Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/77/20240688

Exploring the horizon of AI development: Navigating constraints of chips and power in the technological landscape

Wenwen Hou

School of Computer Science, Guangzhou Institute of Science and Technology, Guangzhou, China

hou.wenwen@outlook.com

Abstract. With the proliferation of AI technology, machine learning has emerged as a cornerstone of AI systems, facilitating pattern recognition and decision-making through robust data analysis. This encompasses various learning paradigms such as supervised, unsupervised, and reinforcement learning, all of which are indispensable for the advancement of artificial intelligence. Nevertheless, the development of AI necessitates substantial computational resources, with specialized chips serving as the linchpin, particularly in demanding tasks such as deep learning. Dedicated chip development, exemplified by GPUs and TPUs, plays a pivotal role in enhancing the performance of AI systems, notwithstanding challenges related to costs and market monopolies. Moreover, AI systems require significant power support, especially during the training of large-scale models. To address these challenges, this paper reviews the existing literature on modeling techniques aimed at enhancing the efficiency of machine learning and reducing energy consumption. This review encompasses optimal algorithm design, hardware optimization, and spatial modeling. Through the implementation of these approaches, the challenges posed by resource constraints in machine learning scenarios can be effectively mitigated, thereby fostering the continued development and application of AI technology.

Keywords: AI, machine learning, deep learning, chip, efficiency, energy consumption.

1. Introduction

In contemporary times, propelled by technological advancements, several global powers are directing their efforts towards the development of high-performance chips, such as GPUs, capable of handling immense arithmetic power. This pursuit aims to establish technological leadership, particularly among tech companies striving to engineer superior AI products. Presently, machine learning commands a substantial portion, nearly 60%, of global investments, primarily driven by the burgeoning AI industry encompassing robotics and speech recognition [1]. IBM defines AI as a system that "utilizes computers and machines to replicate the problem-solving and decision-making capabilities inherent in the human mind"[2]. The emergence of artificial intelligence has ignited a fierce competition between the United States and China in the realm of chip development, given the pivotal role of semiconductors. Semiconductors not only serve as the cornerstone of the modern economy and the digital realm but also constitute essential hardware for the development and operation of AI systems [3]. China, with its substantial semiconductor demand, commands a notable 34.4% market share, yet it accounts for only 16% of the world's total manufacturing capacity[4]. Lawrence & VerWey [4] critique China's

DOI: 10.54254/2755-2721/77/20240688

semiconductor industry plan for lacking clear objectives and a coherent implementation strategy, hindered by bureaucratic hurdles. Furthermore, the veil of secrecy surrounding technology in Japan, South Korea, and even within China and Taiwan has impeded China's quest for significant technological breakthroughs [5]. In addition to the on-chip pressures, there exists a significant strain on environmental resources. The energy demands induced by the large-scale model arithmetic employed in the AI industry are considerable. Even before the widespread adoption of machine learning models, the proliferation of big data centers saw a six-fold increase in server numbers, reaching 30 million. Consequently, the energy consumption per server far surpasses that of earlier models. The continual evolution and refinement of big-model algorithmic techniques have imposed formidable challenges on the energy sector at a global scale. These challenges encompass escalating consumption rates, efficiency dilemmas, volatile trends in supply and demand, and a dearth of comprehensive analyses essential for effective management. The gravity of these issues is particularly pronounced in emerging markets, where unauthorized "grid access" is prevalent, leading to substantial energy losses and heightened levels of CO2 emissions. This exacerbates the already critical efficiency predicaments [6].

This paper presents a comprehensive review of the contemporary state-of-the-art modeling literature aimed at enhancing the efficiency of machine learning algorithms and mitigating energy consumption. We concentrate on four primary areas of optimization: algorithm and model design, hardware optimization, and spatial modeling techniques. Our objective is to offer engineers and scientists in computer science and sustainable technology a fresh perspective within the context of chip constraints and to delineate avenues for future research endeavors. The paper unfolds as follows: Section 2 elucidates the design of optimization algorithms and model structures tailored for processing large-scale datasets. Section 3 explores current advancements in hardware optimization, providing actionable recommendations. In Section 4, we delve into spatial modeling techniques, addressing both their merits and limitations. Lastly, Section 5 presents a contemporary outlook on the future of AI, advocating for reduced dependence on chips.

