DeepFake 检测实验报告

马江岩

2021 年 6 月 23 日

摘要

本次实验, 我们首先从理论上证明了: 生成对抗网络常采用的上采样 (up-sampling) 操作, 会扰乱图像的功率谱特征 (Power Spectrum); 随后我们利用这一原理, 分别利用支持向量机 (SVM) 和逻辑回归 (Logistic Regression) 训练二分类器, 用于鉴别图片是真实照片还是由生成对抗网络的 DeepFake 图片. 我们的方法在 CelebA 和 Faces-HQ 数据集上均达到了 100% 的检测成功率.

1 引言

近年来, 随着人工智能技术的发展, 生成对抗网络 (GAN) 被广泛运用于图像的生成和编辑. 生成对抗网络的出色性能, 在给我们的生活带来便利的同时, 也伴随着一些负面的运用. 例如, 用于篡改并生成虚假图像的 AI 换脸技术 (DeepFake), 被用于制作明星的色情图片和视频、报复性色情视频、假新闻、恶作剧、金融欺诈等, 为我们的社会带来了潜在的风险和危害. 并且, 相比于传统的、需要大量精力和高超技术的图片修改手段, 利用 GAN 生成假图片, 成本低、门槛低, 即使是对计算机视觉缺乏了解的用户, 也能利用生成对抗网络, 依据他们的意愿轻易生成肉眼难以分辨的假照片, 从而损害他人的利益. 因此, 寻找有效的手段, 对 GAN 生成的假照片进行鉴别, 是当前计算机视觉领域迫在眉睫的、意义重大的工作.

我们指出, 生成对抗网络中常用的上采样 (up-sampling) 操作, 会对图像的功率谱特征 (Power Spectrum) 造成一定的扰动, 从而可以被识别出来. 以 Faces-HQ 数据集为例. 如图 1 所示, Faces-HQ 数据集由 4 个数据集构成: `www.thispersondoesnotexist.com`, 100K Faces Project, Flickr-Faces-HQ 和 CelebA-HQ, 每个数据集各含 10000 张照片, 其中前两个数据集为 GAN 生成的假照片, 后两个数据集为真实的图片. 可以看出, 假照片的功率谱特征高于真实图片的功率谱特征, 这一点在高频率区间上尤为明显. 我们可以据此鉴别真假照片.

我们的主要工作如下:

- 我们指出, 生成对抗网络常用的基于卷积的上采样操作, 会扰乱所生成图像的功率谱特征, 使 DeepFake 图片的功率谱特征在高频区间相较于真实图片偏高或偏低.
- 我们从理论上分析并证明了不同的上采样操作 (上卷积、反卷积) 对于图像的功率谱特征

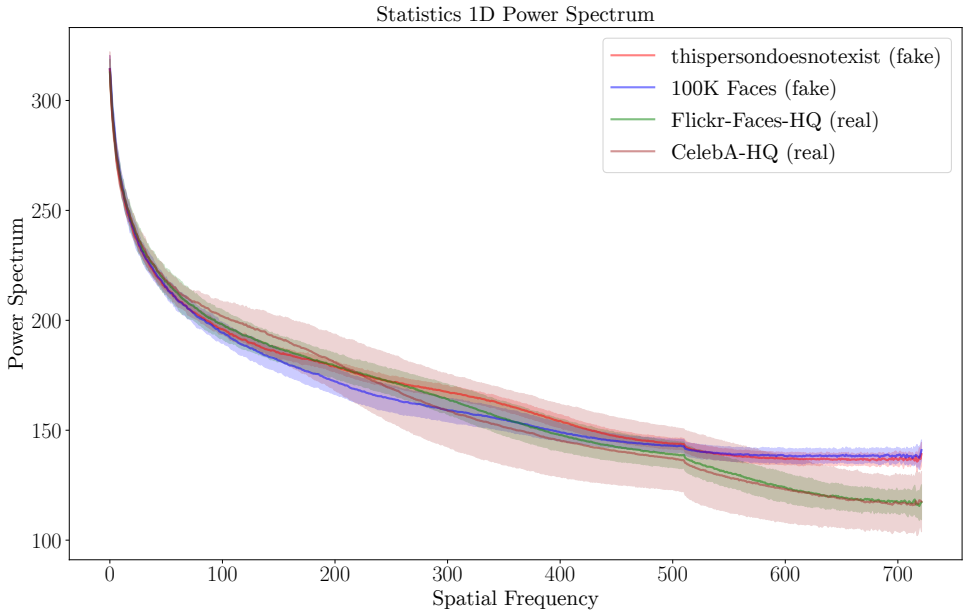

图 1: Faces-HQ 数据集的各个子数据集的图片的功率谱特征. 其中, thispersondoesnotexist 和 100K Faces 数据集是 DeepFake 图片, Flickr-Faces-HQ 和 CelebA-HQ 数据集是真实图片.

