1 2 9 0

Universidade de Coimbra
Faculdade de Ciências e Tecnologia
Departamento de Engenharia Informática

Human-Centered Artificial Intelligence

Master in Data Science and Engineering

Explainable AI: Other Methods

Hugo Gonçalo Oliveira
hroliv@dei.uc.pt

Example-based Explanations

- Explain the behaviour of machine learning models or the underlying data distribution by **selecting particular instances of the dataset**
- More suitable when instances of the data are representable in a human-understandable way (e.g., images, text)
- We often use examples in our jobs and daily lives [Molnar, 2019]

A physician sees a patient with an unusual cough and a mild fever. Patient's symptoms remind her of another patient she had with similar symptoms. She suspects that her current patient could have the same disease and she takes a blood sample to test for this specific disease.

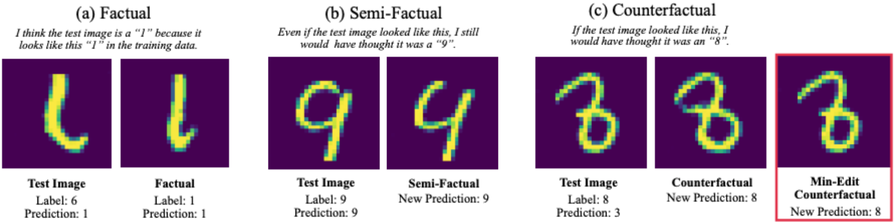
A data scientist works on a new project for one of his clients: Analysis of risk factors that lead to failure of production machines for keyboards. The data scientist remembers a similar project he worked on and reuses parts of the code from the old project because he thinks the client wants the same analysis.

A kitten sits on the window ledge of a burning and uninhabited house. The fire department has already arrived and one of the firefighters ponders for a second whether he can risk going into the building to save the kitten. He remembers similar cases in his life as a firefighter: Old wooden houses that have been burning slowly for some time were often unstable and eventually collapsed. Because of the similarity of this case, he decides not to enter, because the risk of the house collapsing is too great.

Example-based Explanations

- Example-based methods for XAI [Molnar, 2019]:
  - **Counterfactual explanations** tell us how an instance has to change to significantly change its prediction. By creating counterfactual instances, we learn about how the model makes its predictions and can explain individual predictions.
    - *If X had not occurred, Y would not have occurred*
  - **Adversarial examples** are counterfactuals used to fool machine learning models. The emphasis is on flipping the prediction and not explaining it.
  - **Prototypes** are a selection of representative instances from the data and **criticisms** are instances that are not well represented by those prototypes.
  - **Influential instances** are the training data points that were the most influential for the parameters of a model or its predictions. Identifying and analysing them helps to find problems with the data, debug the model and better understand the model's behaviour.
  - **k-Nearest Neighbors** is an (interpretable) machine learning model based on examples

Counterfactuals



(a) Factual

I think the test image is a "1" because it looks like this "1" in the training data.

Test Image
Label: 6
Prediction: 1

Factual
Label: 1
Prediction: 1

(b) Semi-Factual

Even if the test image looked like this, I still would have thought it was a "9".

Test Image
Label: 9
Prediction: 9

Semi-Factual
New Prediction: 9

(c) Counterfactual

If the test image looked like this, I would have thought it was an "8".

Test Image
Label: 8
Prediction: 3

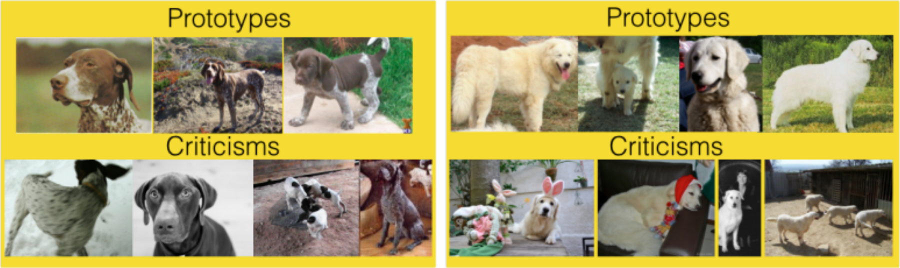
Counterfactual
New Prediction: 8

Min-Edit Counterfactual
New Prediction: 8

Figure 3: *Post-Hoc* Factual, Semi-Factual, and Counterfactual Explanations on MNIST: (a) a *factual explanation* for a misclassification of "6" as "1" found using the twin-system approach (Kenny and Keane 2019), (b) a *semi-factual explanation* for the correct classification of a "9", that shows a synthetic instance with meaningful feature changes that would *not* alter its classification, and (c) a *counterfactual explanation* for the misclassification of an "8" as a "3", that shows a synthetic instance with meaningful feature changes that would cause the CNN to correct its classification (n.b., for comparison a counterfactual using the *Min-Edit* method (see Expt. 1) is shown with its human-undetectable feature-changes).

