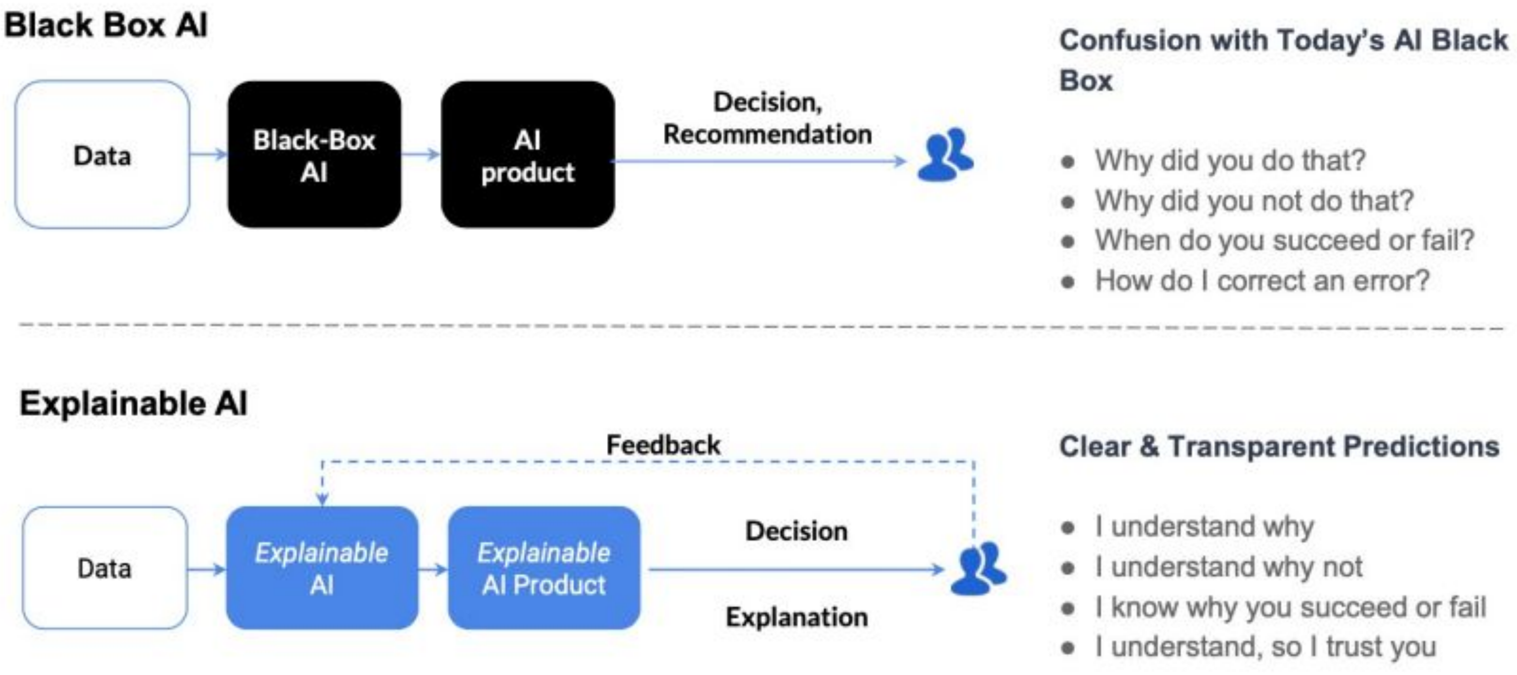
Explainable AI

Introduction to AI
May 10, 2022

Outline

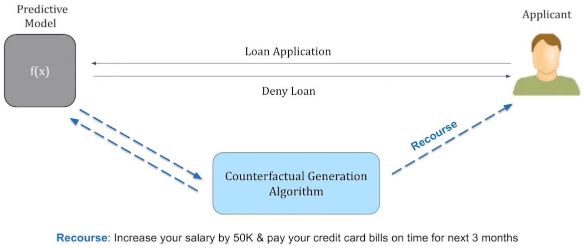
- **Explainable AI**
  - Categorization of explanations
    - Local v.s Global
    - Self-explaining v.s Post-hoc
  - Generating and presenting explanations
    - Explainability techniques
    - Visualization techniques
  - Evaluation of explanations
- **Attacks in NLP**

