



華東師範大學
EAST CHINA NORMAL UNIVERSITY

第 12章 人工智能安全

- 基于人工智能的信息安全技术
- 人工智能的安全

基于人工智能的信息安全技术

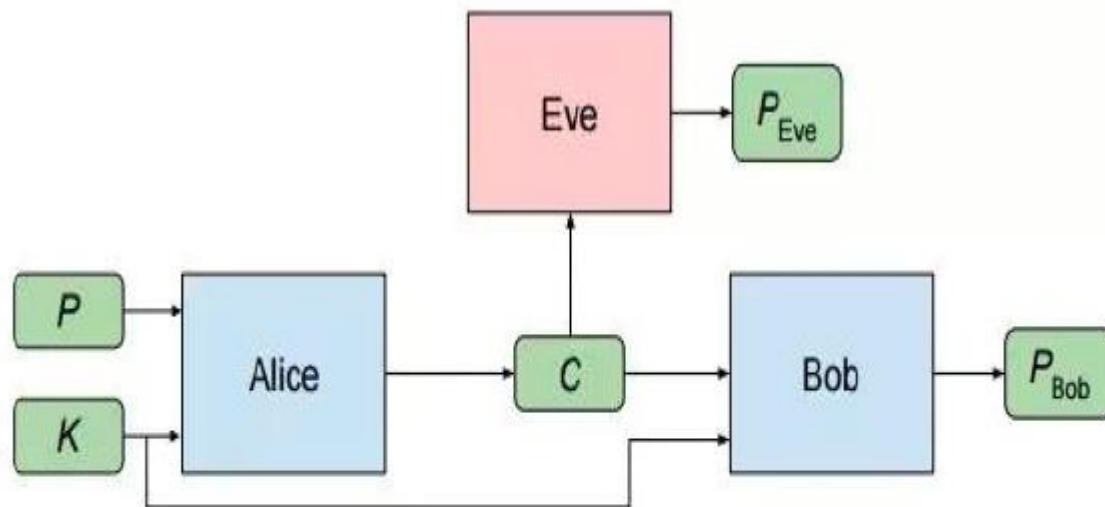
■ 加密技术

- 将明文信息处理为难以读取的密文内容，使之不可读。
- 在网络环境中保障通信安全，保证数据的完整性
- 目前常用的加密算法有安全哈希算法（Secure Hash Algorithm, SHA）和高级加密标准（Advanced Encryption Standard, AES）

■ 使用神经网络的加密算法

- 2016年谷歌大脑的研究团队提出了使用对抗生成网络生成的一个加密算法，其使用了三个神经网络分别完成加密、解密和攻击的工作，以保证通信双方信息的无损传输以及第三方无法破译通信内容。

基于人工智能的信息安全技术



加密系统架构。P =输入的明文，K =共享密钥，C =加密文本，PEve和PBob 为经过计算后得出的明文输出。

learning to protect communications with
adversarial neural cryptography

基于人工智能的信息安全技术

■ 数字水印

- 将特定信息（版权信息等）嵌入在数字信号中，数字信号可能是音频、视频、图片等。
- 当拷贝信息时，水印内容会被同时拷贝，所以水印内容可以作为版权信息的证明，这样能避免或阻止数字媒体未经授权的复制和拷贝

