[21] Qiu, Y., 2019. Estimation of tail risk measures in finance: Approaches to extreme value mixture modeling (Doctoral dissertation, Johns Hopkins University).

[22] Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[23] Farquad, M.A.H., Ravi, V., Raju, S.B., 2014. Churn prediction using comprehensible support vector machine: An analytical CRM application. Applied Soft Computing. 19, 31-40.

[24] Qiu, Y., Chen, P., Lin, Z., et al., 2020. Clustering Analysis for Silent Telecom Customers Based on K-means++. In 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC) .1, 1023-1027.

[25] Tsai, C.F., Lu, Y.H., 2009. Customer churn prediction by hybrid neural networks. Expert Systems with Applications. 36(10), 12547-12553.

[26] Verbeke, W., Martens, D., Mues, C., et al., 2011. Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert systems with applications. 38(3), 2354-2364.

[27] Lu, N., Lin, H., Lu, J., et al., 2012. A customer churn prediction model in telecom industry using boosting. IEEE Transactions on Industrial Informatics. 10(2), 1659-1665.

[28] He, B., Shi, Y., Wan, Q., et al., 2014. Prediction of customer attrition of commercial banks based on SVM model. Procedia computer science. 31, 423-430.

[29] Liu, Z., Sun, M., Zhou, T., et al., 2018. Rethinking the value of network pruning. arXiv preprint arXiv:1810.05270.

[30] Polino, A., Pascanu, R., Alistarh, D., 2018. Model compression via distillation and quantization. arXiv preprint arXiv:1802.05668.

[31] Luo, Z., Xu, H., Chen, F., 2018. Utterance-based audio sentiment analysis learned by a parallel combination of cnn and lstm. arXiv preprint arXiv:1811.08065.

[32] Chen, F., Luo, Z., 2018. Learning robust heterogeneous signal features from parallel neural network for audio sentiment analysis. arXiv preprint arXiv:1811.08065.

[33] Chen, F., Luo, Z., 2019. Sentiment Analysis using Deep Robust Complementary Fusion of Multi-Features and Multi-Modalities. CoRR.

[34] Liu, Y., Bao, Y., 2021. Review of electromagnetic waves-based distance measurement technologies for remote monitoring of civil engineering structures. Measurement. 176, 109193.

[35] Liu, Y., Bao, Y., 2022. Review on automated condition assessment of pipelines with machine learning. Advanced Engineering Informatics. 53, 101687.

[36] Liu, Y., Hajj, M., Bao, Y., 2022. Review of

robot-based damage assessment for offshore wind turbines. Renewable and Sustainable Energy Reviews. 158, 112187.

[37] Sugaya, T., Deng, X., 2019. Resonant Frequency Tuning Of Terahertz Plasmonic Structures Based On Solid Immersion Method. 2019 44th International Conference On Infrared, Millimeter, And Terahertz Waves, pp.1-2.

[38] Yu, F., Milord, J., Orton, S., et al., 2021. Students' Evaluation Toward Online Teaching Strategies for Engineering Courses during COVID. In 2021 ASEE Midwest Section Conference.

[39] Yu, F., Milord, J. O., Flores, L. Y., et al., 2022. Work in Progress: Faculty choice and reflection on teaching strategies to improve engineering self-efficacy. In 2022 ASEE Annual Conference.

[40] Deng, X., Li, L., Enomoto, M., et al., 2019. Continuously Frequency-Tuneable Plasmonic Structures For Terahertz Bio-Sensing And Spectroscopy. Scientific Reports. 9(1), 3498.

[41] Deng, X., Simanullang, M., Kawano, Y., 2018. Ge-Core/A-Si-Shell Nanowire-Based Field-Effect Transistor For Sensitive Terahertz Detection. Photonics. 5(2), 13.

[42] Liu, Y., Liu, L., Yang, L., et al., 2021. Measuring Distance Using Ultra-Wideband Radio Technology Enhanced By Extreme Gradient Boosting Decision Tree (Xgboost). Automation In Construction. 126, 103678.

[43] Yu, F., Strobel, J., 2021. Work-in-Progress: Pre-college Teachers' Metaphorical Beliefs about Engineering. In 2021 IEEE Global Engineering Education Conference (EDUCON) . pp. 1497-1501.

[44] Kundu, S., Fu, Y., Ye, B., et al., 2022. Toward Adversary-aware Non-iterative Model Pruning through D ynamic N etwork R ewiring of DNNs. ACM Transactions on Embedded Computing Systems. 21(5), 1-24.

[45] Milord, J., Yu, F., Orton, S., et al., 2021. Impact of COVID Transition to Remote Learning on Engineering Self-Efficacy and Outcome Expectations. In 2021 ASEE Virtual Annual Conference.