EXPLORING COMPRESSION STRATEGIES FOR LARGE LANGUAGE MODELS TOWARDS EFFICIENT ARTIFICIAL INTELLIGENCE IMPLEMENTATIONS

Doinița ȘENDRE[1]

Dana-Mihaela PETROŞANU[2]

Alexandru PÎRJAN[3]

Abstract

The rapid advancements of Artificial Intelligence (AI) technologies, particularly Large Language Models (LLMs), have brought and accelerated significant innovations across various domains. Regardless of their widespread usefulness, the scalability of LLMs poses considerable challenges, primarily due to their substantial demands on computational and energy resources. This article explores the importance of developing and applying effective compression techniques to mitigate these numerous challenges. Techniques such as pruning, quantization, and knowledge distillation are analyzed for their potential to decrease a LLM's size and its associated computational demands, while striving to maintain performance integrity. Each technique inherently presents unique trade-offs between model efficiency and accuracy, requiring a nuanced understanding of their applications. We have made an in-depth analysis into the complexities of implementing these techniques, highlighting the balance required between performance and compression, along with the complex process of customization to specific LLM architectures. The article further analyzes the very important validation and testing phases that are much needed for ensuring that compressed models perform adequately in real-world applications. We have also considered the future adaptability of compression techniques to evolving AI models and architectures. The conducted study emphasizes the ongoing need for innovative research in model compression in order to make AI technologies more sustainable and accessible across various sectors, thereby expanding their potential benefits while addressing the limitations and risks associated with their deployment.

Keywords: Large Language Models, Model Compression, Pruning Techniques, Quantization, Knowledge Distillation, Computational Efficiency, Neural Network Optimization, Artificial Intelligence Scalability

[1] PhD Professor, School of Management-Marketing, Romanian-American University, 1B, Expozitiei Blvd., district 1, code 012101, Bucharest, Romania, doinita.sendre@rau.ro

[2] PhD Lecturer, Department of Mathematics-Informatics, National University of Science and Technology Politehnica Bucharest, 313, Splaiul Independentei, district 6, code 060042, Bucharest, Romania, dana.petrosanu@upb.ro

[3] PhD Hab. Professor, School of Computer Science for Business Management, Romanian-American University, 1B, Expozitiei Blvd., district 1, code 012101, Bucharest, Romania, alexandru.pirjan@rau.ro

JEL Classification: O3, O33, O34, O35, O36

1. Introduction

In the contemporary landscape of AI, LLMs have emerged as very important tools, driving innovations across a multitude of domains. These models, which include prominent examples such as Generative Pre-trained Transformer (GPT) and Bidirectional Encoder Representations from Transformers (BERT), leverage vast datasets to understand and generate human-like text, offering capabilities that extend into Natural Language Processing (NLP), Machine Learning (ML), and beyond. The applications of LLMs are numerous, covering various areas such as automated content generation [1–3], real-time language translation [4–7], sentiment analysis [8–11], and even aiding in medicines discovery [4,12–15], neurosciences [16], geology [17] or in legal document analysis [18–20]. This wide-ranging applicability emphasizes the models' growing importance in both academic research and industry.

Nevertheless, the scalability of LLMs presents significant challenges. As the size and complexity of these models increase, so do their demands for computational and energy resources. Training state-of-the-art LLMs often requires extensive hardware setups, including multiple high-end Graphics Processing Units (GPUs) or Tensor Processing Units (TPUs), which are cost-prohibitive and also raise environmental concerns due to their high energy consumption. The intensive computational requirements can limit the accessibility of cutting-edge AI technologies, particularly for researchers and organizations with limited resources [1,3,13,21,22].

Given these constraints, the development and application of efficient compression techniques for LLMs assume critical importance. Compression techniques aim to reduce the size of neural networks without significantly compromising their performance. Techniques such as knowledge distillation, quantization, and pruning are employed to create lighter models that retain the efficacy of their larger counterparts while being more economical and environmentally sustainable. The implementation of such compression techniques is not devoid of challenges. Compressed models often face trade-offs between size, speed, and accuracy. While a smaller model size may result in faster computation times and lower energy usage, it might also lead to a decrease in the model's ability to generalize across tasks or maintain the same level of accuracy as the original model. Conversely, maintaining high accuracy can limit the degree of achievable compression. In addition, the process of model compression can be complex and requires careful tuning and validation to ensure that the reduced model still adheres to the performance standards necessary for practical applications.

Despite these challenges, the advantages of model compression are significant, offering a pathway towards more sustainable and accessible AI technologies. As this field progresses, understanding the nuances of various compression techniques and their impact on model performance will be essential. This article aims to make an in-depth analysis into these aspects, presenting an overview of the current methodologies, challenges, and potential future directions in the compression of LLMs.

2. Research methodology

The research methodology section of this study plays a very important role in establishing the analytical approach used to understand compression strategies for LLMs. Given the field's rapid growth and the increasing focus on computational efficiency, this study aims to identify and analyze relevant scientific literature through a well-defined and systematic process. The chosen methodology ensures comprehensive coverage of the relevant scientific literature research and addresses important issues concerning compression techniques in AI.

The Clarivate Web of Science (WoS) database has been selected for its extensive indexing of high-quality scientific literature across multiple disciplines. WoS provides a curated database of peer-reviewed articles, ensuring that the retrieved scientific works maintain the highest standards of academic rigor. The usage of the WoS database has enabled us to draw from a reliable source that is widely recognized in academic circles for its authority and breadth, thereby enhancing the credibility of the research findings.

Consequently, the selected database has been chosen for this study due to its comprehensive and authoritative collection of scientific literature across various disciplines. This choice has offered us curated indexing, allowing access to high-quality, reliable information. By employing the query "TS=((LLM* OR LARGE LANGUAGE MODEL*) AND (AI OR ARTIFICIAL INTELLIGENCE OR MACHINE LEARNING) AND (SIZE COMPRES* OR SIZE REDUC*))", we have ensured the retrieval of relevant scientific articles that explicitly discuss scientifically compression methods applied to LLMs.

The rationale for using WoS extends beyond its exhaustive scope to its advanced search capabilities. The specific query structure uses truncation and Boolean operators, ensuring inclusivity by capturing all forms and variations of keywords related to LLMs and their compression, within the context of AI. By employing the TS field (Topic Search), the search identifies these terms across article titles, abstracts, and keywords, enhancing the likelihood of retrieving relevant research. The inclusion of multiple terms and logical connectors ensures that the search is neither too broad nor too narrow, focusing on specific scientific discussions around LLM compression.

Further filtering the search results to include only scientific research articles, while excluding review articles and conference proceedings, has been necessary for ensuring the relevance and rigor of the findings. Research articles present primary research, providing original data, insights, and methodologies that directly contribute to the understanding and advancement of model compression in LLMs. This type of content is foundational, offering empirical evidence that other forms of literature may not provide in its final finished form. Excluding review articles helps minimize the bias in the selection process, ensuring that the analysis relies on new and innovative research contributions.

Similarly, excluding conference proceedings was required as these papers often represent preliminary findings. While conference proceedings are valuable, they may lack the depth and methodological transparency of fully developed research articles. This further ensures that the collected scientific literature pool for our conducted study consists of in-depth, thoroughly vetted studies that meet high academic standards, providing a solid foundation for advancing knowledge in this domain.

3. Analysis of trends over time and the main research areas in the scientific literature regarding the compressing techniques of LLMs

The exploration of trends over time in the publication of scientific literature related to the compression of LLMs reveals valuable insights into the evolving interests and advancements within this important area of AI research.

By examining the distribution of publications from 2020 throughout to April 20, 2024, we can discern shifts in focus, emerging themes, and the overall growth of the field. There has been a significant rise in publications from 2020 through 2022, peaking in 2022 with 16 publications, a slight decrease in 2023 and an early count of 7 publications for 2024 as of 20-April-2024. This suggests a growing interest and development in the field of model compression techniques.

Starting with 2020, the scientific community began to place an increased emphasis on the scalability and efficiency of LLMs [5,23–25]. The year saw a modest number of publications, totaling four, which reflects the emerging stage of awareness and technological development concerning the compression of these complex models. As LLMs like GPT-3 and others began to demonstrate potent capabilities in various domains, ranging from NLP to automated content generation, the computational and environmental costs associated with these models started to draw significant attention (Figure 1).

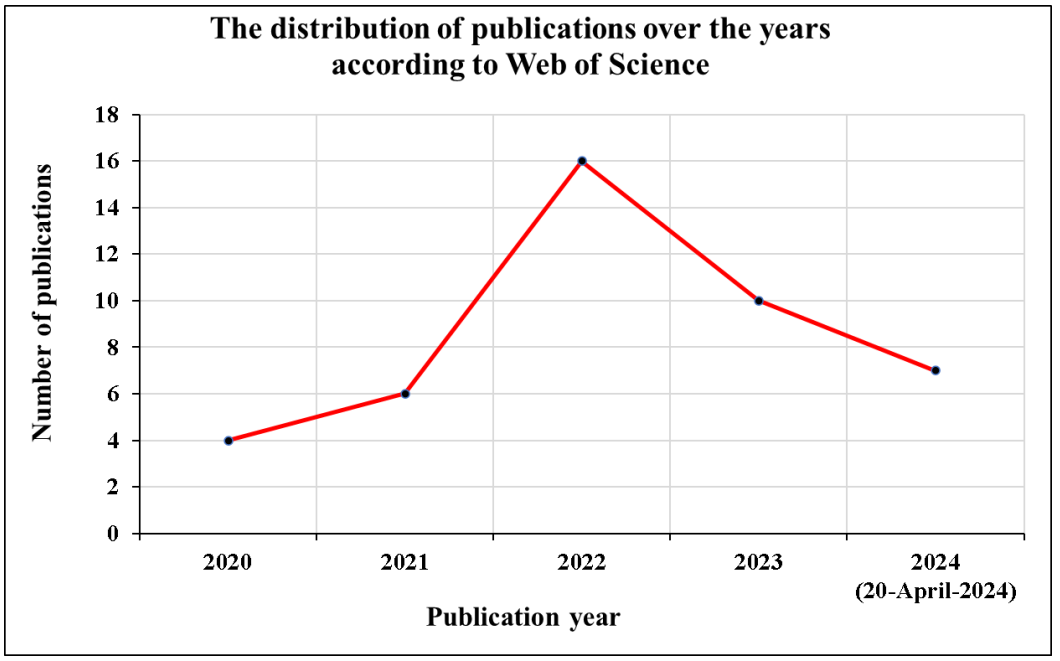

Figure 1. Exploration of trends over time in the publication of scientific literature related to the compression of LLMs[4]

In 2021, the publication count slightly increased to six [26–31]. This rise coincides with a broader comprehension within the AI research community and industry regarding the practical limitations imposed by the massive size and resource demands of state-of-the-art LLMs. The year 2021 saw enhancements in compression techniques such as knowledge distillation and quantization, tailored to mitigate these limitations. The slight increase in publications could be attributed to the consolidation of earlier findings and the initiation of more focused research projects aiming to refine and apply these emerging compression methods more effectively.

The year 2022 marked a significant peak with 16 publications, showcasing a robust interest and concerted effort in tackling the challenges associated with LLMs [2,4,9–11,15,32–41]. This surge can be interpreted as a response to the very important need for more sustainable AI practices, as the AI field grapples with the dual challenges of advancing technology and reducing its carbon footprint. During this period, research studies likely expanded into exploring the efficiency of individual compression techniques and especially their hybrid forms, such as integrating pruning with quantization, to achieve even greater reductions in model size and computational overhead.

