软件学报 ISSN 1000-9825, CODEN RUXUEW
*Journal of Software*,2019,30(2):469−480 [doi: 10.13328/j.cnki.jos.005568]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

# 一种面向人脸活体检测的对抗样本生成算法*

马玉琨[1,2], 毋立芳[1], 简 萌[1], 刘方昊[3], 杨 洲[1]

[1](北京工业大学 信息学部,北京 100124)

[2](河南科技学院 信息工程学院,河南 新乡 453000)

[3](Courant Institute of Mathematics, New York University, New York, NY 10012, USA)

通讯作者: 毋立芳, E-mail: lfwu@bjut.edu.cn

**摘 要**: 近年来,基于深度卷积神经网络的人脸活体检测技术取得了较好的性能.然而,深度神经网络被证明容易受到对抗样本的攻击,影响了人脸系统的安全性.为了建立更好的防范机制,需充分研究活体检测任务对抗样本的生成机理.相对于普通分类问题,活体检测任务具有类间距离小,且扰动操作难度大等特性.在此基础上,提出了基于最小扰动维度和人眼视觉特性的活体检测对抗样本生成算法,将扰动集中在少数几个维度上,并充分考虑人眼的视觉连带集中特性,加入扰动点的间距约束,以便最后生成的对抗样本更不易被人类察觉.该方法只需平均改变输入向量总维度的 1.36%,即可成功地欺骗网络,使网络输出想要的分类结果.通过志愿者的辨认,该方法的人眼感知率比 DeepFool 方法降低了 20%.

**关键词**: 人脸活体检测;对抗样本;卷积神经网络;对抗扰动;视觉集中性

**中图法分类号**: TP391

中文引用格式: 马玉琨,毋立芳,简萌,刘方昊,杨洲.一种面向人脸活体检测的对抗样本生成算法.软件学报,2019,30(2):469−480. http://www.jos.org.cn/1000-9825/5568.htm

英文引用格式: Ma YK, Wu LF, Jian M, Liu FH, Yang Z. Algorithm to generate adversarial examples for face-spoofing detection. Ruan Jian Xue Bao/Journal of Software, 2019,30(2):469−480 (in Chinese). http://www.jos.org.cn/1000-9825/5568.htm

## Algorithm to Generate Adversarial Examples for Face-spoofing Detection

MA Yu-Kun[1,2], WU Li-Fang[1], JIAN Meng[1], LIU Fang-Hao[3], YANG Zhou[1]

[1](Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

[2](School of Information Engineering, He'nan Institute of Science and Technology, Xinxiang 453000, China)

[3](Courant Institute of Mathematics, New York University, New York, NY 10012, USA)

**Abstract**: Face-spoofing detection based on deep convolutional neural networks has achieved good performance in recent years. However, deep neural networks are vulnerable to adversarial examples, which will reduce the safety of the face based application systems. Therefore, it is necessary to analyze the mechanism of generating the adversarial examples, so that the face-spoofing detection algorithms will be more robust. Compared with the general classification problems, face-spoofing detection has the smaller inter-class distance, and the perturbation is difficulty to assign. Motivated by the above, this study proposes an approach to generate the adversarial examples for face-spoofing detection by combining the minimum perturbation dimensions and visual concentration. In the proposed approach, perturbation is concentrated on a few pixels in a single component, and the intervals between pixels are constrained—according to the visual concentration. With such constraints, the generated adversarial examples can be perceived by human with low probability. The

* 基金项目: 北京市教委科技创新项目(KZ201510005012); 国家自然科学基金(61702022); 中国博士后科学基金(2017M610026, 2017M610027)

Foundation item: Science and Technology Innovation Project of Beijing Municipal Education Commission (KZ201510005012); National Natural Science Foundation of China (61702022); China Postdoctoral Science Foundation (2017M610026, 2017M610027)

收稿时间: 2017-09-13; 修改时间: 2017-10-30; 采用时间: 2018-02-09; jos 在线出版时间: 2018-03-13

CNKI 网络优先出版: 2018-03-14 09:18:15, http://kns.cnki.net/kcms/detail/11.2560.TP.20180314.0918.004.html

adversarial examples generated from the proposed approach can defraud the deep neural networks based classifier with only 1.36% changed pixels on average. Furthermore, human vision perception rate of the proposed approach decreases about 20% compared with DeepFool.

**Key words**: face-spoofing detection; adversarial example; convolutional neural network; adversarial perturbation; visual concentration

近年来,随着深度学习技术的发展,卷积神经网络(convolutional neural network,简称 CNN)在视觉领域的应用越来越广泛,如人脸识别、图像分类、分割等[1−3].然而 2014 年,Szegedy 等人提出了深度神经网络(deep neural network,简称 DNN)易受对抗样本攻击的特性,即通过对输入进行不可察觉的细微的扰动,可使深度神经网络以较高的信任度输出任意想要的分类,这样的输入称为对抗样本[4].进而,Goodfellow 等人解释了对抗样本的生成原因以及治理方法[5],在熊猫的图片中加入一个微小的噪声,在人眼不易察觉的情况下,使神经网络以高置信度分类为长臂猿.这对基于深度学习的应用领域安全性构成了一定威胁,如对于人脸识别系统,攻击者通过对人脸图像做精心设计的改动,使系统误认为是攻击者想要的用户身份;或者对于无人驾驶系统,稍微改动“STOP”标志使得深度神经网络识别为别的标志而不及时停车,造成交通事故[6].研究对抗样本的生成过程、分析基于深度学习的系统存在的安全漏洞,有助于建立针对此类攻击的更好的防范机制.

