物理学报 Acta Physica Sinica
Chinese Physical Society
Institute of Physics, CAS

基于深度学习的相位截断傅里叶变换非对称加密系统攻击方法

徐昭　周昕　白星　李聪　陈洁　倪洋

Attacking asymmetric cryptosystem based on phase truncated Fourier fransform by deep learning

Xu Zhao　Zhou Xin　Bai Xing　Li Cong　Chen Jie　Ni Yang

引用信息 Citation: Acta Physica Sinica, 70, 144202 (2021)　DOI: 10.7498/aps.70.20202075
在线阅读 View online: https://doi.org/10.7498/aps.70.20202075
当期内容 View table of contents: http://wulixb.iphy.ac.cn

基于深度学习的相位截断傅里叶变换非对称加密系统攻击方法*

徐昭 周昕† 白星 李聪 陈洁 倪洋

(四川大学电子信息学院, 成都 610065)

(2020 年 12 月 8 日收到; 2021 年 2 月 24 日收到修改稿)

大多数光学加密系统都是对称加密系统, 在光学图像加密中明文和密文之间具有线性关系, 其系统的安全性有待加强. 而基于相位截断傅里叶变换 (phase-truncated Fourier transform, PTFT) 的非对称加密系统, 其非线性的相位截断操作使加密系统的安全性得到了极大提升. 本文提出使用深度学习方法攻击 PTFT 加密系统, 通过 PTFT 加密系统构造出明密文对图像数据集, 然后将其输入残差网络 (residual network, ResNet) 中进行训练, ResNet 自动学习该加密系统的解密特性. 最后应用测试集对训练好的模型进行解密性能测试, 数据表明该模型能够较好地恢复图像并且该模型具有一定的抗噪声能力. 与两步迭代振幅恢复算法相比, 本文所提出方法恢复的图像质量更好.

关键词: 光学加密, 相位截断傅里叶变换, 深度学习, 残差网络

PACS: 42.30.Va, 42.30.Wb, 07.05.Mh **DOI:** 10.7498/aps.70.20202075

1 引 言

在互联网高速发展的时代, 信息安全的重要性不言而喻. 在多种多样的传递信息方式中, 图像是一种重要的载体, 因此保证图像传输的安全得到广泛关注. 数字图像加密是保证图像安全的一种重要技术, 这种技术通常使用计算机等电子手段实现. 但是往往受到速度和成本的限制, 基于传统的计算机加密技术需要耗费较长的时间. 随着光信息技术的发展, 光学图像加密技术也取得了长足的进步. 光学图像加密技术属于并行加密, 它通过干涉、衍射、傅里叶变换等操作使明文图像变成类似噪声的密文图像, 以光速对图像的所有像素进行加密, 速度快效率高. 光的波长、相位、振幅等都能够作为密钥, 和计算机图像加密算法相比, 光学图像加密具有更高的复杂度和加密自由度.

1995 年, Refregier 和 Javidi[1] 首次提出了基于双随机相位编码的光学图像加密算法, 该算法是通过使用两块随机相位板和傅里叶变换来实现. 不过由于该加密算法是对称加密系统, 其线性特性降低了系统的安全性, 因此 Qin 和 Peng[2] 在 2010 年提出了基于相位截断傅里叶变换 (phase-truncated Fourier transform, PTFT) 的非对称加密系统. 该系统在加密过程中通过相位截断和两块随机相位板生成密文, 而在解密过程中则用与加密密钥不同的解密密钥来恢复密文, 此解密密钥是在加密过程中通过截断处理所产生的. 由于该加密系统是非线性的, 因此系统安全性得到很大提高, 攻击难度较大.

随着人工智能技术的飞速发展以及计算机算力的提升和数据量的膨胀, 深度学习在计算机视觉、生物医学、模式识别等领域都取得了突破进展, 很多算法纷纷涌现并投入到实际应用中. 在深

* 国家自然科学基金 (批准号: 61475104, 61177009) 资助的课题.

