**Adversarial Neural Network Inversion via Auxiliary**

**Knowledge Alignment**

Ziqi Yang, Ee-Chien Chang, Zhenkai Liang

**Abstract**

本文关注的是对抗背景下的模型反演攻击。（对抗背景指敌手企图通过模型预测值推断目标模型的训练、测试数据信息）。

本文训练一个second neural network对目标模型进行反演。训练方式是对目标模型黑盒访问，具体使用两种方式：（1）利用敌手的背景知识，合成一个**辅助集**用以训练反演模型，这种情况不需要访问原始训练数据。（2）设计了一种基于**截断**的技术对齐（align）反演模型，当敌手从目标模型得到一部分预测值时，可以进行有效反演。

实验表明，对目标模型训练集无充分了解且只能得到部分预测值时，本文的反演方法可以进行有效反演，且优于以往的方法。

**Introduction**

模型反演的目标是从模型的预测值获取训练数据的信息。

本文中：

**1.如何解决“无法访问原始训练数据”：**

以往工作中，敌手会采用与被害者模型同分布的训练数据，本文的数据集则是从基于背景知识的更一般的数据分布中获得，因为这对敌手来说更好获得。（打个比方，针对一个面部识别分类器，敌手在不清楚确切训练数据分布的情况下，可在网上随机地抓取人脸的图像，构成辅助集）。使用辅助集效果很好，因为保留了一般特征（也就是其特征更加普遍通用）。实验表示，即便可得到训练数据，用这种通用的样本扩展数据集，可以提高反演精度。

**2.如何解决“当敌手只能获取受害者数据的部分预测结果，反演模型依然正常工作”：**

提出一种截断方式来训练反演模型。主要思想是训练时向反演模型提供辅助样本的截断预测。这种截断使反演模型基于截断预测的基础上最大限度重构样本，使反演模型与截断值保持一致。实验显示这种方式很有效。

截断：指对受害模型的预测值，保留其中部分最大值，其余的截断为0。

**3.敌手角色不同时，训练反演模型的方式有何区别？**

敌手是用户，对训练数据一无所知，对分类器只可进行**黑盒**访问（只得到预测值）：训练一个**独立**于原模型的反演模型。

敌手是训练分类器的开发人员：反演模型与分类器在相同数据上**联合**训练，得到的模型精确度更高。

**Contributions：**

1. ·在敌手只能得到部分预测值情况下，提出截断方式，使反演模型与阶段预测的空间保持一致。
2. ·本文的反演方法适用于training class inference attack，且优于现有方式。（这种攻击目标在于为每个类恢复有意义的数据，常通过反演攻击为特定类生成一个代表性样本来实现）。
3. ·发现即使敌手拥有分类器的训练数据，用辅助集中的一般性样本对训练集进行扩充是可以提高反演模型精度的。

**Background**

模型反演可分为两类：optimization-based inversion和training-based inversion（后者和本文的方式很相似，本文在此基础加上了截断）。

Optimization-based inversion：对输入空间上使用基于梯度的优化，找到预测值最接近于给定值的图像（也是通过最小化损失函数实现）。要求是白盒。

Training-based inversion：训练另外一个神经网络*Gθ*（反演模型）来反演原始模型*Fw*。（*Gθ*也就是Abstract中提到的second neural network）。*Gθ*的输入是预测值，输出是一个图像。

**Adversarial model inversion**

根据敌手的知识和角色不同，可细分为三种场景：

1、好奇的用户，想要从受害者*Fw*的截断预测中重构出受害者的输入数据。

2、好奇的用户，想要推断出受害模型*Fw*的功能。

3、恶意开发者，想构建出一个反演模型，之后使用它从截断预测值中重建受害者的输入。

**场景1（Data reconstruction with blackbox classififier）**

敌手知识：从更普遍的分布中得到的样本；*Fw*的输入格式；*Fw*的输出格式（也即预测值的维度）（如何理解第三条：尽管只能得到截断预测值，可以通过一系列问询，收集不同的类作为输出维度）。

