Methods of Reducing Computational Requirements for Large Language Models

Oleksandr Kononov *

* *South East Technological University, Cork Road, Waterford, Ireland*
(e-mail: 20071032@mail.wit.ie).

Abstract: TODO

Keywords: Artificial intelligence, Neural networks,

1. INTRODUCTION

1.1 Background

Large Language Model (LLM) is neural network model which is very capable at natural language tasks such as text generation, text summarization, translation, and more. Vaswani et al. (2017) proposed the novel Transformer architecture which has revolutionized the field by introducing more efficient multi-headed self-attention mechanism compared to Recurrent Neural Networks that came before. With the improvements in training times, better parallelization on GPUs and overall quality of output. Following this, OpenAI have used this novel Transformer architecture to design and develop their Generative Pre-Trained Transformer (GPT) LLMs the following years. In particular, the release of GPT-3 in 2020, has sparked a global interest in the continued development of LLMs from various companies such as Meta with LLaMa, Google with Gemma, Antropic with Claude, etc.

Touvron et al. (2023) developed a series of open-weight LLMs called LLaMa, ranging from 7B parameters to 65B parameters. Their paper demonstrates, using various benchmark tests, we can draw a correlation between the increase in LLM parameters and the scores that it can achieve from the benchmark tests. There are challenges with regards to the computational requirements necessary for inferencing LLMs,“the compute and memory requirements of state-of-the-art language models have grown by three orders of magnitude in the last three years, and are projected to continue growing far faster than hardware capabilities” (Bommasani et al., 2022, p. 97).

1.2 Problem Statement and Motivation

1.3 Research Objectives

1.4 Research Questions

This paper aims to answer the following Research Questions (RQ):

- **RQ1** How the various quantization methods that are in-use work?
- **RQ2** What are quality impacts of quantization on LLMs?
- **RQ3** What are the performance improvements that can be expected of a quantized LLM on low-end hardware?

2. PRELIMINARY LITERATURE REVIEW

3. WORKING THEORY

4. RESEARCH DESIGN

4.1 Introduction

4.2 Design

5. CONCLUSION

ACKNOWLEDGEMENTS

Place acknowledgments here. (left-over from IFAC template, not sure if I will need it yet)

REFERENCES

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J.Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D.E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P.W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X.L., Li, X., Ma, T., Malik, A., Manning, C.D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J.C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J.S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A.W., Tramèr, F., Wang, R.E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S.M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. (2022). On the opportunities and risks of foundation models. URL https://arxiv.org/abs/2108.07258.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). Llama: Open and efficient foundation language models. URL https://arxiv.org/abs/2302.13971.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. URL https://arxiv.org/abs/1706.03762.

Appendix A. APPENDIX A