# WEB2020 大作业

王渭森 2019101408

## 1 引言

近年来，随着互联网技术的发展，智能手机的普及，以及各种网络信息平台的涌现，人们无时无刻不被海量的信息包围。然而，这些信息是良莠不齐、真假难辨的。虚假信息的肆意传播经常会给个人乃至社会造成不良的影响，甚至可能引发网络霸凌、社会动荡等恶性事件。合成图像是虚假信息的一种表现形式，相比于虚假的文字消息，制作精良的虚假图像利用了人们“眼见为实”的普遍心理，具有更大的影响力与破坏力。其中，人脸合成图像已被滥用于各类虚假信息中，如制造虚假新闻、色情视频换明星脸等。随着深度学习技术的广泛研究与发展，使用该技术合成的人脸图像愈发真实精细，人眼难以对此准确的做出判断。这种利用深度学习合成虚假人脸的技术被称之为 Deepfake。

本次作业，我着眼于 deepfake 的检测。我使用了深度学习技术，用深度卷积神经网络（CNN）来对抗 deepfake。另外，我借鉴了隐写检测领域被广泛使用的局部残差特征来探究合成图像与真实图像的局部细节差异。

## 2 方法

给定一张人脸图像 I，经过预处理后得到输入 x。将 x 输入模型 h 中，得到对于 I 是否为合成图像的二分类判断。

x= preprocessing(I)

c ← h(I)

c ∈ {真实图像，合成图像}

模型 h 可以选用各种 SOTA 的 CNN 模型作为基本框架。在本次作业中我使用了 Resnet18，因为该模型体量小，易于训练，被广泛用于各种图像分类任务中。

合成图像检测的一个思路是检测图像局部的异常细节。然而，普通的卷积网络相较之下更加关注于图像的内容而非图像的细节。根据隐写分析领域的研究，使用高通滤波器可以捕捉到图像中的局部残差特征，为一种图像细节特征，有助于隐写检测。因此，我将模型的第一层卷积的卷积核替换为隐写检测常用模型 Spatial Rich Model 中的一组卷积核来捕捉局部细节特征。使用的卷积核如下图：

$$\frac{1}{4}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 2 & -4 & 2 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \frac{1}{12}\begin{bmatrix} -1 & 2 & -2 & 2 & -1 \\ 2 & -6 & 8 & -6 & 2 \\ -2 & 8 & -12 & 8 & -2 \\ 2 & -6 & 8 & -6 & 2 \\ -1 & 2 & -2 & 2 & -1 \end{bmatrix} \quad \frac{1}{2}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

该卷积层被称为 SRMConv。在训练过程中，SRMConv 的参数不做更新。

最后，我利用一个双流卷积神经网络，融合图像内容特征与图像细节特征来进行合成检测。一流的第一层卷积仍使用普通卷积，一层改为 SRM 卷积，经过全剧平均池化后拼接两个流的特征向量，再送入最后一层的二分类器。

