Redefining Efficiency in AI: The Impact of 1.58-bit LLMs on the Future of Computing

Submitted by:
Gregor Lyttek

March 02, 2024

Introduction

The rise of Large Language Models (LLMs) in the Artificial Intelligence (AI) realm signifies a major milestone. These sophisticated models have demonstrated remarkable proficiency in various natural language processing tasks. Nevertheless, their escalating size and complexity have led to concerns about deployment issues and the environmental and economic repercussions, chiefly due to their high energy consumption.

Typically, LLMs function with 16-bit floating values, with matrix multiplication forming the core of their computational workload. Relying on floating-point operations has resulted in significant energy expenses, constraining their efficiency and practicality. To address these issues, the industry has shifted towards post-training quantization, reducing the precision of weights and activations to craft low-bit models for inference. Although widely embraced, this method, particularly with 4-bit variants, has not been entirely optimal.

In this context, the recent introduction of 1-bit model architectures, like BitNet, signifies a groundbreaking advancement. BitNet transforms the computational framework of LLMs by substituting energy-intensive floating-point operations with integer addition. This shift in approach offers a notable decrease in energy consumption, tackling the power limitations that often impede computational performance. Consequently, the adoption of 1-bit LLMs not only aims to boost the efficiency of current models but also aims to redefine AI computational strategies, paving the way for advanced applications and sustainable AI progress.

Numerical Formats in LLMs

Exploring Large Language Models (LLMs) involves grasping the fundamental aspects of numerical representations they rely on. Traditionally, LLMs have utilized 16-bit floating values (FP16/BF16) for computations, striking a balance between capturing diverse numerical values and maintaining computational efficiency.

In the pursuit of more efficient AI models, there has been a shift towards exploring lower-bit variants like 4-bit models, which notably decrease memory and computational demands. This transition aligns with the broader trend in AI towards optimizing model performance while conserving resources.

In the upcoming sections, we will delve into these numerical formats, examining their features, applications in LLMs, and the trade-offs involved. This comprehension is crucial as we progress towards exploring the innovative 1.58-bit models and their potential to revolutionize the AI landscape.

Explanation of 16-bit Floating Values and 4-bit Variants

16-bit Floating Values (FP16/BF16)

- Definition: 16-bit floating-point representation (FP16/BF16) is a format used in computing that occupies 16 bits in computer memory. It is commonly used in deep learning models, including LLMs.
- Components: It includes a sign bit, an exponent, and a fraction, allowing a wide range of values to be represented.
- Usage in LLMs: In LLMs, 16-bit floating-point values are standard for representing weights and activations. They balance computational efficiency and numerical precision.

4-bit Variants

- Definition: 4-bit variant quantization involves reducing the precision of model weights and activations to 4 bits.
- Purpose: This significantly reduces memory and computational requirements, enhancing efficiency in inference tasks.
- Trade-offs: While more efficient, 4-bit quantization can lead to performance reductions compared to higher bit representations.

Hardware Optimization

Innovations like 1.58-bit LLMs, such as BitNet b1.58, mark a new era of computational efficiency and performance in AI. However, to maximize these models' potential, specialized hardware systems tailored to their unique computational needs are essential.

Projects like Groq5 have demonstrated progress by developing dedicated hardware units like Language Processing Units (LPUs) for LLMs. These solutions optimize LLM performance by aligning with their computation patterns and energy efficiency requirements.

The advent of 1.58-bit models presents an opportunity to create hardware and systems optimized for these models. BitNet's unique computation approach, with minimal multiplication operations for matrix multiplication, necessitates a fresh perspective on hardware design. Tailored hardware could significantly enhance energy efficiency, processing speed, and overall computational performance.

Designing hardware specifically for 1.58-bit LLMs not only boosts model capabilities but also drives sustainable and cost-effective AI solutions. This path represents a critical advancement in AI technology, harmonizing computational efficiency with the escalating demands of advanced AI models.

Enhanced Model Deployment 1.58-bit LLMs

The emergence of 1.58-bit LLMs, like BitNet b1.58, marks a notable advancement in the AI domain, especially in deploying language models on edge and mobile devices. Traditionally, these devices faced limitations due to their restricted memory and computing power, hindering the performance and scalability of LLMs in such environments.

However, the introduction of 1.58-bit LLMs signifies a significant shift in this landscape. With their reduced memory usage and energy consumption, these models can now operate effectively on edge and mobile devices, surpassing previous constraints. This progress unlocks a plethora of new applications that were previously unattainable, significantly enhancing the capabilities of these devices.

