DeepFake Detection

马江岩(2000012707)

联系邮箱：georgem@stu.pku.edu.cn

指导教师：王奕森
小组编号：39

北京大学, 人工智能引论课程
2020-2021，春季学期

摘要

我们的项目主题是 DeepFake Detection，即训练一个模型，用于区分真实照片和由神经网络生成的假照片(DeepFake)。通过提取图片的功率谱特征(power spectrum)，我们训练了一个简单的二分类器，在 Faces-HQ 和 CelebA 数据集上均达到了 100% 的检测成功率。

关键词： AI换脸，生成对抗网络，DeepFake

引言

随着人工智能的发展，诸如 AI 换脸、对图片进行人为修改的技术也日趋成熟。一方面，它能为我们的生活带来便利；但另一方面，它也带来了潜在的风险，如制作明星的色情视频、虚假新闻、金融诈骗等。对 AI 生成的假图片进行鉴别，意义重大。

方法

1. 首先对 $M\times N$ 的图像数据 I 按如下公式进行离散傅里叶变换(Discrete Fourier Transform)：

$$F(I)(k,l)=\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\mathrm{e}^{-2\pi\mathrm{i}\cdot\frac{jk}{M}}\mathrm{e}^{-2\pi\mathrm{i}\cdot\frac{jl}{N}}\cdot I(m,n),\ \text{for } k=0,\dots,M-1,l=0,\dots,N-1.$$

2. 然后利用方位角平均(Azimuthal Integration)，将变换后的 2 维的数据降为 1 维：

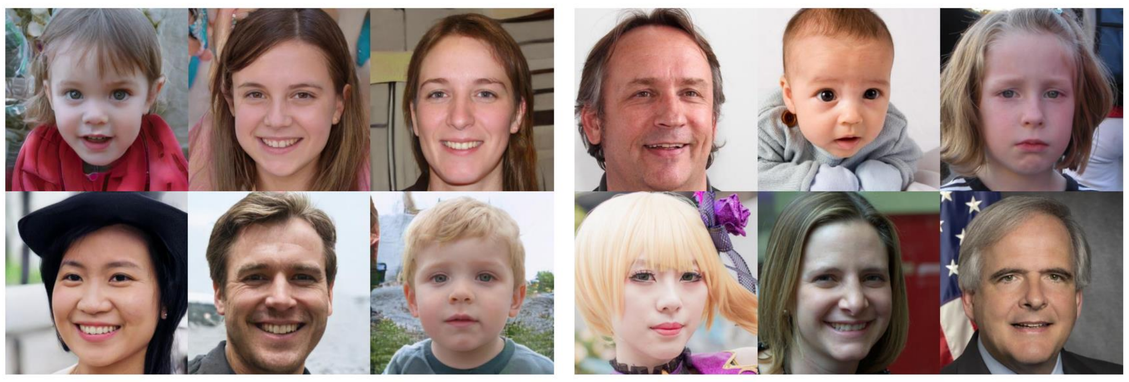
$$AI(\omega_k)=\int_0^{2\pi}\|F(I)(\omega_k\sin\varphi,\omega_k\cos\varphi)\|^2\,\mathrm{d}\varphi,\ \text{for } k=0,\dots,\frac{M}{2}-1.$$

3. 于是我们得到了图像的功率谱特征(Power Spectrum)。由于生成对抗网络(GAN)在生成图片时往往需要经过上采样(up-sampling)操作，这一操作会导致生成图片的频谱分布(spectral distribution)与真实图片不同，故我们可以根据这一现象对图像的真实性进行鉴别。
4. 以真实图像和虚假图像的功率谱特征作为训练数据，利用支持向量机(SVM)或逻辑回归(Logistic Regression)进行训练，即得到可用于鉴别 DeepFake 的机器学习模型。

实验

我们分别在 CelebA、FaceForensics、Faces-HQ 三个数据集上进行了训练和测试，测试集中部分图片如下所示，其中左边为假图片，右边为真图片。可以看出，我们的数据集的图片质量较高，可以达到以假乱真的效果，肉眼很难对图片的真假进行鉴别。

Fake **Real**



最终，我们在不同的数据集上采用不同方法得到的鉴别成功率如下表所示。

	CelebA	FaceForensics	Faces-HQ
SVM	100%	86%	100%
Logistic Regression	100%	77%	100%

总结

相比于传统的用 DNN 通过大量数据进行训练的方法，我们通过分析图像的功率谱特征进行鉴别，所需的数据量更少，且对算力的要求极少（只需几秒钟的训练和测试时间），并达到了 100% 的鉴别成功率。我们的方法也具有较高的普适性，凡是用 GAN 生成的图片，大都采用了上采样的操作，也大都具有功率谱特征扰动的问题，可以用我们的模型进行鉴别。

参考文献

[1] Durall, R., Keuper, M., Pfreundt, F. J., & Keuper, J. (2019). Unmasking deepfakes with simple features.
[2] Durall, R., Keuper, M., & Keuper, J. (2020). Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions.