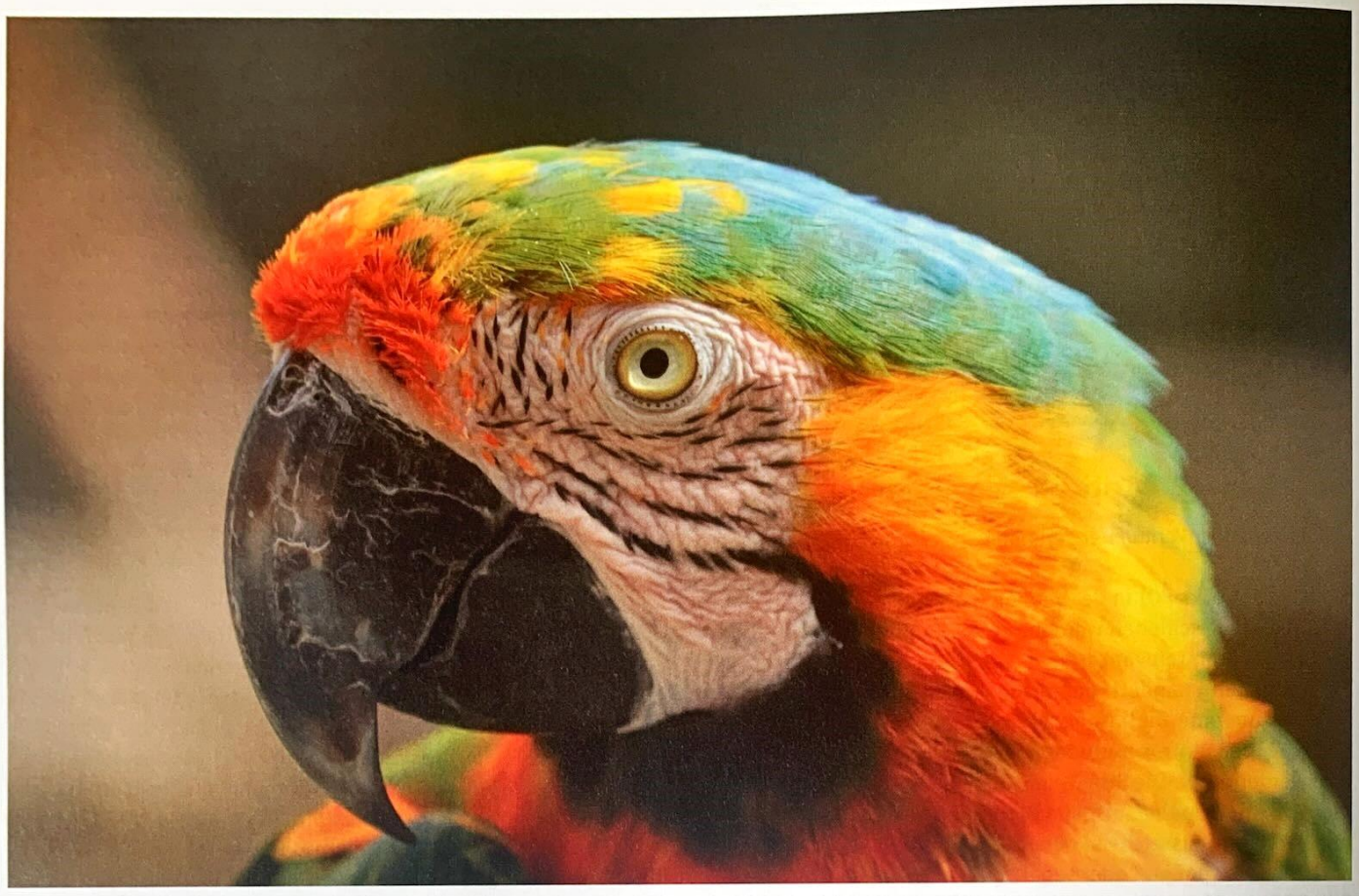


MLOps im Lichte von Generative AI

Große Sprachmodelle

Lothar Wieske

Zum Hype-Thema ChatGPT laden einschlägige Medien zur Meinungsbildung ein - querbeet von Begeisterung bis Empörung. Was als iPhone-Moment für Machine Learning gefeiert wird, wirft ja auch die Frage auf, ob die Mechanismen von MLOps unverändert greifen oder andere und neue Konzepte aufrufen, die in Anpassungen münden (müssen). Der Artikel zeichnet Entwicklungslinien nach und leitet daraus Paradigmenwechsel und Arbeitshypothesen für große Sprachmodelle (Large Language Models) ab.

