E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2023)03-0850-14

论文引用格式:Zhao J J, Wang J W and Wu J F. 2023. Adversarial attack method identification model based on multi-factor compression error. Journal of Image and Graphics,28(03):0850-0863(赵俊杰,王金伟,吴俊凤. 2023. 基于多质量因子压缩误差的对抗样本攻击方法识别. 中国图象图形学报,28(03):0850-0863)[DOI:10.11834/jig.220516]

基于多质量因子压缩误差的对抗样本攻击方法识别

赵俊杰[1],王金伟[2,3*],吴俊凤[2]

1. 南京信息工程大学电子与信息工程学院,南京　210044;2. 南京信息工程大学计算机学院,南京　210044;
3. 数字取证教育部工程研究中心,南京　210044

摘　要:**目的**　对抗样本严重干扰了深度神经网络的正常工作。现有的对抗样本检测方案虽然能准确区分正常样本与对抗样本,但是无法判断具体的对抗攻击方法。对此,提出一种基于多质量因子压缩误差的对抗样本攻击方法识别方案,利用对抗噪声对JPEG压缩的敏感性实现攻击方法的识别。**方法**　首先使用卷积层模拟JPEG压缩、解压缩过程中的颜色转换和空频域变换,实现JPEG误差在图形处理器(graphic processing unit,GPU)上的并行提取。提出多因子误差注意力机制,在计算多个质量因子压缩误差的同时,依据样本差异自适应调整各质量因子误差分支的权重。以特征统计层为基础提出注意力特征统计层。多因子误差分支的输出经融合卷积后,获取卷积层多维特征的同时计算特征权重,从而形成高并行对抗攻击方法识别模型。**结果**　本文以ImageNet图像分类数据集为基础,使用8种攻击方法生成了15个子数据集,攻击方法识别率在91%以上;在快速梯度符号法(fast gradient sign method,FGSM)和基本迭代法(basic iterative method,BIM)数据集上,噪声强度识别成功率超过96%;在对抗样本检测任务中,检测准确率达到96%。**结论**　所提出的多因子误差注意力模型综合利用了对抗噪声的分布差异及其对JPEG压缩的敏感性,不仅取得了优异的对抗攻击方法识别效果,而且对于对抗噪声强度识别、对抗样本检测等任务有着优越表现。

关键词:图像处理;卷积神经网络(CNN);对抗样本;图像分类;压缩误差

Adversarial attack method identification model based on multi-factor compression error

Zhao Junjie[1], Wang Jinwei[2,3*], Wu Junfeng[2]

1. School of Electronics and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China;
2. School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China;
3. Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing 210044, China

Abstract: **Objective** Artificial intelligence (AI) technique based deep neural networks (DNNs) have facilitated image classification and human-facial recognition intensively. However, recent studies have shown that DNNs is vulunerable to small changes for input images. However, DNN-misclassified is caused derived of injecting small adversarial noise into the

收稿日期:2022-05-30;**修回日期**:2022-09-16;**预印本日期**:2022-09-23

*通信作者:王金伟　wjwei_2004@163.com

基金项目:国家自然科学基金项目(62072250)

Supported by: National Natural Science Foundation of China (62072250)