的影响.

- 我们提出了一种简单高效的鉴别 DeepFake 的方法, 利用支持向量机和逻辑回归, 在 CelebA, FaceForensics 和 Faces-HQ 数据集上分别训练了两个二分类器, 并测试其效果. 在 CelebA 和 Faces-HQ 数据集上, 我们的两种分类器均达到了 100% 的鉴别成功率; 在分辨率较低的 FaceForensics 数据集上, 我们基于支持向量机的二分类器的分类成功率为 86%, 基于逻辑回归的二分类器的分类成功率为 77%.

2 基于卷积的上采样

上采样的目的是将低分辨率的图像转化为高分辨率的图像. 这里我们简要介绍两种上采样的方法: 转置卷积 (transposed convolution) 和上采样 + 卷积 (up-sampling+convolution). 它们的示意图如图 2 所示.

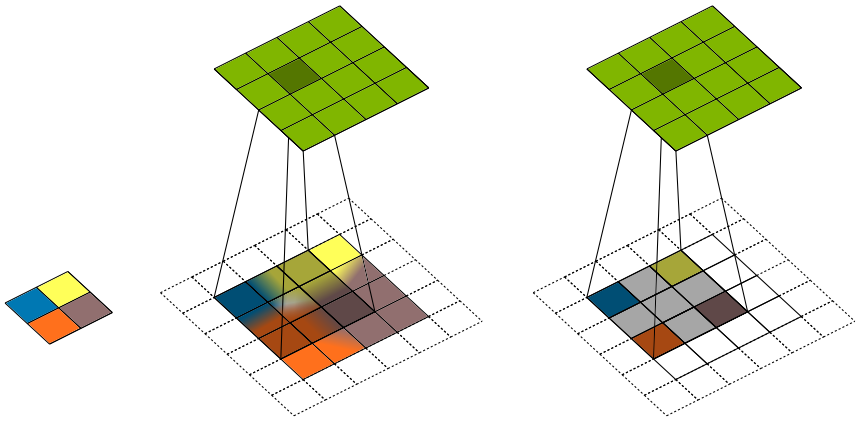

图 2: 基于卷积的上采样方法的示意图. 左: 低分辨率的输入图像 (图中为 2×2); 中: 利用插值的上卷积 (up+conv)—图像先通过插值进行放大, 再进行常规的卷积操作; 右: 转置卷积 (transconv)—图像先用零填充, 再进行卷积操作.

2.1 转置卷积

在转置卷积过程中, 设输入图像的尺寸为 $C_{\text{in}} \cdot H_{\text{in}} \cdot W_{\text{in}}$, 输出图像的尺寸为 $C_{\text{out}} \cdot H_{\text{out}} \cdot W_{\text{out}}$, 权值的大小为 $C_{\text{in}} \cdot C_{\text{out}} \cdot K \cdot K$. 则转置卷积的运算流程图如图 3 所示.

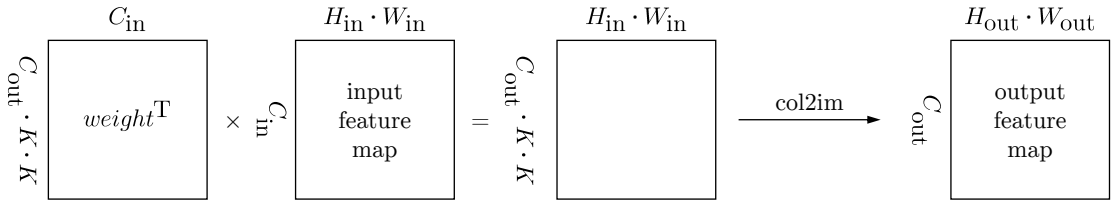

图 3: 转置卷积的运算流程图. 输入图像的尺寸为 $C_{\text{in}} \cdot H_{\text{in}} \cdot W_{\text{in}}$, 通过与尺寸为 $C_{\text{out}} \cdot K \cdot K \cdot C_{\text{in}}$ 的权值的转置做矩阵乘法, 再经过 col2im 操作, 得到尺寸为 $C_{\text{out}} \cdot H_{\text{out}} \cdot W_{\text{out}}$ 的输出图像.