[Kenny and Keane, 2021]

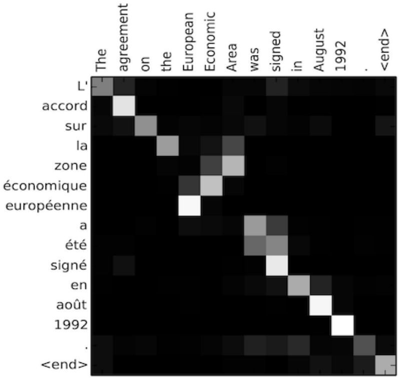
Prototypes and Criticisms

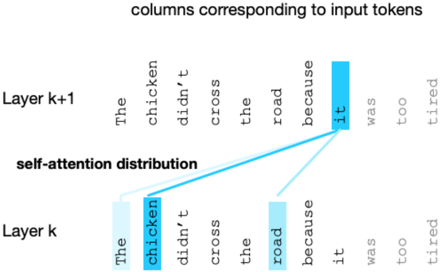


Learned prototypes and criticisms from Imagenet dataset (two types of dog breeds) [Kim et al., 2016]

Attention

- Attention determines the relative importance of each component in a sequence relative to the other components in that sequence.
- (Self-)attention is a key feature of the Transformer [Vaswani et al., 2017]





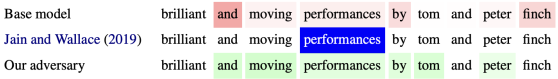
Explaining Large Language Models

Attention is not Explanation [Jain and Wallace, 2019]

- Consistency: to what extent do induced attention weights correlate with measures of feature importance – specifically, those resulting from gradients and leave-one-out methods?
  - Standard **attention methods do not provide meaningful explanations** and should not be treated as though they do.
  - Learned attention weights are **frequently uncorrelated with gradient-based measures of feature importance**.
- Counterfactual attention weights: would alternative attention weights necessarily yield different predictions?
  - One can identify **very different attention distributions that nonetheless yield equivalent predictions**.

Explaining Large Language Models

Attention is not not Explanation [Wiegreffe and Pinter, 2019]



Base model	brilliant	and	moving	performances	by	tom	and	peter	finch
Jain and Wallace (2019)	brilliant	and	moving	performances	by	tom	and	peter	finch
Our adversary	brilliant	and	moving	performances	by	tom	and	peter	finch

Figure 2: Attention maps for an IMDb instance (all predicted as positive with score > 0.998), showing that in practice it is difficult to learn a distant adversary which is consistent on all instances in the training set.

- Attention is not Explanation? In fact, it depends on one's definition of explanation!
  - Plausible explanation? Faithful explanation?
    - Humans often invent a story that plausibly justifies their actions, even if it is not an entirely accurate reconstruction of the neural processes that produced their behaviour at the time
    - Explainability *vs* Interpretability
- Attention Distribution is not a Primitive, i.e., detaching the attention scores obtained by parts of the model degrades the model itself
- Existence does not Entail Exclusivity, i.e., attention scores do provide *an* explanation, not *the* explanation.

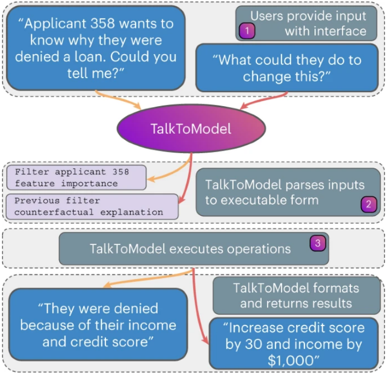
Recent Approaches to XAI

- With the advances in Natural Language Processing a range of alternative explanation methods has been explored
  - Human-readable output, i.e., natural language
  - Many aim at explaining **Large Language Models**
    - Broadly used nowadays, despite being black-boxes
    - Increased trustworthiness is necessary!

