Ivan Miller - Problem Set 4 for Adversarial AI. Tuesday, Oct. 18, 2022

1. *Suppose that you are collecting audio files from a bunch of different individuals to train a model that will map an audio file to a text word. Each file contains the sound of an individual saying a single word. You want to assure the people contributing their voice files that the output of your process won't reveal anything about their individual speech patterns. Do you think you can reasonably promise this? Can you think of an example of something that might go wrong?*

I think it could be possible to support a promise of not revealing individual speech patterns when accepting audio files. If we were to change the parameters of the voice signal in each audio recording such as pitch, frequency etc. to make the recording sound completely different. As far as I know, there is a whole range of parameters that could be adjusted to improve privacy of the original speaker, however, it would be even better to use a text to speech synthesizer (for example, https://cloud.google.com/text-to-speech) to modify the input files before making them accessible to the model.

2. *Consider the following proposal to build an image classifier from private images that supposedly prevents answers of the final model from leaking any private info about an individual image. The classifier outputs 0 or 1.*

- *Suppose there are N private images, x_1, …, x_N, in the training.*
- *The training process starts by generating N training sets each containing N-1 images.*
- *The i-th training set is formed by simply removing the i-th private image, x_i. Now a model is trained on each of the N training sets, and we'll call the i-th model M_i. (So M_i is trained on all the private images except for x_i.)*
- *When classifying an image, the classifier applies all of the N trained models to get N binary answers.*
- *It then takes the mode as its final answer. For example, if N = 10 and 8 of the models say "1" and 2 of the models say "0", it will output 1.*

Do you think this approach might provide any meaningful privacy for the underlying images? What concerns might you have? Please discuss.

I believe that, despite ephemeral sense of security given by the fact that the final output of the model is given by a voting classifier, the proposed approach could potentially lead to data leakage in forms of labels of the training data as well as leaking the features of the training data that may not necessarily be visible to a human eye such as distributions of pixels, their contrast, brightness etc that could be used to reproduce original images. Additionally, if each M_i model produces activation maps that may be accessible to an adversary, it will become possible to an adversary to discover the features of the images that the model is paying attention to, which could be used make inferences about the training images which will make privacy of underlying images impossible.