Enhancing Nuclear Medicine Through Innovative AI Models: Integrating Large Language Models and SmoothQuant for Advanced Imaging Precision

SANSKAR SAXENA[1], VINOD KUMAR[2], JATIN CHOPRA[3], PRANJAL GUPTA[4], DAAMINI BATRA[5], BHAUMIK TYAGI[6]

[1] *Undergraduate Student, (Computer Science Engineering), MIT, Uttar Pradesh, India*

[2, 3, 4] *Undergraduate Student, (Computer Science Engineering), ADGITM, Delhi, India*

[5] *Undergraduate Student, (Information Technology), ADGITM, Delhi, India*

[6] *Jr. Research Scientist, Delhi, India*

Abstract—Large language models (LLMs) exhibit superior performance but demand substantial computational and memory resources. Quantization emerges as a viable strategy to alleviate memory requirements and expedite inference. Nonetheless, extant quantization methods encounter challenges in preserving accuracy and ensuring hardware efficiency. This research introduces SmoothQuant, an innovative, training-free, and accuracy-preserving post-training (PTQ) approach, specifically designed to facilitate 8-bit weight and 8-bit activation (W8A8) quantization for LLMs. Leveraging the insight that weights are amenable to quantization while activations pose challenges, SmoothQuant addresses this imbalance by mitigating activation outliers. With the continuous evolution of artificial intelligence (AI) technologies, integrating innovative AI models into the field of nuclear medicine holds immense promise. This research paper explores the synergistic potential of combining Large Language Models (LLMs) and SmoothQuant, a state-of-the-art post-training quantization technique, to achieve advanced imaging precision in nuclear medicine. This is achieved through an offline migration of the quantization complexity from activations to weights, employing a mathematically equivalent transformation. This research presents a turn-key solution that not only optimizes hardware utilization but also contributes to the democratization of LLMs by mitigating associated costs. The study investigates the impact of this integration on data analysis, interpretation, and overall diagnostic accuracy.

Indexed Terms—Nuclear Medicine, Artificial Intelligence, Large Language Models, Imaging Precision, Quantization.

I. INTRODUCTION

The introduction provides a comprehensive overview of the current landscape within the domain of nuclear medicine, introducing a pioneering methodology that integrates Large Language Models (LLMs) and SmoothQuant. The research elucidates the aims, objectives, and significance of this integrated approach. Trustworthiness emerges as a pivotal concept within the academic discourse on artificial intelligence (AI), underscored by concerns about misrepresentation and hidden biases in the existing literature [1]. As practitioners and researchers within a clinical nuclear medicine department, the question arises regarding the reliability of the nuclear medicine content generated by Chat-GPT. A critical consideration lies in assessing whether the outputs of AI tools, particularly in the context of nuclear medicine, can be deemed trustworthy. The milestone of physician training often culminates in summative board or licensing examinations designed to ensure public protection, uphold professional standards, and evaluate physicians based on a defined body of knowledge [2]. In tort law, the benchmark for determining negligent practice is often rooted in the knowledge and practices accepted by professionals in common law jurisdictions [3]. If AI tools aspire to function as potential aides or substitutes for physicians, their performance may reasonably be held to a comparable standard. Recent endeavors in this direction have yielded varied outcomes. For instance, Shelmerdine et al. conducted a study evaluating the performance of a commercially available AI tool, with a CE-conformity label, in the radiograph reporting section of the United Kingdom Fellowship of the Royal College of Radiologists (FRCR). The tool exhibited subpar performance, struggling to pass two out of ten mock examinations and ranking last among its 26 human peers, requiring special dispensation [4].

Conversely, ChatGPT demonstrated the capability to pass or approach passing all three segments of the United States Medical Licensing Exam (USMLE) without additional training or prompts. Similarly, a Chinese AI tool named Xiaoyi, meaning "little doctor," demonstrated the potential, with training, to pass the Chinese Medical Licensing Exam [5]. These instances underscore the nuanced landscape of AI performance in medical assessments and highlight the need for meticulous scrutiny and evaluation to ensure alignment with established professional standards.

