Maschinelle Sprachverarbeitung mit Large Language Models

Von Bag of Words bis Agents

Tim König, Universität Hildesheim, 27. & 28.2.2025

Recap

- Bag of Words
- Static Word Embeddings
- Dynamische Word Embeddings
- Attention-Mechanismus & Transformerarchitektur
- Classification mit BERT und verwandten Modellen

Tag 1

- Bag of Words & Word Embeddings
- Transformer
- Classification mit Transformer-Modellen
- Model Evaluation

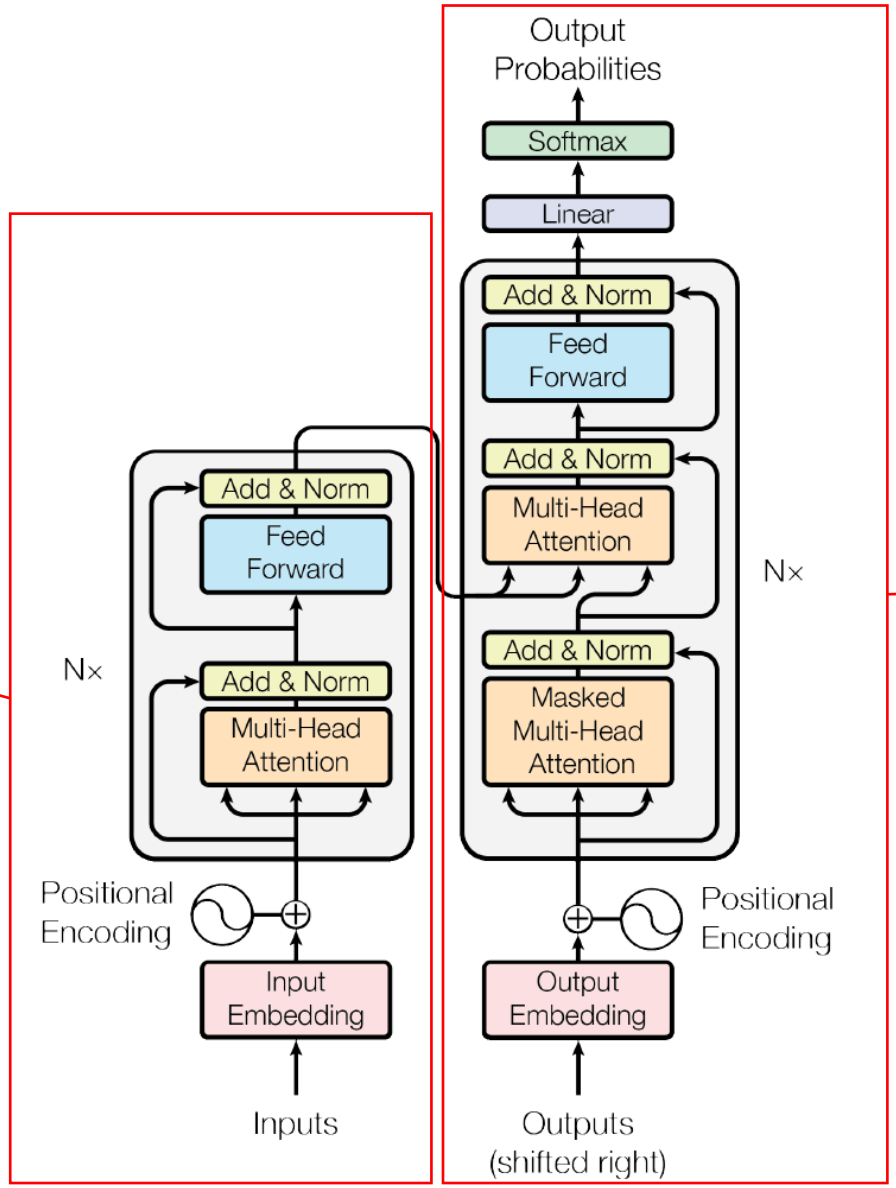
Tag 2

- Classification mit Large Language Models
  - OpenAIs GPT-Modelle
  - Ollama-Framework
- Agents & kombinierte Modelle mit Langchain

Large Language Models

Transformer

BERT, RoBERTa
(Encoder-only)



GPT, Llama
(Decoder-only)

