# Week 8 - Benchmarking and Evaluation

Thursday, January 23, 2025 1:09 PM

## Motivation

1. How do we know whether an interpretability technique actually works? We've seen many proofs-of-concepts in the papers that we've read so far. But compared to other areas of machine learning, where a few widely used benchmark drive most progress, the evaluation of interpretability methods hasn't been so systematic.

2. Part of the reason is that interpretability means many at once, with many types of methodology lumped into the same literature. Moreover, even its more precise objectives are challenging to quantitatively evaluate -- whether a method is useful can be very context dependent.

3. This week's papers aim to formalize the evaluation of interpretability methods. The hope is that by clarifying which methods do well and when, it will be possible to better focus the research community's effort. And as we will see, formal evaluation highlights some major challenges -- many methods don't work, aren't understood properly, and can even be misleading!!

4. Formal evaluation relies on controlled, proxy tasks. There is a range in how realistic they are (or need to be). I think of it like how we evaluate whether a drug works. We don't start with full clinical trials, but instead many smaller computational and laboratory studies.

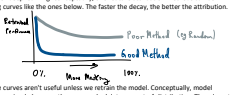
`Computational Experiment --> User Study (simplified tasks) --> User Study (real tasks)`

`[Easier, Less Realistic]` `[Harder, More Realistic]`

## Computational Evaluation

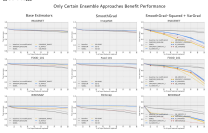
1. The first evaluation strategy we will discuss is related to the lazy retraining idea we discussed in our variable importance notes. As we saw there, we could measure importance by replacing a feature by its mean ("dropout") and measuring performance deterioration. But it was better to measure performance deterioration after giving the model a chance to at least partially adapt to the new, masked data ("retraining").

2. Hooker et al. (2019) bring this observation to the problem of computationally evaluating feature attribution methods for deep learning. Conceptually, two feature attribution methods can be compared by drawing curves like the ones below. The faster the decay, the better the attribution.

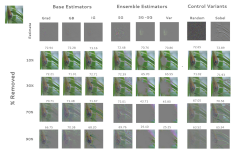



3. They argue that these curves aren't useful unless we retrain the model. Conceptually, model performance could drop simply because the new masked data are out-of-distribution. They do not look like the data used during training, so of course their performance is bad. Empirically, they use a simulation experiment with a linear model to show that opposite ranking strategies (largest to smallest coefficient and vice versa) can't be easily distinguished from one another, even though this is the most extreme difference between attribution methods that you could imagine.

4. The retraining approach is able to highlight the effectiveness of squared SmoothGrad and VarGrad explainability techniques. That these methods would outperform the others so clearly (and that the others would perform so poorly) comes as a surprise and highlights the value of a validated benchmarking strategy.




6. In retrospect, the effectiveness of these methods is suggested by the fact that progressive masking with these methods tends to remove entire objects, rather than scattered pixels like other approaches.




## User Observation

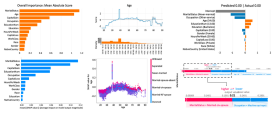
1. How are interpretability methods used in the "real world." It's one thing for authors to demonstrate value in a case study and another for the method to be used well by people who weren't involved in development. The idea of Kaur et al. (2020) is to run a user study to see how data scientists apply methods in a small data analysis task that they have control over. The results read a little like an absurdist comedy...

2. Their study design has two parts, a contextual inquiry and an online survey. For both parts, they prepare questions related to the following data science context:

   - **Data**: A version of the UCI Adult Income dataset that has been deliberately manipulated to introduce problems that are typical in data analysis. For example, they added features that are recoded versions of the original features. They also took 10% of samples in the high income class and replaced their ages with the mean (as if from an imputation).
   - **Models**: They train GAM and LightGBM models. Ideally they would use just one model, but it like their explanation packages didn't support that.
   - **Explanations**: Tutorials on GAM and SHAP-based explanations. The methods are chosen because they are from popular packages and support both local and global explanations. The authors are also developers of InterpretML, which implements GAM explanations.

   Participants for the inquiry (N = 11) were recruited from "a large technology company" (5 of the authors are from MSR). The survey participants (N = 253) were recruited from the internet.

3. In the inquiry, users were given a Jupyter notebook with the data, model, and explanations available. They were also shown the example explanations in the figure below (top row is GAM, bottom is SHAP). The authors didn't simply observe the participants, like in a "fly on the wall" user study -- they asked 10 questions about the data and models.




4. Many users seem to have been impressed by the visualizations without learning much from them. The authors had manipulated the data to see if the explanations helped users discover the quality issues. Instead of making participants more critical about whether the analysis made sense, they ended up overtrusting the results, simply because they were visualized.

> "Age 38 seems to have the highest positive influence on income based on the plot. Not sure why, but the explanation clearly shows it... makes sense." (P9, GAMs)
>
> tools to rationalize suspicious observations instead. After conducting several exploratory tests on the dataset, P8 said "Test of means says the same thing as SHAP about Age. All's good!" (P8, SHAP), and gave confidence ratings of 7 (extremely).

5. In the online survey, participants couldn't analyze the data, but they were shown some of the visualizations that were created in the inquiry. They also added a new manipulation, this time to the explanations. For half of the participants, the order of the feature importances was reversed. They wanted to see whether participants would be skeptical of explanations (or the underlying data/model) when they don't make any sense considering the problem context.