Utilizing Graphics Processing Units (GPUs) or a set of 5 NVIDIA A100 GPUs, each with a 80GB capacity, is commonplace for inference tasks. However, the substantial computational and communication overhead associated with this approach may render the inference latency impractical for real-world applications. A promising avenue to alleviate the computational burden of Large Language Models (LLMs) involves quantization, as documented in the literature [6,7]. This process entails representing weights and activations with low-bit integers, resulting in diminished GPU memory requirements in both size and bandwidth. Additionally, quantization serves to expedite compute-intensive operations, such as General Matrix Multiplication (GEMM) in linear layers and Binary Matrix Multiplication (BMM) in attention mechanisms. For instance, the adoption of INT8 quantization for weights and activations holds the potential to reduce GPU memory usage by half and nearly double the throughput of matrix multiplications compared to the conventional Floating Point 16 (FP16) representation.

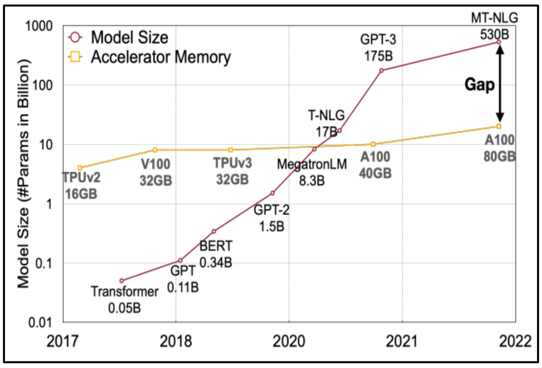

Fig. 1: The evolving model size of large language models.

The evolving model size of large language models has outpaced the growth rate of GPU memory in recent years, resulting in a substantial disparity between the burgeoning demand for memory and the available supply. To address this discrepancy, the application of quantization and model compression techniques emerges as a viable strategy. These approaches play a pivotal role in narrowing the gap between the escalating requirements of large language models and the limited capacity of GPU memory.

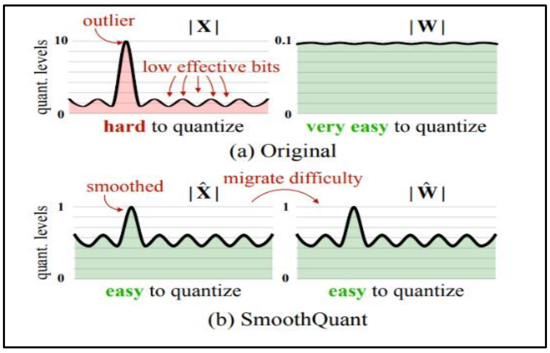

Fig. 2: SmoothQuant's intuition

SmoothQuant addresses the challenge in quantizing activations, as depicted in Figure 2. The activation variable, denoted as X, poses difficulty in quantization due to outliers that extend the quantization range, leaving a limited number of effective bits for the majority of values. To mitigate this, we strategically transfer the scale variance from activations (X) to weights (W) during an offline process. This adjustment effectively reduces the quantization complexity of activations. The resulting smoothed activation, denoted as Xˆ, and the modified weight, denoted as Wˆ, both exhibit ease of quantization, thereby enhancing the efficiency of the quantization process.

Introducing SmoothQuant, a post-training quantization (PTQ) solution designed for Large Language Models (LLMs). A crucial insight underpinning SmoothQuant is the recognition that, while quantizing activations proves challenging due to the influence of outliers [6], tokens across channels demonstrate similar variations. Leveraging this observation, SmoothQuant strategically shifts the quantization complexity from activations to weights during an offline process (Figure 2). The methodology introduces a per-channel scaling transformation that is mathematically equivalent, thereby markedly smoothing magnitudes across channels and rendering the model more amenable to the quantization process.

