Richtlinie zum „Software-Sprint“

RevLLM – Reverse Engineering Werkzeuge für Sprachmodelle

Schlussbericht

Zuwendungsempfänger:

Julian Wergieluk

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS23S42 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Kurze Darstellung der Aufgabenstellung und Motivation

Was war Deine Motivation? Welches Problem wolltest Du mit Deinem Projekt lösen? Wie war die geplante Vorgehensweise zur Problemlösung (auch Angabe der wichtigsten Meilensteine)?

Die Künstliche Intelligenz (KI) und die sogenannten “Großen Sprachmodelle” (Large Language Models / LLMs), wie GPT-4, haben bereits ein beispielloses Potenzial in einer Vielzahl von Anwendungen gezeigt. Von der Unterstützung der Forschung, Entwicklung neuer Produkte und Automatisierung der Prozesse in Unternehmen bis hin zur Umgestaltung des Bildungswesens - die Auswirkungen dieser Technologien auf unsere Gesellschaft werden immens sein.

Die Entwicklung der LLMs wurde in einem rasanten Tempo von einer Handvoll großer Technologieunternehmen vorangetrieben. Dabei sind Entwicklungsprozesse zunehmend intransparent geworden und zielen vor allem darauf ab, die Modelle möglichst leistungsfähig und vielseitig anwendbar zu machen. Die produzierten Modelle sind weitestgehend als Black-Box Systeme anzusehen. Um die technologische Unabhängigkeit unserer Gesellschaft zu gewährleisten, die Innovation zu fördern und die demokratischen Grundsätze zu wahren, ist es notwendig, besser zu verstehen, wie diese Modelle “denken”. Wir müssen die Funktionsweise der LLMs transparent, erklärbar und interpretierbar machen, um sicherzustellen, dass sie nach klar definierten und von der Gesellschaft akzeptierten Prinzipien agieren. Dieses Wissen und die Werkzeuge, die die Erforschung der LLMs ermöglichen, sollen frei zugänglich und verifizierbar gemacht werden.

In diesem Projekt ging es darum, die inneren Entscheidungsprozesse der Sprachmodelle, die auf der Transformer-Architektur basieren (wie z.B. GPT-4), besser zu verstehen und sie

transparenter und interpretierbarer zu machen. Es wurde eine Python-Bibliothek und eine darauf basierende Web-App entwickelt, die eine Reihe von Methoden für die Analyse der “Denkweise” der Sprachmodelle zur Verfügung stellt. Dabei wird anhand von speziell dafür konstruierten Beispiel-Eingaben kombiniert mit den Gewichten des neuronalen Netzes (=der “Modell-Code”) untersucht, welchen Algorithmus das Sprachmodell "implementiert", um die Ausgabe zu produzieren. Die Methoden orientieren sich an aktuellen Forschungsartikeln, wurden über ein einheitliches API und ein Web-UI nutzbar und leicht zugänglich gemacht.

Die Arbeit an dem Projekt erfolgte nach den Prinzipien der agilen Softwareentwicklung. Die Umsetzung der folgenden Meilensteine wurde flexibel und iterativ angegangen und die Fertigstellung mehrerer Meilensteine wurde gleichzeitig vorantreiben, da es nur wenige harte Abhängigkeiten gab.

Meilensteine:

- Literaturrecherche und Experimente mit den wichtigsten Reverse-Engineering Methoden. Erarbeitung der Abgrenzung und Synergien zur existierenden Software und ähnlichen Initiativen.
- Einteilung der Methoden in must-have, nice-to-have, out-of-scope Gruppen.
- Design und Kick-off der Entwicklung der Bibliothek API.
- Implementierung der Methoden gemäß der Priorisierung.
- Tests und weitere Experimente mit mehreren vortrainierten Sprachmodellen.
- Entwicklung der Web-App und Erstellung der Dokumentation.

Beitrag des Projektes zu den Zielen der Förderinitative „Software-Sprint“

Welche Bezüge gibt es zu den Themenfeldern „Civic Tech“ und „Data Literacy“ des Software Sprints oder zu weiteren gesellschaftlich relevanten Zielen bzw. Lösungsansätzen?

