Explainable Machine Learning in NLP: Methods and Evaluation

Peter Hase
UNC Chapel Hill
peter@cs.unc.edu

Collaborators: Mohit Bansal, Shiyue Zhang, Harry Xie, Han Guo, Nazneen Rajani, Caiming Xiong, Owen Shen, and many others

This talk is based on...

- Four recent papers
  - The Out-of-Distribution Problem in Explainability and Search Methods for Feature Importance Explanations (2021)
  - FastIF: Scalable Influence Functions for Efficient Model Interpretation and Debugging (2021)
  - Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language? (2020)
  - Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? (2020)
- This blog post: “Opinions on Interpretable Machine Learning and 70 Summaries of Recent Papers” (2021)
- A lot of great work in the area from others

Outline

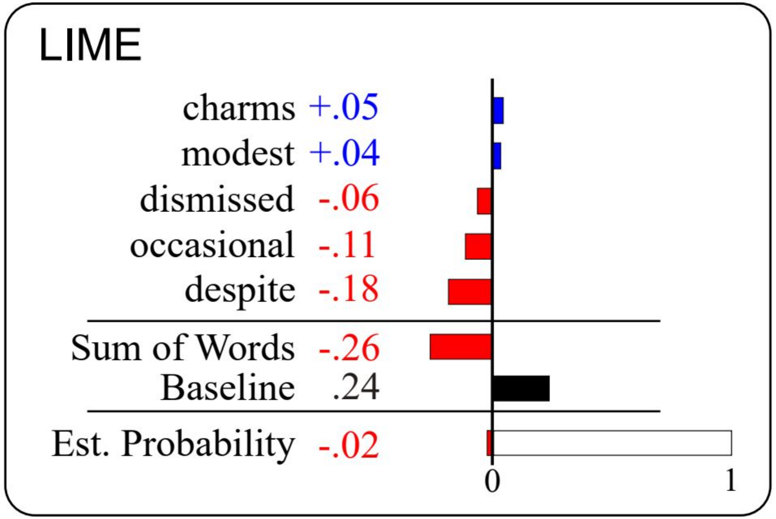
- Goals of explainable artificial intelligence (XAI)
  - Why build understanding of models?
- Measuring progress in XAI
  - Measuring model understanding, or explanation utility for downstream use cases
- Methods for explaining ML models
  - Discuss our work in the area
- Future directions for methods and evaluation procedures
  - What's hard about explaining NLP models?
  - Setting clear and achievable goals for XAI

Concrete Example

Input, Label, and Model Output

$x =$ Despite modest aspirations its occasional charms are not to be dismissed.

$y =$ Positive $\quad \hat{y} =$ Negative



(Ribeiro et al., 2016)

UNC NLP

XAI Goals

- There is a lot of healthy discussion about what XAI might be used for
- Scientific vs. instrumental uses
- Scientific:
  - Find a method for improving an expert's understanding of model behavior
  - Use it to create scientific knowledge about how models work
- Instrumental:
  - Verify model behavior is acceptable (correct, fair, etc.)
  - Fix undesirable model behavior (errors, unfair outputs, etc.)
  - Make more informed model deployment decisions
  - Calibrate people's trust in models (users, engineers, managers, other stakeholders)
  - Collaborate better with AI in human-AI teams
  - Improve our ability to design good tests for models (figure out right questions to ask)
  - etc.

XAI Goals

- Understanding can be instrumental, but not all goals require understanding
  - Verify model behavior is acceptable - do more testing
  - Fix undesirable model behavior - retrain with better data, better objective terms
  - Collaborate better with AI in human-AI teams - make a better GUI, more predictable system, etc.
- But I’m optimistic about usefulness of understanding, especially for goals like: “Improve our ability to design good tests for models”
  - Many input spaces are naturally very high dimensional and it’s hard to test every corner case
  - Narrow the space of inputs to be tested by figuring out where the model might plausibly fail
  - Hopefully uncover “unknown unknowns,” situations we didn’t even know we wanted to test for

