AI-Werkstatt (UV & SE)

Allgemeines

Für alle Themen gelten folgende Richtlinien:

- Bearbeitung der Themen in **2er** oder **3er** Gruppen

- **Realistisch.** Für die Bearbeitung sollen ausdrücklich *nicht-fundamentale* Primitive, Basisalgorithmen etc. selbst nachimplementiert werden. Ziel ist vielmehr, realistische Anwendungen von KI-Verfahren in Unternehmen zu repräsentieren. Daher ist es ausdrücklich erwünscht, auf existierende Standard-Bibliotheken (od. zB GitHub Quellen), Basismodelle, usw. zurückzugreifen und auf deren Basis möglichst praxisnahe, AI-basierte Anwendungen umzusetzen

- **Systemkontext.** Jedes Thema soll im Kontext eines realistischen Systemkontextes mit konkretem Anwendungsbezug bearbeitet werden. Die Bearbeitung beinhaltet daher auch die Spezifikation des angenommenen Anwendungsszenarios sowie des / von Use-Cases. In der Umsetzung soll neben dem AI-Kern auch dessen stimmige Integration in größere Anwendungskontexte und entsprechende Softwarestacks realisiert werden.

- **Experimentelle Bewertung.** In der Praxis ist auch die experimentelle Bewertung umgesetzter Lösungen besonders wichtig (wie gut sind die zu erwartende Ergebnisse? Mit welcher Performance müssen wir bei unterschiedlichen Lastsituationen rechnen? Wie viel mehr Ressourcen sind beispielsweise für eine größere Variante eines bestimmten LLM-Modells ggü. einer kleineren Variante notwendig und ist die höhere Qualität diesen zusätzlichen Zusatzaufwand wert? Wie wirken sich unterschiedliche Konfigurations-optionen auf Faktoren wie Qualität, Performance, etc. aus. Daher sollen geeignete experimentelle Bewertungen immer Bestandteil der erfolgreichen Themenbearbeitung sein.

Benotungskriterien

Folgend sind die Benotungskriterien bzw. Leistungsanforderungen für die UV und das SE genauer spezifiziert.

Übung mit Vorlesung (UV). Für die Benotung der UV relevant sind:

1. Erstellung eines Reports (ca. 3 Seiten) im IEEE 2-Spalten-Format (Vorlagen werden bereitgestellt). **Abgabetermin** voraussichtlich Mitte des Semesters (wird noch rechtzeitig bekanntgegeben, ca. Mitte Dez. 2024).
2. Erstellung eines kurzen Artikels (ca. 6 Seiten) im IEEE 2-Spalten-Format (Vorlagen werden bereitgestellt). Idealerweise wird hierzu der Report (sieht Punkt 1) mit den entsprechenden neuen Erkenntnissen erweitert, und nach den Prinzipien des guten wissenschaftlichen Arbeitens in eine potentiell veröffentlichbare Version gebracht. **Abgabetermin** ist Semesterende (ein genauer Termin wird noch fixiert).

Die Gewichtung der beiden Teile für die Gesamtnote beträgt 30/70 (also 30% Report, 70% Short-Paper).

Seminar (SE). Für die Benotung des SE relevant sind:

- Qualität des Quellcodes und der Auswertung (Reproduzierbarkeit, Systematik der Evaluierung etc.)
- **Presentation 1:** *Initial outline*
  Was wollen wir machen? Wie sieht der Plan zur Realisierung und Auswertung aus? Etc.
- **Presentation 2:** *Intermediate status*
  Was ist der aktuelle Stand? Gibt es bereits Resultate? Roadblocks? Etc.
- **Presentation 3:** *Final status*
  Zusammenfassung der Arbeit und der Resultate + evtl. Demo!
- Aktive Teilnahme an der Diskussion während den Präsentationen

Die Gewichtung einzelnen Teile für die Gesamtnote beträgt (in der Reihenfolge wie oben) 50/10/10/20/10 (also 50% Code, 10% Presentation 1, 10% Presentation 2, 20% Presentation 3, 10% aktive Teilnahme).

Themenauswahl

Die folgend dargestellten Themen sind Vorschläge und können nach Absprache mit den LV-Leitern erweitert bzw. modifiziert werden.

Wie verzerrt (biased) sind frei verfügbare, vortrainierte LLMs?

- **Kontext**: Für das eigene Betreiben sprachbasierter KI-Anwendungen existiert eine Vielzahl frei nutzbarer und vortrainierter LLMs, die in Ihrer Leistungsfähigkeit teilweise durchaus an LLM-Dienste großer Anbieter (Meta, OpenAI, etc.) heranreichen. Je nach Anwendungsfall kann es allerdings sein, dass ein möglicher Bias für die Entscheidung zwischen unterschiedlichen solcher Modelle maßgeblich ist. Ein Unternehmen, dessen Geschäftsmodell z.B. auf Empfehlungen zur Studien- oder Berufswahl basiert, kann kaum wollen, dass Menschen nur aufgrund ihres Vornamens bestimmte Berufe oder Studiengänge (z.B. KI, Veterinärwesen, Maschinenbau, …) nicht empfohlen bekommen.