Das RevLLM-Projekt positioniert sich an der Schnittstelle von Technologie, Gesellschaft und Bildung und trägt direkt zu den Themenfeldern Civic Tech und Data Literacy bei, während es gleichzeitig weitere gesellschaftlich relevante Ziele unterstützt. Durch die Schaffung von Werkzeugen zur Analyse und Interpretation der Arbeitsweise von großen Sprachmodellen (LLMs), fördert das Projekt ein tieferes Verständnis dieser Technologien und trägt zur Verbreitung des Wissens um Künstliche Intelligenz (KI) bei.

Civic Tech

Durch das RevLLM-Projekt werden Bürgerinnen und Bürger, die über die notwendigen Vorkenntnisse verfügen, befähigt, die Funktionsweise von großen Sprachmodellen (LLMs) besser zu verstehen. Dieses tiefere Verständnis ist entscheidend, da es ihnen ermöglicht, effektiver an Diskussionen und Entscheidungsfindungsprozessen zur Regulierung dieser Technologien teilzunehmen. Wenn die Öffentlichkeit informierter über die Mechanismen und potenziellen Auswirkungen von LLMs ist, kann sie eine aktivere Rolle in der Gestaltung von Richtlinien und Regulierungen einnehmen, die sowohl den technologischen Fortschritt als auch das Gemeinwohl im Auge behalten. Das Projekt leistet somit einen wesentlichen Beitrag dazu, die Kluft zwischen technologischem Know-how und gesellschaftlicher Teilhabe

zu überbrücken, und unterstützt eine demokratische Auseinandersetzung mit den Herausforderungen und Chancen, die diese KI-Technologie bietet.

Data Literacy

Die Art und Weise, wie die großen Sprachmodelle produziert werden, erinnert an Datenkompression. Durch das Bereitstellen von Werkzeugen und Methoden, mit denen auch Nicht-Experten die „Denkweise" von LLMs untersuchen können, fördert das RevLLM-Projekt das Verständnis für datengetriebene Technologien. Nutzer lernen, wie Daten und Algorithmen zu Entscheidungen führen, was essentiell ist, um in einer zunehmend von Daten geprägten Welt kompetent agieren zu können. Die Fähigkeit, die Arbeitsweise und Entscheidungsfindung von KI-Modellen kritisch zu hinterfragen, ist ein Aspekt von Data Literacy.

Bildung

Ein spezieller Fokus des RevLLM-Projektes ist die einfach verständliche und interaktive Vermittlung des Wissens über die Funktionsweise der LLMs. Das Projekt hat das Potenzial, Lehrmaterialien und Ressourcen für die Bildung im Bereich KI und Maschinelles Lernen zu liefern.

Ausführliche Darstellung der Ergebnisse

Welche konkreten Ergebnisse hast Du erzielt?

Im Rahmen des RevLLM-Projekts habe ich eine Machine Learning Bibliothek und darauf aufbauende WebApp entwickelt. RevLLM baut auf *nanoGPT* - einer Implementierung der GPT-2-Modellfamilie auf, die von Andrej Karpathy entwickelt wurde, und hält sich an deren Geist der Einfachheit und Transparenz. Die RevLLM Bibliothek hat minimale externe Abhängigkeiten und besteht aus einem sauberen und einfachen Code, der leicht zu verstehen und wiederzuverwenden ist.

Die RevLLM-Bibliothek implementiert verschiedene Methoden zur Analyse des internen Datenflusses der Transformer-Modelle, die nur aus einem Decoder-Block bestehen (decoder-only transformer models). Insbesondere zeigt RevLLM,

- wie Transformer-Modelle einen Prompt-Satz auf eine Folge von Tokens abbilden.
- wie der Raum der Token-Einbettungen mit statistischen Methoden untersucht werden kann.
- implementiert verschiedene Methoden zur Bestimmung der Wichtigkeit von Prompts, um die Token zu identifizieren, denen das Modell die größte Aufmerksamkeit schenkt.
- implementiert Logit-Lens-Interpretationsmethoden, um den Informationsfluss zu erläutern, der in der Transformator-Blocksequenz erfolgt.

Zur Erleichterung der Nutzung und um ein problemloses Experimentieren zu ermöglichen, wurde die Bibliothek mit einer interaktiven Streamlit-App (Web-App) ergänzt und bietet somit eine Webschnittstelle für den Zugriff auf die Bibliotheksfunktionen. Die WebApp lädt automatisch ein ausgewähltes Modell aus der GPT-2-Familie herunter, instanziiert es und

stellt die Methoden der RevLLM-Bibliothek über eine praktische benutzerfreundliche Oberfläche bereit.

