BEIRUT ARAB UNIVERSITY

Unveiling the Black-Box

An Introduction to Explainable AI

Julia El Zini, American University of Beirut

March 29, 2022

juliaelzini@gmail.com

linkedin.com/in/juliaelzini

AGENDA

Terminology

What are we explaining and at which stage of training machine learning (ML) models?

Local Explainability Methods

Explaining local predictions including easy-to-use LIME and SHAP methods

Contrastive Explainability

Explaining models by contrasting their explanations

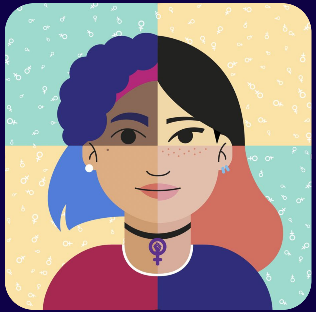
Explaining hidden knowledge

Interpreting learned knowledge by deep learning models in each layer/neuron? tailoring the study to language models

Motivation

Why Explainable AI is important

now more than ever?



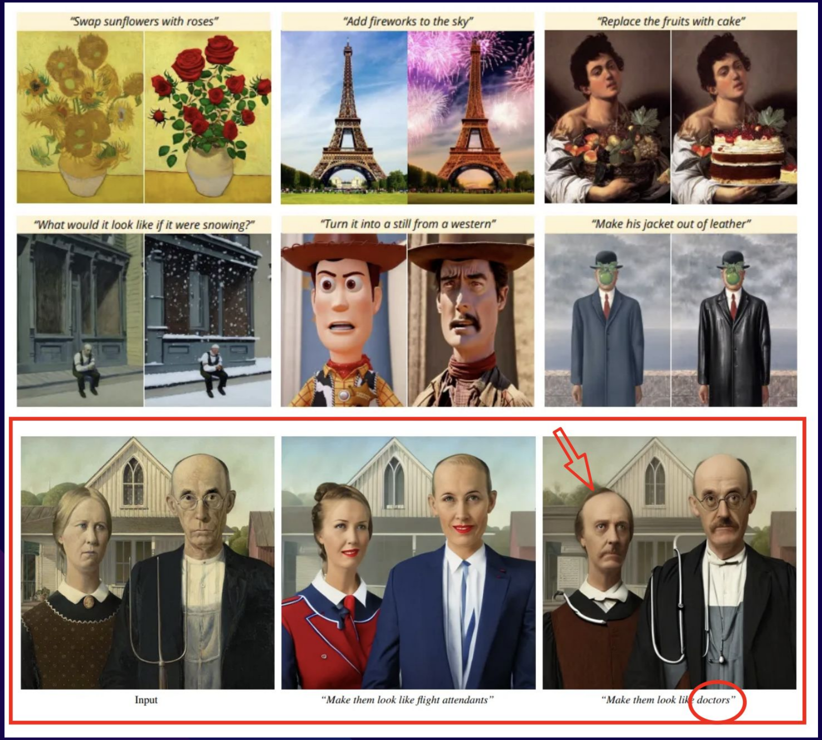
ML affecting almost every aspect in our daily lives → need to ensure accountability and responsibility.

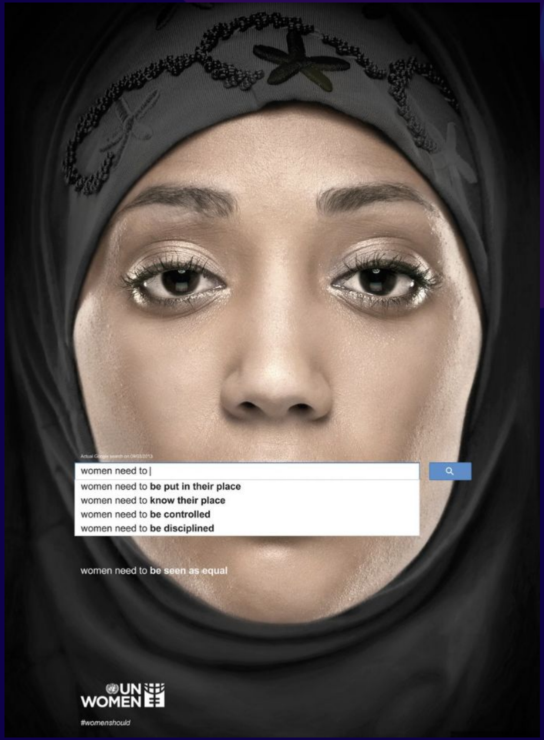
Global standards and regulations are imposed to regulate risks posed by AI.

The deeper the models are; the harder it is to inspect their predictions → need trustworthiness not only performance

ML models might learn wrong correlations yielding discriminatory behavior.

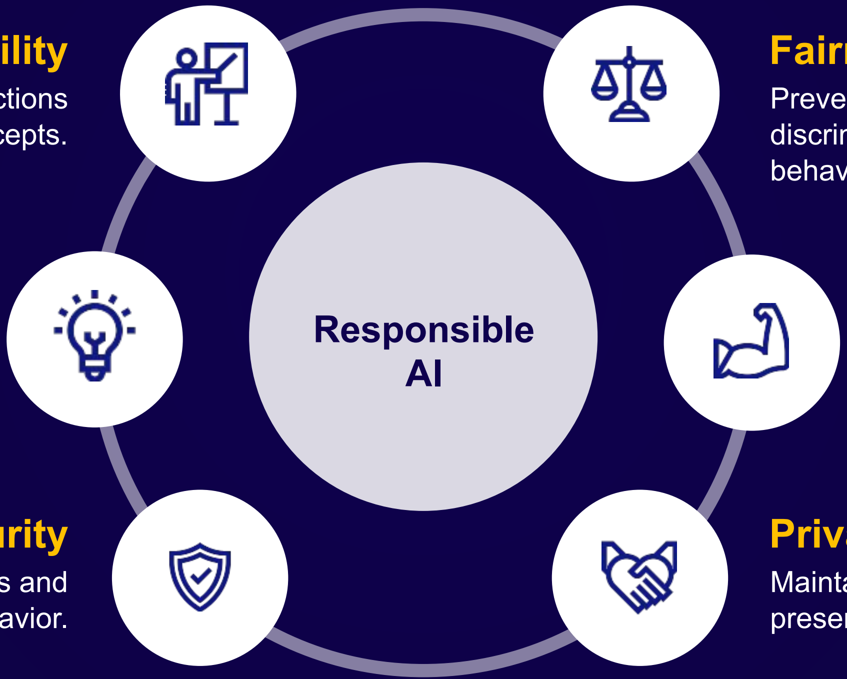
Discrimination can be prevented with explainability





- https://arxiv.org/abs/2211.09800
- https://www.dandad.org/awards/professional/2014/integrated-earned-media/23061/the-autocomplete-truth/

Explainability as key part of responsible AI development



Explainability
Explaining individual/global predictions in human understandable concepts.

Fairness
Preventing any learned bias, discrimination and stereotypical behaviors.

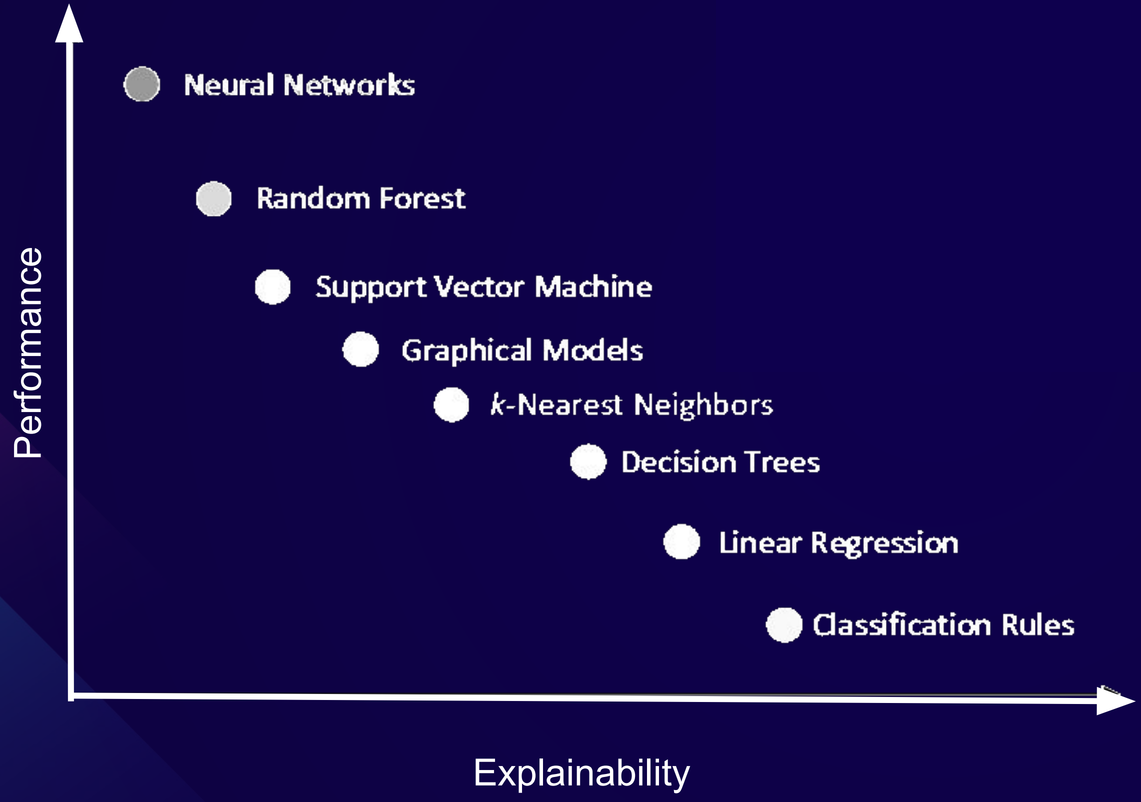
Transparency
Deciphering learned knowledge and inner workings of ML models.

Robustness
Generalizing to unseen cases and maintaining performance in uncontrolled settings.

Security
Protecting ML models and uncovering suspicious behavior.

Privacy
Maintaining confidentiality and preserving trust.

Accuracy-interpretability trade-off



Prior to ExAI, we could either have deep or **explainable** models!

