Ivan Miller. Reading Summary (Attack on IntsaHide). 10.18.2022

1. Describe in your own words a few of the main points this paper is making.
The paper first theoretically argues that image encodings produced by the InstaHide algorithm are distinguishable. Then the authors exploit that "distinguishability" of encodings to solve the InstaHide challenge and recover original training images.
The authors formalize the notion of instance encoding and its privacy and discuss reasons why providing indistinguishability (privacy) guarantees with instance encoding that allows for training accurate models is impossible because of the following reasons: 1) Leakage in a form of adversary's ability to predict the label of an instance from its encoding implies a distinguishing attack between encodings; 2) Theoretically, encodings that are actually indistinguishable from each other must contain almost all the information (to be extracted by the model), which will make it impossible to learn a given problem privately while maintaining reasonable accuracy.

2. What kinds of definitions do they provide for what privacy of image encodings should mean?
Unlike the InstaHide paper, the attack paper introduced several formal definitions for instance encoding: 1) *Dataset encoding* - as an algorithm that takes a dataset as input and outputs an encoded dataset; 2) *Instance encoding* - as a special form of general dataset encoding. The authors also outlined the privacy of instance encoding by providing two types of attacks: 1) *Distinguishing attacks* that have a goal of determining with a higher than 50% probability whether a certain dataset was used to generate the encoded dataset; 2) *Recovering attacks* that are a stronger, and more devastating forms of attack that aim to recover individual training examples, which means that the goal of making that type of attacks impossible is the weakest form of expected privacy.
In addition to physical privacy (preventing extraction of any information about the data that was used for training), the paper also differentiates between the importance of functional privacy - making sure that an adversary won't be able to infer sensitive information about original training data based on the output of the model (leakage). The paper additionally concludes that a given attack on the encoding algorithm could be considered efficient if it runs in polynomial time.

3. Describe the extent to which they attack InstaHide. Do you think this is a devastating attack on the InstaHide design? Discuss why or why not.
In my opinion, the attack presented in the paper is absolutely devastating and it effectively renders InstaHide instance encoding algorithm useless. Not only were the authors able to solve the InstaHide challenge and produce high quality reconstructions of the encoded images that were nearly identical visually to the original training data but they were able to do that at an extremely low cost and with the amount of effort that is unacceptable for a algorithm that claims privacy of the encoded results.
The fact that encoded images were distinguishable (with the help of a similarity function learned using a neural network) allowed the authors to merge the encodings of the same image produced for each training epoch and solve a noisy linear system to recover the original images. The authors also proved that, as opposed to the claim made by InstaHide, mixing more images into each encoding offers no additional privacy and, on the contrary, makes the reconstruction task easier. Moreover, the paper proves that a nearly perfect reconstruction is possible based on just one encoded image, as long as public parameters of InstaHide are known to the attacker.

4. What's a question you would like to see the authors address?
Understanding the importance of the task of privately training neural networks without sacrificing performance I would have liked reading about the authors' general suggestion/opinion as to whether there may exist a solution to that task.