对抗性图像干扰保护隐私的博弈论视角

摘要：用户喜欢通过社交媒体与他人共享个人照片。同时，他们可能想使自动识别这类照片变得困难甚至不可能。诸如模糊之类的经典混淆方法不仅令人不快，而且也没有人们期望的那样有效[28，37，18]。关于对抗性图像扰动（AIP）的最新研究表明，可以有效地混淆识别系统而不会带来令人不快的伪影。但是，在存在针对AIP的对策的情况下[7]，尚不清楚AIP的有效性如何，特别是在对策选择未知的情况下。博弈论为研究战略中不确定因素之间的相互作用提供了工具。我们介绍了用于用户识别器动力学的通用博弈理论框架，并提出了一个案例研究，其中涉及了最新的AIP和人员识别技术。我们为用户提供了一种最佳策略，可确保识别率的上限不受识别者的对策影响。可以从https://goo.gl/hgvbNK获得代码。

## 1.Introduction

如今，人们通过社交媒体分享大量的个人照片。 个人照片包含丰富的私人信息，例如 有关家庭成员，旅行目的地和政治活动的信息。 再加上计算机视觉技术的最新发展[4、11、8、27、34]，这导致人们越来越担心采用计算机视觉技术的恶意实体可能会从视觉数据中提取私人信息。

传统的混淆技术（例如面部模糊和像素化）不仅令人不适，而且对基于卷积的识别器也无效[28，37，18]。

最近有关于对抗性图像扰动（AIP）的研究：在图像上精心制作的加性扰动使卷积混淆而人眼几乎看不见[36，6，21，20]。 AIP确实有希望作为混淆技术。

但是，采取反措施时，AIP是否仍然有效仍然是一个问题。 例如，[7]提出了简单的图像处理策略来应对AIP效果（例如少量模糊）。 如果此外还不清楚对策的具体选择，则最佳策略对用户而言并不明显。

当每个玩家的策略存在不确定性时，博弈论提供了有用的分析工具。 我们提出了一个博弈理论框架来描述一个系统，在该系统中，用户和识别器力求实现对立目标：禁用/启用识别器。 该框架可以从明确制定的假设中独立于识别者的对策来获得用户隐私级别的保证。 我们包括一个人识别博弈的案例研究，根据当前的最新AIP和人识别方法得出用户的隐私保证。

本文展示了博弈论在理解用户识别器动力学方面的效用。 该框架可以扩展到所考虑的特定设置之外。 我们认为，该框架将进一步帮助用户识别器进行更多不同的任务和设置分析。

我们列出的贡献如下：

* 一个用于研究用户识别器动力学的博弈理论框架。
* 对抗性图像摄动（AIP）作为对人迷惑的一种有效且美观的技术的应用。
* 新颖的健壮和识别器选择性AIP。
* 博弈理论框架的经验案例研究，为用户提供了隐私保障。

## 2.Related Work

**隐私和计算机视觉。** 传统上，由安全界主导的关于用户隐私的研究很多[22，39，23，19]，而关于视觉数据中的私人内容的研究只是最近才开始的[37，28，18]。

威尔伯等。 [37]研究了在多种面部混淆方法（模糊，变黑，伪装眼镜等）下商用面部检测器的性能。 哦，等。 [28]和McPherson等。 [18]研究了人脸识别性能。 特别是，[28]表明，当前的识别器可以适应混淆模式。 以上工作得出的结论是，识别器可以抵抗简单的混淆方法（例如面部模糊）。 在这项工作中，我们研究了更强的混淆类型：对抗图像扰动。

**对抗图像扰动（AIP）。** Szegedy等。 [36]首先研究了卷积网络对抗不稳定的现象：有可能产生完全使识别器蒙蔽的看不见的累加扰动。 最初的制作算法基于L-BFGS [36]。 已经提出了更有效的一阶算法[6、31、21、12]。 我们从概念和经验上回顾了现有的AIP算法和我们的新颖变体。

**针对AIP的稳健分类。** 一些卷积前的工作考虑通过训练对抗性数据来增强分类器的总体鲁棒性。 Lanckriet等。 文献[13]对对抗性数据进行了线性分类器的训练，这些数据被固定为每个类别的固定均值和协方差。 Brückner等。 [3]引入了博弈论的概念来规范对抗训练程序。 但是，他们将注意力集中在更简单的模型上：线性[13]或凸[3]。 这项工作建立在一个博弈理论框架上，该框架容纳了最先进的卷积模型。