input sample, such an artificially designed anomalous example, called an adversarial example. Recent detection of adversarial examples can be used to get higher accuracy. But, to determine the level of deep neural network security and implement targeted defense strategies, the classification of attack methods is required to be developed further. The adversarial examples mainly consist two categories: 1) white-box and 2) black-box attacks. A white-box attack is oriented for all information about the target neural network-prior. The attacker can obtain information about the gradient of the loss function of the example and query the output of the target neural network and other information. A black-box attack concerns that the attacker can query the input and output information of the target neural network only. The white-box attack method is mainly implemented by querying the gradient of the network. Black-box attacks are mainly divided into two approaches: 1) bounded query and 2) gradient estimation. It is still challenged for the adversarial attack method used by attackers although existing adversarial example detection schemes can distinguish adversarial examples from natural ones accurately. JPEG compression is a commonly used lossy compression method for images processing. Its compression and decompression process can be linked to errors in relevant to truncation, rounding, color space conversion, as well as quantization. To deal with the heterogeneity for compression, the quantization step uses different quantization tables and a large variation is produced in the magnitude of the error. **Method** To classify adversarial examples' generation methods, we develop a multi-factor error attention model. To classify examples from multiple attack methods, the JPEG errors are injected into a neural network. To achieve parallel extraction of JPEG errors on graphic processing unit (GPU), JPEG compression and decompression processes are simulated in terms of DNN components. Multiple error branches are employed to alleviate multiple attempts of quality factors. A multi-factor error attention mechanism is proposed, which can balance the multisample-differentiated weights of each quality factor error branch. The feature statistical layer is used to calculate the high-dimensional feature vectors of the feature map. An attention mechanism is added to the feature statistical layer, and a attention-based feature statistical layer is proposed. The attention mechanism is beneficial for the feature values to adaptively modify the ratio between them. The peak-convolutional layer-derived feature map output is fed to the attention-based feature statistical layer for each channel. To optimize an efficient model for classifying adversarial examples' generation methods, the output of the multi-factor error branches is fused and sent into convolutional layers, then input into the attention-related feature statistical layer. **Result** We develop 15 ImageNet image classification dataset-based sub-datasets in terms of 8 popular attack methods. The fast gradient sign method (FGSM) and basic iterative method (BIM)-generated adversarial examples are composed of 4 sub-datasets of perturbation coefficients of 2, 4, 6, and 8. The Bandits-based adversarial examples are organized by two sub-datasets of versions L_2 and L_∞. Each sub-dataset is involved of 10 000 training examples and 2 000 test examples. The overall dataset consists of 15 sub-datasets, the attack method recognition rate is above 91%. The accuracy of noise intensity detection is above 96% on the FGSM and BIM datasets. In the adversarial sample detection task, the detection accuracy reaches 96%. The experiments show that the multi-factor feature attention network can not only classify adversarial attack methods in high accuracy, but also has its potentials for noise intensity recognition and adversarial examples' detection tasks. The comparative analysis demonstrate that our model proposed is not significantly degraded from existing schemes for the adversarial example detection task. **Conclusion** A multi-factor error attention model is developed for adversarial example classification. Our initial is dominated to the JPEG errors-aided for adversarial sample detection. The proposed model can simplify the extraction of JPEG compression-decompression errors and puts them on the GPU for parallel extraction. The error branch attention mechanism can be used to balance the weights adaptively between the error branches. The attention-linked feature statistical layer enriches the feature types and balances them adaptively.

Key words: image processing; convolutional neural network (CNN); adversarial example; image classification; compression error

0　引　言

深度神经网络给图像分类(Gao 等,2021)、人脸识别(胡蓝青 等,2022)以及风格转换(Chen 等,2021)等领域带来了巨大的技术变革和性能提升。然而研究表明,深度神经网络对于输入样本的微小变化非常敏感(Goodfellow 等,2015)。在原始样本上添加微小的对抗噪声,就可以引起深度神经网络的误判(邹军华 等,2022)。这种人为设计的异常样

本称为对抗样本。目前,对抗样本的检测已经可以达到较高的准确率(Wang 等,2022)。然而,为判断已部署的深度神经网络安全现状及采取针对性防御策略,还需要进一步识别具体的攻击方法。

