



清华大学
Tsinghua University

基于对抗样本的人脸攻击系统 ——IID交叉创新实践项目结题答辩

成员：张凌霄、徐哲、廖文惠、郭雨萌、吴思瑾

指导老师：李秀教授

目录

CONTENT

项目背景及意义

PART ONE



项目内容

PART TWO



产品Demo

PART THREE



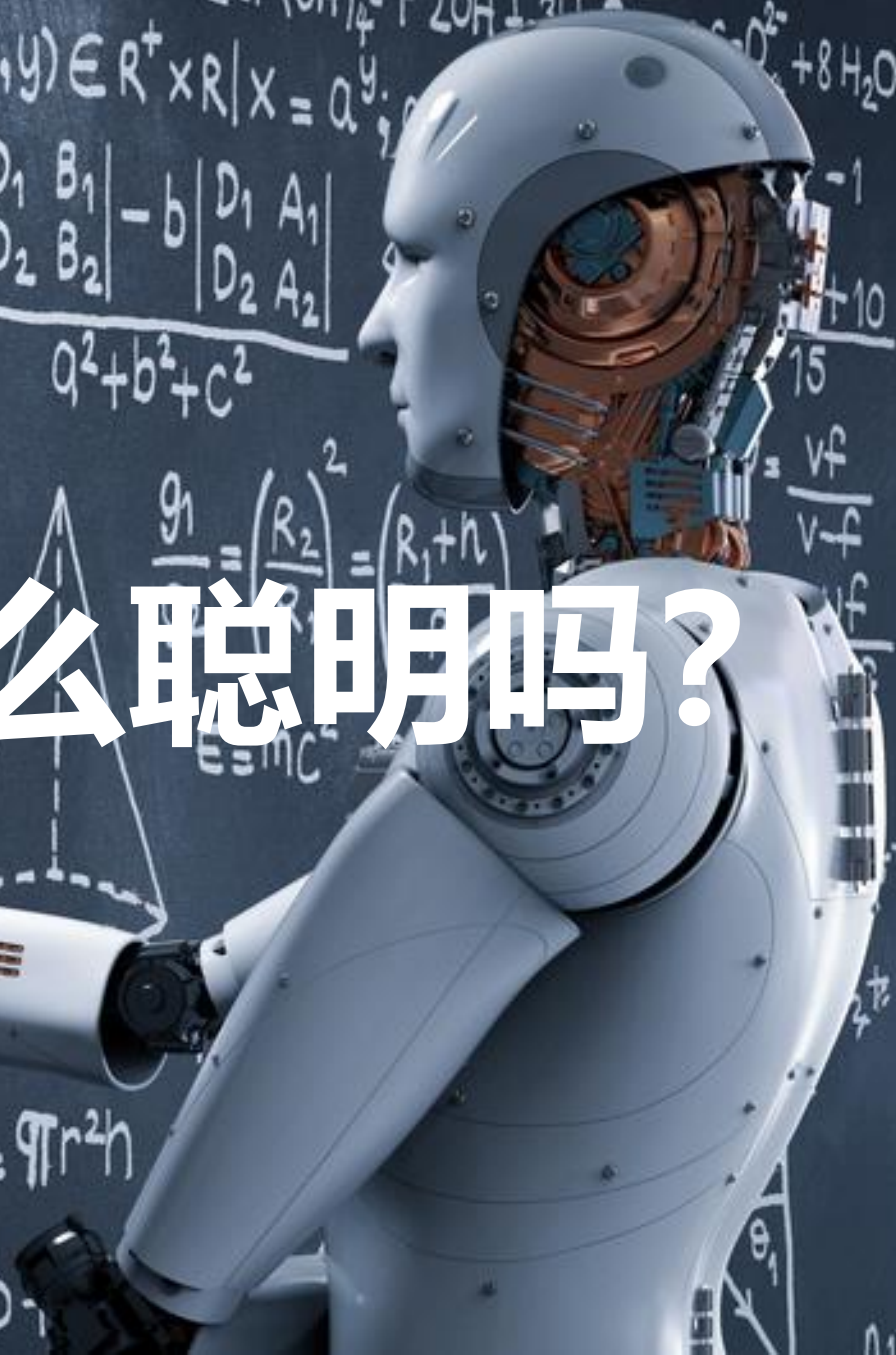
总结与展望

PART FOUR



项目背景及意义

人工智能真的有那么聪明吗？

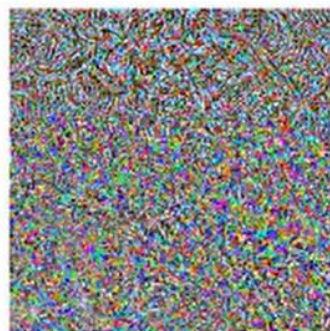


对抗样本

对抗样本是指那些经过特定优化，使得对模型的输入进行了错误分类。如果输入样本是一个自然得到的样本，比如来自 ImageNet 数据集的照片，我们称之为「干净样本」。如果攻击者修改了样本，目的是使得该样本会被误分类，我们称之为「对抗样本」。



