Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (b) Did you describe the limitations of your work? Yes, see § 8
   (c) Did you discuss any potential negative societal impacts of your work? Yes, see § 8
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? N/A
   (b) Did you include complete proofs of all theoretical results? N/A
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, with unspecified ones included in the code
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? No
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? No
   (c) Did you include any new assets either in the supplemental material or as a URL? Yes
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? No
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? Yes, see Appendix B
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? N/A
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? N/A (Volunteers)

A Explainer Details

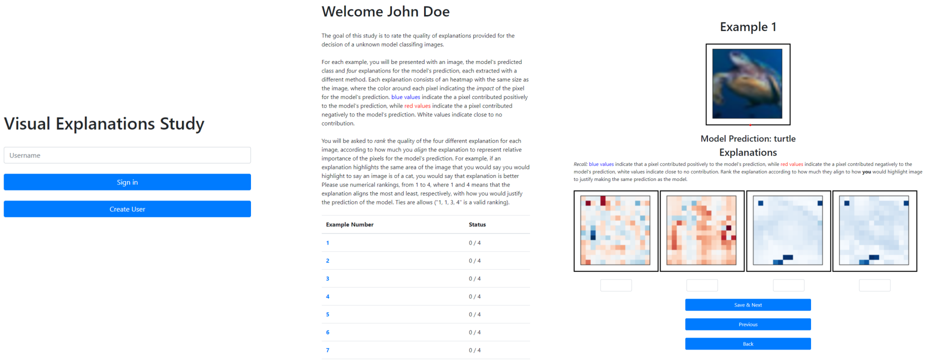
With the *integrated gradients* explainer [Sundararajan et al., 2017], we use 10 iterations for the integral in the *simulability* experiments (due to the computation costs) and 50 iterations for the *plausability* experiments. We use zero vectors as baseline embeddings, since we found little variation in changing this. For both gradients-based explainers, we project into the simplex by using the SOFTMAX function, similar to the attention-based explainers. This results in very negative values having low probability values. Moreover, for evaluating plausibility on text classification and translation quality estimation, we computed the explanation score of a single word by summing the scores of its word pieces.

We would like to note that, unlike the setting in Pruthi et al. [2020], we **do not** apply a *top-k* post-processing heuristic on gradients/attention logits, instead directly projecting them to the simplex. This might explain the difference in results to the original paper, particularly for the low simulability performance of static explainers.

B Human Study for Visual Explanations

The annotations were collected through an annotation webpage, built on top of Flask. Figure 6 shows the three pages of the site. During the annotation, users were asked to rank four explanations, unnamed and in random order. After collecting the ratings, we computed the *TrueSkill* rating, with an initial rating for each method of $\mu = 0, \sigma = 0.5$. After learning the ratings, we then compute the *ranks* by obtaining the 95% confidence interval for the rating each method, and constructing a partial ordering of methods based on this.

The volunteers were a mixture of graduates or graduate students known by the authors. However we would like to point out that due to blind nature of the method annotation, the chance of bias is low.



Visual Explanations Study

Welcome John Doe

The goal of this study is to rate the quality of explanations provided for the decision of a unknown model classifing images.

For each example, you will be presented with an image, the model's predicted class and *four* explanations for the model's prediction, each extracted with a different method. Each explanation consists of an heatmap with the same size as the image, where the color around each pixel indicating the *impact* of the pixel for the model's prediction. blue values indicate the a pixel contributed positively to the model's prediction, while red values indicate the a pixel contributed negatively to the model's prediction. White values indicate close to no contribution.

You will be asked to *rank* the quality of the four different explanation for each image, according to how much you *align* the explanation to represent relative importance of the pixels for the model's prediction. For example, if an explanation highlights the same area of the image that you would say you would highlight to say an image is of a cat, you would say that explanation is better Please use numerical rankings, from 1 to 4, where 1 and 4 means that the explanation aligns the most and least, respectively, with how you would justify the prediction of the model. Ties are allows ("1, 1, 3, 4" is a valid ranking).

Example Number	Status
1	0 / 4
2	0 / 4
3	0 / 4
4	0 / 4
5	0 / 4
6	0 / 4
7	0 / 4

Example 1

Model Prediction: turtle

Explanations

Recall: blue values indicate that a pixel contributed positively to the model's prediction, while red values indicate the a pixel contributed negatively to the model's prediction, white values indicate close to no contribution. Rank the explanation according to how much they align to how **you** would highlight image to justify making the same prediction as the model.

Figure 6: Login page (left), dashboard (middle) and annotation page (right)

In Figures 7 and 8 we show raw attention explanations extracted from all layers (rows) and heads (columns) of the teacher transformer used in our CIFAR-100 experiments. Cross-checking the explanations with the most relevant heads selected by SMaT for image classification, we can see that most selected heads produce plausible explanations (e.g., attention heads from the last layers).

