UNIVERSITÄT
LEIPZIG

Medizinische Fakultät

Vortrag - 2

Question Answering auf dem Lehrbuch „Health Information Systems“ mit Hilfe von unüberwachtem Training eines Pretrained Transformers

Leipzig, Juli 2023

Paul Keller

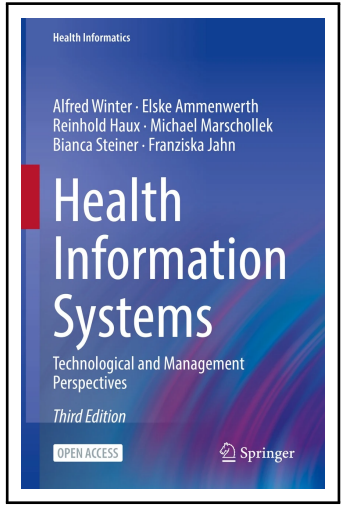
imise.

Überblick

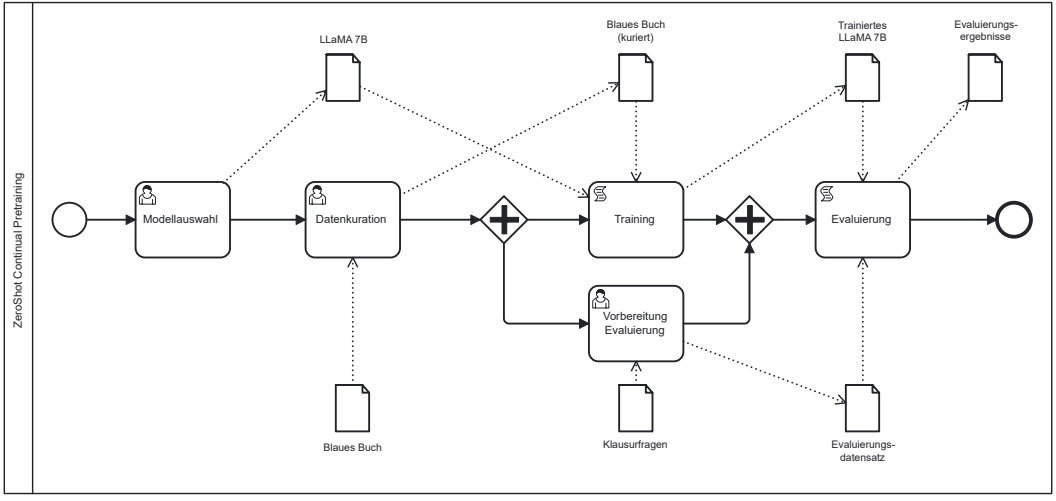
- Einleitung
  - Ziel
  - Aufgaben
- Lösungsansatz
  - Auswahl eines Sprachmodells
  - Datenkuration
  - Training des Modells
  - Klausurfragen
  - Modellevaluation

Einleitung - Ziel

- Fragen beantworten zu „Health Information Systems“ mit einer Konversations-KI
- Beispielklausuren lösen aus dem Modul „Architektur von Informationssystemen im Gesundheitswesen“ mit einer Konversations-KI



Einleitung - Aufgaben



Nicht zur Auswahl stehende Modelle

1. BERT, RoBERTa, DistillBERT
2. BART, T5, Flan-T5
3. DialoGPT, BlenderBot, Opus

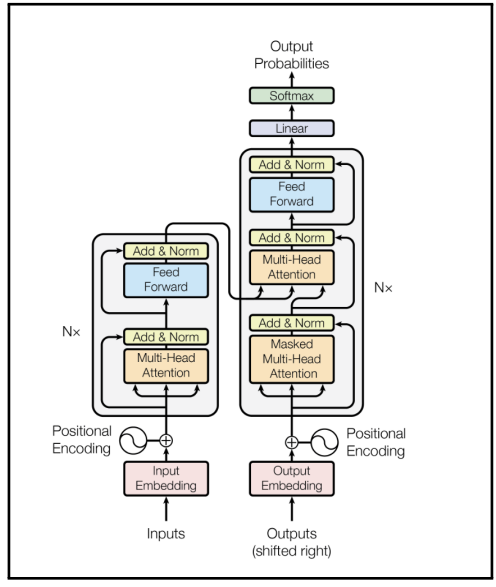


Abbildung: Transformer-Model Architektur (Vaswani u. a. (2017), S. 3)