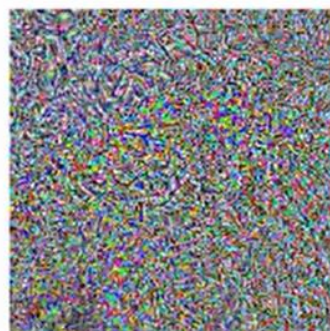
Alps: 94.39%



Dog: 99.99%



Puffer: 97.99%



Crab: 100.00%

对抗样本



x

“panda”

57.7% confidence

+ .007 ×



$\text{sign}(\nabla_x J(\theta, x, y))$

“nematode”

8.2% confidence

=



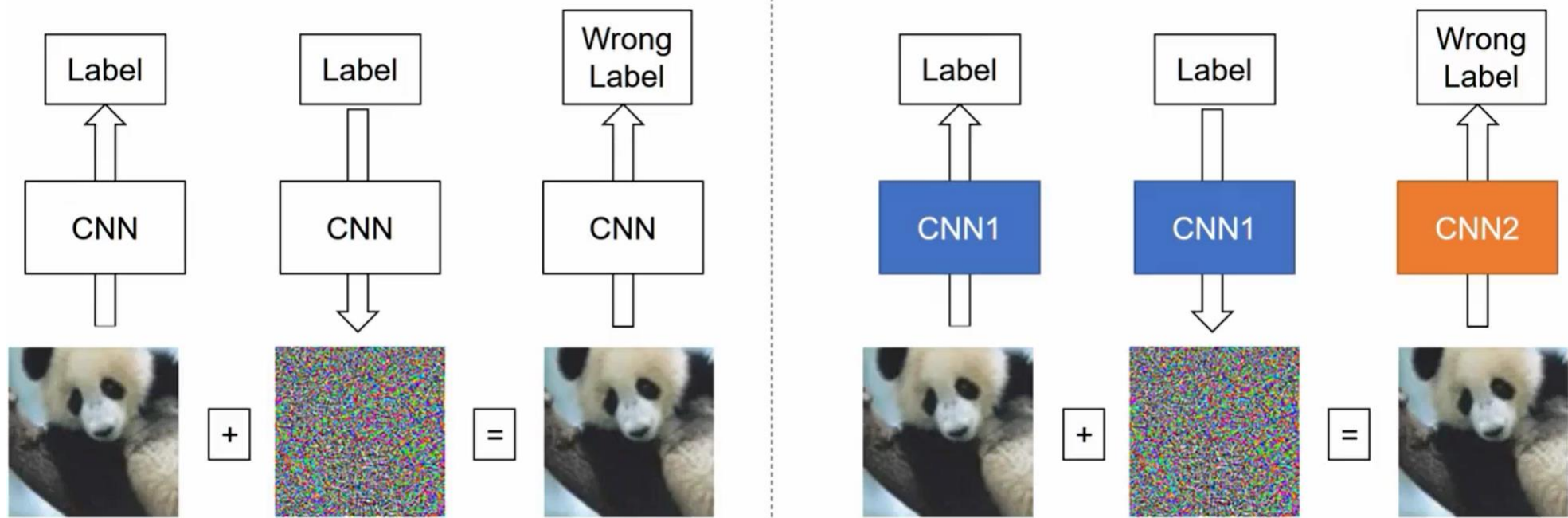
$x +$

$\epsilon \text{sign}(\nabla_x J(\theta, x, y))$

“gibbon”

