Tencent AI Lab

Evaluation Metrics For Explainable AI (XAI)

2020.10.15 Paper Reading

zelongyang

Introduction

Tencent AI Lab

- Explainable AI (XAI) is now a widely discussed topic.
- One of the key challenges of XAI is about defining—and evaluating—what constitutes a quality interpretation.
- There are many aspects/dimensions of interpretability: such as readability, plausibility and faithfulness.

readability, human-interpretability

plausibility, persuasiveness

simulatability

faithfulness, accountability, transparency, fidelity, explainability

Two Papers

Tencent AI Lab

- **Paper 1**: Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? – ACL2020
- **Paper 2**: Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness? – ACL2020

Tencent AI Lab

Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?

Peter Hase and **Mohit Bansal**
UNC Chapel Hill
peter@cs.unc.edu, mbansal@cs.unc.edu

ACL 2020, Citation 5

A key aspect of model interpretability: simulatability

Tencent AI Lab

- **Human subject tests** are carried out on a key aspect of model interpretability, **simulatability**.
- A model is **simulatable** when **a person can predict its behavior on new inputs**.
- This paper performs **two kinds of simulation tests** involving **text and tabular data**, and evaluate **five explanations methods** (LIME, Anchor, Decision Boundary, a Prototype model, and a Composite approach).

Two human-subject tasks for simulatability

Tencent AI Lab

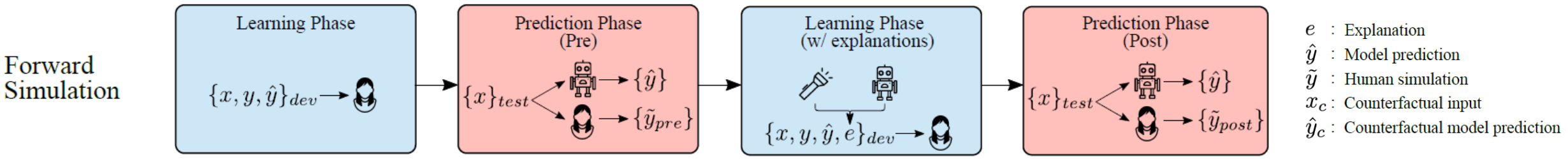
- A model is **simulatable** when a person can predict its behavior on new inputs. This property is especially useful since it indicates that a person understands why a model produces the outputs it does.

- The first task is termed **forward simulation**: given an **input** and an "**explanation**," users must predict what a model would output for the given input.

- The second is **counterfactual simulation**: users are given an **input**, a model's **output** for that input, and an "**explanation**" of that output, and then they must predict what the model will output **when given a perturbation of the original input**.

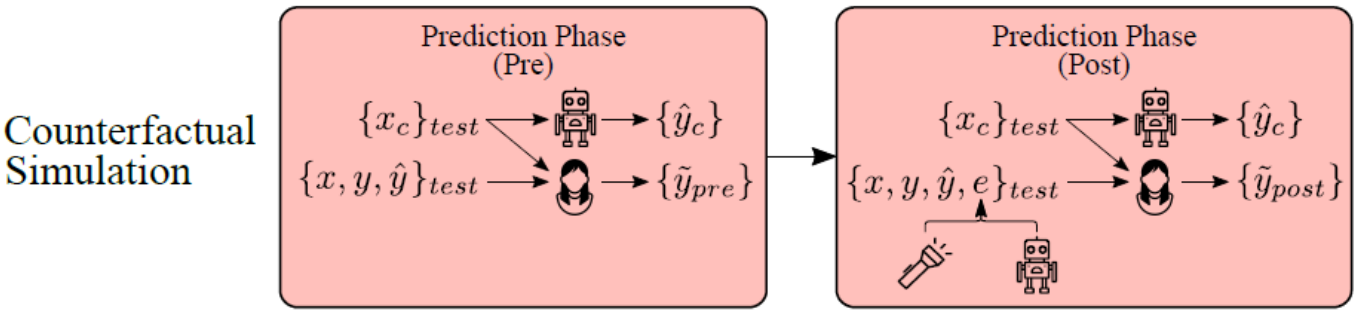
Task 1: Forward Simulation

Tencent AI Lab



- To begin, users are given **N examples with labels and model predictions** but no explanations.
- Then they are asked to predict the model output for **N new inputs**.
- Next, they return to the **same learning examples**, now **with explanations** included.
- Finally, they predict model behavior again on the same instances from the first prediction round.

Task 2: Counterfactual Simulation



e : Explanation
$\hat{y}$: Model prediction
$\tilde{y}$: Human simulation
x_c : Counterfactual input
$\hat{y}_c$: Counterfactual model prediction

- In the Pre Round, users are presented with **N inputs**, their **ground truth labels**, the **model's prediction**, and **a perturbation of the input**. Users then predict model behavior on the perturbations.