Moreover, 1.58-bit LLMs align well with CPU devices commonly found in edge and mobile technologies. This alignment ensures more efficient execution of models like BitNet b1.58 on these platforms, further boosting their performance and functionalities. By deploying 1.58-bit LLMs on such devices, access to advanced AI capabilities becomes more inclusive, fostering innovative applications across various fields, from personalized AI assistants to advanced data analysis tools in mobile environments.

The implications of this enhanced deployment are extensive, promising to broaden the influence and accessibility of LLMs significantly. By enabling the use of these powerful models on a broader range of devices, 1.58-bit LLMs are poised to play a pivotal role in the forthcoming wave of AI-driven advancements.

Mixture-of-Experts (MoE) Explained

In machine learning, the Mixture-of-Experts (MoE) approach involves a model comprising multiple specialized 'expert' sub-models. Each expert is skilled in handling distinct data types or tasks. Within Large Language Models (LLMs), MoE enhances efficiency by delegating computational tasks to these experts.

Traditionally, integrating MoE into LLMs has been hindered by high memory usage and substantial inter-chip communication overhead. Nonetheless, the emergence of 1.58-bit LLMs, such as BitNet b1.58, addresses these issues. The reduced memory requirements of 1.58-bit LLMs decrease the devices needed for deploying MoE models and significantly lessen the activation transfer overhead across networks. Ideally, if the entire MoE model could fit on a single chip, it would eradicate the overhead entirely, making MoE models more viable and effective.

Potential of 1.58-bit LLMs in Handling Long Sequences

In the era of Large Language Models (LLMs), efficiently managing long data sequences is crucial. Handling long sequences poses a challenge due to high memory usage, especially with Key-Value (KV) caches in use. This challenge has limited the expansion of LLM capabilities.

The innovation of BitNet b1.58, an example of 1.58-bit LLMs, is a significant leap in overcoming this hurdle. By reducing activation size from 16 bits to 8 bits, BitNet b1.58 doubles the context length manageable with the same resources. This advancement allows processing of longer sequences without a proportional increase in memory usage.

Future optimization potential exists for 1.58-bit LLMs. The ability to compress these activations to 4 bits or lower presents opportunities for more efficient handling of extended sequences. This optimization is crucial to meet the increasing demand for processing extensive data sequences in AI applications, from natural language understanding to complex data analysis.

The progress in 1.58-bit LLMs like BitNet b1.58 signifies a significant advancement towards supporting long sequences effectively. This not only enhances current LLM capabilities but also paves the way for more robust and adaptable models in the future.

Risk and Cybersecurity Perspective

- Evolving Cyber Threat Landscape: With advancements in AI models like BitNet b1.58, the potential for intricate cyber threats increases, emphasizing the necessity for advanced cybersecurity protocols to identify and counter threats leveraging AI capabilities.
- AI's Role in Cyber Defense: Utilizing 1.58-bit LLMs in defense mechanisms significantly enhances threat detection and response capabilities, allowing for real-time analysis and proactive threat identification that traditional systems may overlook.
- Ethical Challenges of AI-driven Cyber Offense: The ethical concerns surrounding the use of AI in offensive cybersecurity, such as automated hacking or misinformation campaigns, require a cautious approach to the integration of these technologies to prevent misuse.
- Striking a Balance Between Efficiency and Security: While 1.58-bit LLMs offer improved efficiency, ensuring their security against potential vulnerabilities is paramount, particularly in sensitive applications where maintaining a balance between efficiency and security is crucial.
- Compliance and Regulatory Implications: Deploying AI in cybersecurity necessitates adherence to regulatory and compliance frameworks, aligning these models with legal and ethical standards in cybersecurity practices to ensure responsible and effective implementation.

Hypothetical Scenario: Next-Generation Flipper Zero

A Peek into the Future:

Imagine entering a realm where cutting-edge technology seamlessly merges with everyday devices. Picture holding a compact device, no larger than a typical smartphone, yet possessing the computing power of a robust computer. Welcome to the innovative Flipper Zero of the future, driven by the groundbreaking BitNet b1.58.

Unleashing Advanced AI in Your Palm:

Instant Language Processing: Communicate with the Flipper Zero, and it comprehends and responds in various languages, effortlessly decoding dialects and slangs. It's like having a personal UN translator in your pocket.

Advanced Data Analysis: With a few simple commands, unlock its capacity to analyze intricate datasets. Whether predicting market trends or understanding environmental patterns, the Flipper Zero provides insights once reserved for supercomputers.