2. Model design for large-scale data

The rapid advancements in text categorization, speech recognition, and image processing have ushered in a myriad of optimization challenges within the domain of machine learning. Text categorization endeavors often encounter convex optimization problems, rooted in the application of algorithms like logistic regression or support vector machines. Conversely, the terrain of speech and image recognition is characterized by highly intricate non-linear and non-convex problems, necessitating the utilization of deep neural networks.

As the volume of large-scale data continues to burgeon, the quest for models adept at efficiently discerning patterns and correlations within datasets becomes imperative. Sustainable data modeling stands as a beacon, striving to optimize two pivotal facets: achieving maximum learning accuracy while minimizing computational overhead and facilitating rapid processing of extensive datasets. This optimization not only fosters enhanced data processing efficiency but also frequently leads to significant cost reductions. This assertion finds validation in the works of esteemed researchers such as Patnaik, Sundaravaradan and Marwah [7,8,9]. Furthermore, streamlining model complexity emerges as a potent strategy to bolster efficiency by curbing computational burden and curtailment of energy consumption. Leveraging domain expertise, researchers embark on the endeavor of simplifying models, rendering them more computationally tractable. Notably, strides have been made across four major domains [10]: Kernel models, Graph models, Deep models, and Tree models. For instance, kernel methods offer an efficient means to compute inner products in high-dimensional feature spaces, thus enabling the modeling of non-linear relationships with heightened efficiency. Kernel methods have proven to be particularly adept at handling datasets with complex structures, as evidenced by simplification techniques like sampling-based approaches [11] and projection-based approximations [12].

DOI: 10.54254/2755-2721/77/20240688

3. Optimization algorithms

The surge in research endeavors within text categorization, speech recognition, and image processing has catalyzed a spectrum of optimization challenges in machine learning. Text categorization research often grapples with convex optimization problems, typically arising from the application of algorithms such as logistic regression or support vector machines. Conversely, studies in speech or image recognition confront highly complex non-linear and non-convex problems, necessitating the utilization of deep neural networks. Diverse optimization techniques abound in machine learning, offering avenues to reduce unnecessary computation and enhance computational efficiency. Noteworthy among these are three commonly employed approximation optimization methods: Mini-batch Gradient Descent[13], Coordinate Gradient Descent[14], and Numerical Integration based on Markov chain Monte Carlo (MCMC)[15].

3.1. Mini-batch gradient descent

Mini-batch gradient descent updates the parameters (θ) of a model iteratively using mini-batches of data (B) from the dataset. In each iteration (t), the parameters are updated according to the formula:

$$[\theta^{(t+1)} = \theta^{(t)} - \alpha \cdot g_B^{(t)}] \tag{1}$$

where (α) is the learning rate and $(g_B^{(t)})$is the gradient of the loss function$\left(\mathrm{J}(\theta)\right)$ computed on mini-batch (B) at iteration (t).

3.2. Coordinate gradient descent

Coordinate gradient descent updates each parameter $\left(\theta_j\right)$ individually while holding the others fixed. In each iteration (t), a parameter $\left(\theta_j\right)$is chosen, and the parameter is updated according to:

$$[\theta_j^{(t+1)} = \mathrm{argmin}_{\theta_j} \mathrm{J}\left(\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_j, \ldots, \theta_n^{(t)}\right)] \tag{2}$$

This process is repeated for each parameter until convergence.

3.3. Numerical integration using MCMC

Numerical integration using MCMC involves sampling from the parameter space using a Markov chain. Given a target distribution $\left(\mathrm{p}(\theta)\right)$and a proposal distribution$\left(\mathrm{q}\left(\theta' \mid \theta\right)\right)$, the parameters are sampled iteratively according to the Metropolis-Hastings algorithm:

1. Given the current parameter $\left(\theta^{(t)}\right)$, propose a new parameter $\left(\theta'\right)$ from $\left(\mathrm{q}\left(\theta' \mid \theta^{(t)}\right)\right)$.
2. Compute the acceptance probability$\left(\mathrm{A} = \min\left(1, \frac{p\left(\theta'\right)}{p\left(\theta^{(t)}\right)}\right)\right)$.
3. Accept the new parameter with probability (A); otherwise, retain the current parameter.

This process is repeated for a large number of iterations to obtain samples from the target distribution$\left(\mathrm{p}(\theta)\right)$, which can be used for approximating integrals and optimizing functions.

Optimization algorithms in machine learning across diverse domains encounter notable challenges, especially in the domain of deep neural networks. Stochastic gradient descent algorithms, integral to deep learning optimization, often present various complexities [16,17]. Notably, some adaptive methods exhibit learning rate fluctuations during later training stages, leading to the potential problem of non-convergence [18]. However, integrating stochastic gradient descent with its variant properties offers a promising avenue for optimization enhancement[19]. Particularly noteworthy is the transition from adaptive algorithms to stochastic gradient descent methods, which can augment algorithmic accuracy and convergence speed [20]. Furthermore, stochastic optimization methods, as proposed by *T. Chen et al.* [21], can be effectively applied to Markov chain Monte Carlo (MCMC) sampling, thereby bolstering efficiency. Another noteworthy technique, stochastic variational inference, enhances algorithmic

DOI: 10.54254/2755-2721/77/20240688

efficiency by introducing natural gradients and extending variational inference to large-scale datasets, outperforming traditional optimization methods [22].

4. Hardware optimization

Deep Neural Networks (DNNs) have revolutionized numerous fields, leveraging innovative model configurations and advancements in hardware platforms. However, the escalating complexity of training DNNs has resulted in substantial energy requirements, posing a significant barrier to their deployment on energy-constrained embedded and mobile devices. For instance, deploying DNNs like AlexNet for image classification on Internet of Things (IoT) nodes and wearable devices can rapidly deplete a smartphone's battery within an hour [23]. To tackle the energy consumption challenge, Y.-H. Chen et al. [24] and others have proposed strategies to minimize data movement and optimize energy utilization. They highlight the detrimental impact of large filter weights and channels in DNNs on energy consumption, advocating for techniques such as data reuse and support for different shapes to reduce data movement. Additionally, leveraging data statistical information to implement zero-skipping/gating helps avoid unnecessary reads and computations. Furthermore, Cai et al. [25] and colleagues have introduced a layer-by-layer prediction framework based on sparse polynomial regression, aiming to forecast the energy consumption of Convolutional Neural Networks (CNNs) during inference on any GPU platform. Efforts to develop models capable of reasoning about energy consumption have yielded promising results. Yang et al. [26] demonstrated that energy-based pruning surpasses FLOPs-based methods in energy efficiency. They developed a prediction model based on measurements from their hardware accelerator Eyeriss, showcasing the potential for improved energy efficiency.In the context of Mobile Crowdsourcing Machine Learning (MCML), Anh et al. [27] and colleagues proposed a deep Q-learning algorithm to enable servers to dynamically adapt and make optimal decisions in uncertain mobile environments. This algorithm outperforms static alternatives in terms of energy consumption and training latency, facilitating more efficient MCML operations.