转置卷积的前向过程首先做了一个矩阵乘法 (或者也可以用 1×1 的卷积实现), 把输入图像的通道数从 C_{in} 变成了 $C_{\text{out}} \cdot K \cdot K$, 这时候图像的大小还是保持 $H_{\text{in}} \cdot W_{\text{in}}$ 的. 接下来的一步 col2im 操作, 相当于卷积里的 im2col 的逆过程, 把中间结果的每一列从一个 $C_{\text{out}} \cdot K \cdot K$ 的向量, 重构成尺寸为 (C_{out}, K, K) 的张量, 然后根据 C_{out} 的索引把对应 (K, K) 的块回填累加到输出图像对应通道的位置上.

2.2 上采样 + 卷积

上采样例如双线性插值, 就是把图像根据上采样率插值出更大的图像, 图像之间的上采样是各自独立的, 所以通道数不会改变, 然后再用一个卷积操作提取特征, 这一步才可以改变通道数.

3 图像的功率谱特征

为了研究生成对抗网络对图像频谱分布的影响, 我们采用一种简单但有效的一维的傅里叶功率谱的表示. 对于 $M \times N$ 的输入图像 I, 我们利用下式计算其离散傅里叶变换 (Discrete Fourier Transform)$\mathcal{F}$:

$$\mathcal{F}(I)(k,\ell) = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} \mathrm{e}^{-2\pi \mathrm{i}\cdot\frac{jk}{M}} \mathrm{e}^{-2\pi \mathrm{i}\cdot\frac{j\ell}{N}} \cdot I(m,n), \text{ for } k = 0,\ldots,M-1, \ell = 0,\ldots,N-1. \quad (1)$$

(1)式得到的数据是二维的. 为了将数据化为一维, 我们对径向频率 ϕ 进行方位角平均 (azimuthal integration):

$$AI(\omega_k) = \int_0^{2\pi} \left\|\mathcal{F}(I)\left(\omega_k \cdot \cos\phi, \omega_k \cdot \sin\phi\right)\right\|^2 \mathrm{d}\phi, \text{ for } k = 0,\ldots,\frac{M}{2} - 1. \quad (2)$$

(2)式假设图像是正方形的. 图像的功率谱特征的示意如图 4 所示.

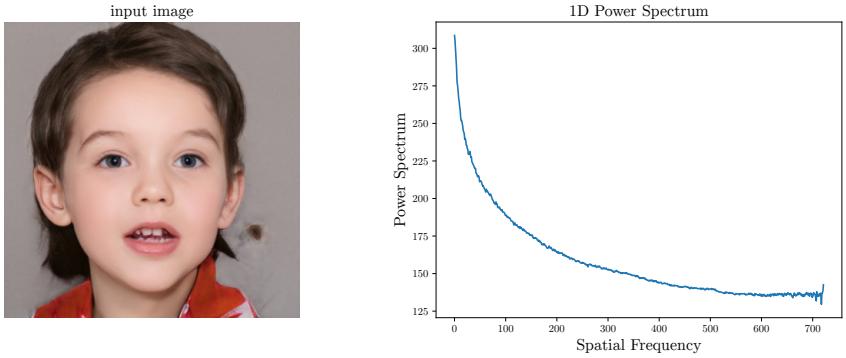

图 4: 图像的功率谱特征的示意图. 左: 输入图像. 右: 图像的功率谱特征在各个频率上的分布.

