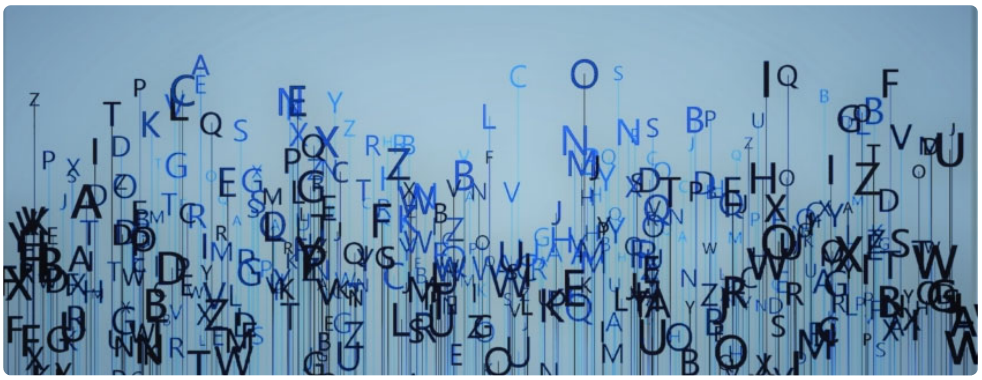
SocietyByte Wissenschaftsmagazin der Berner Fachhochschule

Mit welchen Techniken sich die Arbeitsweise von KI entschlüsseln lässt

Von Alexandre Puttick (BFH Technik & Informatik) | 0 Kommentare



Grosse Durchbrüche in der Künstlichen Intelligenz (KI) beschleunigen die Einführung von KI-gestützten Werkzeugen in Industrie, Forschung und Verwaltung. Der Denkprozess moderner KI-Modelle findet jedoch in einer Blackbox statt. In diesem Artikel werden Techniken untersucht, die entwickelt wurden, um dieses Problem im Bereich der NLP zu entschärfen.

Natürliche Sprachverarbeitung (Natural Language Processing, NLP) bezieht sich auf den Bereich der Techniken, die für die automatische Verarbeitung der menschlichen Sprache in Text- oder Sprachdaten entwickelt wurden. In den letzten Jahren wurde NLP von auf maschinellem Lernen basierenden Methoden dominiert, was in der Einführung grosser Sprachmodelle wie ChatGPT gipfelte. Im weiteren Verlauf dieses Artikels konzentrieren wir uns auf den aktuellen Stand der Forschung im Bereich Explainable AI (XAI) im NLP-Kontext, wobei wir uns auf einige der am weitesten verbreiteten Ansätze konzentrieren. Für einen umfassenderen und detaillierteren Überblick über erklärbare NLP siehe (Danilevsky et al. 2020).

Das Merkmal Wichtigkeit

Feature Importance -Techniken versuchen, die Merkmale in einer Texteingabeprobe zu identifizieren, die am meisten zum endgültigen Ergebnis des Modells beitragen. In der Regel bedeutet dies, dass die wichtigsten Wörter identifiziert werden, aber je nachdem, wie Textproben in mathematische Vektoren kodiert werden, können Verfahren zur Merkmalsbedeutung auch verwendet werden, um wichtige Phrasen oder Sätze zu identifizieren. Diese Methode wird häufig visuell in Form einer Salienzkarte dargestellt, die die wichtigsten Wörter mit einer Intensität hervorhebt, die der Wichtigkeit des jeweiligen Wortes entspricht. Abbildung 1 zeigt ein Beispiel für eine Salienzkarte für eine binäre Klassifizierungsaufgabe. Beliebte Verfahren zur Bestimmung der Wichtigkeit von Merkmalen sind LIME (Ribeiro et al. 2016), SHAP (Lundberg und Lee 2017) und die erstableitende Salienz (Li et al. 2015).

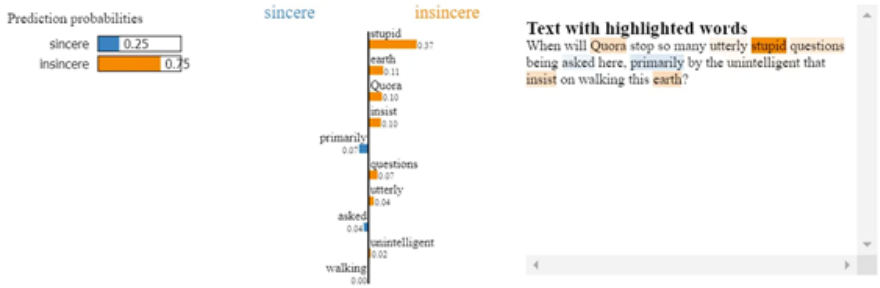


Abbildung 1: Eine Salienzkarte, die unter Verwendung des LIME-Erklärers auf einem Modell erstellt wurde, das trainiert wurde, um auf Quora gestellte Fragen als aufrichtig oder unaufrichtig zu klassifizieren. Blau hervorgehobene Wörter zeigen an, dass die Frage aufrichtig ist, während orange hervorgehobene Wörter das Gegenteil anzeigen. Die undurchsichtige Hervorhebung zeigt die Wörter mit dem grössten Beitrag an(Bildquelle [https://towardsdatascience.com/what-makes-your-question-insincere-in-quora-26ee7658b010]).

Beispielgesteuert

Beispielgesteuerte Interpretierbarkeitstechniken liefern keine expliziten Erklärungen für die Entscheidung des Modells. Stattdessen besteht das Ziel darin, andere Beispiele zu identifizieren, die aus der Sicht des Modells als ähnlich angesehen werden. Dies ermöglicht es einem externen Prüfer, die ähnlichen Stichproben zu überprüfen und festzustellen, welche gemeinsamen Faktoren und Unterschiede wahrscheinlich eine wichtige Rolle gespielt haben. Abbildung 2 zeigt die Ergebnisse der in (Croce et al. 2019) entwickelten beispielbasierten Methoden.

Class	Questions (q_i)	$k(q_1, q_2)$
LOC	*"What is the capital of Ethiopia?"*	0.98
NUM	*"What is the population of Nigeria?"*	
ENTY	*"What was FDR 's dog 's name?"*	0.97
HUM	*"What was J.F.K.'s wife 's name?"*	
ENTY	*"What is the Ohio state bird?"*	0.90
ENTY	*"What is the pH scale?"*	
ENTY	*"What was the first satellite to go into space?"*	0.83
HUM	*"Who was the first American to walk in space?*	
NUM	*"What was the last year that the Chicago Cubs won the World Series?"*	0.73
NUM	*"What is the average speed of the horses at the Kentucky Derby?"*	

Abbildung 2: In diesem Fall bestand die Aufgabe darin, Fragen in die Kategorie zu sortieren, die dem Thema der Frage entspricht (z. B. Ort, Zahl, Entität...). Jedes der obigen Fragenpaare wird vom Modell als ähnlich angesehen, aber die Fragen gehören nicht immer zur selben Klasse.

Generierte Erklärungen

Eine dritte Technik besteht darin, generative Sprachmodelle wie GPT-3 zu trainieren, um natürlichsprachliche Erklärungen für die gegebene Aufgabe zu generieren (z. B. «Generated Text: Der Kandidat ist gut, weil er einen Abschluss von einer führenden Universität hat.»). Um ein Modell zu trainieren, das in der Lage ist, solche Erklärungen zu generieren, ist in der Regel ein ausreichend großer Datensatz erforderlich, der mit von Menschen geschriebenen Erklärungen annotiert ist. Um die generierten Erklärungen mit der Ausgabe zu korrelieren, sollte das Modell gleichzeitig trainiert werden, um die Zielaufgabe (z. B. die Klassifizierung von Bewerbungen) zu erfüllen und eine Erklärung zu generieren, wobei eine kombinierte Verlustfunktion verwendet wird, die sowohl die Modellausgabe als auch die generierten Erklärungen mit Stichproben aus den Trainingsdaten vergleicht. Solche Techniken wurden in (Camburu et al. 2018) erforscht.

Hindernisse und Unzulänglichkeiten

Jeder der drei oben beschriebenen Ansätze ist vielversprechend, weist aber auch einige Mängel auf. Zum Beispiel ist in vielen Fällen nicht klar, wie gut die Erklärungen mit dem tatsächlichen Entscheidungsprozess des Modells übereinstimmen. Die generierten Erklärungen liefern wertvolle Informationen über die Überlegungen der Datenkommentatoren und nicht über die des Modells. Vielleicht gibt es einen Kompromiss zwischen der getreuen Erklärung des Modells und der Erstellung von Erklärungen, die leicht verständlich sind und zur Überprüfung oder Anfechtung unfairer algorithmischer Entscheidungen verwendet werden können. Oftmals ist Letzteres wichtiger.

Es hat sich auch gezeigt, dass verschiedene Techniken oft zu unterschiedlichen, sogar widersprüchlichen Erklärungen führen, ein Beispiel für das so genannte unstimmigkeitsproblem (Krishna et al. 2022). Dies macht die Entwicklung von Metriken erforderlich, die versuchen, die Qualität verschiedener Erklärungstechniken (DeYoung et al. 2020) sowie den Grad der Übereinstimmung zwischen verschiedenen Erklärern zu messen. Idealerweise könnten interpretierbare Modelle anhand ihrer Leistung bei einer Vielzahl von Metriken identifiziert werden, die verschiedene Aspekte der Interpretierbarkeit messen. Letztendlich wird eine Kombination von Ansätzen, wie sie in diesem Artikel behandelt wurden, notwendig sein, um verschiedene Schattierungen des Verständnisses zu erhalten und zufriedenstellend interpretierbare Modelle zu schaffen.

Referenzen

1. Camburu, Oana-Maria, et al. «e-snli: Natural language inference with natural language explanations» Advances in Neural Information Processing Systems 31 (2018).
2. Croce, Danilo, Daniele Rossini, and Roberto Basili. «Auditing deep learning processes through kernel-based explanatory models.» Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019.
3. Danilevsky, Marina, et al. «A survey of the state of explainable AI for natural language processing.» arXiv preprint arXiv:2010.00711 (2020).
4. DeYoung, Jay, et al. «ERASER: A benchmark to evaluate rationalized NLP models.» arXiv preprint arXiv:1911.03429 (2019).
5. Krishna, Satyapriya, et al. «The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective.» arXiv preprint arXiv:2202.01602 (2022).
6. Li, Jiwei, et al. «Visualizing and understanding neural models in nlp.» arXiv preprint arXiv:1506.01066 (2015).
7. Lundberg, Scott M., and Su-In Lee. «A unified approach to interpreting model predictions.» Advances in neural information processing systems 30 (2017).
8. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. «» Why should I trust you?» Explaining the predictions of any classifier.» Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.

AUTOR/AUTORIN: ALEXANDRE PUTTICK

Dr. Alexandre Puttick ist Post-Doktorand in der Forschungsgruppe Angewandte Maschinelle Intelligenz an der Berner Fachhochschule. Seine aktuelle Forschung befasst sich mit der Entwicklung von klinischen Tools für die psychische Gesundheit sowie mit der Erkennung und Abschwächung von Verzerrungen in KI-gesteuerten Rekrutierungs-Tools.

Posts von Alexandre Puttick

Ähnliche Beiträge

Gesellschaftliche Stereotypen in vortrainierten Sprachmodellen

Hi ChatGPT, hast du Vorurteile?

Inklusive und vielfältige Sprache? - Ein Sprachmodell zeigt, wie es geht

Wenn Mehmet und Peter nicht gleich sind - Vorurteile auf Grund der Namensherkunft in Wortvektoren

«Wir müssen Bias aus Sprachmodellen entfernen» - eine Podcastfolge über KI in Verwaltung und Justiz

0

KOMMENTARE