Ivan Miller. Reading Summary (IntsaHide). 10.11.2022

1. Describe in your own words the main problem this paper is trying to address.

The paper introduces an algorithm designed for encryption of images that could be used to train a machine learning classifier. The proposed algorithm claims to offer sufficient security of training data while maintaining high accuracy of the model without slowing down the training process.

The need for such an algorithm comes from the fact that there are applications where private parties have limited amounts of their own data but sharing or pooling that data is not allowed due to existing privacy regulations (e.g. HIPAA). While, according to the paper, all existing frameworks have limitations, for example, differential privacy (a solution when noise is added to gradients of the model) leads to a big drop in accuracy together with some decrease in efficiency, and cryptography-based solutions lead to a big loss in efficiency due to a high computational cost, *the solution proposed in the paper [allegedly] allows to increase privacy with only a minor loss in accuracy of the final model.*

2. Describe one of the primary ideas behind the design of InstaHide.

The main idea introduced in the paper is to encode private images by 1) using a linear combination of training images (mixup) from a private and a large public dataset followed by a 2) random pattern of the sign-flips on the pixel values of this composite image. Both, the images used for producing the mixed-up composite and the random sign-flipped mask are specific to each training image and are only used once. The paper also states that the security of the "encrypted" image could be increased by using several images during the mixup stage: we could combine K different images with K/2 coming from the private dataset and another K/2 from the public dataset (recommended $K \geq 4$).

3. Identify a few good and a few problematic things about the way the paper is written/constructed.

3.1.Positive:

3.1.1. The paper contains helpful introduction into related work covering the basics of relevant frameworks such as mixup, differential privacy, and cryptography-based solutions.
3.1.2. Inclusion of the section on InstaHide deployment (#6) seemed like a good way of reiterating best practices to ensure successful implementation to achieve highest security as a result. However, some of the important recommendations highlighted in the paper itself aren't mentioned in that section (e.g. use random patches of the public dataset for mixup and filter out "flat" patches that have less than 40 key points using SIFT).
3.1.3. Introduction of a challenge dataset and a public invitation to design attacks to recover original images encoded with InstaHide allows readers to put the proposed algorithm to a real life test.

3.2. Negative:

3.2.1. After the paper introduced the second part of the algorithm where a pixel-wise sign-flipping mask is applied to composite images, the authors made a note that that step could alternatively be explained as "retaining only absolute value of each pixel in the mixed image", which seems like a much simpler way to achieve the same result, however, the authors continued with a more complex sign-flipping mask throughout the paper.
3.2.1. Some of the claims made in the paper seem to be rather vague. For example, in section 3.3. "Inference with InstaHide" the authors compared their algorithm with cryptographic frameworks and stated that "distributions of encryptions of different images are indistinguishable", however, aside from providing a set of histograms, no proof of that claim was offered.
3.2.3. When covering different attacks against InstaHide (section 5.3) the language used by authors may be suggesting that they weren't trying to challenge the proposed algorithm in a serious fashion.

For example, when describing demasking using GAN they settle with “... this doesn’t appear enough to allow further attacks that recover the encrypted image”, which does not instill confidence in the security of the proposed algorithm and its ability to withstand a serious adversarial attack.

4.What’s a question you would like to see the authors address?
As a point of reference I would have liked seeing detailed information on how much the best of existing cryptography-based solutions slow down the training of a comparable deep learning model, while maintaining, let’s say, at least 90% test accuracy on CIFAR-10 dataset.