自从有效的卷积网络[11]和相应的AIP算法[36]出现以来，一些工作[6，10]考虑使用AIP训练卷积网络，从而在一定程度上实现了针对AIP的鲁棒性。 另一方面，Graese等。 [7]认为简单的测试时间图像处理，例如翻译，高斯噪声，模糊和调整大小，可以等同地抵消AIP的影响，而无需重新训练卷积网络。 在我们的案例研究中，我们将这些图像处理方法包括在识别器的策略空间中。

**针对分类器的强大AIP。** Sharif等。 [32]提出了一种通过针对一组图像而不是单个图像优化AIP来实现鲁棒性的方法。 Moosavi等人也提出了这种方法。 [20]产生普遍的扰动。 在我们的工作中，我们考虑针对一组目标输入的抖动版本优化AIP。 我们将凭经验证明这可以针对图像处理策略进行有针对性的防御。

**AIP用于身份混淆。** 本文主张将AIP作为禁用识别的有效且美观的手段。 先前Sharif等人。 [32]也使用对抗性优化来欺骗人识别器。 与他们的有限设置（固定姿势，固定识别策略）相比，我们的案例研究涵盖了具有用户识别器动态的大规模社交媒体设置。

**人员识别任务。** 我们的案例研究考虑了社交媒体设置中的人识别任务[5，38，27]，而不是人脸识别[9]（正面人脸，光线充足）或行人重新识别[2，1]（低分辨率，固定 上下文）。 社交媒体照片可以捕捉各种视角，姿势，服装和事件中出现的主题。 张等。 引入了PIPA [38]，这是第一个大规模的社交媒体人识别数据集和基准。 我们的经验研究是建立在这个数据集上的。

**人识别模型。** 多个研究人员在社交媒体照片中提出了人识别技术。 张等。 [38]提出结合从体态检测获得的多个身体部位的线索。 哦，等。 [27]大大简化了[38]，同时实现了最先进的性能。 我们基于[27]建立识别器模型，可能使用更高级的网络体系结构。 Liu等人的并发工作。 [17]声称通过度量学习目标改进了该方法。 还有其他作品[14，28，15]，它们利用社交媒体元数据。

## 3.User-Recogniser Game（用户识别博弈）

本节提供了研究用户识别器博弈的通用框架。 该框架提供了一种工具，用于系统化从一组对玩家的明确假设到博弈论结论的路径。

我们的用户识别器博弈框架如图2所示。用户U根据策略i 2 u干扰原始图像x，旨在阻止识别。 识别器R根据策略j 2 r处理被摄动的图像ri（x），目的是抵消图像摄动的影响。 所得图像nj（ri（x））被传递到模型f进行预测。 博弈的产生是因为每个玩家虽然知道对方的策略空间，但他们并不知道对手的策略。

我们在第3.1节中介绍了相关的博弈论概念和关键理论结果，以帮助规范第3.2节中的框架。 我们将在第3.3节中讨论可能的扩展。

### 3.1. Two-Person Constant-Sum Games（两人恒和博弈）

我们将我们的系统描述为一个**两人博弈**[26]，它由两个玩家，用户U和具有**指定策略空间**0u和0r的识别器R组成。

由于每个玩家都参与策略i∈0u和j∈0r，因此R收到**回报**pij，即识别率； 然后，U收到误识别率1-pij的收益。

博弈论表明，有时将策略随机化会更好。 U可以采用**混合（随机）策略**0u =（0u i）i 0u，定义为策略空间0u上的分布，对于R同样如此。在滥用符号的情况下，我们写出p（0u; 0r）：= Pi; j0ui 当采用混合策略0u和0r aPre时，R的预期收益为0rj pij。 得出U的收益并将其定义为i; j0ui rj（1- pij）= 1-p（0u; 0r）=：p0（0u; 0r）。

我们说两人博弈是一个**常数和博弈**，如果玩家的收益和与策略无关就等于常数p。 在我们的情况下，识别率和误识别率总和为1（p = 1）。 如果策略空间是有限的，则博弈是**有限的**。 我们有以下最优性定理。

