The Intriguing Properties of Model Explanations

Maruan Al-Shedivat Avinava Dubey Eric P. Xing

Machine Learning Department, Carnegie Mellon University

Abstract

Explanation is a simple model (*e.g.*, linear) that approximates the decision boundary of a complex model of interest (*e.g.*, deep neural net).

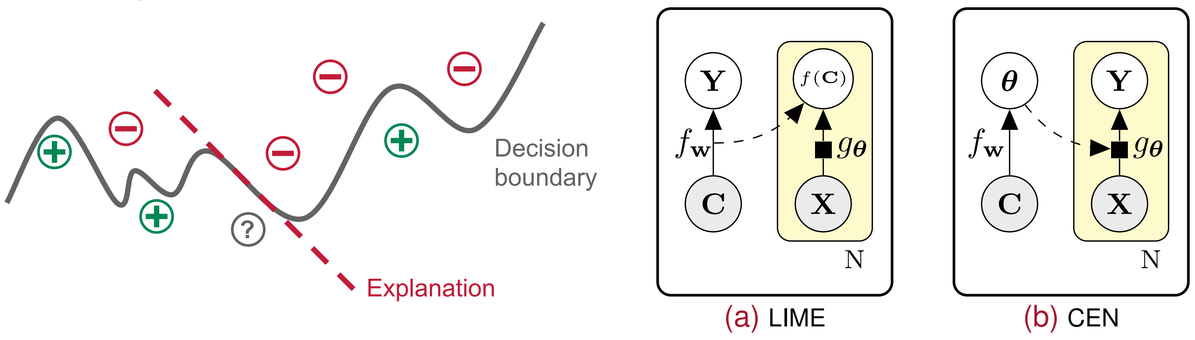
Questions:

Consistency. Explanations are as good as the features they use to explain predictions. How do feature selection and feature noise affect explanations?

Performance. When explanation is a part of the learning and prediction process, how does that affect performance of the predictive model?

Insights. What insights we can gain by visualizing and inspecting explanations?

1. Background



(a) LIME (b) CEN

Figure 1: (a) Local interpretable model-agnostic explanations (LIME). (b) Contextual explanation networks.

Data representations:

- $\mathbf{C}$: low-level or unstructured features (e.g., text, image pixels, sensory inputs, etc.)
- $\mathbf{X}$: high-level or human-interpretable features (e.g., categorical variables).

Constructing explanations *post-hoc* [Ribeiro et al., 2016]

$$\text{model: } f : \mathbf{C} \mapsto \mathbf{Y}, \quad \text{explanation: } g_{\mathbf{c}} := \operatorname*{argmin}_{g \in \mathcal{G}} \mathcal{L}(f, g, \pi_{\mathbf{c}}) + \Omega(g)$$

where

- $\mathcal{L}(f, g, \pi_{\mathbf{c}})$ – loss that measures how well g approximates f in the neighborhood, $\pi_{\mathbf{c}}$
- Linear explanation: $g_{\mathbf{c}}(\mathbf{x}) := b_{\mathbf{c}} + \boldsymbol{\theta}_{\mathbf{c}} \cdot \mathbf{x}$, parametrized by $(b_{\mathbf{c}}, \boldsymbol{\theta}_{\mathbf{c}})$
- $\Omega(g)$ – explanation complexity penalty (typically, L_1)