XAI Evaluation

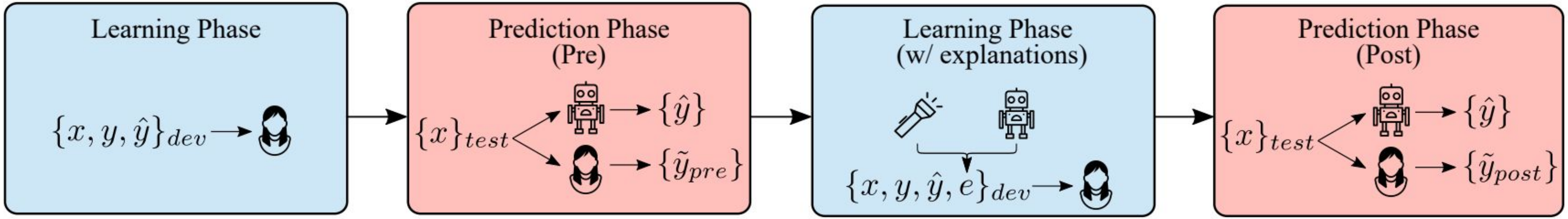
- Let’s focus on the scientific use: improving model understanding
- You understand a model when you have accurate knowledge of the causal chains that lead to model behavior over given inputs
  - Complete understanding is to know the complete causal chain behind all possible model behavior
  - Many levels of description, some better than others
- How to check for understanding?
  - Accurate causal models → accurate predictions of model behavior
  - Ask people what models will do for given inputs
- This is called *simulation* – we measure model *simulatability*
  - Accuracy of a specific explainee’s mental model
- What about faithfulness?
  - Faithful explanations contain accurate information about causal chains describing model behavior
  - So they should improve simulatability

XAI Evaluation

- We ran a human study measuring simulatability
  - Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?
- Give undergrads explanations from a given method (like LIME) and check if it improves their simulation accuracy, for neural models of text/tabular data
- Important experimental controls:
  - Separate explained instances from test instances
  - Evaluate the effect of explanations against a baseline of unexplained examples
  - Balance data by model correctness and model output
  - Force user predictions on all inputs (or penalize abstention)

XAI Evaluation

- Test 1: forward simulation



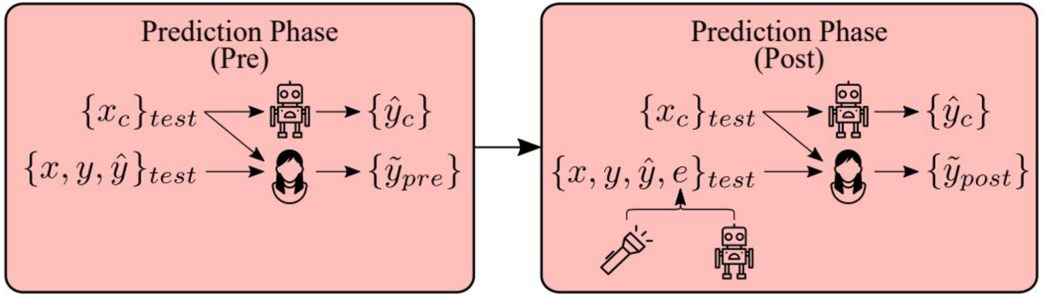
e : Explanation

$\hat{y}$: Model prediction

$\tilde{y}$: Human simulation

XAI Evaluation

- Test 2: counterfactual simulation



e : Explanation

$\hat{y}$: Model prediction

$\tilde{y}$: Human simulation

x_c : Counterfactual input

$\hat{y}_c$: Counterfactual model prediction

XAI Evaluation

- We tested four explanation methods
  - LIME (local linear model)
  - Anchors (probabilistic if-then rules)
  - Prototype explanations (explanation by similar example)
  - Counterfactual explanations (explanation by counterfactual example)
  - + combining them all
- Main results:
  - LIME helps with tabular data
  - Prototype explanations helped with counterfactual simulation
  - Did not get statistically significant results for *any other condition*
  - …including for every method on text data

UNC NLP

XAI Evaluation

- People did not even *realize* the methods weren’t helping
- We asked users to give scores of 1-7 for each explanation
  - “Does this explanation show me why the system thought what it did?”
  - Specifically during counterfactual simulation (explanations side-by-side with test data)
- Scores did not correlate with simulation accuracy!