† 通信作者. E-mail: zhoxn@21cn.com

度学习领域有许多经典的神经网络, 如 LeNet-5[3], AlexNet[4], VGGNet[5], GoogleNet[6], ResNet[7] 等, 众多学者在此基础上做出了很多重要工作. 深度学习的特点在于它能使神经网络自动分析数据和数据之间的关系, 因此完全可以利用神经网络的这一特性来研究光学图像加密中明文和密文之间的对应关系, 也就是说可以通过分析它们之间的关系来成功攻击加密系统. Hai 等[8] 就曾使用卷积神经网络 (convolutional neural networks, CNN)成功破解了双随机相位编码的光学图像加密系统, 不过正如前面所说的, 此加密系统是线性且对称的, 安全性较低, 同时破解难度也较低.

本文旨在使用深度学习的方法攻击 PTFT 加密系统, 该加密系统为非线性且非对称的系统, 破解难度较大. 通过以残差网络 (residual network, ResNet) 为基础训练大量的明密文对, 使神经网络拟合明文和密文之间的对应关系, 将神经网络训练好之后, 输入密文将可以得到恢复出的原文图像. 本文首先介绍了 PTFT 加密系统原理和 ResNet 工作原理, 然后阐述了利用 ResNet 训练图像数据的方法, 最后通过实验验证该方法的可行性, 并且和其他算法所恢复的图像质量进行比较.

2 基本原理

2.1 基于 PTFT 的非对称加密系统

在基于 PTFT 的非对称光学图像加密系统中, 两块随机相位板 R_1, R_2 用作加密密钥, 傅里叶频谱面和输出平面上的截断相位 P_1, P_2 用作解密密钥. 加密密钥是公开的, 解密密钥只有特定的人才能拥有. 图 1 为系统的加密解密原理图.

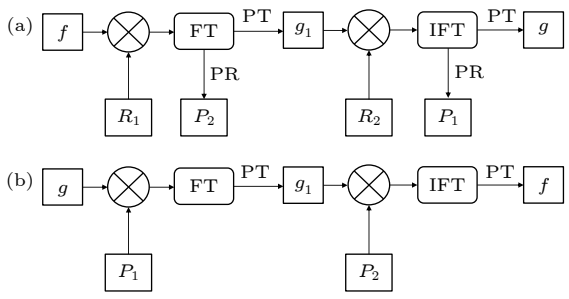

图 1 基于 PTFT 的加密系统原理图 (a) 加密过程; (b) 解密过程

Fig. 1. Schematic diagrams of PTFT system: (a) Encryption; (b) decryption.

加密时, 输入图像 $f(x,y)$ 乘以第一个随机相位板 $R_1(x,y)=\exp[\mathrm{j}2\pi\varphi(x,y)]$, 然后对其进行二维傅里叶变换, 将得到的傅里叶频谱进行相位截断, 截断后的振幅为 $g_1(\mu,\nu)$, g_1 乘以第二块随机相位板 $R_2(\mu,\nu)=\exp[\mathrm{j}2\pi\varphi(\mu,\nu)]$, 再进行一次二维傅里叶逆变换得到一个复振幅, 将复振幅进行相位截断便得到了密文图像 $g(x,y)$. 加密公式如下所示:

$$g_1(\mu,\nu)=\mathrm{PT}\{\mathrm{FT}[f(x,y)\cdot R_1(x,y)]\}, \tag{1}$$

$$g(x,y)=\mathrm{PT}\left\{\mathrm{FT}^{-1}[g_1(\mu,\nu)\cdot R_2(\mu,\nu)]\right\}, \tag{2}$$

其中 PT 表示相位截断, FT 表示傅里叶变换, FT^{-1} 表示逆傅里叶变换.

解密过程如图 1(b) 所示, 相应公式如下:

$$g_1(\mu,\nu)=\mathrm{PT}\{\mathrm{FT}[g(x,y)\cdot P_1(\mu,\nu)]\}, \tag{3}$$

$$f(x,y)=\mathrm{PT}\left\{\mathrm{FT}^{-1}[g_1(\mu,\nu)\cdot P_2(\mu,\nu)]\right\}. \tag{4}$$

通过上述公式可知, 该加密系统的加密密钥和解密密钥并不相同, 加密过程和解密过程不可逆, 为非对称加密系统. 在没有得到解密密钥的情况下无法恢复出原文图像, 因此安全性得到提升, 更容易抵抗传统攻击方法.

