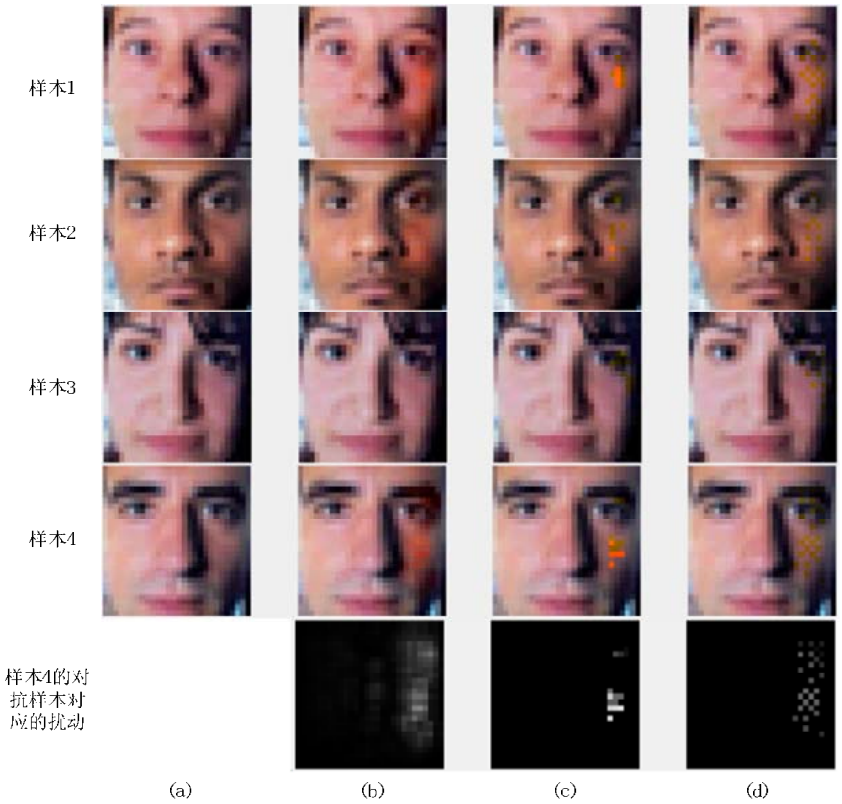
图片为一个原图以及生成的 11 个对抗样本图,共 3 600 张图片.并且评价是在不同的显示设备上完成的,放大尺寸由志愿者自行调整,每张图片的观察时间控制在 3s 左右.图 6(c)表示人眼感知率和扰动幅值之间的关系,由图可知,当μ=90 时,人眼感知率最低,为 30%;μ值增大或者缩小,都会导致人眼感知率增加.分析其原因,人眼可感知性受扰动维度数和扰动幅值两个因素共同影响,而μ=90 的设置使得两者进行了折中,较不易被感知.为了验证该评价方法的合理性和有效性,在测试图片中加入了原图,其人眼感知率低于 5%,即原图很少被误认为人工修改过,证明该主观评价方法是合理而有效的.综合对抗成功率、扰动像素数以及人眼感知率等数据,选择μ=90 为最佳扰动幅度,并作为后续实验及结果比较的参数.

4.3 和相关工作的结果比较

为了进一步与相关工作比较,分别利用 DeepFool 方法和本文的算法生成测试集图像的对抗样本,如图 7 所示.其中,图 7(a)所示为原始的假体图片,且网络能够成功地检测为假,实验对其做扰动,生成对抗样本以欺骗分类器网络判断为活体;图 7(b)所示为 DeepFool 方法生成的对抗样本;图 7(c)所示为本文算法未加扰动间距约束时生成的对抗样本;图 7(d)所示为加入扰动间距约束的算法生成的对抗样本.从图 7 中可以看出,加入扰动间距约束的算法生成结果在视觉上更接近原始图像,更不易被人眼感知.图 7 第 5 排为第 4 排中各对抗样本对应的扰动,本文算法生成的扰动避免了同一像素中 RGB 分量被同时扰动,且被扰动的像素较分散,互相不连续,有较好的视觉效果.利用主观评价方法得到 DeepFool 和 FGS 算法生成结果的人眼感知率分别为 50%和 51%,本文方法的人眼感知率比 Deepfool 和 FGS 方法降低了 20%和 21%(注:图 7 所列例子均为测试集中随机选取的结果,并非精心挑选).