4 上采样对图像功率谱特征的扰动

本节我们来分析生成对抗网络中常用的上采样操作对图像的功率谱特征的影响. 不失一般性, 我们来考虑一个一维的信号 a 和它的离散傅里叶变换 $\hat{a}$:

$$\hat{a}_k = \sum_{j=0}^{N-1} \mathrm{e}^{-2\pi \mathrm{i} \cdot \frac{jk}{N}} \cdot a_j, \text{ for } k = 0, \ldots, N-1. \tag{3}$$

若我们将 a 的频谱特征的分辨率增大 1 倍, 我们得到

$$\hat{a}^{up}_{\bar{k}} = \sum_{j=0}^{2N-1} \mathrm{e}^{-2\pi \mathrm{i} \cdot \frac{j\bar{k}}{2N}} \cdot a^{up}_j \tag{4}$$

$$= \sum_{j=0}^{N-1} \mathrm{e}^{-2\pi \mathrm{i} \cdot \frac{2j\bar{k}}{2N}} \cdot a_j + \sum_{j=0}^{N-1} \mathrm{e}^{-2\pi \mathrm{i} \cdot \frac{2(j+1)\bar{k}}{2N}} \cdot b_j, \tag{5}$$

$$\text{for } \bar{k} = 0, \ldots, 2N-1,$$

其中, 对于转置卷积, $b_j = 0$; 对于上采样 + 卷积, $b_j = \dfrac{a_{j-1} + a_j}{2}$.

现在我们首先考虑 $b_j = 0$ 的情况. 于是, (5)式中的第二项为 0. 第一项与原来的傅里叶变换得到的结果相似, 但参数 k 变成了 $\bar{k}$. 因此, 将频率的分辨率增大至 2 倍将导致频率的轴压缩至 $\dfrac{1}{2}$. 从抽样理论的角度考虑, 我们得到

$$\hat{a}^{up}_{\bar{k}} = \sum_{j=0}^{2N-1} \mathrm{e}^{-2\pi \mathrm{i} \cdot \frac{j\bar{k}}{2N}} \cdot a^{up}_j \tag{6}$$

$$= \sum_{j=0}^{2N-1} \mathrm{e}^{-2\pi \mathrm{i} \cdot \frac{j\bar{k}}{2N}} \cdot \sum_{t=-\infty}^{\infty} a^{up}_j \cdot \delta(j - 2t). \tag{7}$$

注意到和狄拉克梳状函数相乘仅去除了满足 $a^{up} = 0$ 的值. 假设信号是周期性的, 运用卷积定理, 有

$$(7) = \frac{1}{2} \cdot \sum_{t=-\infty}^{\infty} \left(\sum_{j=-\infty}^{\infty} \mathrm{e}^{-2\pi \mathrm{i} \cdot \frac{j\bar{k}}{2N}} \cdot a^{up}_j \right) \left(\bar{k} - \frac{t}{2} \right), \tag{8}$$

根据(6), 它等于

$$\frac{1}{2} \cdot \sum_{t=-\infty}^{\infty} \left(\sum_{j=-\infty}^{\infty} \mathrm{e}^{-2\pi \mathrm{i} \cdot \frac{j\bar{k}}{N}} \cdot a_j \right) \left(\bar{k} - \frac{t}{2} \right). \tag{9}$$

故用零插值 (bed of nails upsampling) 会产生信号 $\hat{a}^{up}$ 的高频副本.

其次, 我们考虑双线性插值的情况. 在(6)式中我们有 $b_j = \dfrac{a_{j-1} + a_j}{2}$, 这等价于用辛格 (sinc) 函数与 $\hat{a}^{up}$ 相乘. 它会抑制高频信号, 故得到的功率谱在高频区间会偏低.

5 DeepFake 检测实验

我们的检测原理如图 5 所示.

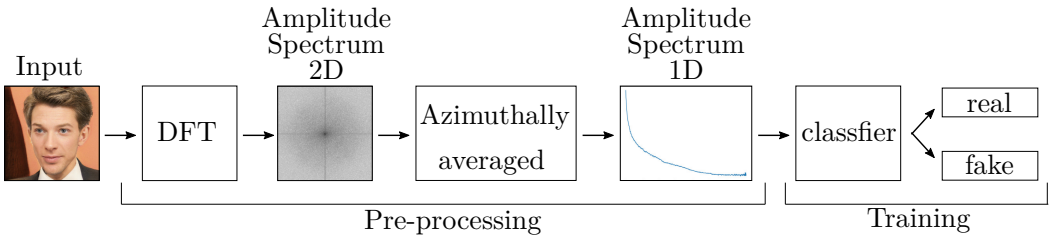

图 5: 我们的实验流程. 它包括两个模块, 即利用离散傅里叶变换的特征提取模块和训练模块. 在训练模块中, 分类器利用变换后的图像特征将人脸图像分类为真或假的.