对抗样本的生成方式主要包含白盒攻击和黑盒攻击两大类(Akhtar 和 Mian,2018)。白盒攻击是指目标神经网络的全部信息已知,攻击者可以获取损失函数关于样本的梯度,并查询目标神经网络的输出等信息。Goodfellow 等人(2015)发现,将逆梯度符号乘以一个自定义系数后,直接添加到原始样本上就可以快速生成对抗样本,从而提出了快速梯度攻击法。Kurakin 等人(2017)将攻击过程分步骤进行,每一步重新查询梯度信息,有效提升了攻击成功率。Carlini 和 Wagner(2017)使用 L_0、L_2、L_∞ 等 3 种范数约束对抗样本与原始样本的空间距离,限制了对抗扰动的幅度,提出了 C&W(Carlini and Wagner)攻击。然而,C&W 攻击通常需要上千次的查询迭代。DDN(decoupling direction and norm)攻击(Rony 等,2019)将对抗扰动的方向和范数进行解耦,以较小的 L_2 范数显著减少了查询次数。黑盒攻击是指攻击者仅能查询到目标神经网络的输入输出信息,而不知道其内部参数。目前,黑盒攻击主要有基于决策的攻击和基于梯度的攻击两种方案。Brendel 等人(2018)首先提出了基于决策的边界攻击方法,使蒸馏防御(Papernot 等,2016)方案失效。Ilyas 等人(2019)使用最小二乘法进行梯度估计,大幅提升了黑盒攻击的查询效率。

常见的对抗样本检测工作主要利用对抗样本与正常样本在统计分布上的差异实现被动式检测。由于深度神经网络对于梯度扰动的敏感性,非常小的对抗噪声就可以引起模型的误判。因此,对抗样本与自然样本之间的差异非常微小,需要针对性设计专用的检测器。隐写分析任务同样存在隐写样本与载体样本之间差异微小的问题(张祎 等,2022)。Liu 等人(2019)发现,使用隐写分析器可以在一定程度上检测出对抗样本,并在空域富模型隐写分析器的基础上提出 ESRM(spatial rich model),实现了对抗样本的高精度检测。

作为一种有损图像压缩方式,JPEG 压缩—解压缩过程中会产生量化误差、截断误差、舍入误差以及转换误差等(Wang 等,2020)。这些误差的数值范围通常较小。研究表明,JPEG 误差可有效实现对抗噪声的放大(Zhao 和 Wang,2021)。实验表明,使用 JPEG 误差可在白盒攻击数据集上达到良好的检测效果。使用质量因子为 100 的 JPEG 误差输入 VGG16(Visual Geometry Group)网络(Simonyan 和 Zisserman,2015),对快速梯度符号法(fast gradient sign method,FGSM)(Goodfellow 等,2015)生成对抗样本的检测率可达 90% 以上。

由于对抗攻击方法的多样性,对抗噪声的幅值存在明显差异,无法人为判断适合当前识别任务的质量因子。此外,多个质量因子压缩—解压缩误差可能都对识别有效。本文设计了多因子误差注意力模型,在并行提取多质量因子误差的同时,利用注意力机制自适应调整每个误差分支的权重。接着以特征统计层(Wang 等,2022)为基础设计了注意力特征统计层,获取最高卷积层输出特征图的多维统计特征,并为每个特征值分配权重。最终将高维带权特征统计向量送入全连接层,实现了对抗攻击方法的有效识别。

1 先验知识与相关工作

1.1 JPEG 误差

彩色图像的 JPEG 压缩过程包括图像分块、颜色空间转换、离散余弦变换(discrete cosine transform,DCT)、量化和编码等步骤,如图 1 所示。编码过程分为 zigzag 编码和霍夫曼编码两个步骤,均为无损编码。