对于人脸识别系统,攻击者往往利用合法用户的照片或视频试图入侵系统,而活体检测任务是检测出请求者为真人还是假体(照片或视频)[7].深度学习技术普及以来,活体检测任务的性能也得到了较大的提升[8−10],然而基于深度学习的活体检测系统同样存在对抗样本攻击问题.图 1 给出了对于人脸活体检测任务的对抗样本实例,其中,图 1(a)所示为活体,图 1(b)所示为照片攻击,图 1(c)所示为针对图 1(b)设计的对抗扰动(为了方便观察,对幅值做了缩放),图 1(d)所示为图 1(b)加上扰动噪声图 1(c)产生的对抗样本.可以看出,通过对假体图片做微小的扰动,使得系统以高置信度分类为活体,从而顺利入侵系统.

(a) 活体 (b) 照片攻击 (c) 扰动 (d) 对抗样本

Fig.1 Generating the adversarial example for face-spoofing detection

图 1 人脸活体检测任务的对抗样本生成示例

对于对抗样本而言,活体检测任务具有其特殊性.活体检测任务最后输出为两类,分别为活体和假体.其中,真、假体反映在图像上的区别微乎其微,例如对于照片攻击,随着成像设备和打印设备的改进,类间距离越来越小,活体检测网络需要具有区分细小差别的能力.另外,假体入侵所使用的照片或视频存在于两次成像之间,受成像设备特性的影响,扰动难度大.以往的对抗样本生成算法将扰动加在输入的每个维度上,而人眼具有视觉连带集中的特性,集中的扰动容易被人眼感知.为了使生成的结果不易被人眼感知,更能迷惑网络分类器,本文提出了基于最小扰动维度的活体检测对抗样本生成方法,所生成的对抗样本具有最小的扰动维度.并且考虑到人眼的视觉连带集中特性,即人眼一旦发现缺陷,视觉立即集中在这片小区域,密集缺陷比较容易被发现[11],通过在生成算法中加入扰动间距约束,生成结果更不易被人类察觉.

本文的主要贡献点在于:

(1) 提出具有最小扰动维度的对抗样本生成算法,只需对输入的极少数维度做扰动,即可生成对抗样本;

(2) 根据人眼的视觉连带集中特性,在对抗样本生成过程中加入了扰动间距约束,生成结果更不易被人眼感知;

(3) 分析了扰动幅度和扰动维度数对人眼感知效果的影响,通过主观评价的方法选取最佳扰动幅度,使算法具有较高的对抗成功率以及生成人眼不易感知的结果.

本文第 1 节介绍相关工作.第 2 节分析简单化的神经网络中对抗样本的特性.第 3 节阐述本文提出的算法.第 4 节为实验设置与结果分析.最后为本文结论.

## 1 相关工作

近年来有多篇文章致力于对抗样本生成算法的研究[4−6],其目的有两个方面:一是能够使分类器在错误分类的前提下降低扰动,二是如何使对抗样本不易被人眼察觉.假设深度神经网络拟合了一个多维函数 $F:X\to Y$,其中,$X$ 为网络输入,$Y$ 为输出向量.对于分类任务,$Y$ 为和 $X$ 对应的分类结果.扰动向量表示为 $\delta_X$,则对抗样本为原始输入加上扰动向量的结果,表示为 $X^*=X+\delta_X$.对抗样本的产生可以表示为以下优化问题:

$$\arg\min_{\delta_X}\|\delta_X\|=\text{s.t. } F(X+\delta_X)=Y^* \tag{1}$$

$Y^*$为对抗样本的目标分类结果.语言描述对抗样本问题,即改动输入使得深度神经网络输出想要的分类的前提下,扰动向量的范数最小.对于二分类问题,可以简化为 $F(X+\delta_X)\neq Y$.2014 年,Goodfellow 等人提出了快速梯度符号法(fast gradient sign method,简称 FGS)以生成对抗样本[5],其基本思想是:设$\theta$是网络参数,$x$ 为输入,$y$ 为与 $x$ 相关的目标输出,$J(\theta,x,y)$为用于训练网络的代价函数,则生成对抗样本所需扰动可用下式求得:

$$\delta_X=\varepsilon\cdot sign(\nabla_x J(\theta,x,y)) \tag{2}$$

其中,$\varepsilon$为一个较小的常数.该方法计算简单,速度快.所生成的扰动和 $x$ 每一维度上的导数符号一致,所有维度的扰动幅度相等.

Goodfellow 等人同时指出,对抗样本在不同的网络结构之间具有泛化性,即用一个网络生成的对抗样本对另一个相同任务的网络具有欺骗性,而无需得知后者的具体结构.分析其原因:不同网络虽然结构不同,但利用相同任务进行训练时,能够学习到相同或相近的输入输出映射关系.对抗样本在网络间的泛化特性意味着,若对分类器网络进行恶意攻击,则攻击者不必已知所攻击的网络结构和参数,可以通过训练自己的网络来产生对抗样本,揭示了深度神经网络分类器防范对抗样本攻击的必要性.

Moosavi-Dezfooli 等人改进了对抗样本的生成方法(称为 DeepFool)[12].该方法将 $x_i$ 维度上的扰动设为

$$r_*(x_i)=-\frac{f(x_i)}{\|\nabla f(x_i)\|_2^2}\nabla f(x_i) \tag{3}$$

并通过迭代的方法生成对抗样本.他们同时提出了使用对抗样本平均扰动幅度定量表示网络对于对抗样本鲁棒性的方法,如式(4)所示.

$$\hat{\rho}_{adv}(f)=\frac{1}{|T|}\sum_{x\in T}\frac{\|\hat{r}(x)\|_2}{\|x\|_2} \tag{4}$$

其中,$\hat{\rho}_{adv}(f)$ 表示平均鲁棒性,$T$ 表示整个测试集,$\hat{r}(x)$ 为生成对抗样本的最小扰动向量.

然而,以上几种方法在扰动幅值最小化的目标函数约束下,将扰动遍布在输入向量的每一维度.如果网络输入为图像,则图像中的每一个像素值会被改动,容易引起人眼视觉集中性的问题,即容易被人眼感知,降低了对抗的可操作性.第 3 节通过对简单神经网络的分析提出改进算法.

## 2 简单化网络分析

深度神经网络通过多个非线性层的叠加,可以学习到比传统方法更具抽象意义的特征[13].神经元通过不同的权值(weight)和偏置(biase)连接,而权值和偏置通过最小化代价函数的训练学习到,网络训练一般基于后向传播的随机梯度下降法进行[14].

深度神经网络没有显式的数学表达式,简化神经网络有助于对网络的分析.此处假设分类器网络的输入为只有两个像素点 $a_1,a_2$ 的灰度图像,$X=(x_1,x_2)$为输入的二维向量,其中,$x_i$ 为 $a_i$ 的灰度值.假设已知一个简单的神经网络实现了如下非线性函数:$F:[-20,20]^2\to[0,1]$,$F(X)=sigmoid(x_1+0.3\times x_2)$,其中,$sigmoid(x)=1/(1+e^{-x})$为神经网络常用的激活函数,$x_i\in[-20,20]$.网络最后根据输出的 $F(X)$值进行分类,对应 $F(X)$的取值更接近于 0 或 1,如图 2 所示,横坐标对应输入的两个维度 $x_1$ 和 $x_2$,纵坐标对应网络的输出值.当输出值大于 0.5 时判断为类别 1,否则判断

为类别 0.图中蓝色表示分类为 0,黄色表示分类为 1.

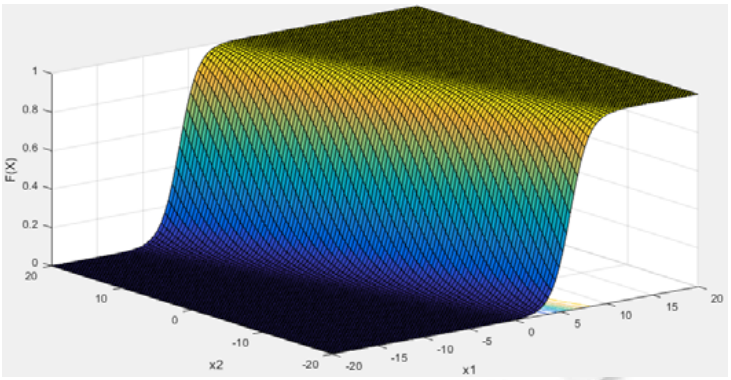



Fig.2 Output of the simplified neural networks

图 2 简单化神经网络的输出

对于一个特定的样本 $X$,网络将其分类为 $Y$.而攻击者考虑对 $X$ 做最小的更改,以改变最后分类结果.以样本 $X=(-2,5)$为例,$F(X)=0.3775$,网络将其分类为 0,对抗样本的目的是改变 $X$,使得 $F(X)>0.5$,使网络分类为 1.而此时 $F(X)$在点$(-2,5)$处的梯度为 $gradF|X_{=(-2,5)}=(0.2350,0.0705)$,梯度在 $x_1$ 或者 $x_2$ 方向上的分量表示函数值随着 $x_1$ 或者 $x_2$ 变化的变化率,因此可知,在输入 $X$ 的两个维度中,$x_1$ 比 $x_2$ 更容易引起输出 $Y$ 的改变,即为了生成对抗样本,改变 $x_1$ 比改变 $x_2$ 更有效.而利用随机梯度下降法训练的一般神经网络分类器,考虑了输出值对于每一输入分量的梯度,因此该分析方法同样适用.

通过对该简化网络的分析可知,由于输入的各个维度连接的权值不同,改变输入向量的不同维度值,对输出值的影响大小不同.对于图像分类的神经网络来说,输出值对于图像中的每一像素的梯度值不同,即每一像素对输出的影响大小不同,将输入向量的各个维度对输出的影响,即梯度值称为显著性映射[15].

图 3 给出了对于某一神经网络分类器 $F$,输入图像每一像素的 $R$ 分量对输出值的显著性映射 $S(\hat{X})$.图中横坐标为输入图像的宽和高 $m\times n$,本例中为 28×28;纵坐标为网络输出值针对某个输入图像的梯度向量,称为显著性映射图.

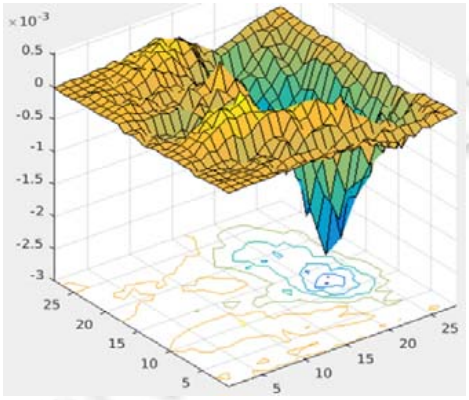


Fig.3 Saliency map of $R$ component in input image for the output of the neural networks

图 3 输入图像的 $R$ 分量对神经网络输出的显著性映射图

由图 3 可知,$S(\hat{X})$ 的某些分量为正,即方向导数 $\frac{\partial F}{\partial x_{i,j,k}}(\hat{x}_{i,j,k})>0$, 说明输出 $F$ 的值会随着图像中该像素 $x_{i,j,k}$

的值增大而增大,而某些且 $S(\hat{X})$ 的分量为负,即方向导数 $\frac{\partial F}{\partial x_{i,j,k}}(\hat{x}_{i,j,k})<0$, 说明增加图像中该像素 $x_{i,j,k}$ 的值反而会降低输出 $F$ 的值.并且 $S(\hat{X})$ 各分量幅值差异较大,即图像中不同坐标点对输出值影响大小不同.梯度分量幅值越大,该坐标点对输出值的影响越大.对于输入 $\hat{X}$, 对显著性接近于 0 的像素做扰动时对输出几乎无影响.为了改变输出分类结果,与在输入的所有维度上做扰动的方法相比,将扰动集中在对输出影响较大的少数几个维度上会更有效.综合以上分析,本文提出了基于最小扰动维度的对抗样本生成方法,将扰动集中在少数几个维度上,并充分考虑人眼的视觉连带集中特性,加入扰动点的间距约束,以便使最后生成的对抗样本更不易被人类察觉.该方法的对抗样本只需改变输入向量总维度的 1.36%,即可成功地欺骗网络;并且通过志愿者的辨认,与 DeepFool 方法相比,该方法更不易被人眼感知.