- **Aufgabe / Fragestellung**: Finden und skizzieren Sie einen LLM-Anwendungsfall, in dem mindestens 3 möglichst unterschiedliche Bias-Risiken bestehen. Demonstrieren Sie die unerwünschten Auswirkungen dieser Biases in einem realistischen Anwendungsfall. Designen Sie einen Ansatz, mit dem sich die Biases für mindestens drei unterschiedliche LLMs experimentell bewerten lassen. Finden oder erstellen Sie ein hierzu geeignetes Test-Datenset und führen Sie die experimentellen Bewertungen durch.

- **Bonus / Stretch-Goal**: Lassen sich die betrachteten Modelle durch nachtrainieren (aka finetuning) mit geeigneten Daten "de-biasen"? Was ist dafür notwendig und ist dies technisch auch realisierbar (mit vertretbarem Resourcenaufwand)? Wie erfolgreich ist dies? Welche möglichen negativen Auswirkungen entstehen dadurch?

Datenschutzfreundliche KI-Nutzung

- **Kontext**: In nahezu allen lernenden KI-Anwendungen mit Bezug zu individuellen Personen stellt sich früher oder später die Frage nach der Privatheit der für das Training genutzten personenbezogenen ("persönlichen") Daten. Etwas vereinfacht dargestellt: Schon ein einfacher Klassifizierer, der z.B. mit Kunden- oder Bevölkerungsdaten trainiert wird, könnte etwa hochspezifische Attributkombinationen, die im Datensatz nur einmal vorkommen (106 Jahre alte Frau, die aus Postleitzahl 5122 kommt und ein steuerpflichtiges Einkommen von 1,23 Mio €/Jahr hat) übermäßig genau abbilden. Wer diese eine 106 Jahre alte Frau kennt, dem würde der Klassifizierer dann ggfs. deren Einkommen offenlegen. Als

Gegenmaßnahme wird in der Wissenschaft eine Vielzahl von Ansätzen zum "Privacy-preserving learning" vorgeschlagen, die aber - aus Gründen - in der Praxis kaum eine Rolle spielen. Alternativ hierzu existiert die (in der Praxis deutlich häufiger genutzte) Möglichkeit, Modelle auf bereits voranonymisierten Daten zu trainieren. Insbesondere kommen hier Verfahren mit Anonymitätsgarantien (k-Anonymität, L-Diversität, …) zum Einsatz. Häufig ist aber unklar, wie diese Anonymisierung die Qualität der KI-Ergebnisse beeinflusst.

- **Aufgabe / Fragestellung**: Finden und skizzieren Sie einen KI-Anwendungsfall, in dem personenbezogene (und idealerweise besonders kritische) Daten notwendigerweise für das Training genutzt werden. Implementieren Sie den Anwendungsfall (inkl. Training mit einem geeigneten Datensatz) und skizzieren Sie das Privatheitsproblem anschaulich. Führen Sie dann parallel auf dem Datensatz eine geeignete Anonymisierung durch und trainieren Sie das genutzte KI-Modell auf dem anonymisierten Datensatz. Wie wirkt sich die Anonymisierung auf die Qualität der Ergebnisse aus? Welchen Einfluss haben unterschiedliche Anonymisierungsparameter (z.B. unterschiedliche Werte für k und l)?

- **Bonus / Stretch-Goal**: Sind im Rahmen der vorgeschalteten Anonymisierung evtl. noch Optimierungen möglich, etwa durch geschickte Auswahl der zur Anonymisierung genutzten Attribute? Für einige Standardverfahren und -implementierungen (insb. scikit-learn) existieren außerdem mittlerweile prototypische Implementierungen z.B. für "differentially private learning". Wie schneiden diese im Vergleich ab?

Bewertung moderner Ansätze zur Zeitreihenvorhersage

- **Kontext**: Nahezu alle Unternehmen werden im Zuge der voranschreitenden Digitalisierung und der breiten Verfügbarkeit von AI-Methoden früher od. später in einem od. mehreren Bereichen Ihres Geschäftsmodells Prognosemodelle einsetzen (müssen), um in ihrem wirtschaftlichen Umfeld kompetitiv zu bleiben. Hierzu muss jedoch gründlich evaluiert werden, (1) welche existierenden Methoden und Modelle für welche Einsatzzwecke gedacht sind, (2) inwieweit komplexe Modelle überhaupt notwendig (/gerechtfertigt) sind (beispielsweise im Vergleich zu „einfachen" statistischen Modellen), und (3) wie unterschiedliche existierende Ansätze auf realitätsnahen Daten abschneiden. Zudem sollte der Resourcenbedarf dieser Ansätze evaluiert werden, bzw. sichergestellt sein, dass bei bereits vortrainierten Modellen (Stichwort: „foundation models") die Möglichkeit zum Finetuning besteht, bzw. die Ansätze entsprechend dokumentiert sind (z.B. Publikationen, etc.).

