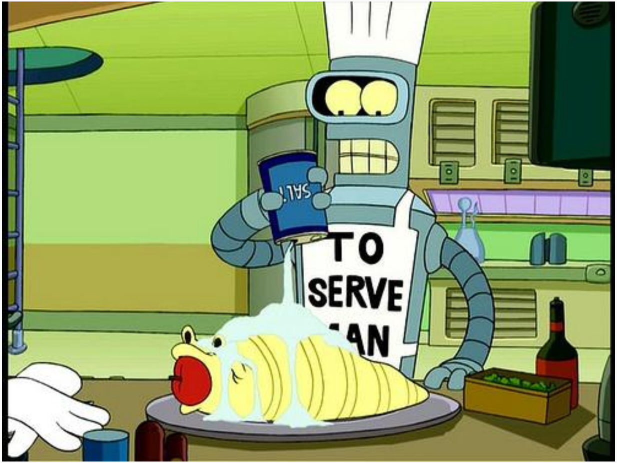
Toward Faithful Explanatory Active Learning with Self-explainable Neural Nets

Stefano Teso, KU Leuven
stefano.teso@cs.kuleuven.be

How can you trust a black-box ML model?



Black-box models can be whimsical and hard to control

How can you trust a human?

But other humans are black boxes too!

We have built-in facilities for determining trust into other agents (*theory of mind*). They rely on:

Understanding: trust involves understanding the other's beliefs & intentions; it depends on the perceived **competence**, **understandability**, **directability** [HJBU13]

→ This is the goal of **explainable machine learning**

Interaction: trust is updated **dynamically**; interactions let you build **expectations** [CDvW^{+}10]

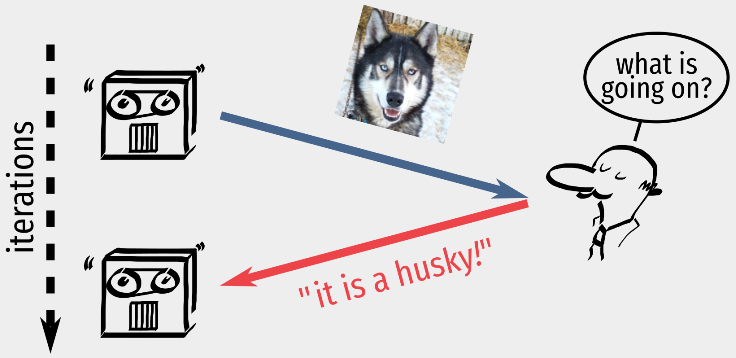
→ This is the goal of **interactive machine learning**

It seems like interactive learning would be a great tool for helping users to justifiably build (or revoke) trust into learned models!

It seems like interactive learning would be a great tool for helping users to justifiably build (or revoke) trust into learned models!

However it is often **opaque**...

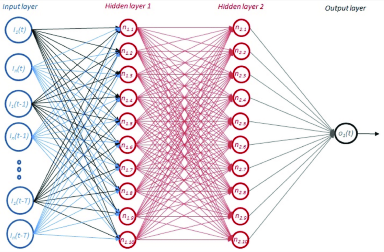
Active Learning is Opaque

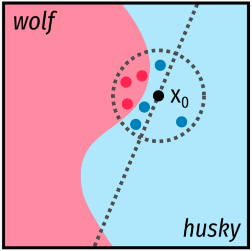


The user **a**) does not know the **model's beliefs**, **b**) cannot **affect** them directly, **c**) has no clue of what his **feedback does**!

Local Model-agnostic Explanations (LIME) [RSG16]

Given a black-box classifier $f : \mathcal{X} \to \{0,1\}$ and interpretable features $\phi_1, \ldots, \phi_m$, LIME explains a prediction $y_0 = f(x_0)$ by **approximating** f around x_0 with an interpretable classifier g_{x_0}:





Example: fitting a linear approximation

$$g_{x_0}(x) \approx \mathrm{sigmoid}(\textstyle\sum_{j=1}^{m} w_j \phi_j(x_0) + b)$$

w_j quantifies the **responsibility** of the jth feature ϕ_j

Husky or wolf?

Consider an example image classification task about discriminating between **husky dogs** and **wolves**

Husky or wolf? ...and why?