5. Spatial modelling

The ascent of AI is undeniably intertwined with endeavors exploring the neural architecture of the brain, exemplified by innovations like Long Short-Term Memory (LSTM) networks [28]. Drawing inspiration from studies elucidating working memory in the neurosciences, AI researchers have seamlessly integrated memory modules into machine learning frameworks, thereby underpinning a plethora of sequential processing tasks. Guided by insights into neocortical plasticity—a cornerstone of continuous learning in the brain—researchers have not only delved into memory mechanisms but also sought inspiration from the brain's attentional faculties. The integration of attention modules into artificial neural networks, whether temporally or spatially, has marked a transformative leap in AI[29,30]. These modules endow networks with the capacity to selectively focus on salient features while disregarding irrelevant elements, thereby bolstering the efficacy of deep neural networks in tasks spanning natural language processing and computer vision. Moreover, this incorporation enhances the efficiency of the training and inference processes, surpassing the capabilities of conventional deep networks. The evolution of such adaptable algorithms has catalyzed the emergence of contemporary big language models [31]. Yet, as the demand for computational power and chips escalates with the blind expansion of sequential algorithms, scholars propose a paradigm shift towards spatial modeling. Wu *et al.* [32] argue that prevailing big language models are rooted in a one-dimensional framework, positing that embracing spatial modeling could alleviate the concomitant chip requirements. Building upon this premise, Bui *et al.* [33] advocate for spatial approaches that enable the simultaneous processing of entire graphs at reduced computational costs. They propose the translation of the Hierarchical Deep Learning Neural Network (HiDeNN) framework, originally designed for addressing 3D problems in structural engineering [34]. This framework leverages multi-scale models within the HiDeNN architecture to address challenges across macro, meso, and microscales. Through collaborative coupling of loss functions, HiDeNN facilitates the modeling and resolution of physical phenomena across varying spatial scales, offering a novel approach to tackle computationally intensive problems.

DOI: 10.54254/2755-2721/77/20240688

6. Conclusions

This review endeavors to elucidate avenues for steering the development of AI technology within the prevailing technological milieu, mindful of the inherent constraints imposed by chip architecture and power consumption. Specifically, we delve into theoretical and empirical dimensions within the ambit of large-scale data-intensive domains, encompassing (1) the conceptualization of model frameworks, (2) the formulation of model optimization algorithms, and (3) an emerging paradigm shift—spatial modeling. The burgeoning influx of big data underscores the imperative for energy-efficient computing prowess, given its palpable impact on human resources. Foreseeably, the trajectory of AI development hinges upon the efficacy of data modeling practices, poised to catalyze advancements in spatial modeling. Novel design paradigms, exemplified by sustainable data modeling, not only hold promise for addressing chip-related challenges but also serve as conduits for maximizing dividends across multifarious scientific domains.

References

[1] Bughin, J., Hazan, E., Sree Ramaswamy, P., DC, W., & Chu, M. (2017). *Artificial intelligence the next digital frontier*.

[2] Lohr, S. (2007). IBM effort to focus on saving energy. *Online, May*.

[3] Miller, C. (2022). *Chip war: The fight for the world's most critical technology.* Simon and Schuster.

[4] Voas, J., Kshetri, N., & DeFranco, J. F. (2021). Scarcity and global insecurity: The semiconductor shortage. *IT Professional, 23*(5), 78–82.

[5] VerWey, J. (2019). Chinese semiconductor industrial policy: Prospects for future success. J. *Int'l Com. & Econ.*, 1.

[6] Mhlanga, D. (2023). Artificial intelligence and machine learning for energy consumption and production in emerging markets: A review. *Energies*, 16(2), 745.

[7] Marwah, M., Shah, A., Bash, C., Patel, C., & Ramakrishnan, N. (2011). Using data mining to help design sustainable products. *Computer, 44*(08), 103–106.

[8] Patnaik, D., Marwah, M., Sharma, R. K., & Ramakrishnan, N. (2010). Data mining for modeling chiller systems in data centers. *Advances in Intelligent Data Analysis IX: 9th International Symposium, IDA 2010, Tucson, AZ, USA, May 19-21, 2010. Proceedings 9*, 125–136.

[9] Sundaravaradan, N., Patnaik, D., Ramakrishnan, N., Marwah, M., & Shah, A. (2011). Discovering life cycle assessment trees from impact factor databases. *Proceedings of the AAAI Conference on Artificial Intelligence, 25*(1), 1415–1420.

[10] Wang, M., Fu, W., He, X., Hao, S., & Wu, X. (2020). A survey on large-scale machine learning. *IEEE Transactions on Knowledge and Data Engineering, 34*(6), 2574–2594.

[11] Kumar, S., Mohri, M., & Talwalkar, A. (2012). Sampling methods for the Nyström method. *The Journal of Machine Learning Research, 13*(1), 981–1006.

