NeuroX:

إطار لتفسير نماذج معالجة اللغات الطبيعية العميقة

يونيو 2023

المحتوى

الملخص

يعد تفسير نماذج معالجة اللغات الطبيعية العميقة (Deep NLP) أمرًا بالغ الأهمية لعدة أسباب. أولاً، النماذج العصبية العميقة، مثل الشبكات العصبية ، معقدة للغاية وتعمل كصناديق سوداء، مما يجعل من الصعب فهم كيفية وصولها إلى اتخاذ قراراتها. من خلال تفسير هذه النماذج، نكتسب رؤى حول طريقة عملها الداخلية، مما يسمح لنا بالكشف عن الأسباب والعوامل التي تؤثر على تنبؤات هذه النماذج. هذه الشفافية ضرورية لبناء الثقة وضمان الاستخدام المسؤول والأخلاقي لأنظمة معالجة اللغات الطبيعية، كما تمكّن القابلية التفسيرية الباحثين والمطورين من تشخيص ومعالجة التحيزات أو الأخطاء أو القيود في هذه النماذج، مما يؤدي إلى تحسين الأداء وتفادي التحيز وبلوغ الإنصاف. علاوة على ذلك، يساعدنا تفسير معالجة اللغات الطبيعية العميقة على فهم حدودها ومخاطرها المحتملة، مما يمكننا من تحديد مجالات التحسين واتخاذ قرارات مناسبة بشأن تطبيقاتها واستعمالاتها. وهذا أمر حيوي لتعزيز الشفافية والإنصاف والمساءلة والتقدم في هذا المجال، مما يؤدي في النهاية إلى تطبيقات ذكاء اصطناعي أكثر شفافية وجديرة بالثقة.

NeuroX هو إطار عمل يهدف إلى تفسير نماذج معالجة اللغات الطبيعية العميقة وزيادة شفافية طريقة عملها الداخلية وكيفية بناء واستخلاص التوقعات. الهدف من إطار العمل والمنهجيات المقترحة هو تجاوز المدخلات والغوص أعمق لتفسير وتقديم تفسيرات أكثر ثراءً لنموذج معين وتوقعاته. وهذا يشمل عدة نواحي، بما في ذلك Neuron Probing الذي يسلط الضوء على مكونات الشبكة العصبية (الطبقات ، ورؤوس الانتباه ، والخلايا العصبية) للشبكة التي تتعلم مفاهيم محددة واكتشاف المفاهيم الكامنة الذي تمثل المفاهيم التي تم تعلمها في التمثيلات المكتسبة.

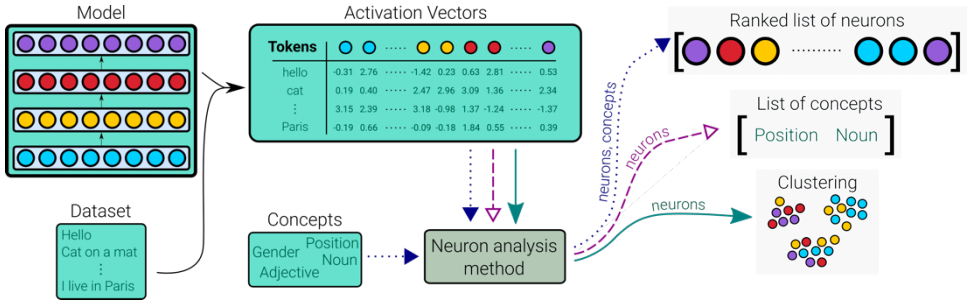
يتم تحقيق تأثير NeuroX من خلال:

- مبادرات دولية تضم شركاء مثل Facebook و MIT-CSAIL وجامعة هارفارد ومعهد ستيفنز للتكنولوجيا وجامعة Dalhausie وجامعة Heinrich Heine Düsseldorf.
- 25 منشورًا عالي الجودة في مجلات ومؤتمرات دولية رفيعة المستوى.
- تنظيم ورش عمل ودروس دولية.
- مجموعات الأدوات
- التغطية الإعلامية: مطبعة معهد ماساتشوستس للتكنولوجيا ، مدونة العلوم (مطبعة معهد ماساتشوستس للتكنولوجيا ، اتصالات إيه سي إم Science Blogs)