# 3 实验设置

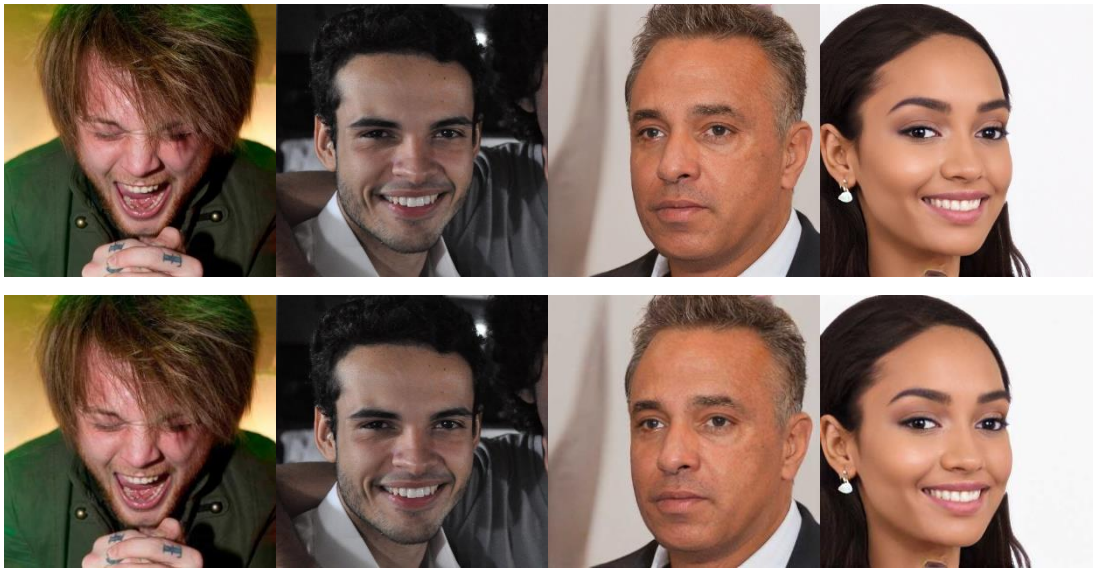
## 3.1 数据

本次作业中我整理了两个数据集。数据集 A 包含了 4000 张人脸图像，图像分辨率为 512×512，其中 2000 张为真实图像，2000 张为合成人脸。真实图像中，1000 张来自于 celebA-HQ 数据集，1000 张来自于 Flickr-Faces-HQ。至于合成图像，1000 张从 http://www.thispersondoesnotexist.com/ 网站采集，该网站对每次访问返回一个由 styleGAN2 模型合成的人脸图像；另 1000 张采集自 https://generated.photos/，该网站收集了合成人脸图像并提供了 API 使用接口。

数据集使用了 4000 张人脸图像。所有的图像原始分辨率为 1024×1024，为了方便上传数据，我将其放缩为 512×512 存入数据集 A 中。

考虑到网上传输图像时会进行压缩，我利用 jpg 压缩技术处理了数据集 A 中的图像，得到数据集 B，压缩率为 50%。

下图中，第一行为数据集 A 中的图像，第二行为相应的数据集 B 图像。从左到右分别来自于 celebA-HQ、Flickr-Faces-HQ、thispersondoesnotexist 和 generated.photos。



数据集 A 按 8：1：1 的比例随机采样，分为训练集，验证集和测试集。数据集 B 按照 A 的划分结果来分配相应的压缩后图像。

## 3.2 实验

对于数据集 A，我分别使用了普通的 CNN，带有 SRMConv 的 CNN，以及双流 CNN 来做合成检测。使用 SGD 以 0.01 的学习率迭代 50 轮后，将学习率降为 0.001 继续迭代五十轮。每一轮迭代后，我在验证集上测试正确率，将正确率最高的模型参数保存下来作为最终