由于 JPEG 是一种图像有损压缩方式,压缩前后会产生误差。为实现数据量的有效压缩,量化后数据以整数形式存储,小数部分直接抛弃(包括四舍五入、向上取整或向下取整等方式)。因此,量化后的数据即使经过反量化,依然与量化前的数据存在差异,从而产生量化误差。量化误差的提取过程如图 2(a)所示。数字图像在 RGB 和 YCbCr(luminance, colour-difference of blue and colour-difference of red)颜色空间中都是以 8 位整数的形式存储的,颜色空间转换和逆变换之后的数据会四舍五入为 0～255 之间的整数,从而产生转换误差。单独提取转换误差需要人为添加二次压缩的过程,如图 2(b)所示。在 JPEG 格式图像解压缩的过程中,反离散余弦变换变换(inverse discrete cosine transform,IDCT)之后的数据依然是小数,需要转换成 8 位整

数,由其产生截断误差和舍入误差。所谓截断是指将大于 255 的数截断为 255,小于 0 的数据截断为 0,防止出现颜色反转;舍入是指将截断后的小数四舍五入为整数。截断误差和舍入误差的提取方式如图 2(c)所示。反 DCT 变换后数据 $\boldsymbol{M}$ 的截断舍入的方式为

$$\bar{m}=\begin{cases}255 & m>255\\ \lceil m\rceil & m\%1\geqslant 0.5\\ \lfloor m\rfloor & m\%1<0.5\\ 0 & m<0\end{cases} \tag{1}$$

