Interpretability Through Invertibility - The User Study (#48681)

Created: 09/30/2020 07:03 PM (PT)
Shared: 10/02/2020 08:08 AM (PT)

1) Have any data been collected for this study already?

No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?

This online study examines if Invertible Neural Network Interpolations can serve as an explanation technique for image classification. We investigate whether they allow MINT graduates to identify patterns and biases a neural network has learned to discriminate between two different types of abstract animals (“Sticky” and “Stretchy”).

Participants (N=60), recruited from Prolific, are randomly assigned to view two different visual explanation techniques (NNWI and Baseline). We will focus our analysis on the following Hypothesis:

H1: Studying the system’s predictions on the validation set (Baseline Explanation technique - referred to as Baseline) allows users to verify that the neural network (NN) is using the blocks spatial arrangement (Pattern 1) for its predictions of the abstract animals.
H2: Baseline does not allow users to detect the NN bias for colour (Pattern 2) and rotation (Pattern 3).
H3: Baseline does not allow users to detect the NN bias for colour (Pattern 2) and rotation (Pattern 3).
H4: Studying the system’s predictions with Invertible Neural Network Weigt Interpolations as explanations (referred to as NNWI) allows users to verify that NN is using Pattern 1.
H5: NNWI allows users to detect Pattern 2 and Pattern 3.
H6: NNWI allows users to verify that NN is neither using the background of the image (Pattern 4) nor the surface structure of objects (Pattern 5).
H7: NNWI allows users to detect Pattern 1 with higher confidence.
H8: NNWI allows users to reject Pattern 4 and Pattern 5 with higher confidence
H9: NNWI allows users to detect Pattern 2 and Pattern 3 with higher confidence.
H10: NNWI leads users to be more confident in the matureness of the system.
H11: Users are more satisfied with NNWI as explanations.

We will evaluate all Hypotheses by measuring the users' agreement to corresponding statements (7-point Likert Scale rating). The average Explanation Satisfaction Scale rating will measure explanation satisfaction. All comparative analyses refer to the Baseline and are tested for significance with non-parametric tests. The second author, Martin Schuessler, will perform the analysis and has not worked with the data or seen it before.

3) Describe the key dependent variable(s) specifying how they will be measured.

1. Average level of agreement to a statement referring to Pattern 1: “The system has learned that the “legs” position relative to the “head” is relevant for the prediction of Stretchy and Sticky.”
2. Average level of agreement to a statement referring to Pattern 2: “The system has learned that colour (type of colour, brightness or intensity) is relevant for the prediction of Stretchy and Sticky.”
3. Average level of agreement to a statement referring to Pattern 3: “The system has learned that rotation of the "building blocks" is relevant for the prediction of Stretchy and Sticky.”
4. Average level of agreement to a statement referring to Pattern 4: The system has learned that the background's brightness is relevant for the prediction of Stretchy and Sticky.
5. Average level of agreement to a statement referring to Pattern 5: The system has learned that the surface structure of the "building blocks" (e.g. circular or rectangular) is relevant for the prediction of Stretchy and Sticky.
6. Average level of agreement to a statement referring to the NN maturity: “The system has learned the right patterns and is ready to be used for this purpose.”
7. Averaged levels of agreement to a subset of four applicable statements taken from the Explanation Satisfaction Scale by Hoffmann et. al.
   1. “From the explanations, I understand how the system works. “
   2. “The explanations of how the system works is satisfying.“
   3. “The explanations of how the system works has sufficient detail.”
   4. “The explanations of how the system works seems complete.”

4) How many and which conditions will participants be assigned to?

The study is a between-subject design. We will randomly but equally assign participants to 1 of 2 possible conditions. In each condition, we will show participants a specific type of explanation after they watched videos that explain Machine Learning, the data set used and the explanation technique they

are about to use.
-> Baseline: The neural networks predictions are explained with a sorted grid of images drawn from the validation set. Each of the five columns of this grid represents a score range. Consequently, similarly rated photos are assigned to the same bucket column.
-> NNWI: This condition uses the same grid layout as the baseline. The difference is that the Neural Network is explained by Weight Vector Interpolations. Each row contains an original image, the remaining cells are filled with weight vector interpolations of this image, which change the prediction of the network so that they fit the designated score range.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
We will conduct a one-way independent Kruskal- Wallis test on all dependent variables followed by a focused comparison of the mean ranks between groups.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We will reject participants who submit low quality answers. For this we will use the Prolific Platform. Low quality answers are defined as follows:
* Failed to finish experiment at all or in under 58 minutes and did not submit completion code
* Undertook experiment on a handheld device smaller than a tablet or the resolution of the device was smaller than 600 px in width or height
* Wrong answer to attention check questions
* Did not watch all videos completely
* Time taken to submit agreement ratings is below 30 seconds
* Participants report technical issues, return the task on the Prolific Platform or withdraw consent for data usage

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
We will recruit 60 Participants from Prolific and they will need to have the following background:
- Fluent in English
- No reported hearing difficulty
- Approval rate on Prolific Platform of at least 95%
- Academic degree in Computer Science, Engineering, Finance, Mathematics, Medicine, Physics, Psychology
- Have not participated in pilot studies

We pay participants 4.00 GBP for their time and a performance-based bonus of 0.30 GBP per correct answer (max. 1.50 GBP)

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
Nothing else to pre-register.