II. LITERATURE REVIEW

A thorough examination of the extant literature concerning artificial intelligence (AI) applications within the domain of nuclear medicine serves as the foundational framework for this research. This review illuminates pivotal advancements, challenges, and discernible gaps in the current comprehension of AI-supported imaging applications within the field of nuclear medicine. Large language models (LLMs) have garnered considerable attention in the literature. Pre-trained language models have demonstrated exceptional performance across diverse benchmarks through scale augmentation. Notably, GPT-3 [8] represents a pioneering instance of an LLM exceeding 100 billion parameters, achieving noteworthy few-shot/zero-shot learning outcomes. Subsequent endeavors [9,10] have further extended the frontiers of scaling, surpassing 500 billion parameters. However, the expanding scale of language models introduces substantial complexities in terms of inference costs and computational demands. This study aims to demonstrate the efficacy of the proposed method in quantizing the three most extensive openly available LLMs, namely OPT-175B [11] and BLOOM-176B [12]. This quantization methodology is devised to alleviate memory costs and expedite the inference process.

Quantization strategies for Large Language Models (LLMs) have been explored in various implementations. GPTQ [13] focuses on quantizing weights exclusively, omitting activations (refer to Appendix A for a brief discussion). Zero-Quant and nuQmm [14] employ per-token and group-wise quantization schemes for LLMs, necessitating customized CUDA kernels. However, these methods are limited in scale, evaluating models up to 20B and 2.7B, respectively, and falter in sustaining the performance levels exhibited by larger LLMs like OPT-175B. LLM.int8() adopts a mixed INT8/FP16 decomposition approach to mitigate activation outliers. Regrettably, this implementation introduces considerable latency overhead, potentially resulting in slower inference than FP16. Outlier Suppression [15] utilizes non-scaling Layer-Norm and token-wise clipping to address activation outliers but is effective only for smaller language models like BERT [16] and BART [17], proving inadequate for maintaining accuracy in the case of LLMs.

III. RELATED BACKGROUND

Quantization maps a high-precision value into discrete levels. We study integer uniform quantization [18] (specifically INT8) for better hardware support and efficiency. The quantization process can be expressed as:

$$\overline{X^{INT8}} = \left\lceil \frac{X^{FP16}}{\Delta} \right\rfloor, \quad \Delta = \frac{\max(|\mathrm{X}|)}{2^{\mathrm{N}-1} - 1}, \qquad (1)$$

where X is the floating-point tensor, X^{-} is the quantized counterpart, Δ is the quantization step size, $\lceil \cdot \rfloor$ is the rounding function, and N is the number of bits (8 in our case). Here we assume the tensor is symmetric at 0 for simplicity; the discussion is similar for asymmetric cases.

Review of Quantization Difficulty

1. Activations are harder to quantify than weights: The weight distribution is quite uniform and flat, which is easy to quantify. Previous work has shown that quantizing the weights of LLMs with INT8 or even with INT4 does not degrade accuracy [6,7], which echoes our observation.

2. Outliers make activation quantization difficult: The scale of outliers in activations is $\sim$ 100× larger than most of the activation values. In the case of per-tensor quantization (Equation 1), the large outliers dominate the maximum magnitude measurement, leading to low effective quantization bits/levels (Figure 2) for non-outlier channels: suppose the maximum magnitude of channel i is mi, and the maximum value of the whole matrix is m, the effective quantization levels of channel i is $2^8 \cdot mi/m$. For non-outlier channels, the effective quantization levels would be very small (2-3), leading to large quantization errors.

3. Outliers persist in fixed channels: Outliers manifest within a limited subset of channels, exhibiting a persistent presence across all tokens. Notably, if a channel contains an outlier, this outlier consistently appears in all tokens. The variation in magnitudes among channels for a specific token is substantial, with some channels containing very large activations while others possess smaller magnitudes. However, the variance in magnitudes across tokens for a given channel is relatively minor, indicating consistent largeness in outlier channels. Given the enduring nature of outliers and the limited variance within each channel, employing per-channel quantization [19], where a distinct quantization step is applied to each channel, presents a viable approach. This strategy is anticipated to yield significantly smaller quantization errors compared to per-tensor quantization, as the small variance within each channel and the persistent presence of outliers can be better accommodated, whereas per-token quantization provides marginal assistance in this context.

