Kalaidos Fachhochschule Schweiz
Die Hochschule für Berufstätige.

Skalierungseffekte der KI: Downscaling (2/2)

Dr. Yoshija Walter
yoshija.walter@kalaidos-fh.ch

Einleitung:
In einem vorherigen Beitrag zu dieser Themenreihe wurde beschrieben, dass die grossen KI-Anbieter ihre Modelle immer weiter hochdimensioniert haben. «Immer noch grösser und noch stärker», schien die Devise zu sein. Während GPT-1 noch aus 117 Millionen Parametern bestand, sprang GPT-2 auf 1,5 Milliarden davon. Bei GPT-3 und 3.5 waren es dann bereits 175 Milliarden Parameter. Mittlerweile wurde die Anzahl Parameter zum kommerziellen Firmengeheimnis, sodass man nur noch durch Whistleblower und Leaks Einsicht in die inneren Werke der grössten Modelle erhält. Laut Quellen soll das Mammut-Modell GPT-4 aus 1,7 Billionen Parametern bestehen und bei dem vergleichbaren Riesen von Anthropic AI namens Claude-2 scheint es ähnlich zu sein. Das bedeutet allerding, dass Programme wie ChatGPT so gross sind, dass ich es nie und nimmer ohne Internetverbindung auf meinem Laptop, geschweige denn auf meinem Handy, laufen lassen könnte. Aber seit kurzem entdecken wir einen neuen Trend: die KI-Community stellt die Modelle auf den Kopf.

Downscaling: Wer ist der stärkste Zwerg?
In den letzten Monaten entstanden KI-Modelle, die ähnlich stark sind, wie ChatGPT und auf normalen Computern laufen. Dies betrifft allerdings nur die Inferenz (Nutzung) und nicht das Trainieren der Modelle (dafür braucht man weiterhin leistungsstarke Computer). Man versucht nun aber die Modelle so stark zu komprimieren, dass man sie sogar auf einer Raspberry Pi (also z.B. einem Mini-Computer der Grösse eines Fingernagels) laufen lassen kann. In der Community spricht man liebevoll von der «KI auf einem Toaster». Eines dieser Modelle nennt sich LLaMA und wurde im Februar 2023 von Meta (der Muttergesellschaft von Facebook und Instagram) veröffentlicht. Dabei kam es zu einem peinlichen Skandal: Das Modell sollte nicht öffentlich gemacht, sondern nur ausgewählten Forschern zur Verfügung

gestellt werden. Wahrscheinlich haben es mit diesem Vorhaben nicht alle so genau genommen und innert weniger Wochen wurden die Details von LLaMA «geleaked». Mit anderen Worten: Das Modell kam ungewollt an die Öffentlichkeit. Mark Zuckerberg hat sich dann damit verteidigt, dass es ja ein eher kleines Modell sei und daher der Schaden nicht beängstigend wäre. Es handelt sich hier um ein LFM («Large Foundation Model»), das gemäss den Bedürfnissen der Forschenden in die eine oder andere Richtung für spezifische Aufgaben weiterentwickelt werden kann. Man spricht bei diesem Prozess vom «Fine-Tuning» eines Modells. LLaMA ist kapazitätstechnisch vergleichbar mit GPT, nur eben deutlich offener und kleiner – nicht zu verwechseln mit LaMDA von Google, welches weitaus grösser ist. Die Outputs sind allerdings noch nicht so alltagstauglich, was zu einem eindrücklichen Projekt von fünf Forschern aus der Stanford University führte. Diese haben das LLaMA-Modell in Richtung ChatGPT weiterentwickelt (mittels Fine-Tuning) und nannten ihr Sprachmodell Alpaca. LLaMA versus Alpaca, das Sprachspiel bei den Sprachmodellen wird schnell ersichtlich. Alpaca wird von vielen über eine Plattform namens Dalai benutzt: Auch hier mag man über den Link zum Dalai-LLaMA schmunzeln. Das Stanford-Team hat GPT-3 insgesamt 175 Fragen gestellt, welches dann gemeinsam mit LLaMA-7B (7B steht für «seven billion», also 7 Milliarden Parameter) eine Menge von 52'000 Frage-Antwort-Paare generiert hat. Anhand dieser durch die KI generierten Daten wurde dann Alpaca-7B erschaffen. Mittlerweile geschieht es immer öfter, dass kleinere Modelle anhand von synthetischen Daten (das sind Daten, die durch eine KI erschaffen wurden) trainiert werden. Man spart sich dadurch viel Zeit und Geld. Gleichzeitig kann man dadurch einfacher sicherstellen, dass man Daten von hoher Qualität benutzt. Mittlerweile hat sich diesbezüglich herausgestellt, dass die Qualität der Daten ein entscheidender Faktor für das Trainieren eines guten Modells ist. Manchmal ist weniger eben doch mehr. Das soll heissen: Weniger Daten, aber dafür bessere, liefern bessere Resultate. Alpaca besitzt nur 7 Milliarden Parameter. Zum Vergleich: GPT-3 (genauer gesagt, das Text-Davinci-003 Modell) läuft mit 175 Milliarden Parametern. Den Stanford-Forschern ist es somit gelungen, mit nur 600 US-Dollar und einer Hand voll Parametern eine leistungsfähige Sprach-KI nachzubauen, welche auf einem normalen Laptop oder gar auf einem neueren Handy zum Laufen gebracht werden kann. Gegen Ende Juli 2023 hat Meta nun eine verbesserte Version namens LLaMA 2 herausgegeben – und zwar komplett gratis und open-source. Erstaunlich dabei ist, dass sich Meta in diesem Fall mit ihrem grössten Konkurrenten, Microsoft, zusammengespannt und das Modell gemeinsam publiziert hat. Microsoft ist bekanntlich der grösste Investor von OpenAI, den Machern von ChatGPT. Damit reagieren sie vermutlich auf zwei Sorgen von ihren grössten Kritikern: Nämlich, (i) dass Closed-Source- Modelle zu einer gefährlichen Machtkonzentration führen können und (ii) dass sich die grossen Konzerne gegenseitig davonrennen, wobei sie die nötigen Sicherheitsbedenken vernachlässigen könnten. Dieser Schachzug könnte also als ein beruhigendes Signal an die Öffentlichkeit gewertet werden.