Black Box AI v.s Explainable AI



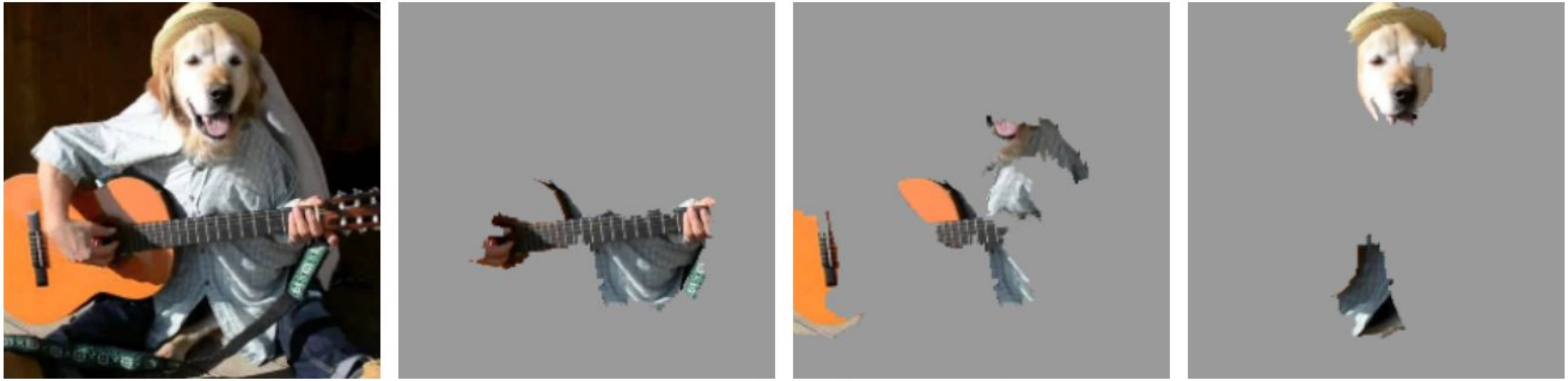
Motivation

- From **business** perspective
    - Medical AI: e.g., diagnostics, anesthesiology
    - Financial AI: e.g., credit scoring, loan approval



- From **model** perspective
    - Debug mispredictions
    - Understand weaknesses and improve ML model
    - Learn new insights