式中,m 表示 $\boldsymbol{M}$ 中的元素,$\bar{m}$ 表示截断舍入之后的 m,% 表示取余符号。

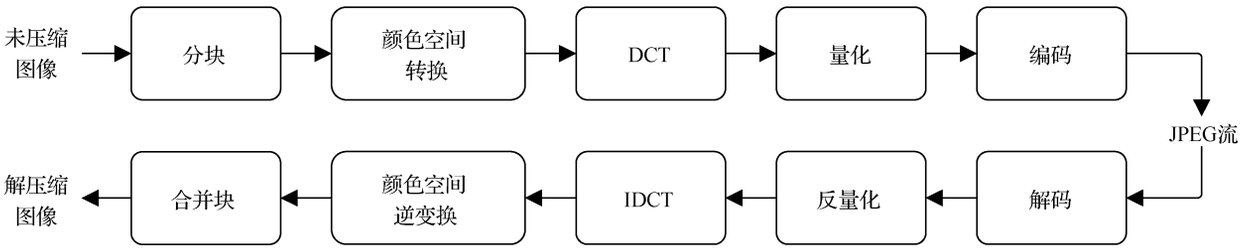

图 1　JPEG 压缩和解压缩流程

Fig. 1　JPEG compression and decompression process

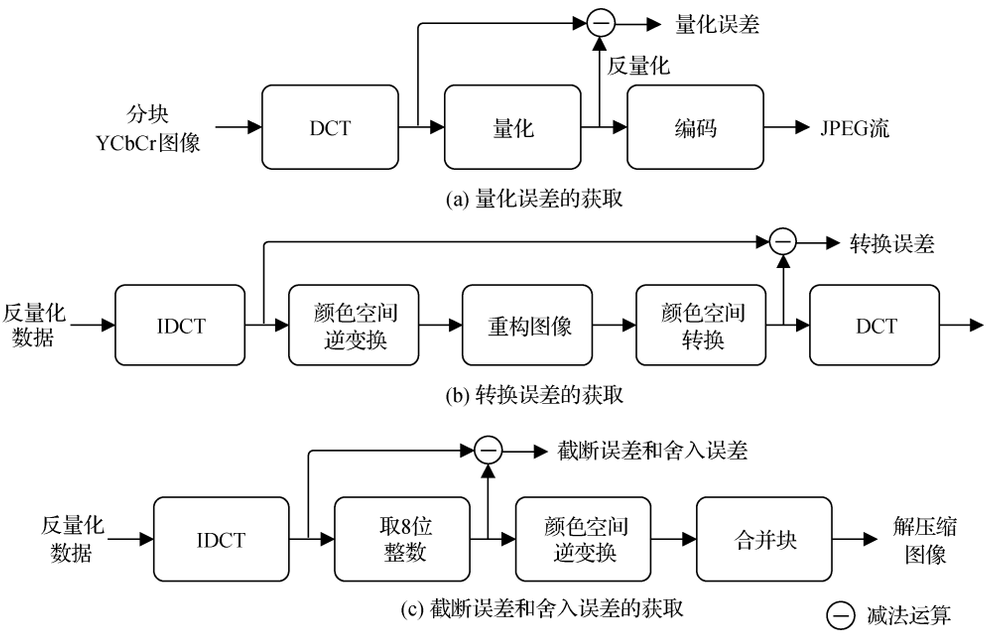

图 2　各类 JPEG 误差的获取

Fig. 2　Extraction of various JPEG errors

((a) extraction of quantification error; (b) extraction of conversion error; (c) extraction of truncation error and rounding error)

1.2　JPEG 误差卷积模拟

转换误差的单独提取需要添加额外的压缩步骤,使误差提取前后的图像处于相同的颜色空间(都处在 YCbCr 颜色空间)。然而,完整的压缩—解压缩过程前后,图像都是处于 RGB 颜色空间的,不需要额外的颜色空间转换来提取整体的 JPEG 压缩误差。JPEG 压缩—解压缩误差的整体提取过程如图 3 所示。

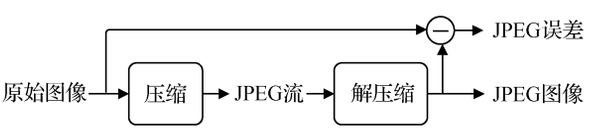

图 3　JPEG 压缩—解压缩误差提取

Fig. 3　Extraction of JPEG compression-decompression error

在图 1 所示的压缩—解压缩过程中,编码和解码过程均是无损的,压缩误差的提取可以省略编解码的步骤。除取整操作外,JPEG 压缩过程均为线性变换。因此,JPEG 压缩误差可以利用卷积神经网络

的组件实现(Zhao 和 Wang,2021),从而实现压缩—解压缩误差的并行提取。分块 DCT 变换过程可以表示为

$$\boldsymbol{X}_f = Flatten(\boldsymbol{X}_b) \times \boldsymbol{M}_t \tag{2}$$

式中,$\boldsymbol{X}_b$ 表示图像 $\boldsymbol{X}$ 的一个小块,$\boldsymbol{X}_f$ 表示频率域的 $\boldsymbol{X}_b$,$Flatten$ 表示将图像块摊平到 1 维,× 表示矩阵乘法,$\boldsymbol{M}_t$ 表示 DCT 变换阵。

令 $\boldsymbol{X}_{fl} = Flatten(\boldsymbol{X}_b)$,矩阵乘法的运算过程可以表示为

$$\boldsymbol{X}_f = \boldsymbol{X}_{fl} \times \boldsymbol{M}_t = [\boldsymbol{X}_{fl} \times \boldsymbol{M}_{t1}, \boldsymbol{X}_{fl} \times \boldsymbol{M}_{t2}, \cdots, \boldsymbol{X}_{fl} \times \boldsymbol{M}_{t64}] \tag{3}$$

式中,$\boldsymbol{M}_{tn}$ 表示 $\boldsymbol{M}_t$ 的第 n 个列向量。由于 $\boldsymbol{X}_{fl}$ 与 $\boldsymbol{M}_{tn}$ 均为 1 维向量,乘法规则为对应元素相乘求和。具体为

$$\boldsymbol{X}_{fl} \times \boldsymbol{M}_{tn} = X_{fl1} \times M_{tn1} + X_{fl2} \times M_{tn2} + \cdots + X_{fl64} \times M_{tn64} \tag{4}$$