- Shallow models are explainable by design | poor performance on complex tasks
- Neural networks increase complexity through non-linearity and back-propagation | high performance
- Practitioners need to compromise performance for interpretability and vice versa
- **ExAI** is the hope to have accurate yet interpretable models

Terminology

The *when* and the *what* of explainability

The *what* of explainability

- Data-centric: Local predictions
- Data-centric: Global predictions
- Network-centric: Learned knowledge
- Network-centric: Learning dynamics

The *what* of explainability

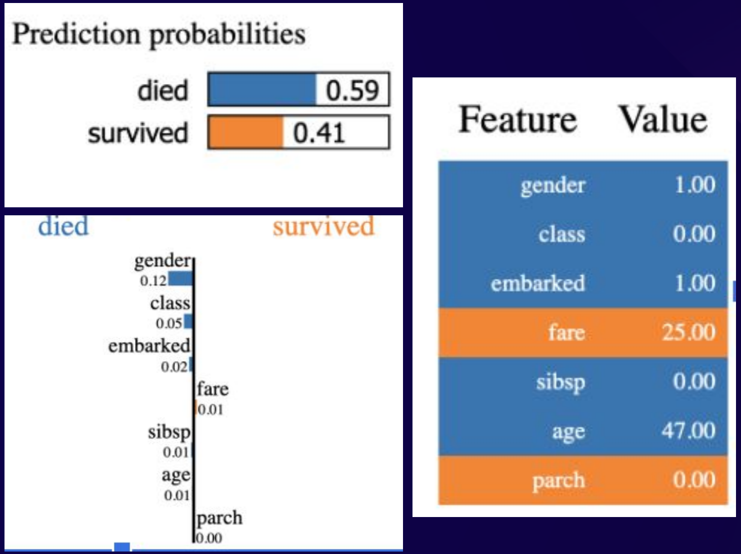
Data-centric: Local predictions

Input features crucial for a particular prediction. *why did the model predict y on input x?*

Data-centric: Global predictions

Network-centric: Learned knowledge

Network-centric: Learning dynamics



LIME explaining a specific prediction in the titanic dataset in terms of input features (individual contributions)

The *what* of explainability

Data-centric: Local predictions

Data-centric: Global predictions

Important features that are common for a label prediction. why does the model predict y in general?

Network-centric: Learned knowledge

Network-centric: Learning dynamics

Common face features, in terms of eigenface, for facial identification

https://towardsdatascience.com/eigenfaces-recovering-humans-from-ghosts-17606c328184?gi=c16b2d045094

The *what* of explainability

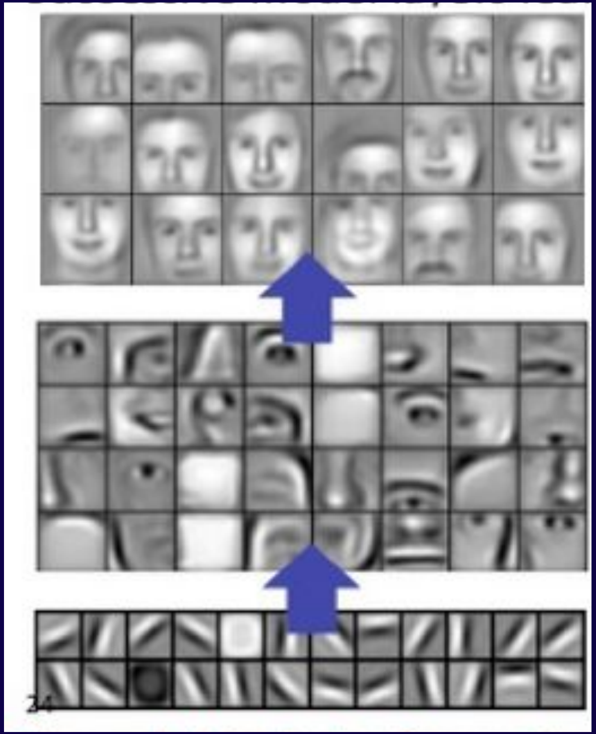
Data-centric: Local predictions

Data-centric: Global predictions

Network-centric: Learned knowledge

Interpretation of the internal state of a neuron or layer. What is the neuron learning?

