idw - Informationsdienst Wissenschaft
Nachrichten, Termine, Experten

Pressemitteilung

Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS
Katrin Berkler

26.11.2024
http://idw-online.de/de/news843614

Forschungs- / Wissenstransfer, Forschungsergebnisse
Informationstechnik
überregional

Multilingual und Open Source: Forschungsprojekt OpenGPT-X veröffentlicht großes KI-Sprachmodell

Das große KI-Sprachmodell des Forschungsprojekts OpenGPT-X steht auf Hugging Face zum Download bereit: »Teuken-7B« wurde mit den 24 Amtssprachen der EU trainiert und umfasst sieben Milliarden Parameter. Akteure aus Forschung und Unternehmen können das kommerziell einsetzbare Open-Source-Modell für ihre KI-Anwendungen nutzen. Damit haben die Partner des vom Bundesministerium für Wirtschaft und Klimaschutz geförderten Konsortialprojekts OpenGPT-X unter der Leitung der Fraunhofer-Institute für Intelligente Analyse- und Informationssysteme IAIS und für Integrierte Schaltungen IIS ein KI-Sprachmodell als frei verwendbares Open-Source-Modell mit europäischer Perspektive auf den Weg gebracht.

»Im Projekt OpenGPT-X haben wir in den vergangenen zwei Jahren mit starken Partnern aus Forschung und Wirtschaft die grundlegende Technologie für große KI-Fundamentalmodelle erforscht und entsprechende Modelle trainiert. Wir freuen uns, dass wir jetzt unser Modell ‚Teuken-7B' weltweit frei zur Verfügung stellen und damit eine aus der öffentlichen Forschung stammende Alternative für Wissenschaft und Unternehmen bieten können«, sagt Prof. Dr. Stefan Wrobel, Institutsleiter am Fraunhofer IAIS. »Unser Modell hat seine Leistungsfähigkeit über eine große Bandbreite an Sprachen gezeigt, und wir hoffen, dass möglichst viele das Modell für eigene Arbeiten und Anwendungen adaptieren oder weiterentwickeln werden. So wollen wir sowohl innerhalb der wissenschaftlichen Community als auch gemeinsam mit Unternehmen unterschiedlicher Branchen einen Beitrag leisten, um den steigenden Bedarf nach transparenten und individuell anpassbaren Lösungen der generativen Künstlichen Intelligenz zu adressieren.«

Teuken-7B ist aktuell eines der wenigen KI-Sprachmodelle, die von Grund auf multilingual entwickelt wurden. Es enthält ca. 50 Prozent nicht-englische Pretraining-Daten und wurde in allen 24 europäischen Amtssprachen trainiert. Es erweist sich über mehrere Sprachen hinweg in seiner Leistung als stabil und zuverlässig. Dies bietet insbesondere internationalen Unternehmen mit mehrsprachigen Kommunikationsbedarfen sowie Produkt- und Serviceangeboten einen Mehrwert. Die Bereitstellung als Open-Source-Modell erlaubt es Unternehmen und Organisationen, eigene angepasste Modelle in realen Anwendungen zu betreiben. Sensible Daten können im Unternehmen verbleiben.

Das OpenGPT-X-Team widmete sich neben dem Modelltraining auch zahlreichen Forschungsfragen, zum Beispiel wie multilinguale KI-Sprachmodelle energie- und kosteneffizienter trainiert und betrieben werden können. Dazu wurde im Projekt ein multilingualer »Tokenizer« entwickelt. Die Aufgabe eines Tokenizers ist es, Wörter in einzelne Wortbestandteile zu zerlegen – je weniger Token, desto (energie-)effizienter und schneller generiert ein Sprachmodell die Antwort. Der entwickelte Tokenizer führte zu einer Reduzierung der Trainingskosten im Vergleich zu anderen multilingualen Tokenizern, wie etwa Llama3 oder Mistral. Dies kommt insbesondere bei europäischen Sprachen mit langen Wörtern wie Deutsch, Finnisch oder Ungarisch zum Tragen. Auch im Betrieb von mehrsprachigen KI-Anwendungen können damit Effizienzsteigerungen erreicht werden.

Das Verbundprojekt OpenGPT-X wurde im Rahmen des BMWK-Förderprogramms »Innovative und praxisnahe Anwendungen und Datenräume im digitalen Ökosystem Gaia-X« gefördert. Somit ist Teuken-7B auch über die Gaia-X Infrastruktur zugänglich. Akteure im Gaia-X-Ökosystem können so innovative Sprachanwendungen entwickeln und in konkrete Anwendungsszenarien in ihren jeweiligen Domänen überführen. Im Gegensatz zu bestehenden Cloud-Lösungen handelt es sich bei Gaia-X um ein föderiertes System, über das sich unterschiedliche Dienstanbieter und Dateneigentümer miteinander verbinden können. Die Daten verbleiben stets beim Eigentümer und werden ausschließlich nach festgelegten Bedingungen geteilt.