■ 近年来通过神经网络来添加水印和提取水印信息的成为学术研究热点。

基于人工智能的信息安全技术

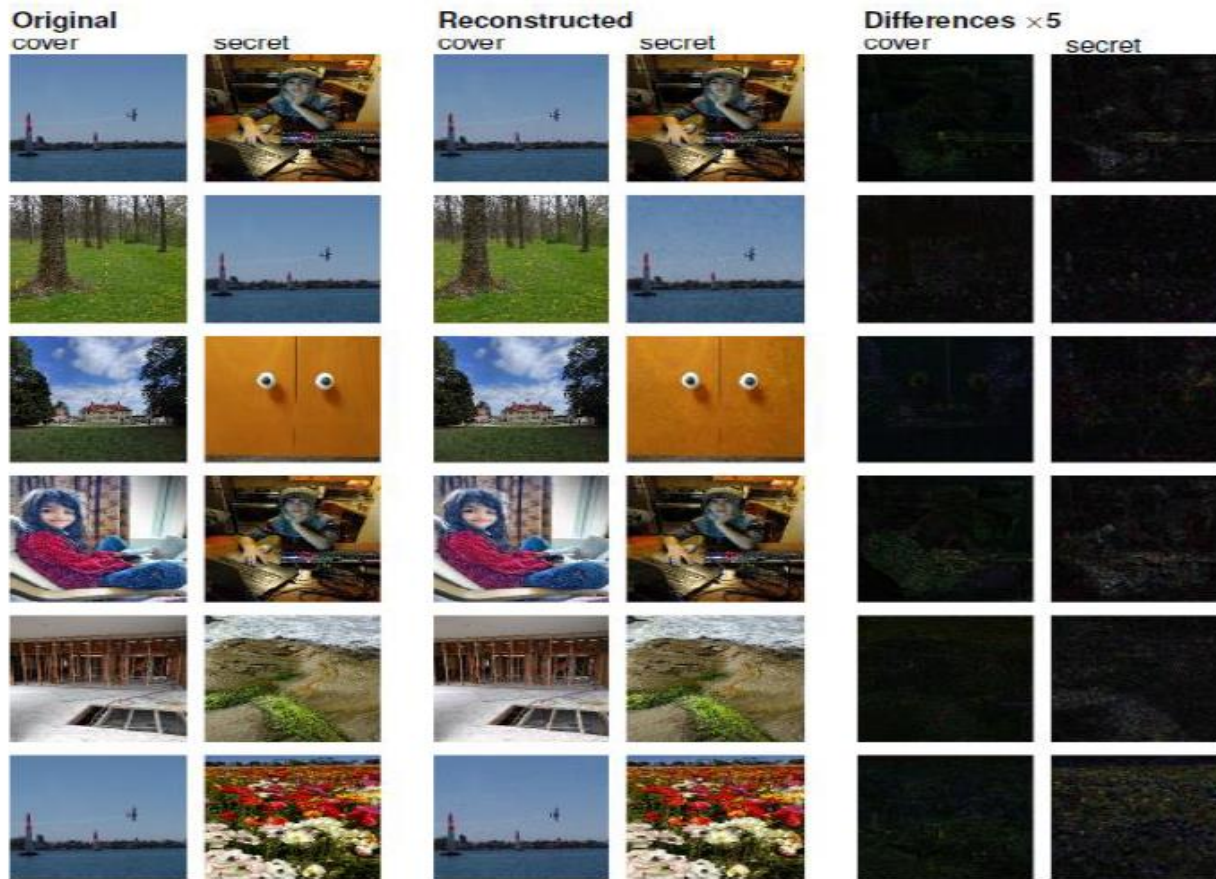


Figure 5: 6 Hiding Results. Left pair of each set: original cover and secret image. Center pair: cover image embedded with the secret image, and the secret image after extraction from the container. Right pair: Residual errors for cover and hidden – enhanced 5 \times . The errors per pixel, per channel are the smallest in the top row: (3.1, 4.5) , and largest in the last (4.5, 7.9).

Hiding Images in Plain Sight: Deep Steganography

人工智能的安全

■ 数据安全

- 人工智能很大程度是依靠数据驱动学习
- 可用性 (availability)
 - ✓ 训练数据是否充足且可靠
 - ✓ 训练数据是否有足够的标注
- 完整性 (completeness)
 - ✓ 数据是否具有代表性
- 隐私性 (privacy)
 - ✓ 数据是否涉及隐私安全问题
 - ✓ 如何保障数据不被窃取

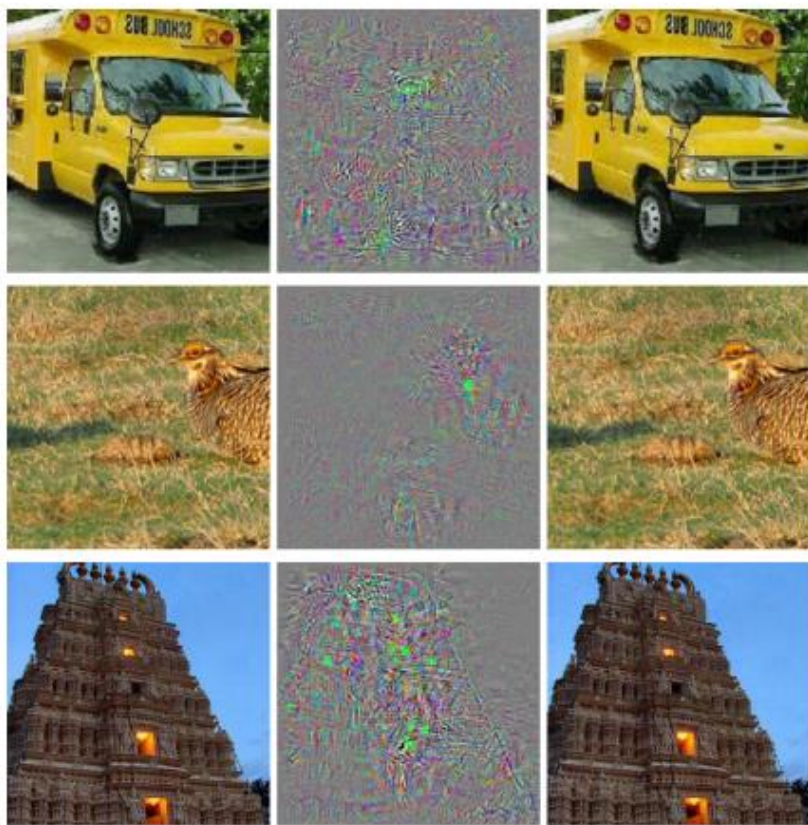
人工智能安全

■ 模型安全

- 人工智能所使用的的模型是由有限的训练数据训练得到的
- 鲁棒性 (robustness)
 - ✓ 模型是否易于受到噪声干扰或攻击
- 正确性 (correctness)
 - ✓ 模型是否正确
- 通用性 (generality)
 - ✓ 模型是否能够应用于现实场景
 - ✓ 模型对输入数据是否有过高的要求

人工智能的安全

■ 对抗样本



(a)



(b)

对抗样本

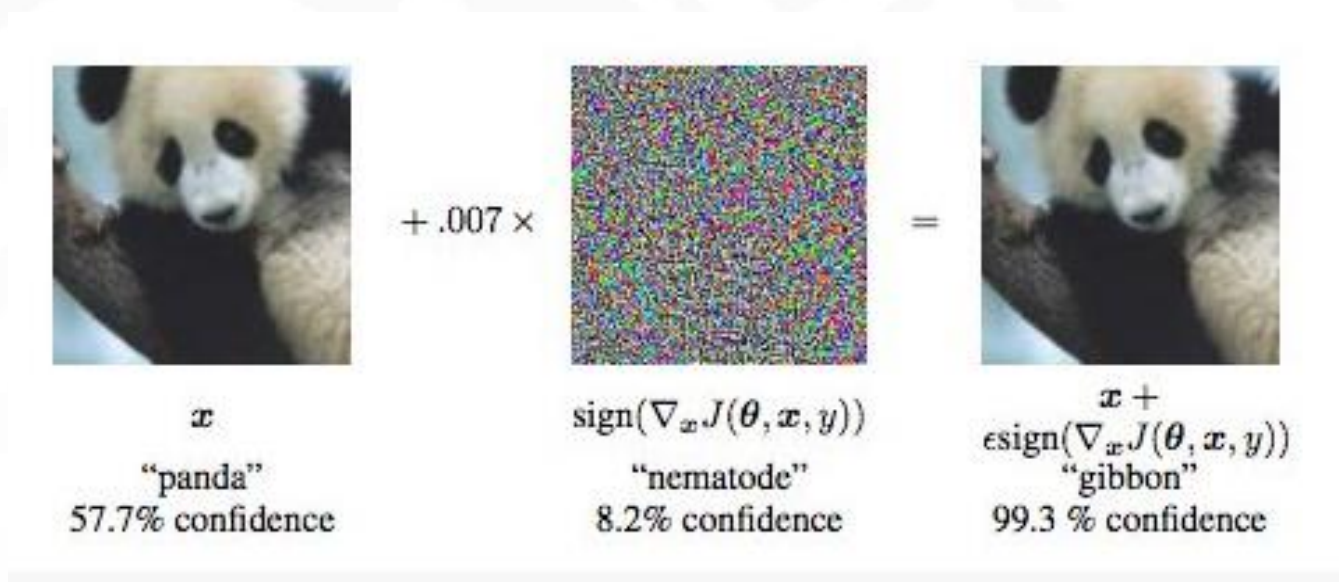
- 从2013年开始，深度学习模型在多种应用上已经达到甚至超过人类水平，比如人脸识别，物体识别。在之前，机器在这些项目的准确率很低，如果机器识别出错了，没人会觉得奇怪。
- 但是现在，深度学习算法好了起来，去研究算法犯的误差变得有价值起来。其中一种误差叫对抗样本（adversarial examples）
- 使用特定技术对输入样本进行微小的修改就可骗过模型而得到错误的结果
- 这种经过修改，使得模型判断错误的样本被称为对抗样本