**定理1**（冯·诺伊曼[26]，1928年）。 对于有限恒和博弈，存在最优或最小极大混合策略0u \*和0r \*，使得p（u ?; r）p（u ?; r？）p（u; r？）8 u; r（1）其中v：= p（u ?; r？）是**博弈的价值**。

公式1表示，当R扮演0r \*时，无论U采取何种策略，都保证R至少具有v的收益； 如果U的下注为0u \*，则保证U的收益为1-v。在我们的方案中，这意味着U的最佳策略可确保一定的误识别率，而与R的策略无关。

通过解决以下问题的线性编程，可以有效地获得U的最佳策略（R的最佳策略可以通过交换最小和最大找到）：arg minumaxrXi;jui rj pij使得0u; 0r是分布的。

如果U对R的策略r有了解，则U可以利用此知识。 如果r达到最大up0（u; r）p0（u ?; r）p0（u ?; r？）= 1 pay，U可以优化她的策略。 v，可能比无知识情形1 v更好的回报。但是，如果R的策略是最优r = r ?，则该知识不会为U带来改进：max up0（u; r ？）= 1􀀀v。

在现实中，并不是所有的玩家都是由于缺乏知识(如对手的战略空间)或纯粹的非理性而发挥最佳。我们称这样的玩家为**非理性玩家**。我们以上的讨论意味着：

**推论1.**如果U知道R的策略0r，并且它不是次优的，那么U可以获得比1-v更好的收益。

### 3.2. Components of the User-Recogniser Game（用户识别器博弈的组件）

我们为用户U和识别器R指定收益，策略空间和信息。

**测试数据。** 我们假设测试数据是根据（^ x; ^ y）D分配的。此数据集是两个参与者竞争的私人信息的来源。

**固定型号。** 我们假设U和R使用固定模型f（例如，公开可用的模型）。 这是一个合理的假设，因为U和R通常没有资源来训练现代卷积网络。

**已知模型。** 每个玩家都知道对手使用f。 这可能是不现实的，但提供了一个很好的起点。 §3.3中讨论了对该假设的放松。

**清偿。** 当玩家采取策略i 2 u和j 2 r时，R的收益是测试集的识别率：

pij = P(^x;^y)\_D\_arg max y fy (nj (ri (^x))) = ^y\_(3)

其中fy表示类别y的模型预测得分。 U收到回报1􀀀pij，即误识别率。

**用户的策略空间0u。** 我们考虑加性扰动，使得对于输入x，ri (x) = x + t(x); jjt(x)jj2 \_ \_ (4) 对于一些常数e>0。当e足够小时，人眼几乎看不到扰动（见图3）。 这些扰动通常称为对抗图像扰动（AIP）。 我们将在§4中讨论现有的AIP及其新颖的变体。

**识别器的策略空间0r。** R旨在抵消AIP的对抗作用。 尽管一些工作建议用AIP重新训练模型，但显示出一定程度的鲁棒性[6，10]，Graese等人。 [7]认为简单的图像处理已经可以廉价有效地抵消AIP效应。 他们已经证明，在MNIST上，翻译（T），高斯加性噪声（N），模糊（B）和裁切和调整大小（C）已将识别率从0％（AIP后）提高到68％， 分别为58％，65％和76％。 在我们的案例研究中，我们将这些转换包含在0r中。 在第3.3节中，我们将讨论扩展策略空间。

**已知的策略空间。** 每个玩家（0u和0r）的策略空间是彼此已知的，而所选策略是未知的。

**多个识别器。** 您可能会遇到一组识别器，但并非全部都是恶意的。 例如，U用识别系统R1将她的个人照片上传到云服务； 她想要一个能够使R1成功识别但禁止恶意系统R2识别的AIP。 我们建议在§4.2中生成选择性AIP的方法，并在§5.5中确认它们的存在。 从理论的角度来看，选择性AIP的存在证明了可能的AIP模式的多样性，这与普遍扰动的存在一致[20]。