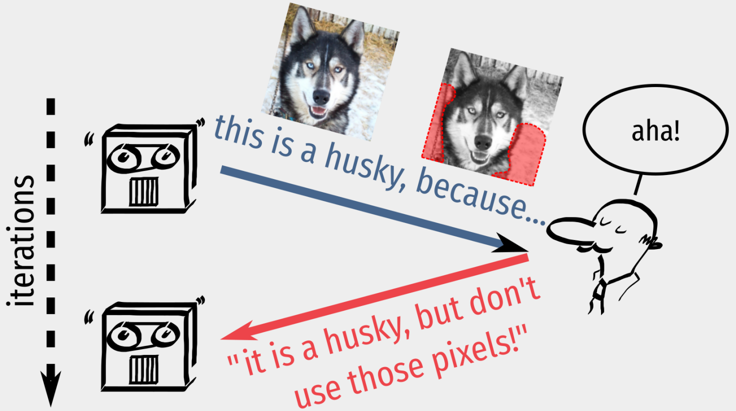
Let $\phi_1, \ldots, \phi_m$ refer to individual pixels

Local explanations allow to spot cases where the model is **right for the wrong reasons**

Remark: this does not suggest any way to fix the issue, though!

Explainable Active Learning

CAIPI(rinhas) turn LIME into trust



a) Explain predictions to user (**competence**, **understandability**),

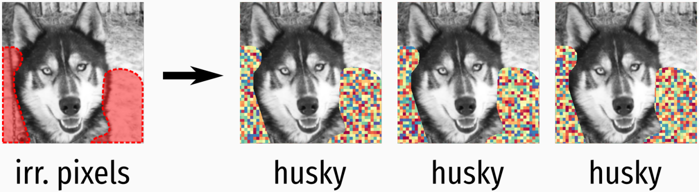
b) Allow user to correct explanations (**directability**)

What is an explanation correction?

1 – The user's correction indicates the **false positive** segments

2 – CAIPI converts the correction into **counterexamples**, e.g., by filling in random values while keeping the same label

Example: husky predicted right for the wrong reasons



Faithful Explainable Active Learning

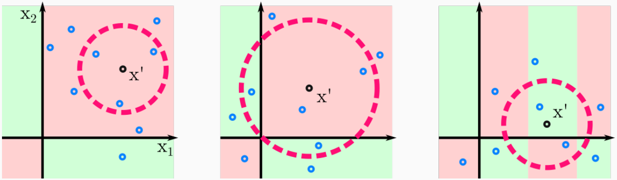
LIME can be unfaithful

Explaining $y^0 = f(\boldsymbol{x}^0)$ with $\phi_1, \ldots, \phi_m$ is **non-trivial**:

- Compute interpretable representation $\boldsymbol{\xi}^0 = \phi(x^0)$
- **Sample** $\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^s$ by perturbing $\boldsymbol{\xi}^0$ at random
- For each $i = 1, \ldots, s$:
  - **Project** $x^i = \phi^{-1}(\boldsymbol{\xi}^i)$
  - **Label** $y^i = f(x^i)$
  - **Weight** $\boldsymbol{\xi}^i$ with a kernel k that represents the neighborhood of $\boldsymbol{\xi}^0$
- **Fit** local model g^0 on $\{(\boldsymbol{\xi}^i, y^i)\}$ via *cost-sensitive learning*
- **Extract** an explanation from g^0

Many of these steps can introduce large amounts of noise

LIME can be unfaithful



Example: the circle is the **kernel** k, only points inside of it have substantial weight.

$\rightarrow$ *the samples fail to capture f regardless of how many*

Unfaithful explanations can confuse (and potentially also persuade) the user. They are contrary to the spirit and goal of explainable interactive learning!

(Unfaithfulness is an issue for other local explainers!)

Self-explainable Neural Networks (SENNs)

Linear models are often considered interpretable:

$$f(x) = \mathrm{sigmoid}\left(\sum_{j=1}^{m} w_j \phi(x)_j + b\right)$$

so long as $\boldsymbol{w}$ is sparse, ϕ interpretable.

Self-explainable Neural Networks (SENNs)

Linear models are often considered interpretable:

$$f(x) = \mathrm{sigmoid}(\sum_{j=1}^{m} w_j \phi(x)_j + b)$$

so long as $\boldsymbol{w}$ is sparse, ϕ interpretable.

SENNs extend linear models to be also deep:

$$f(x) = \mathrm{sigmoid}(\sum_{j=1}^{m} w(x)_j \phi(x)_j + b(x))$$

The "explanation" $\boldsymbol{w}(x)$ varies with x—but its regularized to vary *slowly* w.r.t. $\phi(x)$.