对抗样本

- 对抗样本是机器学习模型的一个有趣现象，攻击者通过在源数据上增加人类难以通过感官辨识到的细微改变，但是却可以让机器学习模型接受并做出错误的分类决定。
- 一个典型的场景就是图像分类模型的对抗样本，通过在图片上叠加精心构造的变化量，在肉眼难以察觉的情况下，让分类模型产生误判。

对抗样本

- 一张原先分类为panda的图片，经过非常细微的扰动后，生成最右边的图片，按我们看这依然是panda，但却被学习器误分成另一个类（gibbon）。



对抗样本

■ 对抗样本的危害



图2 路标识别的攻击示意图^[32]

对抗样本

■ 对抗样本产生的原因

- 神经网络中包含语义信息的部分并不是在每个独立的神经元，而是整个空间
- 神经网络学习到的从输入到输出的映射在很大程度上是不连续的
- 深度学习在高维空间的线性特性已经足以产生这种攻击行为

对人工智能模型的攻击

■ 白盒攻击

- 攻击者熟知人工智能模型的算法和模型参数，生成对抗样本的过程可以与模型的每一部分进行交互

■ 黑盒攻击

- 攻击者只能给定输入去获得模型输出，但并不知道被攻击模型所使用的算法和参数
- 黑盒攻击可以针对任何一个人工智能模型

白盒攻击

- 对人工智能模型的白盒攻击通常会对模型的每一部分进行逐层分解，然后对每一部分添加一定的扰动，使得模型的结果逐步向误判目标类别偏移
- 这是一种非常隐蔽的攻击手段，通过限制扰动的大小可以使得对抗样本看起来与原样本差别很小



识别结果：

女性

男性

白盒攻击的防御策略

■ 生成对抗网络

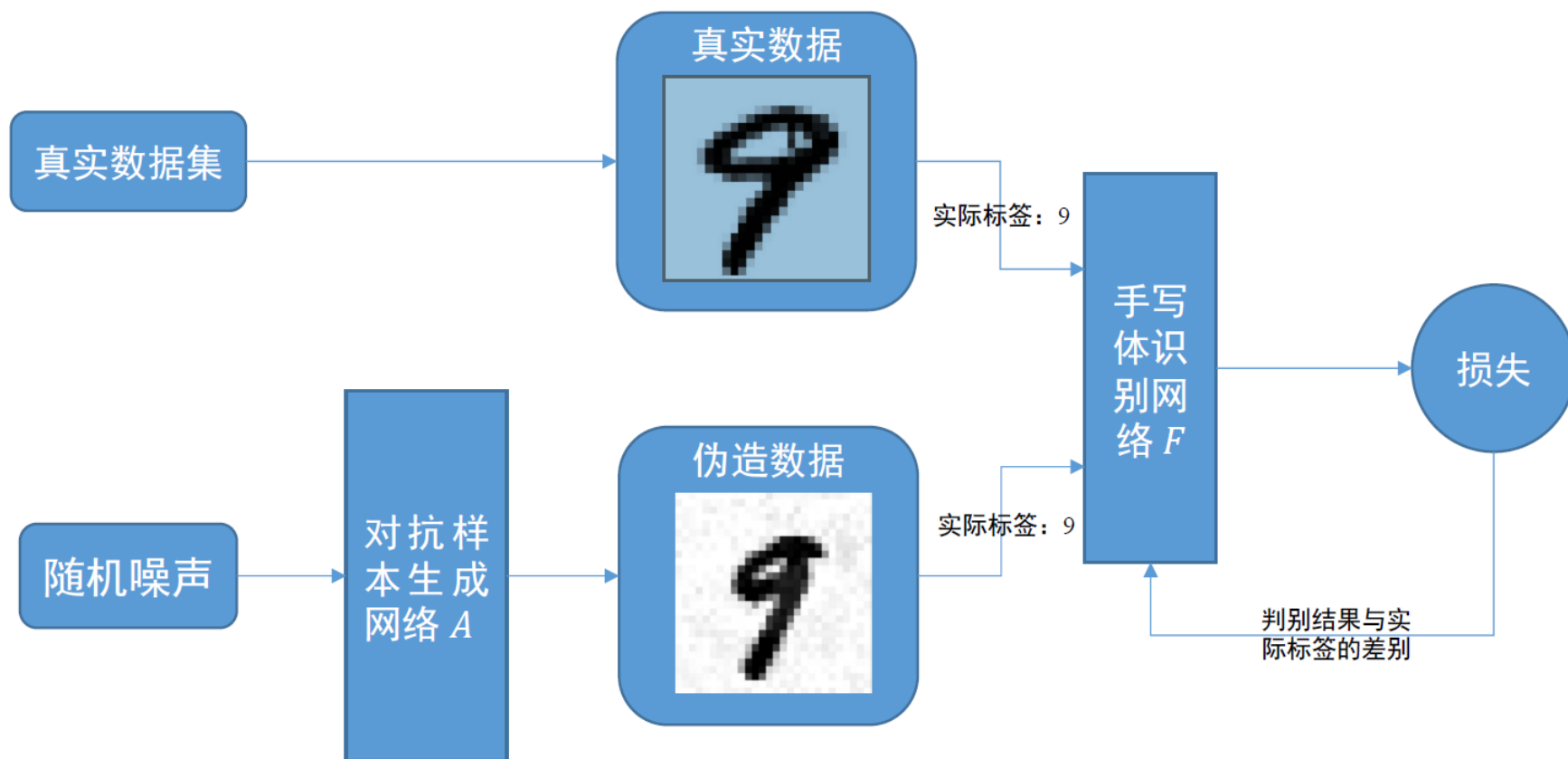
- 由于白盒攻击的过程是对模型内部增添扰动实现的，所以在训练时可以使用同样的方法增强模型训练的鲁棒性
- 如生成对抗网络（generative adversarial network, GAN）就是一种有效的抵御白盒攻击的手段

