Effiziente Reduktion großer Sprachmodelle: Methodenvergleich und Anwendungsstrategien für Energieoptimierung und Hardware-Kompatibilität

Effizient reduction of large language models: Comparing methods and develop strategies to optimize energy footprint and hardware compatibility

Exposé zur Masterarbeit

Dr. Thomas Schmitt
Fachhochschule Südwestfalen
18. April 2024

Autor: Dr. Thomas Schmitt
Referent: Prof. Dr. Heiner Giefers
Korreferent: Prof. Dr. ???
Eingereicht: 18. April 2024

1 Exposé zur geplanten Masterarbeit

1.1 Einleitung - Überblick

Die geplante Arbeit untersucht aktuelle Verfahren zur Effizienzsteigerung von Large Language Models, vergleicht diese miteinander und versucht sie (ggf.) sinnvoll zu kombinieren. Effizienzsteigerung bedeutet in diesem Kontext, die Verringerung von Hardware Anforderungen, insbesondere an die Größe des Hauptspeichers und die Geschwindigkeit der GPU bzw. CPU. Ziel ist Verfahren zu entwickeln bzw. Möglichkeiten zu beschreiben, welche es erlauben, bei akzeptabler Qualität solche Modelle auf weniger performanten Hardware Infrastrukturen wie z.Bsp. embedded Computern lokal zu betreiben.
Als kritisches Maß gilt hier, neben der Qualität der Inferenzergebnisse, vor allem die Latenzzeit bis das Modell eine Antwort liefert. Hierbei gelten für die jeweiligen Anwendungsfälle unterschiedlich akzeptable Zeiträume. Ein aktuelles LLM bietet allerdings auf einem System, welches nicht mit einer (NVIDIA) GPU ausgestattet ist, Latenzzeiten im Bereich von Minuten oder sogar Stunden. Dies ist in fast allen denkbaren Anwendungsfällen inakzeptabel.

1.2 Einleitung - Verfahren zur Steigerung der Effizienz

Es existieren verschiedene Verfahren, um LLMs in Größe und Komplexität zu reduzieren, die in der Einleitung der Arbeit genauer besprochen und untersucht werden sollen:

- Pruning - Umfasst verschiedene Verfahren zum ausdünnen von Neuronalen Netzwerken. Häufig werden 0 Werte und/oder redundante Werte entfernt. Innerhalb der Arbeit werden verschieden Varianten diskutiert und ausgewählt. [ZG17] [Sun+23] [FA23]
- Knowledge Distilliation - Hierbei wird ein Schüler Modell mit Hilfe eines Lehrer Modells (In meinem Fall Llama2) trainiert. Das Training erfolgt dabei auf die Vorhersagen des Lehrer Modells. Die zugrunde liegende Intention ist, dass damit das implizit erworbene Wissen des Lehrer Modells übernommen wird, welches selbst mit einer sehr großen Anzahl von Daten trainiert wurde [Xu+24]
- Low Rank Factorization (Lora) - Lora benutzt die Zerlegung von komplexeren Matrizen in niederrangigere Matrizen, [Xu+15] [Hu+21]. Mit diesem Verfahren kann ein Feintuning eines LLMs erfolgen, um dessen Effizienz zu steigern.
- Quantisierung - Mittels Quantisierung kann ein Modell auf eine geringere Bitbreite transformiert werden. Dies führt zu einer Effizienzsteigerung und erlaubt es darüber hinaus ein Modell auf Hardware Plattformen mit geringerer Bitbreite zu nutzen. Der damit einhergehende Genauigkeitsverlust, der einzelnen Parametern, kann zu schlechteren Inferenzergebnissen führen. [Fra+23]