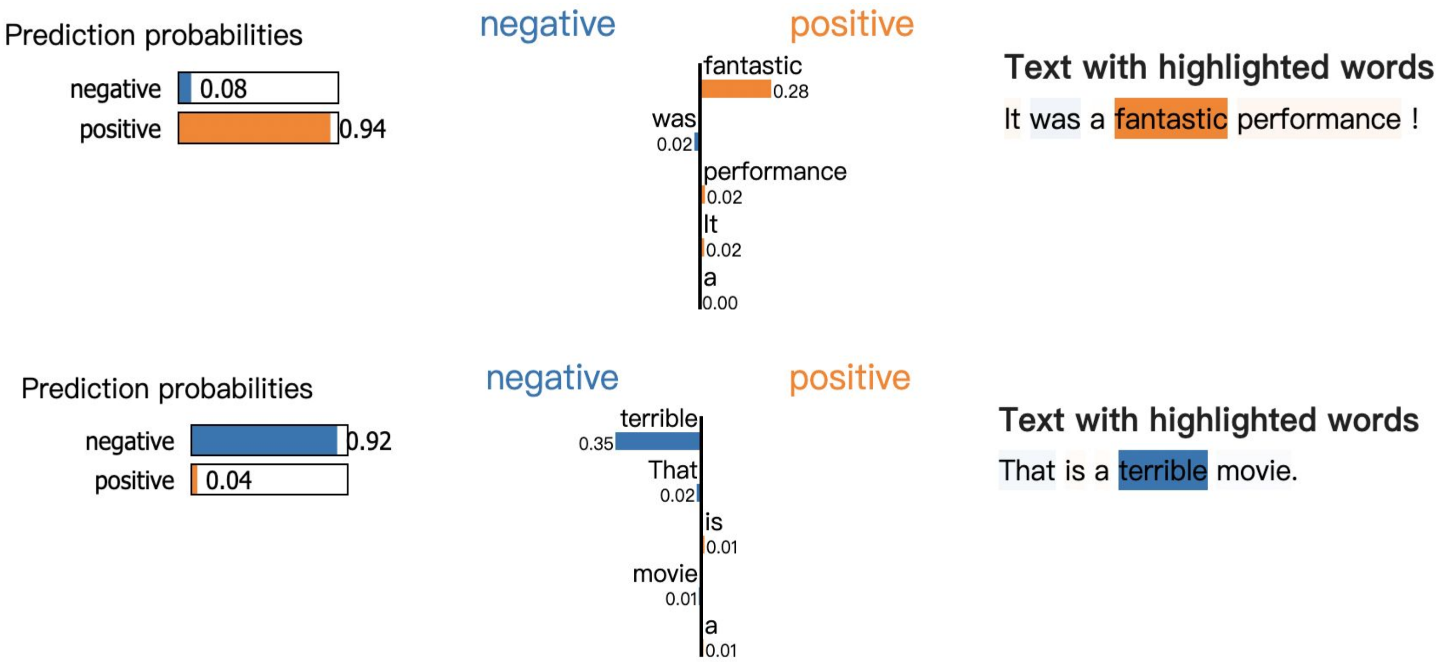
Example: Explanability for Computer Vision



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

[“Why Should I Trust You?” Explaining the Predictions of Any Classifier. (KDD 2016)]

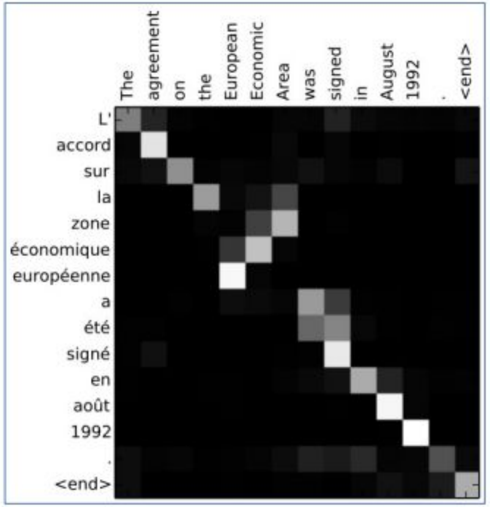
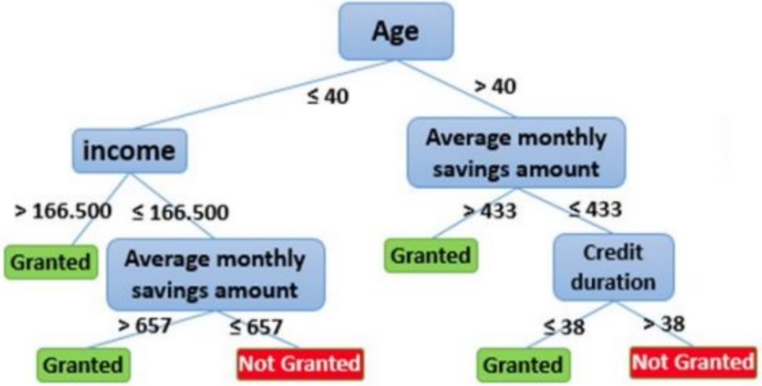
Example: Explanability for NLP