Konnten alle Meilensteine erreicht werden?

Alle im Antrag angeführten Meilensteine wurden erreicht. Darüber hinaus wurde zusätzlich eine Web-App entwickelt, die es den Benutzer*innen ermöglicht, auf die Funktionalität der Bibliothek einfach zuzugreifen.

*Welche zusätzlichen **Erkenntnisse** hast Du aus der Projektarbeit gewonnen, auch im Hinblick auf die Begleitung durch die Open Knowledge Foundation?*

Während der Arbeit an dem Projekt habe ich sehr viel über die Funktionsweise der Transformer-Sprachmodelle, sowie über die Methoden zu deren Analyse (insbesondere über die Erklärbarkeitsmethoden) gelernt.

Ich habe auch besser verstanden, wie das Open Source Ökosystem funktioniert und wie deren Funktionsweise von verschiedenen Organisationen (wie OKF) und Menschen vorangetrieben wird.

Zielgruppe, Nutzen und mögliche Weiterentwicklungen

Welcher Nutzen ergibt sich für die Zielgruppe aus den Ergebnissen Deines Projekts?

Das RevLLM Projekt zielt darauf ab, ein tieferes Verständnis und Transparenz über die Funktionsweise und Entscheidungsprozesse großer Sprachmodelle (LLMs) zu schaffen. Dieser Ansatz bietet der Zielgruppe – bestehend aus Data Scientists, KI-Forschern sowie technisch versierten Ingenieuren und Enthusiasten – vielfältigen Nutzen:

Verbessertes Verständnis der LLM-Innenarchitektur
Die Bereitstellung von Werkzeugen zur Analyse des internen Datenflusses und der Entscheidungsprozesse in LLMs ermöglicht es Fachleuten, ein tieferes Verständnis der inneren Arbeitsweise dieser komplexen Modelle zu entwickeln. Dies ist besonders wichtig, da die Modelle zunehmend in kritischen und einflussreichen Bereichen eingesetzt werden.

Erleichterte Modellinterpretation
Durch die Implementierung von Methoden zur Bestimmung der Wichtigkeit von Prompts und die Anwendung von Logit-Lens-Interpretationstechniken wird die Interpretierbarkeit der Modelle verbessert. Die Fähigkeit, die "Denkweise" eines Modells zu interpretieren, ist essentiell für die Entwicklung sicherer und ethisch verantwortungsvoller KI-Anwendungen.

Beschleunigung der Forschung und Entwicklung
RevLLM erleichtert die Forschung und Entwicklung neuer Modelle und Anwendungen, indem es ein Framework für das Experimentieren und Testen mit bestehenden LLMs bietet. Durch die einfache Wiederverwendung und Anpassung des Codes können Innovationen schneller vorangetrieben und neue Ideen leichter umgesetzt werden.

Welche weitergehenden Effekte ergeben sich aus der Open-Source-Stellung der Ergebnisse?

Die Open-Source-Stellung der Ergebnisse des RevLLM Projekts hat weitreichende positive Effekte, die sich sowohl auf die technische Gemeinschaft als auch auf die Gesellschaft im Allgemeinen erstrecken. Diese Effekte fördern Innovation, Transparenz und eine inklusivere Entwicklung im Bereich der künstlichen Intelligenz (KI). Hier sind einige Schlüsseleffekte:

Förderung der Innovation und Kollaboration
Durch die Bereitstellung von Code und Werkzeugen als Open Source wird eine Umgebung geschaffen, in der Forscher, Entwickler und Interessierte weltweit beitragen, die Software verbessern und auf neue Anwendungsfälle anpassen können. Diese kollaborative Natur beschleunigt die Forschung und Entwicklung neuer Technologien und Anwendungen.

Transparenz und Vertrauen
Open-Source-Software trägt zur Transparenz bei, indem sie es jedem ermöglicht, den Code zu überprüfen und zu verstehen, wie die Ergebnisse zustande kommen. Dies ist besonders in der KI-Entwicklung von Bedeutung, da es hilft, Vertrauen in die Technologie aufzubauen, indem es Einblicke in ihre Funktionsweise und Entscheidungsprozesse bietet.