عرض نظام NeuroX

أنتجت المجموعة البحثية عددًا من الأنظمة التوضيحية لقدرات NeuroX:

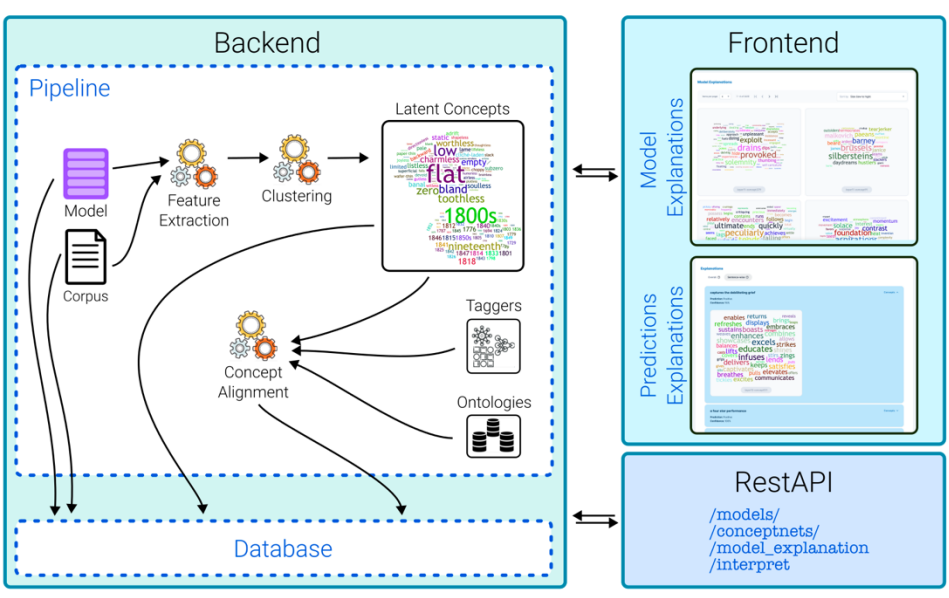
- مجموعة أدوات NeuroX لتحليل الخلايا العصبية الفردية في نماذج معالجة اللغات الطبيعية العميقة
- مكتبة Python تضم طرقًا مختلفة لتفسير وتحليل الخلايا العصبية، موجهة نحو نماذج Deep NLP. المكتبة توفر إمكانات لاستخراج تنشيط الخلايا العصبية، تدريب المسبار، تحليل المجموعات، اختيار الخلايا العصبية إضافة إلى المزيد من القدرات المتوفرة للمستخدمين.



الشكل 1. مجموعة أدوات NeuroX لتحليل الخلايا العصبية الفردية في نماذج معالجة اللغات الطبيعية العميقة