Nicht zur Auswahl stehende Modelle

1. BERT, RoBERTa, DistillBERT
2. BART, T5, Flan-T5
3. DialoGPT, BlenderBot, Opus

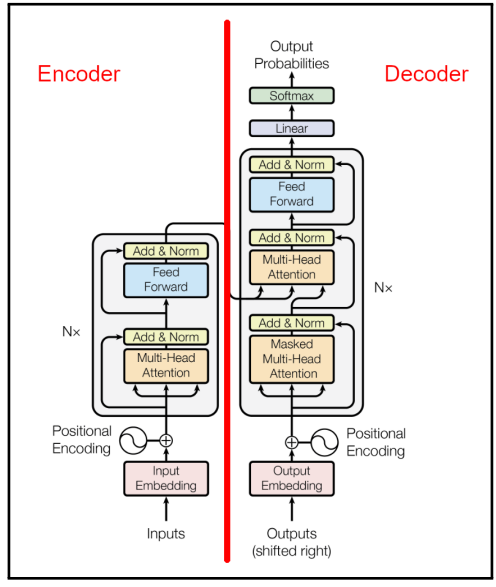


Abbildung: Transformer-Model Architektur (Vaswani u. a. (2017), S. 3)

Zur Auswahl stehende Modelle

Modell	Verfügbarkeit	ZeroShot	Größe	Training	Inferenz
GPT-2	frei	schlecht	1.5 Mrd Parameter	24 GB	6 GB
GPT-3	nur Inferenz	gut	175 Mrd Parameter	1.4 TB	350 GB
GPT-4	nur Inferenz	sehr gut	—	—	—
GPT-J	frei	gut	6.7 Mrd Parameter	104 GB	26 GB
GPT-NeoX	frei	gut	20 Mrd Parameter	160 GB	40 GB
LLaMA	limitiert	gut	7 Mrd Parameter	56 GB	14 GB
	limitiert	gut	13 Mrd Parameter	104 GB	23 GB
	limitiert	sehr gut	33 Mrd Parameter	264 GB	66 GB
	limitiert	sehr gut	65 Mrd Parameter	520 GB	130 GB

Tabelle: Training Speicherbedarf = Anzahl Parameter $\cdot \frac{\text{Parameterformat}}{8} \cdot 4$

Das LLaMA Modell

		BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA
GPT-3	175B	60.5	81.0	-	78.9	70.2	68.8	51.4	57.6
Gopher	280B	79.3	81.8	50.6	79.2	70.1	-	-	-
Chinchilla	70B	83.7	81.8	51.3	80.8	74.9	-	-	-
PaLM	62B	84.8	80.5	-	79.7	77.0	75.2	52.5	50.4
PaLM-cont	62B	83.9	81.4	-	80.6	77.0	-	-	-
PaLM	540B	**88,0**	82.3	-	83.4	**81,1**	76.6	53.0	53.4
LLaMA	7B	76.5	79.8	48.9	76.1	70.1	72.8	47.6	57.2
	13B	78.1	80.1	50.4	79.2	73.0	74.8	52.7	56.4
	33B	83.1	82.3	50.4	82.8	76.0	**80,0**	**57,8**	58.6
	65B	85.3	**82,8**	**52,3**	**84,2**	77.0	78.9	56.0	**60,2**