1.3 Einleitung - Auswahl des Modells

Als geeignetes LLM wird das Modell LLama 2 ausgewählt, welches im Juli 2024 von der Firma Meta AI veröffentlicht wurde [Tou+23]. Das Modell ist nach Open-Source-Lizenz GPL3 lizenziert

und liegt in drei Varianten mit unterschiedlicher Parameteranzahl vor.
Für diese Arbeit wird die kleinste der drei verfügbaren Varianten mit sieben Milliarden Parametern ausgewählt. Neben dieser existieren noch zwei weitere Varianten des Modells, eine mit dreizehn und eine mit siebzig Milliarden Parametern.
Die Auswahl des kleinsten Modells erfolgt vor allem unter dem Gesichtspunkt der notwendigen Ressourcen und des damit einhergehenden zeitlichen Aufwands, für die verschiedenen zu untersuchenden Verfahren. Bei größeren Modellen ist davon auszugehen, dass die Verfahren zur Effizienz Optimierung deutlich mehr Zeit benötigen. Es wird erwogen, die an dem kleineren Modell erprobten vielversprechendsten Strategien, in einem zweiten Schritt, auch auf die größeren Varianten des LLama 2 Modells zu übertragen.

1.4 Ansatz für die Analyse

Die Analyse soll auf dem FH-Cluster der Fachhochschule Iserlohn für Machine Learning erfolgen. Erste Schritte zur Umsetzbarkeit wurden bereits unternommen und es konnten bis jetzt keine Probleme identifiziert werden.
Die Analyse gliedert sich in die folgenden Bereiche:

- Auswahl bzw. Entwicklung eines geeigneten Testszenarios
- Einzelanalyse der Effizienz Verfahren ggf. mit unterschiedlicher Parametrisierung
- Kombination der Technologien zur weiteren Effizienzsteigerung
- Vergleich der Ergebnisse von Einzel und Kombinations-Analyse
- Anwendung auf die größeren Varianten des LLama Modells

1.5 Auswertung

In der Auswertung sollen die Ergebnisse der einzelnen Technologien und der Kombination von Verfahren mit dem Ursprungsmodell verglichen werden. Insbesondere sollen auch die sinnvollen Grenzen der Effizienzsteigerung aufgezeigt und der Erwartungshorizont (ggf. anhand von Beispielen) skizziert werden.
Da das AI Umfeld einer schnellen Wandlung unterliegt, soll am Ende der Arbeit noch ein Status zu ggf. neuen oder verbesserten Technologien zur Effizienzsteigerung erfasst werden. Ebenfalls soll eine Einschätzung/Abgrenzung im Bezug zu ähnlichen Arbeiten in diesem Umfeld, abhängig von den Ergebnissen weiterer Recherche, vorgenommen werden.
Abschließend erfolgt ein Zusammenfassung mit dem Ausblick auf den weiteren Verlauf in diesem Forschungsfeld.

Literatur

[FA23] Elias Frantar und Dan Alistarh. *SparseGPT: Massive Language Models Can Be Accurately Pruned in One-Shot*. 2023. arXiv: 2301.00774 [cs.LG].

[Fra+23] Elias Frantar u. a. *GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers*. 2023. arXiv: 2210.17323 [cs.LG].

[Hu+21] Edward J. Hu u. a. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. arXiv: 2106.09685 [cs.CL].

[Knu97] Donald E. Knuth. *The Art of Computer Programming. Fundamental Algorithms*. 3. Aufl. Bd. 1. Reading, Massachusetts: Addison-Wesley, 1997.

[SCG13] Alex Shinn, John Cowan und Arthur A. Gleckler. *Scheme Reports Process*. 2013. URL: http://www.scheme-reports.org/ (besucht am 13. 07. 2020).

[Sun+23] Mingjie Sun u. a. *A Simple and Effective Pruning Approach for Large Language Models*. 2023. arXiv: 2306.11695 [cs.CL].

[Tou+23] Hugo Touvron u. a. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. 2023. arXiv: 2307.09288 [cs.CL].

[Xu+15] Yangyang Xu u. a. „Parallel matrix factorization for low-rank tensor completion“. In: *Inverse Problems Imaging* 9.2 (2015), S. 601–624. ISSN: 1930-8345. DOI: 10.3934/ipi.2015.9.601. URL: http://dx.doi.org/10.3934/ipi.2015.9.601.

[Xu+24] Xiaohan Xu u. a. *A Survey on Knowledge Distillation of Large Language Models*. 2024. arXiv: 2402.13116 [cs.CL].

[ZG17] Michael Zhu und Suyog Gupta. *To prune, or not to prune: exploring the efficacy of pruning for model compression*. 2017. arXiv: 1710.01878 [stat.ML].