### 3.3. Extensions（扩展）

在上一节中，我们介绍了用户识别器博弈框架，并在本文中探讨了特定的假设。 在本节中，我们展示了可以将框架扩展到此设置之外。

**未知型号。** 许多AIP技术都假定对模型f有完整的了解，但是黑盒AIP的计算是另一个活跃的研究领域[29，30，24，16]；但是，许多关于AIX的技术仍在研究中。 U可以潜在地采用这些方法。

**非恒定金额。** 如果U和R为不同的测试样本分配不同的权重，则收益可能不等于1。对于这种非恒定和博弈，每个参与者都有Nash均衡策略[25]。 最佳策略和收益分析仍然是可能的。

**非可加性AIP。** 该框架允许ri是在图像上引起不可见变化的任何函数。 当前对等式4的限制排除了 整个图像的一像素平移。 大多数（如果不是全部）在AIP上的先前工作是在添加剂设置中完成的。 拟定非加性AIP将是未来的有趣工作。

**非固定型号。** 具有足够计算资源的R可以使用AIP重新训练模型f。 将我们的框架扩展到这种设置的一种选择是将模型参数合并到0r中。 Brückner等。 [3]研究了这种设置，但是假设了凸损失函数。 了解具有连续策略空间和非凸面收益（例如convnet损失）的博弈对于计算机视觉和博弈论研究都是一个悬而未决的问题。

**未知的策略空间。** 对手可能不知道确切的策略集。 随着技术的改进，各个策略空间甚至可能随着时间增长。 该框架不能对未知的策略做很多事情，但是可以根据技术发展自适应地扩展策略空间。

## 4. Adversarial Image Perturbation Strategies（对抗性图像扰动策略）

本节回顾了使用一阶优化方案的现有对抗图像扰动（AIP）算法，并提出了我们的新颖变体。

我们将AIP计算为具有L2范数约束（等式4）的加法变换。 可以将AIP的计算公式化为损失最大化问题max t L(f (x + t) ; y) s.t. jjtjj2 \_ \_ (5) 其中x是输入图像，y是地面真相标签； 将指定损失函数L。

### 4.1. Existing AIP methods

根据损失函数L和优化算法，我们恢复了大多数现有的AIP方法，例如快速梯度向量[31]，快速梯度符号[6]，基本迭代[12]和DeepFool [21]。 Moosavi等人介绍的普遍扰动。 [20]也可以看作是等式5的一种特例，其中在整个测试集上计算损耗，并且在图像之间共享扰动t。 摘要，请参见表1。

**快速梯度向量（FGV）**[31]。 FGV在等式5中采用softmaxlog损失L = -log ^ fy，通过一步梯度上升来求解：t？ =-rL（x）对于某些常数> 0。

**快速梯度符号（FGS）**[6]。 FGS与FGV相同，除了rL（x）被符号（rL（x））代替。

**梯度上升（GA）**。 这是FGV的多步骤变体。 扰动在t（0）= 0处初始化。对损失函数进行梯度上升：对于m = 0，t（m + 1）= t（m）-rL（x + t（m））。 ; K用于某些固定步长> 0和最大迭代次数K 1。

**基本迭代（BI）**[12]。 除了将rL（x）替换为符号（rL（x））外，BI与GA相同。

**DeepFool（DF）**[21]。 DF算法解决了目标：min t jjtjj2 s.t. arg max y fy (x + t) 6= y (6) 找到最小的扰动，使得预测是错误的。 尽管目标不同，但我们表明DF算法也可以看作是求解某些损失函数的方程5的一阶方法。

DF首先找到具有最近决策超平面的类，用c表示。 为了简化搜索，可在f围绕x的线性近似值（切线函数）上找到c。 决策超平面的法向向量由rfc􀀀rfy给出。 在每次迭代中，算法都会计算沿该方向的最小步长，以到达决策超平面。 由于f不是线性的，该算法可能需要多次迭代才能越过决策超平面。

我们观察到，如果将损失函数设置为L = fc􀀀fy，则梯度上升方向与DF阶跃方向rfc􀀀rfy匹配。 因此，我们将DF作为一种梯度上升算法，将每个步长最小化，以仅引起错误的预测。