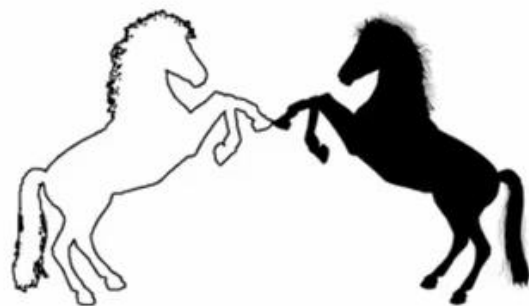
99.3 % confidence

模型了解程度：白盒与黑盒攻击



产生图片所用的CNN和需要攻击的CNN是同一个，我们称为白盒攻击。与之相反的攻击类型称为黑盒攻击，也就是对需要攻击的模型一无所知，但由于其具有一定的迁移性，当我们在ResNet, ImageNet等训练后，是可以迁移到未知模型的。

按攻击目标：有目标攻击和无目标攻击

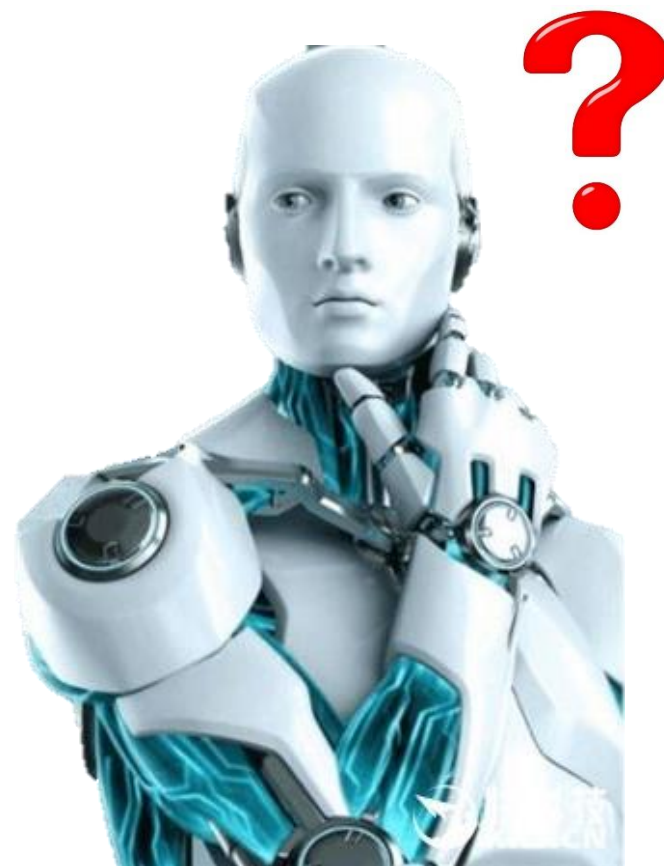
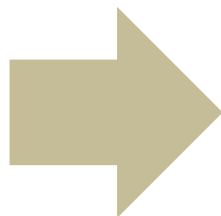
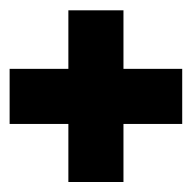


**TARGETED
ATTACK**



**NON-TARGETED
ATTACK**

人脸攻击?



各种人脸识别的场所：上班考勤、清华刷脸系统...

项目内容

训练数据集：LWF公开数据集



24000 位名人人脸

测试平台——AWS名人识别系统(Rekognition)



<https://us-east-2.console.aws.amazon.com/rekognition/home?region=us-east-2#/celebrity-detection>

传统的攻击方法 - FGSM

- Objective

$$\operatorname{argmax}_{x^*} L(x^*, y) \quad \text{s.t.} \quad \|x^* - x\|_{\infty} \leq \epsilon$$

