Kalaidos Fachhochschule Schweiz
Die Hochschule für Berufstätige.

Skalierungseffekte der KI: Upscaling (1/2)

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

Einleitung (ohne Titel):
Immer grösser, schneller, besser. So soll es sein in der Technik, nicht? Man stellt sich ein Rennauto vor, das immer schneller fahren soll. Der Motor wird moderner, die Zylinder grösser, die Geschwindigkeit nimmt zu. Allerdings ist grösser nicht immer besser, oder? Die Frage «Wann ist etwas besser?» lässt sich weder in der wirtschaftlichen Geschäftsmodellierung noch in der Technologie eindimensional beantworten. Effektivität («Das Richtige tun») und Effizienz («Die Dinge richtig tun») sind gleichermassen relevant.

«Grösser und besser» versus «kleiner und dynamisch» in der KI
In den letzten Jahren erlebten wir in der Entwicklung der KI vor allem die Tendenz in eine Richtung: grösser heisst besser. Und so wurden die Modelle immer weiter ausgebaut. Sprachmodelle wie GPT-4 von OpenAI, OPT-175 von Meta, Claude-2 von Anthopic AI oder LaMDA, PaLM-2 und Bard von Google sind so gross, dass sie riesige Rechenfarmen benötigen, um sie zu trainieren und äusserst leistungsfähige Server, um sie anschliessend zu benutzen. Unsere normalen Laptops hätten keine Chance, damit umzugehen, und unsere Handys wären erst recht überfordert. Das bedeutet, dass die besten KI-Modelle im Moment kaum für den Hausgebrauch taugen. Das heisst natürlich nicht, dass wir sie nicht benutzen können, wir haben schliesslich alle über Web-Schnittstellen Zugriff auf Applikationen wie ChatGPT (chat.openai.com) oder Bard (bard.google.com). Aber diese grossen Systeme sind rechnerisch sehr schwerfällig und wir können sie daher nicht auf unseren eigenen Rechnern durchführen. Ein sogenanntes «edge AI» wird damit nicht möglich sein (der Begriff kommt von «edge computing» und bedeutet, dass Berechnungen nicht in einer Cloud, sondern lokal auf dem Computer durchgeführt werden, was wiederum die ganzen Übertragungsprozesse beschleunigt).

Seit Kurzem gibt es aber nun ein neuer Trend, der die KI-Community beschäftigt: Die Modelle sollen schlanker und kleiner gemacht werden – wenn möglich ohne grosse Leistungseinbussen. Damit sollen sie günstiger, schneller und portabler werden. In der Folge dessen erleben wir im Moment zwei interessante Strömungen: Auf der einen Seite bauen die grossen Konzerne immer grössere Modelle, die massiv besser werden. Auf der anderen Seite rennen ihnen die kleineren Forschergruppen hinterher, um wesentlich kleinere Modelle zu bauen, die versuchen, den Fähigkeiten der grossen Systeme nachzuahmen.

Upscaling: Lass uns einen Riesen bauen

Für kleinere Forschergruppen, Start-Ups und KMUs ist es weder erstrebenswert noch finanziell möglich, derart grosse Modelle wie GPT-4 und Co. zu bauen. Die «Normalen» unter uns haben gar nicht das nötige Kapital, um diese zu entwickeln. Laut Sam Altman, dem CEO von OpenAI, kostete es über 100 Millionen US-Dollar, um GPT-4 zu trainieren. Und das gilt lediglich für das Training: Dazu kommen Forschungskosten, Mitarbeiterlöhne, usw. Zum einen beruhigt dies die Dystopen, die Angst vor einer KI haben, die uns alle umbringt oder versklavt – es kann sich schliesslich nicht jeder so eine Maschine bauen. Zum anderen gibt es wiederum viele Kritiker:innen, die Respekt vor der Zentralisierung von so viel Macht haben.

Daten und Anzahl Parameter sind entscheidend

Wenn es um die «Grösse» von KI-Modellen geht, dann gibt es zwei Werte, die interessant sind: (i) die Anzahl der Daten für das Training und (ii) die Zahl der Parameter.

Starten wir mit den Daten. Eigentlich klingt es logisch: Je mehr Daten für das Training (oder die «Lernphase») des Modells verwendet werden, desto mehr «weiss» die KI letztlich auch und umso besser ist sie oft in der Bewältigung ihrer Aufgaben. Die grossen Sprachmodelle wurden anhand eines Grossteils des Internets trainiert und wurden zudem mit der Gesamtheit von Wikipedia, Millionen von Büchern, sowie wissenschaftlichen Artikeln gefüttert. Deshalb ist die KI damit nur bis zu einem gewissen Datum aktuell, weshalb ChatGPT bei der Frage «Welches Wetter haben wir heute?» jeweils antwortet: «Sorry, ich habe nur Daten bis 2021». Das ändert sich auch nicht, wenn es direkt mit dem Internet verbunden wird, was z.B. bei dem Code Interpreter von GPT-4, dem BingChat oder bei Bard der Fall ist (das System kann zwar dann auf aktuelle Inhalte zugreifen, aber «lernen» kann es dadurch nicht, es sei denn, es wird einer erneuten Trainingsphase ausgesetzt). Da es im Internet auch viel Schwachsinn gibt, führt dies auch zu Modellen, die regelmässig unerwünschte Inhalte produzieren. Die Antworten können falsch, sexistisch, rassistisch, politisch oder religiös gefärbt sein. Um dem entgegenzuwirken, versucht man der Maschine beizubringen, welche Form von Antworten erwünscht und erlaubt sind. OpenAI war der erste Akteur, der das sog.