**投影和剪裁。** 范数约束jj jj2以及RGB值约束为[0; 255]必须在解决方案上强制执行。 [16，12]建议在每次迭代后应用投影。 我们遵循这种做法。 对于BW图像，我们平均每个RGB通道的梯度。

### 4.2. Our AIP methods

正如我们将在§5.2中演示的那样，上述方法对于简单的图像处理技术是脆弱的。 我们在这里提出新的AIP方法，重点是鲁棒性。

**梯度上升-非GT（GAMAN1）中的最大值**。 即使通过AIP更改了预测标签，如果被干扰的输入仍接近决策边界，这也将不可靠。 DeepFool（DF）不会很健壮，因为一旦达到决策边界，它就会停止迭代。 另一方面，DF将解决方案引导至最接近的决策边界。 如果我们让DF以固定步长和固定迭代次数在决策边界之外进行迭代，则解决方案可能会更深入地进入错误标签的范围，从而提高鲁棒性。

这激励了我们的GAMAN变体。 代替每次迭代的c的昂贵计算，我们近似c y \*：= arg min y06 = y fy0，这是错误标签中最可能的预测。 我们将损失函数设置为L = fy \*-fy，并对K次迭代以固定步长r进行梯度上升。 这种方法与先前考虑的模拟AIP [32，16]类似，但与模拟AIP不同，后者将解决方案推向固定的模拟目标y。 相反，y \*可能会在迭代过程中发生变化。

**针对图像处理的疫苗接种。** 上述方法相对于固定识别器最大化分类损失函数。 为了应对AIP中和图像处理技术nj，我们考虑将图像处理步骤包括在损失函数L（nj（x + t））中。 可以使用上面考虑的任何一阶方法，只要nj是可微的。 如果处理函数是随机的，我们将平均多个样本的梯度。 我们将此技术称为疫苗接种。 注意，该技术是上述方法的补充。

**选择性AIP。** 我们提出了另一种互补技术，用于生成针对识别器的选定子集的AIP。 为了避免在授权B进行识别的同时获得M的识别，我们建议将混合损失最大化Xk2M\_kLk 􀀀Xk02B\_k0Lk0 (7) 与k; k0> 0。

## 5. Empirical Studies（实证研究）

我们已经建立了一个博弈理论框架来研究用户U和识别器R之间的动态。特别是，研究了以前的对抗性图像扰动（AIP）技术，并提出了新的变体。

在本节中，我们将对人识别框架进行案例研究。 在介绍博弈理论分析之前，我们先评估现有和新提出的AIP技术（第5.2节）的性能以及R的图像处理策略0r（第5.3节）的有效性。 在指定U的策略空间后，将引入完整博弈（第5.4节）； 我们将深入研究该系统。 最后，我们在识别器选择性AIP（第5.5节）上显示了结果。

### 5.1. Dataset and Experimental Setup

**数据集。** 我们基于PIPA（相册中的人物）[38]建立分析，PIPA是从Flickr抓取的大规模社交媒体照片数据集。 我们使用PIPA的val1子集作为测试集，该子集由366个身份（在[27]术语中为val-original split 1）的4820个实例组成。 我们假设用户将裁剪后的头像上传到社交媒体； PIPA提供了GT流浆箱。

**人识别器。** 人识别模型f建立在最先进的框架上[27]。 首先，它在庞大的随机身份数据库上训练用于人识别任务的卷积网络； 然后，使用每个身份大约十个示例，将最终分类层调整为测试身份。 在我们的例子中，我们使用了val0，它的大小和身份集与val1相同。 尽管[27]仅考虑了AlexNet [11]，但我们也考虑了VGG [33]，Google-Net [35]和ResNet152 [8]。 它们显示出更好的识别率（表2）。

**评估。** 我们根据测试集中正确识别的实例的比率来评估R的收益。 U的收益是1减去R的收益。 在所有表中，R是列播放器，U是行播放器。 对于每列（行），U（R）的最佳策略都标记为橙色（蓝色）。

### 5.2. Comparison of Perturbation Methods（摄动方法比较）

**AIP参数。** 除非另有说明，否则我们在所有实验中均将e设置为1000。 对于GoogleNet输入224x224，这相当于被1/256干扰的像素的2％。 对于梯度上升（GA）和基本迭代（BI），步长r设置为104； 对于GAMAN，5 103。 我们将最大迭代次数设置为K = 100，这样可以确定大多数测试样本的范数在K次迭代中达到e = 1000。