Network-centric: Learning dynamics



Features learned by a face identification model across different layers

https://wiki.pathmind.com/neural-network

The *what* of explainability

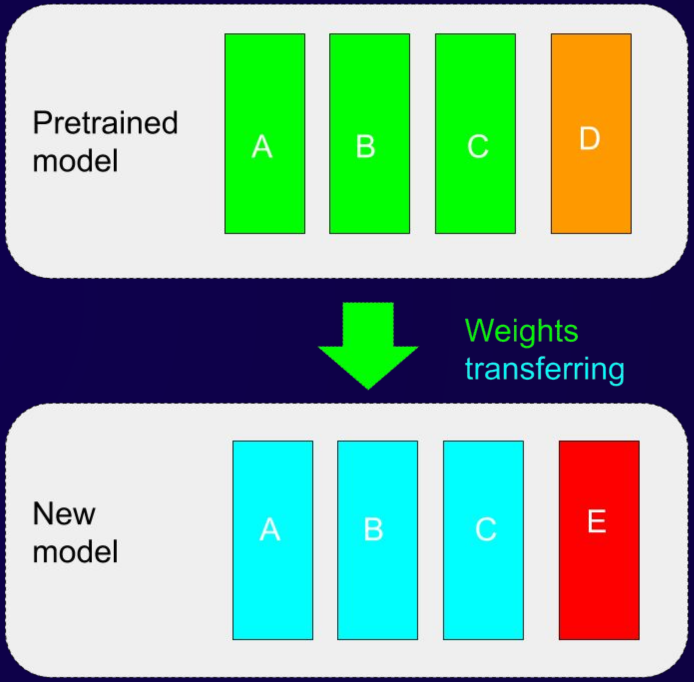
Data-centric: Local predictions

Data-centric: Global predictions

Network-centric: Learned knowledge

Network-centric: Learning dynamics

Interpretation of how different learning paradigms affect the performance. When is the-class specific information formed?



Study on the effect of layer freezing in transfer learning

The *when* of explainability

relative to training ML models

Post-hoc explainability

- Operates on trained networks with pre-defined architectures
- If no assumptions are made on the model → model-agnostic
- Otherwise → model-specific

- *Example: LIME and SHAP*

Inherent explainability

- Builds interpretable models from the ground up
- Supporting evidence while processing the input

- *Example: Tree-based (info gain), generative AI (min-max optimization)*

Post-hoc vs. inherent explainability

Concerns and considerations

01 Inherent explainability with deep models is **computationally heavy** – requires retraining.

02 Retraining to achieve inherent interpretability is not feasible with **privacy constraints** – access to training data

03 Post-hoc explanations do not present a perfect fidelity to the model being explained - **faithfulness concerns**

04 Post-hoc interpretability **exploits existing state-ofthe-art models** – training inherently explainable models do not

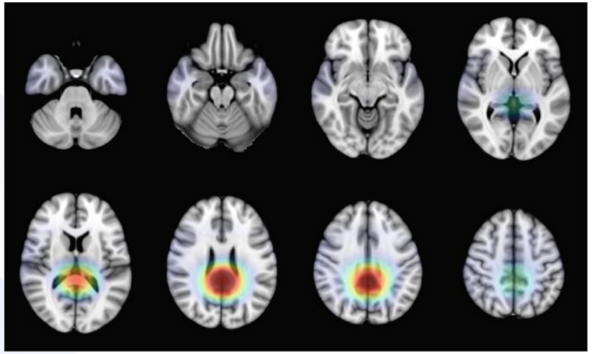
ExAI Methods

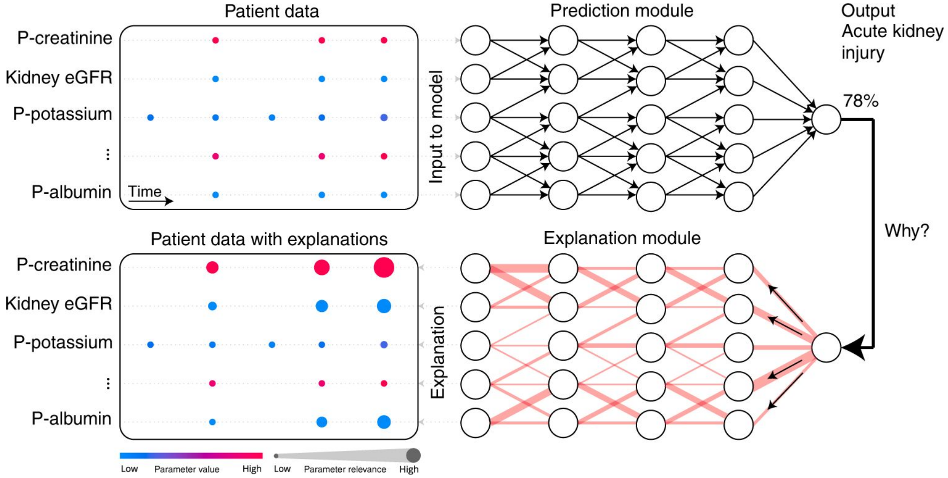
Local explanations

Gradient-based, saliency maps, LIME and SHAP

Gradient-based methods

- Given a network with N-dimensional input $x = \{x_i\}_{i=1}^{N} \in R^N$
- and a Cdimensional output $S(x) = \{S_c\}_{c=1}^{C} \in R^C$ where:
    - C is the total number of classes and
    - $S_c(x)$ represents the network's score function.
- We use the term "gradient" for $\frac{\partial S_C(x)}{\partial x}$ to capture the importance of each input feature for a specific output class c.

