# 对MTCNN的对抗攻击

王思远

June 24, 2021

## 1 引言

人脸检测指从一张任意大小的图片中检测出任意大小的人脸，并标定出其大致范围和人脸地标位置的任务。它是卷积神经网络(CNN)应用中的一个重要问题。多任务卷积神经网络Multi-task convolutional neural network (MTCNN) [1] 是2016年由中国科学院深圳研究院提出的一种神经网络模型，可以快速高效地进行人脸检测，并且由于其级联特性具有较好的鲁棒性。本文使用传统对抗攻击算法对MTCNN进行了白盒攻击，目标为阻止MTCNN成功检测出目标人脸，取得了较好的攻击成果。

## 2 相关研究

### 2.1 MTCNN

MTCNN是一个级联的神经网络模型。它包含了三个网络P-Net、R-Net和O-Net，其中上一个网络的输出是下一个网络的输入。其中，P-Net主要负责快速生成不同大小的候选窗口， R-Net主要负责对候选窗口进行过滤选择，O-Net主要负责生成最终边界框和人脸地标。这三个子网络的网络结构如下图所示：

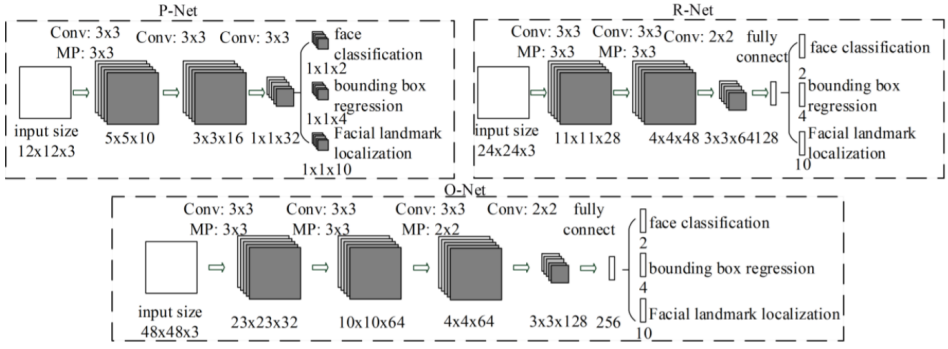


Figure 1: MTCNN的网络结构

在检测时，先基于一个确定的尺寸缩小因子，对图像进行不同尺寸的变换生成图片金字塔，再送入P-Net这个全卷积网络。对P-Net 输出的候选框进行非极大值抑制，再送入R-Net，滤除效果较差的候选框。同样对R-Net 输出的候选框再进行一次非极大值抑制，送入O-Net来标定最终的人脸边界框和面部地标。

### 2.2 对抗攻击

对抗攻击是一种针对CNN的攻击方法。它通过对输入添加微小的扰动使得卷积神经网络无法正常工作（分类器分类错误、回归问题损失函数值过大）。传统的对抗攻击算法都是基于FGSM (Fast Gradient Sign Method)算法的。 FGSM是一种基础的白盒图像攻击算法：

$$x^{adv} = x + \epsilon \cdot \mathrm{sgn}(\nabla_x \mathcal{L}(\theta, x, y))$$

本文中使用的是PGD(Project Gradient Descent)对抗攻击算法。 PGD是一种较强的白盒图像攻击算法，是一种对FGSM算法的改进。考虑到梯度的方向随着梯度上升会进行改变，PGD算法把FGSM的一步梯度上升分成了许多小步，并且在最开始进行了随机初始化：

$$||x_0^{adv} - x||_\infty < \epsilon$$
$$x_t^{adv} = \Pi_\epsilon(x_{t-1}^{adv} + \alpha \cdot \mathrm{sgn}(\nabla_x \mathcal{L}(\theta, x_{t-1}^{adv}, \hat{y})))$$

其中$\mathcal{L}$为损失函数，$\theta$为网络参数，$x$表示输入，$\hat{y}$表示进行扰动之前正确的网络输出，$t$为PGD的步数，$\alpha$为每步的步长。

2019年，华为莫斯科实验室首次成功使用对抗攻击干扰了MTCNN系统。[2]

## 3 攻击方法

### 3.1 选取攻击对象

MTCNN是一个级联的模型，因此对整个模型的对抗攻击难以实现。P-Net、R-Net、O-Net各有三个输出层：人脸分类层、人脸边界框回归层、人脸地标回归层，一共9个可供选择的攻击对象。本文选择攻击P-Net的人脸分类层。