2.2 基于 ResNet 的图像恢复算法框架与实现

传统的 CNN 中, 随着网络层数的叠加, 特征也越来越丰富, 因此神经网络层数越深, 理论上得到的效果越卓越. 但是简单地增加网络层数会导致两个问题, 一个是在后向传播的过程中不能有效地把梯度传递到前面的网络层, 因此会出现梯度消失或者梯度爆炸的问题, 导致模型无法收敛, 影响网络的性能. 初始化和正则化虽可以保证几十层的网络能够正常收敛, 但是在更深层次的网络中, 准确率达到饱和后效果反而变差. 另一个问题是退化现象, 模型精度下降, 训练错误的频率非常高, 相应的现象在 CIFAR-10 和 ImageNet[9] 中都有出现.

因此对于这种问题, He 等[7] 在 2015 年提出了残差网络模型 (ResNet), 和传统 CNN 相比, ResNet 通过旁路支线将输入传递到后面的网络层, 这种结构称为跳转连接 (skip connection)[10], 传统的 CNN 在传递信息的时候, 会存在信息丢失的问题, ResNet 这种直接将输入与输出连接的结构使得神经网络只需要学习输入与输出差别的那部分, 学习更容易了, 这种模型缓解了梯度问题, 让网络学习到的内容更丰富. 其结构如图 2 所示.

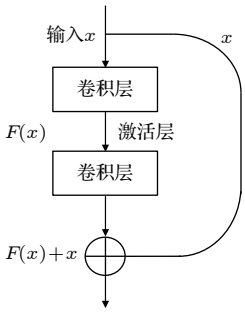

图 2 残差网络模块

Fig. 2. Residual module of ResNet.

本文提出的基于深度学习的攻击 PTFT 加密系统的方法是在不知道密钥的情况下对明文-密文对进行选择明文攻击, 因此需要大量的明文-密文对来制作训练集, 将明文-密文对放进神经网络中进行训练, 让神经网络拟合从密文到明文的过程, 从而达到解密的效果.

如图 3 所示, 我们设计的基于 ResNet 神经网络是一个端到端的结构, 图像加密的过程可以描述为:

$$I = F(O), \tag{5}$$

其中 O 是明文图像, I 是密文图像, 尽管该加密系统不是一个对称加密系统, 不能简单地进行逆变换进行恢复, 但是神经网络的训练过程可以看作是加密的逆过程, 表示为:

$$P = F^{-1}(I), \tag{6}$$

P 代表经过神经网络恢复出来的图像.

如图 3 所示, 尺寸为 64×64 的灰度图像先进入卷积层, 卷积层的卷积核的尺寸为 3×3, 步长为 1×1, 用于对输入图像进行特征提取. 此后经过卷积运算的图像数据经过激活层, 激活函数为 ReLU[11,12], 如下所示:

$$\mathrm{ReLU}(x) = \begin{cases} x, & x > 0, \\ 0, & x \leqslant 0, \end{cases} \tag{7}$$

x 代表神经网络中上一层的输出, 接下来经过残差模块, 残差模块中包括卷积层、批量归一化 (batch normalization, BN) 层[13] 和激活层, BN 层可以加快收敛速度, 提高模型准确性并控制过拟合. 卷积核保持和之前的卷积层一致并输出 64 张特征图, 图 3 中设置了 5 个相同的残差模块, 经过 5 个残差模块后数据又依次被输入到相应的层中, 最后, 我们可以从神经网络中获得 $64\times64\times1$ 的图像.