- Linear:

$$L(x^*, y) = L(x, y) + (x^* - x) \cdot \nabla_x L(x, y)$$

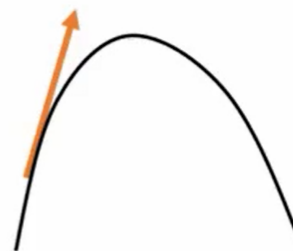
$$\Rightarrow x^* = x + \epsilon \cdot \operatorname{sign}(\nabla_x L(x, y))$$

x^* 是对抗样本, x 是原始样本, y 是正确的prediction, L 是loss function(用交叉熵), 即最大化 x^* 与 x 的差别输进去, 最小化在正确label下的概率 (Loss Function越大, 概率越小)。

PGD - 迭代FGSM

- Fast Gradient Sign Method (FGSM)

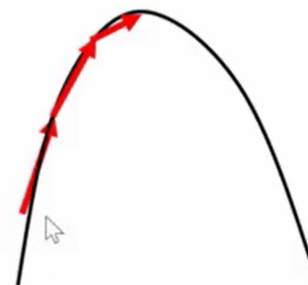
$$x^* = x + \epsilon \text{sign}(\nabla_x L(x, y))$$



- Projected Gradient Descent
(Iterative FGSM)

$$x_0^* = x,$$

$$x_{t+1}^* = \text{clip}(x + \alpha \text{sign}(\nabla_x L(x_t^*, y)))$$