MLOps (Machine Learning Operations) entwirft und bietet Ordnungs- und Gestaltungsrahmen für die Kommunikation und Kooperation in komplexen Machine Learning-Teams. Beim Deep Learning mit seiner Evolution von CNNs und RNNs hin zu LLMs (Large Language Models) fordern neue und andere Arbeitsweisen und Denkstile zur Öffnung und Einbindung auf.

ILSVRC-Fehlerrate - weniger ist mehr

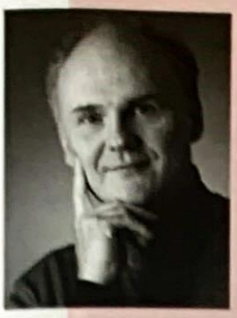
AlexNet, ein Convolutional Neural Network (CNN), wurde an der Universität Toronto entworfen, nahm 2012 am ILSVPR-Wettbewerb zur Bildklassifikation teil und deklassierte die Konkurrenz. Drei wesentliche Punkte sind dazu im Gedächtnis geblieben:

- AlexNet erreichte eine Top-5 Fehlerrate von 15,3 Prozent; sie lag mehr als 10,8 Prozent unter der des Zweitplatzierten im gleichen Wettbewerb im gleichen Jahr.
- AlexNet demonstrierte die wesentliche Bedeutung einer hohen Anzahl von Schichten/Parametern des Modells für seine überragenden Leistungen.
- Das enorm aufwendige Training von AlexNet wurde praktisch nur möglich durch die Nutzung von Grafikprozessor (Graphics Processing Unit - GPU).

In den Jahren von 2012 bis 2017 gab es mit CNNs beim ILSVRC einen regelrechten Wettlauf um immer kleinere Fehlerraten bei immer größeren Anzahlen von Schichten/Parametern. Beim letzten ILSVRC im Jahr 2017 glänzte das beste CNN mit einer Top-5 Fehlerrate von 2,251 Prozent.

Transformers verändern die Welt - im Film wie im Leben

Mitte 2017 erschien der Artikel "Attention Is All You Need" von Mitarbeitern von Google und der Universität Toronto [Tra17].



Lothar Wieske arbeitet im Bereich Enterprise/Solution Architektur. Das Thema Neuronale Netze verfolgt ihn seit den Achtzigerjahren - aber er ist schneller :-) Im Ernst: nach diversen AI-Wintern haben ILSRVC-Wettlauf, Transformers und Foundation Models sein Interesse wieder erweckt. E-Mail: lothar.wieske@web.de

Schon im dritten Satz im Abstract befand sich eine regelrechte Kampfansage: "We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely." Die vorgeschlagene Transformer-Architektur verbesserte für Englisch-Deutsch-Übersetzungen beim populären BLEU (BLEU - Bilingual Evaluation Understudy) den bisherigen Bestwert (SOTA - State Of The Art) um ca. zehn Prozent. BLEU ist ein Algorithmus zur Bewertung der Qualität von maschinell übersetzten Texten, der die Übereinstimmung zwischen der maschinellen und der menschlichen Übersetzung bewertet; je höher die Übereinstimmung, desto besser ist die maschinelle Übersetzung.

Die Transformer-Architektur bestand aus einem Strang von sechs Encoder-Blöcken am Eingang und einem damit quergekoppelten Strang von ebenfalls sechs Decoder-Blöcken hin zum Ausgang. Ein Aufmerksamkeitsmechanismus lieferte für jedes Wort innerhalb eines Satzes die Eingabefolge und Beziehungen zu anderen Wörtern im gleichen Satz. Es zeigte sich, dass Aufmerksamkeitsmechanismen mächtig genug sind, die bisher bei RNNs notwendige Sequentialität der Verarbeitung aufzugeben. Die parallele Verarbeitung aller Token gleichzeitig im Transformer mit seiner rein vorwärts gerichteten Encoder/Decoder-Architektur führte zu bedeutenden Geschwindigkeitssteigerungen gegenüber den vorher für NLP (Natural Language Processing) eingesetzten Recurrent Neural Networks (RNN).

Für die große Transformer-Architektur mit ca. 60 Millionen Parametern (60 M) lag die Trainingszeit mit acht NVIDIA P100 GPUs trotzdem bei etwa 3,5 Tagen.