2. Contextual Explanation Networks (CENs)

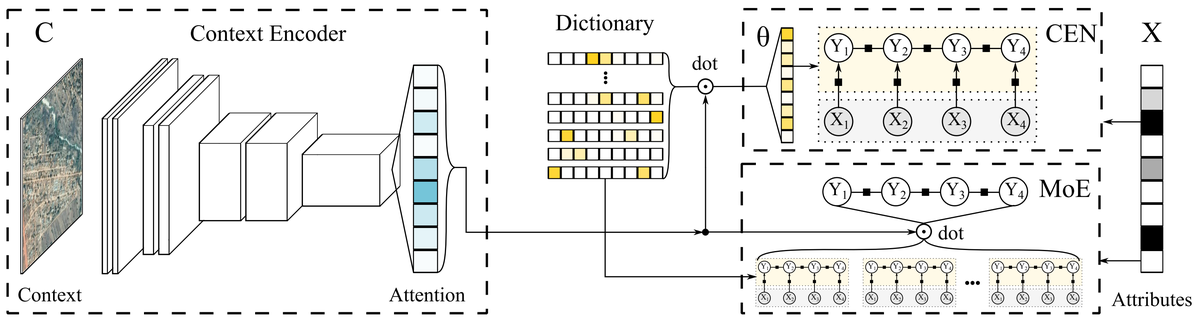


Figure 2: An example of CEN and MoE architectures with CNN context-encoder and structured explanations.

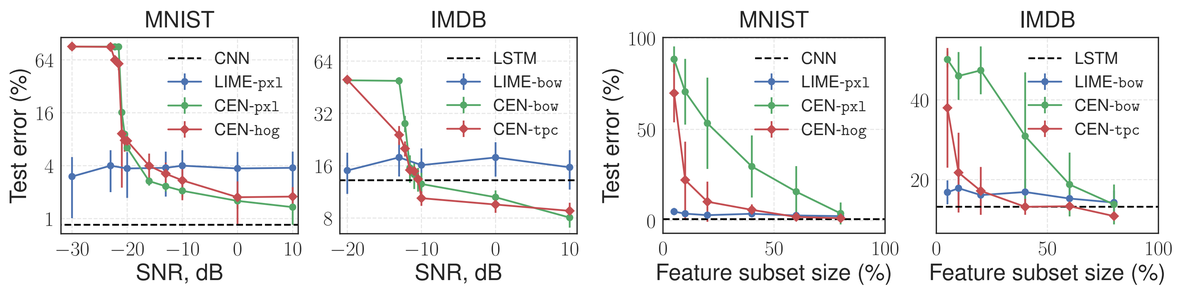
Learning to predict and explain jointly [our work]

$$\text{model: } f(\mathbf{c}, \mathbf{x}) := g_{\mathbf{c}}(\mathbf{x}), \quad \text{explanation: } g_{\mathbf{c}} := \mathrm{enc}(\mathbf{c})$$

where

- Linear explanation: $g_{\mathbf{c}}(\mathbf{x}) := b + \boldsymbol{\theta}_{\mathbf{c}} \cdot \mathbf{x}$, parametrized by $\boldsymbol{\theta} := (b, \boldsymbol{\theta}_{\mathbf{c}})$
- $\mathrm{enc}(\mathbf{c})$ – encoder that generates context-specific explanations, $g_{\mathbf{c}}$
- $g_{\mathbf{c}}$ are applied to $\mathbf{x}$ to make a prediction

3. Consistency of Explanations



(a) Noisy interpretable features (b) Incomplete interpretable features

Figure 3: Predictive performance of LIME- and CEN-generated explanations applied to corrupted features.

Setup

- $\mathbf{X}$ is subsampled or corrupted with additive noise. $\mathbf{C}$ is kept the same.
- LIME and CEN produce explanations using incomplete/corrupted features.
- This simulates a real-life situation when interpretable features are selected imperfectly.

Takeaways

- **LIME overfits explanations.** Despite the loss of information, it almost perfectly approximates the decision boundary of a deep net. *Can we trust such explanations?*
- **CEN-generated explanations are consistent.** Performance of CENs is proportional to the quality of the interpretable features used by explanations.