5.1 频域分析

频域分析在信号处理和应用中具有重要的地位. 在计算机视觉领域, 我们可以在被傅里叶变换定义的空间上研究图像的频率特征. 通过对图像数据进行频率的分解, 我们可以看到信号的强度在各个频段上是如何分布的. 基于频域分析的方法已在图像处理中获得了广泛的应用.

实验中, 我们首先利用(1)式对图像进行离散傅里叶变换, 再利用(2)式将所得数据变化为一维的.

5.2 分类算法

本次实验中, 我们采用了两种分类算法.

首先是逻辑回归. 对于输入 $\mathbf{x}$, 逻辑回归的概率公式定义如下:

$$h_{\mathbf{w}}(\mathbf{x}) = \frac{1}{1 + \mathrm{e}^{-\mathbf{w}^{\mathrm{T}}\mathbf{x}}}. \tag{10}$$

极大似然估计所采用的算法决定了回归的参数 $\mathbf{w}$. 算法在达到收敛标准或达到最大迭代次数时停止.

其次是支持向量机. 支持向量机是数据分类的最常用算法之一. 算法的目的是找到一个最优的超平面, 使得不同类别到这个平面的距离最大. 给定一组数据-标签训练集 $(x_i, y_i), i =$

$1,\ldots,l$, 其中 $x_i \in \mathbb{R}^n$ 且 $\mathbf{y} \in \{1,-1\}^l$, 支持向量机的训练通过求解下面的问题而实现:

$$\min_{\mathbf{w},b,\boldsymbol{xi}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l}\xi_i, \tag{11}$$

$$\text{s.t.} \quad y_i\left(\mathbf{w}^{\mathrm{T}}\phi(x_i)+b\right) \geqslant 1-\xi_i, \xi_i \geqslant 0, \tag{12}$$

其中 $\mathbf{w}$ 和 b 是分类器的参数, ξ 为松弛变量, $C>0$ 是错误项的惩罚参数. 这里训练向量 x_i 被函数 ϕ 映射到更高维的空间. 支持向量机的目的是找到一个超平面, 使得在更高维空间中各类别到超平面的距离最大.

5.3 实验

首先我们在 Faces-HQ 数据集上进行实验. Faces-HQ 数据集由 4 个数据集构成, 即 CelebA-HQ, Flickr-Faces-HQ, 100K Faces project 和`www.thispersondoesnotexist.com`, 每个子数据集各 10000 张图片. 其中, 前两个数据集为真实图片, 后两个数据集为 GAN 生成的假图片. 数据集中的部分图像如图 6 所示.

图 6: Faces-HQ 数据集中的部分图片, 其中左侧的为 GAN 生成的假图片, 右侧为真实图片.

如图 5 所示, 我们的实验流程分为两部分. 一方面, 在预处理阶段, 我们将训练集中的每张图像从空间域转化为一维的频域, 将 $1024\times1024\times3$ 的高分辨率彩色图像转化为 722 个特征 (一维功率谱特征). 方法即为离散傅里叶变换及方位角平均. 这一变换可通过快速傅里叶变换 (Fast Fourier Transform) 优化. 注意到在变换之后, 我们将只使用图像的功率谱特征, 因为它以及包含了训练分类器所需的足够的信息. 另一方面, 预处理阶段结束后, 我们开始训练分类器. 首先, 我们将数据集分为训练集和测试集, 其中训练集占 80%, 测试集占 20%. 然后, 我们

用训练集训练分类器, 在测试集上测试效果. 我们的目标是区分真假图片, 故我们使用一个二分类器.

Faces-HQ 数据集图片的功率谱特征如图 7 所示.