Nach dem LLM ist vor dem LLM

Das Transformer-Modell wurde 2017 veröffentlicht. Schon im Folgejahr entstanden die Modelle GPT und BERT. Beide Modelle gehörten zu den frühen Vertretern einer neuen Klasse großer Sprachmodelle:

- BERT reduzierte die Transformer-Architektur auf den Encoder-Strang;
- GPT beließ es dagegen beim Decoder-Strang.

Beide Modelle gestalteten das Training zweiphasig: erst Vortraining (Pre-Training), dann Feinabstimmung (Fine-Tuning). Das Vortraining benutzte große Mengen an nicht gekennzeichneten Texten für nichtüberwachtes Lernen und diente dem Aufbau grundsätzlicher sprachlicher Grundfähigkeiten im vortrainierten Modell; für unterschiedliche spezifische Aufgabenstellungen wie maschinelle Übersetzung, Textzusammenfassung wurde das vortrainierte Modell dann mit kleineren Mengen an gekennzeichnetem Text weiter trainiert zum feinabgestimmten Modell.

GPT brillierte beispielsweise bei der maschinellen Übersetzung. BERT beeindruckte wiederum bei der Beantwortung von Fragen. GPT wurde vortrainiert mit BooksCorpus (800 M words); BERT wurde zusätzlich zum BooksCorpus mit Wikipedia (2500 M words) vortrainiert. Tabelle 1 zeigt quantitative Eckdaten der beiden Architekturen.

Das Vortraining verfolgte für die Decoder-Architektur von GPT andere Ziele als für die Encoder-Architektur von BERT. GPT baute ein unidirektionales kausales Sprachmodell auf (CLM - Causal Language Modeling); ein kausales Sprachmodell (CLM) sagt das nächste Token basierend auf den bisher gesehenen Token vorher und GPT kann damit gut fließenden Text generieren. BERT baute dagegen ein bidirektioneles, maskiertes Sprachmodell auf (MLM - Masked Language Modeling); ein solches MLM sagt Token mittendrin vorher, aufgrund von Token, die sowohl links als auch rechts in der Eingabe vorkommen. BERT kann damit gut interne Repräsentationen der gesamten Eingabe lernen beispielsweise für Frage-Antwort-Systeme.

GPT-Olympiade - schneller, höher, weiter

Citius, altius, fortius (lateinisch, deutsch: Schneller, höher, stärker) als Motto der Olympischen Spiele müsste für den aktuellen Wettbewerb bei LLMs eigentlich umgeschrieben werden zu: größer, teurer, belastender. Mit GPT – oder oft synonym GPT-1 – fiel der Startschuss für Erweiterungen und Verbesserungen aus dem Haus OpenAI im Jahrestakt (s. Tabelle 2).

GPT-4 fällt nicht nur hinsichtlich der Jahreszahl aus der Reihe; es gibt gar keine wissenschaftliche Veröffentlichung dazu, sondern nur eine Werbebroschüre zum Herunterladen. Nur so lässt sich eine PDF-Datei ohne Autorenliste beschreiben, in der es gleich im zweiten Abschnitt heißt: "Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."

Model	# Transformer Blocks	# Parameters	# Training Words
GPT	12 Decoder	117M	800M
BERT (LARGE)	24 Encoder	340M	2500M

Tabelle 1: Eckdaten der beiden Transformer-Architekturen

Model	Year	# Decoders	# Context Tokens	Batch Size	# Parameters	Training Set
GPT-1	2018	12	512	64	117 M	4,5 GiB
GPT-2	2019	48	1024	512	1,5 B	40 GiB
GPT-3	2020	96	2048	3,2 M	176 B	570 GiB
GPT-4	2023	N/A	N/A	N/A	N/A	N/A

Tabelle 2: Modellentwicklung

GPT-2 wurde mit einem sehr großen Datenset trainiert; es bestand aus 40 GiB Daten mit 8 Millionen Dokumenten, Reddit Posts mit Top-Bewertungen (Karma>=3). Die dabei zu verprobende Schlüsselhypothese war, dass das Modell mit einer steigenden Parameterzahl und zunehmender Trainingsmenge immer besser würde. Das war nicht selbstverständlich, denn andere Sprachmodelle - wie BERT - wurden ab einer bestimmten Trainingsmenge schlechter.