## 3 融合最小扰动维度和扰动间距约束的对抗样本生成算法描述

算法阐述了本文提出的融合最小扰动维度和扰动间距约束的对抗样本生成的具体步骤.该算法首先计算输入的当前值对输出的梯度矩阵,即显著性映射,选择梯度幅值最大的维度进行扰动,并重复迭代剩余未被扰动的维度,直到能够成功地欺骗网络输出想要的分类结果为止.

该算法的一个关键问题是扰动幅度的大小选取.原则上,在梯度 $S(\hat{X})$ 分量符号不改变的情况下,在该方向上做相应的扰动会对目标分类输出起促进作用.而当梯度 $S(\hat{X})$ 分量的符号改变(由正变负或由负变正)时,若继续做扰动,反而会对目标分类输出起抑制作用.因此,为了选择最优的扰动值,需要分析输入向量的某一维度值在改变时对输出的影响.由于研究对象输入为图像,因此将输入范围设为[0,255],即 $x_{i,j,k}\in[0,255]$.具体做法是在输入的图中随机选择一个维度 $x_{i,j,k}$(一个维度对应某一像素点的 RGB 三分量之一),保持其他维度值不变,在[0,255]范围内改变其中一个维度值时输出值的变化情况,得到的输入输出曲线如图 4 所示.横坐标为区间[0,255],为图像像素点的取值范围;纵坐标为输出值 $Y$,即目标类别的置信度.本文中对抗样本的目的是使分类模型判断假样本为真人、目标类别为真,因此纵坐标为真类别的置信度.每一条曲线代表一个维度值 $x_{i,j,k}$ 和输出值 $Y$ 的关系,其中,曲线 1~曲线 4 表示梯度为负的几个维度,为单调递增;曲线 5~曲线 7 表示梯度为正的几个维度,为单调递减.

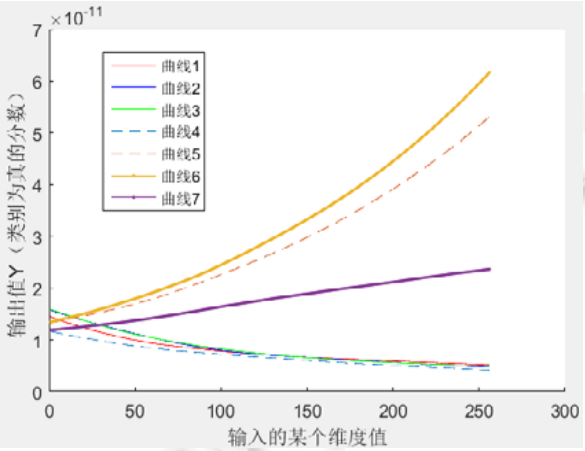



Fig.4 Input-output curves when one dimension of input changes

图 4 输入某一维度值改变时的输入-输出曲线

从图中可以看出,虽然输入-输出曲线为非线性关系,但均为单调递增或单调递减,且每条曲线在[0,255]范围内梯度符号不变.因此,设扰动幅度最大值为$\mu$,并考虑图像取值范围为[0,255],则设置具体扰动幅值 $r$ 遵循以下公式:

$$r_{i,j,k}=\begin{cases}\min(\mu,255-X(i,j,k)), & S(i,j,k)>0\\ \max(-\mu,-X(i,j,k)), & S(i,j,k)<0\end{cases} \tag{6}$$

研究对抗样本生成算法的另一个目的是不易被人眼感知.而人眼具有视觉连带集中性,即针对此特性,为了避免扰动发生在集中的大片区域内,为本文方法加上扰动间距约束,且一个像素内 RGB 三分量只允许最多 1 个分量扰动,以避免颜色跳跃过于明显.具体做法为:将已经被扰动的像素坐标加入到点集 $P$ 中,即 $P=\{(x_1,y_1),(x_2,y_2),\ldots,(x_k,y_k)\}$表示有 $k$ 个像素点已经被扰动.设扰动间距阈值为 $d_{\text{threshold}}$,则当扰动下一个像素点时,首先计算该像素与 $P$ 中每个点的曼哈顿街区距离,当 $k$ 个距离均不小于 $d_{\text{threshold}}$ 时才可对其进行扰动,否则不对其进行操作,进而尝试别的像素点,具体过程如算法 1 所描述.后面的实验结果表明,加入扰动间距约束后,对抗样本更不易被人眼感知.

**算法 1**. 融合最小扰动维度和扰动间距约束的对抗样本生成算法.

1. 输入:$X,Y^*,F,r$.
2. 输出:$X^*,\delta_X$.
3. $X^*=X$
4. $P=\{\},Q$ 为 $m\times n\times 3$ 的全零矩阵
5. **while** $F(X^*)\neq Y^*$
6. 计算显著性映射 $S=\nabla F(X^*)$
7. 在$\delta_X(i,j,k)==0$ 的条件下,找出$(i,j,k)_{\max}=\text{argmax}S(i,j,k)$
8. 计算$(i,j)_{\max}$ 与 $P$ 中每一点的距离 $D$
9. **if** $\forall d(d\in D)\geqslant d_{\text{threshold}}$
10. 在 $X^*_{(i,j,k)_{\max}}$ 上加 $r$.
11. 将点$(i,j)$加入集合 $P$