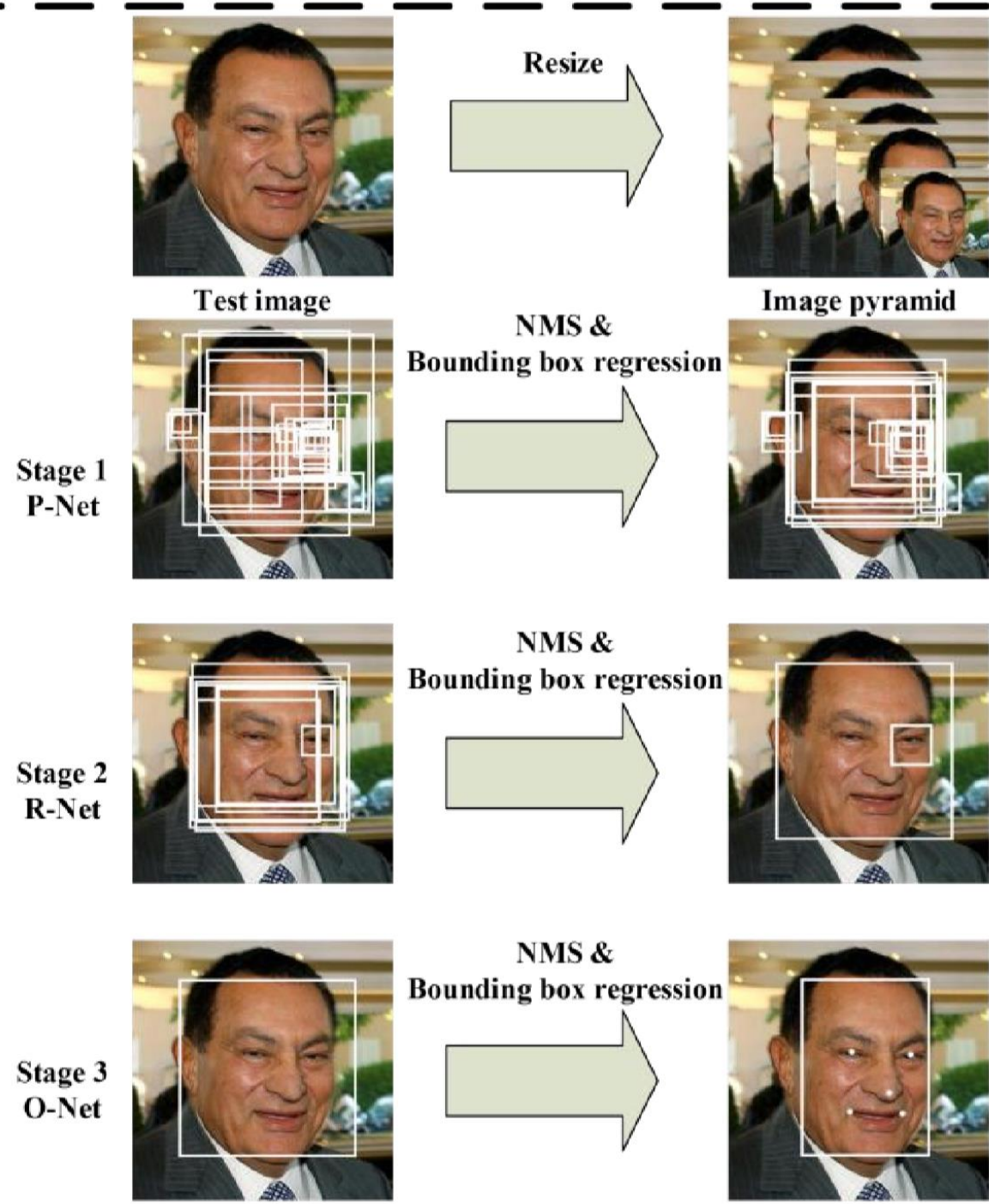
单步的步长比较大，迭代多步则缩短步长。

MTCNN – 人脸检测

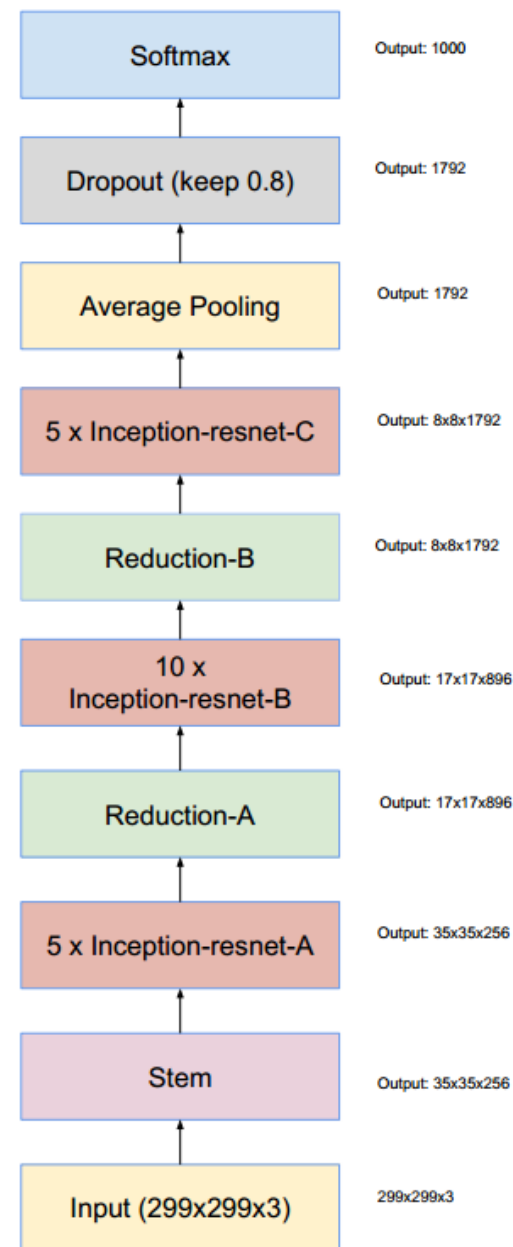
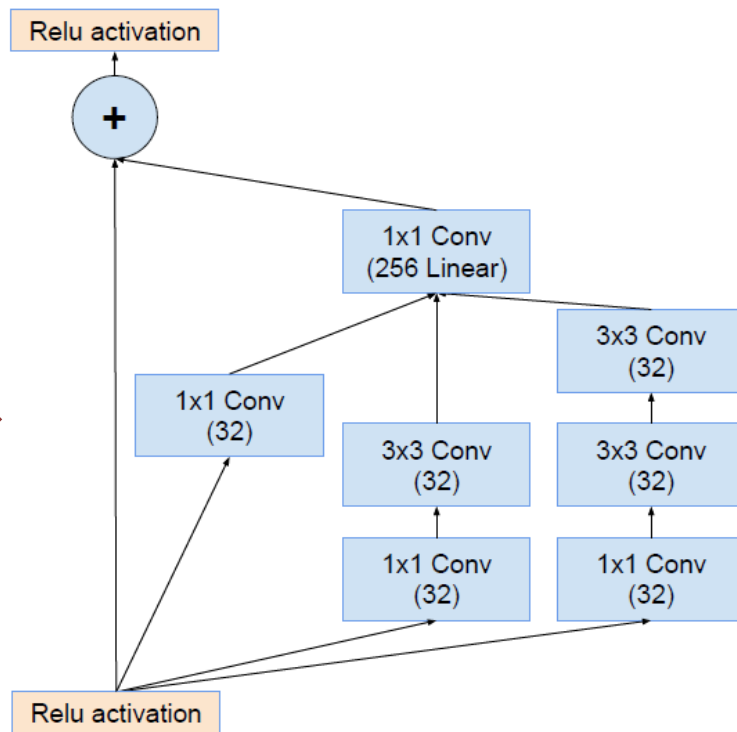
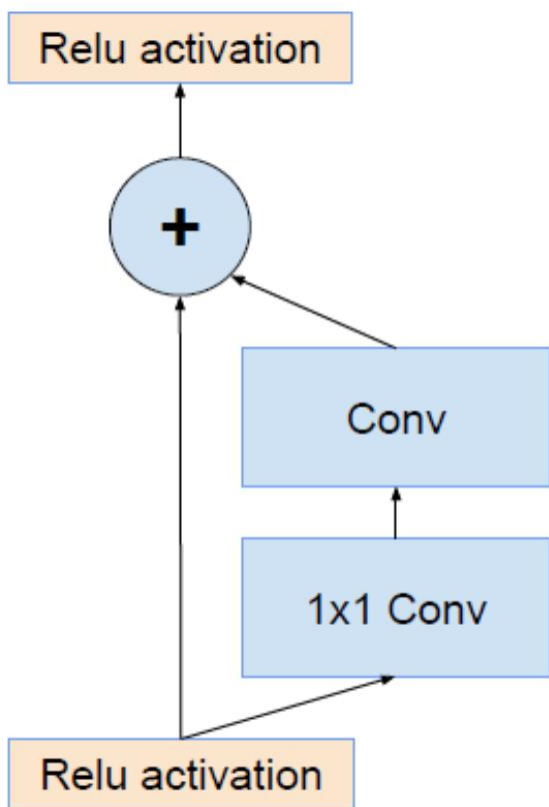
Multi-task convolutional neural network (多任务卷积神经网络)，将人脸区域检测与人脸关键点检测放在了一起，基于cascade框架，总体可分为PNet、RNet、和ONet三层网络结构。