RLHF grossflächig dazu einsetzte (das bedeutet «Reinforcement Learning from Human Feedback», wo Menschen jeweils bewerten, wie zufrieden sie mit der Antwort sind).

Genauso wichtig ist die Anzahl der Parameter, die die Grösse des Modells definiert. Es handelt sich dabei um sog. «künstliche Neuronen», die miteinander verknüpft sind. Im Grunde genommen geht es hier um Matrixmultiplikationen, die aufeinander folgen. Je mehr Parameter, desto mehr und komplexere Berechnungen werden durchgeführt. Die Rechenkomplexität nimmt dabei bei mehr Parametern exponentiell zu, was bedeutet, dass eine Verdoppelung der Anzahl Parameter nicht doppelt so viel Rechenleistung braucht, sondern ein Vielfaches davon. Damit wird auch verständlich, warum unsere normalen Laptops mit den grössten Modellen hoffnungslos überfordert wären. In den letzten Jahren wurden die KI-Modelle immer grösser. Zum Vergleich: Als OpenAI im Jahre 2018 GPT-1 veröffentlicht hatte, bestand das Modell aus 117 Millionen Parametern (was damals bereits atemberaubend gross war). Im Jahr 2019 wurde dann GPT-2 publiziert mit einer Grösse von sage-und-schreibe 1,5 Milliarden Parametern. Es wird noch besser: Als GPT-3 im Juni 2020 das Licht der Welt erblickte wurde sie mit 175 Milliarden Parametern versehen. Das ist eine 116-fache Vervielfältigung der Parameter! Und jetzt kommt es: Gemäss einigen Quellen soll GPT-4 mit einer satten Anzahl von 1,7 Billionen Parametern ausgestattet sein. Lesen Sie die Zahl noch einmal: 1,7 Billionen Parameter. Noch vor kurzer Zeit hätte man dies für einen utopischen Scherz gehalten. Das ist über 1’000-mal mehr als bei GPT-2. Genauer gesagt handelt es sich um ein Geflächt aus acht sog. Experten-Netzwerken à je 220 Milliarden Parametern (man nennt diese Systemarchitektur ein «Mixture of Expert»-Netzwerk).

Das Interessante ist, dass bei der Vergrösserung der KI-Modelle überraschend emergente Fähigkeiten auftauchen. GPT war ursprünglich nichts anderes als ein Satz-Vervollständigungs-Modell. Bei der Hoch-Skalierung wurde auf einmal bemerkt, dass die Systeme erstaunlicherweise weitaus mehr können, als lediglich Sätze zu vervollständigen. Der erste Use-Case bei GPT-3 war, dass die KI aus natürlicher Sprache einen Programmiercode erstellen konnte (daraus wurde eine Applikation namens Codex gebaut). Schnell wurde klar, dass die Möglichkeiten fast endlos sind: Wenn man mit einem Computer in natürlicher Sprache (Deutsch, Englisch, usw.) sprechen kann, dann liegen die Begrenzungen nebst den technischen Limitierungen lediglich in der eigenen Vorstellungskraft. Daraus entstand die Idee, dass die konstante Vergrösserung der Modelle zu stets besseren Systemen führen würde. Im «KI-Wettbewerb», der nach der Veröffentlichung von ChatGPT im November 2022 vom Zaun brach, haben die grossen Konzerne versucht, sich gegenseitig durch immer grössere künstliche neuronale Netzwerke zu überbieten.

Welche Probleme KI-Upscaling mit sich bringt

Dabei gibt es aber durchaus auch Probleme. KI-Systeme brauchen Wochen und Monate, bis sie trainiert sind. Sie brauchen riesige Rechenfarmen, die eine Unmenge an Geld und Energie verbrennen. Das ist nicht nur Kapital-intensiv, sondern auch belastend für die Umwelt. Gleichzeitig wird damit aber auch der Weg für Newcomer versperrt, die nicht die nötigen Kapazitäten haben, um sich innovativ dieser Entwicklung anzuschliessen. Aber vielleicht gäbe es doch auch Möglichkeiten, ähnlich gute Modelle zu erschaffen, die wesentlich kleiner und somit zugänglicher sind? Wie gross müssen die Modelle denn mindestens sein, um gewisse Fähigkeiten zu generieren? Wie stark kann man die Modelle denn überhaupt zufriedenstellend verkleinern? Genau solche Fragen werden demnächst im zweiten Teil dieses Beitrags diskutiert.

Quellen und weiterführende Referenzen

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