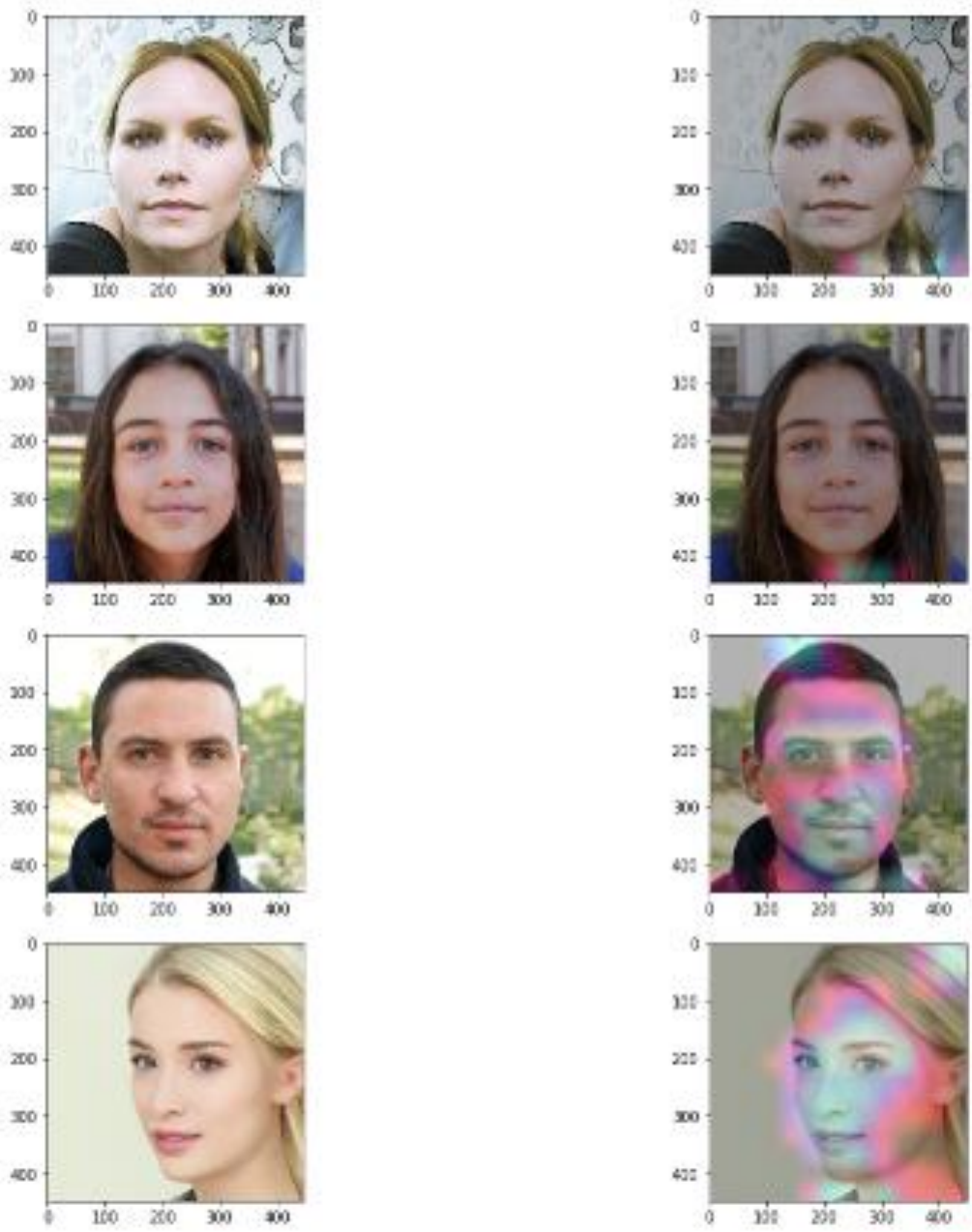
的最优模型参数。

下表为在测试集上的正确率：

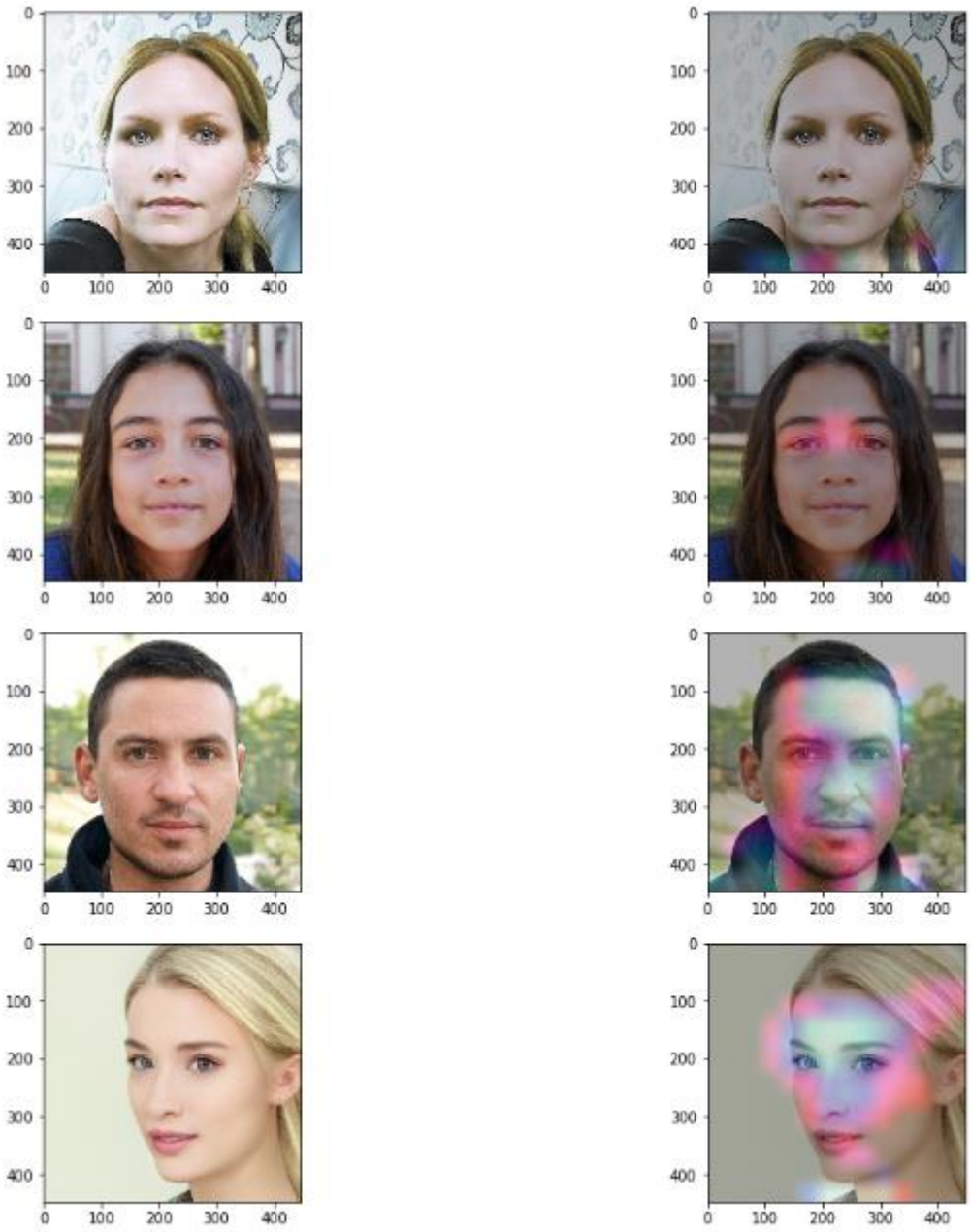
| | 普通 CNN | SRMConv CNN | 双流 CNN |
|---|---|---|---|
| 测试集正确率 | 1 | 1 | 1 |
| 验证集正确率 | 0.995 | 1 | 1 |