**基线摄动方法。** 我们考虑三种常用的混淆类型：噪声，模糊和眼线。 噪声加上方差iid高斯噪声； 模糊执行ab大小的高斯核的卷积； 眼图条在上部1/3位置放置了一条灰色的水平条e，其厚度为e。 即使使用较小的an，ab和ae，它们也会导致距原始图像较大的L2距离（> 1000）。 在表2中，我们通过计算jj'jj2> 1000时的识别率，报告了jj'jj2 = 1000时识别率的下限。

**AIP性能。** 我们首先针对所有网络变体评估所有考虑的AIP方法。 表2显示了结果。 我们观察到噪声，模糊和眼图条对小L2扰动的识别性能几乎没有影响。 AIP变体表现出更好的混淆性能。 总体而言，香草渐变比带符号的版本具有更好的混淆效果； 在AlexNet上，快速渐变矢量（FGV）上的识别率降低到13.3，而快速渐变符号（FGS）则为23.6； 多迭代类似物的行为与Gradient Ascent（GA）达到0.2的行为相似，而Basic Iterative（BI）则为1.2。 最后，我们观察到DeepFool（DF）和GAMAN（第4.2节）非常有效，将识别率降低到零。

**网络性能。** 比较体系结构，我们发现与最新的体系结构相比，AlexNet在AIP方面出奇地强大。 例如，与没有AIP的Alexnet相比，GoogleNet的性能要好得多（分别为83.8和87.8）； 当使用FGS时，AlexNet执行23.6，而Google-Net执行5.9。 使用多迭代AIP时，体系结构选择不会产生重大影响。 在下一个实验中，我们选择GoogleNet。 它的性能合理，但比ResNet快得多。

### 5.3. Robustness of AIPs

**基本处理**Proc. 甚至在R的图像处理策略开始之前，就需要（1）将受干扰的图像调整为原始图像的大小（根据网络输入大小），并（2）将其量化为整数值（例如24位真彩色）。 我们将以上两个基本处理步骤表示为Proc。

**图像处理策略**为0r。 我们为案例研究完全指定了R的策略空间。 继Graese等。 [7]，我们考虑0r = fProc; T; N; B; C; TNBCg。 Proc是上面描述的基本处理，所有其他策略都应用于Proc。 T是在图像边长的10％以内的随机偏移量进行平移。 N添加方差为a2 = 102的iid高斯噪声。B使用从f1中选择的宽度的高斯核来模糊。 3; 5; 7; 9g均匀随机。 C裁剪时会在图像边长的10％以内产生随机偏移，然后将其调整为原始大小。 对于每种策略，识别器将来自五个随机样本的分数进行汇总。 我们还考虑了所有四个的组合（TNBC）。 它在每个处理的图像上运行模型四次，在原始图像上运行一次； 然后将分数平均。

**AIP的鲁棒性。** 表3显示了使用R的处理策略时GoogleNet的识别率。 尽管多迭代AIP无需任何处理即可产生零识别率，但Proc已展现出强大的中和效果：梯度上升（GA）和DeepFool（DF）的识别率分别从零跃升至8.6和51.8。 DF的不稳定是由于提前停止（第4.1节）。 R的处理策略进一步提高了识别率。 一般而言，模糊B和裁剪C策略对AIP的危害比翻译T和噪声N更大。 在AIP方面进行比较，我们表明，我们的新变体GAMAN（第4.2节）在除N之外的所有处理策略中均占主导地位。 在这种情况下，GA的效果更好，但数量很少（分别为14.1和15.0）。 随后的分析基于GAMAN。

**定性的。** 该方法的定性示例如图3所示。图像和预测结果在Proc之后。 与模糊处理相比，GA和GAMAN可靠地导致了错误识别，而没有牺牲美观性。

### 5.4. User-Recogniser Games

**疫苗接种策略**0u。 响应识别器R的处理策略，用户U可以针对预期的处理类型对AIP进行疫苗接种（第4.2节）。 我们考虑六个变体0u = fGAMAN; = T; = N; = B; = C; = TNBCg。 我们使用斜杠=表示对GAMAN进行了疫苗接种。 对于= T; = N; = B; = C，来自5个随机函数样本的梯度在每次迭代中平均。 组合策略= TNBC分别从各个方法中得到4个梯度和1个原始梯度，从而对所有疫苗变体计算出相同数量的梯度。