    **end**
12. $Q(i,j,k)=1$
13. **end while**
14. $\delta_X=X^*-X$
15. 返回 $X^*,\delta_X$

算法解释:

1. $X$ 为原始输入样本,是 $m\times n\times 3$ 的图像.$Y^*$为目标输出类别(真或假).$F$ 为分类器网络的映射函数.$r$ 为扰动幅度.
2. $\delta_X$ 为总的扰动向量,是初始值为 $m\times n\times 3$ 的全 0 矩阵.$X^*$为被扰动后的输入,$X^*=X+\delta_X$.
3. 将 $X$ 赋值给 $X^*$.
4. $P$ 为被扰动的像素点的集合,初始值为空集.$Q$ 为 $m\times n\times 3$ 的矩阵,用来记录某个像素点是否被遍历过,是初始值为 $m\times n\times 3$ 的全零矩阵.
5. 当 $X^*$对应的输出 $F(X^*)$不等于目标输出 $Y^*$时,循环执行第 6 步~第 12 步.
6. 计算显著性映射 $S=\nabla F(X^*)$.
7. 在$\delta_X==0$ 的条件下,寻找 $S$ 中的最大幅值对应的像素点$(i,j)_{\max}$.约束$\delta_X==0$ 是为了避免重复扰动同一像素点.
8. 当 $P$ 包含 $l$ 个点时,$D$ 为包含 $l$ 个元素的集合.
9. 当$(i,j)_{\max}$ 与已扰动的所有点的距离不低于 $d_{\text{threshold}}$ 时,执行第 10 步、第 11 步.
10. 给图像上对应的像素点 $X^*_{(i,j)_{\max}}$ 加上扰动 $r$,$r$ 值由公式(6)决定.
11. 将点$(i,j)$加入集合 $P$,表示点$(i,j)$处已被扰动.
12. $Q(i,j)=1$,表示$(i,j)$像素点被遍历过,下次循环不再考虑,避免算法陷入死循环.此处只记录平面坐标,不

记录 RGB 分量,同一像素点的 RGB 三分量不允许被同时扰动.

13. 当第 5 步条件满足时,结束 while 循环.
14. 循环 $l$ 次以后,若 $\delta_X$ 中有 $l$ 维不为 0,则表示有 $l$ 维度被扰动过.
15. 返回 $X^*,\delta_X$.

## 4 实 验

### 4.1 实验设置

本文所用分类器为卷积神经网络,网络结构见表 1.网络包含 4 个卷积层和 1 个 softmax 层,前两个卷积层后接 pooling 层,前 3 个卷积层后加 Batch Normalization 层.网络最后输出两类:真/假.

**Table 1** Architecture of the CNN based face-spoofing detection classifier

**表 1** 基于卷积神经网络的活体检测分类器结构

| 类型 | 滤波器尺寸/步长,填充 | 输出尺寸 | 参数 |
|---|---|---|---|
| 输入 | – | 28×28×3 | – |
| Conv1 | 3×3×3/1,0 | 26×26×20 | 540 |
| Batch normalization | – | 26×26×20 | – |
| Pool1 | 2×2/2 | 13×13×20 | – |
| Conv2 | 4×4×20/1,0 | 10×10×30 | 9.6k |
| Batch normalization | – | 10×10×30 | – |
| Pool2 | 2×2/2 | 5×5×30 | – |
| Conv3 | 4×4×40/1,0 | 2×2×40 | 25.6k |
| Batch normalization | – | 2×2×40 | – |
| Conv4 | 2×2×40/1,0 | 1×1×50 | 8k |
| Fully connected | – | 2 | 100 |
| Softmax | – | 2 | – |

训练网络所用活体检测数据库为 2011 年公开的 Print Attack[16].该数据库采集 50 个志愿者的两种视频片段,分别为真人、打印照片攻击,每一类有 200 个视频,每个视频长度为 10s 左右.数据库给出了视频中每一帧的人脸位置信息.作为数据预处理,读取视频中的每一帧,并裁剪出人脸区域,保存为图片,用于训练和测试网络,最终得到 72 806 个训练图片、48 451 个测试图片.训练结束后的网络活体检测半错误率(half total error rate,简称 HTER)为 1.72%.

### 4.2 扰动幅度$\mu$的影响

设置不同的扰动幅值$\mu$会对生成结果有较大影响,为了寻找最佳的$\mu$值,分别设置$\mu$=30,60,90,120,150,180,210,240,255,利用本文算法生成对抗样本,生成的对抗样本结果以及所对应扰动如图 5 所示.实验中设$d_{\text{threshold}}$=2,即任意两个扰动点的曼哈顿街区距离不小于 2,避免了两个扰动点相连.由于输入图像为 28×28 的大小,如果扰动间距阈值设定过高,可扰动的像素数急剧下降,会导致对抗成功率下降.图中第 1 列为假体攻击的原图,后面 9 列为$\mu$取不同值时的对抗样本.由图 5 可以看出,当$\mu$值较小时,扰动点数较多,视觉效果也更模糊;当$\mu$值逐渐增大时,扰动点数逐渐减少,随之而来的是扰动点的颜色跳跃越来越明显.扰动图的黑色部分表示未被更改,带颜色部分表示该位置对应的某分量被更改,红、绿、蓝颜色分别对应 RGB 分量.

图 6 展示了不同的扰动幅度所生成的对抗样本结果分析,其中,

- 图 6(a)曲线为生成成功率,由于加入了扰动间距约束,可扰动的像素数有限,当扰动幅度设置较小时,容易出现遍历完所有像素仍无法生成对抗样本的情况,即对抗失败.由实验结果可知,当$\mu$=30 时,生成对抗样本的成功率较低;当$\mu$=90 时,生成成功率可达 97%以上.
- 图 6(b)曲线为扰动像素数,扰动幅度越大,所需扰动像素数越少.
- 图 6(c)曲线为人眼感知率.