根据实验结果，预测是比较准确的。我使用了类激活图对网络的关注进行了可视化。

下图为普通 CNN 的预测结果，前两张为真实图像，后两张为合成图像。

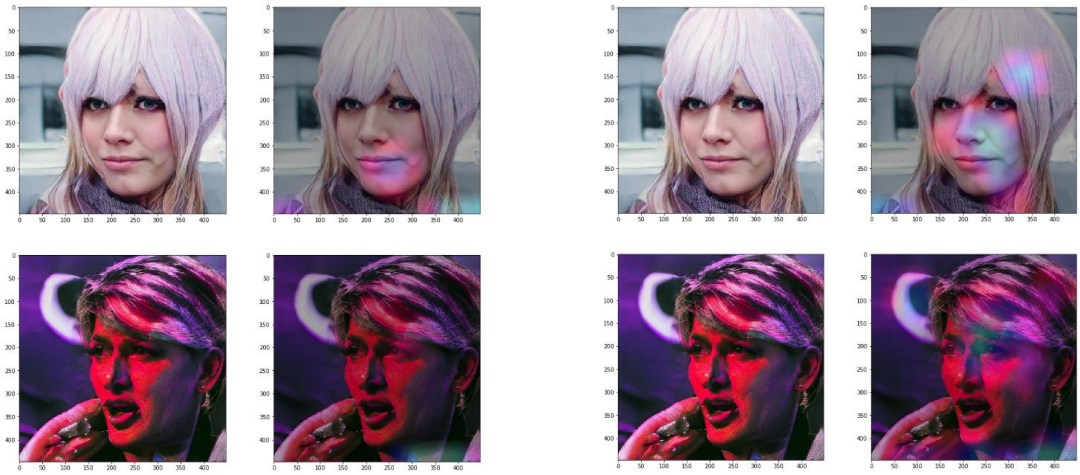


下图为 SRMconv CNN 的预测结果，前两张为真实图像，后两张为合成图像。




我们发现在合成图像人脸出产生了高激活区域，尤其是眼睛、鼻梁与嘴巴。这说明网络成功学习到了合成图像中面部的异常，与我的期望相符合。

另外，注意到在验证集中普通 CNN 有两张漏判，即将合成图像认成正常图像。我打印了这两张图的类激活图。下图中左侧为普通 CNN 生成的，右侧为 SRMConv CNN 生成的。



可以看出，SRMConv 相比于普通 CNN，发现了头发处的异常，或许 SRMConv 对于纹理细节更敏感。

对于数据集 B，我使用了普通 CNN 和 SRMConv CNN，探究 jpg 压缩对于合成检测的影响。最优模型保存方法与上述相同。在测试集上正确率见下表。

| | 普通 CNN | SRMConv CNN |
|---|---|---|
| 测试集正确率 | 0.975 | 0.9475 |
| 验证集正确率 | 0.9825 | 0.965 |

Jpg 压缩严重影响了局部的噪声特征，使得 SRMConv CNN 相对于普通 CNN 的优势不复存在。相较于不压缩的图像，压缩数据集的检测准确率也有所降低。

# 4 讨论

本次作业对使用卷积神经网络做合成图像检测任务进行了探究。通过实验观察，推测出以下几个结论：

1. CNN 在 deepfake 检测任务中有较好的表现。
2. 对于不压缩的图像，SRMConv 能捕捉到图像细节异常，比普通卷积有更好的表现。
3. 对于压缩图像，SRMConv 并不适用。

本次实验的局限性：

1. 数据集有偏差。由于所有的图像都来自于四个数据集，图像可能带有数据集的特征分布，导致网络学习到了数据及分布差异，而非合成人脸的异常部分。
2. 对于上一点，一个好的解决方法是在第五个数据集中做测试，然而由于时间与精力有限，我并没有进一步做这方面的探究。