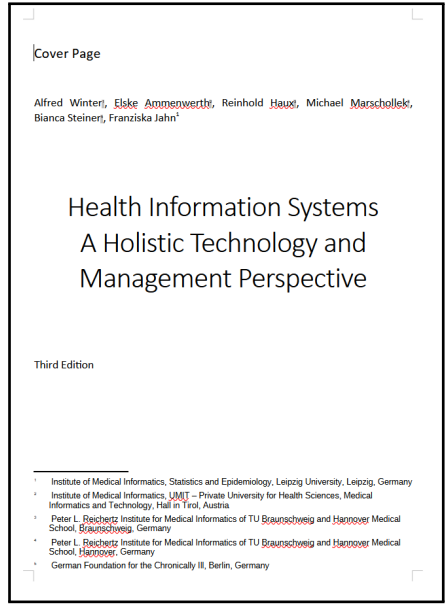
Tabelle: Zero-Shot Leistung verschiedener Modelle in Begründungsaufgaben für gesunden Menschenverstand. Veröffentlicht in Touvron u. a. (2023)

Datenkuration - Extraktion

- Vorliegend als Word-Datei
- Umwandlung in Markdown-Format
- Übungsabschnitte + Lösungen = Fragen

Unverständliche Formate

- Bilder
- Bildunterschriften und Referenzen auf Bilder
- Kopf- und Fußzeilen
- Seitenzahlen



Cover Page

Alfred Winter[1], Elske Ammenwerth[2], Reinhold Haux[3], Michael Marschollek[4], Bianca Steiner[5], Franziska Jahn[5]

Health Information Systems
A Holistic Technology and
Management Perspective

Third Edition

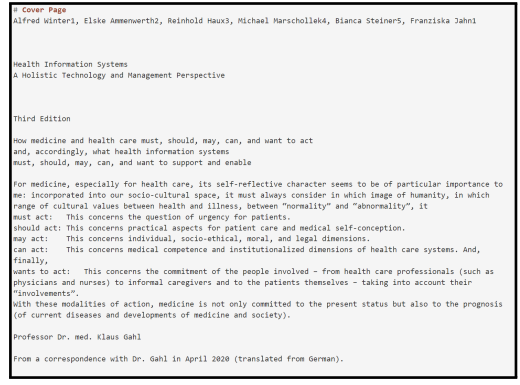
[1] Institute of Medical Informatics, Statistics and Epidemiology, Leipzig University, Leipzig, Germany
[2] Institute of Medical Informatics, UMIT – Private University for Health Sciences, Medical Informatics and Technology, Hall in Tirol, Austria
[3] Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Brautschweig, Germany
[4] Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Hannover, Germany
[5] German Foundation for the Chronically Ill, Berlin, Germany

Datenkuration - Extraktion

- Vorliegend als Word-Datei
- Umwandlung in Markdown-Format
- Übungsabschnitte + Lösungen = Fragen

Unverständliche Formate

- Bilder
- Bildunterschriften und Referenzen auf Bilder
- Kopf- und Fußzeilen
- Seitenzahlen



```
# Cover Page
Alfred Winter1, Elske Ammenwerth2, Reinhold Haux3, Michael Marschollek4, Bianca Steiner5, Franziska Jahn1



Health Information Systems
A Holistic Technology and Management Perspective



Third Edition

How medicine and health care must, should, may, can, and want to act
and, accordingly, what health information systems
must, should, may, can, and want to support and enable

For medicine, especially for health care, its self-reflective character seems to be of particular importance to
me: incorporated into our socio-cultural space, it must always consider in which image of humanity, in which
range of cultural values between health and illness, between "normality" and "abnormality", it
must act:   This concerns the question of urgency for patients.
should act: This concerns practical aspects for patient care and medical self-conception.
may act:    This concerns individual, socio-ethical, moral, and legal dimensions.
can act:    This concerns medical competence and institutionalized dimensions of health care systems. And,
finally,
wants to act:   This concerns the commitment of the people involved - from health care professionals (such as
physicians and nurses) to informal caregivers and to the patients themselves - taking into account their
"involvements".
With these modalities of action, medicine is not only committed to the present status but also to the prognosis
(of current diseases and developments of medicine and society).

Professor Dr. med. Klaus Gahl

From a correspondence with Dr. Gahl in April 2020 (translated from German).
```

Datenkuration - Textpassagen ohne Wissen

- Titelseite
- Vorwort
- Vorstellung der Autoren
- Tabellenverzeichnis
- Fehlerhafte Textpassagen (fehlende Verweise)

Training des Modells

1. Text in Trainings- und Validierungsdatensatz aufteilen
2. Datensätze in Token umwandeln
3. Token in Batches aufteilen
4. Batches mit Modell berechnen
5. Backpropagation auf das Modell anwenden

Definition

Overfitting beschreibt den Prozess, bei dem ein Modell die Trainingsdaten auswendig lernt und nicht mehr in der Lage ist, neue Daten korrekt zu klassifizieren. Overfitting kann durch einen Leistungsabfall zwischen Trainings- und Testdaten erkannt werden.

Training des Modells

1. Text in Trainings- und Validierungsdatensatz aufteilen
2. Datensätze in Token umwandeln
3. Token in Batches aufteilen
4. Batches mit Modell berechnen
5. Backpropagation auf das Modell anwenden

Definition

Ein **Token** repräsentiert eine Teilmenge eines Wortes und wird in Transformer-Modellen verwendet, um die Eingabe in logische Einheiten zu unterteilen.

Training des Modells

1. Text in Trainings- und Validierungsdatensatz aufteilen
2. Datensätze in Token umwandeln
3. Token in Batches aufteilen
4. Batches mit Modell berechnen
5. Backpropagation auf das Modell anwenden

Definition

Ein **Batch** repräsentiert eine Teilmenge von Eingaben, welche geschlossen verarbeitet werden, bevor die Gewichte und Bias angepasst werden.

Training des Modells

1. Text in Trainings- und Validierungsdatensatz aufteilen
2. Datensätze in Token umwandeln
3. Token in Batches aufteilen
4. Batches mit Modell berechnen
5. Backpropagation auf das Modell anwenden

Definition

Backpropagation beschreibt den Prozess der iterativen Anpassung von Gewichten und Bias auf Basis des Gradienten einer Fehlerfunktion.

Training des Modells - Bibliotheken

- PyTorch
- Transformers (Huggingface)
- DeepSpeed (Microsoft)

Parameter	**Wert**
Lernrate	$3e^{-4}$
AdamW β_1	0.9
AdamW β_2	0.95
Gewichtsreduktion	0.1
Gradientenlimitierung	1.0
Aufwärmphase	0
Epochen	2

Tabelle: Parameter für das Training des LLaMA 7B-Modells

Training des Modells - Ausführung

- Rechenzentrum Universität Leipzig
- Supercomputer Paula und Clara
- SLURM Skripte
- Gemeinsam genutzte Ressourcen

Paula	**Clara**
12 Nodes	31 Nodes
Nvidia A30 GPU (24GB)	Nvidia V100 GPU (32GB) Nvidia RTX 2080 TI GPU (11GB)
10757 CUDA Cores	9472 CUDA Cores

Evaluierungsdatensatz - Klausurfragen

Quellen

- Mündliche Prüfung „Architektur von Informationssystemen im Gesundheitswesen“
- Schriftliche Prüfung „Informationssysteme in medizinischer Versorgung und Forschung“
- Übungsaufgaben aus „Health Information Systems“

Fragenkategorien

1. **Einzelfragen**, die einen bestimmten Sachverhalt abfragen
2. **Multi-Fakten-Fragen**, die mehrere Fakten abfragen
3. **Transferfragen**, die einen Sachverhalt von einem Kontext in einen anderen übertragen

Klausurfragen - Umformulierung

- Kein Fine-Tuning möglich
- Fragen resultieren in Nachahmung
- „Welche Farbe hat der Himmel?“ → „Welche Farbe hat der Strand?“
- Umformulierung in unvollständige Aussagen: „Der Himmel hat die Farbe...“

- Aufbauende Fragen → In sich geschlossene Fragen
- Zeichnungen, Verbindung von Begriffen mit Pfeilen
  - Systembeschreibung, Begriffslisten