设分类器为*Fw*，其输入为x，预测向量为g，敌手可得到一个m-截断的向量f，用trunc*m*(**g**)表示。得到f的方式是将g中最大的m个值保留，其余截断为零（m是受害者设置的），如：

clipboard.png

f中保留m=2个最大值，其余截断为0。

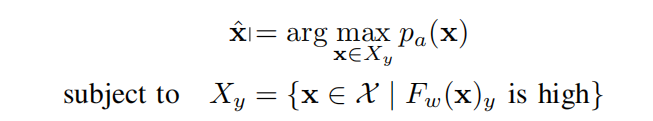
敌手目标是在普遍分布中找到一个样本x'使之满足：

clipboard.png

**场景2（Training class inference）**

敌手知识和场景1相同，不同的是敌手目标不在于数据重构，而在于找到类的**代表性样本**。

设y为目标类，敌手要在普遍分布*pa*中找到一个x尖满足：

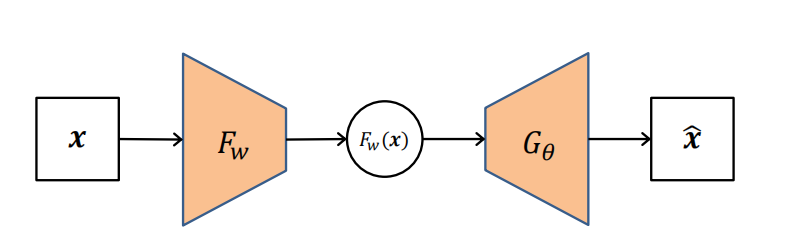
（*Fw*(**x**)*y*指x被*Fw*分为y类的置信度）。

**场景3（Joint classifier and model inversion）**

敌手有*Fw*的全部知识，包括训练数据、结构、参数，可以自由决定*Fw*是什么样的。本文假设在*Fw*发布后，敌手可从用户获得截断预测值，并希望重构出用户输入。敌手将训练一个满足精度要求的分类器*Fw*（这个也是原来分类器的任务），并提高数据重构的质量。

**Approach**

框架：



输入x到*Fw*得到预测向量*Fw*(**x**)，训练反演模型*Gθ*, 将预测向量作为*Gθ*的输入，输出重构数据x尖。

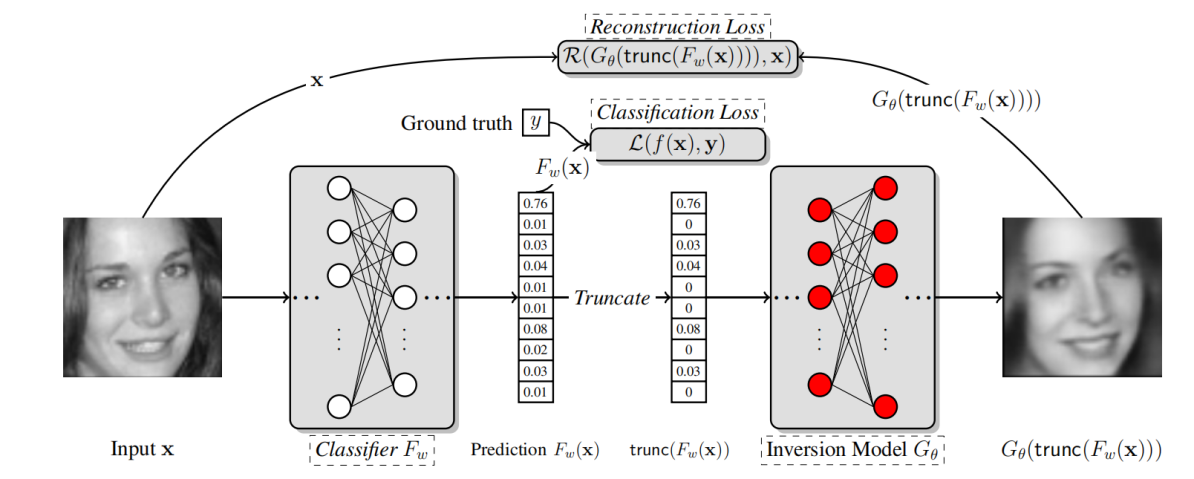
接下来分三部分详细阐述本文的方法：（1）在场景1,2中*Fw*是固定的，因此需要找到训练*Gθ*的数据，在A部分讲如何找到这部分训练数据(辅助集)；（2）敌手只可获得截断预测，因此要找到方法使之尽量“对齐”原始预测值，B部分讲截断方法；（3）场景3中*Fw*不固定，采用联合训练*Fw*和*Gθ*，在C部分讲训练方式。

