Appendix

A. Extension to Multi-Target Model Inversion Attacks

Existing MI attacks aim to extract sensitive features (or entirely construct samples) from a private dataset. So far, all previous works on MI attacks only consider a single target model. However, multiple models trained on the same private dataset are often available in a real scenario. Performing MI attack on multiple different target models that are trained on the same private dataset could be an interesting and practical problem in real-world applications. Will the attacker gain more information about this private dataset in this case?

In our experiments, we extend our attack algorithm over single-target setting to the multi-target setting by following the same attacking procedure. First, an inversion-specific GAN is built under the guidance of multiple target models, where the supervised loss of discriminator $L_{\text{multi\_sup}}$ is a weighted combination of cross-entropy losses between the discriminator output distributions and soft labels given by different target models; supervised loss and generator loss are exactly same as under the single-target setting:

$$L_{\text{multi\_G}} = \|\mathbb{E}_{x\sim p_{\text{data}}}\mathbf{f}(x) - \mathbb{E}_{z\sim \text{noise}}\,\mathbf{f}(G(z))\|_2^2 + \lambda_h L_{\text{entropy}} \tag{1}$$

$$L_{\text{multi\_D}} = L_{\text{multi\_sup}} + L_{\text{multi\_unsup}} \tag{2}$$

where

$$L_{\text{multi\_sup}} = -\mathbb{E}_{x\sim p_{\text{data}}\,(x)} \sum_{m=1}^{M} w_m \sum_{k=1}^{K} T_k^m(x) \log p_{\text{disc}}(y = k \mid x) \tag{3}$$

and

$$\begin{aligned} L_{\text{multi\_unsup}} = -\{&\mathbb{E}_{x\sim p_{\text{data}}\,(x)} \log D(x) + && (4)\\ &\mathbb{E}_{z\sim \text{noise}} \log(1 - D(G(z)))\} && (5)\end{aligned}$$

Here we use the same notation as the single-target setting. M is the number of target models, w_m is the weight on the loss over the mth target model T^m.

Then the learned GAN over multiple target models are applied for distributional recovery where the new identity loss is a weighted combination of identity loss over multiple target models, and prior loss remain unchanged:

$$L_{\text{multi\_id}} = -\sum_{m=1}^{M} \frac{w_m}{L} \sum_{l=1}^{L} \log T_k^m(G(\sigma\epsilon_l + \mu)) \tag{6}$$

$$L_{\text{multi\_prior}} = -\mathbb{E}_{z'\sim p_{\text{gen}}} \log D(G(z')) \tag{7}$$

$$\tag{8}$$

and overall loss for distributional recovery under multi-attack setting is:

$$L_{\text{multi}} = L_{\text{multi\_prior}} + \lambda_i L_{\text{multi\_id}} \tag{9}$$

In our experiments, the same set of training hyperparameters such as learning rate and weight decay are used. In addition, we set w_m in (3) and (6) to $\frac{1}{M}$ so that the discriminator supervised loss and the identity loss are averaged over all target models.