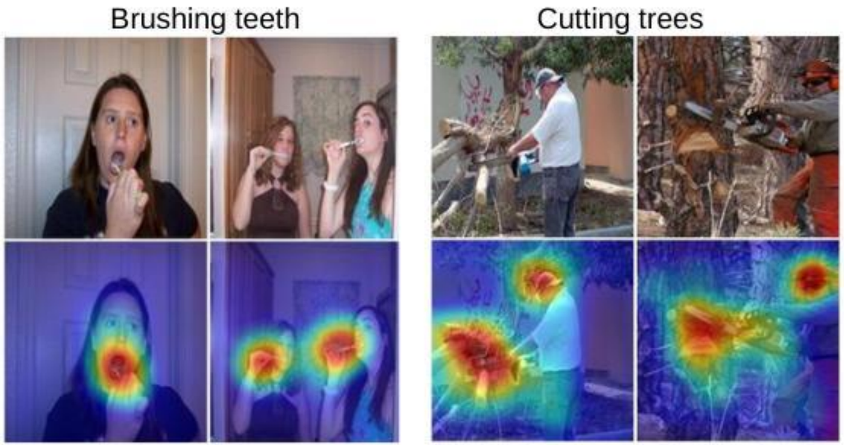




https://www.nature.com/articles/s41467-020-17431-x

Saliency maps

- Application of gradient-based methods to images
  - Each pixel is guarded as a feature
- Such techniques are used to explain visual content with relevant pixels being masked or highlighted (mostly with different intensities)
- Example: Class Activation Mapping (CAM) method and its variants (Grad-CAM++)



Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2921-2929).

Locally Interpretable Explanations (LIME)

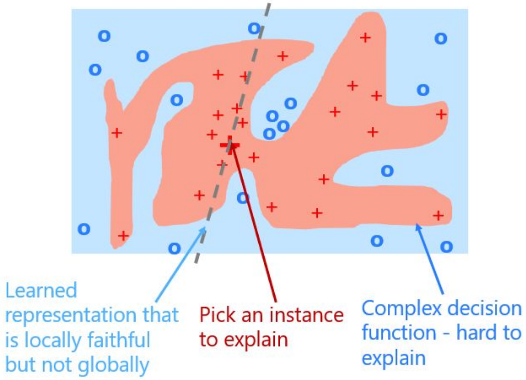
Algorithm 1 Sparse Linear Explanations using LIME

Require: Classifier f, Number of samples N
Require: Instance x, and its interpretable version x'
Require: Similarity kernel π_x, Length of explanation K
  $\mathcal{Z} \leftarrow \{\}$
  for $i \in \{1, 2, 3, ..., N\}$ **do**
    $z'_i \leftarrow sample\_around(x')$
    $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
  end for
  $w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$ $\triangleright$ with z'_i as features, $f(z)$ as target
  return w

Weights of linear/logistic regression → feature importance

Linear approximation

Local neighbors



Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

Demo

- ❏ LIME for tabular data
- ❏ LIME with image classifiers

SHAP: A unified approach to interpreting model predictions

- Cooperative game
- Players
- Gain

- Features
- Prediction
- Model

In game theory, assume you have a **cooperative game**. The gain is 19$. Shapley values specify how to split the gain across players.

In ExAI, SHAP quantifies the contribution that each feature brings to the prediction made by the model

Number of combinations before A

Number of combinations after A

A's contribution to this coalition

A B C
A C B
B A C
C A B
B C A
C B A

$$\phi_A = \sum_{S \subseteq F \setminus \{A\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f(S \cup \{A\}) - f(S) \right]$$

Summation is over these 4 coalition blocks

Total number of combinations/ coalitions

Coalition's value after A is added

Coalition's value before A is added

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Demo

SHAP documentation

Contrastive ExAI Methods

CARLA and CEnt

Contrastive explanations: Problem formulation

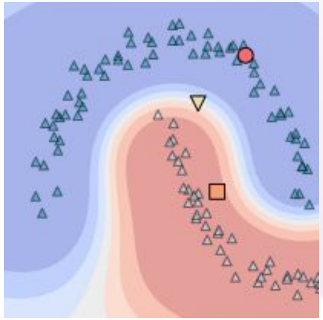
Explain a prediction by finding a contrastive example x' (close to x) such that:

$$\begin{array}{ll} \underset{x'}{\arg\min} & \delta(x, x') \\ \text{subject to} & f(x') = y_{\text{contrast}} \end{array}$$

Where f is classifier and δ is an edit distance

Challenges

1. Integrate immutability and semi-immutability in δ
   a. *Immutability: cannot change race, gender…*
   b. *Semi-immutable: has_degree can go in one direction*
2. Combine δ with a custom edit function
   a. *Relocating is twice as hard as changing jobs*
3. Accommodate for attainability and plausibility of x'

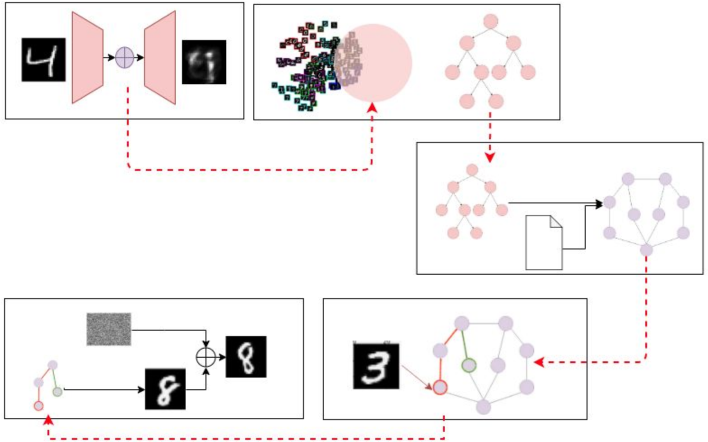


Our work: Entropy-based Contrastive Explanation - CEnt

Given a proximity measure π and a local approximator g, the optimization problem can be written as:

$$\begin{aligned} \arg\min_{x'} \quad & \delta(x, x') \\ \text{subject to} \quad & f(x') = y_{\text{contrast}} \end{aligned} \longrightarrow \begin{aligned} \arg\min_{x'} \quad & \mathcal{L}_{\pi_x}(f, g) + \lambda_1 R(g) + \lambda_2 \delta_g(x, x') \\ \text{subject to} \quad & f(x') = y_{\text{contrast}} \end{aligned}$$