式中,X_{fli} 表示 $\boldsymbol{X}_{fl}$ 的第 i 个元素,M_{tni} 表示 $\boldsymbol{M}_{tn}$ 的第 i 个元素。式(4)的计算过程与卷积运算完全一致,且与向量 $\boldsymbol{X}_{fl}$ 和 $\boldsymbol{M}_{tn}$ 的形状无关。使用式(2)的变换过程需要每次将图像分割为 8 × 8 的小块,并将每个小块内的数据摊平到 1 维才能运算。而式(4)表明,将 $\boldsymbol{M}_t$ 与形变为 8 × 8 的小块再与 $\boldsymbol{X}_b$ 进行,结果与式(2)完全一致。具体为

$$Flatten(\boldsymbol{X}_b) \times \boldsymbol{M}_t = \boldsymbol{X}_f = \boldsymbol{X}_b * Reshape(\boldsymbol{M}_t) \tag{5}$$

式中,$Reshape(\,)$ 表示将 $\boldsymbol{M}_{tn}$ 形变为 8 × 8 的卷积核。使用式(5)的方式实现 DCT 变换,不需要将 $\boldsymbol{X}_b$ 进行摊平。此外,只需要将卷积的步长设置为 8,就可实现输入图像 $\boldsymbol{X}$ 不同小块之间的切换。因此,图像分块的过程也可以相应避免。

IDCT 变换的过程和 DCT 变换基本一致,只是采用的 $\boldsymbol{M}_t$ 不同。因此,DCT 变换和 IDCT 变换都可以用卷积运算实现。

颜色空间转换及其逆变换是以像素为单位进行的,分块—颜色空间转换的过程发生调换,并不会对 JPEG 压缩的输出产生影响。分块 DCT 变换的过程可以共同由式(5)所示的卷积过程实现。同理,调换颜色空间逆变换—合并块的过程也不会对 JPEG 解压缩的输出产生影响。图 3 所示的 JPEG 误差获取过程可以简化为图 4 所示的过程。

使用图 4 所示的 JPEG 误差模拟提取方法和常用的 JPEG 压缩—解压缩方法(Wang 等,2020)在计算结果上完全一致。本文使用 10 000 幅 224 × 224 像素的图像比较了它们的性能,常用的 JPEG 压缩—解压缩方式的耗时约为 12.73 s,本文中的模拟方案耗时约为 4.32 s。

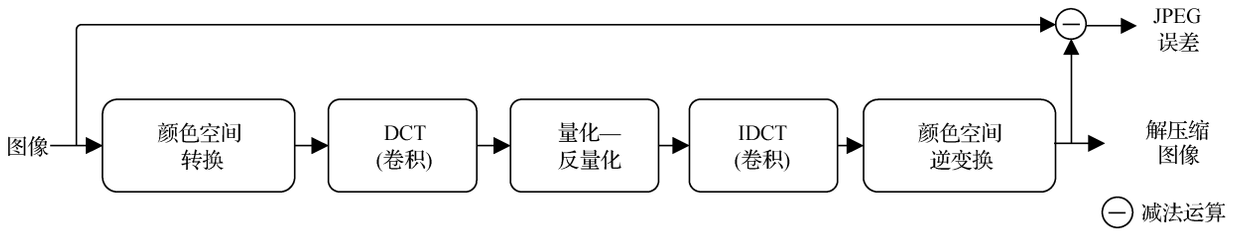

图 4 JPEG 误差提取简化

Fig. 4 Simplified JPEG error extraction

1.3 对抗样本检测

由于对抗噪声的添加,对抗样本与自然样本在数据分布上存在差异。对抗样本检测工作的本质就是检测对抗噪声给图像带来的分布差异。

目前的检测方案主要分为基于样本自身特征和基于原始分类器中间特征的两大类方法。基于样本自身特征的检测方案从对抗样本本身的统计特征出发,利用隐写分析器、深度神经网络等工具分析差异实现检测。Grosse 等人(2017)发现,对抗样本与自然样本的最大值、平均值与能量分布等统计特征存在差异,并使用这些特征进行了对抗样本的检测。由于对抗样本检测与隐写分析任务在检测微小噪声方面的相似性,隐写分析器对于对抗样本检测任务同样有效(Schöttle 等,2018)。Schöttle 等人(2018)基于隐写分析方法实现了对投影梯度下降方法(Madry 等,2018)生成对抗样本的检测。投影梯度下降攻击方法的基本步骤与基本迭代法(Kurakin 等,2017)相似,但由于每次迭代步幅很小,攻击成功率极高且生成对抗样本的噪声幅值很小。Liu 等人(2019)结合隐写分析方法,提出了基于空域富模