在MTCNN检测图像时，它会将其按数个尺寸缩放以生成图像金字塔，再将这些不同尺寸的图像送入P-Net进行检测。也就是说，该网络有多个输入。如果对每个输入都进行对抗攻击，一方面耗时过长，另一方面造成的扰动也过大，过于容易被肉眼发现。

因此在选择攻击的图像时，本文使用了如下的策略：先检测一次获得人脸边界框，然后找到缩放后该人脸框最接近12*12（PNet的检测大小）的缩放比例。在获得这个比例之后，选择攻击对象为图像金字塔中的该比例以及比该比例大或小一级、两级的比例，一共五个图像。程序自动地由这五个图像生成图像金字塔，再对其进行攻击。

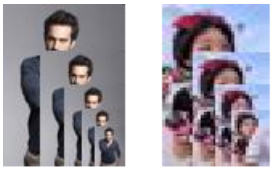

Figure 2: 部分待攻击的图像金字塔

### 3.2 损失函数

生成扰动的过程中使用了两个损失函数的和：

**分类损失函数** 为了使得网络无法正确识别人脸，本文使用了有目标的攻击，目标为让P-Net人脸分类层输出为0。使用了二分类交叉熵损失函数，该损失函数用于生成扰动：

$$\begin{aligned}\mathcal{L}_{BCE}(y, \mathbf{0}) &= -\frac{1}{n}\sum_{i=1}^{n} 0 \cdot log(y_i) + (1-0) \cdot log(1-y_i) \\ &= -\frac{1}{n}\sum_{i=1}^{n} log(1-y_i)\end{aligned}$$

**全变分损失函数(Total variation loss)** 该损失函数用于减少生成的扰动中的尖锐的颜色变化、鬼影、过于明显的噪声等：

$$\mathcal{L}_{TV}(y) = \frac{1}{xyz}\sum_{i,j,k}\sqrt{(y_{k,i,j} - y^2_{k,i,j+1}) + (y_{k,i,j} - y^2_{k,i+1,j})}$$

最终使用的损失函数为$\mathcal{L}(y) = \mathcal{L}_{BCE}(y, \mathbf{0}) + \mathcal{L}_{TV}(y)$。对x进行扰动时，目标为最小化$\mathcal{L}(y)$。

### 3.3 对扰动插值

在攻击之后，需要把得到的小尺寸扰动重新放大成原图尺寸，再加到原图上。使用最近邻插值法和三线性插值法得到的对抗样本分别如下，可以看到三线性插值法中没有特别明显的色块，较为适合本任务。

Figure 3: 最近邻插值法（左）和三线性插值法（右）生成的对抗样本

# 4 实验与分析

## 4.1 训练MTCNN

### 4.1.1 生成数据

本文使用了CelebA来生成MTCNN的训练数据。CelebFaces Attribute (CelebA)是一个名人人脸属性数据集，包含10177个名人的202599张人脸图片，每张图片都做好了特征标记，包含人脸边界标注框、5个人脸地标坐标以及40个属性标记。

生成数据时，在CelebA标定好的面部框周围随机偏移人脸框，计算该随机框和原框的IOU（交并比）。IOU>0.66的，归入正样本，视为有人脸； 0.35<IOU<0.58的，归入部分样本，视为有部分人脸；IOU<0.1的，归入负样本，视为没有人脸。以上三种样本比例约为1:1:3。随机框中包含全部五个地标点（双眼，鼻子，双嘴角）的，归入地标样本。生成的数据集大小如下表所示：

| Train | Positive | Part | Negative | Landmark |
|---|---|---|---|---|
| 12*12(PNet) | 61362 | 53911 | 144732 | 154558 |
| 24*24(RNet) | 60585 | 54003 | 145179 | 153706 |
| 48*48(ONet) | 60669 | 53555 | 145136 | 153434 |
| Test | Positive | Part | Negative | Landmark |
| 12*12(PNet) | 10833 | 9401 | 25017 | 27401 |
| 24*24(RNet) | 10563 | 9302 | 25378 | 26712 |
| 48*48(ONet) | 10604 | 9672 | 24948 | 27217 |