UNC NLP

XAI Evaluation

- Results corroborated by follow-up studies:
  - Explain, Edit, and Understand: Rethinking User Study Design for Evaluating Model Explanations
  - What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods

UNC NLP

XAI Methods: Overview

- So how are people explaining models?
- There many, many taxonomies of explanation methods
- I’m going to go by families of approaches

XAI Methods: Overview

- Feature importance/attribution
- Approximator models
- Interpreting model weights and representations
- Finding influential training data
- Counterfactual explanations (w.r.t. input)
- Prototype/exemplar explanations
- Natural language explanations
- Unit testing
- “Don’t use black box models”

XAI Methods: Overview

- **Feature importance/attribution**
- Approximator models
- Interpreting model weights and representations
- **Finding influential training data**
- Counterfactual explanations (w.r.t. input)
- Prototype/exemplar explanations
- **Natural language explanations**
- Unit testing
- “Don’t use black box models”

XAI Methods: Feature Importance

- Based on: The Out-of-Distribution Problem in Explainability and Search Methods for Feature Importance Explanations (2021)
    - Metrics from DeYoung et al. (2020) ERASER paper

XAI Methods: Feature Importance

- One way of formalizing importance: *comprehensiveness*
  - If you remove important features, you expect model confidence to decline

$$\mathrm{Comp}(f, x, e) = f(x)_{\hat{y}} - f(\mathtt{Replace}(x, e))_{\hat{y}}$$

- Want to find explanations that indicate which features are most important
  - Using a fixed “budget” – can only select up to 5/10/20/50% of features
- *Sufficiency*: keeping important features maintains model confidence

XAI Methods: Feature Importance

- People use comp/suff metrics to evaluate LIME, Integrated Gradients, etc.
- But those methods don't optimize for comprehensiveness or sufficiency
- Let's optimize for those things directly!

$$\arg\max_{E} \frac{1}{|S|} \sum_{i=1}^{|S|} \mathrm{Suff}(f, x, e_i, s_i) \quad \text{s.t. } e_i \in \{0,1\}^d \text{ and } \sum_{d} e_i^{(d)} \leq \mathrm{ceiling}(s_i \cdot d)$$

Get a **set of explanations** (of varying sparsity)

Indicate features to keep/remove

Limit on # features (sparsity)

- Search for a solution with a *local, greedy search* starting from random point(s), called Parallel Local Search (PLS)

XAI Methods: Feature Importance

- We run experiments for BERT/RoBERTa models on six benchmark NLP datasets
- Keep compute budget fixed across methods
  - LIME uses forward passes
  - Integrated Gradients uses forward and backward passes
  - Parallel Local Search uses forward passes
- Compare with Anchors, which can be thought of as search method
- Propose a few other more complicated search methods and a random search

XAI Methods: Feature Importance

		Sufficiency ↓		Comprehensiveness ↑	
Dataset	Method	Standard Model	CT Model	Standard Model	CT Model
SNLI	LIME	20.00 (2.02)	27.08 (1.68)	82.18 (2.82)	75.34 (1.93)
	Int-Grad	43.76 (3.27)	32.91 (2.36)	34.01 (2.55)	43.22 (2.28)
	Anchors	11.93 (1.52)	30.96 (1.87)	55.72 (2.62)	48.86 (2.37)
	Gradient Search	17.55 (1.47)	33.98 (1.43)	53.15 (2.53)	49.36 (1.95)
	Taylor Search	6.91 (1.10)	28.00 (1.46)	73.20 (2.57)	66.76 (2.12)
	Ordered Search	-1.45 (0.93)	15.06 (1.37)	87.78 (2.41)	84.67 (1.61)
	Random Search	-1.54 (0.96)	15.38 (1.39)	87.36 (2.47)	84.63 (1.68)
	Parallel Local Search	**-1.65** (1.07)	**14.16** (1.38)	**87.95** (2.55)	**86.18** (1.45)

- **PLS is best in 21 of 24 conditions** (at *p*=.05), **by up to 17.6 points** over next best
- LIME is the best salience method, but it is best overall only once and is outperformed by Random Search on Sufficiency 9/10 times
- Search outperforms LIME 2/3 of the time with only 1/4 of the compute budget