(a) 原图 (b) DeepFool 方法 (c) 本文算法(未加扰动间距约束) (d) 本文算法(加入扰动间距约束)

Fig.7 Comparisons of adversarial examples from different algorithms

图 7 不同算法生成的对抗样本比较

表 2 分析了不同算法所需的扰动维度和平均扰动幅度等.平均扰动幅度的计算方法如公式(4)所示,即扰动向量的 L2 范数与原始数据 L2 范数的比值.由表 2 可知,本文算法未加扰动间距约束时,平均扰动 30.3 个像素点,占原始输入维度的 1.29%,即只需改动原始输入向量的 1.29%,即可成功地欺骗网络;加入间距约束后,算法平均扰动像素点为 31.9,占原始输入维度的 1.36%,略微有增加,但视觉效果明显提升.

Table 2 Comparison with other related methods

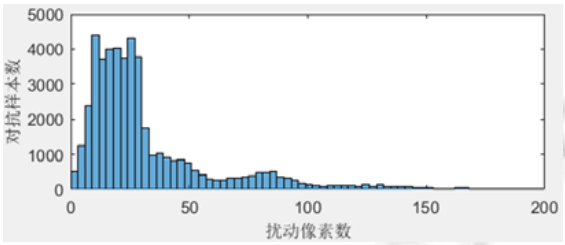
表 2 与其他相关方法的比较

方法	DeepFool	本文算法未加扰动间距约束	本文算法加入扰动间距约束
扰动维度数	2 351	30.3	31.9
扰动维度比例(%)	99.99	1.29	1.36
平均扰动幅度	0.0382	0.140 7	0.142 5
网络间的泛化性(%)	87.63	78.44	85.75

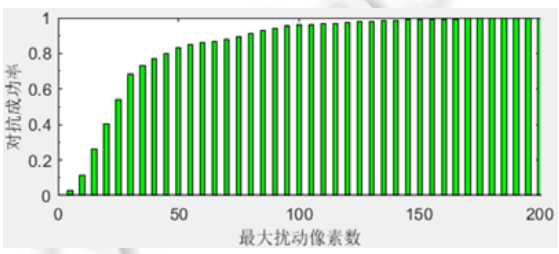
为了验证本文算法生成结果的泛化性,用相同的数据库训练另一个网络 LeNet-5[17],即卷积神经网络中具有代表性的一个结构.LeNet-5 网络的活体检测正确率为 98.52%.利用表 1 所示网络生成的对抗样本集针对 LeNet-5 网络的欺骗成功率为 85.75%,即表 1 所示网络生成的对抗样本中有 85.75%能够欺骗 LeNet-5,证明本文算法生成的对抗样本在不同网络之间具有较好的泛化性,与 DeepFool 方法 87.63%的泛化性相当.

文章《DeepFool: A simple and accurate method to fool deep neural networks》指出:网络层数越高,分类性能就越好,对于对抗样本的鲁棒性也就越好.加入一些技巧,如 Batch normalization 和 dropout 之后,可以在一定程度上提高模型鲁棒性,但对抗样本问题仍然存在.本文在表 1 所示网络中全连接层上加入 dropout,活体检测的等错误率 HTER 为 2.12%,与未加 dropout 时相当;利用本文方法重新生成对抗样本集,其平均扰动幅度为 0.122,即与未加 dropout 时相比,鲁棒性没有明显的提高.Dropout 的主要作用在训练阶段加速收敛和防止过拟合,但对于对抗样本不鲁棒.

图 8(a)示出了对 48 451 个测试图片利用算法 2 生成对抗样本所需扰动的维度数直方图,扰动维度集中在[0,50]左右.有少量样本的扰动维度较大.通过限制最大扰动维度数,分析对抗样本生成的成功率如图 8(b)所示,即若扰动维度数大于某一阈值,则停止迭代,测试该时刻生成样本是否能够成功地欺骗.由分析结果可知,当最大维度设置为 100(输入维度的 4.25%)时,生成对抗样本的成功率为 97%.