Categorization of different types of explanation

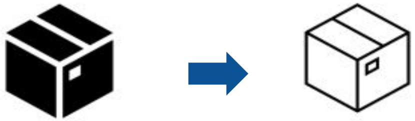
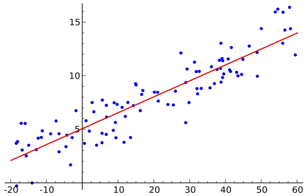
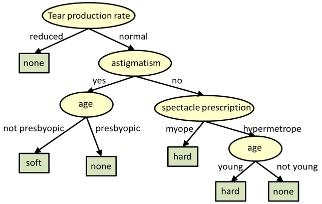
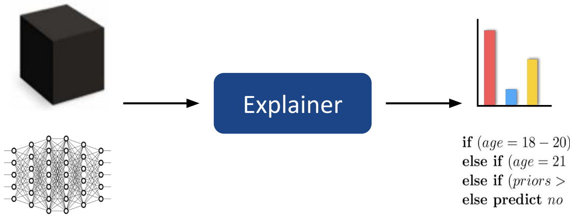
- Local v.s Global
- Self-explaining vs. Post-hoc explanation

Local explanation v.s Global explanation

Local explanation	Global explanation
Explain a **single prediction**	Understand **the whole logic** of a model
	

Self-explaining vs. Post-hoc explanation

<table>
<tr><th>Self-explaining</th><th>Post-hoc explanation</th></tr>
<tr><td>Directly get the explanation with the prediction

Only one model</td><td>Does not come directly with the prediction

Model 1
Explainer</td></tr>
<tr><td></td><td>
if $(age = 18 - 20)$ and $(sex = male)$ then predict yes
else if $(age = 21 - 23)$ and $(priors = 2 - 3)$ then predict yes
else if $(priors > 3)$ then predict yes
else predict no</td></tr>
</table>

Explaining machine learning predictions: State-of-the-art, challenges, andopportunities. (NeurIPS 2020 Tutorial)

Explainability Techniques in NLP

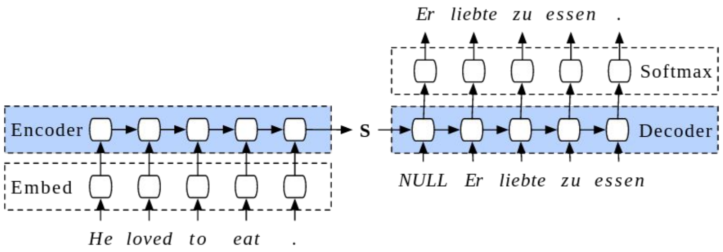
- Feature importance
- Surrogate model
- Example-driven
- Provenance-based
- Declarative induction

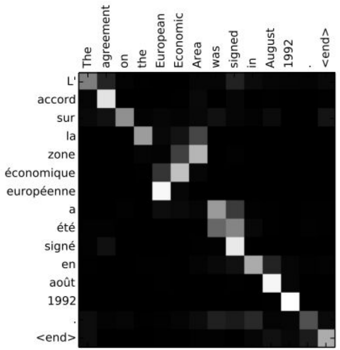
Explainability - 1. Feature Importance

- The main idea of feature importance is to derive explanation by investigating the **importance scores of different features** used to output the final prediction
- **3 types of features**
    - **Manual features** from feature engineering
    - **Lexical features** including words/tokens and N-gram
    - **Latent features** learned by neural nets