Table 1: 生成的数据集大小

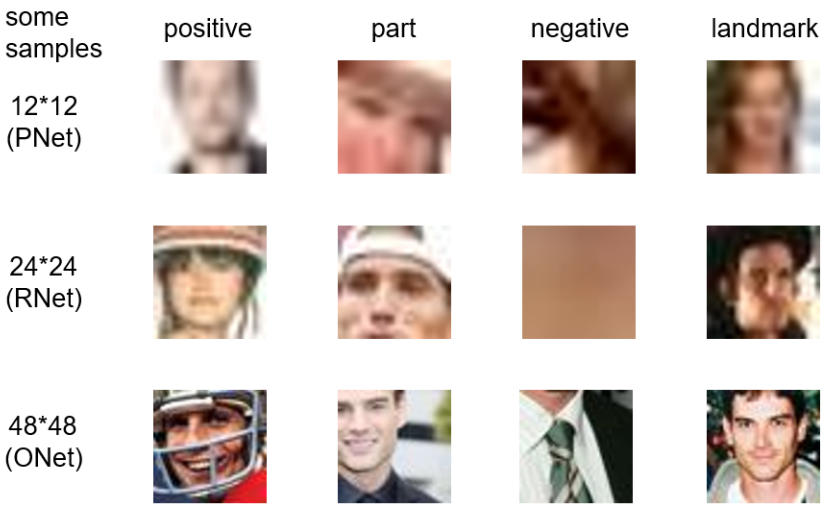


Figure 4: 数据集样本示例

#### 4.1.2 训练三个子网络

本实验中训练MTCNN时选择的batch size为64，learning rate为0.001，优化器为Adam。分类层使用的损失函数为二分类交叉熵损失函数(BCE)：

$$\mathcal{L}_{BCE}(y, \hat{y}) = -\frac{1}{n}\sum_{i=1}^{n} \hat{y}_i \cdot log(y_i) + (1 - \hat{y}_i) \cdot log(1 - y_i)$$

回归层使用的损失函数为均方损失函数(MSE)：

$$\mathcal{L}_{MSE}(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

三个子网络训练中损失函数值变化见下图。最终选择为P-Net训练17个epoch，R-Net 训练 45个epoch， O-Net训练44个epoch。该模型的人脸检测成功率达到了94.7%。