■ 生成对抗网络实际上由两个不同的网络组成：

- 生成网络：通过神经网络将输入的一个服从简单分布的随机变量转化为能够欺骗判别网络的对抗样本
- 判别网络：通过神经网络判断输入样本的真实类别
- 训练时两个网络交替进行参数优化，在对抗过程中共同提升性能

在模型训练时，生成网络负责生成对抗样本、判别网络（即我们真正需要的网络）对样本类别进行判断。在这一过程中，生成网络所生成的试图欺骗判别网络的对抗样本会被判别网络识破，从而达到防御白盒攻击的目的

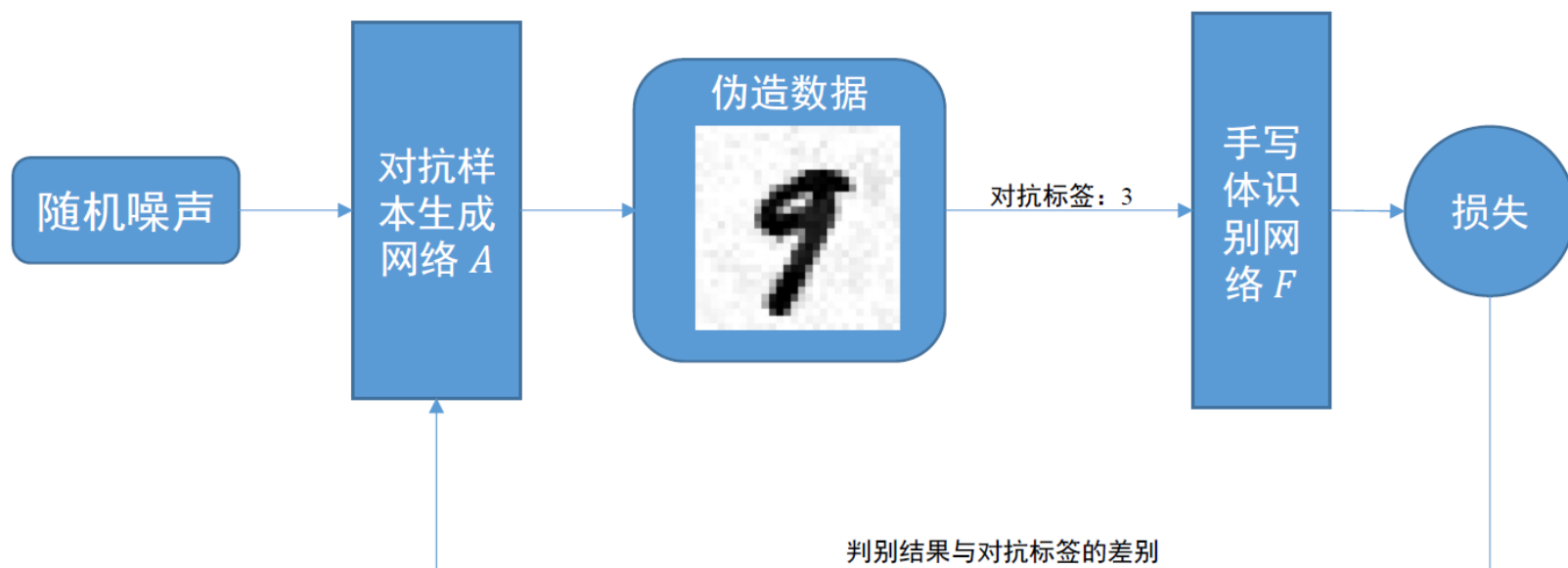
判别网络训练过程



生成网络训练过程

- 生成网络的训练过程是判别网络的对抗过程
 - 根据判别网络识别的结果，不断提升对抗样本生成网络合成对抗样本的能力，从而使其能够产生更具有误导性的对抗样本
- 此时希望合成的伪造数据被识别为伪造的对抗标签而不是合成数据所对应的实际标签
- 在该过程中，判别网络参数保持不变。