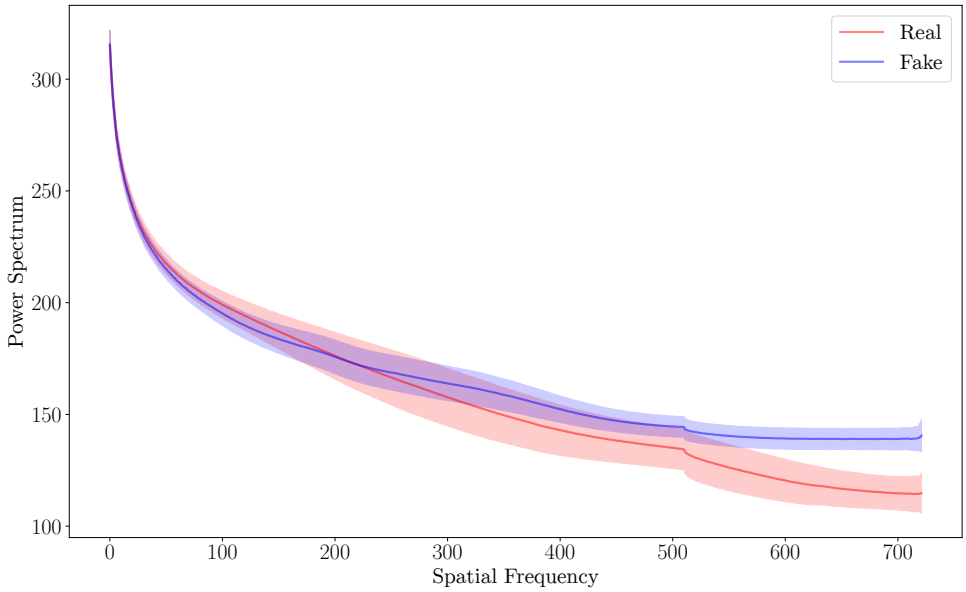

图 7: Faces-HQ 数据集中真图像和假图像的功率谱特征.

最后, 在这一数据集上, 我们用支持向量机和逻辑回归均达到了 100% 的分类成功率.

其次是 CelebA 数据集. 为了训练 DeepFake 检测器, 我们需要真的和假的图片. 其中, 我们使用 CelebA 数据集中的图片作为真实图片, 并训练了一个 DCGAN 来生成逼真的假图片. 我们将数据集中 162,770 张图片用于训练, 39,829 张图片用于测试.

CelebA 数据集图片的功率谱特征如图 8 所示.

最后, 基于支持向量机和逻辑回归的二分类器在 CelebA 数据集上均达到了 100% 的分类成功率.

然后是 FaceForensics++ 数据集. 这是一个图片分辨率相对较低的数据集. 我们的实验流程与前两个数据集基本相同, 但多了一个模块. 由于 FaceForensics++ 数据集由视频构成, 我们需要首先提取一帧的图像, 并裁取包含面部的部分. 由于视频的内容不同, 这些图片有着不同的大小. 我们训练流程的预处理阶段不依赖于图片大小, 故不需要改动. 但我们的分类器需要相同大小的图片作为输入. 因此, 我们需要添加一个额外的处理模块, 将一维的功率谱特征插值为固定的大小 (300), 并通过除以第 0 项的方法进行正则化. 其余流程不变.

FacesForensics++ 数据集中图片的功率谱特征如图 9 所示.

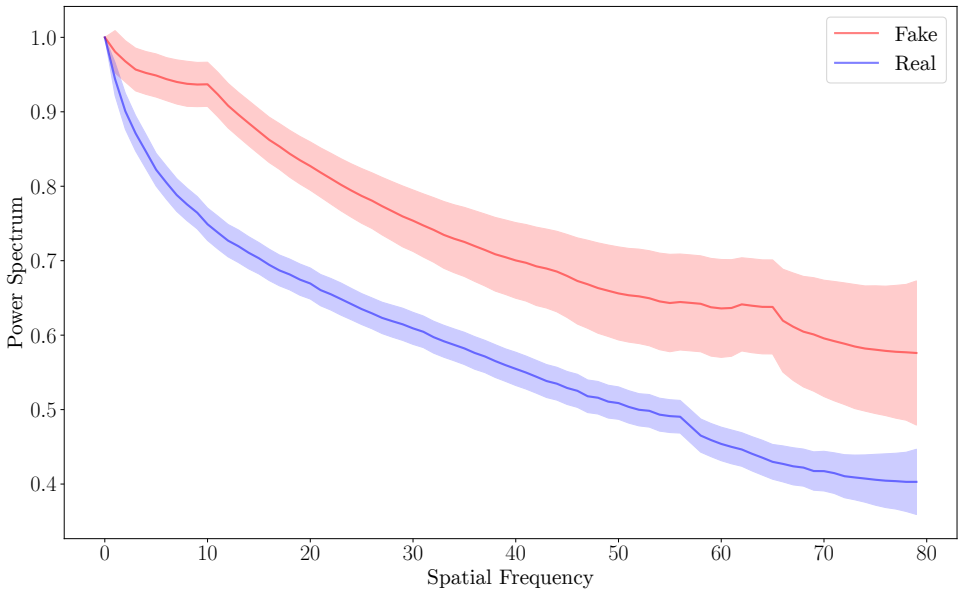

图 8: CelebA 数据集中真图像和假图像的功率谱特征.