Calimocho = CAIPI - LIME + SENN

Given a dataset with instances x, labels y, **model explanations** z and their **corrections** $\bar{z}$, Calimocho learns SENNs using:

$$\min_f \lambda \ell_Y(f) + (1-\lambda)\ell_Z(f) + \alpha\Omega(f)$$

$$\ell_Y(f) = \sum \ell_Y(f(x), y) \qquad \text{\# label loss}$$

$$\ell_Z(f) = \sum \langle \boldsymbol{w}(x), \boldsymbol{z} - \bar{\boldsymbol{z}} \rangle \qquad \text{\# explanation loss}$$

Take-away message:

– LIME is approximate and slow, while SENNs are **exact** and fast

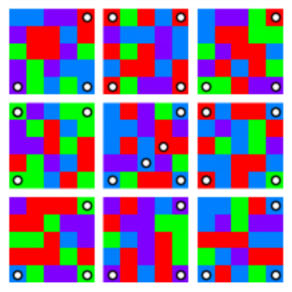
– CAIPI converts corrections into counterexamples, while Calimocho learns $\boldsymbol{w}(\boldsymbol{x})$ directly from explanation corrections

Experiment: Colors

5×5 images can be positive for **two reasons**:

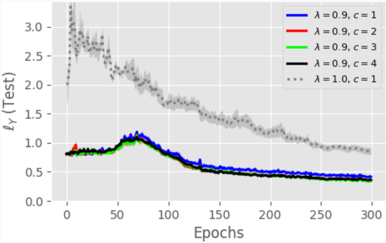
Rule 0: four corner pixels have the same colors

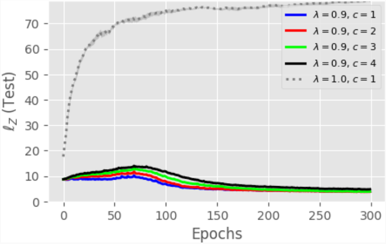
Rule 1: three top middle pixels have different colors



In training set either **both** rules hold or **none** does; in the test set only one of them applies [RHDV17]

Q1: Does CALI learn from corrections?

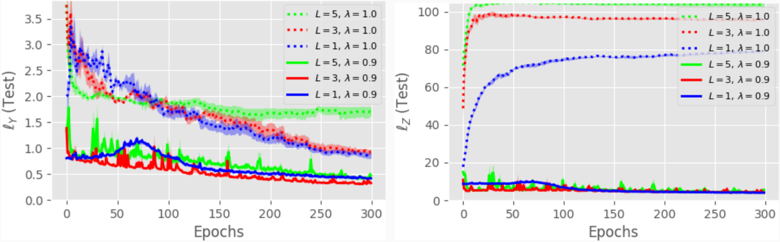




Label loss (left) and **explanation loss** (right) on the test set as more queries are asked (x axis)

Take-away: when no corrections are given (gray line), label loss decreases slowly and explanation loss **increases**!

Q2: Can CALI learn deeper SENNs?



Label loss (left) and **explanation loss** (right) on the test set as more queries are asked (x axis)

Take-away: explanation corrections can help enormously to learn deeper nets with L layers!

Take-away message

1. Trust $\approx$ Interaction + Explanations

2. Explainable active learning with Calimocho:
   – Explain predictions over time $\rightarrow$ mental model
   – Acquire explanation corrections $\rightarrow$ directability
   – Use self-explainable model $\rightarrow$ faithfulness

3. Preliminary experiments show promise:
   – Corrections keep explanations under control
   – Might be key in applying AL to deeper nets

4. Much more work needed!

Thank you! Questions?

Luke J Chang, Bradley B Doll, Mascha van't Wout, Michael J Frank, and Alan G Sanfey.
Seeing is believing: Trustworthiness as a dynamic belief.
Cognitive psychology, 61(2):87–105, 2010.

Robert R Hoffman, Matthew Johnson, Jeffrey M Bradshaw, and Al Underbrink.
Trust in automation.
IEEE Intelligent Systems, 28(1):84–88, 2013.

Isaac Lage, Andrew Ross, Samuel J Gershman, Been Kim, and Finale Doshi-Velez.
Human-in-the-loop interpretability prior.
In *Advances in Neural Information Processing Systems*, pages 10159–10168, 2018.

Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez.
Right for the right reasons: training differentiable models by constraining their explanations.
In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2662–2670. AAAI Press, 2017.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin.
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*, pages 1135–1144. ACM, 2016.