Decision-Making Assistance: In critical moments, whether navigating a crisis or making swift business decisions, the Flipper Zero provides real-time decision support, evaluating scenarios with remarkable depth and speed.

Revolutionizing Portable Computing

Powerful Compactness: The new Flipper Zero breaks away from conventional portable devices. Its advanced processing and AI features deliver superior computational abilities at your fingertips.

Reshaping IoT and Mobile Computing: The implications for IoT are substantial. From intelligent homes that anticipate residents' needs to adaptive urban infrastructure, integrating potent AI into compact devices opens up a world of opportunities.

A Day with the Flipper Zero

Envision a day enriched by this device. In the morning, it analyzes traffic data to recommend the optimal route to work. Throughout the day, it organizes your schedule, predicts your requirements, and facilitates your interactions, becoming an essential part of your routine. In professional settings, it provides data-driven insights, while at home, it serves as a central hub for managing a smart, intuitive living space.

This envisioned scenario doesn't just showcase a futuristic gadget; it exemplifies the potential held by 1.58-bit LLMs like BitNet b1.58. It signifies a shift into an era where AI is not merely a tool but an integral aspect of daily life, transforming our relationship with technology at its core.

Conclusion

In our whitepaper titled "Redefining Efficiency in AI: The Impact of 1.58-bit LLMs on the Future of Computing," we have explored the transformative potential of 1.58-bit Large Language Models (LLMs), focusing on BitNet b1.58. Our examination covered the technical advancements in numerical formats within LLMs, the implications of these models for hardware optimization, improved deployment, and their integration with Mixture-of-Experts (MoE) models.

We also addressed the potential cybersecurity risks and opportunities, emphasizing how these models can play a dual role in defense and offense scenarios. Furthermore, we envisioned a scenario involving a next-generation Flipper Zero device powered by a 1.58-bit LLM, showcasing the significant impact these models could have on portable and IoT devices.

Our discussions highlighted that 1.58-bit LLMs, such as BitNet b1.58, are not just small improvements but rather a substantial advancement in AI efficiency and capability. These models create new possibilities for AI applications, enhancing accessibility and efficiency, particularly in environments with limited resources. The shift towards 1.58-bit LLMs represents a crucial milestone in developing more sustainable and potent AI models.

Looking ahead, the ongoing progress of these models and their integration across various sectors will undoubtedly influence the future of AI development. It is essential to approach this future with a deep understanding of the vast potential and ethical implications associated with such advanced technology.

References

[BZB+19] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: reasoning about physical commonsense in natural language. CoRR abs/1911.11641, 2019【65†source】.

[CCKS23] Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. QuIP: 2-bit quantization of large language models with guarantees. CoRR abs/2307.13304, 2023【66†source】.

[CLC+19] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. CoRR abs/1905.10044, 2019【67†source】.

[Com23] Together Computer. Redpajama: an open dataset for training large language models, 2023【68†source】.

[FAHA23] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. OPTQ: accurate quantization for generative pre-trained transformers. In The Eleventh International Conference on Learning Representations, 2023【69†source】.

[HCB+19] Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Xu Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V. Le, Yonghui Wu, and Zhifeng Chen. Gpipe: Efficient training of giant neural networks using pipeline parallelism. In Advances in Neural Information Processing Systems, pages 103–112, 2019【70†source】.

[Hor14] Mark Horowitz. 1.1 computing's energy problem (and what we can do about it). In 2014 IEEE International Conference on Solid-State Circuits Conference ISSCC 2014 Digest of Technical Papers, San Francisco, CA, USA, February 9-13, 2014, pages 10–14, 2014【71†source】.

[KLZ+23] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles, 2023【72†source】.

[LTT+23] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, and Song Han. AWQ: activation-aware weight quantization for LLM compression and acceleration. CoRR abs/2306.00978, 2023【73†source】.

[MCKS18] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? A new dataset for open book question answering. CoRR abs/1809.02789, 2018【74†source】.

[MXBS16] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016【75†source】.

[PKL+16] Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers. The Association for Computer Linguistics, 2016【76†source】.

[WLG17] Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In Leon Derczynski, Wei Xu, Alan Ritter, and Tim Baldwin, editors, Proceedings of the 3rd Workshop on Noisy User-generated Text NUT@EMNLP 2017, Copenhagen, Denmark, September 7, 2017, pages 94–106. Association for Computational Linguistics, 2017【77†source】.