[12] Martinsson, P.-G., Rokhlin, V., & Tygert, M. (2011). A randomized algorithm for the decomposition of matrices. *Applied and Computational Harmonic Analysis, 30*(1), 47–68.

[13] Smith, S. L., Kindermans, P.-J., Ying, C., & Le, Q. V. (2017). Don't decay the learning rate, increase the batch size. *ArXiv Preprint ArXiv:1711.00489*.

[14] Nutini, J., Schmidt, M., Laradji, I., Friedlander, M., & Koepke, H. (2015). Coordinate descent converges faster with the gauss-southwell rule than random selection. *International Conference on Machine Learning*, 1632–1641.

[15] Jiang, J., Cui, B., Zhang, C., & Fu, F. (2018). Dimboost: Boosting gradient boosting decision tree to higher dimensions. *Proceedings of the 2018 International Conference on Management of Data*, 1363–1376.

[16] Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research, 12*(7).

[17] Zeiler, M. D. (2012). Adadelta: An adaptive learning rate method. *ArXiv Preprint ArXiv:1212.5701*.

DOI: 10.54254/2755-2721/77/20240688

[18] Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *ArXiv Preprint ArXiv:1511.06434*
[19] Johnson, R., & Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. *Advances in Neural Information Processing Systems, 26*.
[20] Keskar, N. S., & Socher, R. (2017). Improving generalization performance by switching from adam to sgd. *ArXiv Preprint ArXiv:1712.07628*.
[21] Chen, T., Fox, E., & Guestrin, C. (2014). Stochastic gradient hamiltonian monte carlo. *International Conference on Machine Learning*, 1683–1691.
[22] Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. (2015). Efficient machine learning for big data: A review. *Big Data Research, 2*(3), 87–93.
[23] Zhang, Z., & Kouzani, A. Z. (2020). Implementation of DNNs on IoT devices. *Neural Computing and Applications, 32*(5), 1327-1356.
[24] Chen, Y.-H., Krishna, T., Emer, J. S., & Sze, V. (2016). Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks. *IEEE Journal of Solid-State Circuits, 52*(1), 127–138.
[25] Cai, E., Juan, D.-C., Stamoulis, D., & Marculescu, D. (2017). Neuralpower: Predict and deploy energy-efficient convolutional neural networks. *Asian Conference on Machine Learning*, 622–637.
[26] Yang, T.-J., Chen, Y.-H., & Sze, V. (2017). Designing energy-efficient convolutional neural networks using energy-aware pruning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5687–5695.
[27] Anh, T. T., Luong, N. C., Niyato, D., Kim, D. I., & Wang, L.-C. (2019). Efficient training management for mobile crowd-machine learning: A deep reinforcement learning approach. *IEEE Wireless Communications Letters, 8*(5), 1345–1348.
[28] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.
[29] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *ArXiv Preprint ArXiv:1409.0473*.
[30] Reed, S. E., Zhang, Y., Zhang, Y., & Lee, H. (2015). Deep visual analogy-making. *Advances in Neural Information Processing Systems, 28*.
[31] Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., & Wang, Y. (2024). *A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology, 15*(3), 1–45.
[32] Wu, R.-T., Liu, T.-W., Jahanshahi, M. R., & Semperlotti, F. (2021). Design of one-dimensional acoustic metamaterials using machine learning and cell concatenation. *Structural and Multidisciplinary Optimization, 63*, 2399–2423.
[33] Bui, K.-H. N., Cho, J., & Yi, H. (2022). Spatial-temporal graph neural network for traffic forecasting: An overview and open research issues. *Applied Intelligence, 52*(3), 2763–2774.
[34] Saha, S., Gan, Z., Cheng, L., Gao, J., Kafka, O. L., Xie, X., Li, H., Tajdari, M., Kim, H. A., & Liu, W. K. (2021). Hierarchical deep learning neural network (HiDeNN): An artificial intelligence (AI) framework for computational science and engineering. *Computer Methods in Applied Mechanics and Engineering, 373*, 113452.

Acknowledgments
Extend my sincere gratitude to the reviewers for their thorough evaluation and insightful comments, which significantly improved the quality of this work. Their expertise and attention to detail have been truly appreciated throughout the review process