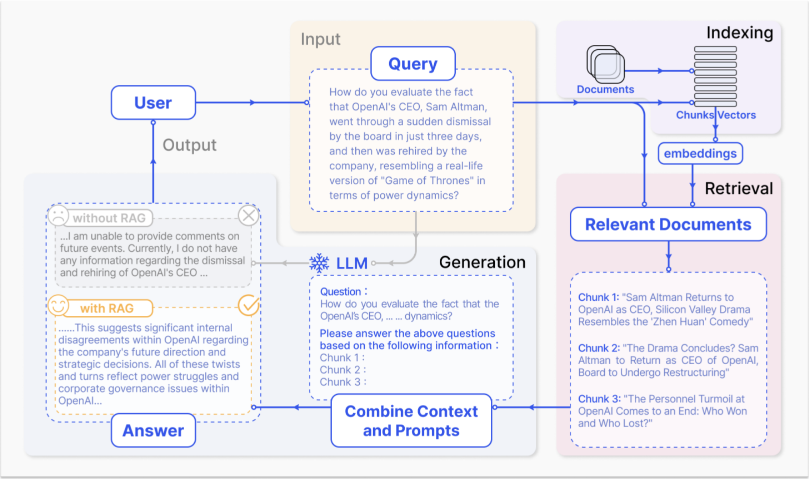
Natural Language Explanations

Talk To Model [Slack et al., 2023]

- UI and conversation tool that makes interpreting models easier for laypeople and ML practitioners alike
  - Uses many post-hoc explanations
  - Chooses the "best" explanation to practitioner
  - Answers user questions with natural language
  - Users only need to provide the model and the data
  - They can communicate via natural language


"Applicant 358 wants to know why they were denied a loan. Could you tell me?"
1 Users provide input with interface
"What could they do to change this?"
TalkToModel
Filter applicant 358 feature importance
Previous filter counterfactual explanation
TalkToModel parses inputs to executable form 2
TalkToModel executes operations 3
TalkToModel formats and returns results
"They were denied because of their income and credit score"
"Increase credit score by 30 and income by $1,000"

Retrieval Augmented Generation



User

Output

Input

Query

How do you evaluate the fact that OpenAI's CEO, Sam Altman, went through a sudden dismissal by the board in just three days, and then was rehired by the company, resembling a real-life version of "Game of Thrones" in terms of power dynamics?

Indexing

Documents

Chunks Vectors

embeddings

Retrieval

Relevant Documents

Chunk 1: "Sam Altman Returns to OpenAI as CEO, Silicon Valley Drama Resembles the 'Zhen Huan' Comedy"

Chunk 2: "The Drama Concludes? Sam Altman to Return as CEO of OpenAI, Board to Undergo Restructuring"

Chunk 3: "The Personnel Turmoil at OpenAI Comes to an End: Who Won and Who Lost?"

LLM

Generation

Question：
How do you evaluate the fact that the OpenAI's CEO, ... dynamics?

Please answer the above questions based on the following information：
Chunk 1 :
Chunk 2 :
Chunk 3 :

Combine Context and Prompts

without RAG

...I am unable to provide comments on future events. Currently, I do not have any information regarding the dismissal and rehiring of OpenAI's CEO ...

with RAG

......This suggests significant internal disagreements within OpenAI regarding the company's future direction and strategic decisions. All of these twists and turns reflect power struggles and corporate governance issues within OpenAI...

Answer

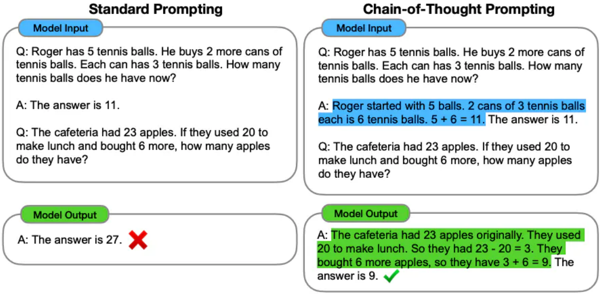
[Gao et al., 2024]

- Retrieved documents as an explanation?

Explaining Large Language Models

Chain of Thought Prompting [Wei et al., 2022]

- Chain of Thought (CoT): series of intermediate natural language reasoning steps that lead to the final output
  - Allows model to decompose multi-step problems into intermediate steps
  - Interpretable method to understand model's response
  - Potentially applicable to any task that require multi-step approach.
  - Can be induced simply by including examples of CoT sequences while prompting.



Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

A: The answer is 27. ✗

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

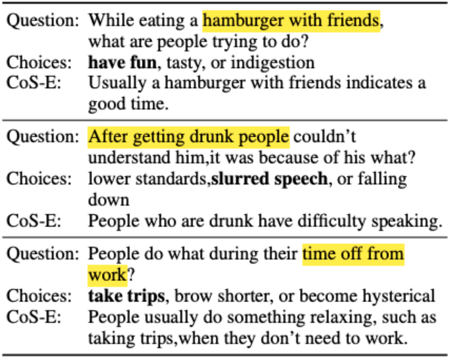
Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

Explaining Large Language Models

Leveraging Language Models for Common Sense Reasoning [Rajani et al., 2019]

- Manually add common-sense explanations (CoS-E) to dataset of common-sense Question Answering.
- Fine-tuned a LM for generating explanations
- Explanations can be used by a classifier
  - Further contributes to better performance is the challenging task of common-sense reasoning!

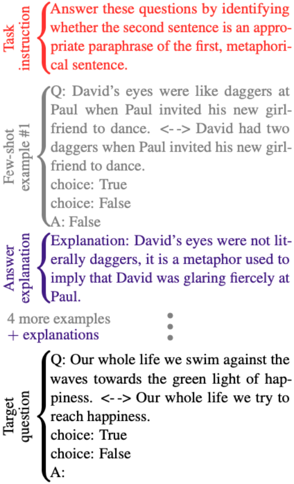


Question:	While eating a hamburger with friends, what are people trying to do?
Choices:	**have fun**, tasty, or indigestion
CoS-E:	Usually a hamburger with friends indicates a good time.
Question:	After getting drunk people couldn't understand him,it was because of his what?
Choices:	lower standards,**slurred speech**, or falling down
CoS-E:	People who are drunk have difficulty speaking.
Question:	People do what during their time off from work?
Choices:	**take trips**, brow shorter, or become hysterical
CoS-E:	People usually do something relaxing, such as taking trips,when they don't need to work.

Explaining Large Language Models

Explanations in Context [Lampinen et al., 2022]

- In-context learning
  - Few-shot:
    questions + answers + explanations
- Explanations further improve performance!


Task instruction
Few-shot example #1
Answer explanation
Target question

Answer these questions by identifying whether the second sentence is an appropriate paraphrase of the first, metaphorical sentence.

Q: David's eyes were like daggers at Paul when Paul invited his new girlfriend to dance. <--> David had two daggers when Paul invited his new girlfriend to dance.
choice: True
choice: False
A: False

Explanation: David's eyes were not literally daggers, it is a metaphor used to imply that David was glaring fiercely at Paul.

4 more examples
+ explanations

Q: Our whole life we swim against the waves towards the green light of happiness. <--> Our whole life we try to reach happiness.
choice: True
choice: False
A:

Discussion

- LLMs are large black boxes, broadly used
  - Explanations are necessary!
- Limitations of presented methods?

Questions?

Bibliography I

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., and Wang, H. (2024).
Retrieval-augmented generation for large language models: A survey.

Jain, S. and Wallace, B. C. (2019).
Attention is not Explanation.
In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Kenny, E. M. and Keane, M. T. (2021).
On generating plausible counterfactual and semi-factual explanations for deep learning.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11575–11585.

Kim, B., Khanna, R., and Koyejo, O. O. (2016).
Examples are not enough, learn to criticize! criticism for interpretability.
Advances in neural information processing systems, 29.

Lampinen, A., Dasgupta, I., Chan, S., Mathewson, K., Tessler, M., Creswell, A., McClelland, J., Wang, J., and Hill, F. (2022).
Can language models learn from explanations in context?
In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 537–563, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Molnar, C. (2019).
Interpretable Machine Learning.
Independently published.
https://christophm.github.io/interpretable-ml-book/.

Bibliography II

Rajani, N. F., McCann, B., Xiong, C., and Socher, R. (2019).
Explain yourself! leveraging language models for commonsense reasoning.
In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Slack, D., Krishna, S., Lakkaraju, H., and Singh, S. (2023).
Explaining machine learning models with interactive natural language conversations using talktomodel.
Nature Machine Intelligence, 5(8):873–883.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017).
Attention is all you need.
In *Advances in neural information processing systems*, pages 5998–6008.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022).
Chain-of-thought prompting elicits reasoning in large language models.
Advances in neural information processing systems, 35:24824–24837.

Wiegreffe, S. and Pinter, Y. (2019).
Attention is not not explanation.
In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.