Microsoft gehört indes zu den führenden Playern, wenn es sowohl um die Vergrösserung als auch um die Verkleinerung der KI-Modelle geht. Mit den milliardenschweren Investitionen in OpenAI hat sich der Konzern in das Wettrennen quasi eingekauft und sich den Marktführer zu eigen gemacht, womit sie nun intensive Forschung betreiben. Besonders atemberaubend war ihre Ankündigung von LongNet, welches ein Kontextfenster von einer Milliarde Tokens hat, wobei es bei ChatGPT lediglich ca. 4'000 sind. Unter Kontextfenster versteht man die Anzahl an Tokens, an die sich das System während einer laufenden Konversation «erinnert». Im März wurde Vicuna als Weiterentwicklung von LLaMA publiziert (in der Zoologie gehören Vicunas zu der gleichen Familie wie die Alpacas). Microsoft hat dann im Sommer 2023 ein besseres aber immer noch schlankes Modell namens Orca veröffentlicht. Ganz im Sinne von «high-quality data» haben sie anschliessend ein schlankes Modell namens Phi-1 kreiert, basierend auf dem Forschungspaper «Textbooks are all you need».

Die magische Diät: Wie «verschlankt» man denn eine KI?

Um ein KI-Modell besser und kleiner zu machen, kann man entweder auf bessere Datenqualität in der Trainingsphase setzen (dazu werden heute auch oft synthetische Daten eingesetzt) oder man reduziert die Modellgrösse. Die letztere Strategie wird «Modellkompression» genannt. Hier gibt es vier Methoden:

1. ***Quantisierung:***
   Da es bei künstlich neuronalen Netzwerken im Prinzip um die Summe von komplexen Berechnungen geht, will man hier die Komplexität dieser Berechnungen reduzieren. Man versucht also grosse Modellgewichte mit kleineren zu ersetzen, damit insgesamt weniger Rechenleistung gebraucht wird. Ziel dessen ist die Reduzierung der Modellgrösse, was die KI zwar etwas weniger präzise aber dafür umso effizienter macht.
2. ***Kürzen («pruning»):***
   Nach dem Trainieren einer KI sind gewisse Parameter und Verbindungen zwischen ihnen wesentlich wichtiger als andere. Man kann sich das ein wenig wie in einem Gehirn vorstellen: Wenn jemand tagtäglich nur Taxi fährt, aber nie Fussball spielt, dann sind z.B. die Neuronen für das Autofahren und für das geografische Verständnis wesentlich relevanter als jene, die für die Sensomotorik zuständig sind. In dieser Methode werden bei der KI nun jene «Neuronen» (bzw. Parameter und Verbindungen) rausgeschnitten, die für das Modell nicht so relevant sind. Damit wird das System wesentlich kleiner und schneller.
3. ***Destillierung des Wissens:***
   Man hat herausgefunden, dass ein Training-Modell in der Regel grösser ist als ein Inferenz-Modell. Es sind halt nicht alle Informationen, die man lernt, für die Praxis

letztlich gleichermassen relevant. Aus diesem Grund benutzt man ein grosses Training-Modell (man nennt dies das «Lehrer-Netzwerk»), um ein kleineres Inferenz-Modell für die Praxis (das sog. «Studenten-Netzwerk») zu trainieren. Das oben genannte Beispiel macht sich diese Methode zu Nutze: Alpaca ist das Studenten-Netzwerk, welches anhand der viel grösseren Modelle GPT-3.5 und GPT-4 trainiert wurde. Alpaca ist zwar nicht so gut wie GPT-3.5, kommt dem besagten Lehrer-Netzwerk aber für viele Aufgaben bereits erstaunlich nahe – und ist lediglich ein Bruchteil so gross.