4. Explanations as a Regularizer

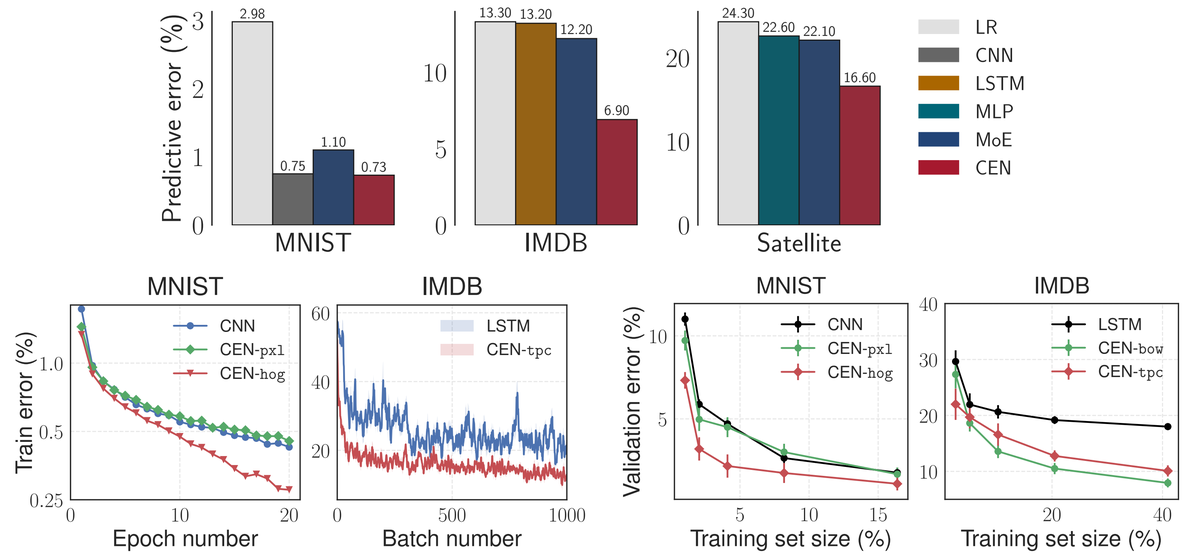


Figure 4: **Upper:** Predictive errors for the baselines, MoE, and CEN. **Lower:** Training error vs. iteration (epoch or batch) and validation error for models trained on random subsets of data of different sizes.

Takeaways

- **Large data regime.** When the data is abundant, prediction by explanation does not hurt the performance (CENs match the performance of vanilla deep nets).
- **Low data regime.** When the data is scarce, explanations can play the role of a good regularizer and boost performance.

5. Visualization & Insights

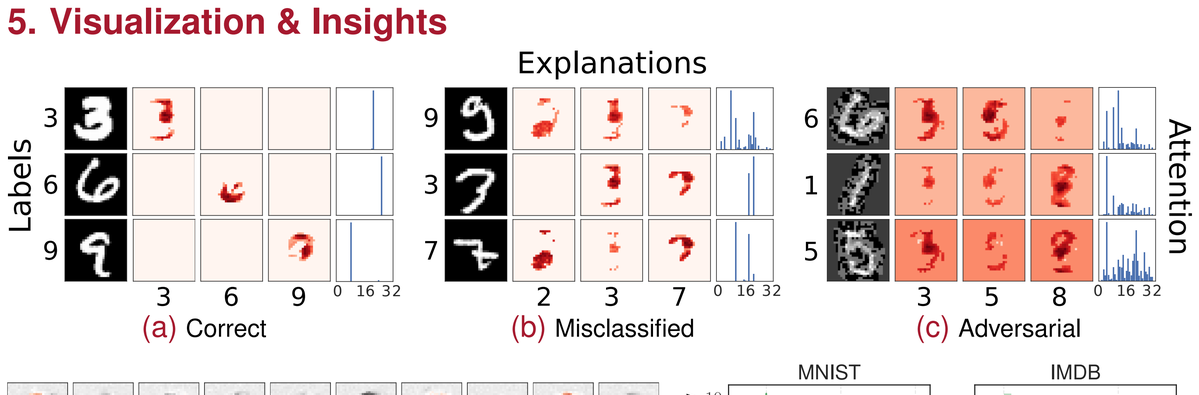


Figure 5: Explanations generated by CEN for the 3 top classes and the corresponding attention vectors for (a) correctly classified, (b) misclassified, and (c) adversarially constructed images. (d) Elements from the learned 32-element dictionary that correspond to different writing styles of 0 digits. (e) Histogram of the attention entropy for correctly and incorrectly classified instances.