Bis zu GPT-2 war die Lehrmeinung, dass vortrainierte Modelle für spezifische Fähigkeiten feinabgestimmt werden mussten. Bei GPT-2 wurde auch untersucht, wie das Modell ohne eine solche Feinabstimmung passende Ergebnisse liefern kann. Es wurden also nicht mehr beim Training mit seinen zwei Phasen Vortraining und Feinabstimmung weitere Modellanpassungen vorgenommen; stattdessen wurden richtige Beispiele für die Ausgabe einfach während der Inferenz ohne Modellanpassungen mitgegeben. ("Mach es halt so ähnlich ...") In der Praxis benötigte GPT-2 dann viele Beispiele; aber es konnte sporadisch auch mit nur wenigen, einem oder sogar nur einer Frage funktionieren.

Nach der Veröffentlichung von GPT-2 wechselte OpenAI seinen Status von gemeinnützig zu gewinnorientiert. Mit dieser Veränderung engagierte sich Microsoft mit einer Investition von etwa einer Milliarde Dollar bei OpenAI und erhielt dafür Exklusivrechte zur kommerziellen Nutzung von GPT-3. Bei GPT-3 wurden Modellgröße und Trainingsmenge nochmals deutlich vergrößert. GPT-3 verwendete die gleichen neuronalen Architekturprinzipien mit dem gleichen Aufmerksamkeitsmechanismus, aber baute mehr und breitere Schichten ein. Beim Training kamen zu den Reddit Posts noch Buchsammlungen, die gesamte Wikipedia und mehr.

Das daraus resultierende Modell GPT-3 kam der von GPT-2 nur angedeuteten Fähigkeit, Zero-Shot, One-Shot oder Few-Shot (kein/ein/mehrere Beispiele während der Inferenz) zu betreiben, sehr viel näher. Das Forschungsprogramm von OpenAI also kam in drei Etappen kurz und knapp in den Titeln der zugehörigen Veröffentlichungen zu einem vorläufigen Abschluss:

- GPT-1 - Improving Language Understanding by Generative Pre-training
- GPT-2 - Language Models are Unsupervised Multitask Learners
- GPT-3 - Language Models are Few Shot Learners

Auch andere Technologieunternehmen haben basierend auf den Erfolgen von GPT-2 und GPT-3 immer wieder noch bessere LLMs gebaut. Die eigentliche Transformer-Architektur hat sich im Kern dabei nicht grundlegend verändert, es ging eigentlich nur um Größenwachstum und eine Differenzierung nach Encoder, Decoder und Encoder/Decoder. Außerdem wurden die Einsatzfelder erweitert hin zu multimodalen Modellen; jenseits von geschriebener Sprache (language) ging es bald auch um Bilder (vision), Filme (video) und gesprochene Sprache (speech).

Schneller, höher, weiter. Der Artikel „On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" [Ben21] geht auf verschiedenste Risiken großer Sprachmodelle ein und hat das vielfach aufgegriffene Meme „Stochastic Parrot" geprägt. Die Veröffentlichung und darauf folgende Ereignisse führten dazu, dass die Koautoren Gebru und Mitchell (Pseudonym: Shmitchell) ihre Jobs bei Google verloren und die Google-Mitarbeiter daraufhin protestierten.

Erst kam das Fressen, nun kommt die Moral (frei nach Brecht)

Mit den Fortschritten bei LLMs zeichnete sich ab, dass da etwas anrollt, das wohl geeignete Leitplanken vertragen könnte. Deswegen entschloss sich das Stanford Institute for Human-Centered Artificial Intelligence (HAI) unter seinem Dach das *Center for Research on Foundation Models* (CRFM) ins Leben zu rufen, um eben solche Modelle eingehender untersuchen und entwickeln zu können. Das CRFM definierte ein Grundlagenmodell (FM - Foundation Model) als ein großes Modell, das auf einer großen Datenmenge trainiert wird und für eine große Zahl von sogenannten Downstream-Aufgaben angepasst werden kann. Der Begriff FM erlaubte es bei zunehmend multimodalen Modellen nicht mehr, weiterhin nur von Sprachmodellen (LLMs) reden zu müssen.

Neben den eindrucksvollen Möglichkeiten von FMs sind auch ihre problematischen Seiten mehr als deutlich geworden; die Modelle erzeugen unerwartete Fehler und verzerren die Welt - und, man versteht sie schlicht nicht.

Das CRFM ist eine interdisziplinäre Gruppe von Professoren, Dozenten, Studenten und Wissenschaftlern aus mehr als 10 Fachbereichen der Universität Stanford. Die Mitglieder verbindet das gemeinsame Interesse an der Untersuchung und Entwicklung von FMs. Jenseits der vorherigen eher technischen Fokussierung börsennotierter und gewinnorientierter unternehmerischer Interessen bringt das CRFM von akademischer Seite nun auch ethische Ansprüche und Forderungen in die Entwicklung und Verwendung von FMs mit ein und schafft dafür Grundlagen für deren Verankerung und Umsetzung.