**疫苗接种有帮助吗？** 表4列出了所讨论的加工和疫苗接种策略相结合的GoogleNet识别率。 我们确实观察到，每种疫苗接种类型均使香草AIP GAMAN相对于各自的加工类型更为稳健：对于B，接种率从22.2降至5.8。 / B是U对抗除N以外的所有处理策略的最有效策略。对于N，相应的疫苗接种/ N产生的U收益最佳。我们推测这是因为噪声N导致高频模式，而其他噪声则使噪声平滑。 输出。 最后，我们观察到，联合接种/ TNBC不能最有效地针对所有加工类型制备AIP。 给定梯度计算数量的预算，很难在所有方面都做得很好。

**最佳确定性策略。** 我们可以将表4视为策略i 2 0u和j 2 0r的R的回报表pij。 首先，假设玩家只选择固定策略。 然后，用确定性约束0u i求解方程2； 0r j 2 f0; 1g可产生U / B的最佳策略，并保证最高8.6的识别率。

**最佳随机策略。** 博弈论表明，有时将策略随机化会更好。 求解无积分约束的方程2时，U和R的最优解分别为0u \* =（/ B：61％; / TNBC：39％）和0r \* =（N：52％; B：48％）。 播放0u \*可确保U最多允许7.3识别率，这比确定性情况8.6更好。

**了解R的策略。** 如第3.1节所述，如果R不能发挥最佳策略，那么了解R的策略可以提高U的收益范围。 让我们考虑一下R扮演的两种可能的非最优策略。（1）如果R提交给B，则U的最优策略是B列中的最小行：/ B，识别率5.8。 （2）如果R在0r上均匀地随机化，则U的最佳策略是列平均数上的最小行：/ B，识别率为3.4。 在这两种情况下，U的识别率都较低。

**关于0r的知识有限**。 假设U不了解R拥有的所有可能的技术。 例如，U不知道策略N。然后，U的最佳解决方案是（/ B：100％），她认为这将确保她的识别率最高为5.8。 然后，R可以用N攻击U，从而获得8.6的识别率。 对对手的战略空间了解有限确实会造成伤害。

### 5.5. Selective AIPs（选择性AIP）

我们假设U希望避免由一组恶意识别器M进行识别，而由良性识别器B授权进行识别。我们在表5中进行了实验。我们将GAMAN在GoogleNet上的性能作为基线（第一行）。 对于所有k ∈ M∪B，我们用Ak = 1求解方程7，以生成选择性AIP。

当M = {GoogleNet}和B = {AlexNet}时，生成的AIP对M（在Proc之后）仅产生8.7的标识，而允许B标识97.9％。 因此，我们确认了选择性。 但是，与仅需要混淆M（4.0）的情况相比，这是以提高M（8.7）的识别率为代价的。

我们还考虑了M = {AlexNet，ResNet}和B = {VGG，GoogleNet}给出的多M，多B情况。 在Proc之后，M的平均性能为17.7，B的平均性能为97.7。 因此，选择性适用于多个模型，但对MARE的识别率仍然很高（17.7）。 我们注意到，通过将摄动大小的预算从1000增加到2000，我们仍然可以达到较低的比率：3.8。

选择性AIP的存在不仅具有实用意义，而且具有理论意义。 他们表明，AIP的空间足够多样化，可以容纳同时阻碍和协助识别的模式。

## 6. Discussion & Conclusion

**博弈论方法。** 博弈论是一种可以克服玩家选择不确定性的工具，可以独立于对手的策略提供收益保证。 博弈论还表明，如果没有一种能最好地应对所有可能的对抗技术的技术，则最好对现有技术进行随机化处理。

如第3.3节所述，本文介绍的博弈论框架可以扩展到其他设置，其中对每个玩家的资源约束较少。 本文是迈向分析用户识别器动力学的有前途研究方向的第一步。

**结论。** 在这项工作中，我们构建了一个博弈论框架来研究具有两个目标，即用户U和识别器R的两个系统，这些系统具有敌对目标（禁用/启用识别）。 我们已经研究了针对U的现有和新的对抗性图像扰动（AIP）技术。作为该框架的案例研究，我们针对社交媒体用户的隐私保障进行了博弈论分析，假设策略空间包括社交媒体用户的状态。 艺术AIP和人识别技术。