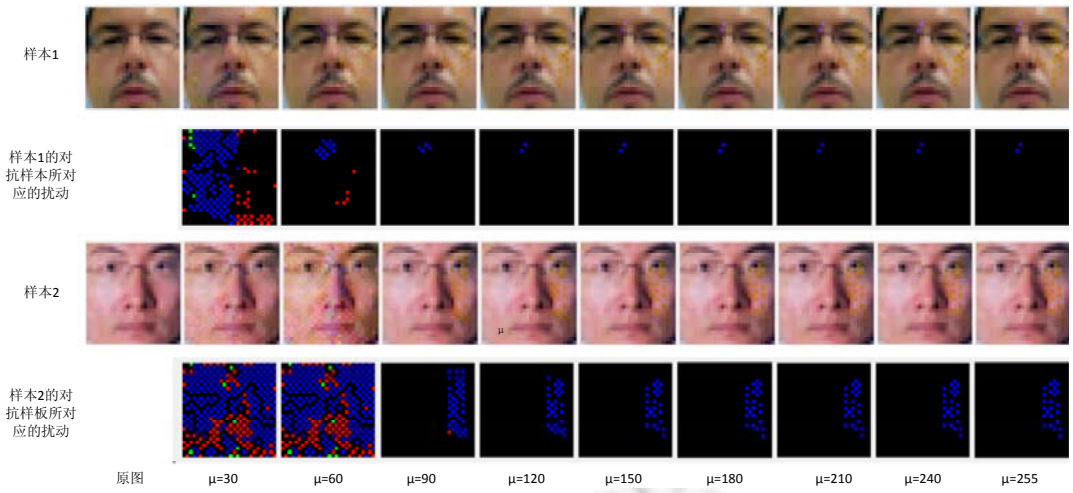
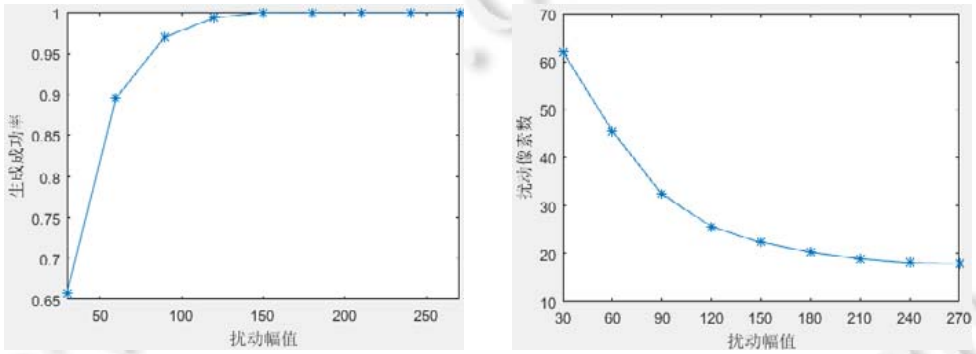
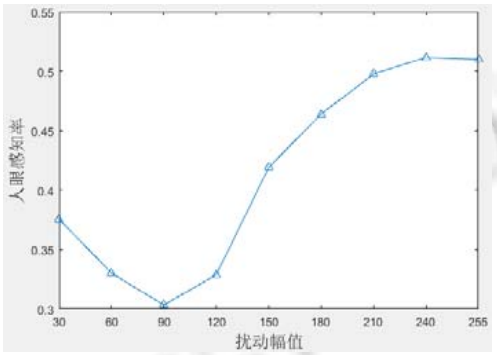



Fig.5 Generated adversarial examples with different perturbation range

图 5 不同的扰动幅度生成的对抗样本




(a) 对抗成功率 (b) 扰动像素数




(c) 人眼感知率

Fig.6 Results analysis with different perturbation range

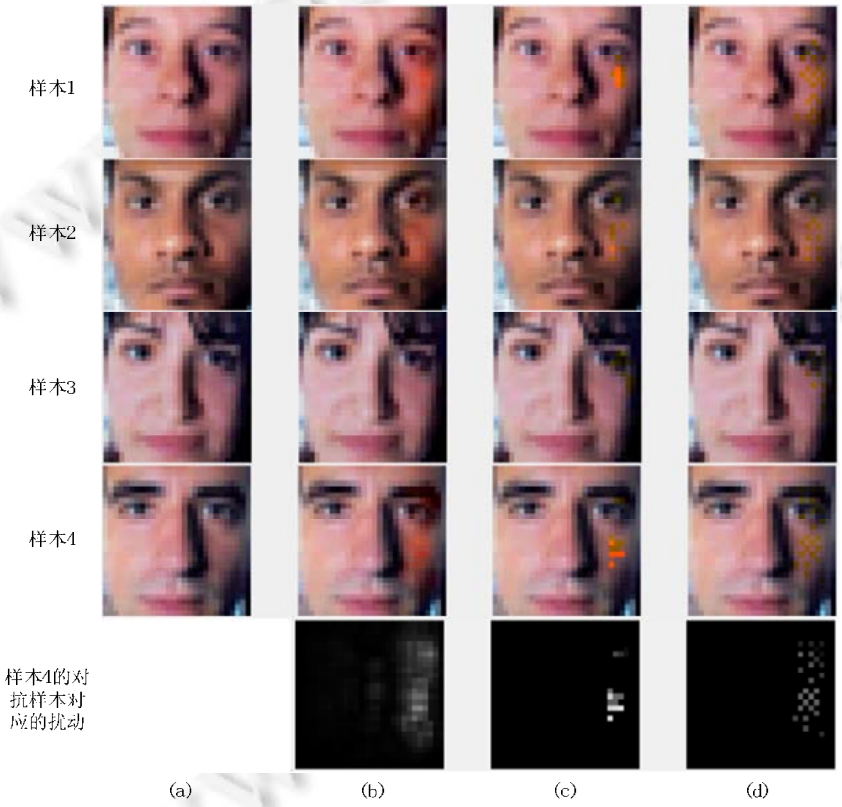
图 6 不同扰动幅度的结果分析

仔细观察图 5 可以发现,扰动的幅值和点数共同影响了人眼感知效果.采用主观评价的方法定量表达此效果.对于每个志愿者,每次给出 1 组图片,包括原图以及 9 个扰动幅值、FGS 方法以及 DeepFool 方法生成的对抗样本,并回答问题:“该组图片哪些被人工修改过?”.共有 20 个志愿者参与评价,每个志愿者评价 300 组图片,一组