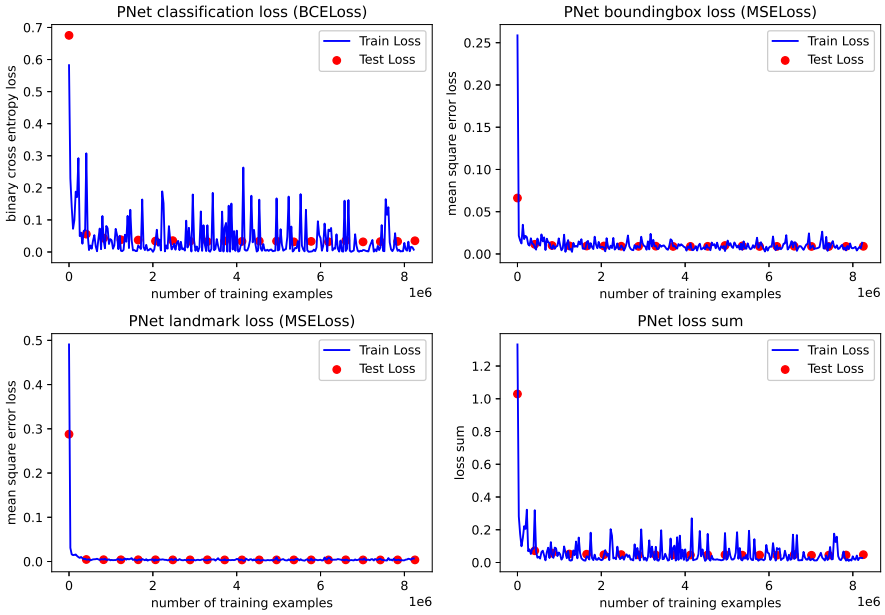


Figure 5: P-Net训练过程

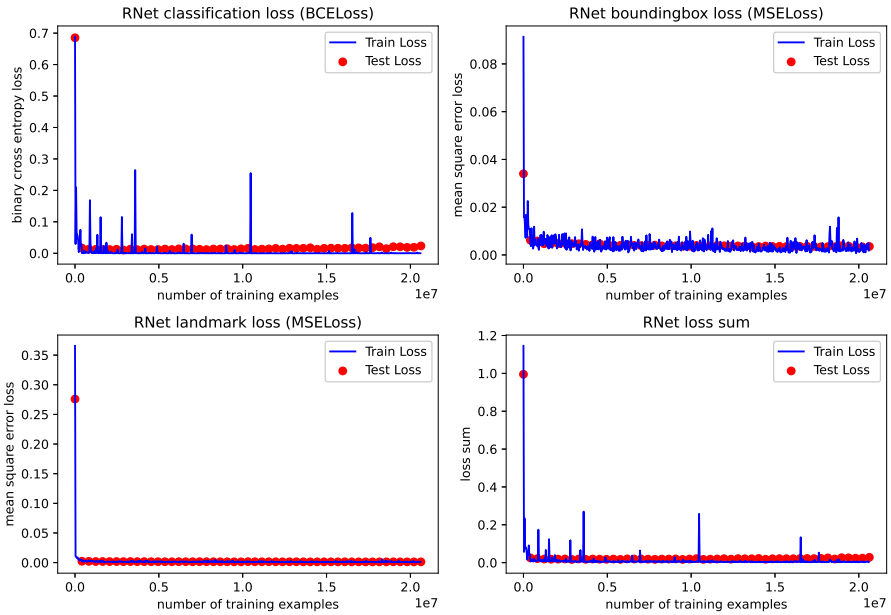


Figure 6: R-Net训练过程

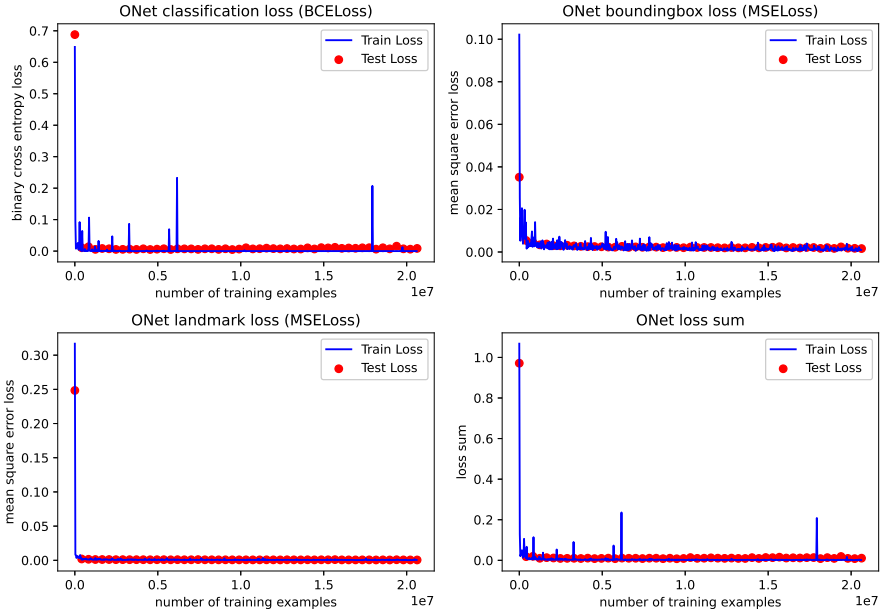


Figure 7: O-Net训练过程

## 4.2 PGD攻击

从训练MTCNN时用于生成测试数据的图片中抽取了1000张，组成了用于测试PGD算法成功率的数据集。本实验中测试了$\epsilon = 0.03$，步数为50，步长从$0.001 \times 2^{0.0}$ 到$0.001 \times 2^{6.0}$时PGD算法的成功率。