In 2023, there was a notable decrease in publications to ten [12,13,18,42–48]. This drop might reflect a phase where the research community began to fit in the rapid advancements

[4] Source: The figure was devised based on the official data retrieved from Clarivate Web of Science in April 2024.

made in the previous years, shifting its focus from pioneering new methods to optimizing and validating existing techniques. It is also plausible that as some of the most accessible problems were being solved, the challenges became more complex, requiring longer cycles of research and development to achieve breakthroughs at the same pace as before.

Moving towards the partial data available for 2024, with seven publications recorded by April 20, the continuing interest in this area is clear, although with a publication rate that suggests a stabilization, or a slight decrease compared to the high mark in 2022 [1,3,21,22,49–51]. This trend could indicate several scenarios like a maturation of the field where major innovations become rarer and more incremental or a shift in focus towards other emerging areas of AI that require foundational research or simply the cyclical nature of research funding and publication outputs.

Overall, the trajectory of publications from 2020 to 2024 emphasizes a significant and growing recognition of the importance of developing effective compression techniques for LLMs. This trend is driven by the ongoing need to make AI technologies more accessible and sustainable, especially as these models find broader applications across industries and sectors. The data also suggests an increasing complexity in tackling the numerous challenges of compressing LLMs, reflecting deeper collaborations across computational and applied sciences.

As we look to the future, it is very important for the research community to continue promoting innovations in this space, particularly as the deployment environments for AI become more diverse and demanding. The adaptability of compression techniques to new model architectures and the integration of AI systems into edge devices and mobile platforms will likely be key areas of focus. Furthermore, as AI continues to integrate into more aspects of everyday life, ensuring the efficiency and sustainability of these systems will remain a major concern, driving ongoing research and interest in the field of LLM compression.

In the following, the research areas involved in the development and application of compression techniques for LLMs have been analyzed. The Computer Science field dominates the research areas with 41 publications, highlighting the central role of this field in the development and application of compression techniques for LLMs [2,5–13,18,21,23–27,29,31–33,36–38,41,42,45–47,49–60]. Engineering follows with 20 publications, indicating significant interdisciplinary work involving practical and technical aspects of implementing these techniques [1,4,6,10,18,27,30,32,33,35,42,43,47,49,51,52,56,60–62]. Physics [1,10,30,31,45,59], Telecommunications [6,27,32,33,47], and Materials Science [1,30,48] show fewer contributions, but emphasize the multi-disciplinary approach involving fundamental principles, data transmission, and possibly the materials used in computational hardware for AI. These findings suggest a robust and interdisciplinary effort in refining AI model efficiency, with a strong concentration in Computer Science. The trends reflect the academic and practical importance of this research area, along with

potential shifts in focus or emerging subfields within the broader AI research community (Figure 2).

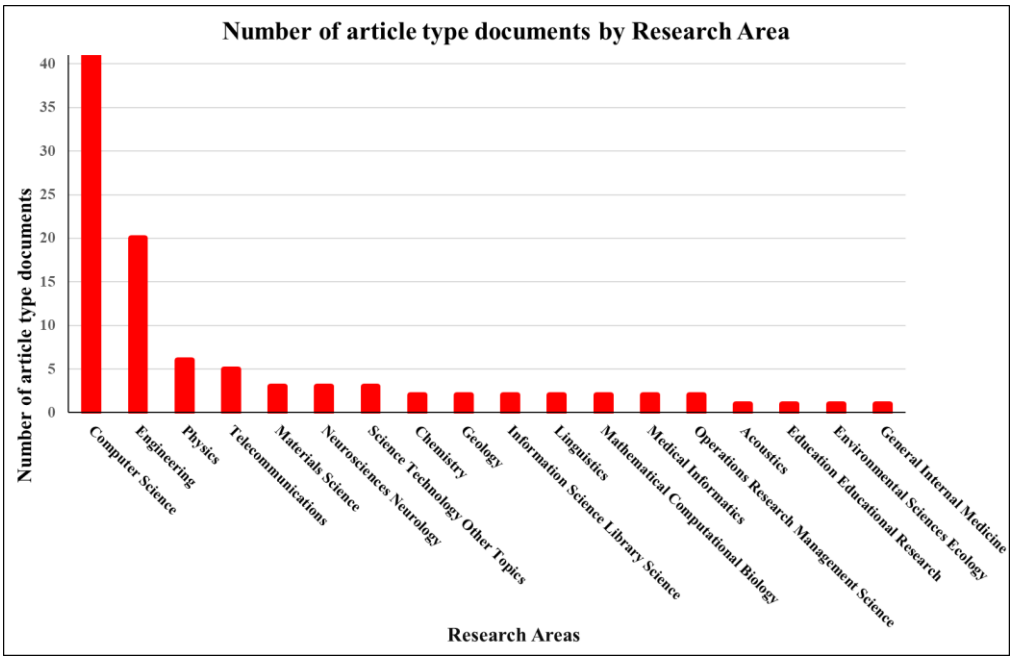

Figure 2. Research areas involved in the development and application of compression techniques for LLMs [5]

In order to obtain a comprehensive analysis of the correlation between different research areas involved in the development and application of compression techniques for LLMs, it is important to study the complex interactions and interdisciplinary efforts that characterize this field. This detailed exploration will highlight the predominant trends, collaborations across disciplines and highlight potential gaps along with opportunities for future research.

The rapid advancement of AI, particularly in the domain of LLMs such as GPT [1] and BERT [10,46,50], has required the exploration and implementation of various compression techniques to make these models more accessible and sustainable. The central challenge consists of reducing the computational and energy demands of these models without significantly compromising their performance. The interdisciplinary nature of this challenge has brought together experts from Computer Science, Engineering, Physics, Telecommunications, and Materials Science, each contributing unique perspectives and methodologies.

Computer Science is of extreme importance for research in AI model compression. It provides the theoretical frameworks, algorithms, and software implementations necessary

[5] Source: The figure was devised based on the official data retrieved from Clarivate Web of Science in April 2024.

for developing effective compression techniques such as pruning, quantization, and knowledge distillation. The field's dominance in the literature, accounting for over 68% of the publications, is indicative of its central role. Researchers in Computer Science work on algorithmic modifications and optimizations that can significantly reduce the size and computational complexity of LLMs. Pruning techniques are developed to remove redundant weights from neural networks, and quantization methods are applied to reduce the precision of the numerical values used in models, thereby decreasing the memory requirements and accelerating computation.

Engineering, with a substantial 33% of the publications, primarily focuses on the practical application and implementation of these compression techniques. Engineering research often bridges the gap between theoretical Computer Science models and real-world applications, addressing challenges related to hardware design, software-hardware integration, and the scalability of AI systems. Engineers work on adapting compression techniques to be compatible with existing and emerging hardware platforms, ensuring that compressed models are theoretically effective and practically viable. Engineering research might explore structured pruning techniques that are more amenable to conventional hardware architectures, therefore enhancing the efficiency of matrix operations that are extremely important for deploying models on general-purpose GPUs and on other accelerators.

Physics and Telecommunications contribute to a lesser extent, representing 10% and 8.33% of the publications, respectively. Nevertheless, their contributions are very important for understanding and improving the physical and network constraints associated with deploying AI models. Physics research might focus on the thermodynamic and quantum properties of materials used in hardware that supports AI computations, potentially leading to innovations in energy-efficient computing architectures. Meanwhile, Telecommunications research addresses the data transmission aspects, necessary for deploying AI models in distributed systems and for real-time applications such as automated content generation and real-time language translation.

Materials Science, though only accounting for 5% of the publications, plays a very important role in the development of new materials that can enhance the performance and efficiency of computational hardware. Research in this area might explore novel semiconductor materials or advanced manufacturing techniques that can be used to build more efficient GPUs and TPUs, which are needed for training and running large-scale AI models.

The correlation between these research areas can be seen in the collaborative efforts that aim to address the complex challenges posed by LLMs. The integration of Computer Science and Engineering is evident in the development of hardware-aware algorithms where compression techniques are tailored to the specific capabilities and limitations of the hardware used to run the models. Similarly, the collaboration between Materials Science

and Physics can lead to breakthroughs in hardware technology, such as the development of energy-efficient neural network processors that could further enhance the viability of compressed models. These interdisciplinary interactions promote innovation, ensure that the solutions developed are robust and applicable in a variety of settings. Advancements in Telecommunications research can enhance the deployment capabilities of AI models by improving the efficiency of data transfer across networks, which is very important for applications like cloud-based AI services and mobile AI applications [63].

Despite the robust collaborative efforts, there are gaps and challenges that need to be addressed to further advance the field of AI model compression. One of the significant challenges is the potential loss of accuracy and model generalizability due to compression. While compression techniques aim to minimize the impact on performance, there is often a trade-off between model size and its ability to perform complex tasks. Future research studies need to focus on developing compression techniques that can maintain high accuracy while achieving substantial reductions in model size and computational requirements.

Another area for potential improvement is the adaptability of compression techniques to new AI architectures and algorithms. As AI continues to evolve, with new models and approaches being developed at a rapid pace, compression techniques also need to be adaptable to these changes. This requires ongoing research and development to ensure that the techniques are effective for current models and also for future AI systems. The complexity of implementing these compression techniques also poses a significant challenge. Effective compression requires a deep understanding of both the architecture of the model and of the underlying algorithms. Customizing compression strategies to fit a specific model without losing essential functionalities demands extensive experimentation, complex engineering, and iterative tuning. This process is further complicated by the need for rigorous validation and testing to ensure that the compressed models perform adequately in real-world applications. Validation involves extensive testing against diverse datasets to identify any potential degradation in performance that may have been introduced during the compression process. This is extremely important for maintaining the trust and reliability of AI technologies in sensitive applications such as healthcare, finance, and autonomous driving.

Furthermore, the integration of emerging technologies such as federated learning and edge computing with model compression techniques could open new avenues for deploying AI in decentralized and privacy-preserving manners. These technologies allow for AI models to be trained and operated directly on user devices, reducing the need for data transmission and central processing, thereby enhancing user privacy and system efficiency. Conversely, this integration presents unique challenges, including the need for models that are compressed, robust enough to handle variable data environments and the computational capabilities of edge devices.

The field also faces ethical considerations, particularly in terms of bias and fairness. Compressed models, by necessity, simplify the representations learned by larger models, which could lead to the amplification of biases present in the training data. Ensuring that compression techniques do not exacerbate these biases requires careful attention to the design of both the model and the training process. This includes implementing strategies for bias detection and mitigation during both the training and compression phases.

Interdisciplinary collaboration will be of utmost importance for overcoming these challenges. As illustrated by the current distribution of research efforts, no single field can address all aspects of model compression alone. Collaborative projects that bring together experts from Computer Science, Engineering, Physics, Materials Science, and Telecommunications can leverage the strengths of each discipline to develop more comprehensive and effective solutions. Combining the theoretical insights from Computer Science with practical implementations from Engineering and cutting-edge materials from Physics can lead to the development of next-generation AI systems that are both powerful and efficient.

Future research studies should also focus on creating standardized frameworks and tools for implementing and evaluating model compression techniques. Such frameworks would help unify the efforts across different research areas and facilitate the sharing of best practices and benchmarks. This could accelerate the development of new compression methods and their adoption in industry and academia.