- In the Post Round, users are given **the same data**, but they are also equipped with **explanations of the model predictions for the original inputs**. Users then predict model behavior on the perturbations.

- Therefore, **any improvement in performance is attributable to the addition of explanations**.

Data Balancing

Tencent AI Lab

- For Forward Test:
  - The data is balanced so that users cannot succeed by guessing the true label: the tp, fp, tn, fn are equally represented in the inputs.
  - User predictions are forced on all inputs, so performance is not biased toward overly specific explanations.
- For Counterfactual Test:
  - The perturbations are sampled such that for any instance, there is a 50% chance that the perturbation receive the same prediction as the original input.

Datasets and Experimental Settings For Subjective Tests

Tencent AI Lab

- Dataset
  - Movie Review Excerpts: 10,662 reviews with binary sentiment labels.
  - *Adult* dataset: records of 15,682 individuals, with labels indicating whether their annual income is more than $50,000.

- Experimental Settings
  - The authors hired 32 trained undergraduates who had taken CS or statistics courses. Each user was randomly assigned to on of the ten dataset-method pairs.

  - In total, the authors collected 1103 forward test and 1063 counterfactual test responses in total.

Results

Tencent AI Lab

	Text					Tabular				
Method	n	Pre	Change	CI	p	n	Pre	Change	CI	p
User Avg.	1144	62.67	-	7.07	-	1022	70.74	-	6.96	-
LIME	190	-	0.99	9.58	.834	179	-	**11.25**	8.83	.014
Anchor	181	-	1.71	9.43	.704	215	-	5.01	8.58	.234
Prototype	223	-	3.68	9.67	.421	192	-	1.68	10.07	.711
DB	230	-	−1.93	13.25	.756	182	-	5.27	10.08	.271
Composite	320	-	3.80	11.09	.486	254	-	0.33	10.30	.952

Table 1: Change in user accuracies after being given explanations of model behavior, relative to the baseline performance (Pre). Data is grouped by domain. CI gives the 95% confidence interval, calculated by bootstrap using n user responses, and we bold results that are significant at a level of $p < .05$. LIME improves simulatability with tabular data. Other methods do not definitively improve simulatability in either domain.

	Forward Simulation					Counterfactual Simulation				
Method	n	Pre	Change	CI	p	n	Pre	Change	CI	p
User Avg.	1103	69.71	-	6.16	-	1063	63.13	-	7.87	-
LIME	190	-	5.70	9.05	.197	179	-	5.25	10.59	.309
Anchor	199	-	0.86	10.48	.869	197	-	5.66	7.91	.140
Prototype	223	-	−2.64	9.59	.566	192	-	**9.53**	8.55	.032
DB	205	-	−0.92	11.87	.876	207	-	2.48	11.62	.667
Composite	286	-	−2.07	8.51	.618	288	-	7.36	9.38	.122

Table 2: Change in user accuracies after being given explanations of model behavior, relative to the baseline performance (Pre). Data is grouped by simulation test type. CI gives the 95% confidence interval, calculated by bootstrap using n user responses. We bold results that are significant at the $p < .05$ level. Prototype explanations improve counterfactual simulatability, while other methods do not definitively improve simulatability for one test.

Results

Tencent AI Lab

- **LIME with tabular data** is the only setting where there is **definitive improvement in forward and counterfactual simulatability**. With no other method and data domain do we find a definitive improvement across tests.

- **The prototype method does reliably well on counterfactual simulation tests** in both data domains, though not forward tests. It may be that the explanations are helpful only when shown side by side with inputs.

- Even with combined explanations in the Composite method, we do not observe definitive effects on model simulatability.

Subjective Simulatability Ratings

Tencent AI Lab

	Text Ratings				Tabular Ratings			
Method	n	μ	CI	σ	n	μ	CI	σ
LIME	144	4.78	1.47	1.76	130	5.36	0.63	1.70
Anchor	133	3.86	0.59	1.79	175	4.99	0.71	1.38
Prototype	191	4.45	1.02	2.08	144	4.20	0.82	1.88
DB	224	3.85	0.60	1.81	144	4.61	1.14	1.86
Composite	240	4.47	0.58	1.70	192	5.10	1.04	1.42

Table 3: User simulatability ratings by data domain, on a scale of 1 to 7. The mean and standard deviation for ratings are given by μ and σ. The 95% confidence interval for the mean is given by CI, as calculated by bootstrap.

- The authors ask users to give **subjective judgments of the explanations**. They rate each method on a 7 point Likert scale, in response to the question, "**Does this explanation show me why the system thought what it did?**"
- The authors **measure how explanation ratings relate to user correctness in the Post phase of the counterfactual simulation test**. The authors do not find evidence that explanation ratings are predictive of user correctness.
- It seems that users rated explanations based on **quality** rather than **model correctness**.