- *Migrate the quantization difficulty from activations to weights.*

We aim to choose a per-channel smoothing factor *s* such that $\hat{X} = X\text{diag}(s)^{-1}$ is easy to quantize. To reduce the quantization error, we should increase the effective quantization bits for all the channels. The total effective quantization bits would be largest when all the channels have the same maximum magnitude. Therefore, a straightforward choice is $s_j = \max(|X_j|)$, j = 1, 2, ..., C_i, where *j* corresponds to *j-th* input channel. This choice ensures that after the division, all the activation channels will have the same maximum value, which is easy to quantify [20]. It is essential to acknowledge that the range of activations is dynamic and exhibits variability across different input samples. In our methodology, we assess the scale of activation channels using calibration samples derived from the pre-training dataset. However, the current formula tends to shift all quantization challenges onto the weights. Consequently, this approach results in suboptimal model performance attributable to activation quantization errors. To address this, there is a necessity to distribute the quantization difficulty more evenly between weights and activations, ensuring that both are amenable to the quantization process. This adjustment is imperative for enhancing the overall efficiency of the quantization methodology.

Here we introduce a hyper-parameter, migration strength α, to control how much difficulty we want to migrate from activation to weights, using the following equation:

$$s_j = max(|Xj|)^{\alpha} / max(|Wj|)^{1-\alpha} \qquad (2)$$

IV. METHODOLOGY

The methodology section outlines the research design, data sources, and experimental setup. It describes how LLMs, particularly Large Language Models like ChatGPT, and SmoothQuant are integrated into the nuclear medicine imaging pipeline. Details of datasets and evaluation metrics are also provided. The paper delves into the intricacies of SmoothQuant as a post-training quantization technique. It provides a detailed explanation of how this method optimizes model performance without compromising imaging precision in nuclear medicine applications. The core of the research investigates the combined impact of LLMs and SmoothQuant on imaging precision in nuclear medicine. This includes an analysis of how these technologies enhance the accuracy of image interpretation, minimize information loss during quantization, and contribute to more precise diagnostic outcomes.

Implementing SmoothQuant within Transformer blocks involves focusing on linear layers, which constitute the majority of parameters and computations in Large Language Models (LLMs). The default approach involves applying scale smoothing to input activations in self-attention and feed-forward layers, quantizing all linear layers using an 8-bit weight and 8-bit activation (W8A8) format. Additionally, quantization is applied to Binary Matrix Multiplication (BMM) operators involved in attention computations.

Figure 3 illustrates the designed quantization flow for transformer blocks. This process encompasses the quantization of inputs and weights within compute-intensive operators such as linear layers and BMM in attention layers, utilizing integer (INT) representation. This approach ensures an efficient and effective application of SmoothQuant within the critical components of transformer blocks in LLMs.

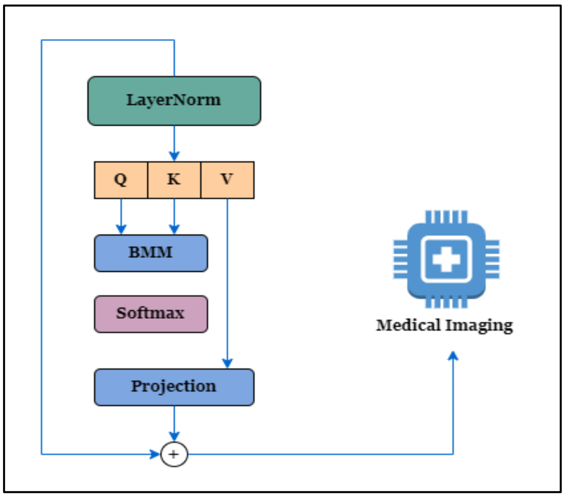

Fig 3: SmoothQuant's precision mapping for a Transformer block integrated with Medical Imaging

For activation smoothing, a migration strength parameter (α) of 0.5 is identified as a generally optimal value for all OPT and BLOOM models. For GLM-130B, where activations present greater quantization challenges, α is set to 0.75. The determination of a suitable α is accomplished through a swift grid search conducted on a subset of the Pile validation set. To derive the activation statistics necessary for calibration of smoothing factors and static quantization step sizes, a single calibration process is conducted using 512 randomly selected sentences from the pre-training dataset, Pile. This calibrated model, with both smoothed and quantized activations, is subsequently employed uniformly across all downstream tasks. This approach enables the evaluation of the generality and zero-shot performance of the quantized Large Language Models (LLMs).

