Woche 14 Theorie: Generative Modelle, Interview

Skript

Erarbeitet von
Joana Grah und Renato Vukovic

Lernziele

- Das grobe Konzept von LLM nachvollziehen können.
- Ein Gefühl für die Komplexität der aktuellen Debatten erhalten.

Inhalt

Generierung von Text

Widmen wir uns jetzt der Generierung von Texten. Dazu habe ich einen Experten hier – Renato Vukovic ist Doktorand am Institut für Informatik an der HHU, in der Arbeitsgruppe für Dialog Systems und Machine Learning.

Hi, Renato!

Hi, vielen Dank für die Einladung!

Danke, dass du hier bist!

1. Seit wann können KIs denn eigentlich Texte generieren, und welche Modelle waren dort am erfolgreichsten?

Ich denke, das Ziel Text zu generieren ist eines der ersten für KI, wie man z. B. auch am Turing Test sehen kann, aber gleichzeitig ist es auch eines der schwierigsten, weil man sehen kann, dass es immer noch nicht gelöst ist.

Angefangen hat es damals mit regelbasierten Chatbots wie ELIZA, die nach und nach von den statistischen Machine-Learning-Methoden abgelöst wurden. Heutzutage wird in der Regel versucht, einem Modell auf Grundlage der Verteilung von Sprache beizubringen, wie sie funktioniert, indem das Model lernt, das nächste Wort in einer Sequenz vorherzusagen.

Diese Idee ist essenziell für die sogenannten Large Language Models, das sind große neuronale Netze, die auf großen Mengen Text mit diesem simplen Ziel trainiert werden: das nächste Wort im Text vorherzusagen.
Diese Modelle haben das Natural Language Processing, wie dieses Feld heißt, revolutioniert, indem sie heutzutage im Prinzip das Rückgrat zu so ziemlich jeder Natural Language Processing Task bilden können: für Übersetzung, Zusammenfassung, Klassifizierung, Generierung von Text, aber auch, was für mich am interessantesten ist, Konversation.

2. Kannst du die Funktionsweise der Modelle mal in ein paar Sätzen grob erklären?

Ja, das Training solcher Modelle verläuft in der Regel in zwei Schritten, die Unsupervised Pre-Training und Fine-Tuning heißen.

Quellen [11-15]

Beim ersten Schritt ist das Ziel des Trainings, dass das Modell in der Lage ist, das nächste Wort in einem Satz, mit den Wörtern davor als Input, richtig vorherzusagen, z. B. im Satz „KI für alle ist sehr ...“ müsste dann ein Adjektiv ergänzt werden, wie z. B. „interessant“. Dadurch, dass das Model während des ersten Trainingsschritts eine große Bandbreite an Text sieht und sehr viele Parameter hat, lernt es eine ziemlich allgemeingültige Repräsentation von Sprache, die im Prinzip auf der Wahrscheinlichkeitsverteilung von Wörtern beruht. Diese Art des Trainings wird konkret Language Modelling genannt, da das Modell lernt Sprache in den Parametern darzustellen. In den Embeddings von Wörtern sind dann also Informationen, die auf dem Kontext beruhen, in dem diese Wörter vorkommen. Durch die schiere Größe haben solche Modelle auch die Kapazität im Training zu lernen, Aufgaben zu verstehen und diese entsprechend zu erfüllen, um bessere Vorhersagen über die kommenden Wörter zu treffen.
Der erste Schritt ist unsupervised, da im Text jederzeit das nächste Wort vorliegt, das dann im Training vorhergesagt werden muss. Das heißt, man muss dafür keine Daten annotieren, weil die Labels – also das nächste Wort – dafür schon implizit in den Daten vorliegen. Da diese Modelle allerdings nach dem ersten Schritt theoretisch jede Art von Wort und alles was