Zugänglichkeit von KI-Technologien
Die Zugänglichkeit von Open-Source-Ressourcen ermöglicht es einem breiteren Spektrum von Menschen, KI-Technologien zu erforschen, zu lernen und anzuwenden. Dies reduziert die Barrieren für den Einstieg und fördert eine breitere Verteilung des Wissens und der Fähigkeiten, was essentiell ist, um eine gerechte technologische Zukunft zu gewährleisten.

Bildung und Selbstbildung
Open-Source-Projekte dienen als wertvolle Lernressourcen für Studierende, Lehrkräfte und Selbstlerner. Sie bieten praktische Einblicke in die Programmierung, das Design von Algorithmen und die Arbeitsweise von KI-Modellen, was zur Entwicklung der technischen Kompetenzen und zur Förderung der Data Literacy beiträgt.

Gibt es Ideen für die Weiterentwicklung Deiner Lösung und Pläne zu deren Umsetzung?

Ich möchte weiterhin an dem RevLLM Projekt arbeiten und möglicherweise auch nach Fördermöglichkeiten für diese Arbeit suchen. Ich habe viele Ideen wie ich die Bibliothek verbessern und für die Benutzer*innen attraktiver machen kann. Hier einige Beispiele:

- Implementierung weiterer Interpretierbarkeitsmethoden.
- Implementierung der Unterstützung für weitere Modelle (insbesondere aktuelle Open Source Modelle aus der Huggingface-Sammlung).

Hat die Arbeit in dem Projekt Dich in Deiner persönlichen, fachlichen Weiterentwicklung unterstützt?

Ja, die Arbeit an dem RevLLM-Projekt war eine gute Gelegenheit, tief in spezifische Technologien einzutauchen, in diesem Fall in die Mechanismen hinter großen Sprachmodellen und Machine Learning-Techniken. Das hat mein Verständnis und die Fähigkeiten in diesen Bereichen erheblich verbessert. Insbesondere habe ich viel Neues über die Interpretierbarkeitsmethoden und über die Bibliothek Pytorch gelernt.

Ich habe auch das offene und eigenständige Arbeiten sehr genossen, was für meine Kompetenzen in Bereichen Projektmanagement und Kommunikation sehr fordernd war.

Kurze Darstellung der Arbeiten, die zu keiner Lösung geführt haben

Gab es Arbeiten bzw. Lösungsansätze, die nicht weiter verfolgt wurden?

Ich habe am Anfang des Projekts das Ziel verfolgt, eine geeignete Modellfamilie für RevLLM zu finden. Ich habe mit vielen verschiedenen Modellen experimentiert bis ich schließlich entschieden habe, die nanoGPT Modelle zu verwenden. Somit gibt es eine Liste von Modellen, die nicht in das Projekt eingeflossen sind (Beispiel: Distilbert).

Was waren die Hintergründe, und wie bist Du alternativ vorgegangen?

Das Hauptproblem rund um die Auswahl der Modelle war die Größe der HuggingFace transformers-Bibliothek (ca. 700.000 Zeilen Code), über die viele Sprachmodelle geladen und instanziert werden. Diese Größe macht es schwierig, den Modellcode zu verstehen und zu modifizieren (Voraussetzung für viele der implementierten Methoden). Ein Ausweg war es, die nanoGPT-Bibliothek als Grundlage für den Modellcode zu wählen. nanoGPT ist sehr klein (im Vergleich zu der transformers-Bibliothek) und besteht aus Code, der einfach nachvollziehbar ist.

Kurze Angabe von Präsentationsmöglichkeiten für mögliche Nutzer

Wo können sich Interessenten detailliert über Deine Projektergebnisse informieren (z.B. Webseite, GitHub, Veröffentlichungen)?

https://github.com/jwergieluk/revllm (Source code)

https://revllm.dev/ (Homepage mit Dokumentation)

Kurze Erläuterung zur Einhaltung der Arbeits- und Kostenplanung

Gab es im Projektverlauf Ereignisse, die eine Anpassung der Planung erforderlich machten – z.B. Mehr- oder Minderaufwand bei der Bearbeitung von Teilaufgaben?

Nein.

Kurze Darstellung von etwaigen Ergebnissen bei anderen Stellen

Gab es Entwicklungen anderer Personen oder Institutionen, die Einfluss auf Deine Arbeiten und die Zielsetzung hatten? Wenn ja, worin bestand dieser und wie bist Du damit umgegangen?

Nein.

Julian Wergieluk: