Transformer Decoder Architektur

1st Stefan Maier
Function and Algorithm
ZF Lifetec
Alfdorf, Deutschland
Stefan.Maier1@zf.com

***Abstract*—Große Sprache Modelle (engl. Large Language Models) wie Chat GPT basieren auf der Transformer Architektur neuronaler Netze. Die Transformer Architektur besteht dabei aus einem Encoder- sowie Decoder-Element. In der Wissenschaft und Forschung existieren auch Architekturen welche nur einen Teil eines Transformers zur Lösung eines Problems verwenden. In dieser Arbeit wird genauer auf Transformer Architekturen eingegange, welche lediglich aus dem Decoder Element bestehen. Eine weit Verbreitete Decoder Architektur ist die von Chat GPT verwendete GPT (kurz: Generative Pretrained Transformer) Architektur. Es wird auf die Architektur und die weiter Entwicklung der Architektur eingegangen.**

***Index Terms*—component, formatting, style, styling, insert.**

I. Einführung

Die Verarbeitung von natürlicher Sprache (engl. Natural Language Processing. kurz NLP) ist ein wichtiger und großer Bestandteil der künstlichen Intelligenz Forschung. Die Forschungsbereiche decken dabei ein breites Spektrum an Aufgaben ab, wie z.B. Beantwortung von Fragen, Semantische Ähnlichkeit, Text Generierung, Dokumenten Klassifikation o.ä. Die ersten Fortschritte wurde im NLP Bereich bereits durch statistische Hidden Markov Modelle erzielt. Die Leistung solcher statistischen Modelle ist jedoch begrenzt und konnte die Komplexität der natürlichen Sprache nicht geeignet abbilden. Mit den Entwicklungen im Bereich des Deep-Learnings gelangen entscheidende Schritte in der NLP Forschung. Die Transformer Architektur, und das damit verbundene vortrainieren großer Sprachmodelle wie BERT, GPT,T5 oder RoBERTa haben große Fortschritte gebracht, die sich nicht nur auf die NLP Forschung beschränken sondern in der allgemeinen Gesellschaft und Wirtschaft Einzug halten.

A. Ziel der Arbeit

Das Ziel der Arbeit ist es die Decoder Architektur welche Häufig von generativen Sprachmodellen verwendet wird genauer zu beleuchten. Dabei wird die Decoder-only Architektur anhand des Generative Pretrained Transformers erklärt. Neben dem Generative Pretrained Transformer existieren ebenfalls weitere Modelle die kurz erläutert werden. Weiterhin soll die Arbeit die Anwendungsgebiete sowie Potenziale und Grenzen der Decoder Architektur aufzeigen. Im Ausblick wird auf aktuelle Forschungsfelder im Bereich der Decoder Architekturen eingegangen.

B. Aufbau und Struktur der Arbeit

Die Arbeit ist wie folgt strukturiert: In Kapitel 2 wird zunächst auf die Grundlagen in Form der Decoder Architektur als Teil der Transformer Architektur eingegangen. Dazu wird die historische Entwicklung, der Aufbau und der Trainingsprozess der Decoder Architektur erläutert und diese vom Encoder Teil der Transformer Architektur abgegrenzt. In Kapitel 3 wird auf unterschiedliche mögliche Anwendungsgebiete aus der Forschung und Entwicklung eingegangen. Aus diesen Erkenntnissen werden in Kapitel 4 die Potenziale und damit einhergehenden Grenzen solcher großen Sprachmodelle eingegangen. Kapitel 5 fasst die Ergebnisse zusammen und gibt einen Ausblick auf die aktuellen Forschungsbereiche.

II. Decoder Architektur

Im folgenden Abschnitt wird die Decoder Architektur am Beispiel des Generative Pretrained Transformers (GPT) erläutert. Dabei wird auf die historische Entwicklung, den Aufbau, den Trainingsprozess und die Abgrenzung zur Encoder Architektur eingegangen.