(a) 测试集对抗样本扰动维度直方图



(b) 扰动维数限制和对抗成功率关系

Fig.8 Analysis of average perturbation dimensions in adversarial examples

图 8 对抗样本平均扰动维度分析

5 结 论

深度学习技术易受对抗样本的攻击,而人脸活体检测任务的对抗样本具有其特殊性:活体检测任务真假体图像相近,类间距离较小,且假体存在于两次成像之间,对其做扰动有一定的局限性.针对以上特点以及人眼对图片的视觉特性,提出了一种基于人眼视觉特性的最小扰动维度对抗样本生成方法.该方法将对输入图像的扰动集中在少数几个维度上,并充分考虑人眼的视觉连带集中特性,加入扰动点的间距约束,以使最后生成的对抗样本更不易被人类察觉.利用人脸活体检测数据库Print Attack对典型CNN分类模型进行对抗,该方法通过平均扰动输入总维度的 1.36%,即可成功地生成对抗样本.并且通过加入扰动间距约束,使对抗样本更加不易被人眼感知,通过志愿者对于对抗样本的主观评价,其人眼感知率比经典 FGS 方法及 DeepFool 方法降低了 20%,证明该方法有更好的欺骗效果.

本文研究了人脸活体检测任务中对抗样本的生成机理,揭示了活体检测任务分类器的安全隐患,为下一步建立合理的防范机制奠定了基础.

References:

[1] Schroff F, Kalenichenko D, Facenet PJ. A unified embedding for face recognition and clustering. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. IEEE Computer Society, 2015. 815−823.

[2] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. Curran Associates Inc., 2012. 1097−1105.

[3] Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In: Proc. of the Int'l Conf. on Representation Learning, 2015. 1−13.

[4] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. Intriguing properties of neural networks. In: Proc. of the Int'l Conf. on Representation Learning. 2014. 1−10.

[5] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proc. of the Int'l Conf. on Representation Learning, 2015. 1−10.

[6] Meier U, Masci J. Multi-column deep neural network for traffic sign classification. Neural Networks the Official Journal of the Int'l Neural Network Society, 2012,32(1):333−338.

[7] Xu X. Research on deep learning based face liveness detection algorithm [MS. Thesis]. Beijing: Beijing University of Technology, 2016 (in Chinese with English abstract).

[8] Lucena O, Junior A, Moia V, Souza R, Valle E, Lotufo R. Transfer learning using convolutional neural networks for face anti-spoofing. In: Proc. of the Int'l Conf. on Image Analysis and Recognition. Springer-Verlag, 2017. 27−34.

[9] Xu Y, Jian M, Xu X, Qi W. Face liveness detection scheme with static and dynamic features. Int'l Journal of Wavelets Multiresolution & Information Processing, 2018,16(2):No.1840001.

[10] Yang J, Lei Z, Li SZ. Learn convolutional neural network for face anti-spoofing. Computer Science, 2014,9218:373−384.

[11] Gonzalez RC, Woods RE. Digital Image Processing. Beijing: Publishing House of Electronics Industry, 2006.

[12] Moosavi-Dezfooli SM, Fawzi A, Frossard P. Deepfool: A simple and accurate method to fool deep neural networks. In: Proc. of the Computer Vision and Pattern Recognition. IEEE, 2016. 2574−2582.

[13] Yin BC, Wang WT, Wang LC. Review of deep learning. Journal of Beijing University of Technology, 2015,41(1):48−59 (in Chinese with English abstract).

[14] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature, 1986,323(6088):533−536.

[15] Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A. The limitations of deep learning in adversarial settings. In: Proc. of the IEEE European Symp. on Security and Privacy. IEEE, 2016. 372−387.

[16] Anjos A, Marcel S. Counter-measures to photo attacks in face recognition: A public database and a baseline. In: Proc. of the Int'l Joint Conf. on Biometrics. IEEE, 2011. 1−7.

[17] Lécun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc. of the IEEE, 1998,86(11): 2278−2324.

附中文参考文献:

[7] 许晓.基于深度学习的活体人脸检测算法研究[硕士学位论文].北京:北京工业大学,2016.

[13] 尹宝才,王文通,王立春.深度学习研究综述.北京工业大学学报,2015,41(1):48–59.

马玉琨(1983－),女,河南新乡人,博士,CCF 专业会员,主要研究领域为数字图像处理.

毋立芳(1970－),女,博士,教授,博士生导师,CCF 专业会员,主要研究领域为数字图像处理,模式 识别.

简萌(1987－),女,博士,讲师,CCF 专业会员,主要研究领域为模式识别,多媒体计算.

刘方昊(1993－),男,硕士,主要研究领域为优化理论.

杨洲(1994－),男,硕士生,主要研究领域为计算机视觉.