6. Fortunately, people did become more skeptical of the models/explanations when shown the reversed importances. E.g., a decrease from 4.8/7 to 3.8/7 for the "model is reasonable" question for the two GAM conditions (Table 2, left)

7. Remember that even the version without the reversed importances has many issues in the underlying data. Few people expressed any concerns about this, even when they saw some of the strange results in the explanation, like the income spike at age 38. The responses about model deployment were especially alarming:

> A subset of these participants also said that they would attempt to convince a customer that this was the right model to deploy by simply asking the customer to trust their judgment.
>
> In contrast, without (partially) accurate mental models of the tools, even the most experienced participants "don't see any red flags as confirmed by the explanations" (P57, GAM-Manipulated, 4 years of ML experience). Worse, in

Though this does raise some questions about what kinds of background these survey respondents really have.

8. Another finding from the survey is that few of the respondents (~ 5%) were able to accurately describe the meaning of the axes in the local explanation plots. Some were very far off, e.g., thinking that the importance measures represented labels in the original data. Yet, they seemed to have been impressed by the visualizations.

## Complementarity

1. In the past, most AI tasks were focused on automation, e.g. reading handwritten digits for USPS. But now, many AI applications are meant to be used with a human in the loop. Developers argue that the tools are not meant to replace people, and that AI assistance will make them more effective. For example, most (but not all...) people think radiologists will still be needed to make diagnoses and identify appropriate treatments. The AI is supposed to make them more accurate. Human and AI abilities are supposed to "complement" one another.

2. A lot of research today is aimed at understanding the extent to which these types of stories are true. Within this context, the focus of Bansal/Wu et al. (2021) is on the specific role of explanations. Are humans and machines better able to "collaborate" when models come with explanations? Their answer is basically, no, at least not with today's interpretability techniques.

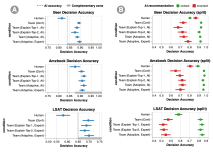
3. This answer contradicts a fair bit of prior research on the topic. They argue that the prior work is not reliable, because in all of those studies, the AI alone did better than the human-AI teams. The tasks were so straightforward that there was no potential for complementarity. In contrast, Bansal/Wu et al. (2021) deliberately pick 50 test set samples so model and human accuracy were comparable (~ 85%), where human accuracy was measured in a separate pilot experiment.

4. The study design has the following components:

   - **Tasks**: They consider three tasks. For two, the goal is to identify the sentiment of an online reviews (Amzbook, Beer). The third is to solve multiple choice problems from the LSAT.
   - **Explanation Method**: For the reviews datasets, they use LIME to highlight the words that are most important for each sample's prediction. For the LSAT problem, they wrote brief text explanations themselves and presented them to participants as if they were from an AI.
   - **Explanation Presentation**: For some participants, an explanation was only shown for the top predicted class. For others, explanations were shown for the top two predicted classes. An intermediate, *adaptive*, explanation was shown to others. This defaults to the top class but shows two when model confidence is low.

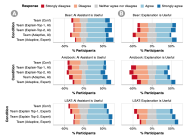
   As a negative control, some of the participants were assigned a version with only the model output and a confidence score, but no explanation. For each of the three tasks, they recruit a little over 500 mechanical turk workers.


5. The results highlight the potential for complementarity when the tasks are appropriately challenging. The team always did better than either the unassisted people or the AI alone (below, left). This seems to have been true across all the tasks. However, the explanations did not help! Their argument is that, at least in the Beer and LSAT tasks, the explanations led to inappropriate reliance on the AI, compared to simply showing confidence. They went along with the AI prediction, even when it was incorrect.




6. What's going on with Amzbook? They authors argue that the explanations highlighted some of the model's failures. Since this is a context where people can do relatively well on their own, it seems these explained failures (see the first figure above) led to people simply ignoring the AI assistance.

5. They followed up the task with a survey about the usefulness of the AI and explanations. People seem to think that the assistance was useful, but except for LSAT, the explanations were not. For LSAT, the explanations might be useful because either (i) the task is inherently more difficult and (ii) the explanations were generated manually, not using any interpretability algorithm.




6. Based on a qualitative survey, they identify the following collaboration strategies: Mostly follow AI (6%), AI as Prior Guide (47%) , AI as a Post Check (25%), Mostly Ignore AI (22%). They found that 42% of participants used the explanations.

## Takeaways

1. In some ways these studies are discouraging. The first shows that many explainability techniques don't work. The second shows that people don't understand them. The third shows that they don't improve task performance. Nonetheless, the purpose of evaluation is to objectively check the stories that might be popular in the community and to highlight the most promising areas for further study, and all the papers were successful from this point-of-view.

2. Some larger themes that emerge are:

   - Explanations shouldn't lead to blind trust. There are opportunities for designing tools (or trainings?) that "activate system 2" in the Kahneman and Tversky sense. It seems too easy to use interpretability outputs to validate snap judgments, when really they should be used as the starting point for more critical investigation.
   - Complementarity might be better achieved through coordination. Rather than having AI and people solve the same subtask together, larger problems can be broken into parts that are safely automated and others that require more in-depth human study.
   - Explaining a model/a model's predictions might be too narrow a focus for interpretability. We would like people who use models to maintain an appropriate level of skepticism and use their models wisely.