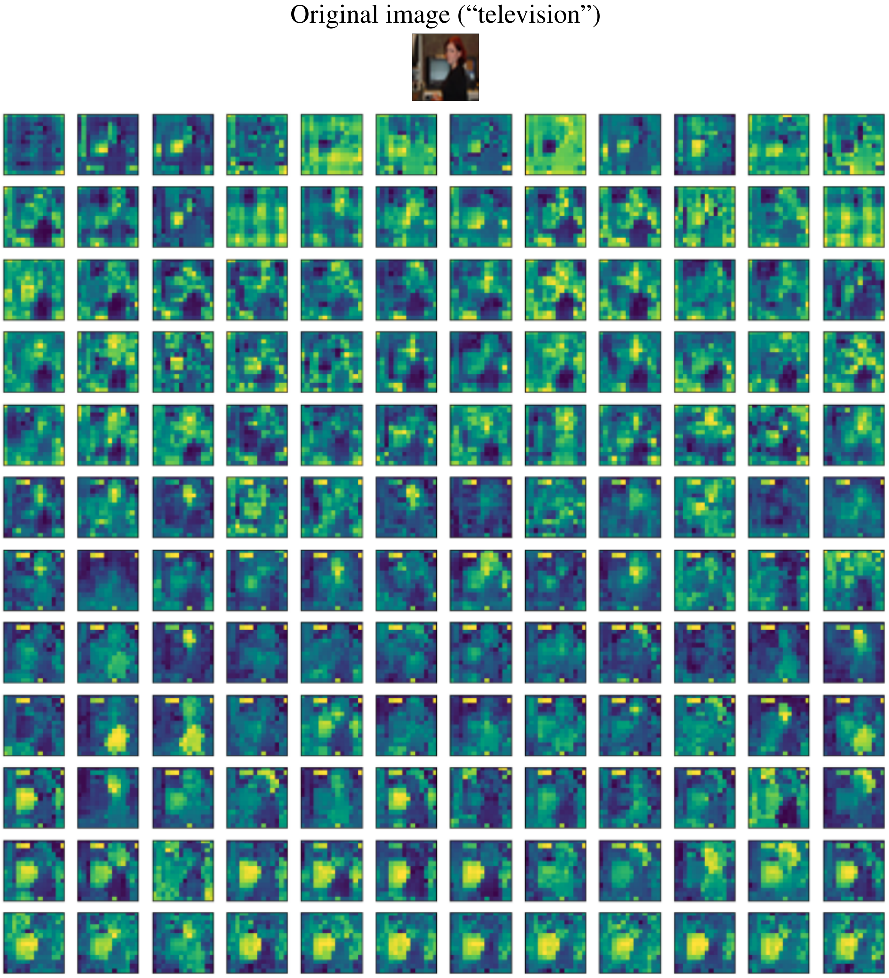


Figure 7: Explanations from all layers (rows) and heads (columns) of the CIFAR-100 teacher model.

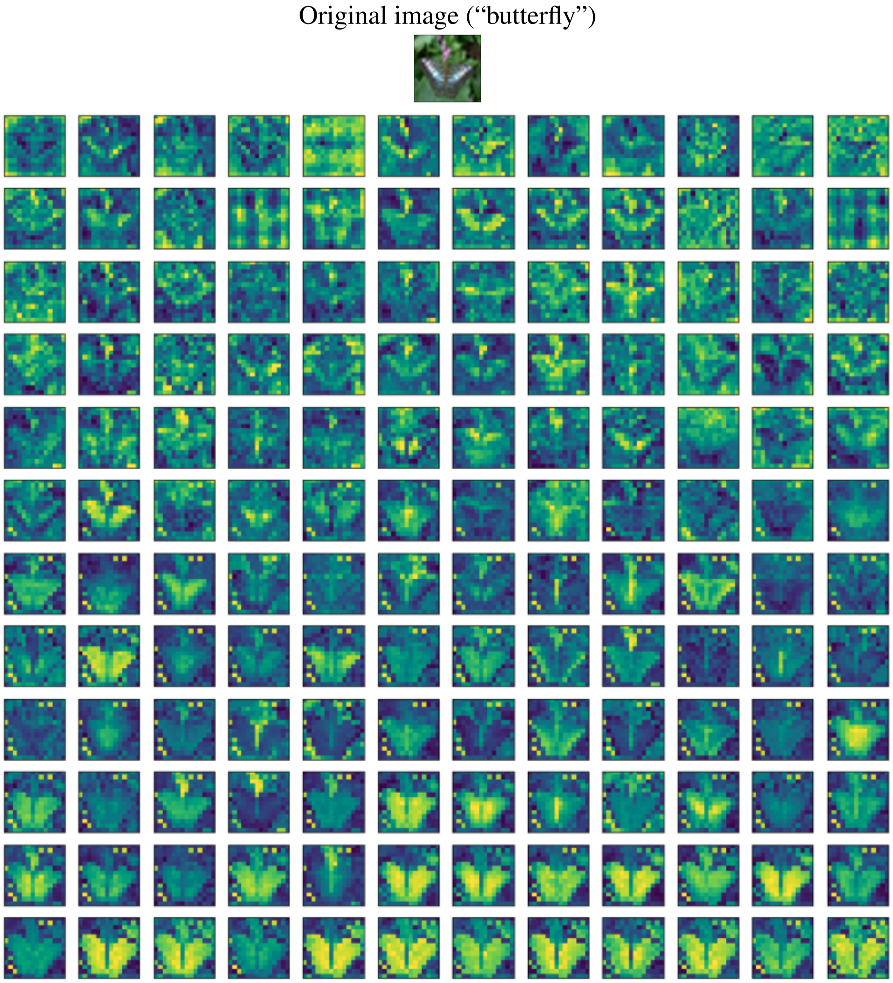


Figure 8: Explanations from all layers (rows) and heads (columns) of the CIFAR-100 teacher model.