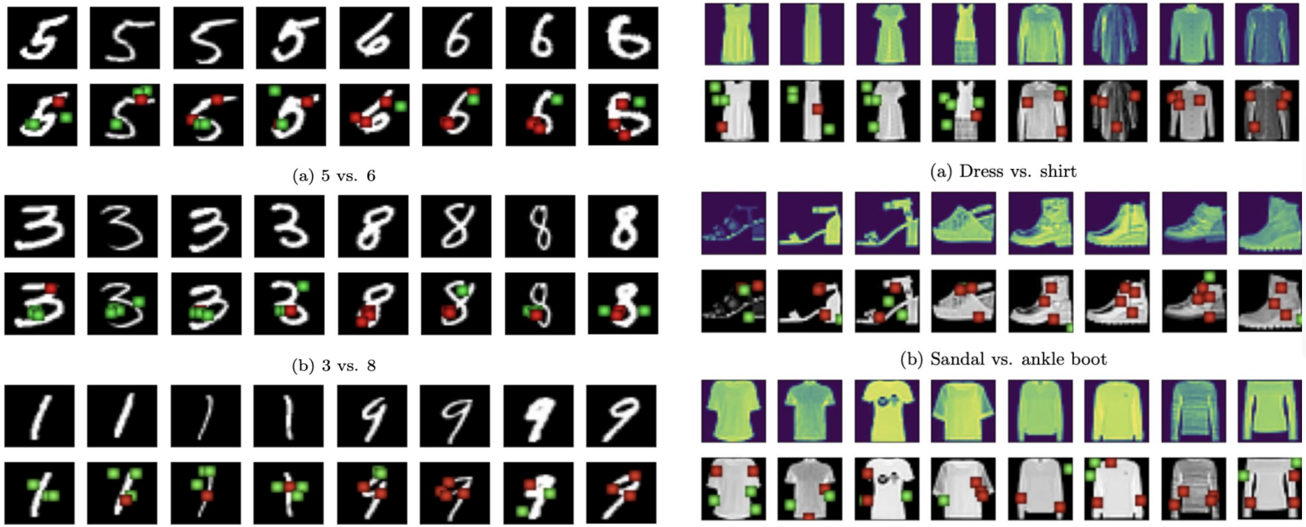
- $\mathcal{L}_{\pi}(f,g)$ is the approximation loss calculated on local neighbors of x
- $R(g)$ is a regularizer
- λ_1 and λ_2 are regularization parameters
- Approximator g is a decision tree → no need for assumptions on f → Highly interpretable with good approximation
  - Several leaves → diverse explanations
  - Immutability and semi-immutability can be modeled → challenge 1 ✅
  - Custom edit function can be integrated with leaves → challenge 2 ✅



Challenge 3: Accommodate for attainability and plausibility of x' → VAE distance

Our work: <u>Ent</u>ropy-based <u>C</u>ontrastive Explanation - CEnt

- f is a CNN trained on MNIST and Fashion MNIST datasets
- A visual contrast is a Gaussian kernel around a pixel whose intensity changed in x’
  - If intensity is amplified in x’ → pertinent negative (green)
  - Otherwise → pertinent positive (red)



(a) 5 vs. 6

(b) 3 vs. 8

(c) 1 vs. 9

(a) Dress vs. shirt

(b) Sandal vs. ankle boot

(c) T-shirt vs. pullover

Julia El Zini, Mohammad Mansour, and Mariette Awad. "CEnt: An Entropy-based Model-agnostic Explainability Framework to Contrast Classifiers' Decisions." *arXiv preprint arXiv:2301.07941* (2023).