»Ich freue mich über die heutige Veröffentlichung des Gaia-X-basierten KI-Sprachmodells Teuken-7B und gratuliere dem Projekt OpenGPT-X, dass es diesen wichtigen Meilenstein erreicht hat. Besonders ist, dass Teuken-7B auch die sichere Nutzung sensibler Unternehmensdaten ermöglicht, da die Gaia-X-Standards die Datenspeicherung und -verarbeitung nach höchsten europäischen Datenschutz- und Sicherheitsbestimmungen garantieren. Innovationen wie diese stärken die digitale Souveränität, die Wettbewerbsfähigkeit und auch die Resilienz Deutschlands und Europas. Deshalb fördert das BMWK das Projekt mit rund 14 Millionen Euro«, sagt Dr. Franziska Brantner, Parlamentarische Staatssekretärin im BMWK.

Prof. Dr.-Ing. Bernhard Grill, Institutsleiter am Fraunhofer IIS, betont die Bedeutung für sicherheitsrelevante Anwendungen: »Mit dem hier veröffentlichten, von Grund auf vollkommen eigenständig trainierten Sprachmodell demonstrieren die Projektpartner ihre Fähigkeit, eigene große Modelle erzeugen zu können. Der damit verbundene Zugriff auf ein großes KI-Sprachmodell ermöglicht Anwendungen, die ohne nicht einsehbare Fremd-Komponenten eine sehr viel bessere Kontrolle über diese Technologie bieten – z. B. für spezifische, besonders auch sicherheitskritische Anwendungen im Automobilbereich, in der Robotik, der Medizin oder dem Finanzwesen. Durch Training mit den für den konkreten Anwendungsfall relevanten Daten und die Verwendung anwendungsspezifischer Architekturen können für Unternehmen so individuelle KI-Lösungen geschaffen werden, die ohne Black-Box-Komponenten auskommen.«

Generative KI aus einem starken Verbund – mit europäischer Perspektive

In die Modellentwicklung sind wichtige Forschungsergebnisse aus dem OpenGPT-X-Projekt eingeflossen, wie beispielsweise Tools und Technologien, um sehr große Datenmengen aufzubereiten, leistungsfähige europäische HPC-Infrastrukturen zu nutzen und ein effizientes Modelltraining durchzuführen. Trainiert wurde Teuken-7B mithilfe des Supercomputers JUWELS am Forschungszentrum Jülich. Neben den beiden Fraunhofer-Instituten und dem Forschungszentrum Jülich haben der KI Bundesverband, die TU Dresden, das Deutsche Forschungszentrum für Künstliche Intelligenz (DFKI), IONOS, Aleph Alpha, ControlExpert sowie der Westdeutsche Rundfunk (WDR) als Partner an OpenGPT-X mitgearbeitet. Die in OpenGPT-X entstandene Technologie bietet den Partnern auch zukünftig die Basis für das Training weiterer eigener Modelle.

»OpenGPT-X dient als Beispiel dafür, wie mit den Mitteln eines öffentlichen Förderprojekts und der gemeinsamen Anstrengung eines breit aufgestellten Konsortiums – von der zugrundeliegenden Infrastruktur über das Training von Modellen bis hin zur produktiven Anwendung – wertvolle Basistechnologie entstehen kann. Im Interesse der Technologie- und Datensouveränität gilt es nun, auf dieser Grundlage aufzubauen: Wir wünschen uns, dass OpenGPT-X als Basis für viele nachfolgende Aktivitäten genutzt werden wird«, betont Daniel Abbou, Geschäftsführer im KI Bundesverband und Präsident des European AI Forum.

Das Anfang 2022 gestartete Forschungsprojekt steht nun kurz vor dem Abschluss. Es läuft noch bis zum 31. März 2025, so dass weitere Optimierungen und Evaluierungen der Modelle erfolgen können.

Der Weg zur Nutzung von Teuken-7B

Interessierte Entwicklerinnen und Entwickler aus der Wissenschaftscommunity oder Unternehmen können Teuken-7B bei Hugging Face kostenfrei herunterladen und in der eigenen Entwicklungsumgebung damit arbeiten. Das Modell

wurde durch ein »Instruction Tuning« bereits fu□r den Chat optimiert. Mit Instruction Tuning werden große KI-Sprachmodelle dahingehend angepasst, dass das Modell Anweisungen von Nutzerinnen und Nutzern richtig versteht, was vor allem fu□r die Anwendung der Modelle in der Praxis relevant ist – zum Beispiel fu□r den Einsatz in einer Chatanwendung.

Teuken-7B steht in zwei Varianten zur Verfu□gung: einer Version, die fu□r Forschungszwecke genutzt werden kann, und einer Version unter der Lizenz »Apache 2.0«, die Unternehmen neben Forschung auch fu□r kommerzielle Zwecke nutzen und in eigene KI-Anwendungen integrieren können. Die Leistungsfähigkeit beider Modelle ist in etwa vergleichbar, einige der fu□r das Instruction Tuning verwendeten Datensätze schließen jedoch eine kommerzielle Nutzung aus und wurden aus diesem Grund in der Apache 2.0-Version nicht verwendet.

