Patrick Fernandes

Explainable AI and How to Evaluate It

Applications to NLP

Carnegie Mellon Portugal
TÉCNICO LISBOA
Language Technologies Institute

Why is it relevant?

❖ Recently deep learning as emerged as the main paradigm in artificial intelligence research

Why is it relevant?

- Recently deep learning as emerged as the main paradigm in artificial intelligence research

- However this great improvements came with a cost: lack of interpretability

Why is it relevant?

- Recently deep learning as emerged as the main paradigm in artificial intelligence research

- However this great improvements came with a cost: lack of interpretability

- Interpretability is necessary for many use cases:

Why is it relevant?

- ❖ Recently deep learning as emerged as the main paradigm in artificial intelligence research
- ❖ However this great improvements came with a cost: lack of interpretability
- ❖ Interpretability is necessary for many use cases:
  - ➢ High Reliability

Why is it relevant?

- Recently deep learning as emerged as the main paradigm in artificial intelligence research
- However this great improvements came with a cost: lack of interpretability
- Interpretability is necessary for many use cases:
  - High Reliability
  - Ethical and Legal Requirements

Why is it relevant?

❖ Recently deep learning as emerged as the main paradigm in artificial intelligence research

❖ However this great improvements came with a cost: lack of interpretability

❖ Interpretability is necessary for many use cases:
  - ➢ High Reliability
  - ➢ Ethical and Legal Requirements
  - ➢ Scientific uses

Why is it relevant?

- Recently deep learning as emerged as the main paradigm in artificial intelligence research

- However this great improvements came with a cost: lack of interpretability

- Interpretability is necessary for many use cases:
  - High Reliability
  - Ethical and Legal Requirements
  - Scientific uses
  - **Model Debugging!**

Why is it relevant?

[1]

The [MASK] ran to the emergency room to see his patient.

Mask 1 Predictions:

36.5% **doctor**
12.7% **man**
2.8% **boy**
2.7% **nurse**
2.0% **patient**

The [MASK] ran to the emergency room to see her patient.

Mask 1 Predictions:

44.9% **nurse**
19.3% **woman**
7.4% **doctor**
5.3% **girl**
3.6% **mother**

[CLS] The [MASK] ran to the emergency room to see her patient . [SEP]

[1] Wallace et al. *EMNLP 2020 Interpretability Tutorial*

What is interpretability?

- Interpretability is the *ability to provide* ***explanations*** *in* ***understandable terms*** *to a human* [2]

[2] Zhang et al. A Survey on Neural Network Interpretability

What is interpretability?

- Interpretability is the *ability to provide* ***explanations*** *in* ***understandable terms*** *to a human* [2]
  - **Explanations:** natural language, logic rules or something else.

[2] Zhang et al. A Survey on Neural Network Interpretability

What is interpretability?

- Interpretability is the *ability to provide **explanations** in **understandable terms** to a human* [2]
  - **Explanations:** natural language, logic rules or something else.
  - **Understandable terms:** terms from domain knowledge related to the task

[2] Zhang et al. A Survey on Neural Network Interpretability

What is interpretability?

TABLE I
SOME INTERPRETABLE "TERMS" USED IN PRACTICE. [2]

Field	Raw input	Understandable terms
Computer vision	Images (pixels)	Super pixels (image patches)[a] Visual concepts[b]
NLP	Word embeddings	Words
Bioinformatics	Sequences	Motifs (position weight matrix)[c]

[a] image patches are usually used in attribution methods [17].
[b] colours, materials, textures, parts, objects and scenes [18].
[c] proposed by [19] and became an essential tool for computational motif discovery.

[2] Zhang et al. A Survey on Neural Network Interpretability [2020]

What is interpretability?

- Explanations should then be
    - **Readable:** They are human-interpretable
    - **Plausible:** How persuasive these explanations are
    - **Faithful:** They represent the underlying model decision process

Some examples of interpretability methods

- ❖ LIME [3]

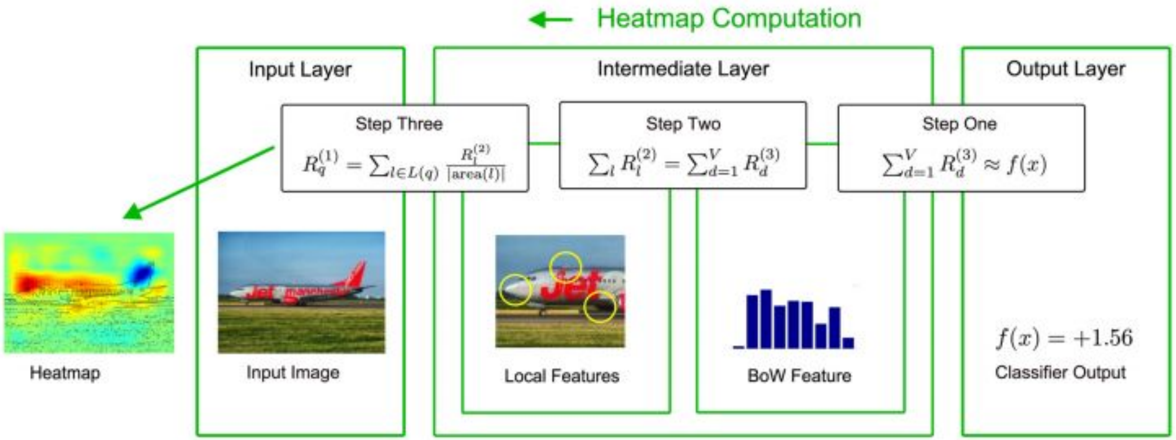
(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

[3] Ribeiro et al. "Why Should I Trust You?" Explaining the Predictions of Any Classifier [2016]

Some examples of interpretability methods

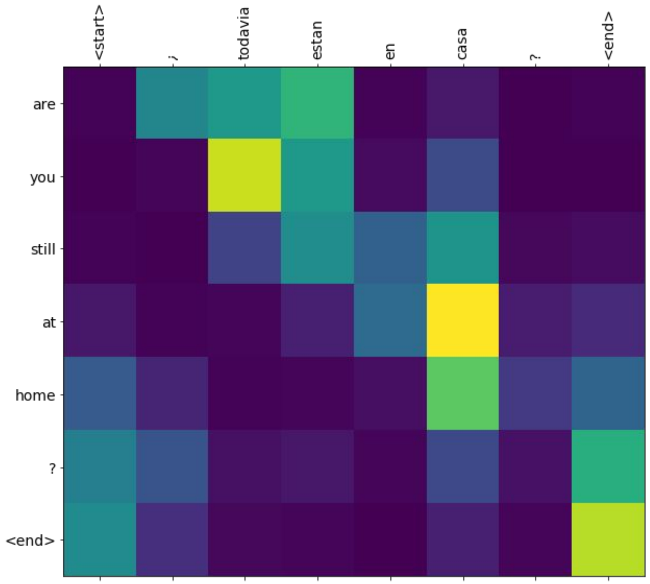
❖ LRP [4], Integrated Gradients [5], ...



[4] Bach et al. *On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation* [2015]
[5] Sundararajan et al. *Axiomatic Attribution for Deep Networks* [2017]

Some examples of interpretability methods

❖ Attention-based Explanations

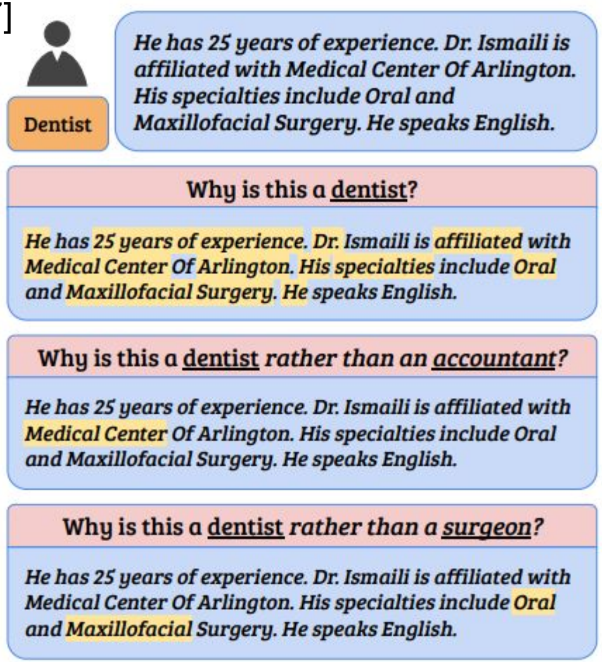


[6]