为了更好地恢复密文图像, 本文使用均方差 (mean squared error, MSE) 损失函数[14,15], 其定义如下:

$$\mathrm{MSE} = \frac{1}{M\times N}\sum_{i=1}^{M\times N}(x_i - y_i)^2, \tag{8}$$

其中 x_i 和 y_i 分别代表最后一层的输出值和原始图像的真值, M 和 N 分别代表图像的宽度和高度, 最后我们期望在训练的过程中最小化损失函数, 当损失函数不再下降时, 训练完成, 模型具有恢复原文图像的能力.

3 实验结果与分析

本次实验的硬件平台采用的 CPU 是 Intel 至强 E5-2630V4, 内存为 16 G, GPU 为 NVIDIA GeForce GTX1080Ti, 编译环境为 Python 3.6, 深度学习平台为 TensorFlow 1.12. 数据集基于 MNIST[16], 将所有 MNIST 数字图像进行 PTFT 加密得到密文, 由明密文对组成的构成所有的数据集, 训练集中有 10000 对明密文对, 验证集和测试集中各有 1000 对, 每张图像的大小为 64×64. 其中验证集的目的是调整网络的超参数, 而测试集的目的则是验证神经网络的泛化能力.

实验中学习率设置为 0.0001, 优化器采用的是 Adam[17], batch size 设置为 16, 实验结果如图 4

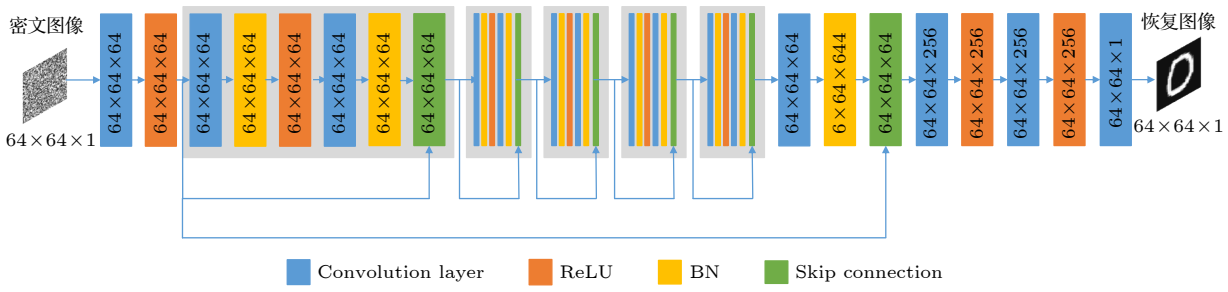

图 3 基于 ResNet 的网络架构

Fig. 3. Neural network based on ResNet.

所示, 图 4(a) 是明文图像, 图 4(b) 是密文图像, 图 4(c) 是通过神经网络恢复的明文图像.

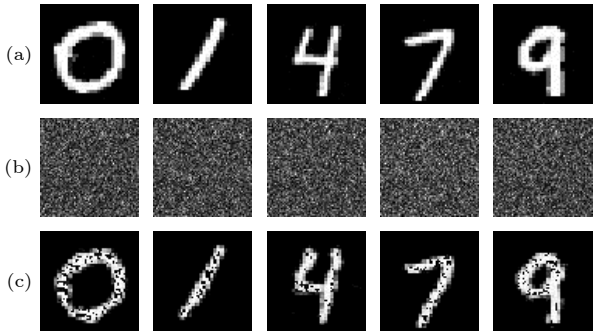

图 4 神经网络重建效果图 (a) 明文图像; (b) 密文图像; (c) 通过神经网络恢复的明文图像

Fig. 4. Images reconstructed by neural network: (a) Plaintext; (b) ciphertext; (c) plaintext reconstructed by neural network.

为了测试模型的鲁棒性, 可以将高斯噪声添加进密文中, 然后检测其重建效果. 本文使用与密文能量比分别为 10%, 20% 和 50% 的高斯噪声得到被污染的密文, 并将其放入测试集中进行测试, 得到图 5 所示结果. 图 5(a) 是没有添加噪声的结果, 图 5(b) 是添加 10% 高斯噪声的结果, 图 5(c) 是添加 20% 高斯噪声的结果, 图 5(d) 是添加 50% 高斯噪声的结果. 可以看到, 即使密文中带有噪声, 但是恢复出来的图像仍然可以清晰地看到轮廓, 这表明本文提出的模型具有很强的抗噪声能力. 但当密文中带有噪声比例较高如 50% 时恢复出来的图像失真严重, 已难以辨认, 可以认为该神经网络对噪声的鲁棒性范围在 50% 以下.