NxPlain: شرح التنبؤات في نماذج معالجة اللغات الطبيعية العميقة

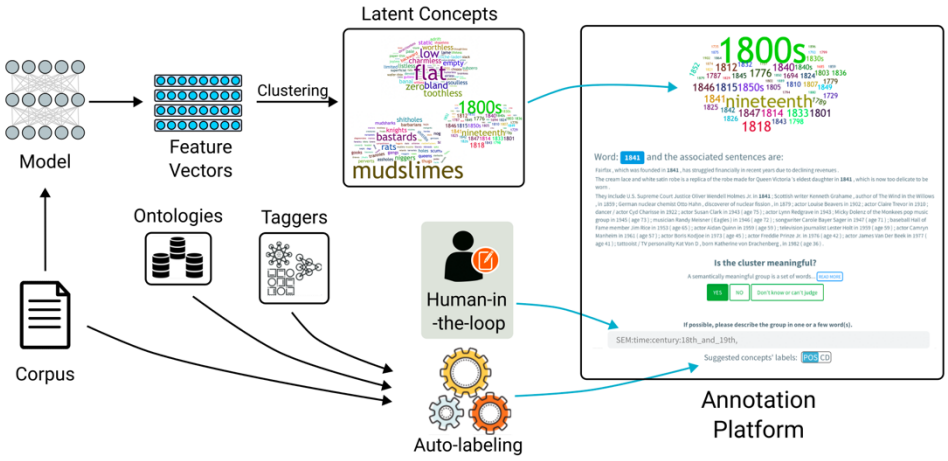
1. NxPlain هو تطبيق ويب يقدم شرحًا لتنبؤ النموذج باستخدام المفاهيم الكامنة. يساعد NxPlain في كشف المفاهيم الكامنة التي تم تعلمها في نموذج معالجة اللغات الطبيعية العميقة، ويوفر تفسيرًا للمعرفة المكتسبة في النموذج، ويشرح تنبؤات النموذج بناءً على المفاهيم المستخدمة. يسمح التطبيق للمستخدمين بتصفح المفاهيم الكامنة بطريقة مبسطة، مما يتيح للمستخدمين بكفاءة معرفة المفاهيم الأكثر بروزًا من خلال عرض على مستوى المجموعة الشاملة وعرض محلي على مستوى الجملة. تُعد أداة NxPlain تطبيقًا مفيدًا لتصحيح الأخطاء، وكشف تحيز النموذج، ولإبراز الارتباطات الزائفة في النموذج.



الشكل 2. NxPlain: شرح التنبؤات في نماذج معالجة اللغات الطبيعية العميقة

ConceptX: إطار عمل: تحليل المفاهيم المكتسبة في نماذج معالجة اللغات الطبيعية العميقة

ConceptX هو إطار عمل يستخدم الإنسان في حلقة تفسير الفضاء التمثيلي الكامن والتعليق عليه في نماذج اللغة المدربة مسبقًا (pLMs). نحن نستخدم طريقة غير خاضعة للإشراف لاكتشاف المفاهيم التي تم تعلمها في هذه النماذج وتمكين المستخدمين لتوليد تفسيرات للمفاهيم من خلال واجهة رسومية.



الشكل 3. ConceptX: إطار عمل: تحليل المفاهيم المكتسبة في نماذج معالجة اللغات الطبيعية العميقة

عمل ريادي في تحليل التمثيل

- إطار بناء مسابر للتحقق

التغطية الإعلامية: مطبعة معهد ماساتشوستس للتكنولوجيا، مدونة العلوم، وغيرها

- تحليل الخلايا العصبية

التغطية الإعلامية: مطبعة معهد ماساتشوستس للتكنولوجيا ، اتصالات إيه سي إم، وغيرها

- تحليل المفاهيم الكامنة

تم إيداع براءة اختراع عن المشروع.

الأثر العلمي

مجلات محكمة:

- Hassan Sajjad, Fahim Dalvi, Nadir Durrani, Preslav Nakov (2022). On the Effect of Dropping Layers of Pre-trained Transformer Models. Computer Speech & Language.
- Hassan Sajjad, Nadir Durrani, Fahim Dalvi (2022). Neuron-level Interpretation of Deep NLP Models: A Survey. Transactions of the Association for Computational Linguistics.
- Prakhar Ganesh, Yao Chen, Xin Lou, Mohammad Ali Khan, Yin Yang, Deming Chen, Marianne Winslett, Hassan Sajjad, Preslav Nakov (2021). Compressing Large-Scale Transformer-Based

Models: A Case Study on BERT. Transactions of the Association for Computational Linguistics.

- Yonatan Belinkov, Nadir Durrani, Hassan Sajjad, Fahim Dalvi, James Glass (2020). On the Linguistic Representational Power of Neural Machine Translation Models. Computational Linguistics.

مؤتمرات عالمية:

- Nadir Durrani, Hassan Sajjad, Fahim Dalvi, Firoj Alam (2022). On the Transformation of Latent Space in Fine-Tuned NLP Models. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).
- David Arps, Younes Samih, Laura Kallmeyer, Hassan Sajjad (2022). Probing for Constituency Structure in Neural Language Models. Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).
- Hassan Sajjad, Firoj Alam, Fahim Dalvi, Nadir Durrani (2022). Effect of Post-processing on Contextualized Word Representations. Proceedings of the 29th International Conference on Computational Linguistics (COLING).
- Hassan Sajjad, Nadir Durrani, Fahim Dalvi, Firoj Alam, Abdul Khan, Jia Xu (2022). Analyzing Encoded Concepts in Transformer Language Models. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT).
- Fahim Dalvi, Abdul Rafae Khan, Firoj Alam, Nadir Durrani, Jia Xu, Hassan Sajjad (2022). Discovering Latent Concepts Learned in BERT. International Conference on Learning Representations (ICLR).
- Nadir Durrani, Hassan Sajjad, Fahim Dalvi (2021). How transfer learning impacts linguistic knowledge in deep NLP models? Findings of the Association for Computational Linguistics (ACL-IJCNLP)).
- Esther Seyffarth, Younes Samih, Laura Kallmeyer, Hassan Sajjad (2021). Implicit Representations of Event Properties within Contextual Language Models:Searching for ``Causativity Neurons''. International Conference on Computational Semantics (IWCS).
- Fahim Dalvi, Hassan Sajjad, Nadir Durrani, Yonatan Belinkov (2020). Analyzing Redundancy in Pretrained Transformer Models. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).
- Nadir Durrani, Hassan Sajjad, Fahim Dalvi, Yonatan Belinkov (2020). Analyzing Individual Neurons in Pre-trained Language Models. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).
- John M. Wu, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, James Glass (2020). Similarity Analysis of Contextual Word Representation Models. Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL).
- Nadir Durrani, Fahim Dalvi, Hassan Sajjad, Yonatan Belinkov, Preslav Nakov (2019). One Size Does Not Fit All: Comparing NMT Representations of Different Granularities. Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL).
- D. Anthony Bau*, Yonatan Belinkov*, Hassan Sajjad, Fahim Dalvi, Nadir Durrani, James Glass (2019). Identifying and Controlling Important Neurons in Neural Machine Translation. International Conference on Learning Representations (ICLR).

- Fahim Dalvi*, Nadir Durrani*, Hassan Sajjad*, Yonatan Belinkov, D. Anthony Bau, James Glass (2019). What is one Grain of Sand in the Desert? Analyzing Individual Neurons in Deep NLP Models. Proceedings of the AAAI Conference on Artificial Intelligence (AAAI).
- Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Stephan Vogel (2017). Understanding and Improving Morphological Learning in the Neural Machine Translation Decoder. Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP).
- Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, James Glass (2017). Evaluating Layers of Representation in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks. Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP).
- Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, James Glass (2017). What do Neural Machine Translation Models Learn about Morphology?. Proceedings of the 55th Conference of the Association for Computational Linguistics (ACL).

ورشات عمل:

- Organizing Workshop of BlackboxNLP 2020-2021
- Ahmed Abdelali, Nadir Durrani, Fahim Dalvi, Hassan Sajjad (2022). Post-hoc analysis of Arabic transformer models. Proceedings of the BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP.

عروض:

- Fahim Dalvi, Hassan Sajjad, and Nadir Durrani (2023). NeuroX Library for Neuron Analysis of Deep NLP Models. Proceedings of the Association for Computational Linguistics (ACL).
- Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Tamim Jaban, Mus'ab Husaini, Ummar Abbas (2023). NxPlain: A Web-based Tool for Discovery of Latent Concepts. Proceedings of the European Chapter of the Association for Computational Linguistics (EACL).
- Firoj Alam, Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Abdul Rafae Khan, Jia Xu (2023). ConceptX: A Framework for Latent Concept Analysis. In Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI)
- Fahim Dalvi, Avery Nortonsmith, D. Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, James Glass (2019). NeuroX: A Toolkit for Analyzing Individual Neurons in Neural Networks. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI). Honolulu, USA, Jan

دورس:

- Hassan Sajjad, Narine Kokhlikyan, Fahim Dalvi, Nadir Durrani (2021). Fine-grained Interpretationand Causation Analysis in Deep NLP Models. North American Chapter of the Association of Computational Linguistics: Human Language Technologies (NAACL-HLT).