对抗样本生成网络训练过程



判别网络训练过程

■ 判别网络的训练过程分为两个方面

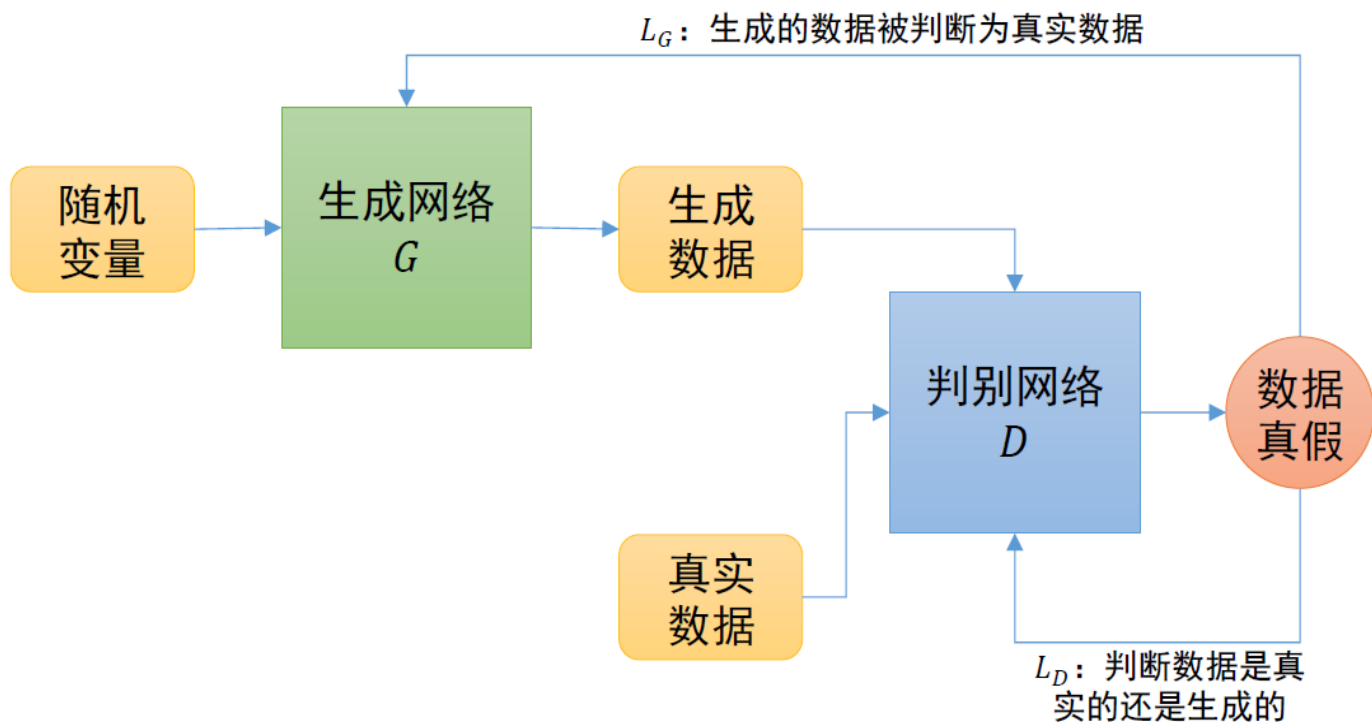
- 根据真实数据集的数据和其真实标签来增强判别网络识别数据的能力
- 根据对抗样本生成网络合成的伪造数据来增强判别网络抵抗干扰的能力

■ 此时不论是真实数据还是对抗样本，算法都希望判别网络输出结果与图片标签一致。

■ 在上述过程中，对抗样本生成网络参数保持不变

生成对抗网络

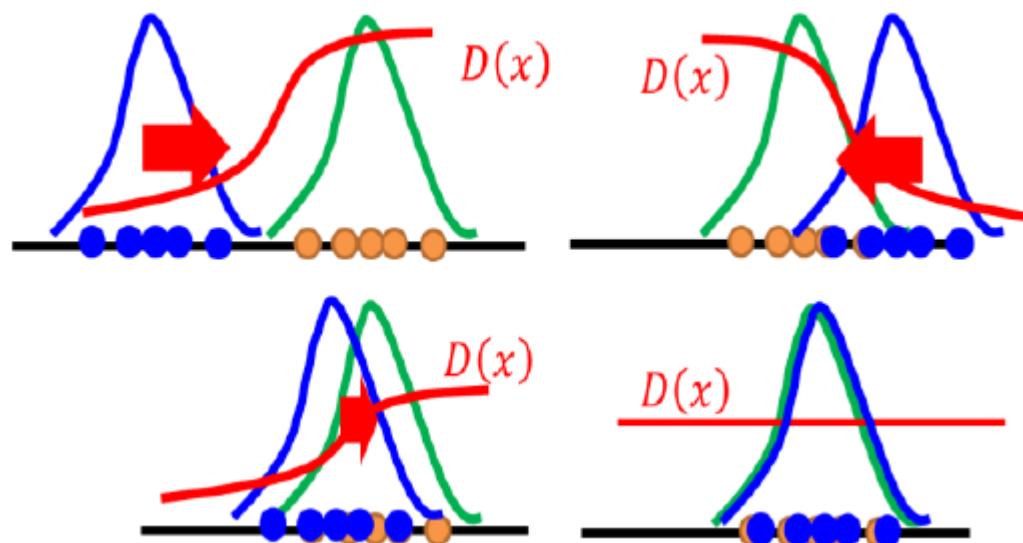
- 生成对抗网络是深度学习中常用的一种生成模型（generative model）
- 生成对抗网络一般可如下表示：






生成对抗网络

- 生成网络通过简单的分布来拟合复杂的分布，并从所拟合的分布中采样得到符合一定要求的样本
- 在训练过程中，判别网络需分辨出真实数据与生成数据的不同、生成网络会逐渐学习得到数据的真实分布情况
- 随着不断优化生成网络，判别网络逐渐无法分辨生成网络所合成数据的真伪
- 最后，生成网络完全模拟出了真实数据的分布情况，使得区别网络无法分辨数据的真伪，开始随机猜测结果，此时对抗网络的训练达到收敛

生成对抗网络



-  分类效果
-  真实数据分布
-  生成数据分布

生成对抗网络

生成对抗网络有两个优化目标：生成网络和判别网络。首先优化判别网络模型参数

- $G^* = \operatorname{argmin}_G \max_D V(G, D)$

- $V = \underbrace{E_{x \sim P_{data}} [\log D(x)]}_{\text{判别网络目标}} + E_{x \sim P_G} [\log(1 - D(x = G(z)))]$