https://kpzhang93.github.io/MTCNN_face_detection_alignment/

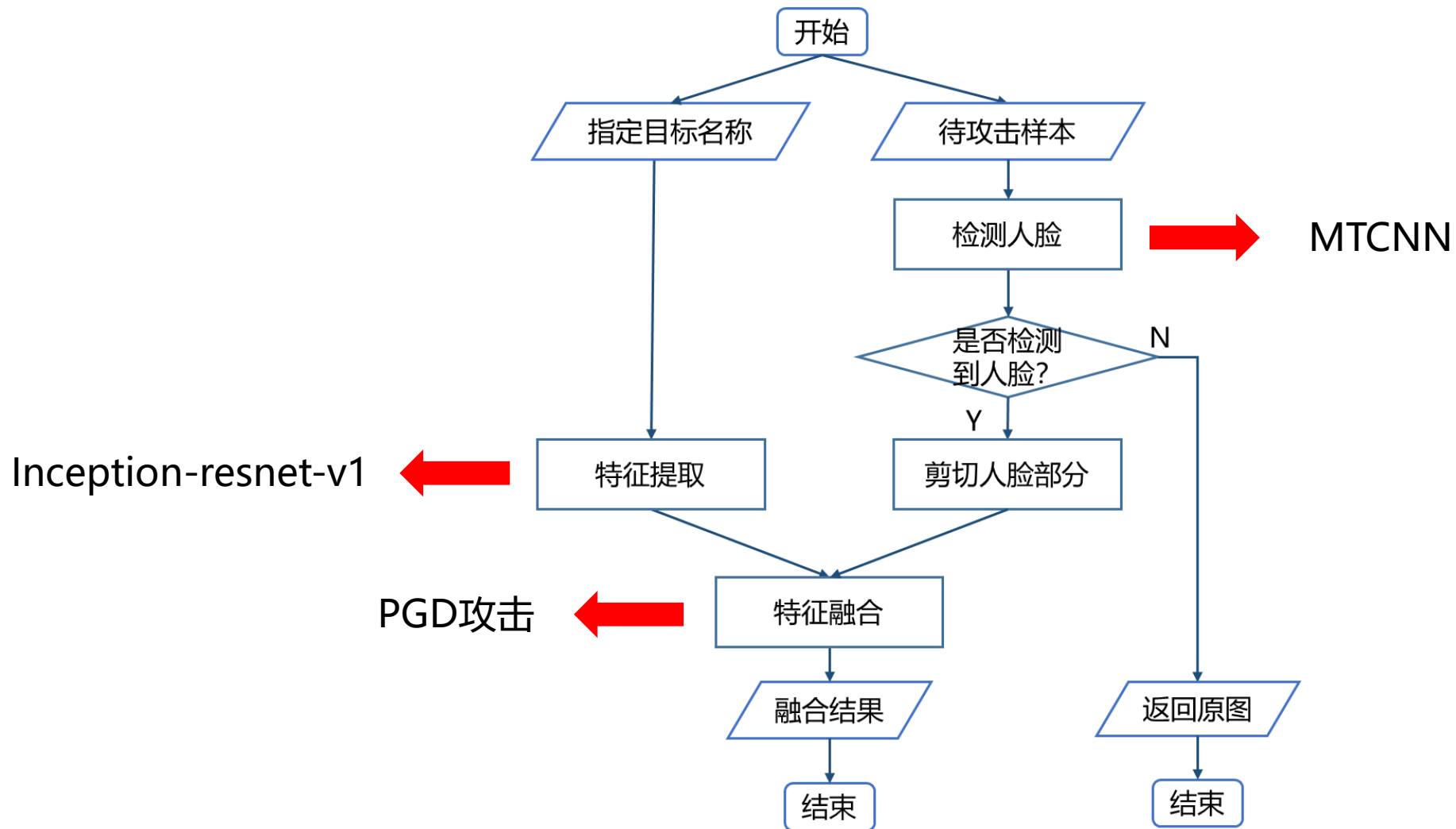


特征提取 – InceptionV4



Inception-ResNet-v1

对抗样本生成框架 - 目标攻击





Demo



Demo QR Code ((到时候我放服务器上让它跑着))

Demo – 目标攻击

aws 服务 资源组


owenzhang 俄亥俄 支持

Amazon Rekognition

- 指标
- 演示
- 对象和场景检测
- 图像监管
- 面部分析
- 名人识别**
- 面部比较
- 图像文本识别
- 视频演示
- 视频分析
- 其他资源
- 入门指南
- 下载开发工具包
- 开发人员资源
- 定价
- 常见问题
- 论坛


名人识别

Rekognition 会自动识别图像中的名人，并提供置信度得分。




演示完成?
[了解更多](#)

结果




Rafael Nadal
[了解详情](#)



Rafael Nadal
[了解详情](#)

选择示例图像



使用您自己的图像
图片必须是 .jpeg 或 .png 格式，
大于5MB。没有存储您的图像。

[上传](#) 或者拖放

[前往](#)

Demo – 目标攻击



纳达



什么变化
纳达尔

Demo ((展示我们网站...演示一下生成的步骤))

Demo – 目标攻击



服务

资源组



owenzhang

俄亥俄

支持

Amazon Rekognition

指标

演示

对象和场景检测

图像监管

面部分析

名人识别

面部比较

图像文本识别

视频演示

视频分析

其他资源

入门指南

下载开发工具包

开发人员资源

定价

常见问题

论坛

名人识别

Rekognition 会自动识别图像中的名人，并提供置信度得分。



演示完成?
[了解更多](#)

结果



Arnold Schwarzenegger
[了解详情](#)

98 %

Arnold Schwarzenegger
[了解详情](#)