Die Forschungsagenda des CRFM zielt also auf Grundlagen für effizientere, robustere, interpretierbare, multimodalere und ethisch unbedenklichere FMs.

BLOOM und BigScience markieren einen Wendepunkt

BLOOM (BigScience Language Open-science Open-access Multilingual) markierte für Generative AI einen fundamentalen Paradigmenwechsel mit dem Einzug von Community und Open-Source in die Arbeit an und mit LLMs. Neben die Kathedralen der börsennotierten und gewinnorientierten BigTechs traten die Basare der Kooperation und Kommunikation von Gleichgesinnten in Organisationen wie BigScience und EleutherAI.

Das amerikanische Unternehmen *Hugging Face* entwickelt Werkzeuge für maschinelles Lernen. Es ist vor allem bekannt für seine Bibliothek (Transformers Library) zur Verarbeitung natürlicher Sprache und für seine Plattform (Huggingface Hub) zur gemeinsamen Nutzung von Modellen und Daten für maschinelles Lernen. Im Jahr 2021 rief das Unternehmen den *BigScience* Research Workshop ins Leben, um mit anderen Forschungsgruppen am offenen großen Sprachmodell BLOOM zusammen zu arbeiten. Im BigScience Research Workshop sind über 1000 Forscher aus mehr als 70 Ländern und mehr als 250 Einrichtungen an der (Weiter-)Entwicklung von BLOOM beteiligt, darunter Mitarbeiter von HuggingFace BigScience, Microsoft DeepSpeed, NVIDIA Megatron-LM, PyTorch und viele andere Freiwillige.

Dank einem Budget von etwa 3 Millionen EUR konnte BLOOM dann Mitte 2022 etwa 117 Tage lang auf dem Supercomputer Jean Zay im Süden von Paris trainiert werden. Als Hardware kam ein Cluster mit 48 Knoten zum Einsatz. Jeder dieser Knoten hatte als CPU einen AMD-Prozessor mit 32 Cores und als RAM einen Hauptspeicher mit 512 GB. Jeder Knoten hatte aber ebenso acht GPUs, und zwar NVIDIA-Beschleuniger A100 (Ampere Architektur) mit je 80 GB HBM2e, dem gegenüber klassischen RAM deutlich schnelleren Speicher auf Grafikkarten. Das gesamte Cluster hatte damit 48 CPUs mit in Summe 1536 Core mit in Summe 16384 GB CPU

Model	# Parameters	Power Consumption	CO2eq Emissions
GPT-3	175 B	1,287 MWh	502 tCO2eq
BLOOM	176 B	433 MWh	25 tCO2eq

Tabelle 3: Ökologische Kosten

RAM und 384 GPUs mit in Summe 30720 GB HBM2e verknüpft durch schnelle NVLink 4-Verbindungen. Mit seinen 176 M Parametern kann BLOOM nun Texte und Programme in 46 natürlichen Sprachen und 13 Programmiersprachen generieren. Das fertige und am 6. Juli 2022 in der Version 1.3 freigegebene Modell wurde und wird im Hugging Face-Ökosystem bereitgestellt unter den Bedingungen von RAIL (Responsible AI License).

Ökonomische und ökologische Kosten

Die Arbeit mit LLMs kostet, ökonomisch und ökologisch. Die ökonomischen Kosten für das Training von SOTA LLMs sind gewaltig und offizielle Zahlen gibt es nicht; sie haben Milliarden Parameter, werden mit Billionen von Tokens trainiert, und das kostet wohl Millionen EUR. Die ökologischen Kosten haben Alexandra Sasha Luccioni, Sylvain Viguier und Anne-Laure Ligozat für BLOOM im Vergleich mit GPT-3 einmal geschätzt und gegenübergestellt [Luc22]. Tabelle 3 zeigt einige Details, in der Veröffentlichung finden sich weitere Daten; beispielsweise müssen ja noch die CO_2-Emissionsfaktoren der Stromnetze und Energie-Effizienz der Rechenzentren (PUE - Power Usage Effectiveness) eingerechnet werden. Da kommt noch mal 'ne Schippe drauf.