**A. Gθ’s Training Data: Auxiliary Set**

用于训练Gθ的数据集称之为辅助集。辅助集是从普遍分布*pa*中得来的（*pa*的来源例如收集公开的人脸图像）。为进一步提高重建数据的质量，本文选辅助集时会有一些偏好（如：面部识别，会尽量选择正面的人脸图片）。

**B. Truncation Method for Model Inversion**

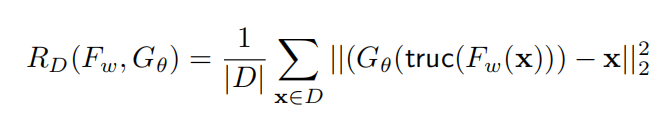
思想：将分类器对辅助样本的预测截断为与受害者部分预测相同的维数，并将其作为输入训练反演模型，迫使反演模型从截断预测中最大限度重构输入数据。（而以往的方式直接将不分预测输入到反演模型中，得到的重构结果是无意义的）。



和上面框架思路基本一致，训练Gθ时通过的是最小化重构损失函数R（最上方），而训练*Fw*用的是最小化分类损失函数L（中间）。

**C. Joint Training of Classififier and Inversion Model**

敌手为开发者时，联合训练分类器和反演模型，以获得更高反演精度。为分类损失函数L加上一个重构损失函数（l2范式）进行正则化：



**Experiment**

**1. 辅助集的选择对精度的影响**

用三种数据集进行实验--

same：和分类器训练集D相同

generic：从更普遍的分布中得到

distinct：和D同分布但不重叠

三类辅助集在FaceScrub数据集上反演模型的输出如下：



可以看出generic精度很高，distinct很低（因为提供的信息不够）。

结论：No knowledge的情况下，通过从更普遍分布中得到的辅助集可以保证反演精度，这样的辅助集也容易获得。

**2. 截断的影响**

m指保留预测向量中最大的m个值。

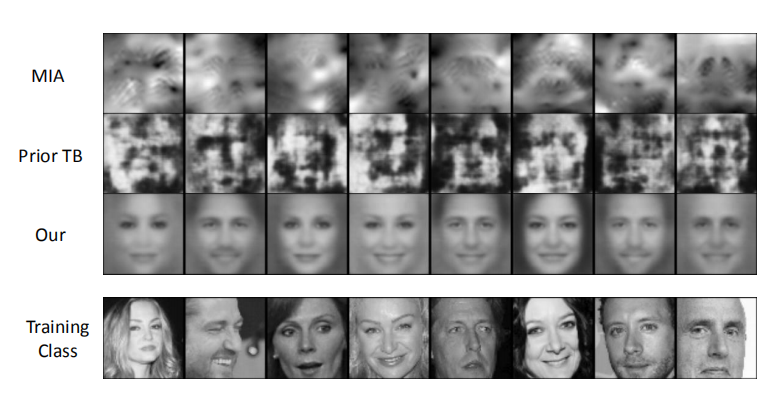
对比本文与之前的方法（用generic辅助集）：



对不同m，本文方法表现的都更好。

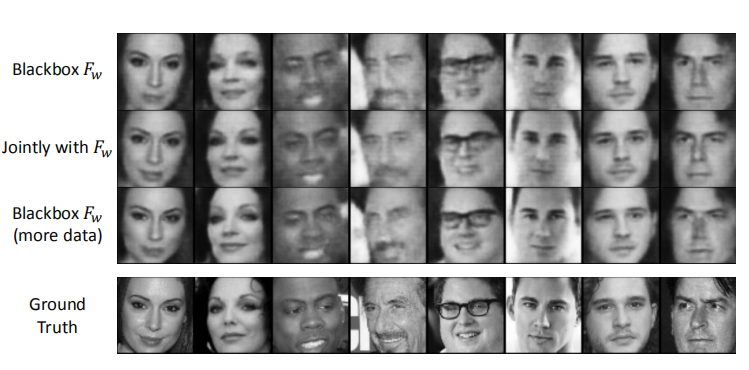
结论：只有部分预测值时，敌手使用本文训练方法依然可以执行精确反演。

**3.和以往反演方式做比较**



结论：本文方式优于其它方式，结果更具识别率（清晰度高）和代表性（都是正脸的图像）。

**4. 反演模型黑盒独立训练（场景1,2）和白盒联合训练（场景3）效果对比**



结论：联合训练效果更好，但仍有精度损失；增大辅助集后，黑盒训练准确率更高。