[6] https://www.tensorflow.org/tutorials/text/nmt_with_attention

Some examples of interpretability methods

- Contrastive Explanations [7]



[7] Jacovi et al. *Compositional Explanations of Neurons* [2021]

Some examples of interpretability methods

❖ Compositional Neurons [8]

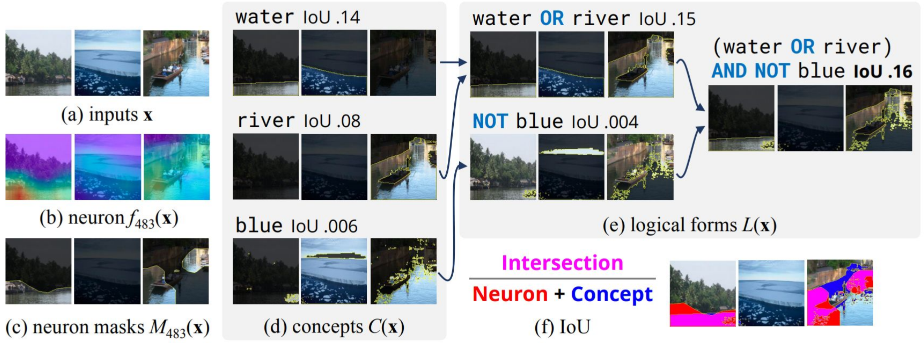


Figure 1: Given a set of inputs (a) and scalar neuron activations (b) converted into binary masks (c), we generate an explanation via beam search, starting with an inventory of primitive concepts (d), then incrementally building up more complex logical forms (e). We attempt to maximize the IoU score of an explanation (f); depicted is the IoU of $M_{483}(\mathbf{x})$ and `(water OR river) AND NOT blue`.

[8] Mu & Andreas. *Compositional Explanations of Neurons* [2020]

The evaluation problem

- ❖ All these methods provide different “perspectives” on the inner workings of neural networks

The evaluation problem

❖ All these methods provide different “perspectives” on the inner workings of neural networks

LIME:

[CLS] It ' s hard to believe that a movie this **bad** could actually be released . The dial ##og was **unnatural** . Especially **poor** was the portrayal of the relationship between the boy and his future step - **father** . I guess you could say that they succeeded in producing awkward dial ##og , **but** what was said seemed false and **artificial** .

*Input * Gradients*

[CLS] It ' s hard to believe that a movie **this** bad could actually be released . The dial ##og **was** unnatural . **Especially** poor **was** the portrayal of the relationship between the boy and his future step - father . I guess you could say that they succeeded in producing awkward dial ##og , but what **was** said seemed false and artificial .

Attention

[CLS] It ' s hard to believe that a movie **this bad** could actually be released . **The dial ##og** was **unnatural** . Especially **poor** was **the** portrayal of the relationship between the boy and his future step - father . I guess you could say that they succeeded in producing awkward dial ##og , but what was said seemed false and artificial .

The evaluation problem

- ❖ It’s unclear which one of these methods “better”

The evaluation problem

- ❖ It’s unclear which one of these methods “better”

- ❖ The community lacks a standard *quantitative* measure of the *faithfulness* of explanations

The communication game [9]

- In this work, they frame the explainability problem as a *communication* problem

[9] Treviso and Martins. The Explanation Game: Towards Prediction Explainability through Sparse Communication [2020]

The communication game [9]

- In this work, they frame the explainability problem as a *communication* problem

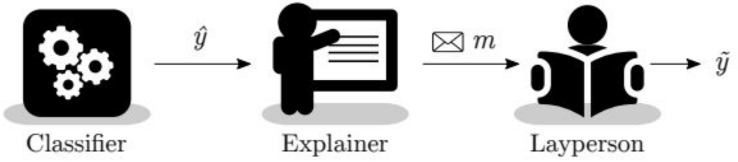


Figure 1: Our framework to model explainability as communication. Predictions $\hat{y}$ are made by a classifier C; an explainer E (either embedded in C or operating post-hoc) accesses these predictions and communicates an explanation (a message m) to the layperson L. Success of the communication is dictated by the ability of L and C to match their predictions: $\tilde{y} \stackrel{?}{=} \hat{y}$. Both the explainer and layperson can be humans or machines.

[9] Treviso and Martins. The Explanation Game: Towards Prediction Explainability through Sparse Communication [2020]

The communication game [9]

- Good explanations are the ones the *communicate* well the information to the layperson

[9] Treviso and Martins. The Explanation Game: Towards Prediction Explainability through Sparse Communication [2020]

The communication game [9]

- Good explanations are the ones the *communicate* well the information to the layperson

Clf.	Explainer	SST CSR	SST ACC$_L$	IMDB CSR	IMDB ACC$_L$	AgNews CSR	AgNews ACC$_L$	Yelp CSR	Yelp ACC$_L$	SNLI CSR	SNLI ACC$_L$
C	Random	69.41	70.07	67.30	66.67	92.38	91.14	58.27	53.06	75.83	68.74
C	Erasure	80.12	81.22	92.17	88.72	97.31	95.41	78.72	68.90	77.88	70.04
C	Top-k gradient	79.35	79.24	86.30	83.93	96.49	94.86	70.54	62.86	76.74	69.40
C	Top-k softmax	84.18	82.43	93.06	89.46	**97.59**	95.61	81.00	70.18	78.66	71.00
C_{ent}	Top-k 1.5-entmax	**85.23**	**83.31**	93.32	89.60	97.29	**95.67**	82.20	70.78	80.23	73.39
C_{sp}	Top-k sparsemax	**85.23**	81.93	93.34	89.57	95.92	94.48	82.50	70.99	**82.89**	**74.76**
C_{ent}	Selec. 1.5-entmax	83.96	82.15	92.55	89.96	97.30	95.66	81.38	70.41	77.25	71.44
C_{sp}	Selec. sparsemax	**85.23**	81.93	93.24	89.66	95.92	94.48	83.55	71.60	82.04	73.46
C_{bern}	Bernoulli	82.37	78.42	91.66	86.13	96.91	94.43	84.93	66.89	76.81	69.65
C_{hk}	HardKuma	85.17	80.40	**94.72**	**90.16**	97.11	95.45	**87.39**	**71.64**	74.98	71.48

[9] Treviso and Martins. The Explanation Game: Towards Prediction Explainability through Sparse Communication [2020]

The communication game [9]

❖ This framework has some downsides *(in my opinion)*:

➢ The explainer has access to the label*

➢ The layperson only has access to the *message*, not the input

➢ The explainer is needed at test time

[9] Treviso and Martins. The Explanation Game: Towards Prediction Explainability through Sparse Communication [2020]

How much do explanations from the teacher aid students? [9]

- They propose evaluating explanations to the degree in which

  “they help a student model in learning to simulate the teacher on future examples”

[9] Pruthi el al. Evaluating Explanations: How much do explanations from the teacher aid students? [2020]

How much do explanations from the teacher aid students? [10]

- They propose evaluating explanations to the degree in which *"they help a student model in learning to simulate the teacher on future examples"*

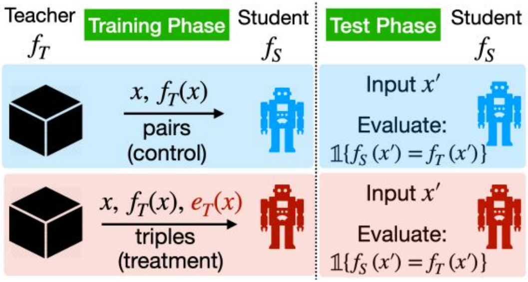


Figure 1: The proposed framework for evaluating explanation quality. Student models learn to mimic the teacher, with and without explanations (provided as "side information" with each example). Explanations are effective if they help students to better approximate the teacher on future test examples *for which such explanations are not available*. Students and teachers could be either models or people.