in den Trainingsdaten vorhanden ist einfach reproduzieren könnten, werden sie im zweiten Schritt noch darauf trainiert, gewissen Regeln zu folgen oder beim Berechnen der Antwort besser den Instruktionen des Users oder der Anfrage zu folgen.
Dies passiert, indem das Model zum einen auf von Menschen geschriebenen Anfrage-Antwort-Paaren trainiert wird und darüber hinaus noch auf menschlichem Feedback darauf, welche Art von Antwort von den Menschen präferiert wird und auch, ob gewisse Regeln der Konversation eingehalten werden, wie z. B. dass man keine Finanztipps geben soll oder nicht aggressiv sein soll. Diese Fine-Tuning-Methode heißt Reinforcement Learning from Human Feedback. Bei diesem Reinforcement Learning geht es darum, einen Reward zu maximieren, indem man entsprechende Antworten als Actions liefert. In diesem Fall besteht der Reward aus den Präferenzen und dem Feedback von Menschen.
Diese Modelle nutzen in der Regel eine sogenannte Transformer-Architektur, also eine Art neuronales Netz, das hauptsächlich einen sogenannten Attention-Mechanismus nutzt, mit dem man die grammatikalischen Verbindungen in Sprache modellieren kann, z. B. so kann man die Verbindung zwischen Subjekt und Objekt in einem Satz gut von dem Modell lernen lassen.

Die stärksten Modelle wie ChatGPT haben sehr viele Parameter, weshalb sowohl Training, als auch Inferenz sehr kosten- und zeitintensiv sind. So haben diese Modelle mehr als 100 Milliarden Parameter, die im Training angepasst werden müssen, und das Pre-Training dauert bis zu mehreren Wochen, und das selbst auf 1000 der besten TPUs – das sind GPUs, die im Prinzip extra für neuronale Netze konzipiert wurden. Und, was man auch noch sich vorstellen muss, ist, dass in diesem Training diese Modelle Billionen von Wörtern sehen, und das sind mehr Textdaten als Menschen in ihrem ganzen Leben sehen.

3. Einer der ersten Chatbots sozusagen ist ja ELIZA gewesen, also ein Modell, das quasi ein Gespräch mit einer Psychotherapeut*in simuliert hat, und es hat zwar nicht den Turing-Test bestanden, wurde aber schon als sehr „menschlich“ empfunden. Jetzt, Jahrzehnte später, wird ChatGPT gehypt und solche Large Language Models sind

super erfolgreich. Für wie gefährlich hältst du eigentlich die Vermenschlichung dieser KI-Systeme und wie weit, glaubst du, sind sie jetzt noch von sogenannter AGI, also Artificial General Intelligence, entfernt?

Ja also wie mit jeder Technologie kommen mit Chancen natürlich auch große Verantwortungen und alles, was für was Gutes genutzt werden kann, kann natürlich auch missbraucht werden. Obwohl es sehr aufregend ist, dass es heutzutage schon etwas wie ChatGPT gibt, bringt es natürlich auch Probleme und Gefahren mit sich. ChatGPT ist z. B. nur so gut wie die Daten, auf denen es trainiert wurde, das heißt, wenn die Daten falsche Aussagen oder Bias – z. B. bezüglich Gender – enthalten, dann ist es wahrscheinlich, dass das Model diese einfach reproduziert. Darüber hinaus lernt es z. B. nicht, wie man richtig mit privaten Daten umgeht, d. h. man muss entsprechend sehr vorsichtig sein, welche Daten oder Informationen man mit so einem Modell teilt. Abgesehen davon ist das Verhalten des Models nicht wirklich erklärbar, also man kann nicht überprüfen, wie das Model jetzt zu einer speziellen Antwort kam. Weiterhin, kann ChatGPT nicht erklären, welche die Quellen für sein Wissen sind, da die ja irgendwo in den Parametern sind. All diese Probleme sind natürlich Bestandteil aktiver Forschungsbereiche und da bin ich gespannt, wie die Lösungen aussehen werden.

Während es natürlich sehr wichtig ist, diese Probleme technologisch zu lösen, ist es mindestens genauso wichtig, den Nutzern, also allen Menschen, beizubringen, damit verantwortungsbewusst umzugehen und entsprechende Standards zu setzen.

Was Artificial General Intelligence angeht, denke ich, dass wir noch weit davon entfernt sind sie zu erreichen. Natürlich ist ChatGPT sehr beeindruckend und man kann durchaus sagen, dass es gewissermaßen übermenschliche Fähigkeiten besitzt, allerdings löst es nicht alle Aufgaben, für die Intelligenzen notwendig sind. Man sollte dabei auch nicht vergessen, wie viel Computerleistung nötig ist, um ChatGPT zu trainieren und noch stärkere Modelle werden noch mehr Leistung benötigen. Das kann so viel werden, dass es nicht klar ist, ob wir die benötigte Leistung überhaupt zur Verfügung stellen können, also entsprechend, ob wir dann noch generellere Modelle trainieren können. Dabei ist auch wichtig zu betonen, dass das ganze Wissen von ChatGPT von Demonstrationen menschlicher Intelligenz kommt, also menschengemachtem Text oder menschlichem Feedback.