The compression of LLMs is a dynamic field that covers multiple disciplines, each contributing valuable and very important insights and technologies to address the challenges associated with these advanced AI systems. The ongoing collaboration and integration of diverse research areas are essential for advancing the state of the art in AI model compression. By continuing to encourage these interdisciplinary efforts and by addressing the technical, practical, and ethical challenges head-on, the research community can ensure that AI technologies become more sustainable, efficient, and accessible. This will expand the potential applications of AI and will also ensure that it is deployed in a manner that is beneficial and equitable for all sectors of society. The continual evolution of AI demands a proactive approach to research and development in model compression, making it an exciting and critical area of study in the expanding field of AI.

4. Challenges associated with the successful application of compression techniques to LLMs

In the following, we make an in-depth analysis into the main challenges associated with the successful application of compression techniques to LLMs. This analysis provides an overview of the complexities and main obstacles that researchers and practitioners face

when striving to reduce the computational and resource demands of these complex AI systems without compromising their performance.

One of the foremost challenges in compressing LLMs is maintaining a delicate balance between the model's size, speed, and accuracy. Compression techniques such as pruning, quantization, and knowledge distillation aim to reduce the physical size of the neural networks and the resources they require. Nevertheless, each method introduces potential trade-offs [24,42,49].

4.1. Pruning compression technique

Pruning is a compression technique applied to LLMs aimed at reducing the model's size and computational demands by eliminating redundant or non-critical parameters. It involves eliminating weights or neurons that contribute least to the model outputs, but it can also inadvertently remove elements that are necessary for certain tasks, leading to a loss in model generalizability or performance on specific benchmarks.

This technique is very important for making neural networks more efficient, especially in scenarios where computational resources are constrained, or costs need to be minimized. The concept of pruning stems from the observation that not all weights in a neural network contribute equally to its performance, suggesting that some can be removed with minimal impact on the model's efficacy.

Pruning can be broadly categorized into two types: structured and unstructured. Unstructured pruning involves the removal of individual weights across the network's matrices, leading to sparse connectivity between neurons. This type of pruning is highly flexible and can result in significant model size reduction. Nonetheless, it requires specialized software and hardware that can efficiently handle sparse matrices to realize computational speedups [64–66].

Structured pruning, on the other hand, removes entire rows, columns, or filters from matrices, leading to a reduced complexity in the network’s architecture. This form of pruning is more amenable to conventional hardware as it maintains the dense matrix structures necessary for optimized GPU utilization.

The implementation of pruning typically follows a three-step process:

I. Training: The neural network is first fully trained to learn the complex patterns in the data.

II. Removal (actual pruning): After training, weights that contribute the least to the output (often those with the smallest magnitudes) are identified and removed. This process can be iterative, involving re-training the network several times to refine which weights are pruned.

III. Fine-tuning: Once pruning is complete, the network undergoes additional training or fine-tuning with the remaining weights to recover any loss in performance due to the pruning process.

Various algorithms and criteria can be used to determine which weights to prune, such as magnitude-based pruning, where weights below a certain threshold are removed, and gradient-based pruning, which considers the sensitivity of the output to changes in each weight. One of the most evident benefits of pruning is the reduction in model size. By removing non-essential weights, the model becomes lighter, which saves storage space and also reduces the bandwidth needed for deploying the model in distributed systems or on edge devices.

Pruned models often require fewer computational resources. This is particularly beneficial in resource-limited environments where reducing the number of operations per inference can lead to faster response times and a lower power consumption. Structured pruning aligns well with existing hardware architectures, potentially increasing the efficiency of matrix operations. This can be particularly advantageous when deploying models on general-purpose GPUs or other accelerators that benefit from dense matrix operations.

The primary drawback of pruning is the potential reduction in model accuracy. Pruning important weights, even if they appear insignificant, can impair the model's ability to generalize from the training data to real-world applications. This requires a careful balance between the degree of pruning and the maintenance of model performance. Determining the optimal strategy for pruning is not a trivial aspect. It requires extensive experimentation with different pruning levels, methods, and fine-tuning cycles, which can be time-consuming and computationally expensive. Moreover, the criteria for pruning must be carefully chosen to avoid removing weights critical for certain tasks. The effectiveness of pruning is highly dependent on the quality of the initial model training. Poorly trained models might retain redundant weights while also lacking sufficient diversity in the weights that contribute to critical decision-making processes within the model.

Pruning presents a viable method for compressing LLMs by reducing unnecessary complexities and enhancing computational efficiency. While it offers considerable advantages in terms of model size and operational speed, it also poses challenges, including potential losses in accuracy and the complexity of its implementation. As the field of AI continues to evolve, further research and development are necessary to refine pruning techniques, ensuring they can reduce resource demands without significantly compromising the performance and adaptability of LLMs. This ongoing advancement will be extremely important in making AI technologies more accessible and sustainable, catering to a broader range of applications and environments.

4.2. Quantization compression technique

Quantization is a very important compression technique applied to LLMs to reduce the computational resources required for their operation while attempting to maintain acceptable levels of accuracy and performance. While this significantly decreases the size and increases the processing speed, it can lead to quantization errors, where the approximation of values causes a drop in accuracy. Quantization, in the context of ML and specifically in LLMs, refers to the process of reducing the precision of the numerical values used in a model. Traditionally, neural networks use floating-point arithmetic to perform calculations. These calculations, while accurate, are computationally expensive and resource intensive. Quantization addresses this aspect by approximating these floating-point numbers into lower-bit representations, typically using integer formats.

The basic principle behind quantization is the mapping of a continuous set of values (like those represented in floating-point) into a discrete set of values (like integers). This mapping reduces the memory requirements and speeds up the computation as integer operations are generally faster and more power-efficient on modern computing hardware than floating-point operations. Quantization can be broadly categorized [21,67,68] into three types:

I. Post-Training Quantization: This technique is applied after a model has been fully trained. The weights and activations, which are originally in floating-point, are converted into a lower-bit format. The main advantage of post-training quantization is its simplicity and ease of implementation as it does not require retraining the model.

II. Quantization-Aware Training (QAT): This method integrates quantization into the training process itself. By simulating the effects of quantization during training, QAT helps the model adjust its parameters to minimize the loss in accuracy that typically occurs when quantization is applied post-training.

III. Dynamic Quantization: This technique primarily quantizes the activations based on their distribution in real-time as they vary from one input to another. It is typically applied at the inference stage and is particularly useful for models where activation ranges can vary significantly.

Quantization incorporates various methodologies for mapping and representing values in a compact form. In uniform quantization, the range between the smallest and largest values is divided evenly, making this method straightforward and well-suited for hardware implementation due to its simplicity. Conversely, non-uniform quantization employs techniques such as logarithmic scaling, where the intervals between quantized values vary, enhancing fidelity particularly in regions near zero, where precision is often most important. Additionally, scalar and vector quantization approaches treat data differently, namely scalar quantization is processing each value independently, while vector quantization handles groups of values collectively, based on their overall distribution.

The application of quantization to LLMs brings significant advantages, primarily reducing the model's size through the use of fewer bits to represent each weight and activation, therefore diminishing the storage requirements. This compression also enables faster inference speeds as computations with lower-bit values are quicker on specialized hardware, a key aspect for applications demanding real-time processing. Furthermore, the reduced computational complexity leads to lower energy consumption, which is very important for models deployed on energy-constrained devices such as mobile phones and embedded systems.

One must take into account that quantization is not without its challenges. The reduction in bit precision can result in accuracy loss, particularly in complex tasks where high precision is essential. Moreover, some quantization techniques may depend on specific hardware capabilities to achieve computational benefits, limiting their usefulness in environments lacking such support. Implementing quantization also adds complexity, even if post-training quantization is relatively straightforward, quantization-aware training (QAT) necessitates adjustments to the training protocol, potentially introducing additional overhead in tuning and validation. Additionally, the effectiveness of quantization can vary significantly with different model architectures, requiring careful evaluation to understand its impact on each unique case.

Quantization presents a viable method for compressing LLMs, offering significant benefits in terms of reduced model size, increased processing speed, and enhanced energy efficiency. Nevertheless, the trade-offs in terms of potential accuracy loss and the complexities involved in its implementation must be carefully managed. Future research is needed to develop more advanced quantization techniques that can minimize accuracy loss while maximizing computational efficiency. This ongoing development is necessary for the broader adoption and application of LLMs in resource-constrained environments, contributing to the advancement of accessible and sustainable AI technologies. In the context of achieving LLM compression by means of quantization, it is important to highlight the extremely important balance between efficiency gains and potential pitfalls. As the field of AI continues to evolve, quantization will play an essential role in enabling the deployment of advanced neural networks in diverse and challenging real-world applications.

4.3. Knowledge Distillation compression technique

Knowledge distillation (KD) is a model compression technique that has gained significant attention in the field of AI, specifically in the context of LLMs. This technique involves transferring knowledge from a larger, often more cumbersome model (referred to as the "teacher") to a smaller, more efficient model (referred to as the "student"). The predominant goal of KD is to enable the student model to perform at par with the teacher model while requiring less computational power and memory, thereby making the deployment of AI

technologies more feasible in resource-constrained environments. The challenge consists in ensuring that the student model captures the nuanced understanding of the teacher model without needing the same computational resources. These techniques, while effective in reducing the size and computational load, must be applied judiciously to avoid undermining the model's ability to perform its intended tasks accurately.

KD operates on the premise that a complex model, which has been extensively trained and has a deep understanding of the data, can impart this knowledge to a simpler model. The process involves two main stages: the training of the teacher model and the distillation phase where the student learns from the teacher. The teacher model is typically a fully trained, high-capacity model that achieves high performance on the tasks for which it is designed. This model's depth and complexity allow it to capture subtle patterns in large datasets, making it an effective but resource-intensive solution.

During distillation, the student model is trained to predict the hard target labels of the training data and also to mimic the output distributions (soft targets) provided by the teacher model [69]. Soft targets are the probabilities or logits produced by the teacher for each class, which carry more information per example than hard labels. An example of this aspect is given by the fact that while a hard label might indicate the correct translation of a sentence, the soft targets could reveal how closely other potential translations compare, according to the teacher's understanding. The distillation loss, typically a form of cross-entropy between the soft targets of the teacher and the outputs of the student, guides the student training. This loss is often combined with the traditional hard target loss, balancing learning from the teacher and adhering to the ground truth [70].

A common method to enhance the effectiveness of KD is temperature scaling. This involves modifying the "softmax" function used during training by introducing a temperature parameter that controls the smoothness of the output probability distribution [71]. A higher temperature results in a softer probability distribution over classes, which provides more informative gradients for the student model during training. Beyond basic temperature scaling, custom strategies may involve adjusting the layers of the student that receive guidance from the teacher or altering the representation forms that the student should learn [69]. Some approaches focus on distilling intermediate representations (features) instead of just output probabilities [20].

One of the primary advantages of KD is the reduction in the size and computational requirements of the student model. This allows the deployment of complex AI models on devices with limited hardware capabilities, such as mobile phones and embedded devices. Despite its smaller size, a well-distilled student model can achieve performance close to that of the teacher model, making this technique particularly valuable for applications where performance cannot be compromised. By learning from the soft probabilities, student models often generalize better to new data compared to training from scratch or from hard

labels alone. This is because the soft labels encode additional information about the relationships between different classes [72].

The success of the student model heavily relies on the quality of the teacher model. A poorly trained teacher model can mislead the student, resulting in worse performance than a model trained directly from the data. The process of KD can be complex, involving careful tuning of the temperature parameter and of the distillation loss. Finding the right balance between learning from soft and hard targets requires extensive experimentation. Although the student model is lighter, the overall training time including the teacher's training can be substantial. Additionally, the resources required for training the teacher model are significant [73].