[10] Pruthi el al. Evaluating Explanations: How much do explanations from the teacher aid students? [2020]

How much do explanations from the teacher aid students? [10]

❖ In this framework, explanations are only used at training time

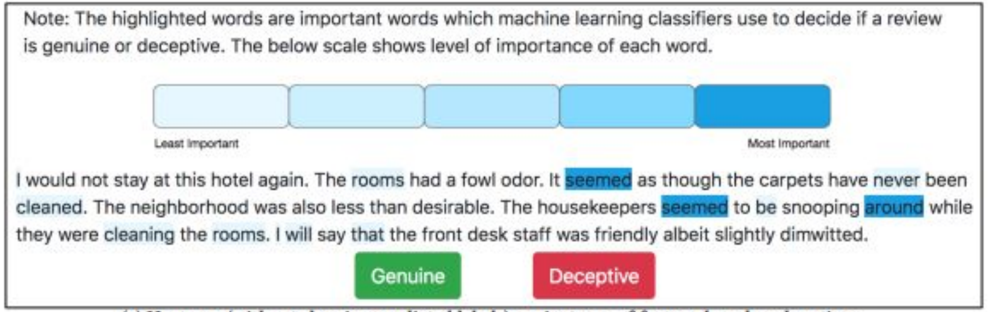
	attention regularization					multi-task learning				
Examples	500	1000	2000	4000	8000	500	1000	2000	4000	8000
No Explanation	90.0	91.5	92.6	93.6	94.9	90.0	91.5	92.6	93.6	94.9
Random Explanation	89.4	90.6	92.4	93.9	94.6	89.6	91.5	92.7	94.1	94.5
Trivial Explanation	78.5	82.8	88.3	92.3	93.5	86.1	90.6	91.5	93.4	93.8
LIME	90.2	91.3	92.6	94.0	94.8	90.2	91.3	92.6	94.0	95.0
Gradient Norm	90.4	91.6	92.4	92.7	93.7	88.8	92.3	93.1	94.3	94.2
Gradient × Input	90.5	91.7	92.2	93.6	94.7	89.3	91.2	92.7	94.4	94.5
Integrated Gradients	92.4	92.6	93.6	94.8	95.7	89.5	91.6	93.3	94.5	95.2
Attention	**92.7**	**93.9**	**95.2**	**96.2**	**97.0**	89.6	91.5	**94.4**	**96.0**	**96.6**

[10] Pruthi el al. Evaluating Explanations: How much do explanations from the teacher aid students? [2020]

Can we "train" humans as well?

- We mostly care about explainability with regard to *humans* (readable) !

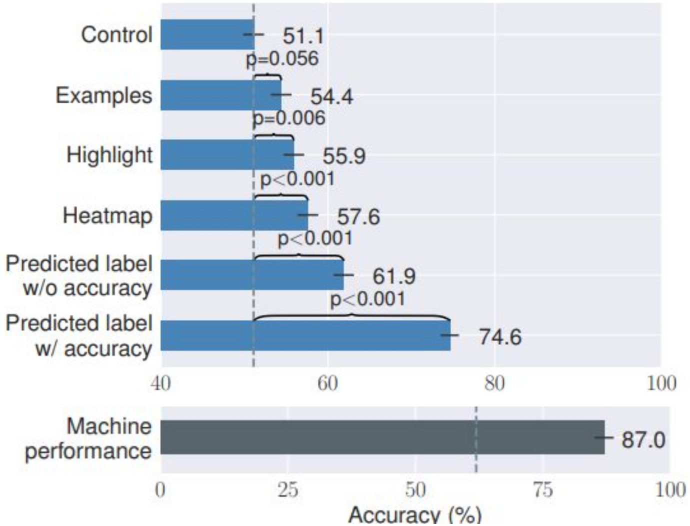
- There has been work that "mimics" this teacher-student evaluation framework with *human* students [11]



(a) Heatmap (without showing predicted labels), an instance of feature-based explanations.

[11] Lai & Tao. On Human Predictions with Explanations and Predictions of Machine Learning Models [2019]

Can we “train” humans as well?



[10] Lai & Tao. On Human Predictions with Explanations and Predictions of Machine Learning Models [2019]

Can we learn better explanations?

- These previous works provided an explicit metric to *evaluate* explanations
  - Good explanations make the student learn *faster/more sample-efficiently*

Can we learn better explanations?

- These previous works provided an explicit metric to *evaluate* explanations
  - Good explanations make the student learn *faster/more sample-efficiently*

- Can we can try to use this as an explicit learning signal to learn better explanations!
  - How? Meta-learning!

Conclusion

- ❖ Interpretability is (very) useful and important
  - ➢ For model debugging, ethical reasons, etc...
- ❖ There are many approaches to bring interpretability to our models based on different principles
- ❖ We are just now scratching the surface of how we can compare these methods with each other!

Thank you