4. ***Low-rank Tensor Dekomposition:***
   Grob gesagt handelt es sich bei dieser Methode um die Wegrationalisierung von Mehrspurigkeiten. Bei Deep-Neural-Networks ist es ein bekanntes Phänomen, dass es zu einer Überparametrisierung kommt. Das bedeutet, dass sich mehrere Parameter um dasselbe Problem kümmern, was zu Redundanzen führt. Man kann sich das in etwa so vorstellen, wie wenn es mehrere Regionen im Gehirn gäbe, die für das Sehen zuständig wären, wir aber eigentlich nur eine einzige solcher Regionen brauchen, weil sie sonst dasselbe (oder etwas sehr Ähnliches) machen würden. Hier geht es nun darum, diese Doppelspurigkeiten zu eliminieren, sodass es pro Aufgabenverarbeitung nur ein dezidiertes Set von Parametern gibt. Damit können «unnötig» grosse Modelle verkleinert werden.

Schlusswort:

Wir leben in einer faszinierenden Zeit, wo sich die Entwicklung der KI exponentiell rasant beschleunigt. Mittlerweile werden die Modelle nicht immer nur grösser. Mit Hilfe der grossen Modelle (ohne diese geht es letztlich nicht) werden kleinere Modelle trainiert, die immer noch sehr leistungsstark sind, aber gleichzeitig auch genug schlank, sodass sie auf normalen Rechnern funktionieren. In einer Welt wie unsere, wo praktisch jede Woche ein neues innovatives KI-Modell die Schlagzeilen erobert, bleibt abzusehen, was in den nächsten Monaten für weitere Neuerungen auf uns zukommen. Diese Neuerungen werden sich wahrscheinlich auf der Achse einer dieser zwei Trends wiederfinden: Entweder werden die Modelle noch grösser und mächtiger oder noch schlanker und effizienter. Aber wer weiss, vielleicht kommt auch schon bald wieder ein ganz neuer Trend auf uns zu?

Wir bleiben gespannt.

Quellen und weiterführende Informationen

Chen, L., Li, S., Yan, J., Wang, H., Gunaratna, K., Yadav, V., Tang, Z., Srinivasan, V., Zhou, T., Huang, H., & Jin, H. (2023). *AlpaGasus: Training A Better Alpaca with Fewer Data* (arXiv:2307.08701). arXiv. https://doi.org/10.48550/arXiv.2307.08701

Gema, A. P., Daines, L., Minervini, P., & Alex, B. (2023). *Parameter-Efficient Fine-Tuning of LLaMA for the Clinical Domain* (arXiv:2307.03042). arXiv. https://doi.org/10.48550/arXiv.2307.03042

Li, Z., Gronke, M., & Steidel, C. (2023, June 19). *ALPACA: A New Semi-Analytic Model for Metal Absorption Lines Emerging from Clumpy Galactic Environments*. arXiv.org. https://arxiv.org/abs/2306.11089v1

Liu, T., & Low, B. K. H. (2023). *Goat: Fine-tuned LLaMA Outperforms GPT-4 on Arithmetic Tasks* (arXiv:2305.14201). arXiv. https://doi.org/10.48550/arXiv.2305.14201

Maeng, K., Colin, A., & Lucia, B. (2019). *Alpaca: Intermittent Execution without Checkpoints* (arXiv:1909.06951). arXiv. https://doi.org/10.48550/arXiv.1909.06951

Peng, B., Li, C., He, P., Galley, M., & Gao, J. (2023). *Instruction Tuning with GPT-4* (arXiv:2304.03277). arXiv. http://arxiv.org/abs/2304.03277

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., & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models* (arXiv:2302.13971). arXiv. https://doi.org/10.48550/arXiv.2302.13971

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Wu, Z., Geiger, A., Potts, C., & Goodman, N. D. (2023). *Interpretability at Scale: Identifying Causal Mechanisms in Alpaca* (arXiv:2305.08809). arXiv. https://doi.org/10.48550/arXiv.2305.08809

Zhang, R., Han, J., Liu, C., Gao, P., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., & Qiao, Y. (2023). *LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention* (arXiv:2303.16199). arXiv. http://arxiv.org/abs/2303.16199