- 判别网络目标最大化 $\log D(x)$

- $P_{data}(x) \log D(x) + P_G(x) \log(1 - D(x))$

- 对 $D(x)$ 求导可得最优分类器：

- $D^*(x) = \frac{P_{data}(x)}{P_{data}(x) + P_G(x)}$



最优的分类器能准确区分真假样本

生成对抗网络

生成对抗网络有两个优化目标：生成网络和判别网络。接着优化生成网络参数

- $G^* = \operatorname{argmin}_G \max_D V(G, D)$

- $V = E_{x \sim P_{data}} [\log D(x)] + E_{x \sim P_G} [\log(1 - D(x = G(z)))]$



- 生成网络的优化目标是最小化 $\log(1 - D(G(z)))$ ，将最优的判别网络代入这一优化目标

- $$\begin{aligned} V(G, D^*) &= E_{x \sim P_{data}} \left[\log \frac{P_{data}(x)}{P_{data}(x) + P_G(x)} \right] + E_{x \sim P_G} \left[\log \frac{P_G(x)}{P_{data}(x) + P_G(x)} \right] \\ &= -2\log 2 + \text{KL} \left(P_{data}(x) \parallel \frac{P_{data}(x) + P_G(x)}{2} \right) + \text{KL} \left(P_G(x) \parallel \frac{P_{data}(x) + P_G(x)}{2} \right) \\ &= -2\log 2 + 2JS(P_{data}(x) \parallel P_G(x)) \end{aligned}$$

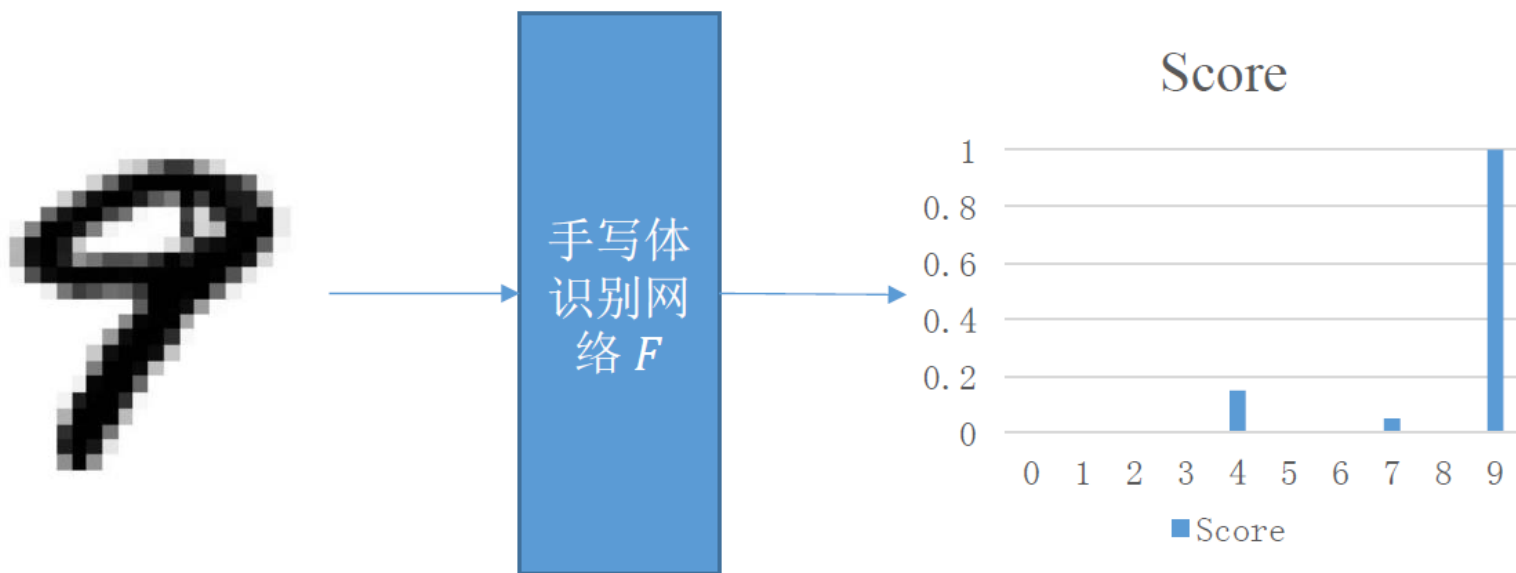
- $G^* = \operatorname{argmin}_G D_f(P_{data} \parallel P_G)$



最优生成网络能生成与真实样本具有相同分布的数据

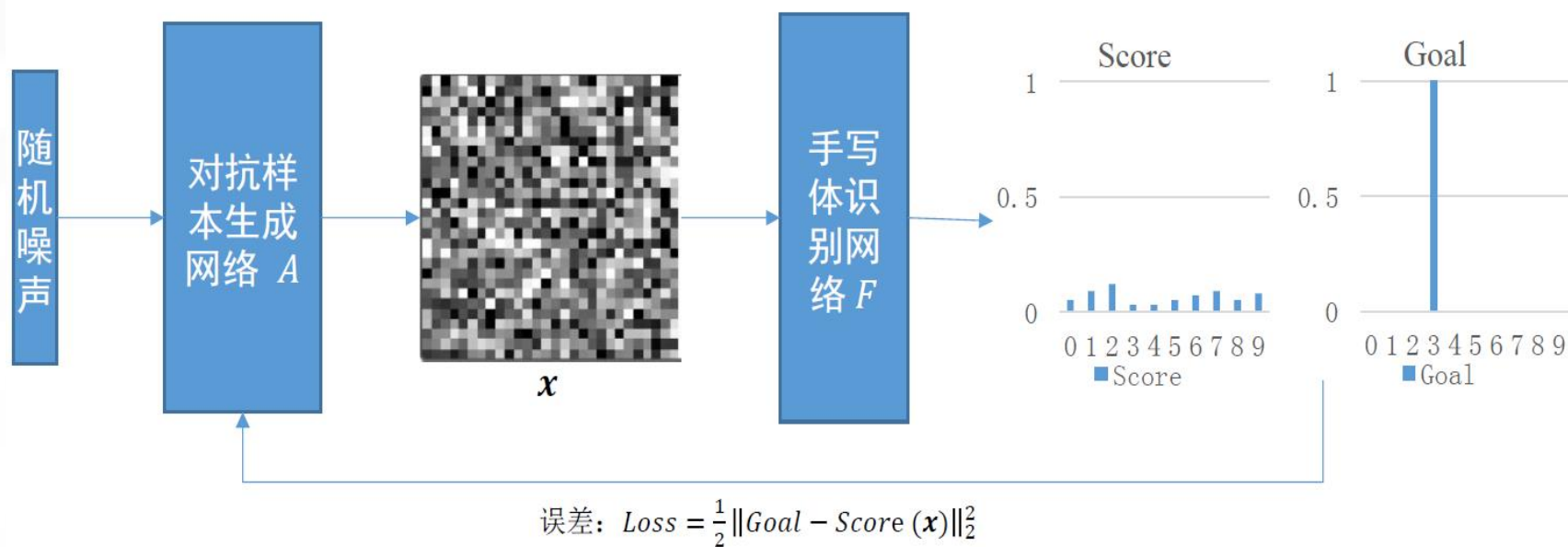
无针对攻击 (Non-Targeted Attack)