匹配置信度

选择示例图像



使用您自己的图像

图片必须是.jpeg 或 .png 格式，不得大于5MB。没有存储您的图像。

[上传](#)

或者拖放

[前往](#)

```
    "BoundingBox": {  
      "Width": 0.14613179862499237,  
      "Height": 0.22030237317085266,  
      "Left": 0.36246418952941895,  
      "Top": 0.1576673835515976
```

Demo – 目标攻击



服务

资源组



owenzhang

俄亥俄

支持

Amazon Rekognition

指标

演示

对象和场景检测

图像监管

面部分析

名人识别

面部比较

图像文本识别

视频演示

视频分析

其他资源

入门指南

下载开发工具包

开发人员资源

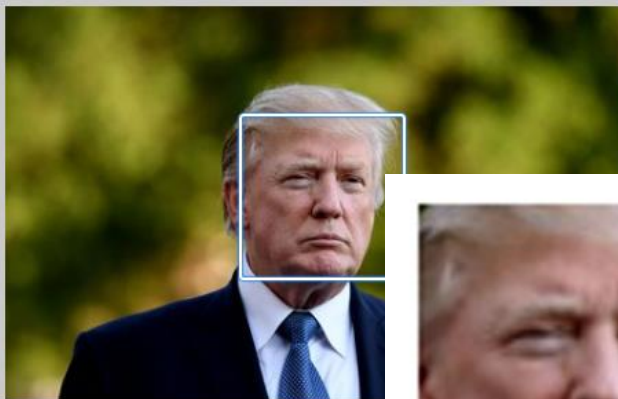
定价

常见问题

论坛

名人识别

Rekognition 会自动识别图像中的名人，并提供置信度得分。



演示完成?
[了解更多](#)

结果



Donald Trump
[了解详情](#)

匹配置信度

95 %

Donald Trump
[了解详情](#)

选择示例图像



使用您自己的图像

图片必须是 jpeg 或 png 格式，不得大于 5MB。没有存储您的图像。

上传

或者拖放

使用图像 URL

[前往](#)

Demo – 目标攻击



还是川普

Demo – 目标攻击

aws 服务 资源组


Amazon Rekognition

- 指标
- 演示
- 对象和场景检测
- 图像监管
- 面部分析
- 名人识别**
- 面部比较
- 图像文本识别
- 视频演示
- 视频分析
- 其他资源
- 入门指南
- 下载开发工具包
- 开发人员资源
- 定价
- 常见问题
- 论坛

名人识别

Rekognition 会自动识别图像中的名人，并提供置信度得分。

结果





Hillary Clinton
了解详情

匹配置信度 **88 %**

81 %

选择



[上传](#) 或者拖放

使用图像 URL [前往](#)

```
"Landmarks": [
  {
    "Type": "eyeLeft",
    "X": 0.4725761115550995,
    "Y": 0.4341590702533722
  },
  ...
]
```

Demo – 目标攻击

名人识别

Rekognition 会自动识别图像中的名人，并提供置信度得分。



选择示例图像



使用您自己的图像

图片必须是.jpeg 或.png 格式，不得大于5MB。没有存储您的图像。

 上传

或者拖放

演示完成?

[了解更多](#)

▼ 结果



Lin Dan

匹配置信度

100 %

▶ 请求

▼ 响应

```
{
  "CelebrityFaces": [
    {
      "Urls": [],
      "Name": "Lin Dan",
      "Id": "13yi8Ha",
      "Face": {
        "BoundingBox": {
          "Width": 0.2666666805744171,
          "Height": 0.2666666805744171,
          "Left": 0.344999988079071,
          "Top": 0.12166666984558105
        },
        "Confidence": 99.99852752685547,
        "Landmarks": [
          {
            "Type": "eyeLeft",
```

▼ 结果



Lin Dan

匹配置信度

100 %

▶ 请求

Demo – 目标攻击 ((明天拿算法攻击下林丹))

总结与展望

总结

算法层面

- 优化loss值，进一步减小对抗样本于原图差异
- 真实场景下，三维人脸攻击

应用层面

- 优化网页功能、外观
 - 与网络安全公司合作
-



人工智能时代 什么是真？什么是假？