V. RESULTS AND DISCUSSIONS

Quantization schemes are evaluated in Figure 4, depicting the inference latency of various quantization approaches utilizing our PyTorch implementation. Notably, the inference latency decreases as the quantization granularity becomes coarser, progressing from O1 to O3. Furthermore, static quantization demonstrates a considerable acceleration in inference compared to dynamic quantization, attributed to the elimination of the need to calculate quantization step sizes at runtime. SmoothQuant consistently outperforms the FP16 baseline across all settings, whereas LLM.int8() typically exhibits slower performance. In light of these observations, a recommendation is made to consider employing a coarser quantization scheme if the acceptable level of accuracy permits, as it contributes to reduced latency without compromising performance.

Determining an appropriate migration strength parameter, denoted as α (as per Equation 2), is crucial for achieving a balanced quantization difficulty between weights and activations. An ablation study is conducted to assess the impact of different α values on OPT-175B with LAMBADA, as depicted in Figure 10. Observations indicate that when α is excessively small (e.g., 0.6), the quantization of weights becomes challenging. Optimal quantization errors for both weights and activations, ensuring the preservation of model performance post-quantization, are achieved only within a specific range of α values, commonly referred to as the "sweet spot" region, typically ranging from 0.4 to 0.6. This underscores the importance of selecting an appropriate α to strike the desired balance in quantization difficulties for weights and activations.

GPU Latency (ms) of different quantization schemes. The coarser the quantization scheme (from per-token to per-tensor, dynamic to static, O1 to O3, the lower the latency. SmoothQuant achieves lower latency compared to FP16 under all settings, while LLM.int8() is mostly slower.

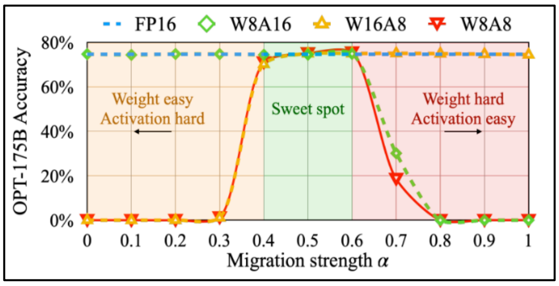

Figure 4: Migration strength α

This makes both activations and weights easy to quantize. If the α is too large, weights will be hard to quantize; if too small, activations will be hard to quantize.

CONCLUSION

Introducing SmoothQuant, a meticulously designed post-training quantization methodology aimed at facilitating seamless 8-bit weight and activation quantization for Large Language Models (LLMs) comprising up to 530 billion parameters.

SmoothQuant operates uniformly across all General Matrix Multiplications (GEMMs) within LLMs, effectively reducing inference latency and minimizing memory usage in comparison to the mixed-precision activation quantization baseline. The seamless integration of SmoothQuant into PyTorch and FasterTransformer yields substantial benefits, achieving up to a 1.56× acceleration in inference speed and a 50% reduction in memory footprint. This integration positions SmoothQuant as a transformative tool, democratizing the application of LLMs by providing a comprehensive and efficient solution to curtail serving costs.

Seeking guidance from a licensed medical professional is advisable for accurate and current information. Furthermore, the information provided is based on the knowledge available up to 2021, and therefore, it may not encompass newer developments or updates in the medical field. This seemingly modest acknowledgment underscores the limitations inherent in Large Language Models (LLMs). However, our assessment contends that this response inadequately addresses the substantive issue at hand – the provision of not merely unhelpful but factually inaccurate and potentially misleading answers delivered with unwarranted confidence. Caution is warranted when considering assertions that LLMs can be employed for tasks such as summarizing medical records, drafting authorization letters for insurers justifying treatment costs, or serving as decision-support tools for diagnosis. In summary, while ChatGPT exhibits a capability to generate content that appears convincing, including abstracts with references that may deceive peer reviewers, our initial analysis indicates a notable shortfall in demonstrating the requisite knowledge expected of a certified nuclear medicine physician in Europe, especially in the context of a standardized examination. Candidates preparing for exams or practicing physicians are advised to independently verify the validity of statements generated by these models, recognizing the potential for unreliability. Based on the performance observed in this preliminary analysis, we currently find no evidence that ChatGPT poses a threat to the integrity of online nuclear medicine examinations. However, given the rapid pace of development, this circumstance may evolve in the near future.

