Author Response for: "Inverting Gradients - How easy is it to break privacy in federated learning"

General Comments: We thank all reviewers for their valuable feedback and interest in this attack. We want to stress that the key points of this work are a surprisingly effective new attack, evaluation of previous work in realistic settings and attack of multi-step federated learning. To the best of our knowledge, our work is the first to investigate the multi-step setting at all.

Some questions arose about the theoretical analysis for fully connected layers. This analysis is not only meaningful in other domains, such as medicine (e.g. Jarrar et al., "MLP neural network classifiers for medical image segmentation") and financial manners (e.g. Kadhim et al. "Prediction of the Performance Related to Financial Capabilities"), but also shows that attacks on gradient data are easier than inversion from feature representations as in [20]. Also, this result directly applies to iRevNets (which have reversible feature representations), showing that even deep CNN architectures exist that leak all information. Finally knowledge of the feature representation already enables attacks like Melis et al. [23], which utilize an auxiliary malicious classifier.

Reviewer 1: Note that our main claim is not only that previous results did only consider shallower networks, but also that previous results consider unrealistic settings (smooth activations, no stride in convolutional layers, untrained parameters), which make reconstruction easier. Our results show that a stronger attack still succeeds in a realistic setting both for a single image, and for multiple images and steps. We directly compare visually to previous works (which both use L-BFGS and Euclidean loss) in Fig.2, showing that previous works struggle in realistic settings. We will extend our supplementary material and show additional visual results for Table 1.

Reviewer 2: Regarding technical novelty, note that we propose a deceptively simple new attack that nevertheless significantly broadens the applicability of gradient inversion attacks. We formalize a realistic threat scenario and evaluate previous works and the new attack in this new setting. We investigate why realistic scenarios differ, covering trained vs. untrained parameters and architecture properties. We then move to federated learning with multiple steps and multiple images, and discuss how to attack this scenario.

Regarding the recovery of batch results, where we show that recovery from a batch of 100 averaged gradients is possible, the key factor here are the privacy implications. Assuming this was a batch of 100 private photos, would we consider it secure if only 5 of 100 private photos were revealed? The surprising revelation of the 100 image experiment is that the distortions arising from batching are not uniform. One could have expected all images to be equally distorted and near-irrecoverable, yet some images are highly distorted and other only to an extend at which the pictured object can still be recognized easily. This non-uniformity is a significant result for the privacy of gradient batches. Also note that Fig.4 of [35] looks better because the attack scenario there is easier. We analyze the attack of [35] in Fig.2 and Tab.1 and find that it struggles in realistic settings.

Reviewer 3: We did conduct experiments on fully-connected networks as a sanity check, reaching a full reconstruction. However, as these merely confirm the proven statement of Prop. 3.1., we did not deem them interesting enough to include.

Our statement about magnitude and direction of gradients relies on intuitions from optimization, i.e., that the negative gradient is the direction of steepest descent, and that for strongly convex functions the gradient magnitude is an upper bound on distance to the optimal solution.

After considering the single-image scenario as in previous works, we do cover more practical scenarios in Sec. 6. The scenario with multiple participants is equivalent to what we discuss, if the server receives contributions from each participant separately. If the contributions are averaged without knowledge of the server (such as in secure aggregation), then recovery of images from multiple participants reduces to recovery from a batch of averaged gradients.

The attack on averaged gradients only has knowledge about the average of gradients, however we assume the number of participating images to be known to the server. The server might request this information anyway (for example to balance heterogeneous data), but even if the exact number of images is unknown, the server (which we assume to have significantly more compute power than the user) could run reconstructions over a range of candidate numbers, given that the number of images is only a small integer value and then select the solution with minimal reconstruction loss.

We will include more information about computational costs in the revised version of this work. An analysis of attacks against differentially private models is highly interesting and a topic of future work for us. The reconstruction on CIFAR for the untrained "LeNet(Zhu)" model is better with L-BFGS because this model is smooth with a large linear layer. This is a less realistic scenario, but ideal for L-BFGS, and the superior convergence speed of L-BFGS is realized. Finally, we're glad that you found the code example working as desired.

Reviewer 4: We thank you for valuing this work and sharing our enthusiasm in it.