DIY - Schöne Ideen zum Selbermachen

Besondere Anerkennung und Erwähnung sollen im Rahmen dieses Überblicks und Einstiegs folgende vortrainierte Modelle mit Links zum Herunterladen und Selbermachen finden:

- bert-base-uncased: https://huggingface.co/bert-base-uncased
- gpt2: https://huggingface.co/gpt2
- t5-base: https://huggingface.co/t5-base
- distilbert-base-uncased: https://huggingface.co/distilbert-base-uncased
- microsoft/codebert-base: https://huggingface.co/microsoft/codebert-base
- bigscience/bloom: https://huggingface.co/bigscience/bloom
- google/vit-base-patch16-224: https://huggingface.co/google/vit-base-patch16-224

Mit der Huggingface-Infrastruktur und insbesondere den dort niedergelegten Anleitungen sollte sich der eine oder andere Appetithappen fürs heimische Generative AI Lab finden.

Zusammenfassung

Da hat sich doch ganz schön was getan mit Transformern. CNNs wurden typischerweise für Aufgaben mit Bildern im Supervised Learning Modus trainiert; gelegentlich kam dann noch Transfer Learning dazu, wenn man von den intern gelernten Repräsentationen profitieren wollte, aber dem Modell letztlich doch eigene Eingabe-/Ausgabe-Paare aufprägen wollte. Ganz ähnlich war die Sachlage bei RNNs; dort gestaltete sich in der Regel der Supervised Learning Modus als richtig aufwendig, einerseits weil es die entsprechende händische Aufbereitung der Trainingsdaten in sich hatte und andererseits auch weil die Durchlaufzeiten der Netze bei Training wie Inferenz durch die sequenzielle Verarbeitung Token für Token sehr beschränkt waren.

Bei Transformern muss nun in drei Phasen gedacht werden:

- Vortraining (Pre-Training),
- Feinabstimmung (Fine-Tuning) und
- Ausführung (Inference).

Und entlang dieser Phasen sind auch andere Arbeitsteilungen und Lieferketten absehbar. Die eigene Erstellung von FMs birgt enorme Einstiegshürden, was Menschen, Daten und Modelle angeht; also werden FMs wohl häufiger extern bezogen und intern veredelt. Zu kommerziellen Bezugswegen kommen vermehrt auch öffentliche Bereitstellungswege, mit denen auch höhere ethische Ansprüche spezifischer abgebildet werden können. Wenn dann ein oder mehrere vortrainierte Modelle im Haus sind, geht es um weitere Entscheidungen und Kompetenzen zum Einsatz von Feinabstimmung und/oder Aufgabenbeschreibung (Prompt Engineering).

Jedenfalls rücken vor dem Hintergrund von Governence/Risk/Compliance (GRC) auch ethische Ansprüche stärker in den Vordergrund. Dabei geht es um Transparenz schon während des Trainings, was Daten und Vorgehen angeht, und ein besseres Verständnis, wie eigentlich Verzerrungen und Verfälschungen in den Modellen entstehen. Aber auch Transparenz während der Ausführung und verbesserte Nachvollziehbarkeit und Erklärbarkeit sind gefordert und erfordern entsprechende Werkzeuge für alle drei Phasen Vortraining, Feinabstimmung und Ausführung.

Es zeichnet sich recht eindeutig ab, dass sich aktuelle MLOps-Entwürfe da nicht einfach und schnell verlängern lassen.

Literatur und Links

[AwLLM] Awesome-LLM, https://github.com/Hannibal046/Awesome-LLM

[Ben21] E. M. Bender et al., On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?, https://dl.acm.org/doi/10.1145/3442188.3445922

[BERT] J. Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, https://arxiv.org/abs/1810.04805

[BLOOM] BigScience Workshop, BLOOM: A 176B-Parameter Open-Access Multilingual Language Model, https://arxiv.org/pdf/2211.05100.pdf

[CRFM] Center for Research on Foundation Models, https://crfm.stanford.edu/

[GPT-1] A. Radford et al., Improving Language Understanding by Generative Pre-Training, https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[GPT-2] A. Radford et al., Language Models are Unsupervised Multitask Learners, https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[GPT-3] T. B. Brown et al., Language Models are Few-Shot Learners, https://arxiv.org/pdf/2005.14165.pdf

[GPT-4] GPT-4 Technical Report, https://cdn.openai.com/papers/gpt-4.pdf

[Luc22] A. Luccioni et al., Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model, https://arxiv.org/pdf/2211.02001.pdf

[Tra17] A. Vaswani et al., Attention Is All You Need, https://arxiv.org/abs/1706.03762