We thank the reviewers for the insightful and helpful comments. Minor remarks will be reflected in the text. Individual responses to the questions are included below.

Reviewer 1

The reviewer raises the very relevant issue of applicability, both theoretical and practical. The paper's framework is general and can be applied to a wide class of problems and ML models. Hence, we agree with the reviewer in that the theoretical applicability is significant. Moreover, we will improve the presentation of experiments, to highlight the practical applicability of the work.

Also, the practical applicability of our work depends on a reasoner for some relevant fragment of first order logic. The same applies to a number of recent works that exploit formal methods. The improvements made to reasoners, e.g. SMT solvers, in recent years offer guarantees that practical applicability will continue to improve.

Furthermore, there is already practical evidence to the relevance and the insights provided by the use of formal methods in ML; some are referenced in our paper.

The duality relationship that our paper reveals will enable researchers to look at adversarial examples and explanations in a new light. This will open new avenues of research, that will foster more efficient exact methods but also better heuristics.

The reviewer is quite right that explanations find important uses in many other settings. We will expand on this.

Reviewer 2

We thank the reviewer for the suggestions regarding presentation.

We will improve readability of the paper, e.g. by clarifying the concepts that are not widely known in the community, including some of the logic-related notation used.

The proof of Theorem 1 will be cleaned up in the revised paper, as suggested by the reviewer. All the minor issues spotted will be addressed. Thank you!

Reviewer 3

The reviewer raises a *'so what'* question.

The importance of having a deep understanding of the connection between adversarial examples and explanations is epitomized by a number of recent works, some of which are cited in our paper.

Also significant is a recent keynote talk by I. Goodfellow at the AAAI 2019 conference on *"Adversarial Machine Learning"* (please see the official video recording starting from 53m55s), which makes a very strong case for relating adversarial examples and explanations. Namely, Goodfellow conjectures that there should be a connection between adversarial attacks and explanations and points out that this connection has not been revealed so far. We believe that our work is a significant step forward in this direction.

We chose an image dataset as we can visually demonstrate the discovered relationship between adversarial examples and explanations to the reader. We want to convey that each explanation hits all adversarial examples and vice versa. This "hitting" process can be highlighted using coloring in the image (please see Figures 1 and 2 in the paper). While our method can be applied to categorical datasets, we feel that such visualization might be less visually appealing to a reader.