Explainability - 1. Feature Importance

- Example 1: Attention mechanism





- Example 2: Integrated Gradients

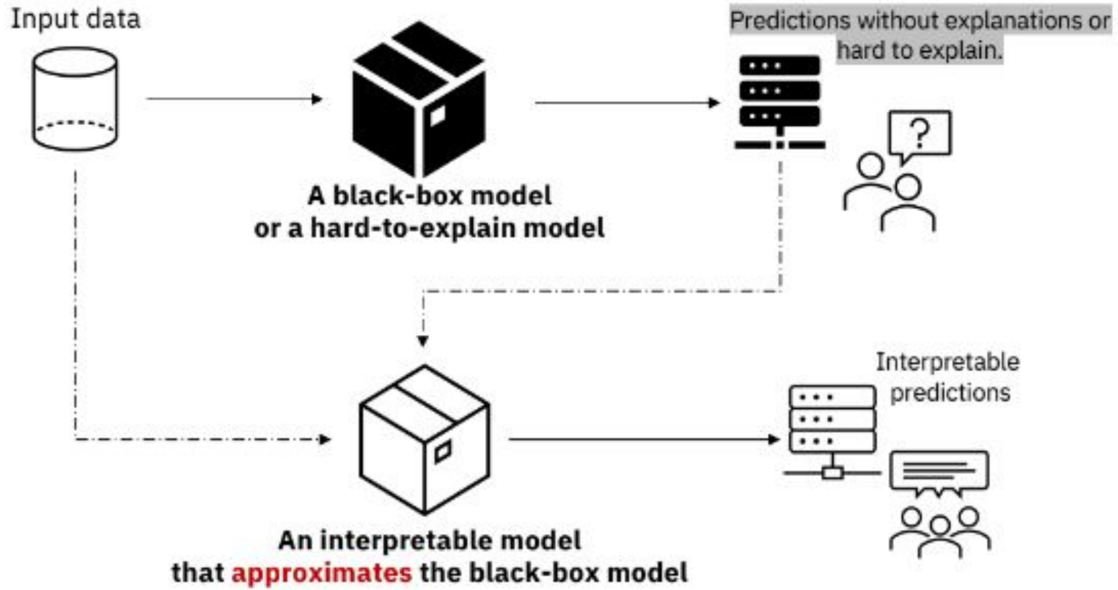
Question: Why was Polonia relegated from the country's top flight in 2013?
Answer: disastrous financial situation

L0	Polonia was relegated from the country's top flight in 2013 because of their disastrous financial situation. They are now playing in the 4th league....	**L9**	Polonia was relegated from the country's top flight in 2013 because of their disastrous financial situation. They are now playing in the 4th league....
L1	Polonia was relegated from the country's top flight in 2013 because of their disastrous financial situation. They are now playing in the 4th league....	**L10**	Polonia was relegated from the country's top flight in 2013 because of their disastrous financial situation. They are now playing in the 4th league....
L2	Polonia was relegated from the country's top flight in 2013 because of their disastrous financial situation. They are now playing in the 4th league....	**L11**	Polonia was relegated from the country's top flight in 2013 because of their disastrous financial situation. They are now playing in the 4th league....

[Neural Machine Translation by Jointly Learning to Align and Translate (ICLR2015)]
[Towards Interpreting BERT for Reading Comprehension Based QA (ACL2020)]

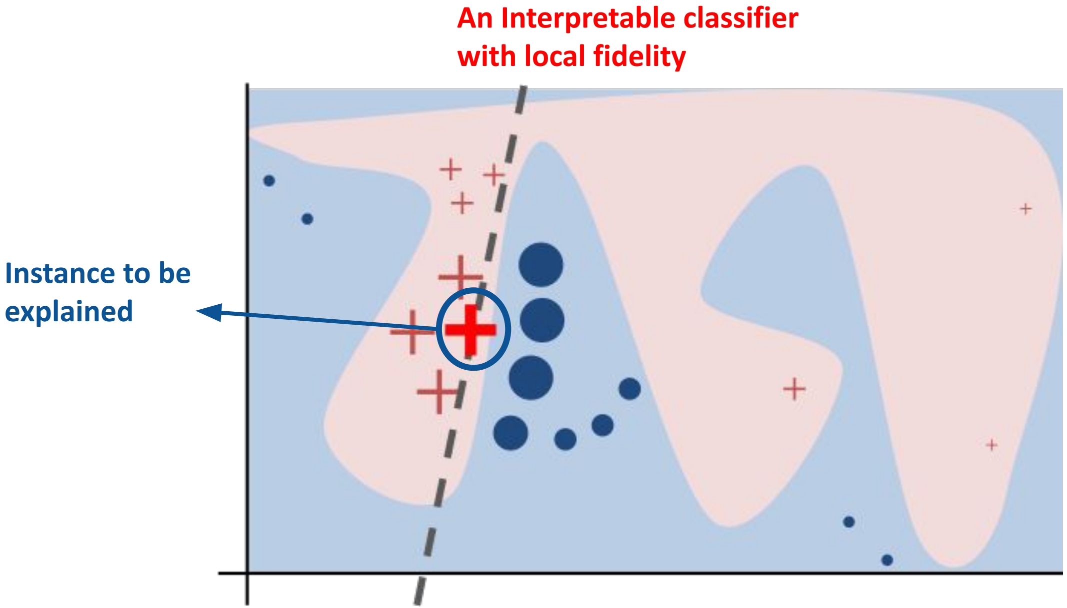
Explainability - 2. Surrogate Model

- Model predictions are explained by **learning a second, usually more explainable model**, as a proxy