KD is particularly useful in scenarios where deploying large models is not feasible. It has been successfully applied in NLP tasks like machine translation [4–6], sentiment analysis [8,10], and question-answering [3] systems. Future research in KD is likely to focus on improving the efficiency of the distillation process, developing more robust student models that can outperform their teacher models, and extending the applicability of this technique to newer and more complex model architectures.

KD stands out as a promising technique for model compression. Its ability to transfer deep knowledge from large, resource-intensive models to more manageable counterparts without significant loss in performance is a major advantage. As AI technologies continue to evolve, optimizing and refining KD will be of extreme importance for advancing the practical deployment of AI systems, particularly in environments where resources are constrained.

The broader implications of KD extend beyond just model size and computational efficiency. By enabling powerful models to be compressed into more manageable forms, KD opens up new possibilities for AI applications in areas that were previously considered impractical due to hardware limitations. This includes real-time applications on mobile devices, such as live language translation and advanced on-device AI assistance, which can now benefit from deep learning insights without connectivity or high-power consumption. In addition, the democratization of access to advanced AI technologies through techniques like KD can help bridge the gap between well-funded, large-scale research institutions and smaller organizations or startups. This could offer equal opportunities and promote innovation across various sectors by making cutting-edge AI tools more accessible and less expensive to deploy.

One must also carefully take into consideration the challenges and limitations of KD. The reliance on a high-quality teacher model means that any inherent biases or errors in the teacher model are likely to be transferred to the student model. This could perpetuate or even amplify undesirable characteristics unless carefully managed. Additionally, the complexity of the distillation process itself may pose barriers to its widespread adoption, as it requires significant expertise and resources to implement effectively. Technical

challenges include the need for careful calibration of the distillation parameters, such as the temperature of the "softmax" function and the balance between different components of the loss function [73]. These parameters can significantly influence the effectiveness of the distillation and require detailed empirical evaluation to optimize.

Future studies might explore automated methods for optimizing these parameters or developing more adaptive distillation techniques that can dynamically adjust based on the student model's performance. Additionally, extending the concept of distillation beyond the teacher-student framework to include multiple teachers or collaborative distillation processes could offer new ways to enhance model performance and efficiency. As AI continues to advance, the role of KD is likely to grow, particularly in the development of AI models that are both powerful and practical for everyday applications.

The exploration of hybrid models that combine KD with other compression techniques such as pruning and quantization could yield even more efficient and robust AI systems. Furthermore, integrating KD into the lifecycle of AI development, from training through deployment, could help in continuously refining models in a resource-efficient manner. The integration of KD with emerging technologies like federated learning, where models are trained across multiple decentralized devices while keeping all the training data local, could further enhance privacy and scalability. This represents a significant step forward in creating AI systems that are both powerful and privacy-preserving.

KD is a powerful tool for model compression that offers significant benefits, including reduced model size, retained performance, and enhanced generalization. Nonetheless, its successful implementation requires careful consideration of various technical and ethical factors. Continued research and development in this area are essential to fully attain its potential and address the ongoing challenges. As the field progresses, KD will play an increasingly important role in making advanced AI technologies more accessible and sustainable, contributing significantly to the advancement of both the science and application of AI.

4.4. Complexity of implementation and trade-offs in model usefulness

The process of implementing compression techniques is inherently complex. It requires a deep understanding of both the architecture of the model and the underlying algorithms. Each LLM has unique characteristics based on its training data, structure, and intended use cases. Customizing compression strategies to fit a specific model without losing essential functionalities demands extensive experimentation, complex engineering, and iterative tuning.

The appropriate level of pruning or the best quantization scheme can vary widely between models. Developers must conduct numerous trials to identify the most effective parameters,

which can be a time-consuming and resource-intensive process. Additionally, each iteration must be rigorously tested to ensure that the compressed model still meets the performance criteria necessary for its application.

Compressed models often face significant trade-offs in their usefulness. A smaller model size generally leads to faster computation times and lower energy usage, which are beneficial for deploying models on edge devices or in environments where computing resources are limited. Nevertheless, these advantages may come at the cost of reduced accuracy or a diminished ability to generalize across different tasks and datasets. A model compressed for efficient translation might struggle with the subtleties of the language that are extremely important for tasks like sentiment analysis or legal document interpretation. Ensuring that a model remains versatile across various applications while being compressed is a significant challenge.

The endeavor to compress LLMs involves an array of complex procedures that require an in-depth comprehension of model architectures and underlying algorithms while also demanding careful customization to harmonize with the model's intrinsic characteristics without undermining its core functionalities. This complexity emerges from the inherent diversity and characteristics of LLMs, which are shaped by their training data, structural configurations, and the specific applications for which they are intended. As such, the implementation of compression techniques is a complex challenge that requires precise engineering, extensive experimentation, and iterative refinement to ensure that the integrity and effectiveness of the model are preserved.

The first step in the compression of a LLM is a thorough understanding of its architecture. LLMs, such as GPT and BERT, are built on complex neural network architectures that involve multiple layers of processing units, each responsible for understanding different aspects of the input data. These models employ mechanisms like attention and transformer architectures, which allow them to handle vast amounts of data and capture complex patterns in language.

Understanding these mechanisms is necessary for effective compression because each component of the architecture plays a specific role in the model's learning and inference processes. The attention mechanism in transformers facilitates the model's ability to focus on relevant parts of the input data, enhancing its understanding and generation of language. Compressing such a model without a detailed understanding of these components could lead to significant losses in functionality and performance, as critical aspects of the model’s capability to process and generate language might be inadvertently diminished.

Given the unique characteristics of each LLM, developing a one-size-fits-all compression approach is impractical. Instead, compression strategies must be tailored to fit specific models. This customization involves modifying existing compression techniques, such as

pruning, quantization, and KD, to align with the model's architecture and the requirements of its application domain.

The choice between structured and unstructured pruning depends on the specific model and the computational resources available. Structured pruning, which involves removing entire neurons or layers, may be suitable for models where computational efficiency is a priority and can be aligned with hardware that benefits from dense matrix operations. Conversely, unstructured pruning, which targets individual weights for removal, might be preferred when minimal impact on model performance is most important, and the available hardware can efficiently handle sparse matrices.

The process of implementing compression techniques is inherently iterative. It often requires multiple cycles of compression, testing, and tuning to find the optimal balance between model size, speed, and accuracy. Each iteration involves applying a compression technique, evaluating the model's performance on a set of benchmarks, and adjusting the parameters of the compression technique based on the outcomes.

This iterative process is very important because it allows for the gradual refinement of the compression strategy, minimizing the risk of degrading the model's performance. An example of this aspect is given by the fact that during the pruning process, an initial round of weight removal might show minimal impact on performance. Subsequent rounds might gradually increase the amount of pruning, with continuous monitoring to ensure that the performance does not fall below acceptable thresholds.

Implementing compression techniques requires advanced engineering skills and effective management of computational resources. Engineers must be adept at both software development and ML, with a deep understanding of how changes to the model’s architecture and parameters affect its behavior. This dual expertise is necessary to modify the model efficiently and to implement the compression techniques without introducing bugs or errors that could lead to unexpected behavior.

Additionally, managing computational resources effectively is extremely important, especially in environments where these resources are limited. Compression techniques can reduce the computational load of LLMs, making it feasible to deploy them on lower-end hardware or in resource-constrained environments. Nevertheless, achieving these reductions without excessive expenditure on computational resources during the compression process itself requires careful planning and management.

Despite the progress in model compression techniques, numerous challenges remain. The complexity of implementation is not merely a technical barrier but also a strategic one, involving decisions about which aspects of a model are essential for its intended use and how those aspects can be preserved during compression. Future studies in this area are likely to focus on developing more automated and adaptive compression techniques that can dynamically adjust to the model's performance during the compression process.

Advances in ML, such as reinforcement learning and meta-learning, could potentially be harnessed to automate the selection and tuning of compression parameters, reducing the need for manual intervention and making the process more efficient.

In addition, as AI and ML continue to evolve, the adaptability of compression techniques to new model architectures and training paradigms will be determinant factors. This includes the ability to compress models that use emerging techniques such as few-shot learning, unsupervised learning, or transfer learning, in situations where traditional compression approaches may not be directly applicable. The compression of LLMs represents a significant technical endeavor that holds the potential to transform the scalability and applicability of AI technologies, particularly in environments constrained by computational resources or hardware capabilities. The intricacies involved in achieving effective compression without compromising the functional integrity of these models require a disciplined approach to understanding, customizing, and iteratively refining the technologies used.

4.5. Validation and testing

An important part of implementing compression techniques consists in the validation phase, where the performance of the compressed model is rigorously assessed against benchmarks that are representative of real-world tasks. This validation is more than a one-time event, is a continuous process that ensures the model remains robust and performs well under different conditions and datasets. In order to attain an effective validation, diverse datasets that cover possible scenarios that the model may encounter in practical applications are used. This extensive testing helps identify any potential degradation in performance or in functionality that might have been introduced during the compression process. Additionally, it provides insights into how the model performs on various tasks, which is very important for understanding the trade-offs made during compression .

Researchers must consider that as new types of neural network architectures are developed, they may introduce different characteristics or sensitivities that need to be considered in the compression process. Architectures that use mechanisms beyond attention, such as those incorporating dynamic neural networks or capsule networks, often require novel approaches to compression that can accommodate their unique properties [74–77]. The complexity of compressing LLMs also necessitates collaboration across various disciplines within both academia and industry. This multidisciplinary approach brings together expertise from areas such as ML, software engineering, hardware design, and application development. Such collaborations can accelerate the refinement of compression techniques and help bridge the gap between theoretical research and practical application.

Collaborative efforts can also facilitate the sharing of best practices, tools, and resources, making it easier for smaller organizations or individuals to adopt and benefit from

compressed models. By democratizing access to advanced AI technologies, the broader AI community can drive innovation and application across a wider array of sectors. As compression techniques become more advanced and widely implemented, it is necessary to consider the ethical and social implications of deploying compressed models. These considerations include the transparency of model behaviors, the fairness of their outputs, and their accessibility to various user groups. Addressing these issues requires careful design of the compression process to ensure that it does not inadvertently introduce biases or reduce the model's ability to handle diverse data inputs fairly. It also involves developing guidelines and standards for the responsible use of compressed models, particularly in sensitive areas such as healthcare, law enforcement, and financial services.

The compression of LLMs is a complex, dynamic, and critically important area of research within AI. The successful implementation of compression techniques requires a deep understanding of the underlying technologies, a commitment to rigorous testing and validation, and a proactive approach to adapting these methods to new developments in the field. As these techniques continue to evolve, they promise to make AI more accessible and sustainable, thereby expanding the potential for these technologies to benefit society. The ongoing research, collaboration, and ethical consideration will be of extreme importance for obtaining these benefits while mitigating the risks associated with AI deployment in diverse environments.

5. Conclusions

The field of AI and ML is rapidly evolving, with new models and techniques being developed continuously. Ensuring that compression techniques remain effective as models evolve is a significant challenge. Compressed models must be robust, effective with current technologies and also adaptable to future developments. Researchers and developers must anticipate changes in hardware capabilities, data availability, and model architectures. This requires a forward-thinking approach to compression, where techniques are designed to optimize current models and adapt to next-generation AI technologies.