Modellevaluation - Antwortkategorien

1. **Richtig**, wenn die Antwort mit der Antwort im Buch übereinstimmt
2. **Falsch**, wenn die Antwort falsches Wissen enthält
3. **Nicht beantwortet**, wenn der generierte Text keinen Bezug zur Frage hat

Modellevaluation - Metriken

1. F1 = Harmonisches Mittel zwischen Präzision und Recall
2. MikroF1 = Berechnung über alle Fragen & Antworten
3. MakroF1 = Mittel der F1-Werte pro Frage
4. MikroF1 instabil bei vielen falschen Antworten $\rightarrow$ MakroF1
5. MakroF1 mehrheitlich bei Evaluierung genutzt (Usbeck u. a. (2019))

$C :=$ korrekt beantwortet

$S :=$ beantwortet

$G :=$ Goldener Standard

$$C = S \cap G$$

$$prec = \frac{|C_q|}{|S_q|}$$

$$recall = \frac{|C_q|}{|G_q|}$$

$$F1 = 2\frac{prec \cdot recall}{prec + recall}$$

Modellevaluation - Kriterien

„ChatGPT versus Traditional Question Answering for Knowledge Graphs[...]“ Omar u. a. (2023)

- Korrektheit: MakroF1
- Determinismus: 3x gleiche Fragen
- Robustheit: Rechtschreibfehler & grammatikalische Fehler
- Erklärbarkeit: Antwort + Erläuterung (zusätzliche Eingabe)
- Fragenverständnis: Antwort bezieht sich inhaltlich auf Frage

Quellen

- Ben Wang (2021). „Mesh-Transformer-JAX: Model-Parallel Implementation of Transformer Language Model with JAX“. In: https://github.com/kingoflolz/mesh-transformer-jax
- Black, Sidney u. a. (Mai 2022). „GPT-NeoX-20B: An Open-Source Autoregressive Language Model“. In: Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models. Association for Computational Linguistics, S. 95–136. doi: 10.18653/v1/2022.bigscience-1.9.
- Brown, Tom u. a. (2020). „Language Models are Few-Shot Learners“. In: Advances in Neural Information Processing Systems. Hrsg. von H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan und H. Lin. Bd. 33. Curran Associates, Inc., S. 1877–1901. url: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf
- Omar, Reham, Omij Mangukiya, Panos Kalnis und Essam Mansour (2023). ChatGPT versus Traditional Question Answering for Knowledge Graphs: Current Status and Future Directions Towards Knowledge Graph Chatbots. Version 1. arXiv: 2302.06466 [cs.CL].
- OpenAI (2023). „GPT-4 Technical Report“. In: arXiv: 2303.08774[cs.CL].
- Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei und Ilya Sutskever (2019). „Language Models are Unsupervised Multitask Learners“. In: url: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.
- Touvron, Hugo u. a. (2023). LLaMA: Open and Efficient Foundation Language Models. arXiv: 2302.13971 [cs.CL].
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser und Illia Polosukhin (2017). „Attention is All you Need“. In: Advances in Neural Information Processing Systems. Hrsg. von I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan und R. Garnett. Bd. 30. Curran Associates, Inc. url: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
- Winter, Alfred, Elske Ammenwerth, Reinhold Haux, Michael Marschollek, Bianca Steiner und Franziska Jahn (2023). Health Information Systems. 3. Aufl. Health Informatics. Springer Cham. isbn: 978-3-031-12310-8. doi: 10.1007/978-3-031-12310-8.
- Usbeck, Ricardo, Michael Röder, Michael Hoffmann, Felix Conrads, Jonathan Huthmann, Axel-Cyrille Ngonga-Ngomo, Christian Demmler und Christina Unger (2019). „Benchmarking question answering systems“. In: Semantic Web 10.2, S. 293–304.

UNIVERSITÄT
LEIPZIG

Medizinische Fakultät

VIELEN DANK!

Paul Keller
Institut für Medizinische Informatik, Statistik und Epidemiologie (IMISE)
www.imise.uni-leipzig.de