Conclusion

Tencent AI Lab

- Many explanation methods may not noticeably help users understand how models will behave.
- Methods that are successful in one domain might not work equally well in another.
- Combining information from explanations does not result in overt improvements in simulatability.
- Users' rates for explanations are not good indicators for simulation correctness. It seems that users rated explanations based on quality rather than model correctness.

Tencent AI Lab

Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness?

Alon Jacovi
Bar Ilan University
alonjacovi@gmail.com

Yoav Goldberg
Bar Ilan University and Allen Institute for AI
yoav.goldberg@gmail.com

ACL 2020, short paper, citation 11

Two concepts: faithfulness vs. plausibility

Tencent AI Lab

- "**Plausibility**" refers to how convincing the interpretation is to humans. (persuasiveness)
- "**Faithfulness**" refers to how accurately it reflects the true reasoning process of the model. (fidelity, explainability)
- It is possible to satisfy one of these properties without the other.
- Despite the difference between the two criteria, many authors do not clearly make the distinction, and sometimes conflate the two.

A key aspect of model interpretability: faithfulness

Tencent AI Lab

- Current works are vague about the definition of interpretability and its evaluation. These conflations are harmful.
- Faithfulness: intuitively, we would like the provided interpretation to reflect **the true reasoning process** of the model when making a decision.
- Among all the aspects of interpretability, **faithfulness** should be defined and evaluated explicitly, and independently from **plausibility**.

An illustration

Tencent AI Lab

- Consider a textual system with explanations behaving in the following way: when the output is correct, the explanation consists of random content words; and when the output is incorrect, it consists of random punctuation marks.

- "**Faithfulness**" refers to how accurately it reflects the true reasoning process of the model. (fidelity, explainability)The former case appears more plausible than the latter case, but neither case is faithful.

It is dangerous to conflate Plausibility and Faithfulness

Tencent AI Lab

- Consider the case of **recidivism prediction**:
- A judge is exposed to a model's prediction and its interpretation, and the judge believes the interpretation to reflect the model's reasoning process.
- Since the **interpretation's faithfulness carries legal consequences**, a plausible but unfaithful interpretation may be the worst-case scenario.

Guidelines for Evaluating Faithfulness

Tencent AI Lab

- Be explicit in what you evaluate and do not conflate plausibility and faithfulness.

- Faithfulness evaluation **should not involve human-judgement** on the quality of interpretation.
  - Human cannot judge if an interpretation is faithful or not. If they understand the model, interpretation would be unnecessary.
  - **Human judgement measures plausibility, not faithfulness.**

- Faithfulness evaluation should not involve human-provided gold labels.

Three assumptions to define faithfulness

Tencent AI Lab

- **Assumption 1 (The Model Assumption)**. Two models will make the same predictions if and only if they use the same reasoning process.

- **Assumption 2 (The Prediction Assumption)**. On similar inputs, the model makes similar decisions if and only if its reasoning is similar.

- **Assumption 3 (The Linearity Assumption)**. Certain parts of the input are more important to the model reasoning than others. Moreover, the contributions of different parts of the input are independent from each other.

Faithfulness evaluation should not be in binary manner

Tencent AI Lab

- Currently, faithfulness evaluation are often done in a binary manner: whether an interpretation is strictly faithful or not.
- In other words, there is a clear trend of proof via counter-example, for various interpretation methods, that they are not globally faithful.
- An interpretation functions is an approximation of the model or decision's true reasoning process, so it **by definition loses information**. We should instead evaluate faithfulness on a more nuanced "**grayscale**" that allows interpretations to be useful even if they are not globally and definitively faithful.

Towards Better Faithfulness Criteria

Tencent AI Lab

- We must develop **formal definition and evaluation** for faithfulness that allows us the freedom to say when a method is sufficiently faithful to be useful in practice.

- The degree (as grayscale) of faithfulness should be evaluated **at the level of specific models and tasks**.

The findings of the two papers are consistent with each other

Tencent AI Lab

Paper 2 by Jacovi

- Faithfulness reflects the true reasoning process of the model when making a decision.
- Faithfulness should not be evaluated in a binary way. We should instead evaluate faithfulness on a more nuanced “grayscale”.
- Faithfulness should be evaluated across models and tasks.
- Faithfulness evaluation should not involve human-judgement on the quality of interpretation.

Paper 1 by Hase

- Simulatability indicates that a person understands why a model produces the output it does.
- In subjective simulatability rating, users are asked to give subjective judgments of the explanations on a 7 point Likert scale.
- Methods that are successful in one domain might not work equally well in another.
- The authors do not find evidence that explanation ratings are predictive of user correctness. It seems that users rated explanations are based on quality rather than model correctness.

- Thanks