The compression of LLMs presents a complex array of challenges that cover technical, practical, and strategic dimensions. Balancing performance trade-offs, managing complex implementation processes, ensuring robust validation, and future-proofing technologies are all critical to the success of these initiatives. As the field progresses, developing more advanced, efficient, and adaptable compression techniques will be essential for making AI technologies more sustainable and accessible to a broader range of users and applications. This ongoing effort will require an intensive collaboration between researchers, engineers, and industry stakeholders in order to overcome these challenges and harness the full potential of compressed LLMs.

Acknowledgment

"The authors would like to express their gratitude for the logistics support received from the Center of Research, Consultancy and Training in Economic Informatics and Information Technology RAU-INFORTIS of the Romanian-American University and the Center for Computational Science and Machine Intelligence of the Romanian-American University."

References

[1] Tuan, N.T.; Moore, P.; Thanh, D.H. V; Pham, H. V A Generative Artificial Intelligence Using Multilingual Large Language Models for ChatGPT Applications. APPLIED SCIENCES-BASEL 2024, 14, doi:10.3390/app14073036.

[2] Lin, J.F.; Liu, Y.L.; Cleland-Huang, J. Information Retrieval versus Deep Learning Approaches for Generating Traceability Links in Bilingual Projects. Empir Softw Eng 2022, 27, doi:10.1007/s10664-021-10050-0.

[3] Zou, Z.; Mubin, O.; Alnajjar, F.; Ali, L. A Pilot Study of Measuring Emotional Response and Perception of LLM-Generated Questionnaire and Human-Generated Questionnaires. Sci Rep 2024, 14, doi:10.1038/s41598-024-53255-1.

[4] An, B.; Long, C.J. Paraphrase Based Data Augmentation For Chinese-English Medical Machine Translation. JOURNAL OF ELECTRONICS & INFORMATION TECHNOLOGY 2022, 44, 118–126, doi:10.11999/JEIT210926.

[5] Vázquez, R.; Raganato, A.; Creutz, M.; Tiedemann, J. A Systematic Study of Inner-Attention-Based Sentence Representations in Multilingual Neural Machine Translation. COMPUTATIONAL LINGUISTICS 2020, 46, 387–424, doi:10.1162/coli_a_00377.

[6] Zhu, X.N.; Yang, M.Y.; Zhao, T.J.; Zhu, C.H. Minimum Bayes-Risk Phrase Table Pruning for Pivot-Based Machine Translation in Internet of Things. IEEE ACCESS 2018, 6, 55754–55764, doi:10.1109/ACCESS.2018.2872773.

[7] Kim, K.; Park, E.J.; Shin, J.H.; Kwon, O.W.; Kim, Y.K. Divergence-Based Fine Pruning of Phrase-Based Statistical Translation Model. Comput Speech Lang 2017, 41, 146–160, doi:10.1016/j.csl.2016.06.006.

[8] Boiy, E.; Moens, M.F. A Machine Learning Approach to Sentiment Analysis in Multilingual Web Texts. Inf Retr Boston 2009, 12, 526–558, doi:10.1007/s10791-008-9070-z.

[9] Karthik, E.; Sethukarasi, T. A Centered Convolutional Restricted Boltzmann Machine Optimized by Hybrid Atom Search Arithmetic Optimization Algorithm for Sentimental Analysis. Neural Process Lett 2022, 54, 4123–4151, doi:10.1007/s11063-022-10797-7.

[10] Catelli, R.; Pelosi, S.; Esposito, M. Lexicon-Based vs. Bert-Based Sentiment Analysis: A Comparative Study in Italian. Electronics (Basel) 2022, 11, doi:10.3390/electronics11030374.

[11] Nedjah, N.; Santos, I.; Mourelle, L.D. Sentiment Analysis Using Convolutional Neural Network via Word Embeddings. Evol Intell 2022, 15, 2295–2319, doi:10.1007/s12065-019-00227-4.

[12] Buonocore, T.M.; Crema, C.; Redolfi, A.; Bellazzi, R.; Parimbelli, E. Localizing In-Domain Adaptation of Transformer-Based Biomedical Language Models. J Biomed Inform 2023, 144, doi:10.1016/j.jbi.2023.104431.

[13] Tan, R.; Lin, Q.; Low, G.H.; Lin, R.X.; Goh, T.C.; Chang, C.C.E.; Lee, F.F.; Chan, W.Y.; Tan, W.C.; Tey, H.J.; et al. Inferring Cancer Disease Response from Radiology Reports Using Large Language Models with Data Augmentation and Prompting. JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION 2023, 30, 1657–1664, doi:10.1093/jamia/ocad133.

[14] Gronsbell, J.; Minnier, J.; Yu, S.; Liao, K.; Cai, T.X. Automated Feature Selection of Predictors in Electronic Medical Records Data. Biometrics 2019, 75, 268–277, doi:10.1111/biom.12987.

[15] Shoulson, I.; Arbatti, L.; Hosamath, A.; Eberly, S.W.; Oakes, D. Longitudinal Cohort Study of Verbatim-Reported Postural Instability Symptoms as Outcomes for Online Parkinson's Disease Trials. J Parkinsons Dis 2022, 12, 1969–1978, doi:10.3233/JPD-223274.

[16] Scheffer, L.K. Analysis Tools for Large Connectomes. Front Neural Circuits 2018, 12, doi:10.3389/fncir.2018.00085.

[17] Brenowitz, N.D.; Bretherton, C.S. Prognostic Validation of a Neural Network Unified Physics Parameterization. Geophys Res Lett 2018, 45, 6289–6298, doi:10.1029/2018GL078510.

[18] Abdelhakim, M.; Liu, B.Q.; Sun, C.G. Ar-PuFi: A Short-Text Dataset to Identify the Offensive Messages towards Public Figures in the Arabian Community. Expert Syst Appl 2023, 233, doi:10.1016/j.eswa.2023.120888.

[19] Petroșanu, D.M.; Pîrjan, A.; Tăbușcă, A. Tracing the Influence of Large Language Models across the Most Impactful Scientific Works. Electronics (Switzerland) 2023, 12.

[20] Kim, M.-Y.; Rabelo, J.; Babiker, H.K.B.; Rahman, M.A.; Goebel, R. Legal Information Retrieval and Entailment Using Transformer-Based Approaches. The Review of Socionetwork Strategies 2024, doi:10.1007/s12626-023-00153-z.

[21] Cho, M.S.; Vahid, K.A.; Fu, Q.C.; Adya, S.; Mundo, C.C.D.; Rastegari, M.; Naik, D.; Zatloukal, P. EDKM: An Efficient and Accurate Train-Time Weight Clustering for Large

Language Models. IEEE COMPUTER ARCHITECTURE LETTERS 2024, 23, 37–40, doi:10.1109/LCA.2024.3363492.

[22] Schonfeld, E.; Pant, A.; Shah, A.R.Y.; Sadeghzadeh, S.; Pangal, D.; Rodrigues, A.; Yoo, K.; Marianayagam, N.; Haider, G.; Veeravagu, A. Evaluating Computer Vision, Large Language, and Genome-Wide Association Models in a Limited Sized Patient Cohort for Pre-Operative Risk Stratification in Adult Spinal Deformity Surgery. J Clin Med 2024, 13, doi:10.3390/jcm13030656.

[23] Dong, L.B.; Guo, Q.M.; Wu, W.L.; Satpute, M.N. A Semantic Relatedness Preserved Subset Extraction Method for Language Corpora Based on Pseudo-Boolean Optimization. Theor Comput Sci 2020, 836, 65–75, doi:10.1016/j.tcs.2020.07.020.

[24] Choudhary, T.; Mishra, V.; Goswami, A.; Sarangapani, J. A Comprehensive Survey on Model Compression and Acceleration. Artif Intell Rev 2020, 53, 5113–5155, doi:10.1007/s10462-020-09816-7.

[25] Holtzen, S.; Van den Broeck, G.; Millstein, T. Scaling Exact Inference for Discrete Probabilistic Programs. PROCEEDINGS OF THE ACM ON PROGRAMMING LANGUAGES-PACMPL 2020, 4, doi:10.1145/3428208.

[26] Rajpal, D.; Garg, A.R.; Mahela, O.P.; Alhelou, H.H.; Siano, P. A Fusion-Based Hybrid-Feature Approach for Recognition of Unconstrained Offline Handwritten Hindi Characters. Future Internet 2021, 13, doi:10.3390/fi13090239.

[27] Kumar, V.; Recupero, D.R.; Riboni, D.; Helaoui, R. Ensembling Classical Machine Learning and Deep Learning Approaches for Morbidity Identification From Clinical Notes. IEEE ACCESS 2021, 9, 7107–7126, doi:10.1109/ACCESS.2020.3043221.

[28] McNorgan, C. The Connectivity Fingerprints of Highly-Skilled and Disordered Reading Persist Across Cognitive Domains. Front Comput Neurosci 2021, 15, doi:10.3389/fncom.2021.590093.

[29] Eronen, J.; Ptaszynski, M.; Masui, F.; Smywinski-Pohl, A.; Leliwa, G.; Wroczynski, M. Improving Classifier Training Efficiency for Automatic Cyberbullying Detection with Feature Density. Inf Process Manag 2021, 58, doi:10.1016/j.ipm.2021.102616.

[30] Li, X.; Luo, D.H.; Cheng, Y.; Wong, K.Y.; Hung, K. Identifying the Primary Odor Perception Descriptors by Multi-Output Linear Regression Models. APPLIED SCIENCES-BASEL 2021, 11, doi:10.3390/app11083320.

[31] Hoppe, N.; Winter, J.M.; Adami, S.; Adams, N.A. ALPACA - a Level-Set Based Sharp-Interface Multiresolution Solver for Conservation Laws. Comput Phys Commun 2022, 272, doi:10.1016/j.cpc.2021.108246.

[32] Wang, H.Z.; Qu, Z.H.; Zhou, Q.H.; Zhang, H.B.; Luo, B.Y.; Xu, W.C.; Guo, S.; Li, R.X. A Comprehensive Survey on Training Acceleration for Large Machine Learning Models in IoT. IEEE Internet Things J 2022, 9, 939–963, doi:10.1109/JIOT.2021.3111624.

[33] Bae, J.; Cheon, B.D.; Kim, H.Y. Pro-Attention: Efficient Probability Distribution Matching-Based Attention Through Feature Space Conversion. IEEE ACCESS 2022, 10, 131192–131201, doi:10.1109/ACCESS.2022.3229055.

[34] Chen, Y.C. Effects of Technology-Enhanced Language Learning on Reducing EFL Learners' Public Speaking Anxiety. Comput Assist Lang Learn 2022, doi:10.1080/09588221.2022.2055083.

[35] Naser, M.Z. Do We Need Exotic Models? Engineering Metrics to Enable Green Machine Learning from Tackling Accuracy-Energy Trade-Offs. J Clean Prod 2023, 382, doi:10.1016/j.jclepro.2022.135334.

[36] Pinto, J.P.; Viana, P.; Teixeira, I.; Andrade, M. Improving Word Embeddings in Portuguese: Increasing Accuracy While Reducing the Size of the Corpus. PeerJ Comput Sci 2022, 8, doi:10.7717/peerj-cs.964.

[37] Tan, J.H.; Tan, Y.H.; Chan, C.S.; Chuah, J.H. ACORT: A Compact Object Relation Transformer for Parameter Efficient Image Captioning. Neurocomputing 2022, 482, 60–72, doi:10.1016/j.neucom.2022.01.081.