Explainability - 2. Surrogate Model

- Example: LIME (Local Interpretable Model-agnostic Explanations)



[“Why Should I Trust You?” Explaining the Predictions of Any Classifier (KDD 2016)]

Explainability - 3. Example-driven

- Such approaches explain the prediction of an input instance **by identifying and presenting other instances that are semantically similar to the input instance**
- Example

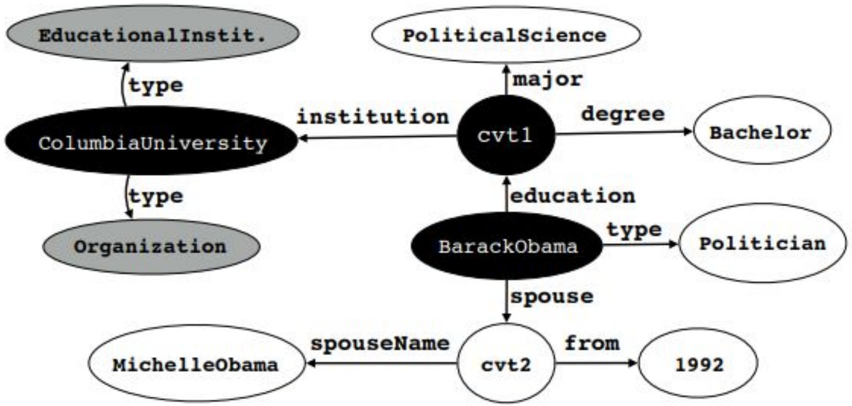
  "What is the capital of Germany?" refers to a Location. **WHY?**

  Because "What is the capital of California?" which also refers to a **Location** in the training data

Auditing Deep Learning processes through Kernel-based Explanatory Models (EMNLP2019)

Explainability - 4. Provenance-based

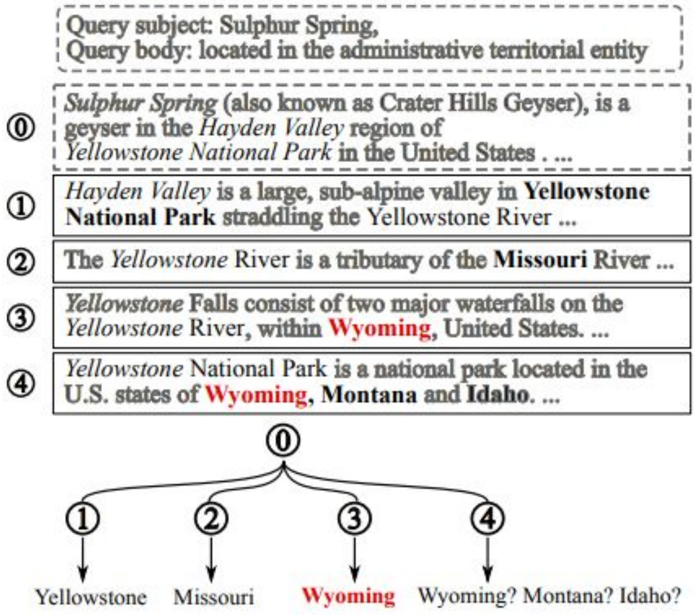
- Explanations are provided by **illustrating (some of) the prediction derivation process**
- Example: QUINT - a live system for question answering over knowledge bases
  - E.g. "Where was Obama educated?" and the answer entity ColumbiaUniversity



[QUINT: Interpretable Question Answering over Knowledge Bases (EMNLP2017)]