- **Aufgabe / Fragestellung**: Identifizieren Sie einen geeigneten Benchmark Datensatz, mit dem Sie arbeiten möchten (*Vorschlag*: der M4 Competition Datensatz zur Evaluierung univariater Zeitreihenvorhersagemodelle). Identifizieren Sie folgend 2-3 Ansätze zur Zeitreihenvorhersage mit frei verfügbaren Implementierungen; suchen Sie die entsprechende Literatur zu diesen Ansätzen und versuchen Sie die Methoden und Modelle zu verstehen (ist die Methode überhaupt für das Problem geeignet? Haben wir genügend Daten? Ist der Ansatz für uni- oder multivariate Zeitreihen, etc.?). Testen und evaluieren Sie die Ansätze mittels geeigneter Performanz Maße (hier gibt es tatsächlich keine „Standard" Methode, sondern viele Möglichkeiten mit entsprechenden Vor- und Nachteilen).

- **Bonus / Stretch-Goal**: Ist es möglich die verschiedenen Ansätze zu kombinieren, um bessere Vorhersagen zu erhalten? Welche Möglichkeiten gibt es hier? Was ist sinnvoll, was nicht? Ist es möglich, bevor eine Vorhersage getroffen wird, zu entscheiden ob überhaupt eine Vorhersage getroffen werden soll?

Maschine Learning Methoden zur Analyse & Modellierung kollektiven Verhaltens

- **Kontext**: Kollektives Verhalten bezieht sich auf die koordinierten Aktionen großer Gruppen von Individuen, wie z.B. Vögel od. Fische in einem Schwarm, die sich als kohärente Einheit ohne zentrale Führung bewegen. Jedes Individuum folgt einfachen Regeln, die auf den Bewegungen seiner nächsten Nachbarn basieren, was zu komplexen, synchronisierten Mustern führt. Dieses Verhalten hilft der Gruppe oft, Raubtieren zu entgehen, Nahrung zu finden oder effizient durch ihre Umgebung zu navigieren. Um solches Verhalten zu verstehen, existieren verschiedene erstaunlicherweise einfache Modelle (im Wesentlichen Bewegungsgleichungen mit entsprechender Miteinbeziehung der Interaktionen) die sich im Laufe der Zeit etabliert haben (z.B. das klassische Vicsek od. d'Dorsogna Modell). Diese Modelle sind über verschiedene Kenngrößen parametrisiert (z.B. wie stark interagieren Vögel mit ihren unmittelbaren Nachbarn in einem Schwarm abhängig von deren Distanz) und können relativ leicht simuliert werden (dazu gibt es auch einige gute Softwarebibliotheken). Zur Simulation startet man beispielsweise von einer zufällig initialisierten Punktewolke (in 2D od. 3D) und beobachtet die entstehenden Muster. Je nach Parametrisierung wird man unterschiedliche Effekte sehen. Obwohl Simulationen od. die formale Analyse interessant sind, stellt sich oft die Frage nach Lösungsansätzen für das „inverse Problem", d.h., aus tatsächlichen Beobachtungen (z.B., Videos von Vogelschwärmen) passende Parameter eines zuvor gewählten Modells zu identifizieren. Würde man dann das System simulieren, sollte man idealerweise die beobachteten Muster (zumindest qualitativ) beobachten können. Existierende Ansätze

basieren aber meist darauf die genauen „Bahnen“ der Individuen zu kennen (also die Bewegungstrajektorien der Individuen).

- **Aufgabe / Fragestellung**: Nutzen Sie existierende Bibliotheken zur Simulation kollektiven Verhaltens mittels verschiedener Modelle. In 2D/3D würden Sie z.B. sich über die Zeit verändernde Punktewolken erhalten. Variieren Sie die Parameter und generieren Sie so eine Menge an „Trainingsdaten“. Folgend identifizieren Sie passende Ansätze solche Sequenzen von Punktwolken verarbeiten zu können (z.B. rekurrente neuronale Netze mit entsprechender Funktionalität Punktewolken als Input zu erhalten). Trainieren und evaluieren Sie folgend, wie gut es möglich ist mittels solcher Ansätze die Parameter (meist 1-4) des Simulationsmodells vorherzusagen. Machen Sie sich auch Gedanken darüber welche Performanz Maße zur Evaluierung verwendet werden sollten, also welche Vor- und Nachteile die jeweils verwendeten Metriken haben.

- **Bonus / Stretch-Goal**: Die Aufgabenstellung kann beliebig erweitert werden, vor allem hinsichtlich der Anzahl an simulierten Modellen und möglichen Ansätzen zur Verarbeitung der Sequenzen.