[38] Tang, X.Y.; Zeng, S.; Yu, F.; Yu, W.; Sheng, Z.Y.; Kang, Z. Self-Supervised Anomaly Pattern Detection for Large Scale Industrial Data. Neurocomputing 2023, 515, 1–12, doi:10.1016/j.neucom.2022.09.069.

[39] Chen, X.; Wang, J.; Gomes, J.; Dubovik, O.; Yang, P.; Saito, M. Analytical Prediction of Scattering Properties of Spheroidal Dust Particles With Machine Learning. Geophys Res Lett 2022, 49, doi:10.1029/2021GL097548.

[40] Romero, R.W.A.; Viallon, M.; Spaltenstein, J.; Petrusca, L.; Bernard, O.; Belle, L.; Clarysse, P.; Croisille, P. CMRSegTools: An Open-Source Software Enabling Reproducible Research in Segmentation of Acute Myocardial Infarct in CMR Images. PLoS One 2022, 17, doi:10.1371/journal.pone.0274491.

[41] Ali, A.; Pinciroli, R.; Yan, F.; Smirni, E. Optimizing Inference Serving on Serverless Platforms. PROCEEDINGS OF THE VLDB ENDOWMENT 2022, 15, 2071–2084, doi:10.14778/3547305.3547313.

[42] Kashikar, P.; Sentieys, O.; Sinha, S. Lossless Neural Network Model Compression Through Exponent Sharing. IEEE Trans Very Large Scale Integr VLSI Syst 2023, 31, 1816–1825, doi:10.1109/TVLSI.2023.3307607.

[43] Zhang, J.F.; Zhang, Z.Y. Machine Learning Hardware Design for Efficiency, Flexibility, and Scalability [Feature]. IEEE CIRCUITS AND SYSTEMS MAGAZINE 2023, 23, 35–53, doi:10.1109/MCAS.2023.3302390.

[44] Woods, L.T.; Rana, Z.A. Constraints on Optimising Encoder-Only Transformers for Modelling Sign Language with Human Pose Estimation Keypoint Data. J Imaging 2023, 9, doi:10.3390/jimaging9110238.

[45] Bati, A.; Bryngelson, S.H. RoseNNa: A Performant, Portable Library for Neural Network Inference with Application to Computational Fluid Dynamics. Comput Phys Commun 2024, 296, doi:10.1016/j.cpc.2023.109052.

[46] Bhanage, D.A.; Pawar, A. V Robust Analysis of IT Infrastructure's Log Data with BERT Language Model. INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS 2023, 14, 705–714.

[47] Li, D.C.; Maulana, M.R.; Chou, L.D. NNSplit-SOREN: Supporting the Model Implementation of Large Neural Networks in a Programmable Data Plane. COMPUTER NETWORKS 2023, 222, doi:10.1016/j.comnet.2022.109537.

[48] Tüzemen, E.S.; Yüksek, A.G.; Demir, I.; Horoz, S.; Altuntas, I. Modeling of Temperature-Dependent Photoluminescence of GaN Epilayer by Artificial Neural Network. JOURNAL OF THE AUSTRALIAN CERAMIC SOCIETY 2023, 59, 1145–1159, doi:10.1007/s41779-023-00911-w.

[49] Zhu, J.; Wang, L.Y.; Han, X.; Liu, A.M.; Xie, T. Safety and Performance, Why Not Both? Bi-Objective Optimized Model Compression Against Heterogeneous Attacks Toward AI Software Deployment. IEEE TRANSACTIONS ON SOFTWARE ENGINEERING 2024, 50, 376–390, doi:10.1109/TSE.2023.3348515.

[50] Pujar, S.; Zheng, Y.H.; Buratti, L.; Lewis, B.; Chen, Y.C.; Laredo, J.; Morari, A.; Epstein, E.; Lin, T.N.; Yang, B.; et al. Analyzing Source Code Vulnerabilities in the D2A Dataset with ML Ensembles and C-BERT. Empir Softw Eng 2024, 29, doi:10.1007/s10664-023-10405-9.

[51] Wu, R.H.; Zhu, X.Y.; Chen, J.S.; Liu, S.; Zheng, T.Y.; Liu, X.; An, H. SWattention: Designing Fast and Memory-Efficient Attention for a New Sunway Supercomputer. JOURNAL OF SUPERCOMPUTING 2024, doi:10.1007/s11227-024-05890-8.

[52] Chung, J.; Shin, T.; Yang, J.S. Simplifying Deep Neural Networks for FPGA-Like Neuromorphic Systems. IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS 2019, 38, 2032–2042, doi:10.1109/TCAD.2018.2877016.

[53] Huang, S.P.; Jin, L.W.; Xue, K.N.; Fang, Y. Online Primal-Dual Learning for a Data-Dependent Multi-Kernel Combination Model with Multiclass Visual Categorization Applications. Inf Sci (N Y) 2015, 320, 75–100, doi:10.1016/j.ins.2015.05.009.

[54] Ablimit, M.; Kawahara, T.; Hamdulla, A. Lexicon Optimization Based on Discriminative Learning for Automatic Speech Recognition of Agglutinative Language. Speech Commun 2014, 60, 78–87, doi:10.1016/j.specom.2013.09.011.

[55] Zahera, H.M.; El-Sisi, A.B. Accelerating Training Process in Logistic Regression Model Using OpenCL Framework. INTERNATIONAL JOURNAL OF GRID AND HIGH PERFORMANCE COMPUTING 2017, 9, 34–45, doi:10.4018/IJGHPC.2017070103.

[56] Tiwari, A.K.; Shreevastava, S.; Som, T.; Shukla, K.K. Tolerance-Based Intuitionistic Fuzzy-Rough Set Approach for Attribute Reduction. Expert Syst Appl 2018, 101, 205–212, doi:10.1016/j.eswa.2018.02.009.

[57] Manojlovic, I.; Svenda, G.; Erdeljan, A.; Gavric, M. Time Series Grouping Algorithm for Load Pattern Recognition. Comput Ind 2019, 111, 140–147, doi:10.1016/j.compind.2019.07.009.

[58] Nguyen, P.T.; Le, A.C.; Ho, T.B.; Nguyen, V.H. Vietnamese Treebank Construction and Entropy-Based Error Detection. Lang Resour Eval 2015, 49, 487–519, doi:10.1007/s10579-015-9308-5.

[59] Antcheva, I.; Ballintijn, M.; Bellenot, B.; Biskup, M.; Brun, R.; Buncic, N.; Canal, P.; Casadei, D.; Couet, O.; Fine, V.; et al. ROOT - A C++ Framework for Petabyte Data Storage, Statistical Analysis and Visualization. Comput Phys Commun 2009, 180, 2499–2512, doi:10.1016/j.cpc.2009.08.005.

[60] Ittiphalin, M.; Arnonkijpanich, B.; Pathumnakul, S. An Artificial Intelligence Model to Estimate the Fat Addition Ratio for the Mixing Process in the Animal Feed Industry. J Intell Manuf 2017, 28, 219–228, doi:10.1007/s10845-014-0972-x.

[61] Tapiador-Morales, R.; Linares-Barranco, A.; Jimenez-Fernandez, A.; Jimenez-Moreno, G. Neuromorphic LIF Row-by-Row Multiconvolution Processor for FPGA. IEEE Trans Biomed Circuits Syst 2019, 13, 159–169, doi:10.1109/TBCAS.2018.2880012.

[62] Abed, S.; Khalil, Y.; Modhaffar, M.; Ahmad, I. High-Performance Low-Power Approximate Wallace Tree Multiplier. INTERNATIONAL JOURNAL OF CIRCUIT THEORY AND APPLICATIONS 2018, 46, 2334–2348, doi:10.1002/cta.2540.

[63] Tăbuscă, A.; Tăbuscă, S.-M. Impact of 5G Technology in Global Economy. Cybersecurity and Legal Issues. Journal of Information Systems & Operations Management 2020, 13, 177–189.

[64] Moayed, H.; Mansoori, E.G.; Moosavi, M.R. An Efficient Pruning Method for Subgraph Matching in Large-Scale Graphs. Journal of Supercomputing 2023, 79, 10511–10532, doi:10.1007/s11227-023-05061-1.

[65] Cinquin, O. ChIP-GPT: A Managed Large Language Model for Robust Data Extraction from Biomedical Database Records. Brief Bioinform 2024, 25, doi:10.1093/bib/bbad535.

[66] Yan, H.; Liu, Y.; Jin, L.; Bai, X. The Development, application, and Future of LLM Similar to ChatGPT. Journal of Image and Graphics 2023, 28, 2749–2762, doi:10.11834/jig.230536.

[67] Li, W.; Hu, A.; Xu, N.; He, G. Quantization and Hardware Architecture Co-Design for Matrix-Vector Multiplications of Large Language Models. IEEE Transactions on Circuits and Systems I: Regular Papers 2024, doi:10.1109/TCSI.2024.3350661.

[68] Luo, Y.; Wei, Z.; Xu, G.; Li, Z.; Xie, Y.; Yin, Y. Enhancing E-Commerce Chatbots with Falcon-7B and 16-Bit Full Quantization. Journal of Theory and Practice of Engineering Science 2024, 4, 52–57, doi:10.53469/jtpes.2024.04(02).08.

[69] Tan, C.; Liu, J. Improving Knowledge Distillation With a Customized Teacher. IEEE Trans Neural Netw Learn Syst 2024, 35, 2290–2299, doi:10.1109/TNNLS.2022.3189680.

[70] Jiang, Y.; Weng, J.; Zhang, X.; Yang, Z.; Hu, W. A CNN-Based Born-Again TSK Fuzzy Classifier Integrating Soft Label Information and Knowledge Distillation. IEEE Transactions on Fuzzy Systems 2023, 31, 1843–1854, doi:10.1109/TFUZZ.2022.3215566.

[71] Yang, Z.; Zhang, Y.; Sui, D.; Ju, Y.; Zhao, J.; Liu, K. Explanation Guided Knowledge Distillation for Pre-Trained Language Model Compression. ACM Transactions on Asian and Low-Resource Language Information Processing 2024, 23, doi:10.1145/3639364.

[72] Du, Y.; Niu, J.; Wang, Y.; Jin, X. Multi-Stage Knowledge Distillation for Sequential Recommendation with Interest Knowledge. Inf Sci (N Y) 2024, 654, doi:10.1016/j.ins.2023.119841.

[73] Rao, J.; Meng, X.; Ding, L.; Qi, S.; Liu, X.; Zhang, M.; Tao, D. Parameter-Efficient and Student-Friendly Knowledge Distillation. IEEE Trans Multimedia 2024, 26, 4230–4241, doi:10.1109/TMM.2023.3321480.

[74] Arnes, J.I.; Horsch, A. Schema-Based Priming of Large Language Model for Data Object Validation Compliance. SSRN Electronic Journal 2023, doi:10.2139/ssrn.4453361.

[75] Sandholm, T.E.; Mukherjee, S.; Huberman, B.A. Randomness Is All You Need: Semantic Traversal of Problem-Solution Spaces with Large Language Models. SSRN Electronic Journal 2024, doi:10.2139/ssrn.4721407.

[76] Yang, J.; Wang, Y. Toward Auto-Modeling of Formal Verification for NextG Protocols: A Multimodal Cross- and Self-Attention Large Language Model Approach. IEEE Access 2024, 12, 27858–27869, doi:10.1109/ACCESS.2024.3366803.

