Survey Instructions

Your task is to evaluate textual explanations of image classification by answering 5 questions on a scale of 1 to 5. You MUST NOT leave any input blank.

What is image classification?

Image classification is an image recognition task. Given an image, the AI algorithm automatically assigns it one of the predefined labels (such as "dog", "car", "house", etc.).

Example:  class: “Basset” (a breed of dog)

What are textual explanations?

The algorithms, in addition to assigning a label to the image, also produce a short text that should serve as a justification for why a particular label was assigned. For example, for an assigned label of "car", the explanation could be "The object in the image has four wheels, several windows and it is located on a road". The explanations can be shorter or longer and have different structures. Your task is to evaluate how good the explanations are according to a human.

How do you rate the explanations?

We will ask you whether you consider the explanations to be

1. **Fluency**: Are the explanations grammatically and linguistically correct?
2. **Clarity**: Are they easy to understand?
3. **Convincing**: Do they convincingly explain why the label was assigned?
4. **Reflecting the decision process**: Do they reflect how the label was assigned?
5. **Overall quality**: What is your overall assessment of the explanation quality?

(Those explanations will be visible next to every input box)

Important note:

the AI models CAN make WRONG DECISIONS. If you think the label is incorrect, but the model does a good job of explaining why it was assigned - such an explanation should be highly rated. The incorrect label assignment should NOT affect your evaluation of the quality of the explanation.

In particular, we would like to emphasise the difference between an explanation being "convincing" and "reflecting the decision-making process".:

Example A
Image: (corner of the wall of a medieval castle, made of bricks)
Predicted class: mosquito net
Textual explanation: This is a mosquito net, because the model has detected vertical and horizontal lines that intersect each other on the right-hand side and in the centre of the image.
Assesment: The quality of the explanation should not be affected by the fact that the model made an incorrect decision. The explanation mentions the detection of vertical and horizontal lines visible in the image (grid created by bricks) and also provides a correct localisation of them. It is conceivable that this reflects the model's decision process well, since mosquito nets also have a grid pattern. This explanation is also quite convincing.
Example B
Image: (corner of the wall of a medieval castle, made of bricks)
Predicted class: mosquito net
Textual explanation: A brick building.
Assesment: Despite the fact that we see 'a brick building' in the image, this does not give any information as to why the model decided that this was a mosquito net. This certainly does not reflect the model's decision-making process and is not convincing.
Example C
Image: (corner of the wall of a medieval castle, made of bricks)
Predicted class: castle
Textual explanation: In the picture we can see a building made of bricks.
Assesment: This explanation can be considered convincing to some extent (it's subjective, which is why we want to measure it in the study). However, the explanation does not give any information about how the model detected "a brick building" and decided that it was a castle.