4. Glaubst du, dass KIs wie ChatGPT bald auch für uns Aufsätze, Hausarbeiten oder vielleicht sogar Bücher schreiben?

Ja also ich kann mir das durchaus vorstellen. Ich denke, es hat einige Vorteile, denn so ein Model macht es einfacher denn je eigene Ideen niederschreiben, und so werden Informationen oder Ideen natürlich auch zugänglicher, wenn sie in gut leserlicher Form vorhanden sind. Am Ende denke ich, dass man ChatGPT nicht wirklich verbieten kann. Natürlich sollte man lernen, selbst Texte zu schreiben, aber die Lehre wird sich den neuen Technologien und den neuen Begebenheiten, die sie mit sich bringen, anpassen müssen. Entsprechend muss man den Leuten beibringen, wie man ChatGPT richtig nutzt, anstatt es zu verteufeln.

Man stelle sich vor, man würde heutzutage Bücher als Informationsquellen verbieten, oder Computer oder gar das Internet. Das ist alles natürlich unvorstellbar. Aber auch wie das damals bei der Einführung solcher Technologien war, ist es natürlich jetzt auch bei ChatGPT. Und deshalb denke ich, statt ChatGPT zu verbieten, sollten wir uns überlegen, wie wir es auf bestmögliche und auch noch verantwortliche Weise nutzen lernen, und so auch natürlich in der Lehre.

Ja, vielen herzlichen Dank!

Ja, danke auch!

Quellen

Quelle [11] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Quelle [12] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Quelle [13] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Quelle [14] Glaese, A., McAleese, N., Trębacz, M., Aslanides, J., Firoiu, V., Ewalds, T., ... & Irving, G. (2022). Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.

Quelle [15] Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H. T., ... & Le, Q. (2022). Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Weiterführendes Material

Elements of AI. Neural networks - Advanced neural network techniques.
https://course.elementsofai.com/5/3

Playing around with GANs.
- https://www.whichfaceisreal.com
- https://thispersondoesnotexist.com
- https://thisartworkdoesnotexist.com
- https://thiscatdoesnotexist.com
- https://thishorsedoesnotexist.com
- https://thischemicaldoesnotexist.com

Blogs von OpenAI, DeepMind und Google AI (als Auszug) um up to date zu bleiben.
- https://openai.com/blog/
- https://www.deepmind.com/blog
- https://ai.googleblog.com/

Blogposts von OpenAI (rückwärts chronologisch).
- ChatGPT https://openai.com/blog/chatgpt/
- InstructGPT https://openai.com/blog/instruction-following/
- GPT-3 https://openai.com/blog/gpt-3-apps/
- GPT-2 https://openai.com/blog/better-language-models/

Blogpost von DeepMind zu Sparrow (ähnliches Modell wie ChatGPT).
- https://www.deepmind.com/blog/building-safer-dialogue-agents

Blogpost von Google zu Bard AI (das auf LaMDA beruht).
- https://blog.google/technology/ai/bard-google-ai-search-updates/
- https://blog.google/technology/ai/lamda/

Interessantes Paper zum Carbon Footprint solcher Modelle.
- Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L. M., Rothchild, D., ... & Dean, J. (2021). Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.

Training eines Language Models auf einer einzigen GPU an einem Tag.
- Geiping, J., & Goldstein, T. (2022). Cramming: Training a Language Model on a Single GPU in One Day. *arXiv preprint arXiv:2212.14034*.

Disclaimer

Transkript zu dem Video „Woche 14 Theorie: Generative Modelle“, Joana Grah und Renato Vukovic. Dieses Transkript wurde im Rahmen des Projekts ai4all des Heine Center for Artificial Intelligence and Data Science (HeiCAD) an der Heinrich-Heine-Universität Düsseldorf unter der Creative Commons Lizenz CC-BY 4.0 veröffentlicht. Ausgenommen von der Lizenz sind die verwendeten Logos, alle in den Quellen ausgewiesenen Fremdmaterialien sowie alle als Quellen gekennzeichneten Elemente.