UNC NLP

XAI Methods: Feature Importance

- If we care about an objective/metric, we should try to directly optimize for it
- Hopefully automatic metrics like suff/comp correlate with simulatability
- …but this might not be the case according to follow-up study (Fel et al., 2021)
- Want to always keep our ultimate goals in mind

XAI Methods: Feature Importance

- Other half of the paper is on the "Out of Distribution Problem"

$$\mathrm{Comp}(f, x, e) = f(x)_{\hat{y}} - f(\boxed{\texttt{Replace}}(x, e))_{\hat{y}}$$

Typically arbitrarily chosen:
1. Delete words
2. Replace tokens with MASK or UNK token
3. Set embedding to 0
4. Impute words
5. Etc.

These are all out of distribution to a model trained on real data!

UNC NLP

XAI Methods: Feature Importance

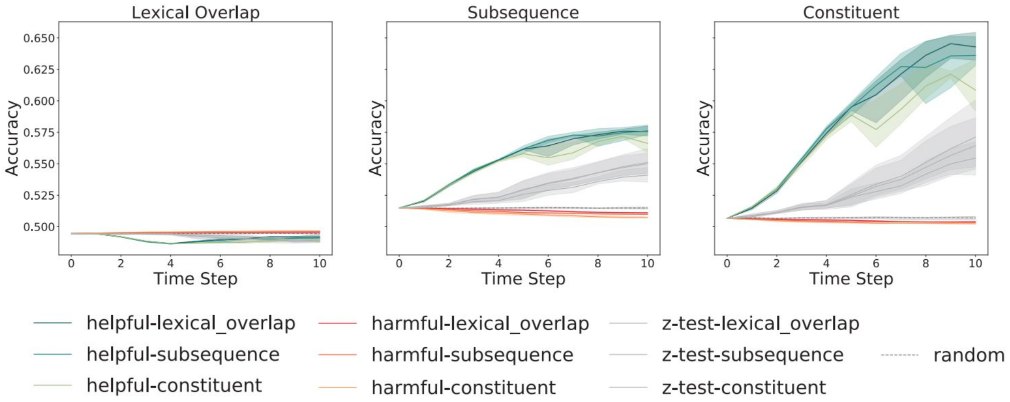
- What happens if you train models so that counterfactuals aren't OOD?
- Suff and Comp metrics look a lot worse
  - As Siwon Kim et al. suggest, importance of features is overestimated
  - "Interpretation of NLP models through input marginalization"
- We give an additional argument that explanations are *socially misaligned* when counterfactuals are OOD
  - "Aligning faithful interpretations with their social attribution" (Jacovi and Goldberg, 2021)

XAI Methods: Finding Influential Training Data

- Based on: FastIF: Scalable Influence Functions for Efficient Model Interpretation and Debugging (2021)
- The influence function estimates the effect of a training point on the model loss for a test point
- We'd like to find the most influential data points *out of the entire train set*
- This would take >2 hrs per test point for a BERT model on MNLI
- We speed up how long it takes to compute the influence function
- And we find a promising subset of train points to look through
- → less than 2 minutes per test point
- Now we can do a lot of things we couldn't before!

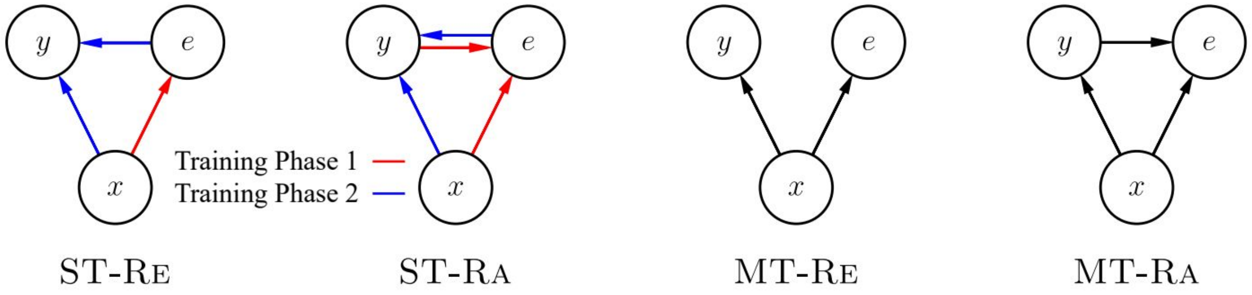
XAI Methods: Finding Influential Training Data

- Treat the "influential training data" as explanations, check simulatability
  - With another model as the explainee, we find that this is a good explanation
- Look at influence between data subsets
  - Identify training data that is usually helpful / usually harmful
- Fix model errors!
  - Short fine-tuning on a small set of positively influential data can improve model generalization!