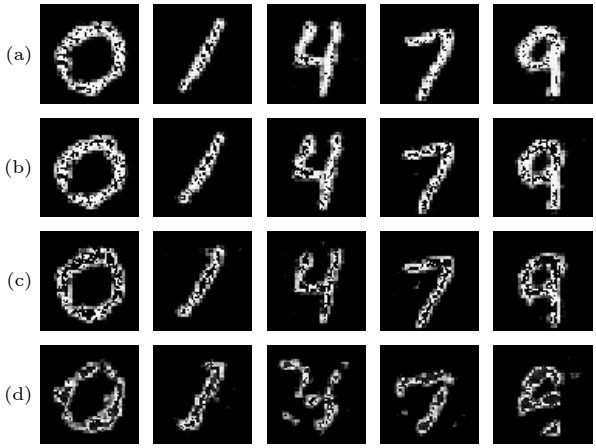

图 5 含有不同能量比高斯噪声的密文解密效果 (a) 0%; (b) 10%; (c) 20%; (d) 50%

Fig. 5. Reconstruction results of ciphertext containing Gaussian noise with different energy ratios: (a) 0%; (b) 10%; (c) 20%; (d) 50%.

此后, 针对密文含有 20% 能量比高斯噪声的情况, 我们尝试让训练集中的密文也被与密文能量比为 20% 的高斯噪声污染, 重新制作数据集并重新训练, 再用训练好的网络去处理含有 20% 能量比高斯噪声的测试集中的密文, 结果如图 6(b) 所示.

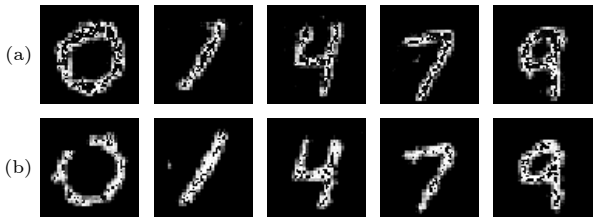

图 6 使用含不同能量比高斯噪声的密文训练集后的测试效果 (a) 0%; (b) 20%

Fig. 6. Test results after using ciphertext groups of Gaussian noise with different energy ratios: (a) 0%; (b) 20%.

作为对比, 图 6(a) 是训练集中密文没有被高斯噪声污染情况下训练好的网络, 对含有 20% 能量比高斯噪声的测试集密文重建结果, 可以看到经过噪声训练的神经网络在抗噪声方面表现更佳. 因此在实际中, 如果待处理的密文含有噪声, 可以考虑通过让训练集中的密文也遭受大致相当的噪声污染的方法, 以提高网络的训练效果.

由于在实际应用中, 密文受到噪声污染的比例通常是无法确定的, 因此针对训练集和测试集噪声比例不同做出的测试如图 7 所示 (测试集噪声比例均为 30%).

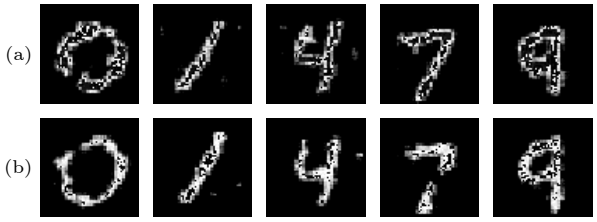

图 7 使用含不同能量比高斯噪声的密文训练集后的测试效果 (a) 0%; (b) 20%

Fig. 7. Test results after using ciphertext groups of Gaussian noise with different energy ratios: (a) 0%; (b) 20%.

可以看到, 即使训练集和测试集的噪声比例不同, 经过噪声训练的神经网络在恢复被噪声污染的密文时仍然效果更好.