Explainability - 5. Declarative Induction

- The main idea of declarative induction is to **construct human-readable representations such as trees, programs, and rules**
- Example: ExplorePropose-Assemble reader (EPAr)



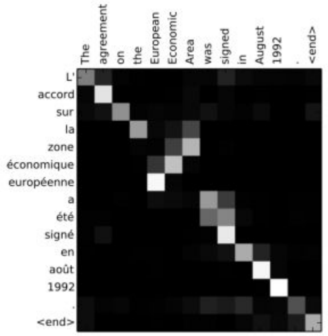
Explore, Propose, and Assemble: An Interpretable Model for Multi-Hop Reading Comprehension (ACL2019)

Visualization Techniques

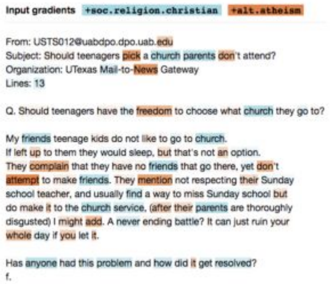
1. Saliency
2. Declarative Representations
3. Natural language explanation
4. Others: raw example…

Visualization - 1. Saliency

- Has been primarily used to visualize the importance scores of different types of elements in XAI learning systems
- **Most used!**
- E.g.
  - input-output word alignment
  - highlighting words in input text
  - displaying extracted relations



(a) Saliency heatmap (Bahdanau et al., 2015)



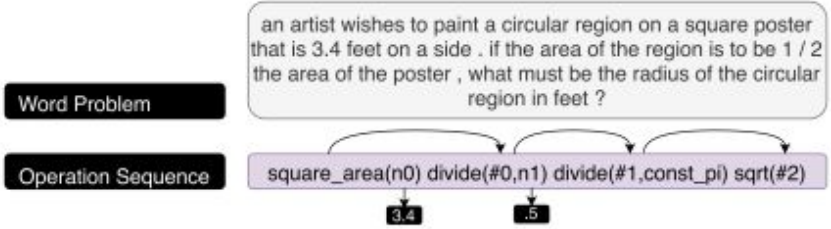
(b) Saliency highlighting (Mullenbach et al., 2018)

Rule Body, $R_1(a,c) \wedge R_2(c,b) \Rightarrow$	**Target,** $R(a,b)$
Common to both	
isConnectedTo$(a,c)\wedge$ isConnectedTo(c,b)	isConnectedTo
isLocatedIn$(a,c)\wedge$ isLocatedIn(c,b)	isLocatedIn
isAffiliatedTo$(a,c)\wedge$ isLocatedIn(c,b)	wasBornIn
isMarriedTo$(a,c)\wedge$ hasChild(c,b)	hasChild
only in DistMult	
playsFor$(a,c)\wedge$ isLocatedIn(c,b)	wasBornIn
dealsWith$(a,c)\wedge$ participatedIn(c,b)	participatedIn
isAffiliatedTo$(a,c)\wedge$ isLocatedIn(c,b)	diedIn
isLocatedIn$(a,c)\wedge$ hasCapital(c,b)	isLocatedIn
only in ConvE	
influences$(a,c)\wedge$ influences(c,b)	influences
isLocatedIn$(a,c)\wedge$ hasNeighbor(c,b)	isLocatedIn
hasCapital$(a,c)\wedge$ isLocatedIn(c,b)	exports
hasAdvisor$(a,c)\wedge$ graduatedFrom(c,b)	graduatedFrom
Extractions from DistMult [Yang et al., 2015]	
isLocatedIn$(a,c) \wedge$ isLocatedIn(c,b)	isLocatedIn
isAffiliatedTo$(a,c) \wedge$ isLocatedIn(c,b)	wasBornIn
playsFor$(a,c) \wedge$ isLocatedIn(c,b)	wasBornIn
isAffiliatedTo$(a,c) \wedge$ isLocatedIn(c,b)	diedIn

