Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики

Интерпретируемый ML

Александров Валентин Валерьевич

Москва, 2021

Содержание

Введение

Современные модели машинного обучения показывают хорошие результаты в своих предсказательных способностях, что их примение в реальных задачах стало обыденным явлением. Однако этого удалось добиться за счет повышения сложности моделей. Это привело к тому, что современные модели машинного обучения являются **черными ящиками** - системами, чье внутреннее устройство и процесс получения ответа являются исключительно трудными для анализа. В противовес прозрачным "простым"моделям вроде линейной регрессии и решающим деревьям, мы зачастую не можем достоверно достоверно ответить на вопрос "как работает эта система"в случае современных нейронных сетей.

Существуют опасения относительно надежности и оправданности предсказаний, которые дают современные ML-системы. Это особенно критичная проблема для областей, в которых принимаются решения с очень большими ставками: здравоохранение, правосудие, финансы и т. д. Для этих областей, существуют повышенные требования к безопасности, этичности и ответственности принимаемых решений. Уже существует обильное множество примеров реальных нейросетевых моделей, которые показывают нежелательные поведения и возникающие с этим этические проблемы. Добавить статей с примерами

Проблема в удовлетворении к возросшим требованиям к безопасности систем машинного обучения в том, что мы не можем доподлинно покрыть юнит-тестами все возможные поведения модели, как это возможно в сфере разработки программного обеспечения. Чтобы убедиться в надежности и безопасности модели вместо этого приходится полагаться на интерпретируемость. Если модель интерпретируема для нас, то мы опосредованно можем судить о её надежноссти.

К сожалению, общепринятого формального определения интерпретации не существует. В контексте ML систем в

A fundamental motivation for interpretability is **accountability**. For example, if a predictive mistake happens which caused harm, it's important to explain why this mistake happened

1 Метрики интерпретируемости

Из-за того, что для понятий "объяснения"и "интерпретации"не существует формального определения, встает очень большая проблема как оценивать качество объяснений, предложенных тем или иным методом, и как эти методы сравнивать между собой. В

— — — — — — Бинарный выбор, при котором человек должен из пары представленных объяснений выбрать более качественное

Симуляция инференса. Человек при заданном объяснении и входном наблюдении должен корректно "просимулировать"модель и угадать её предсказание

Симуляция контрфактов. Человек при заданом объясненнии, входе и выходе модели должен ответить, как следует поменять вход так, чтобы изменилось предсказание модели.

Functually-grounded оценивание. Выбирается простая вычислимая прокси-метрика, по которой мы опосредованно судим об интерпретируемости модели. Так как её можно считать автоматически, без участия ручного труда человека, её очень удобно использовать для первичного оценивания методик. Её недостаток в том, что зачастую очень сложно выбрать работающую прокси-метрику. Для каждой выбранной прокси-метрики требуется веское обоснование, почему это можно использовать. Пример прокси-метрики - разреженность эмбеддингов, если ранее было показано, что разреженные модели лучше интерпретируются

Заключение

Проблема интерпретируемости для современных моделей машинного обучения является актуальной проблемой и вызывает интерес как со стороны исследователей, так и со стороны регулирующих органов.

Список использованных источников

1. *Madsen, Andreas.* Post-hoc Interpretability for Neural NLP: A Survey / Andreas Madsen, Siva Reddy, Sarath Chandar // *arXiv:2108.04840 [cs].* — 2021. — . — arXiv: 2108.04840. http://arxiv.org/abs/2108.04840.

2. *Lipton, Zachary C.* The mythos of model interpretability / Zachary C. Lipton // *Communications of the ACM.* — 2018. — . — Vol. 61, no. 10. — Pp. 36–43. https://doi.org/10.1145/3233231.

3. Concept2vec: Metrics for Evaluating Quality of Embeddings for Ontological Concepts / Faisal Alshargi, Saeedeh Shekarpour, Tommaso Soru, Amit Sheth // *arXiv:1803.04488 [cs].* — 2020. — . — arXiv: 1803.04488. http://arxiv.org/abs/1803.04488.