Prinzip

- Auto-regressives Language Modeling: Modell lernt, aus einer Sequenz von Wörtern das nächste vorauszusagen

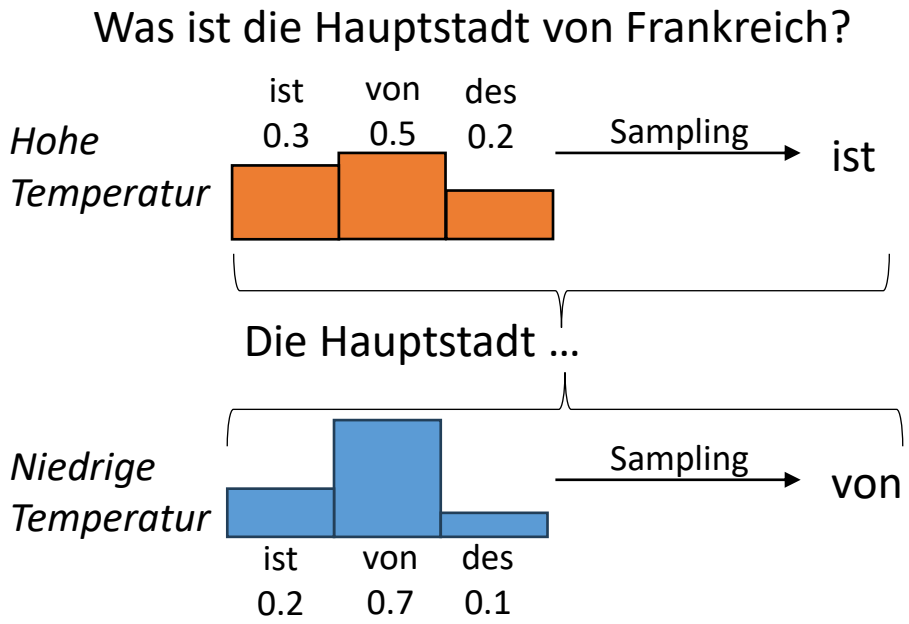
  GPT is a model popular in [MASKED].

- Decoder: Generiert Output-Sequenz auf Grundlage einer Input-Sequenz
- Generiert immer *Token für Token,* auf Grundlage von Input und bereits erzeugtem Output:

  Was ist die Hauptstadt von Frankreich? Die Hauptstadt von Frankreich ist Paris.

Temperature

Der „Temperature“-Parameter variiert die Wahrscheinlichkeitsverteilung der Textgeneration



Unterschiedliche Sampling-Techniken, z.B. *greedy* (wahrscheinlichste Wort), *top-k* (sampling aus k wahrscheinlichsten Worten)…

➡ Textgeneration durch LLMs ist ***immer*** probabilistisch!
Dadurch sind *Halluzinationen* möglich!

ChatGPT & Co.

- „Chat“-Modelle sind Decoder-Modelle, welche mittels Supervised Learning auf Frage-Antwort-Muster trainiert wurde (*Alignment*) – ähnlich dem Fine-Tuning von BERT-Modellen
- Daten für das Supervised Learning („*Gold Standard*“) idR. durch unterbezahlte Gigworker bereitgestellt – insb. in Ländern des globalen Südens
- LLMs benötigen große Mengen an Daten, um Sprache möglichst umfassend statistisch zu repräsentieren und große Mengen menschlicher Codierarbeit, um sie für unsere Zwecke nutzbar zu machen
- „Artificial Intelligence“ ist ein Marketingbegriff – das Prinzip von LLMs unterscheidet sich nicht grundlegend von den NLP-Techniken, die wir bereits kennengelernt haben

LLMs in der Forschung

- Mächtige Tools, deren kostenintensives Training bereits durchgeführt wurde
- Chat-Logik macht Modelle sehr flexibel – es müssen „nur” Anweisungen gegeben werden
- Zugriff über APIs skalierbar – kein Chatfenster!
- Modelle sind häufig proprietär, was insb. bei der Verarbeitung personenbezogener Daten problematisch sein kann (Verarbeitung auf Servern im Ausland etc.)
- Zunehmend auch leistungsstarke Open Source Modelle verfügbar

Ollama

- Framework, das es mittels verschiedener Techniken ermöglicht, LLMs mit geringerem Hardware-Aufwand laufen zu lassen
- *Quantization*: unterschiedliche Techniken, um die Größe von LLMs zu verringern, indem die Präzision (Nachkommastellen) der Vektorenwerte verringert oder komprimiert werden
- LLMs können auch ohne GPU verwendet werden
- Nur Open Source Modelle verfügbar
- Aber: Große Modelle sind immernoch hardwareintensiv!