[77] Insuasti, J.; Roa, F.; Zapata-Jaramillo, C.M. Computers' Interpretations of Knowledge Representation Using Pre-Conceptual Schemas: An Approach Based on the BERT and Llama 2-Chat Models. Big Data and Cognitive Computing 2023, 7, doi:10.3390/bdcc7040182.

Bibliography

Abdelhakim, M.; Liu, B.Q.; Sun, C.G. Ar-PuFi: A Short-Text Dataset to Identify the Offensive Messages towards Public Figures in the Arabian Community. Expert Syst Appl 2023, 233, doi:10.1016/j.eswa.2023.120888.

Abed, S.; Khalil, Y.; Modhaffar, M.; Ahmad, I. High-Performance Low-Power Approximate Wallace Tree Multiplier. INTERNATIONAL JOURNAL OF CIRCUIT THEORY AND APPLICATIONS 2018, 46, 2334–2348, doi:10.1002/cta.2540.

Ablimit, M.; Kawahara, T.; Hamdulla, A. Lexicon Optimization Based on Discriminative Learning for Automatic Speech Recognition of Agglutinative Language. Speech Commun 2014, 60, 78–87, doi:10.1016/j.specom.2013.09.011.

Ali, A.; Pinciroli, R.; Yan, F.; Smirni, E. Optimizing Inference Serving on Serverless Platforms. PROCEEDINGS OF THE VLDB ENDOWMENT 2022, 15, 2071–2084, doi:10.14778/3547305.3547313.

An, B.; Long, C.J. Paraphrase Based Data Augmentation For Chinese-English Medical Machine Translation. JOURNAL OF ELECTRONICS & INFORMATION TECHNOLOGY 2022, 44, 118–126, doi:10.11999/JEIT210926.

Antcheva, I.; Ballintijn, M.; Bellenot, B.; Biskup, M.; Brun, R.; Buncic, N.; Canal, P.; Casadei, D.; Couet, O.; Fine, V.; et al. ROOT - A C++ Framework for Petabyte Data Storage, Statistical Analysis and Visualization. Comput Phys Commun 2009, 180, 2499–2512, doi:10.1016/j.cpc.2009.08.005.

Arnes, J.I.; Horsch, A. Schema-Based Priming of Large Language Model for Data Object Validation Compliance. SSRN Electronic Journal 2023, doi:10.2139/ssrn.4453361.

Bae, J.; Cheon, B.D.; Kim, H.Y. Pro-Attention: Efficient Probability Distribution Matching-Based Attention Through Feature Space Conversion. IEEE ACCESS 2022, 10, 131192–131201, doi:10.1109/ACCESS.2022.3229055.

Bati, A.; Bryngelson, S.H. RoseNNa: A Performant, Portable Library for Neural Network Inference with Application to Computational Fluid Dynamics. Comput Phys Commun 2024, 296, doi:10.1016/j.cpc.2023.109052.

Bhanage, D.A.; Pawar, A. V Robust Analysis of IT Infrastructure's Log Data with BERT Language Model. INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS 2023, 14, 705–714.

Boiy, E.; Moens, M.F. A Machine Learning Approach to Sentiment Analysis in Multilingual Web Texts. Inf Retr Boston 2009, 12, 526–558, doi:10.1007/s10791-008-9070-z.

Brenowitz, N.D.; Bretherton, C.S. Prognostic Validation of a Neural Network Unified Physics Parameterization. Geophys Res Lett 2018, 45, 6289–6298, doi:10.1029/2018GL078510.

Buonocore, T.M.; Crema, C.; Redolfi, A.; Bellazzi, R.; Parimbelli, E. Localizing In-Domain Adaptation of Transformer-Based Biomedical Language Models. J Biomed Inform 2023, 144, doi:10.1016/j.jbi.2023.104431.

Buttrick, N. Studying Large Language Models as Compression Algorithms for Human Culture. Trends Cogn Sci 2024, 28, 187–189.

Catelli, R.; Pelosi, S.; Esposito, M. Lexicon-Based vs. Bert-Based Sentiment Analysis: A Comparative Study in Italian. Electronics (Basel) 2022, 11, doi:10.3390/electronics11030374.

Chen, X.; Wang, J.; Gomes, J.; Dubovik, O.; Yang, P.; Saito, M. Analytical Prediction of Scattering Properties of Spheroidal Dust Particles With Machine Learning. Geophys Res Lett 2022, 49, doi:10.1029/2021GL097548.

Chen, Y.C. Effects of Technology-Enhanced Language Learning on Reducing EFL Learners' Public Speaking Anxiety. Comput Assist Lang Learn 2022, doi:10.1080/09588221.2022.2055083.

Cho, M.S.; Vahid, K.A.; Fu, Q.C.; Adya, S.; Mundo, C.C.D.; Rastegari, M.; Naik, D.; Zatloukal, P. EDKM: An Efficient and Accurate Train-Time Weight Clustering for Large Language Models. IEEE COMPUTER ARCHITECTURE LETTERS 2024, 23, 37–40, doi:10.1109/LCA.2024.3363492.

Choudhary, T.; Mishra, V.; Goswami, A.; Sarangapani, J. A Comprehensive Survey on Model Compression and Acceleration. Artif Intell Rev 2020, 53, 5113–5155, doi:10.1007/s10462-020-09816-7.

Chung, J.; Shin, T.; Yang, J.S. Simplifying Deep Neural Networks for FPGA-Like Neuromorphic Systems. IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS 2019, 38, 2032–2042, doi:10.1109/TCAD.2018.2877016.

Cinquin, O. ChIP-GPT: A Managed Large Language Model for Robust Data Extraction from Biomedical Database Records. Brief Bioinform 2024, 25, doi:10.1093/bib/bbad535.

Dong, G.; Chen, W. Blockwise Compression of Transformer-Based Models without Retraining. Neural Networks 2024, 171, 423–428, doi:10.1016/j.neunet.2023.12.001.

Dong, L.B.; Guo, Q.M.; Wu, W.L.; Satpute, M.N. A Semantic Relatedness Preserved Subset Extraction Method for Language Corpora Based on Pseudo-Boolean Optimization. Theor Comput Sci 2020, 836, 65–75, doi:10.1016/j.tcs.2020.07.020.

Du, Y.; Niu, J.; Wang, Y.; Jin, X. Multi-Stage Knowledge Distillation for Sequential Recommendation with Interest Knowledge. Inf Sci (N Y) 2024, 654, doi:10.1016/j.ins.2023.119841.

Eronen, J.; Ptaszynski, M.; Masui, F.; Smywinski-Pohl, A.; Leliwa, G.; Wroczynski, M. Improving Classifier Training Efficiency for Automatic Cyberbullying Detection with Feature Density. Inf Process Manag 2021, 58, doi:10.1016/j.ipm.2021.102616.

Gronsbell, J.; Minnier, J.; Yu, S.; Liao, K.; Cai, T.X. Automated Feature Selection of Predictors in Electronic Medical Records Data. Biometrics 2019, 75, 268–277, doi:10.1111/biom.12987.

Holtzen, S.; Van den Broeck, G.; Millstein, T. Scaling Exact Inference for Discrete Probabilistic Programs. PROCEEDINGS OF THE ACM ON PROGRAMMING LANGUAGES-PACMPL 2020, 4, doi:10.1145/3428208.

Hoppe, N.; Winter, J.M.; Adami, S.; Adams, N.A. ALPACA - a Level-Set Based Sharp-Interface Multiresolution Solver for Conservation Laws. Comput Phys Commun 2022, 272, doi:10.1016/j.cpc.2021.108246.

Huang, S.P.; Jin, L.W.; Xue, K.N.; Fang, Y. Online Primal-Dual Learning for a Data-Dependent Multi-Kernel Combination Model with Multiclass Visual Categorization Applications. Inf Sci (N Y) 2015, 320, 75–100, doi:10.1016/j.ins.2015.05.009.

Insuasti, J.; Roa, F.; Zapata-Jaramillo, C.M. Computers' Interpretations of Knowledge Representation Using Pre-Conceptual Schemas: An Approach Based on the BERT and Llama 2-Chat Models. Big Data and Cognitive Computing 2023, 7, doi:10.3390/bdcc7040182.

Ittiphalin, M.; Arnonkijpanich, B.; Pathumnakul, S. An Artificial Intelligence Model to Estimate the Fat Addition Ratio for the Mixing Process in the Animal Feed Industry. J Intell Manuf 2017, 28, 219–228, doi:10.1007/s10845-014-0972-x.

Jiang, Y.; Weng, J.; Zhang, X.; Yang, Z.; Hu, W. A CNN-Based Born-Again TSK Fuzzy Classifier Integrating Soft Label Information and Knowledge Distillation. IEEE Transactions on Fuzzy Systems 2023, 31, 1843–1854, doi:10.1109/TFUZZ.2022.3215566.

Kane, M.J.; King, C.; Esserman, D.; Latham, N.K.; Greene, E.J.; Ganz, D.A. A Compressed Large Language Model Embedding Dataset of ICD 10 CM Descriptions. BMC Bioinformatics 2023, 24, doi:10.1186/s12859-023-05597-2.

Karthik, E.; Sethukarasi, T. A Centered Convolutional Restricted Boltzmann Machine Optimized by Hybrid Atom Search Arithmetic Optimization Algorithm for Sentimental Analysis. Neural Process Lett 2022, 54, 4123–4151, doi:10.1007/s11063-022-10797-7.

Kashikar, P.; Sentieys, O.; Sinha, S. Lossless Neural Network Model Compression Through Exponent Sharing. IEEE Trans Very Large Scale Integr VLSI Syst 2023, 31, 1816–1825, doi:10.1109/TVLSI.2023.3307607.

Kim, K.; Park, E.J.; Shin, J.H.; Kwon, O.W.; Kim, Y.K. Divergence-Based Fine Pruning of Phrase-Based Statistical Translation Model. Comput Speech Lang 2017, 41, 146–160, doi:10.1016/j.csl.2016.06.006.

Kim, M.-Y.; Rabelo, J.; Babiker, H.K.B.; Rahman, M.A.; Goebel, R. Legal Information Retrieval and Entailment Using Transformer-Based Approaches. The Review of Socionetwork Strategies 2024, doi:10.1007/s12626-023-00153-z.

Kumar, V.; Recupero, D.R.; Riboni, D.; Helaoui, R. Ensembling Classical Machine Learning and Deep Learning Approaches for Morbidity Identification From Clinical Notes. IEEE ACCESS 2021, 9, 7107–7126, doi:10.1109/ACCESS.2020.3043221.

Li, D.C.; Maulana, M.R.; Chou, L.D. NNSplit-SOREN: Supporting the Model Implementation of Large Neural Networks in a Programmable Data Plane. COMPUTER NETWORKS 2023, 222, doi:10.1016/j.comnet.2022.109537.

Li, W.; Hu, A.; Xu, N.; He, G. Quantization and Hardware Architecture Co-Design for Matrix-Vector Multiplications of Large Language Models. IEEE Transactions on Circuits and Systems I: Regular Papers 2024, doi:10.1109/TCSI.2024.3350661.

Li, X.; Luo, D.H.; Cheng, Y.; Wong, K.Y.; Hung, K. Identifying the Primary Odor Perception Descriptors by Multi-Output Linear Regression Models. APPLIED SCIENCES-BASEL 2021, 11, doi:10.3390/app11083320.

Lin, J.F.; Liu, Y.L.; Cleland-Huang, J. Information Retrieval versus Deep Learning Approaches for Generating Traceability Links in Bilingual Projects. Empir Softw Eng 2022, 27, doi:10.1007/s10664-021-10050-0.