XAI Methods: Natural Language Explanations

- Based on: Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language? (2020)
- Previous work had trained models to generate NL explanations for predictions
- But there had not been a faithfulness evaluation for these explanations
- We conduct a faithfulness evaluation for a few graphical models using eSNLI



XAI Methods: Natural Language Explanations

- We automate a simulation experiment using a model as the explainee
- Try to avoid separating data from their explanations in this experiment
- Introduce a “leakage-adjusted simulatability” (LAS) metric for this
  - When explanations leak the label, the explainee should accurately simulate the task model
  - When explanations do not leak the label, would be good if explainee accurate simulates task model
  - Take a raw average of the effect of explanations on simulation accuracy in these two cases
- Results:
  - Several kinds of explanations have a positive effect on simulation accuracy (by raw average across two cases)
  - Namely humans and rationalizing models
- A human simulation study would be a good follow-up to this

Explanations	LAS Score (CI)
HUMAN	4.31 (1.97)
MT-RE	-15.83 (1.81)
MT-RA	4.34 (4.12)
ST-RE	0.55 (0.87)
ST-RA	6.74 (4.53)

XAI Methods: Natural Language Explanations

- We also optimized explanations for simulatability in a message-passing game
- Specifically, optimized for our LAS metric using Gumbel-Softmax or RL
- Mixed results with this approach
- How do we efficiently optimize explanations for improving simulatability while avoiding *pragmatic drift*?
    - “Multi-agent Communication meets Natural Language: Synergies between Functional and Structural Language Learning” (Lazaridou et al., 2020)

XAI Future Directions (for NLP Methods)

- **Feature importance/attribution**
- Approximator models
- Interpreting model weights and representations
- **Finding influential training data**
- Counterfactual explanations (w.r.t. input)
- Prototype/exemplar explanations
- **Natural language explanations**
- Unit testing
- "Don't use black box models"

XAI Future Directions (for NLP Methods)

- Feature importance/attribution
    - Better feature spaces: want to manipulate abstract features of text
    - What are reasonable counterfactual (perturbation) distributions?
    - Better understanding of how proxy metrics connect to ultimate goals

XAI Future Directions (for NLP Methods)

- Finding influential training data
    - Would be interested in more work on fixing model errors with positively influential data
    - Can we find influential pretraining data?
    - Can this help build/refine pretraining datasets?

UNC NLP

XAI Future Directions (for NLP Methods)

- Natural language explanations
  - A simulation study would be good here
  - Should help to find similar data points that you'd expect explanations to help generalize across
  - How do we efficiently improve explanation quality while avoiding pragmatic drift?
  - Interactivity/dialogue is a natural next step to single-turn explanations

UNC NLP

XAI Future Directions (for NLP Methods)

- Should be clear about our goals when designing methods
- Scientific:
  - Find a method for improving an expert's understanding of model behavior
  - Use it to create scientific knowledge about how models work
- Instrumental:
  - Verify model behavior is acceptable (correct, fair, etc.)
  - Fix undesirable model behavior (errors, unfair outputs, etc.)
  - Make more informed model deployment decisions
  - Calibrate people's trust in models (users, engineers, managers, other stakeholders)
  - Collaborate better with AI in human-AI teams
  - Improve our ability to design good tests for models (figure out right questions to ask)
  - etc.
- Doesn't mean methods have to be specialized, but specialized evaluations may suggest they should be

Thank You!

Code: https://github.com/peterbhase/

Contact Info:

Peter Hase, UNC Chapel Hill
peter@cs.unc.edu
https://peterbhase.github.io

UNC NLP