Prompt Engineering

- „Kunst“ generativen Modellen möglichst gute & konsistente Antworten zu entlocken
- Unterschiedliche Techniken, z.T. modellspezifisch:
  - Klare & ausführliche Aufgabenbeschreibung
  - Klare Trennung von Input und Aufgabe
  - Positive Formulierungen („du sollst“ statt „du sollst nicht“)
  - Einzelschritte ausformulieren
  - Das Modell anhalten, Einzelschritte zu nennen (*Reasoning / Chain of Thought*)
  - Beispiele nennen (*Few-Shot Classification*)
  - Persona („Du bist ein*e Politkwissenschaftler*in…“)
  - …
- Zunehmend Versuche, Prompt Templates zu standardisieren (z.B. hier)

Fine-Tuning von Large Language Models

- LLMs haben keinen Classification Layer – daher muss immer das ***gesamte*** Modell gefinetuned werden!
- Große Modelle brauchen entsprechend viele Trainingsdaten, damit Fine-Tuning einen Effekt hat
- Techniken wie Quantization oder PEFT (Parameter Efficient Fine Tuning) erlauben effizienteres Fine-Tuning*
- Fine-Tuning von LLMs idR. hardwareintensiv
- *Fine-Tuning auf einen Task verschlechtert meist die Performance in allen anderen Tasks!*

*Siehe dazu auch das *Ludwig* Framework: https://ludwig.ai/

Vor- und Nachteile

Vorteile

- Sehr flexibel, benötigen idR. kein Fine-Tuning
- Intuitive „Bedienung“ über Freitextanweisungen
- Unterschiedliche Techniken wie Prompt Engineering oder Few-Shot Classification können Ergebnisse verbessern
- Kleinere Modelle können auch lokal laufen (Datenschutz!)

Nachteile

- Inhärent probabilistisch – Modelle generieren den wahrscheinlichsten, nicht unbedingt den richtigen Output (Halluzinationen!)
- Prompting mitunter esoterisch und (noch) wenig standardisiert
- Große Modelle benötigen idR. (kostenaufwändige) Compute Power oder proprietäre APIs
- Fine-Tuning von LLMs ist aufwändig und bei großen Modellen (fast) unmöglich

Coding Time…

Agents

Agents

- Unterschiedliche Definitionen von „Agents“, aber im Kern: LLM-basierte Systeme, welche komplexe Aufgaben bewältigen, indem sie Teilschritte an andere Systeme auslagern
- Andere Systeme sind idR. *Tools* (z.B. ein Python-Terminal oder eine Datenbank) oder andere LLMs (um die Stärken unterschiedlicher Modelle zu kombinieren)
- LLMs wie ChatGPT werden idR. eingesetzt, um zu entscheiden, wann ein Tool (z.B. Python für mathematische Berechnungen) oder ein anderes Modell (z.B. DALL-E für die Bildgeneration) verwendet werden sollen

Anwendungen

- Tools kompensieren die Schwächen von LLMs – z.B. bei mathematischen Berechnungen
- Die meisten „intelligenten“ Anwendungen von z.B. ChatGPT, etwa das erstellen von Plots, werden durch *Tools* (z.B. Python-Zugriff) ermöglicht
- Multimodale Systeme, z.B. Bildgeneration in ChatGPT, werden idR. durch verknüpfte Modelle (*Agent Swarms*) erzeugt
- Anwendung in der Forschung?

Retrieval Augmented Generation (RAG)

- LLMs können sog. Retriever als *Tool* nutzen, um auf Datenbanken zuzugreifen
- Retriever können auf Textdatenbanken zugreifen, aber z.B. auch Google oder die Wikipedia API nutzen
- Die erhaltenen Informationen werden idR. als Kontext hinzugefügt, um die Antwort des LLMs zu verbessern
- RAG zielt v.a. darauf ab, faktische Informationen in Antworten zu verbessern – und somit Halluzinationen des Modells zu verringern