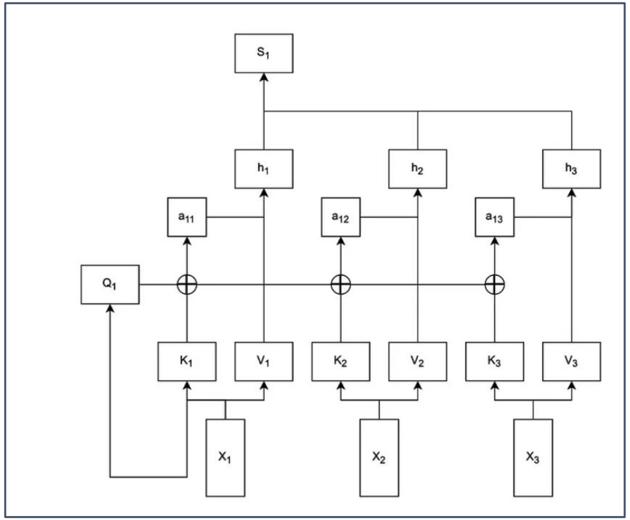
Demo

- [ ] CARLA on GitHub
- [ ] CEnt on GitHub

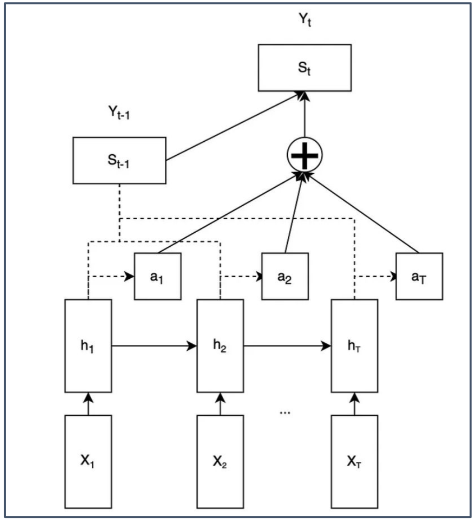
ExAI Methods

Learned Knowledge

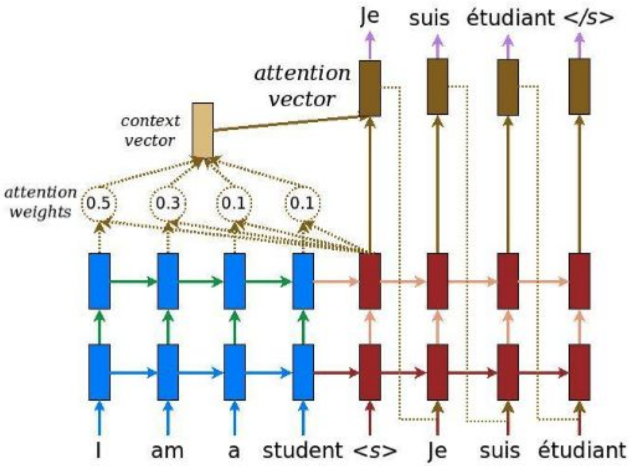
Language models (BERT, GPT...) are attention-based!



Encoder with self-attention mechanism replacing recurrence.
Each input t gets encoded into vector ht

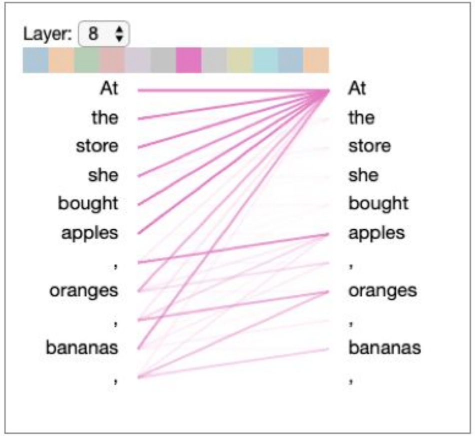


Decoder using attention to produce output Yt
from encoder-created vectors h

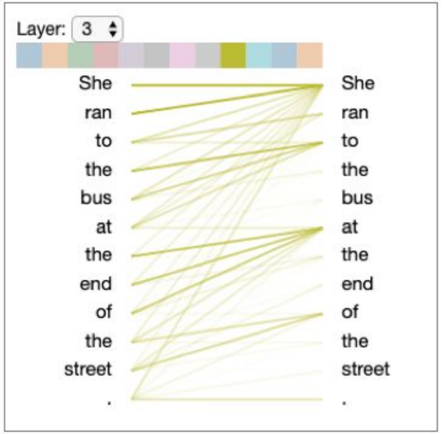


Attention refers to the ability of a transformer model to attend to different parts of **another sequence** when making predictions.