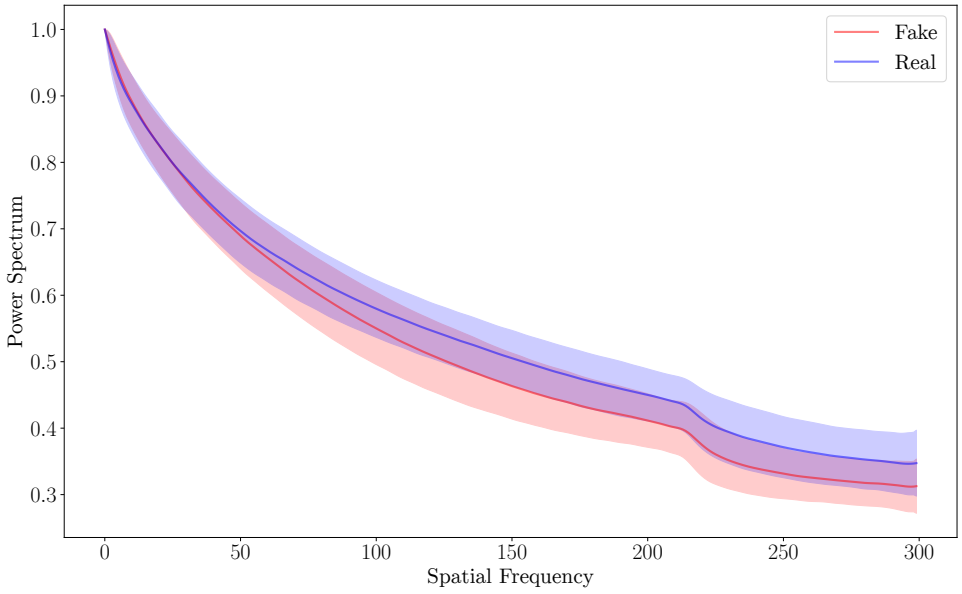

图 9: FaceForensics++ 数据集中真图像和假图像的功率谱特征.

最后, 在 FaceForensics++ 数据集上, 我们基于支持向量机的分类器的分类成功率为 86%, 基于逻辑回归的分类器的分类成功率为 77%.

6 总结

我们通过分析图像的功率谱特征, 实现了一个简单的 DeepFake 检测的方法. 和传统的基于 DNN 的 DeepFake 检测方法相比, 我们的方法具有如下优势:

- 我们的方法所需的数据更少. 由于真实图像和假图像的功率谱特征差别较明显 (尤其是在高频频段), 故我们的方法只需很少的数据就完成达到鉴别 DeepFake 图像的任务.
- 我们的方法所需算力更少. 传统的 DNN 往往需要用 GPU 训练大量时间, 而我们的方法几乎不需要什么算力, 在一台普通的笔记本电脑上仅需几秒就能完成训练和测试的全过程.
- 我们的方法成功率较高. 传统方法经过大量训练往往也只能达到百分之九十多的成功率, 而我们的方法在高分辨率和中等分辨率的数据集上均可达到 100% 的分类成功率.

在 AI 换脸技术飞速发展、日趋成熟的今天, 其所带来的风险也不容忽视. 不同于以往的图像处理方法, 利用生成对抗网络, 即使是对计算机视觉没有什么了解的人, 也能根据自己的愿望, 生成任何人的假图片和假视频. 相信我们的识别 DeepFake 的方法, 能为抵御恶意换脸的行为提供一些参考, 降低 AI 技术被不法分子利用的可能性.

参考文献

[1] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7890–7899, 2020.

[2] Ricard Durall, Margret Keuper, Franz-Josef Pfreundt, and Janis Keuper. Unmasking deepfakes with simple features. *arXiv preprint arXiv:1911.00686*, 2019.