[WMC+23] Lei Wang, Lingxiao Ma, Shijie Cao, Ningxin Zheng, Quanlu Zhang, Jilong Xue, Ziming Miao, Ting Cao, and Yuqing Yang. Ladder: Efficient tensor compilation on customized data format. In OSDI 2023【78†source】.

[WMD+23] Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Huaijie Wang, Lingxiao Ma, Fan Yang, Ruiping Wang, Yi Wu, and Furu Wei. Bitnet: Scaling 1-bit transformers for large language models. CoRR abs/2310.11453, 2023【79†source】.

[XLS+23] Guangxuan Xiao, Ji Lin, Mickaël Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: accurate and efficient post-training quantization for large language models. In International Conference on Machine Learning ICML 2023, July 23-29, 2023, Honolulu, Hawaii, USA, 2023【80†source】.

[YBS19] Vikas Yadav, Steven Bethard, and Mihai Surdeanu. Quick and (not so) dirty: Unsupervised selection of justification sentences for multi-hop question answering. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, EMNLP-IJCNLP 2019【81†source】.

[ZHB+19] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: can a machine really finish your sentence? In Proceedings of the 57th Conference of the Association for Computational Linguistics, pages 4791–4800, 2019【82†source】.

[ZS19] Biao Zhang and Rico Sennrich. Root mean square layer normalization. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, Advances in Neural Information Processing Systems, pages 12360–12371, 2019【83†source】.

[ZZL22] Yichi Zhang, Zhiru Zhang, and Lukasz Lew. PokeBNN: A binary pursuit of lightweight accuracy. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12465–12475. IEEE, 2022【84†source】.

Introduction

The rise of Large Language Models (LLMs) in the Artificial Intelligence (AI) realm signifies a major milestone. These sophisticated models have demonstrated remarkable proficiency in various natural language processing tasks. Nevertheless, their escalating size and complexity have led to concerns about deployment issues and the environmental and economic repercussions, chiefly due to their high energy consumption.

Typically, LLMs function with 16-bit floating values, with matrix multiplication forming the core of their computational workload. Relying on floating-point operations has resulted in significant energy expenses, constraining their efficiency and practicality. To address these issues, the industry has shifted towards post-training quantization, reducing the precision of weights and activations to craft low-bit models for inference. Although widely embraced, this method, particularly with 4-bit variants, has not been entirely optimal.

In this context, the recent introduction of 1-bit model architectures, like BitNet, signifies a groundbreaking advancement. BitNet transforms the computational framework of LLMs by substituting energy-intensive floating-point operations with integer addition. This shift in approach offers a notable decrease in energy consumption, tackling the power limitations that often impede computational performance. Consequently, the adoption of 1-bit LLMs not only aims to boost the efficiency of current models but also aims to redefine AI computational strategies, paving the way for advanced applications and sustainable AI progress.

Hardware Optimization for 1.58-bit LLMs

Innovations like 1.58-bit LLMs, such as BitNet b1.58, mark a new era of computational efficiency and performance in AI. However, to maximize these models' potential, specialized hardware systems tailored to their unique computational needs are essential.

Projects like Groq5 have demonstrated progress by developing dedicated hardware units like Language Processing Units (LPUs) for LLMs. These solutions optimize LLM performance by aligning with their computation patterns and energy efficiency requirements.

The advent of 1.58-bit models presents an opportunity to create hardware and systems optimized for these models. BitNet's unique computation approach, with minimal multiplication operations for matrix multiplication, necessitates a fresh perspective on hardware design. Tailored hardware could significantly enhance energy efficiency, processing speed, and overall computational performance.

Designing hardware specifically for 1.58-bit LLMs not only boosts model capabilities but also drives sustainable and cost-effective AI solutions. This path represents a critical advancement in AI technology, harmonizing computational efficiency with the escalating demands of advanced AI models.

Integration of Mixture-of-Experts (MoE) Models with 1.58-bit LLMs

Mixture-of-Experts (MoE) models have become a cost-effective strategy within the domain of Large Language Models (LLMs). These models assign computational tasks to different 'expert' networks, enhancing efficiency and specialization. Yet, the widespread use of MoE models has been hindered by high memory usage and significant inter-chip communication challenges.

The introduction of 1.58-bit LLMs, exemplified by BitNet b1.58, offers a remedy to these issues. The reduced memory requirements of 1.58-bit LLMs decrease the devices necessary for deploying MoE models, addressing a key limitation in their implementation. Moreover, this memory reduction notably reduces the overhead associated with transferring activations between networks.