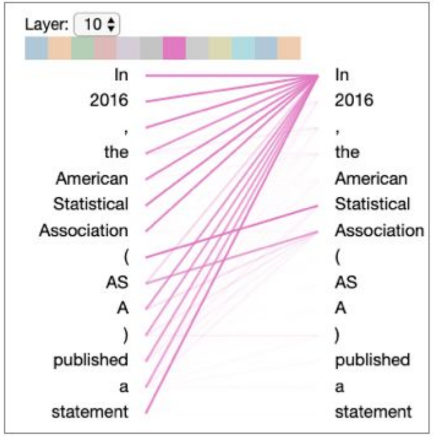
Understanding attention layers in GPT-2



Listing

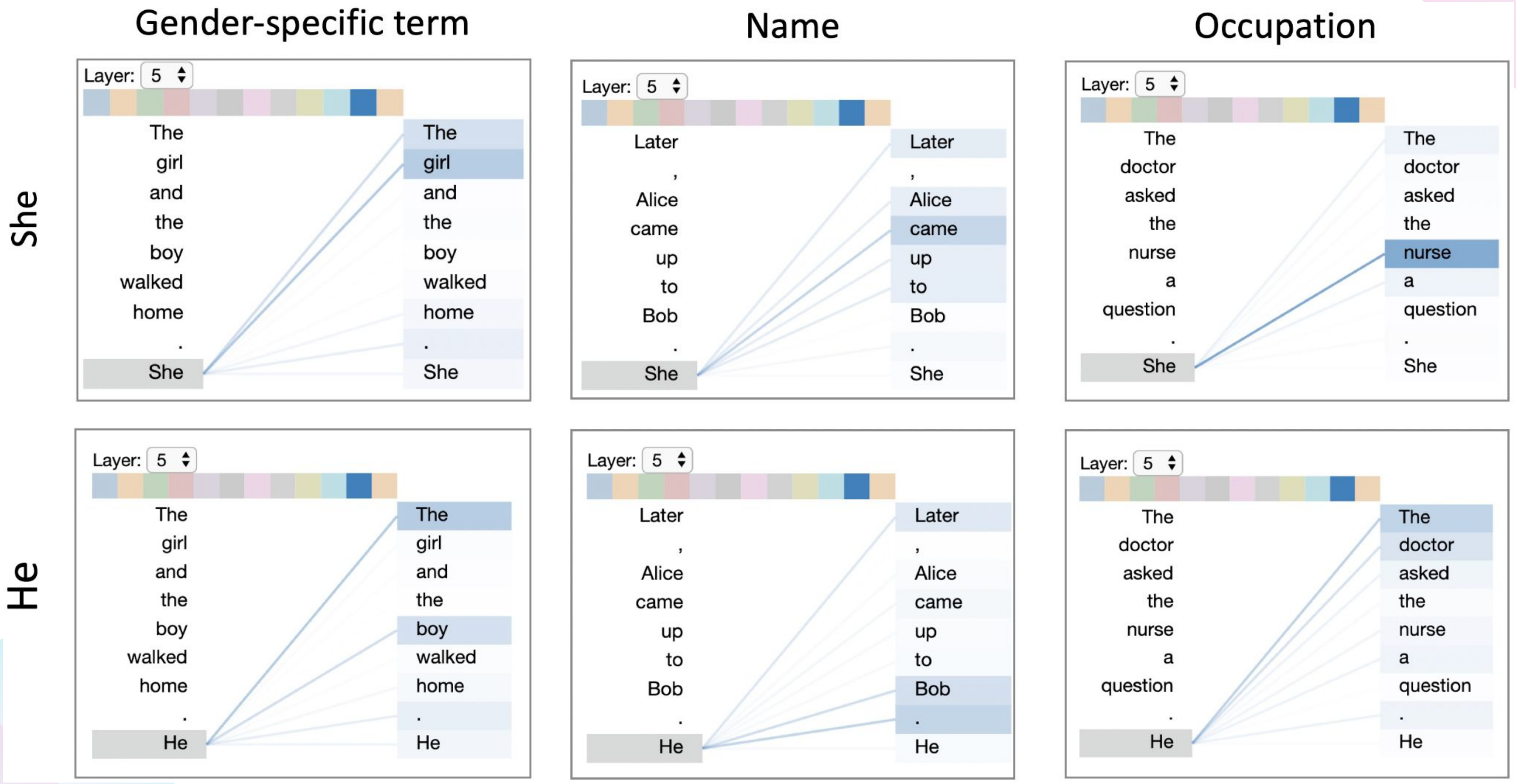


Preposition dependency resolution



Acronyms

Understanding attention layers in GPT-2: bias?



Vig, Jesse. "A multiscale visualization of attention in the transformer model." *arXiv preprint arXiv:1906.05714* (2019).

Demo

- BERTVis tutorial

Limitations & Research Directions

Challenges and limitations

Lifelong learning and changing models

AI models are constantly evolving as new data is collected and new algorithms are developed

New era of Generative AI

Such models are trained on huge datasets to learn complex patterns → not clear how to explain their outputs.

Incomplete information with transfer learning

Almost all deep models do not learn from scratch, pretraining intensifies opacity

Limited applicability to Deep Learning

Deep layer stacking hinders explainability especially with large language models

Risk of reverse-engineering and model theft

By understanding inner-workings of deep models, one can seamlessly alter the input to manipulate their behavior

Explanations as human- understandable concepts

Explanations can generated in forms of neuron activations which does not necessarily map to an understandable concept

Research Directions

01.

Explaining Generative AI

After the revolution caused by such models, we need to invest in explaining their learned knowledge, understanding their vulnerabilities and making them bias-free.

02.

Evaluating ExAI methods

With the increase of number of ExAI methods, researchers are collaborating to create common evaluation frameworks and benchmarks.

03.

Enhancing human-machine interaction

If humans are enabled to work collaboratively with AI systems, explanations will be improved by integrating feedback and model adjustment.

04.

Addressing Reinforcement Learning (RL) challenges

it can be difficult to provide explanations for actions of an RL agent. Future research will focus on developing methods for explaining the decision-making process of RL agents.

Conclusion

- Introduced ExAI **Terminology**
  - Four modes of explanations: local, global, learned knowledge and learning paradigms
  - Two types of explanations: post-hoc and inherent
- Described the methodology of local prediction methods
  - gradient-based, saliency maps, LIME and SHAP
- Introduced contrastive explainability
  - CARLA framework and CEnt method
- Learned how large language models can be dissected

Take-home message

“

Explainable AI is not just a matter of transparency, but also of trust and responsible innovation.

Let's strive for a future where AI is not a black-box, but a tool that empowers human decision-making and enhances our collective well-being.

”

Thank you!

linkedin.com/in/juliaelzini

Scan this and let's connect!