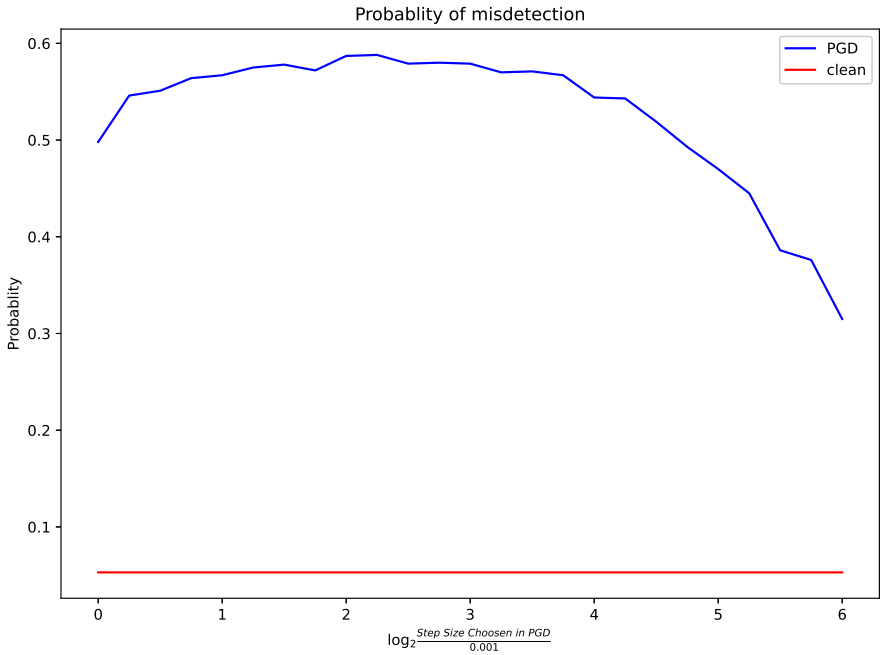


Figure 8: PGD成功率与步长关系

随着步长增大，成功率因扰动幅度快速增大而快速增大。一定程度后，因为步长过大， PGD算法在很少的几步内就走到了$\epsilon$球面上，PGD算法逐渐接近于使用了随机初始化的FGSM算法，因而精度降低，成功率下降。步长为 $0.001 \times 2^{2.25} \approx 0.0047568$时，攻击成功率达到最大值58.8%。

## 4.3 MTCNN的鲁棒性

MTCNN由于其级联特性拥有较好的鲁棒性。以下是对MTCNN鲁棒性好的具体原因的一些猜想。

### 4.3.1 P-Net输出大量候选框

在检测时，P-Net输出得到大量候选框之后才做非极大值抑制，消除特别近似的候选框。这些被消除的候选框中很多都可以变换成结果。在对P-Net的人脸分类层进行攻击时，对抗样本需要使

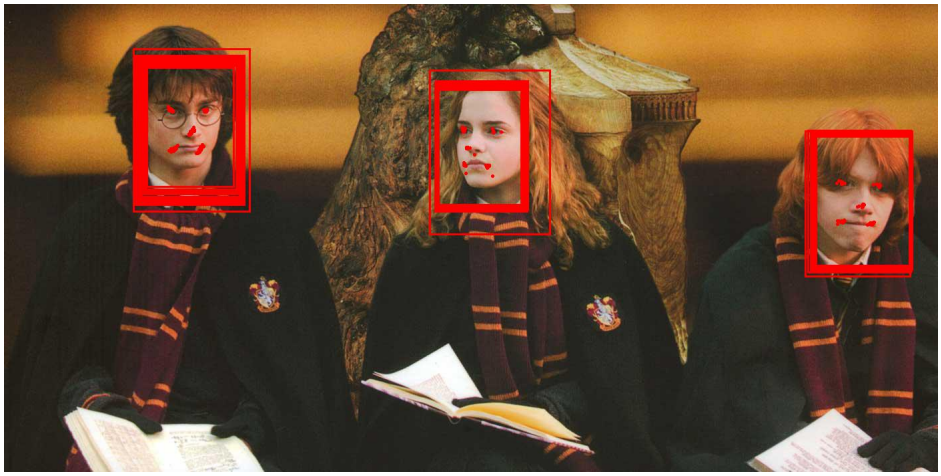

Figure 9: 去除非极大值抑制后的MTCNN输出

得这些候选框经过P-Net后都小于阈值，不能遗漏任何一个，否则MTCNN 依然能成功检测人脸。上图为去掉非极大值抑制后得到的人脸检测结果，该图佐证了这个观点。

#### 4.3.2 难以定位攻击对象

在攻击对象选取错误时，MTCNN自然难以攻击。如果手动选取攻击对象，可以只攻击三张不同大小的图片，而本文需要攻击五张图片，才能保证其中包含真正需要的攻击对象。

#### 4.3.3 图像金字塔

一方面，图像的缩放容易破坏扰动特征；另一方面，缩小图片、制造扰动再插值放大，导致扰动中一定会产生或深或浅的色块，这也导致其对抗样本容易被肉眼识别。

## 5 展望

在改进攻击MTCNN的算法时，有两个值得尝试的方向：一是改进选取攻击对象的算法，使得攻击更为精准；二是改进损失函数，使得对抗样本更难被肉眼识别、攻击成功率更高。

## References


[1] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016.

[2] Edgar Kaziakhmedov, Klim Kireev, Grigorii Melnikov, Mikhail Pautov, and Aleksandr Petiushko. Real-world attack on mtcnn face detection system. *arXiv preprint arXiv:1910.06261*, 2019.