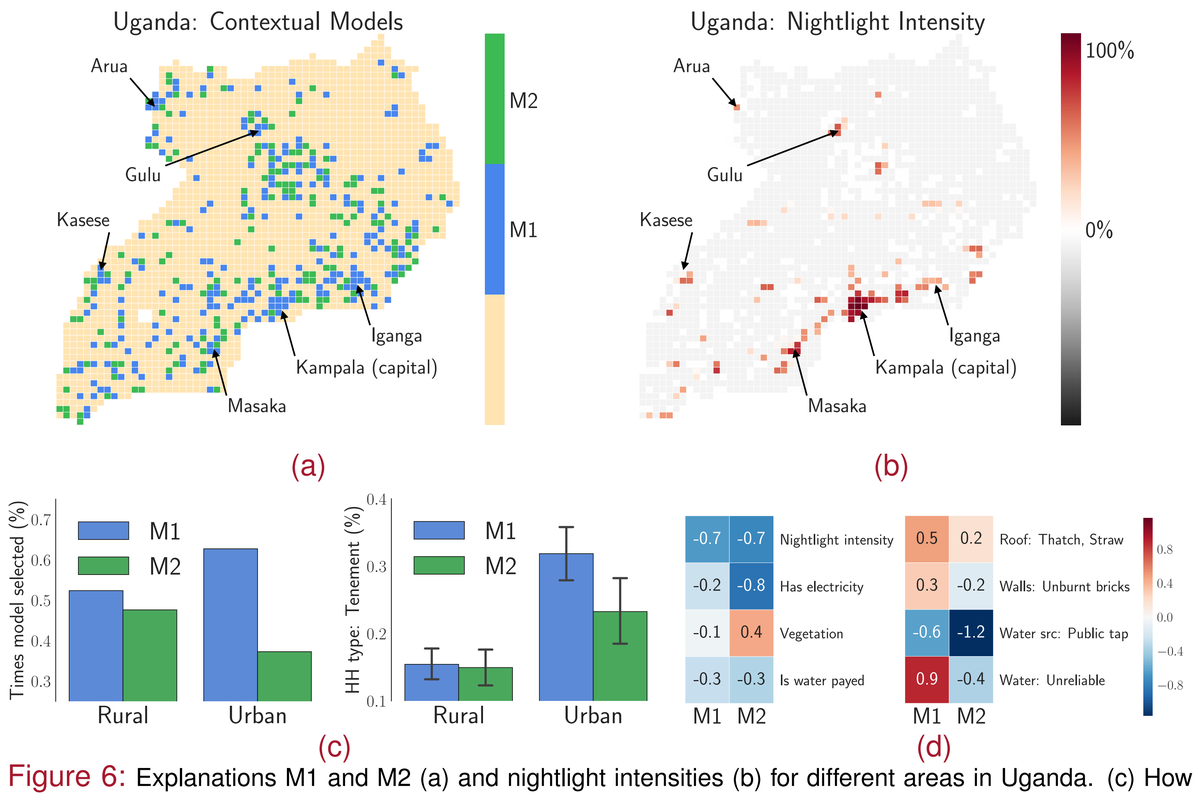


Figure 6: Explanations M1 and M2 (a) and nightlight intensities (b) for different areas in Uganda. (c) How frequently M1 and M2 are selected for rural/urban areas (left) and the average proportion of Tenement-type households in an urban/rural area for which M1 or M2 was selected (right). (d) Weights given to the survey features by the two explanations (M1 and M2) discovered by CEN.

Summary

Consistency. Be careful using post-hoc explanations as they can be silently misleading when the interpretable features are selected imperfectly. Learning to predict and to explain jointly helps to detect such problems.

Performance. Explanations can regularize the model and boost performance.

Insights. Visualizing explanations helps understanding the patterns in the data used by the model to make predictions.

Full paper: http://arxiv.org/abs/1705.10301

Interpretable ML Symposium, NIPS 2017, Long Beach, USA

Correspondence: {alshedivat, akdubey, epxing}@cs.cmu.edu