All Reviewers We appreciate the helpful feedback of our reviewers and the pointers to relevant recent work. We agree that some of our tables and layout could be better designed and are grateful for the suggestions for doing so. We apologize for the typos identified and will make sure that the paper is suitably polished in the final version.

Reviewer #1 **Parameter Sweeping.** We agree that a parameter sweep in the adversarial training experiments will produce fairer results. When choosing ϵ, we used a common setting ($\ell_2, \epsilon = 0.25$) for data scaled to $[0, 1]$, and in the final version, can instead sweep in increments of 0.1. Experiment II is computationally intensive so we are unable produce these results before the rebuttal deadline but we expect that increasing ϵ or maximum attack steps in each iteration may increase the robustness but also increase the training budget for adversarial training. We are working on additional experiments with different hyper-parameters to build scatter plots of the robustness against accuracy. **Novelty Concerns**. 1) We appreciate that the reviewer brings Yeh et al (2019) to our attention. We find one major difference is that Yeh et al (2019) is motivated by evaluating attribution methods (i.e., no attacks are evaluated in their work), whereas our focus is on robustness to attacks. Nonetheless, in theory SENS-MAX is indeed another way to incorporate our local robustness λ. However, using Monte Carlo to approximate SENS-MAX is different from performing gradient descent to attack attributions, and especially in high-dimensional image spaces, an adversarial example may not be easily sampled. Another difference is that while Yeh et al (2019) proposes Hessian regularization as a possible remedy, we implement and evaluate its effectiveness in the adversarial setting. We will update the paper to reflect this discussion. 2) We find the 3rd version of Singh et al (2020) is the one we cited. There are significant differences between the newest version, which was not available the time of submission, and the one we cited. We will update the paper to discuss the latest version, as well as adding the relevant comparisons to our experiments.

Reviewer #2 **Comparison with IG-NORM** We agree that IG-NORM tends to show greater robustness, but this comes at an appreciable cost in accuracy. In preparing our experiments, we also found that it is quite sensitive to initialization, leading to variability in model fitness for a given allocation of training time in epochs (mentioned on lines 261-263). These tradeoffs may be significant for practical settings. **Appendix J of Levine et al (2019).** We appreciate bringing this appendix to our attention. We believe our work can be viewed as a further exploration of the results shown in in Figure 13, giving geometric intuitions along with new theory and experiments that might shed fresh light on those results. **Novelty of Experiment I.** Although we are curious about which references the reviewer has in mind, to the best of our knoweldge prior work has not shown the effectiveness of stochastic smoothing on the Dombrowski et al (2019) attack, and we are not aware of references that report measurements of a technique similar to our Uniform Gradient. If we are mistaken, we would appreciate pointers to the relevant work.

Reviewer #3 **Analysis for IG.** We agree that including more analysis of IG is an interesting direction, with relevance to practice given the popularity of that method. However, characterizing IG under the same assumptions that we made for SG requires considering the global geometry of the model between an arbitrary baseline and the input. This is part of our ongoing work, but is a significant addition that would be difficult to present adequately in a single short paper. **Other Hessian Approximations.** 1) Thank you for making us aware of Moosavi-Dezfooli et al (2018). If we had known of it at the time of submission, we might still have used Singla et al (2019), as the two-step approximation used by Moosavi-Dezfooli et al (2018), along with the fact that we do not require approximation over a Gaussian but instead a single point, suggests that Singla et al (2019) is a better fit in terms of efficiency for our needs. 2) We expect that Singla et al (2019) provides a closer approximation than Yoshida et al (2017), which regularizes the aggregate spectral norms of weights at each layer, giving a loose upper-bound of the input Hessian spectrum. However, we will include discussion of these methods in the paper, as well as appendices containing experimental comparisons. **Generality Experiments.** We suspect that a black-box attribution adversary may not be as meaningful in practice as a black-box label adversary, except when it is reasonable to assume that models trained on the same distribution are expected to have (approximately) identical attributions for given test points. However, we agree that these experiments may generate interesting results, and will report on them in an appendix in future versions of the paper.

Reviewer #4 **Variability Measurement.** We appreciate the discussion on variability tests and agree that they will strengthen our conclusions. We note that some of the attacks we measured (e.g., SM and IG) are deterministic and we have not observed significant variability in those parts of our empirical analysis when hyper-parameters are fixed. As mentioned in the review, our experiments are computationally intensive. Therefore we were unable to include multiple trials at the time of submission; we will report results over multiple trials in the final version along with a discussion of variability. **Results and Prior Work.** In the experiment section, we agree that more analysis and interpretations about our empirical results can make the conclusions more convincing. Also, as mentioned by previous reviewers, we notice that there are more interesting work of which we are not aware at the moment of submission. We will re-structure the "Related Work" section based on feedback from all reviewers. **Writing and Layout.** We appreciate your (very) helpful suggestions about reorganizing some of our figures and tables; we agree that Table 2 can be confusing, and we plan to re-design Tables 1 and 2 based on your feedback.