为了评价图像质量, 我们使用峰值信噪比 (peak signal to noise ratio, PSNR) 和结构相似性 (structural similarity index, SSIM) 作为参考[18]:

$$\mathrm{PSNR} = 10 \times \log_{10} \frac{255^2}{\mathrm{MSE}}, \tag{9}$$

$$\mathrm{SSIM}\,(X,Y) = \frac{(2\mu_X\mu_Y + C_1)\,(2\sigma_{XY} + C_2)}{(\mu_X^2 + \mu_Y^2 + C_1)\,(\sigma_X^2 + \sigma_Y^2 + C_2)}, \tag{10}$$

$$C_1 = (k_1 L)^2, \quad C_2 = (k_2 L)^2, \tag{11}$$

其中 $X(i,j)$ 和 $Y(i,j)$ 分别代表原始图像和恢复图像的像素值, μ_X 和 μ_Y 分别是 $X(i,j)$ 和 $Y(i,j)$ 的均值, σ_X 和 σ_Y 分别是 $X(i,j)$ 和 $Y(i,j)$ 的标准差, σ_{XY} 是 $X(i,j)$ 和 $Y(i,j)$ 的协方差, C_1 和 C_2 用来保持稳定性, 使 (10) 式的分母任意时刻都不为 0, $k_1 = 0.01$, $k_2 = 0.03$ [19], L 是像素值的动态范围, 对于实验中的图像, $L = 255$.

图 8 分别是图 6(a), (b) 的 PSNR 和 SSIM, 可以从数据上得知, 加入噪声训练过的神经网络, 在恢复有噪声的密文图像时的质量确实比没有经过噪声训练的神经网络恢复的图像质量好.

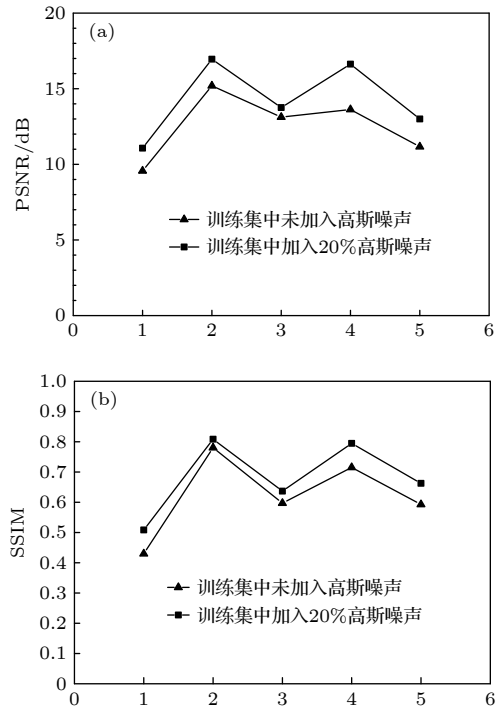

图 8 图 6 中图像的 PSNR 和 SSIM (a) PSNR; (b) SSIM

Fig. 8. PSNR and SSIM in Fig.6: (a) PSNR; (b) SSIM.

除了深度学习能够恢复该加密系统加密后的图像, 此前 Wang 和 Zhao[20] 曾提出使用两步迭代振幅恢复算法来恢复加密图像. 该算法建立在密文和加密密钥公开的前提下, 可以简单分为两步进行, 第一步求 PTFT 加密系统中傅里叶平面上振幅分布的近似值, 给输出平面上的相位分布即解密密钥赋一个任意初始值, 根据密文求出迭代第一次的傅里叶平面上的振幅分布值, 由此值求出密文, 与真正的密文的均方差表示迭代算法的收敛性. 第二步与第一步类似, 由第一步求出傅里叶平面振幅分布的近似值后, 根据此值和空域的随机相位板迭代出明文图像, 迭代出来的明文和原始明文的均方差表示迭代算法的收敛性. 根据此算法迭代 10 万次得出图 9(b) 结果.