Über OpenGPT-X

Das OpenGPT-X-Projekt startete am 1. Januar 2022 mit einer Förderung des Bundesministeriums fu□r Wirtschaft und Klimaschutz (BMWK) in Höhe von rund 14 Millionen Euro und endet am 31. März 2025. Die zehn Projektpartner sind Fraunhofer IAIS, Fraunhofer IIS, Forschungszentrum Ju□lich, KI Bundesverband, TU Dresden, DFKI, IONOS, Aleph Alpha, ControlExpert und WDR. Unter der Leitung von Fraunhofer IAIS und Fraunhofer IIS erforscht das Projekt die gesamte Wertschöpfungskette der Generativen KI: Von der hochskalierbaren, GPU-basierten Infrastruktur und den Daten fu□r das Training großer Sprachmodelle, u□ber die Entwicklung der Modelle, bis hin zur produktiven Anwendung in Form von Prototypen und Proof of Concepts (PoCs). Übergreifendes Ziel des Projektes war es, ein eigenes großes KI-Sprachmodell zu entwickeln, das fu□r Forschung und Unternehmen Open Source zur Verfu□gung gestellt und insbesondere auf die multilingualen Bedu□rfnisse Europas ausgerichtet wird. Mit der Veröffentlichung von Teuken-7B hat das Projekt dieses Ziel erreicht und stellt damit eine aus der öffentlichen Forschung stammende Alternative fu□r zuku□nftige wissenschaftliche Untersuchungen und wirtschaftliche Anwendungen der Generativen KI zur Verfu□gung.

Über Fraunhofer IAIS

Als Teil der größten Organisation fu□r anwendungsorientierte Forschung in Europa ist das Fraunhofer-Institut fu□r Intelligente Analyse- und Informationssysteme IAIS mit Sitz in Sankt Augustin/Bonn und einem Standort in Dresden eines der fu□hrenden Wissenschaftsinstitute auf den Gebieten Ku□nstliche Intelligenz (KI), Maschinelles Lernen und Big Data in Deutschland und Europa. Rund 380 Mitarbeitende unterstu□tzen Unternehmen bei der Optimierung von Produkten, Dienstleistungen und Prozessen sowie bei der Entwicklung neuer digitaler Geschäftsmodelle. Das Fraunhofer IAIS gestaltet die digitale Transformation unserer Arbeits- und Lebenswelt: mit innovativen KI-Anwendungen fu□r Industrie, Gesundheit und Nachhaltigkeit, mit zukunftsweisenden Technologien wie großen KI-Sprachmodellen oder Quantum Machine Learning, mit Angeboten fu□r die Aus- und Weiterbildung oder fu□r die Pru□fung von KI-Anwendungen auf Sicherheit und Vertrauenswu□rdigkeit.

Über Fraunhofer IIS

Der Bereich Audio und Medientechnologien des Fraunhofer IIS prägt seit u□ber 30 Jahren die weltweit eingesetzten Standards und Technologien in der Audio- und Filmindustrie. Angefangen bei der Erfindung von mp3 und fortgesetzt in der Entwicklung von AAC und dem Testplan der Digital Cinema Initiative, finden sich heute Systeme und Technologien aus Erlangen in fast allen Geräten der Unterhaltungselektronik und der (mobilen) Kommunikation. Unsere neueste Generation an Medientechnologien wie MPEG-H Audio, xHE-AAC, LC3/LC3plus, Symphoria und upHear sind ebenfalls bereits weltweit verbreitet. Seit u□ber 20 Jahren beschäftigen wir uns zudem mit Sprachtechnologien. Zuletzt entstand der EVS-Standard von dem alle 5G-Sprachdienste profitieren. Heute bauen wir unsere Aktivitäten in Richtung Sprachsignalverarbeitung und Sprachassistenzsysteme aus.

wissenschaftliche Ansprechpartner:
Pressekontakt pr@iais.fraunhofer.de

URL zur Pressemitteilung: https://huggingface.co/openGPT-X Modell-Download und Model Card

URL zur Pressemitteilung: https://opengpt-x.de/en/models/teuken-7b Technische Infos und Benchmarks zum Modell

URL zur Pressemitteilung: https://opengpt-x.de/news-de Fachpublikationen aus OpenGPT-X

URL zur Pressemitteilung: https://huggingface.co/spaces/openGPT-X/european-llm-leaderboard European LLM Leaderboard

URL zur Pressemitteilung: https://discord.gg/RvdHpGMvB3 Fachcommunity – Feedback und technische Fragen

URL zur Pressemitteilung: https://www.iais.fraunhofer.de/opengpt-x Demo-Termin vereinbaren