Characterizing Human Explanation Strategies to Inform the Design of Explainable AI for Building Damage Assessment

Donghoon Shin Seoul National University
Sachin Grover Carnegie Mellon University
Kenneth Holstein Carnegie Mellon University
Adam Perer Carnegie Mellon University

NEURAL INFORMATION PROCESSING SYSTEMS
arXiv: 2111.02626
Project website

Background

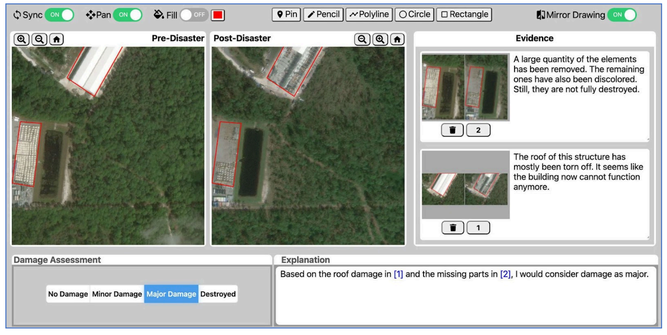
- Public satellite image datasets of natural disasters, such as xBD, have been used to develop AI tools for assessing building damage from satellite imagery
- Yet, fully-automated approaches are unlikely to be perfectly safe and reliable
- On such an account, explainable AI (XAI) is a promising means of supporting human-AI collaborations for high-stakes visual detection tasks
- However, most existing XAI techniques are not informed by the understandings of task-specific needs of humans for explanations

Research method

- As a first step toward understanding what humans require from XAI in building damage assessment tasks, we begin by characterizing **how humans generate explanations** for their own assessments in such tasks
- We conducted an online crowdsourcing study (N=60) to collect data on how people explain their own assessments in contexts of building damage detection

Annotation system

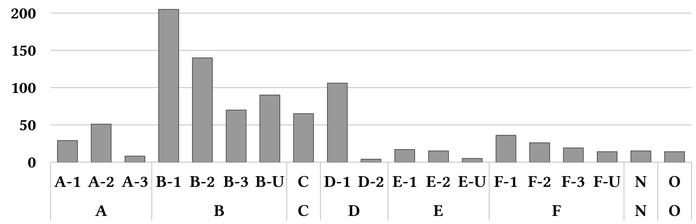
- We developed a web-based annotation system, where users can draw markups directly on pre- and post-disaster image using various drawing tools, along with explaining them as a text



Annotation system we designed and used during the study

Analysis & Results

- We used an iterative, open coding approach to identify categories among the explanations that participants generated
- From the study, we could surface 6 major strategies (A - F), along with several minor explanation methods (O) and “No damage” (N)
- A total of 929 codes were derived
- Summary of the codes and the descriptions of each code are as follows:



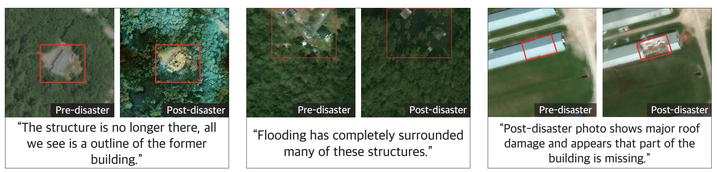
Summary of the codes

Code A: Constructing a causal argument to explain building damage

- **A-1**. Pointing to visual evidence of a natural disaster in a building’s surroundings to explain their assessment of building damage (e.g., *“From the evidence of flooding, I would say the building seems to have been affected”*)
- **A-2**. Inferring that a particular type of natural disaster had occurred based on evidence of damage to a building, then explaining their overall assessment of building damage with reference to the type of disaster (e.g., *“The building has roof damage. Probably a hurricane came and hit it”*)
- **A-3**. Constructing more complex, multi-step causal arguments (e.g., *“(Step 1) There was a fire and (Step 2) it was a wildfire that took everything from the building. (Step 3) You can only see the outline of the building”*)

Code B: Contrasting pre- and post-disaster imagery

- **B-1**. Referencing contrasts in the appearance of a specific building between the pre- and post-disaster images
- **B-2**. Comparing the pre- and post-disaster appearance of the area surrounding a building
- **B-3**. Highlighting contrasts in the appearance of specific sub-structures
- **B-U**. Ambiguous cases in which people generated contrast-based explanations, without clearly specify which elements they were comparing



(a) Example of B-1 **(b) Example of B-2** **(c) Example of B-3 and C**

Examples of code B-1, B-2, and B-3

Code C: Highlighting affected part of a building

- Rather than drawing markup around the whole building, some participants referenced specific affected parts of the building, but without necessarily comparing pre- and post-images in their explanations

Code D: Explanations based on the extent of damage to a specific building

- **D-1**. Explaining their assessment of the level of damage to a given building based on the proportion of the building that appears to be damaged (e.g., *“Approximately a half of the building was collapsed”*)
- **D-2**. Lowering assessments of damage by arguing that the damage appeared repairable (e.g., *“One part (of the building) was hit ... seems like it could be rebuilt”*)

Code E: Explaining reasons for lack of confidence in their own assessment

- **E-1**. Signaling their lack of confidence with reference to properties of a satellite image, such as low-resolution, visual distortion, or small buildings being obscured by shadows or taller buildings (e.g., *“The imagery has great distortion and is difficult to judge”*)
- **E-2**. Pointing out that they saw other changes between the pre- and post-disaster images, which made it challenging to precisely assess building damage (e.g., *“Oddly, it appears this building has been newly built up since the disaster”*)
- **E-U**. Noting that it was difficult to assess building damage, without necessarily providing a clear reason (e.g., *“This area is hard to judge”*)

Code F: Using the number of damaged structures in an image as the measure for severity of the disaster

- **F-1**. Explaining the building damage assessment with reference to the number of other buildings that appeared to be affected (e.g., *“It appears that one building has disappeared, leading me to believe it was destroyed. However, the remaining buildings seen are unharmed”*)
- **F-2, F-3**. Explaining their assessment with reference to the extent of damage visible in the surrounding area, by including (e.g., *“None of the large buildings appear to be damaged, but there is evidence of a large mud patch (in the surrounding area), indicating some minor flood damage”*) or excluding (e.g., *“All trees have been damaged or destroyed”*) building damage
- **F-U**. Ambiguous cases (e.g., *““Every area was totally destroyed”*)

Conclusion & Future work

- Participants often made use of contextual information (e.g., surrounding area and building) with direct evidences, weaving these into a coherent story
- Participants also frequently made reference to the visual contrast between pre- and post-disaster images, while arguing for its causal interpretations
- Finally, participants sometimes signaled their level of confidence in their own damage assessments within their explanations along with the reasons
- Future studies should explore how different types of explanations may impact HADR decision-makers in practice