Lin, Y.J.; Chen, K.Y.; Kao, H.Y. LAD: Layer-Wise Adaptive Distillation for BERT Model Compression. Sensors 2023, 23, doi:10.3390/s23031483.

Luo, Y.; Wei, Z.; Xu, G.; Li, Z.; Xie, Y.; Yin, Y. Enhancing E-Commerce Chatbots with Falcon-7B and 16-Bit Full Quantization. Journal of Theory and Practice of Engineering Science 2024, 4, 52–57, doi:10.53469/jtpes.2024.04(02).08.

Manojlovic, I.; Svenda, G.; Erdeljan, A.; Gavric, M. Time Series Grouping Algorithm for Load Pattern Recognition. Comput Ind 2019, 111, 140–147, doi:10.1016/j.compind.2019.07.009.

McNorgan, C. The Connectivity Fingerprints of Highly-Skilled and Disordered Reading Persist Across Cognitive Domains. Front Comput Neurosci 2021, 15, doi:10.3389/fncom.2021.590093.

Moayed, H.; Mansoori, E.G.; Moosavi, M.R. An Efficient Pruning Method for Subgraph Matching in Large-Scale Graphs. Journal of Supercomputing 2023, 79, 10511–10532, doi:10.1007/s11227-023-05061-1.

Naser, M.Z. Do We Need Exotic Models? Engineering Metrics to Enable Green Machine Learning from Tackling Accuracy-Energy Trade-Offs. J Clean Prod 2023, 382, doi:10.1016/j.jclepro.2022.135334.

Nedjah, N.; Santos, I.; Mourelle, L.D. Sentiment Analysis Using Convolutional Neural Network via Word Embeddings. Evol Intell 2022, 15, 2295–2319, doi:10.1007/s12065-019-00227-4.

Nguyen, P.T.; Le, A.C.; Ho, T.B.; Nguyen, V.H. Vietnamese Treebank Construction and Entropy-Based Error Detection. Lang Resour Eval 2015, 49, 487–519, doi:10.1007/s10579-015-9308-5.

Petroșanu, D.M.; Pîrjan, A.; Tăbușcă, A. Tracing the Influence of Large Language Models across the Most Impactful Scientific Works. Electronics (Switzerland) 2023, 12.

Pinto, J.P.; Viana, P.; Teixeira, I.; Andrade, M. Improving Word Embeddings in Portuguese: Increasing Accuracy While Reducing the Size of the Corpus. PeerJ Comput Sci 2022, 8, doi:10.7717/peerj-cs.964.

Pujar, S.; Zheng, Y.H.; Buratti, L.; Lewis, B.; Chen, Y.C.; Laredo, J.; Morari, A.; Epstein, E.; Lin, T.N.; Yang, B.; et al. Analyzing Source Code Vulnerabilities in the D2A Dataset with ML Ensembles and C-BERT. Empir Softw Eng 2024, 29, doi:10.1007/s10664-023-10405-9.

Qi, H.; Cao, J.; Chen, S.; Zhou, J. Compressing Recurrent Neural Network Models through Principal Component Analysis. Stat Interface 2023, 16, 397–407, doi:10.4310/22-SII727.

Rajpal, D.; Garg, A.R.; Mahela, O.P.; Alhelou, H.H.; Siano, P. A Fusion-Based Hybrid-Feature Approach for Recognition of Unconstrained Offline Handwritten Hindi Characters. Future Internet 2021, 13, doi:10.3390/fi13090239.

Rao, J.; Meng, X.; Ding, L.; Qi, S.; Liu, X.; Zhang, M.; Tao, D. Parameter-Efficient and Student-Friendly Knowledge Distillation. IEEE Trans Multimedia 2024, 26, 4230–4241, doi:10.1109/TMM.2023.3321480.

Romero, R.W.A.; Viallon, M.; Spaltenstein, J.; Petrusca, L.; Bernard, O.; Belle, L.; Clarysse, P.; Croisille, P. CMRSegTools: An Open-Source Software Enabling Reproducible Research in Segmentation of Acute Myocardial Infarct in CMR Images. PLoS One 2022, 17, doi:10.1371/journal.pone.0274491.

Sandholm, T.E.; Mukherjee, S.; Huberman, B.A. Randomness Is All You Need: Semantic Traversal of Problem-Solution Spaces with Large Language Models. SSRN Electronic Journal 2024, doi:10.2139/ssrn.4721407.

Scheffer, L.K. Analysis Tools for Large Connectomes. Front Neural Circuits 2018, 12, doi:10.3389/fncir.2018.00085.

Schonfeld, E.; Pant, A.; Shah, A.R.Y.; Sadeghzadeh, S.; Pangal, D.; Rodrigues, A.; Yoo, K.; Marianayagam, N.; Haider, G.; Veeravagu, A. Evaluating Computer Vision, Large Language, and Genome-Wide Association Models in a Limited Sized Patient Cohort for Pre-Operative Risk Stratification in Adult Spinal Deformity Surgery. J Clin Med 2024, 13, doi:10.3390/jcm13030656.

Shoulson, I.; Arbatti, L.; Hosamath, A.; Eberly, S.W.; Oakes, D. Longitudinal Cohort Study of Verbatim-Reported Postural Instability Symptoms as Outcomes for Online Parkinson's Disease Trials. J Parkinsons Dis 2022, 12, 1969–1978, doi:10.3233/JPD-223274.

Tăbuscă, A.; Tăbuscă, S.-M. Impact of 5G Technology in Global Economy. Cybersecurity and Legal Issues. Journal of Information Systems & Operations Management 2020, 13, 177–189.

Tan, C.; Liu, J. Improving Knowledge Distillation With a Customized Teacher. IEEE Trans Neural Netw Learn Syst 2024, 35, 2290–2299, doi:10.1109/TNNLS.2022.3189680.

Tan, J.H.; Tan, Y.H.; Chan, C.S.; Chuah, J.H. ACORT: A Compact Object Relation Transformer for Parameter Efficient Image Captioning. Neurocomputing 2022, 482, 60–72, doi:10.1016/j.neucom.2022.01.081.

Tan, R.; Lin, Q.; Low, G.H.; Lin, R.X.; Goh, T.C.; Chang, C.C.E.; Lee, F.F.; Chan, W.Y.; Tan, W.C.; Tey, H.J.; et al. Inferring Cancer Disease Response from Radiology Reports Using Large Language Models with Data Augmentation and Prompting. JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION 2023, 30, 1657–1664, doi:10.1093/jamia/ocad133.

Tang, X.Y.; Zeng, S.; Yu, F.; Yu, W.; Sheng, Z.Y.; Kang, Z. Self-Supervised Anomaly Pattern Detection for Large Scale Industrial Data. Neurocomputing 2023, 515, 1–12, doi:10.1016/j.neucom.2022.09.069.

Tapiador-Morales, R.; Linares-Barranco, A.; Jimenez-Fernandez, A.; Jimenez-Moreno, G. Neuromorphic LIF Row-by-Row Multiconvolution Processor for FPGA. IEEE Trans Biomed Circuits Syst 2019, 13, 159–169, doi:10.1109/TBCAS.2018.2880012.

Tiwari, A.K.; Shreevastava, S.; Som, T.; Shukla, K.K. Tolerance-Based Intuitionistic Fuzzy-Rough Set Approach for Attribute Reduction. Expert Syst Appl 2018, 101, 205–212, doi:10.1016/j.eswa.2018.02.009.

Tuan, N.T.; Moore, P.; Thanh, D.H. V; Pham, H. V A Generative Artificial Intelligence Using Multilingual Large Language Models for ChatGPT Applications. APPLIED SCIENCES-BASEL 2024, 14, doi:10.3390/app14073036.

Tüzemen, E.S.; Yüksek, A.G.; Demir, I.; Horoz, S.; Altuntas, I. Modeling of Temperature-Dependent Photoluminescence of GaN Epilayer by Artificial Neural Network. JOURNAL OF THE AUSTRALIAN CERAMIC SOCIETY 2023, 59, 1145–1159, doi:10.1007/s41779-023-00911-w.

Vázquez, R.; Raganato, A.; Creutz, M.; Tiedemann, J. A Systematic Study of Inner-Attention-Based Sentence Representations in Multilingual Neural Machine Translation. COMPUTATIONAL LINGUISTICS 2020, 46, 387–424, doi:10.1162/coli_a_00377.

Wang, H.Z.; Qu, Z.H.; Zhou, Q.H.; Zhang, H.B.; Luo, B.Y.; Xu, W.C.; Guo, S.; Li, R.X. A Comprehensive Survey on Training Acceleration for Large Machine Learning Models in IoT. IEEE Internet Things J 2022, 9, 939–963, doi:10.1109/JIOT.2021.3111624.

Woods, L.T.; Rana, Z.A. Constraints on Optimising Encoder-Only Transformers for Modelling Sign Language with Human Pose Estimation Keypoint Data. J Imaging 2023, 9, doi:10.3390/jimaging9110238.

Wu, R.H.; Zhu, X.Y.; Chen, J.S.; Liu, S.; Zheng, T.Y.; Liu, X.; An, H. SWattention: Designing Fast and Memory-Efficient Attention for a New Sunway Supercomputer. JOURNAL OF SUPERCOMPUTING 2024, doi:10.1007/s11227-024-05890-8.

Yan, H.; Liu, Y.; Jin, L.; Bai, X. The Development，application，and Future of LLM Similar to ChatGPT. Journal of Image and Graphics 2023, 28, 2749–2762, doi:10.11834/jig.230536.

Yang, J.; Wang, Y. Toward Auto-Modeling of Formal Verification for NextG Protocols: A Multimodal Cross- and Self-Attention Large Language Model Approach. IEEE Access 2024, 12, 27858–27869, doi:10.1109/ACCESS.2024.3366803.

Yang, Z.; Zhang, Y.; Sui, D.; Ju, Y.; Zhao, J.; Liu, K. Explanation Guided Knowledge Distillation for Pre-Trained Language Model Compression. ACM Transactions on Asian and Low-Resource Language Information Processing 2024, 23, doi:10.1145/3639364.

Zahera, H.M.; El-Sisi, A.B. Accelerating Training Process in Logistic Regression Model Using OpenCL Framework. INTERNATIONAL JOURNAL OF GRID AND HIGH PERFORMANCE COMPUTING 2017, 9, 34–45, doi:10.4018/IJGHPC.2017070103.

Zhang, J.F.; Zhang, Z.Y. Machine Learning Hardware Design for Efficiency, Flexibility, and Scalability [Feature]. IEEE CIRCUITS AND SYSTEMS MAGAZINE 2023, 23, 35–53, doi:10.1109/MCAS.2023.3302390.

Zhu, J.; Wang, L.Y.; Han, X.; Liu, A.M.; Xie, T. Safety and Performance, Why Not Both? Bi-Objective Optimized Model Compression Against Heterogeneous Attacks Toward AI Software Deployment. IEEE TRANSACTIONS ON SOFTWARE ENGINEERING 2024, 50, 376–390, doi:10.1109/TSE.2023.3348515.

Zhu, X.N.; Yang, M.Y.; Zhao, T.J.; Zhu, C.H. Minimum Bayes-Risk Phrase Table Pruning for Pivot-Based Machine Translation in Internet of Things. IEEE ACCESS 2018, 6, 55754–55764, doi:10.1109/ACCESS.2018.2872773.