A significant benefit of incorporating 1.58-bit LLMs in MoE models is the potential elimination of inter-chip communication overhead. If the entire model could fit on a single chip, it would transform the deployment and utility of MoE models, enhancing their viability and effectiveness across various applications. This advancement could lead to more sophisticated, specialized, and efficient AI systems by combining the strengths of MoE models with the efficiency gains of 1.58-bit LLMs.

Potential of 1.58-bit LLMs in Handling Long Sequences

In the era of Large Language Models (LLMs), efficiently managing long data sequences is crucial. Handling long sequences poses a challenge due to high memory usage, especially with Key-Value (KV) caches in use. This challenge has limited the expansion of LLM capabilities.

The innovation of BitNet b1.58, an example of 1.58-bit LLMs, is a significant leap in overcoming this hurdle. By reducing activation size from 16 bits to 8 bits, BitNet b1.58 doubles the context length manageable with the same resources. This advancement allows processing of longer sequences without a proportional increase in memory usage.

Future optimization potential exists for 1.58-bit LLMs. The ability to compress these activations to 4 bits or lower presents opportunities for more efficient handling of extended sequences. This optimization is crucial to meet the increasing demand for processing extensive data sequences in AI applications, from natural language understanding to complex data analysis.

The progress in 1.58-bit LLMs like BitNet b1.58 signifies a significant advancement towards supporting long sequences effectively. This not only enhances current LLM capabilities but also paves the way for more robust and adaptable models in the future.

Conclusion

The development of BitNet b1.58 and the evolution of 1.58-bit LLMs represent a paradigm shift in the efficiency and application of Large Language Models. These models are redefining what is possible in AI by enabling new scaling laws that align model performance with inference cost. This breakthrough is evident in comparisons showing that a 13B BitNet b1.58 model outperforms a 3B FP16 LLM in terms of latency, memory usage, and energy consumption. Similarly, 30B and 70B BitNet b1.58 models demonstrate superior efficiency against 7B and 13B FP16 LLMs, respectively.

Crucially, the scalability of BitNet b1.58 has been proven through rigorous training with 2 trillion tokens, following methodologies of leading open-source models like StableLM-3B. The impressive performance of BitNet b1.58 across a range of benchmarks, including Winogrande, PIQA, SciQ, LAMBADA, and ARC-easy, underscores its strong generalization capabilities.

The implications of these findings are profound. 1.58-bit LLMs like BitNet b1.58 are not only more efficient but also open new possibilities for AI applications in areas previously constrained by computational and energy limitations. This includes enhanced deployment on edge and mobile devices, integration with MoE models, and superior handling of long sequences.

Looking ahead, the ongoing evolution of 1.58-bit LLMs presents an exciting frontier in AI research and application. The development of specialized hardware, further optimization of model architectures, and exploration of new applications in diverse fields are areas ripe for innovation. As we stand on the cusp of this new era in AI, it is clear that 1.58-bit LLMs will play a pivotal role in shaping the future of computing and artificial intelligence.

References

[BZB+19] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: reasoning about physical commonsense in natural language. CoRR abs/1911.11641, 2019【65†source】.

[CCKS23] Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. QuIP: 2-bit quantization of large language models with guarantees. CoRR abs/2307.13304, 2023【66†source】.

[CLC+19] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. CoRR abs/1905.10044, 2019【67†source】.

[Com23] Together Computer. Redpajama: an open dataset for training large language models, 2023【68†source】.

[FAHA23] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. OPTQ: accurate quantization for generative pre-trained transformers. In The Eleventh International Conference on Learning Representations, 2023【69†source】.

[HCB+19] Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Xu Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V. Le, Yonghui Wu, and Zhifeng Chen. Gpipe: Efficient training of giant neural networks using pipeline parallelism. In Advances in Neural Information Processing Systems, pages 103–112, 2019【70†source】.

[Hor14] Mark Horowitz. 1.1 computing's energy problem (and what we can do about it). In 2014 IEEE International Conference on Solid-State Circuits Conference ISSCC 2014 Digest of Technical Papers, San Francisco, CA, USA, February 9-13, 2014, pages 10–14, 2014【71†source】.

[KLZ+23] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles, 2023【72†source】.

[LTT+23] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, and Song Han. AWQ: activation-aware weight quantization for LLM compression and acceleration. CoRR abs/2306.00978, 2023【73†source】.

[MCKS18] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? A new dataset for open book question answering. CoRR abs/1809.02789, 2018【74†source】.

[MXBS16] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016【75†source】.

[PKL+16] Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers. The Association for Computer Linguistics, 2016【76†source】.