A. Historische Entwicklung der GPT-Decoder Architektur

Transformer bilden das Grundgerüst moderne großer Sprachmodelle. Diese wurden 2017 in [1] beschrieben. 2018 wurde die erste Version von GPT veröffentlicht, welche auf dem Decoder Part der Transformer Architektur aufbaut. Auf Basis der GPT Modell entstanden weitere Decoder-only Modelle welche die folgende Figure 1 gut veranschaulicht. Die Figure 1 zeigt die Entwicklung unterschiedlicher Transformer Modell bis zum letzten Jahr. Im Rahmen dieser Seminararbeit zeigt Decoder-Only Zweig die Entwicklung von reinen Decoder Architekturen. Damit ist GPT1 eines der ersten Modelle welche auf eine reine Decoder Architekutur setzt. Daraus entsprange viele unterschiedliche Decoder-Only Modelle von Google, Meta oder anderen Forschungseinrichtungen. Die aktuellsten versionen von Llama, GPT-4 oder Bard stellen dabei bis heute die neusten Entwicklungen dar.

B. Aufbau der GPT-Decoder Architektur

Wie im vorherigen Abschnit beschrieben legt das GPT Modell den Grundstein für die Decoder-only Architektur Entwicklungen. Dies wurde durch das Paper [3] beschrieben und basiert dabei auf den Grundzügen der Transformer Architekturen welche bereits in Paper [1] beschrieben wurden. Die

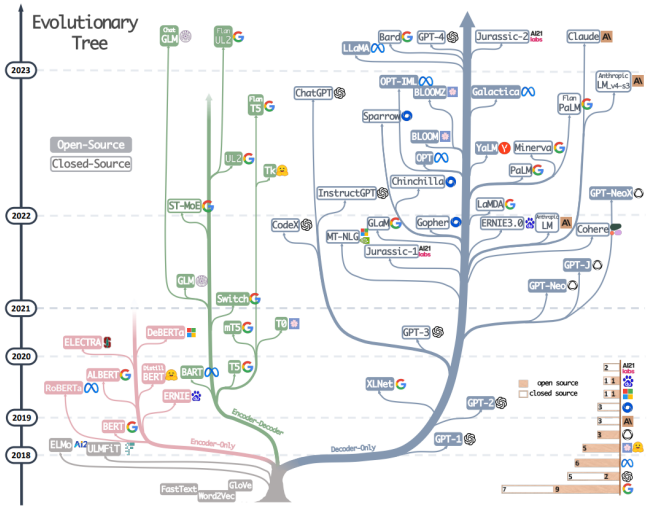

Fig. 1. Historische Entwicklung von Transformer Modellen aus [2]

folgende Abbildung illustriert dabei den Aufbau einer Decoder Schicht des GPT Modells. Die Figure 2 zeigt dabei die

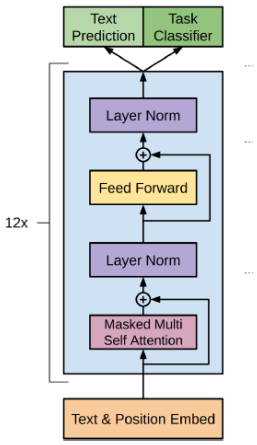

Fig. 2. Decoder Architektur nach [3]

Architektur Elemente einer Decoder Schicht. Diese Schichten werden in der ersten GPT Architektur 12 mal hintereinander geschaltet. Jedes dieser Decoder Schichten besteht dabei aus einem Masked MUlti-Self Attention Layer, Layer Normalisation, Feed Forwar Netzwerk und einem weiteren Layer Normalisation. Die genaue Funktionsweise dieser Schichten wird in [1] beschrieben. Abschließend wird in der GPT Architektur ein Layer je nach Aufgabe angehängt, welche in Figure 2 durch Text Vorhersage oder Aufgaben Klassifizierung repräsentiert wird.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," Aug. 2023.

[2] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, B. Yin, and X. Hu, "Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond."

[3] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training."