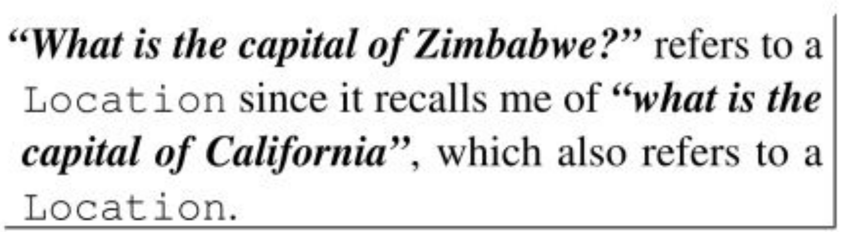
(c) Raw declarative rules (Pezeshkpour et al., 2019b)

Visualization - 2. Declarative Representations

- Directly present the **learned declarative representations**
- E.g. logic rules, trees, and programs



(d) Raw declarative program (Amini et al., 2019)



(e) Raw examples (Croce et al., 2019)

Visualization - 3. Natural Language Explanation

- The explanation is **verbalized in human-comprehensible language**
- Generated by using
  - sophisticated deep learning approaches
  - simple template-based approaches

Evaluation - Challenge

Unlike in traditional Machine Learning (ML), the task of explaining inherently **lacks “ground-truth” data** — there is no universally accepted definition of what constitutes a “correct” explanation.

Evaluation

- **Informal Examination**
  - **High-level discussions** of how examples of generated explanations align with human intuition
  - Compared to other explainable approaches
- **Comparison to Ground Truth**
  - Quantify the performance of explainability techniques
- **Human Evaluation**
  - Humans directly evaluate the effectiveness of the generated explanations via one or more user studies
  - Pros: avoid the assumption that **there is only one good explanation that could serve as ground truth**

Summary of XAI

- Categorization of explanations
  - Local v.s Global
  - Self-explaining v.s Post-hoc
- Generating and presenting explanations
  - Explainability techniques
  - Visualization techniques
- Evaluation of explanations

DistilBERT model in HW2 can be attacked?

- Results of attacking against various fine-tuned BERT models.

Dataset	Method	Original Acc	Attacked Acc	Perturb %	Query Number	Avg Len	Semantic Sim
IMDB	BERT-Attack(ours)	90.9	**11.4**	**4.4**	**454**	215	**0.86**
	TextFooler		13.6	6.1	1134		**0.86**
	GA		45.7	4.9	6493		-

[BERT-ATTACK: Adversarial Attack Against BERT Using BERT (EMNLP 2020)]

Attacks in NLP

Introduction to AI
May 10, 2022

Why we should care about adversarial attack?

- The man posted a picture of himself leaning against a bulldozer with the caption **“يصبحهم”, or “yusbihuhum”, which translates as “good morning”**. But Facebook’s artificial intelligence-powered translation service, which it built after parting ways with Microsoft’s Bing translation in 2016, instead **translated the word into “hurt them” in English or “attack them” in Hebrew**

[News reference]

White box attack v.s Black box attack

White box attack	Black box attack
The attacker has access to the **model's parameters**	The attacker has **no access to these parameters**, i.e., it uses a different model or no model at all to generate adversarial images with the hope that these will transfer to the target model

NLP Attacks

- Useful toolkit: Textattack

1. Goal	Stipulate **the goal of the attack**, like to change the prediction score of a classification model, or to change all of the words in a translation output
2. Constrains	**Determine if a potential perturbation is valid** with respect to the original input
3. Transformation	Take a text input and **transform it by inserting and deleting characters, words, and/or phrases**
4. Search method	Explore the space of possible transformations within the defined constraints and attempt to **find a successful perturbation which satisfies the goal function**

Attacks In HW4

1. Goal	Change the prediction, i.e., positive → negative, negative → positive
2. Constrains	No constrain. But it will be better if you minimum the difference between original sentence and attacked sentence
3. Transformation	Try it by yourself
4. Search method	You can based on the result of LIME and SHAP

Summary of Attacks in NLP

- Model robustness in NLP is important, instead of encouraging you to attack online APIs or release toxic datasets, we should think about **how to prevent the attack**