从图 9 中可以直观地看出, 由本文提出的深度学习算法恢复出来的图像质量好于两步迭代振幅恢复算法恢复的图像质量. 而从图 10 中可以得知深度学习算法恢复的图像的 PSNR 高于两步迭

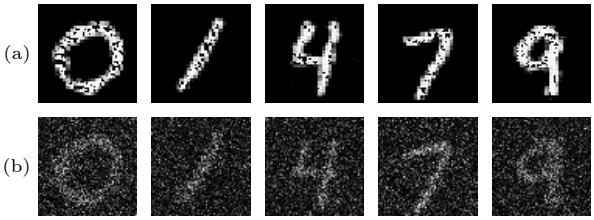

图 9 (a) 深度学习算法恢复结果; (b) 两步迭代振幅恢复算法恢复结果

Fig. 9. (a) Reconstruction results by deep learning; (b) reconstruction results by two-step iterative amplitude retrieval approach.

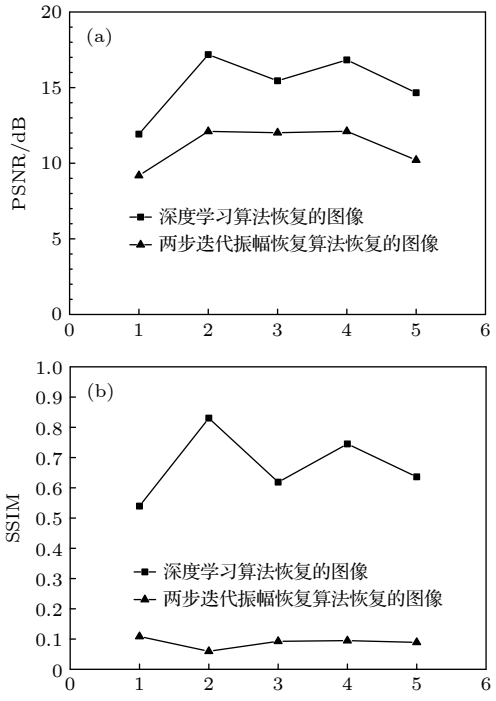

图 10 图 9 中图像的 PSNR 和 SSIM (a) PSNR; (b) SSIM

Fig. 10. PSNR and SSIM in Fig. 9: (a) PSNR; (b) SSIM.

代法恢复图像的 PSNR, 深度学习算法恢复的图像的 SSIM 均高于 0.5, 而两步迭代法恢复图像的 SSIM 均不高于 0.1.

4 结 论

本文提出使用深度学习的方法攻击基于 PTFT 的非对称加密系统, 由于该加密系统在双随机加密系统的基础上进行相位截断, 因此该系统是非线性的加密系统, 攻击难度大. 我们从深度学习角度对 PTFT 系统进行研究, 使用基于 ResNet 的神经网络训练明密文对数据集, 该网络自动学习密文到明文的拟合过程. 实验结果表明, 通过深度学习的方法能够对密文图像进行有效的恢复, 并且该网络具有一定的抗噪声能力. 与两步迭代振幅恢复算法相比, 本文提出的方法恢复的图像质量也远好于它. 在下一步的研究工作中将尝试提高恢复图像的质量, 并对更复杂的加密系统进行攻击研究.