- 无针对攻击：任意生成输入数据，使得模型输出为指定结果
- 假设已经获得一个训练好的神经网络 F ，能够识别手写数字。现在希望生成能够干扰神经网络 F 的对抗样本 i ，使得对抗样本 i 被错误识别为数字3



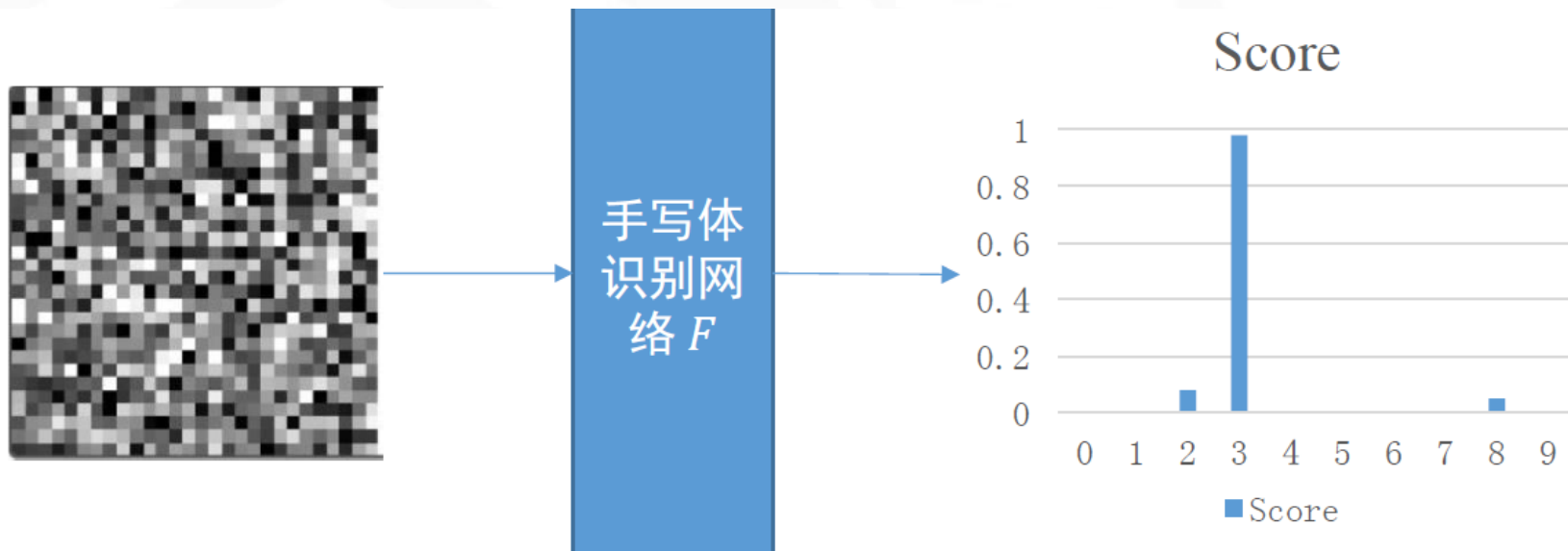
无针对攻击 (Non-Targeted Attack)

- 训练一个能够生成对抗样本的生成网络 A ，其能够将随机噪声转化为一副对抗样本图片
- 将对抗样本输入手写体识别网络 F ，使用 F 的输出与预设目标之间的误差来优化对抗样本生成网络 A



无针对攻击 (Non-Targeted Attack)