图片为一个原图以及生成的 11 个对抗样本图,共 3 600 张图片.并且评价是在不同的显示设备上完成的,放大尺寸由志愿者自行调整,每张图片的观察时间控制在 3s 左右.图 6(c)表示人眼感知率和扰动幅值之间的关系,由图可知,当$\mu$=90 时,人眼感知率最低,为 30%;$\mu$值增大或者缩小,都会导致人眼感知率增加.分析其原因,人眼可感知性受扰动维度数和扰动幅值两个因素共同影响,而$\mu$=90 的设置使得两者进行了折中,较不易被感知.为了验证该评价方法的合理性和有效性,在测试图片中加入了原图,其人眼感知率低于 5%,即原图很少被误认为人工修改过,证明该主观评价方法是合理而有效的.综合对抗成功率、扰动像素数以及人眼感知率等数据,选择$\mu$=90 为最佳扰动幅度,并作为后续实验及结果比较的参数.

### 4.3 和相关工作的结果比较

为了进一步与相关工作比较,分别利用 DeepFool 方法和本文的算法生成测试集图像的对抗样本,如图 7 所示.其中,图 7(a)所示为原始的假体图片,且网络能够成功地检测为假,实验对其做扰动,生成对抗样本以欺骗分类器网络判断为活体;图 7(b)所示为 DeepFool 方法生成的对抗样本;图 7(c)所示为本文算法未加扰动间距约束时生成的对抗样本;图 7(d)所示为加入扰动间距约束的算法生成的对抗样本.从图 7 中可以看出,加入扰动间距约束的算法生成结果在视觉上更接近原始图像,更不易被人眼感知.图 7 第 5 排为第 4 排中各对抗样本对应的扰动,本文算法生成的扰动避免了同一像素中 RGB 分量被同时扰动,且被扰动的像素较分散,互相不连续,有较好的视觉效果.利用主观评价方法得到 DeepFool 和 FGS 算法生成结果的人眼感知率分别为 50%和 51%,本文方法的人眼感知率比 Deepfool 和 FGS 方法降低了 20%和 21%(注:图 7 所列例子均为测试集中随机选取的结果,并非精心挑选).




(a) 原图 (b) DeepFool 方法 (c) 本文算法(未加扰动间距约束) (d) 本文算法(加入扰动间距约束)

Fig.7 Comparisons of adversarial examples from different algorithms

图 7 不同算法生成的对抗样本比较

表 2 分析了不同算法所需的扰动维度和平均扰动幅度等.平均扰动幅度的计算方法如公式(4)所示,即扰动向量的 L2 范数与原始数据 L2 范数的比值.由表 2 可知,本文算法未加扰动间距约束时,平均扰动 30.3 个像素点,占原始输入维度的 1.29%,即只需改动原始输入向量的 1.29%,即可成功地欺骗网络;加入间距约束后,算法平均扰动像素点为 31.9,占原始输入维度的 1.36%,略微有增加,但视觉效果明显提升.

**Table 2** Comparison with other related methods

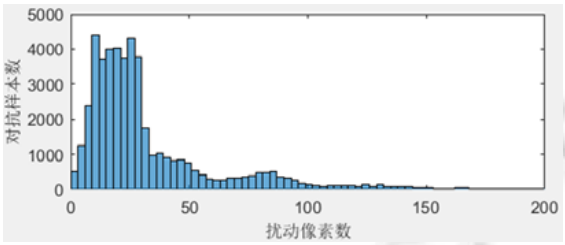
**表 2** 与其他相关方法的比较

| 方法 | DeepFool | 本文算法未加扰动间距约束 | 本文算法加入扰动间距约束 |
|---|---|---|---|
| 扰动维度数 | 2 351 | 30.3 | 31.9 |
| 扰动维度比例(%) | 99.99 | 1.29 | 1.36 |
| 平均扰动幅度 | 0.0382 | 0.140 7 | 0.142 5 |
| 网络间的泛化性(%) | 87.63 | 78.44 | 85.75 |

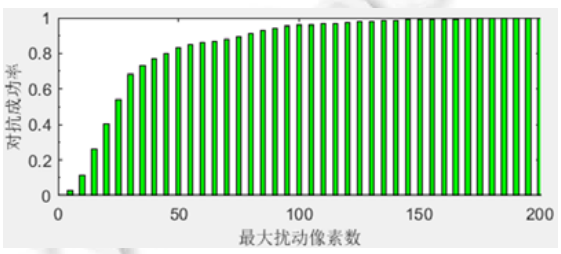
为了验证本文算法生成结果的泛化性,用相同的数据库训练另一个网络 LeNet-5[17],即卷积神经网络中具有代表性的一个结构.LeNet-5 网络的活体检测正确率为 98.52%.利用表 1 所示网络生成的对抗样本集针对 LeNet-5 网络的欺骗成功率为 85.75%,即表 1 所示网络生成的对抗样本中有 85.75%能够欺骗 LeNet-5,证明本文算法生成的对抗样本在不同网络之间具有较好的泛化性,与 DeepFool 方法 87.63%的泛化性相当.

