Woche 12 Theorie: Textverarbeitung

Skript

Erarbeitet von
Marc Feger

Lernziele

- Die Grundlagen von Transformermodellen erklären können
- BERT als Beispiel für ein Transformermodell nennen können
- Den Unterschied zwischen einfachen Textrepräsentationen wie dem Worttaschenmodell und fortgeschrittenen Textrepräsentationen wie Transformern erklären können

Inhalt

Einstieg

Stell dir vor, du sitzt in einem Café und hörst die Gespräche der Menschen um dich herum.

Projekt „KI für alle"
HeiCAD – Heine Center for Artificial Intelligence and Data Science

Du hörst heraus, dass das Paar an deinem Nachbartisch bald in den Urlaub nach Bangkok fliegt und du erinnerst dich an deine letzte Reise nach Thailand.

Vielleicht hörst du einige Worte auf Französisch von einem Paar am Nachbartisch.

Obwohl du die Sprache selbst nicht sprichst, kannst du dennoch eine Menge darüber erfahren, worüber die Menschen sprechen und wie sie sich dabei fühlen.

Das zeigt, dass unsere Fähigkeit, Sprache zu verstehen, weit über das bloße Verständnis von Wörtern hinausgeht.

Wir können ihren Kontext nutzen, um ihre grundlegende Bedeutung, die dahinter liegenden Emotionen und Stimmungen zu erfassen, selbst wenn wir die Sprache selbst nicht sprechen können.

Denn für uns scheint es ganz selbstverständlich, all diese Informationen in Echtzeit zu verarbeiten und verstehen zu können.

Jetzt stell dir vor, du willst versuchen, einem Computer beizubringen, das Gleiche zu tun.

Genau das ist das Ziel der Sprachverarbeitung oder Natural Language Processing, einem Feld der Künstlichen Intelligenz, das sich mit dem Verarbeiten von Sprache beschäftigt.

Ein zentrales Konzept dabei ist die Textrepräsentation, also die Art und Weise, wie wir Texte in eine Form bringen, um von Maschinen verstanden werden zu können.

Einfache Text-Repräsentation

Eine sehr einfache Methode, die zur Text-Repräsentation eingesetzt wird, basiert auf der sogenannten „Bag-of-Words“-Darstellung.

Dabei wird ein Vokabular aus allen Wörtern in einem Korpus erstellt.

Einblendung Grafik Dokumente und Wörter

Stell dir vor, man hat ein Büro voller Dokumente, aber jedes Dokument wird auf eine Liste von einzigartigen Wörtern reduziert.

Um jetzt die Dokumente darzustellen, wird jedes Dokument in seine einzelnen Wörter zerlegt und ein Vektor erstellt, der so groß ist wie das gesamte Vokabular.

Anschließend wird für jedes Wort in einem Vektor eine 1 oder 0 eingetragen, abhängig davon, ob das Wort im Dokument vorkommt oder nicht.

Diese Methode ist wie ein Mensch, der zwar sehen kann, dass ein Wort in einem Text vorkommt, aber keine Ahnung hat, was es bedeutet und wie es mit den anderen Worten zusammenhängt.

Dabei erfasst diese Darstellung keine Beziehungen zwischen den Wörtern, wie sie in der Sprache tatsächlich existieren.

Zum anderen sind die Vektoren für jedes Wort in einem solchen One-Hot-Encoding extrem spärlich, da fast alle Einträge Nullen sind.

Stell dir vor, das Vokabular hat eintausend Wörter und das Dokument, das du betrachtest, nur 10 oder weniger.

Der Vektor, um diesen kurzen Text darzustellen, würde trotzdem eintausend Einträge haben, wovon lediglich ein Prozent der Einträge eine 1 haben würden.

Daher ist diese Vorgehensweise nicht nur für Speicher und Rechenzeit sehr kostspielig, sondern macht sie auch blind dafür, semantische Ähnlichkeiten oder Unterschiede zwischen Wörtern zu erfassen.

Fortgeschrittene Text-Repräsentation

Bestimmt hast du schon einmal von ChatGPT gehört oder es bereits getestet.

Quelle [1]

Sicherlich warst du davon überrascht, wie gut die Dialoge sind und du hattest bestimmt auch das Gefühl, dass die AI das Geschriebene wirklich versteht!

Der Grund dafür, dass Modelle wie ChatGPT solch überzeugende Sprachkompetenzen erwerben konnten, liegt daran, dass es auf einer Technologie namens Transformer basiert.

Doch einen Schritt zurück:

Ein Transformer ist eine Art neuronales Netzwerk, das in der Lage ist, die Bedeutung von Wörtern und Sätzen zu verstehen, indem es die Beziehungen zwischen ihnen untersucht.

Ein klassisches Beispiel für Transformer in der Text-Repräsentation ist BERT, den man eigentlich aus der Sesamstraße kennt.

Quelle [2]

Aber im Ernst, BERT oder “Bidirectional Encoder Representations from Transformers” hat einen Meilenstein in der Text-Repräsentation gesetzt, indem dieses Sprachmodell erstmals zeigen konnte, wie mächtig Transformer-Modelle tatsächlich sind.

Einblendung Grafik BERT / Transformer

An dieser Stelle könnten wir Stunden darüber sprechen, wie BERT oder Transformer aufgebaut sind, grundlegend kannst du dir diese Transformer-Modelle aber als einen Stapel von sogenannten Encodern vorstellen.

Diese Encoder bestehen aus mehrteiligen neuronalen Netzen, die speziell für die Verarbeitung von Texten entwickelt wurden und in jeder Schicht ein feineres Verständnis für Wörter entwickeln.

Im Gegensatz zur spärlichen Repräsentation, setzen Transformer-Modelle auf dichte Wort-Einbettungen, die Informationen über die Bedeutung der Wörter und ihren Kontext enthalten.

Ein wesentliches Merkmal von Transformer-Modellen ist ihre Verwendung von Attention-Mechanismen, die es ihnen ermöglichen, sich auf bestimmte Teile des Textes zu konzentrieren.

Kurz gesagt: Modelle wie BERT schaffen es, den Kontext von Worten zu erfassen.

Einblendung Grafik Wörter

BERT berücksichtigt den Kontext, indem es den gesamten Satz, in dem ein Wort vorkommt, in Betracht zieht und analysiert, wie jedes Wort mit allen anderen Worten zusammenhängt.

Take-Home Message

Wie du siehst, BERT ist ein absolutes Powerpaket in der Welt der Textverarbeitung!

Mit seiner bahnbrechenden Technologie, die auf den modernsten Methoden der künstlichen Intelligenz basiert, kann es Texte lesen und verstehen.

Letztendlich eröffnet der Einsatz von Transformern wie BERT völlig neue Möglichkeiten in der Textanalyse, Frage-Antwort-Spielen oder Textgenerierung, die ohne so nicht möglich wären.

Quellen

Quelle [1] https://chat.openai.com/chat

Quelle [2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Weiterführendes Material

Towards Data Science.

Introduction to Text Representations for Language Processing – Part 1, Sundaresh Chandran, 12. Juni 2020. https://towardsdatascience.com/introduction-to-text-representations-for-language-processing-part-1-dc6e8068b8a4

An Overview for Text Representations in NLP, jiawei hu, 4. März 2020. https://towardsdatascience.com/an-overview-for-text-representations-in-nlp-311253730af1

Machine Learning – Text Processing, Javaid Nabi, 13. September 2018. https://towardsdatascience.com/machine-learning-text-processing-1d5a2d638958

Machine Learning Mastery.

Archive | Deep Learning for Natural Language Processing. https://machinelearningmastery.com/category/natural-language-processing/

Podcast InsideHeiCAD.

Staffel 2, Folge 3. https://www.heicad.hhu.de/aktivitaeten/der-heicadpodcast

Grah, J. (Moderatorin), Behrendt, M. (Gästin). (2022, 6. April). #3: Wie kann man politische Entscheidungen mit Hilfe von Methoden des Natural Language Processing unterstützen? (PhD Pitches) [Audio-Podcast]. In *InsideHeiCAD*. Heine Center for Artificial Intelligence and Data Science.

Disclaimer

Transkript zu dem Video „Woche 12 Theorie: Textverarbeitung“, Marc Feger.