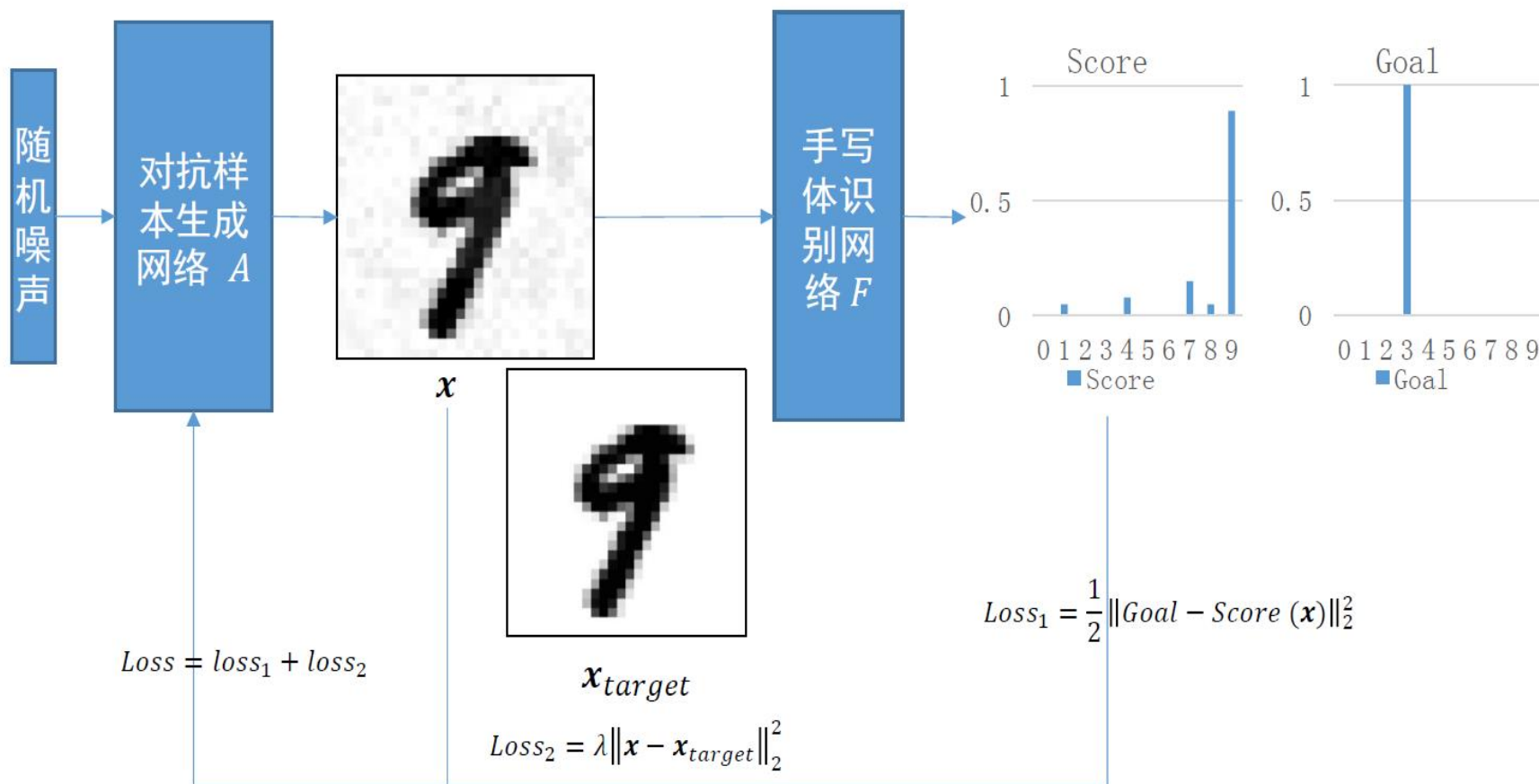
- 通过迭代训练，使得对抗样本生成网络 A 可生成众多被手写体识别网络 F 错误分类为3的对抗样本。
- 这样，手写体识别网络 F 被攻击成功。
- 在对抗样本的生成过程中，没有用到攻击模型的内部结构知识，所以这是一次黑盒攻击



有针对攻击 (Targeted Attack)

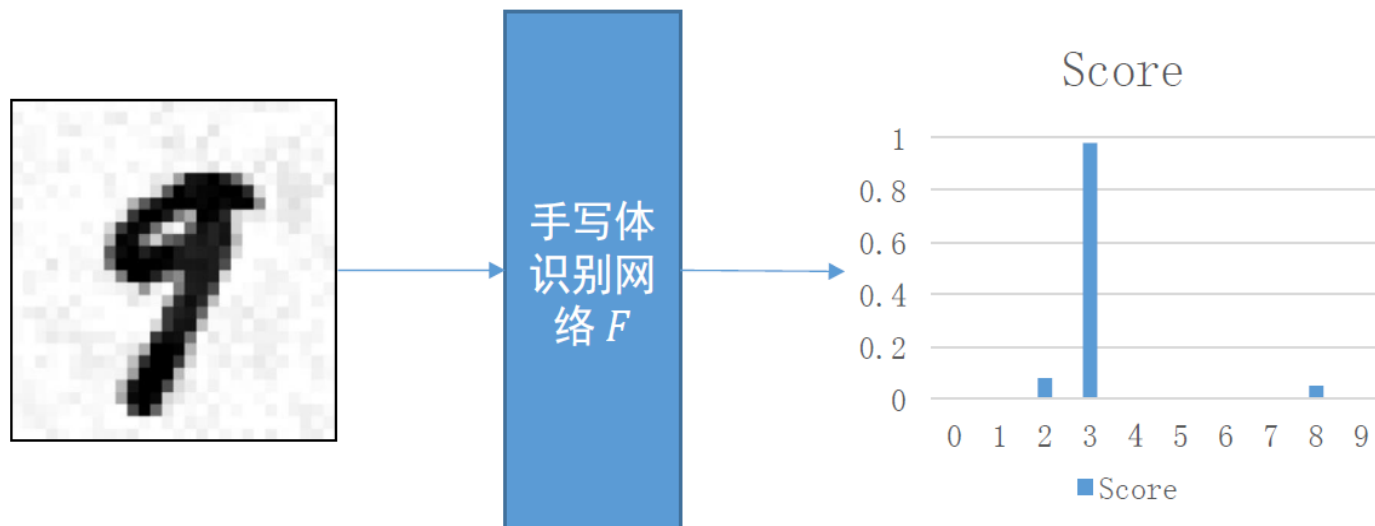
- 有针对攻击：生成人类与模型判断相互迥异的对抗样本
- 假设已经获得一个训练好的神经网络 F ，其能够识别手写体数字，现在想生成能够干扰神经网络 F 的有针对的对抗样本 i 。
- 如：对抗样本 i 被人识别为9，但被 F 错误识别为3。

有针对攻击 (Targeted Attack)



有针对攻击 (Targeted Attack)

- 这种攻击方式同样也是黑盒攻击。可见，手写体识别网络 F 被攻击成功。



黑盒攻击的防御策略

■ 常用的黑盒攻击防御策略有：

- 数据压缩：通过对输入数据进行压缩或者降维，在保证识别准确率的情况下提升模型对干扰攻击的鲁棒性
- 数据随机化：对训练数据进行随机缩放、增强等操作，提升模型的鲁棒性
- 训练额外的网络来判断训练数据是否为攻击样本



THE END