文章《DeepFool: A simple and accurate method to fool deep neural networks》指出:网络层数越高,分类性能就越好,对于对抗样本的鲁棒性也就越好.加入一些技巧,如 Batch normalization 和 dropout 之后,可以在一定程度上提高模型鲁棒性,但对抗样本问题仍然存在.本文在表 1 所示网络中全连接层上加入 dropout,活体检测的等错误率 HTER 为 2.12%,与未加 dropout 时相当;利用本文方法重新生成对抗样本集,其平均扰动幅度为 0.122,即与未加 dropout 时相比,鲁棒性没有明显的提高.Dropout 的主要作用在训练阶段加速收敛和防止过拟合,但对于对抗样本不鲁棒.

图 8(a)示出了对 48 451 个测试图片利用算法 2 生成对抗样本所需扰动的维度数直方图,扰动维度集中在[0,50]左右.有少量样本的扰动维度较大.通过限制最大扰动维度数,分析对抗样本生成的成功率如图 8(b)所示,即若扰动维度数大于某一阈值,则停止迭代,测试该时刻生成样本是否能够成功地欺骗.由分析结果可知,当最大维度设置为 100(输入维度的 4.25%)时,生成对抗样本的成功率为 97%.




(a) 测试集对抗样本扰动维度直方图




(b) 扰动维数限制和对抗成功率关系

Fig.8 Analysis of average perturbation dimensions in adversarial examples

图 8 对抗样本平均扰动维度分析

## 5 结 论

深度学习技术易受对抗样本的攻击,而人脸活体检测任务的对抗样本具有其特殊性:活体检测任务真假体图像相近,类间距离较小,且假体存在于两次成像之间,对其做扰动有一定的局限性.针对以上特点以及人眼对图片的视觉特性,提出了一种基于人眼视觉特性的最小扰动维度对抗样本生成方法.该方法将对输入图像的扰动集中在少数几个维度上,并充分考虑人眼的视觉连带集中特性,加入扰动点的间距约束,以使最后生成的对抗样本更不易被人类察觉.利用人脸活体检测数据库Print Attack对典型CNN分类模型进行对抗,该方法通过平均扰动输入总维度的 1.36%,即可成功地生成对抗样本.并且通过加入扰动间距约束,使对抗样本更加不易被人眼感知,通过志愿者对于对抗样本的主观评价,其人眼感知率比经典 FGS 方法及 DeepFool 方法降低了 20%,证明该方法有更好的欺骗效果.

本文研究了人脸活体检测任务中对抗样本的生成机理,揭示了活体检测任务分类器的安全隐患,为下一步建立合理的防范机制奠定了基础.

**References**:


[1] Schroff F, Kalenichenko D, Facenet PJ. A unified embedding for face recognition and clustering. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. IEEE Computer Society, 2015. 815−823.

[2] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. Curran Associates Inc., 2012. 1097−1105.

[3] Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In: Proc. of the Int'l Conf. on Representation Learning, 2015. 1−13.

[4] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. Intriguing properties of neural networks. In: Proc. of the Int'l Conf. on Representation Learning. 2014. 1−10.

[5] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proc. of the Int'l Conf. on Representation Learning, 2015. 1−10.

[6] Meier U, Masci J. Multi-column deep neural network for traffic sign classification. Neural Networks the Official Journal of the Int'l Neural Network Society, 2012,32(1):333−338.

[7] Xu X. Research on deep learning based face liveness detection algorithm [MS. Thesis]. Beijing: Beijing University of Technology, 2016 (in Chinese with English abstract).

[8] Lucena O, Junior A, Moia V, Souza R, Valle E, Lotufo R. Transfer learning using convolutional neural networks for face anti-spoofing. In: Proc. of the Int'l Conf. on Image Analysis and Recognition. Springer-Verlag, 2017. 27−34.

[9] Xu Y, Jian M, Xu X, Qi W. Face liveness detection scheme with static and dynamic features. Int'l Journal of Wavelets Multiresolution & Information Processing, 2018,16(2):No.1840001.

[10] Yang J, Lei Z, Li SZ. Learn convolutional neural network for face anti-spoofing. Computer Science, 2014,9218:373−384.

[11] Gonzalez RC, Woods RE. Digital Image Processing. Beijing: Publishing House of Electronics Industry, 2006.

[12] Moosavi-Dezfooli SM, Fawzi A, Frossard P. Deepfool: A simple and accurate method to fool deep neural networks. In: Proc. of the Computer Vision and Pattern Recognition. IEEE, 2016. 2574−2582.

[13] Yin BC, Wang WT, Wang LC. Review of deep learning. Journal of Beijing University of Technology, 2015,41(1):48−59 (in Chinese with English abstract).

[14] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature, 1986,323(6088):533−536.

[15] Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A. The limitations of deep learning in adversarial settings. In: Proc. of the IEEE European Symp. on Security and Privacy. IEEE, 2016. 372−387.

[16] Anjos A, Marcel S. Counter-measures to photo attacks in face recognition: A public database and a baseline. In: Proc. of the Int'l Joint Conf. on Biometrics. IEEE, 2011. 1−7.

[17] Lécun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc. of the IEEE, 1998,86(11): 2278−2324.

**附中文参考文献:**

[7] 许晓.基于深度学习的活体人脸检测算法研究[硕士学位论文].北京:北京工业大学,2016.

[13] 尹宝才,王文通,王立春.深度学习研究综述.北京工业大学学报,2015,41(1):48–59.

**马玉琨**(1983－),女,河南新乡人,博士,CCF 专业会员,主要研究领域为数字图像处理.

**毋立芳**(1970－),女,博士,教授,博士生导师,CCF 专业会员,主要研究领域为数字图像处理,模式 识别.

**简萌**(1987－),女,博士,讲师,CCF 专业会员,主要研究领域为模式识别,多媒体计算.

**刘方昊**(1993－),男,硕士,主要研究领域为优化理论.

**杨洲**(1994－),男,硕士生,主要研究领域为计算机视觉.