Nevertheless, we assert that the current capabilities, or lack thereof, of ChatGPT highlight an immediate need to systematically address the ethical challenges associated with such systems. The education and training of clinicians must adapt to accommodate the extent of agency wielded by tools like ChatGPT in the medical field. In a somewhat whimsical manner, and in contrast to Hinton's advice, maintaining the training of nuclear medicine physicians and radiologists is deemed prudent, at least for the foreseeable future.

REFERENCES

[1] Kitamura FC, Marques O. Trustworthiness of artificial intelligence models in radiology and the role of explainability. J Am Coll Radiol. 2021;18:1160–2. https://doi.org/10.1016/j.jacr.2021.02.008.

[2] Mirzaei S, Hustinx R, Prior JO, Ozcan Z, Boubaker A, Farsad M, et al. Improving nuclear medicine practice with UEMS/ EBNM committees. J Nucl Med: Off Publ Soc Nucl Med. 2020;61:18N-20N.

[3] Sokol DK. How good a doctor do you need to be? BMJ Br Med J. 2012;345:e7858. https://doi.org/10.1136/bmj.e7858.

[4] Shelmerdine SC, Martin H, Shirodkar K, Shamshuddin S, Weir-McCall JR, Collaborators F-AS. Can artificial intelligence pass the Fellowship of the Royal College of Radiologists examination? Multi-reader diagnostic accuracy study. BMJ. 2022;379:e072826. https://doi.org/10.1136/bmj-2022-072826.

[5] Rampton V, Ko A. Robots, radiologists, and results. BMJ. 2022;379:o2853. https://doi.org/10.1136/bmj.o2853.

[6] Dettmers, T., Lewis, M., Belkada, Y., and Zettlemoyer, L. Llm.int8(): 8-bit matrix multiplication for transformers at scale. arXiv preprint https://arXiv:2208.07339, 2022.

[7] Yao, Z., Aminabadi, R. Y., Zhang, M., Wu, X., Li, C., and He, Y. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers, 2022. URL https://arxiv.org/abs/2206.01861.

[8] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. Advances in neural information processing systems, 33: 1877–1901, 2020b.

[9] Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al. Scaling language models: Methods, analysis & insights from training gopher. arXiv preprint arXiv:2112.11446, 2021.

[10] Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G. Korthikanti, V., et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. arXiv preprint arXiv:2201.11990, 2022.

[11] Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. Opt: Open pre-trained transformer language models, 2022. URL https://arxiv.org/abs/2205.01068

[12] Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilic, S., Hesslow, ´ D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., et al. Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint https://arXiv:2211.05100, 2022.

[13] Frantar, E., Ashkboos, S., Hoefler, T., and Alistarh, D. Gptq: Accurate post-training quantization for generative pretrained transformers. arXiv preprint https://arXiv:2210.17323, 2022.

[14] Park, G., Park, B., Kwon, S. J., Kim, B., Lee, Y., and Lee, D. nuqmm: Quantized matmul for efficient inference of large-scale generative language models. arXiv preprint https://arXiv:2206.09557, 2022.

[15] Wei, X., Zhang, Y., Zhang, X., Gong, R., Zhang, S., Zhang, Q., Yu, F., and Liu, X. Outlier suppression: Pushing the limit of low-bit transformer language models, 2022. URL https://arxiv.org/abs/2209.13325.

[16] Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT 2019, pp. 4171– 4186. Association for Computational Linguistics, 2019.

[17] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461, 2019

[18] Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., and Kalenichenko, D. Quantization and training of neural networks for efficient integerarithmetic-only inference. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2704–2713, 2018.

[19] Bondarenko, Y., Nagel, M., and Blankevoort, T. Understanding and overcoming the challenges of efficient transformer quantization. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 7947–7969, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021. emnlp-main.627

[20] Roemmele, M., Bejan, C. A., and Gordon, A. S. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In Logical Formalizations of Common-sense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011. AAAI, 2011. URL http://www. aaai.org/ocs/index.php/SSS/SSS11/paper/view /2418.