参考文献

[1] Refregier P, Javidi B 1995 *Opt. Lett.* **20** 767
[2] Qin W, Peng X 2010 *Opt. Lett.* **35** 118
[3] Lecun Y, Bottou L, Bengio Y, Haffner P 1998 *Proc. IEEE.* **86** 2278
[4] Krizhevsky A, Sutskever I, Hinton G 2017 *Commun. ACM.* **60** 84
[5] Simonyan K, Zisserman A 2014 *arXiv e-prints* arXiv: 1409.1556
[6] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A 2015 *2015 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*) Boston, USA, June 7–12, 2015 p1
[7] He K, Zhang X, Ren S, Sun J 2015 *arXiv e-prints* arXiv: 1512.03385
[8] Hai H, Pan S, Liao M, Lu D, He W, Peng X 2019 *Opt. Express* **27** 21204
[9] Srivastava R, Greff K, Schmidhuber J 2015 *Proceedings of the 28th International Conference on Neural Information Processing Systems* Montreal, Canada, December 7–10, 2015 p2377
[10] Drozdzal M, Vorontsov E, Chartrand G, Kadoury S, Pal C 2016 *arXiv e-prints* arXiv: 1608.04117
[11] Glorot X, Bordes A, Bengio Y 2011 *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics* (*AISTATS*) Fort Lauderdale, USA, April 11–13, 2011 p315
[12] Nair V, Hinton G 2010 *Proceedings of the 27th International Conference on International Conference on Machine Learning* Madison, USA, June 21–24, 2010 p807
[13] Ioffe S, Szegedy C 2015 *Proceedings of the 32nd International Conference on International Conference on Machine Learning* Lille, France, July 6–11, 2015 p448
[14] Dong C, Loy C C, He K, Tang X 2014 *Proceedings of the 13th European Conference on Computer Vision* Zurich, Switzerland, September 6–12, 2014 p184
[15] Ishikawa M 1996 *Neural Networks.* **9** 509
[16] Ciregan D, Meier U, Schmidhuber J 2012 *2012 IEEE Conference on Computer Vision and Pattern Recognition* Providence, USA, June 16–21, 2012 p3642
[17] Kingma D P, Ba J 2014 *arXiv e-prints* arXiv: 1412.6980
[18] Horé A, Ziou D 2010 *20th International Conference on Pattern Recognition* Istanbul, Turkey, Auguest 23–26, 2010 p2366
[19] Wang Z, Bovik A C, Sheikh H R, Simoncelli E P 2004 *IEEE Trans. Image Process.* **13** 600
[20] Wang X, Zhao D 2012 *Opt. Commun.* **285** 1078

Attacking asymmetric cryptosystem based on phase truncated Fourier fransform by deep learning*

Xu Zhao Zhou Xin † Bai Xing Li Cong Chen Jie Ni Yang

(*College of Electronics and Information Engineering, Sichuan University, Chengdu* 610065, *China*)

(Received 8 December 2020; revised manuscript received 24 February 2021)

Abstract

Most of optical encryption systems are symmetric cryptosystems. The plaintext and the ciphertext in optical image encryption are related linearly. The security of the system needs to be strengthened. The asymmetric cryptosystem based on phase truncated Fourier transforms (PTFT) makes the security of the encryption system greatly improved by its nonlinear phase truncation. Deep learning (DL) as a method of machine learning was proposed decades ago. With the development of computer's performance, the practicality of deep learning proves to be more and more obvious. Recently, deep learning has been effectively used in many fields such as biomedicine, object detection, etc. The good results have been achieved. In this article proposed is the attack to the PTFT encryption system by deep learning. Through the PTFT encryption system, we construct a plaintext-ciphertext paired image dataset and then train it by residual network (ResNet). There are two problems encountered by the traditional neural network model. One is vanishing or named exploding gradient, which makes training effect difficult to converge and the other is a degradation phenomenon. When continuing to increase the number of layers for a suitable depth model, the model accuracy will decline which is not caused by overfitting. This problem can be solved by the ResNet to a certain extent by directly bypassing and then taking the input information to the output to protect the integrity of the information. The biggest difference between ordinary directly connected convolutional neural networks and ResNet is that the ResNet has many bypass branches that directly connect the input to the subsequent layers, so that the subsequent layers can directly learn the residuals. The ResNet can automatically learn the decryption characteristics of the encryption system. Finally, the test set is used to test the decryption performance of the trained model. The data show that the model can restore the image with high quality and the model has a certain anti-noise ability. Compared with the two-step iterative amplitude recovery algorithm, the the method proposed in this paper can recover high quality image.

Keywords: optical encryption, phase truncated Fourier transforms, deep learning, residual network

PACS: 42.30.Va, 42.30.Wb, 07.05.Mh **DOI:** 10.7498/aps.70.20202075

* Project supported by the National Natural Science Foundation of China (Grant Nos. 61475104, 61177009).

† Corresponding author. E-mail: zhoxn@21cn.com