4. *Alatrish, Emhimed.* Building ontologies for different natural languages / Emhimed Alatrish, Dusan Tosic, Nikola Milenkovic // *Computer Science and Information Systems.* — 2014. — Vol. 11, no. 2. — Pp. 623–644. http://www.doiserbia.nb.rs/Article.aspx?ID=1820-02141400023A.

5. DBpedia: A Nucleus for a Web of Open Data / Sören Auer, Christian Bizer, Georgi Kobilarov at al. // The Semantic Web / edited byDavid Hutchison, Takeo Kanade, Josef Kittler at al. — Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. — Vol. 4825. — Pp. 722–735. — Series Title: Lecture Notes in Computer Science. http://link.springer.com/10.1007/978-3-540-76298-0_52.

6. *Doshi-Velez, Finale.* Towards A Rigorous Science of Interpretable Machine Learning / Finale Doshi-Velez, Been Kim // *arXiv:1702.08608 [cs, stat].* — 2017. — . — arXiv: 1702.08608. http://arxiv.org/abs/1702.08608.

7. Ontology-based Interpretable Machine Learning for Textual Data / Phung Lai, NhatHai Phan, Han Hu et al. // *arXiv:2004.00204 [cs, stat].* — 2020. — . — arXiv: 2004.00204. http://arxiv.org/abs/2004.00204.

8. Improving Disentangled Text Representation Learning with Information-Theoretic Guidance / Pengyu Cheng, Martin Renqiang Min, Dinghan Shen et al. // *arXiv:2006.00693 [cs, stat].* — 2020. — . — arXiv: 2006.00693. http://arxiv.org/abs/2006.00693.

9. *Zhang, Xiongyi.* Disentangling Representations of Text by Masking Transformers / Xiongyi Zhang, Jan-Willem van de Meent, Byron C. Wallace // *arXiv:2104.07155 [cs].* — 2021. — . — arXiv: 2104.07155. http://arxiv.org/abs/2104.07155.

10. An Interpretability Illusion for BERT / Tolga Bolukbasi, Adam Pearce, Ann Yuan et al. // *arXiv:2104.07143 [cs].* — 2021. — . — arXiv: 2104.07143. http://arxiv.org/abs/2104.07143.

11. CEVO: Comprehensive EVent Ontology Enhancing Cognitive Annotation / Saeedeh Shekarpour, Faisal Alshargi, Valerie Shalin et al. // *arXiv:1701.05625 [cs].* — 2018. — . — arXiv: 1701.05625 version: 2. http://arxiv.org/abs/1701.05625.

12. *Paccanaro, Alberto.* Learning distributed representations of concepts from relational data using linear relational embedd / Alberto Paccanaro, G. Hinton // *IEEE Transactions on Knowledge and Data Engineering - TKDE.* — 2000. — .

13. *HE, Shilin.* Interpretable NLP. — 2021. — . — original-date: 2019-07-11T01:35:34Z. https://github.com/ShilinHe/interpretableNLP.

14. Learning to Deceive with Attention-Based Explanations / Danish Pruthi, Mansi Gupta, Bhuwan Dhingra et al. // *arXiv:1909.07913 [cs].* — 2020. — . — arXiv: 1909.07913. http://arxiv.org/abs/1909.07913.

15. DeBERTa: Decoding-enhanced BERT with Disentangled Attention / Pengcheng He, Xiaodong Liu, Jianfeng Gao, Weizhu Chen // *arXiv:2006.03654 [cs].* — 2021. — . — arXiv: 2006.03654. http://arxiv.org/abs/2006.03654.

16. Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models / Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, Daniel S. Weld // *arXiv:2101.00288 [cs].* — 2021. — . — arXiv: 2101.00288. http://arxiv.org/abs/2101.00288.

17. *Molnar, Christoph.* Interpretable Machine Learning / Christoph Molnar. https://christophm.github.io/interpretable-ml-book/index.html.