型(Fridrich 和 Kodovsky,2012)的增强版 ESRM,实现了对抗样本的检测并取得较高的精确度。Wang 等人(2022)在神经网络中引入了特征统计层,并使用级联方式对每个卷积层输出的特征图统计多维特征,实现了对抗样本的端到端检测。

基于原始分类器中间特征的检测方法无法单独检测对抗样本,需要将样本输入到被攻击的原始分类器中,利用原始分类器神经元的输出变换实现检测。Feinman 等人(2017)提出使用原始分类器中间层输出的核密度估计来衡量对抗样本与自然样本的距离差异。Li 和 Li(2017)使用支持向量机(support vector machine,SVM),结合级联特征对对抗样本进行检测,直接获取原始分类器每个卷积层的输出特征图,将这些特征图的统计特征送入支持向量机进行分类,然而检测精度较低。Carrara 等人(2019)提出一种先提取分类器隐藏层神经元的输出,再使用长短期记忆网络对对抗样本进行检测的方法。

1.4　对抗样本与 JPEG 压缩

在人为生成对抗样本的过程中,对抗噪声的添加强度以刚好欺骗受攻击模型为标准。对抗样本的攻击效果具有脆弱性,JPEG 压缩可有效抵御对抗攻击(Kurakin 等,2017)。Kurakin 等人(2017)的实验表明,在扰动系数为 16 的 FGSM(Goodfellow 等,2015)和基本迭代法(basic iterative method,BIM)(Kurakin 等,2017)对抗样本上,JPEG 压缩能起到良好的抵御效果。

对抗噪声的幅值较小,使对抗样本检测困难。对抗样本的生成过程只考虑到噪声在空间域的分布情况,没有关注变换域。而 JPEG 误差大多是在 DCT 域产生的,且误差幅值较小。Zhao 和 Wang(2021)模拟 JPEG 压缩过程,将对抗样本分块变换到离散余弦域,从而构建了轻量级的对抗样本检测模型。他们的研究表明,JPEG 压缩—解压缩误差在对抗样本检测任务中可以有效放大对抗噪声。

1.5　注意力机制

在神经网络中引入注意力机制的目的是定位重要区域,并抑制无用信息。图像识别领域的注意力机制主要分为空间注意力(Jaderberg 等,2015)和通道注意力(Hu 等,2018)两种模式。空间注意力在通道维度进行池化,进而使用卷积层获取关于空间位置的权重。通道注意力机制将特征图分层获取最大值和均值,分别输入到辅助的多层感知机获取每个通道的权重。Woo 等人(2018)将两种注意力机制结合在一起,提出了空间通道混合注意力机制。

2　攻击方法识别模型

2.1　多因子误差注意力模型

JPEG 压缩—解压缩过程中产生的误差对于对抗样本的检测可以起到重要作用(Zhao 和 Wang,2021)。然而,在对抗攻击方法的识别任务中,不仅需要检测到对抗噪声的存在,还需要识别出不同攻击方法生成噪声的分布模式。虽然压缩误差可以起到放大噪声的作用,但是不同攻击方法生成的对抗噪声在分布、幅值等方面存在差异,无法有效确定适合的压缩质量因子。

为避免单因子压缩误差引起的有效识别特征丢失,本文联合使用多质量因子提取压缩误差,设计了多因子误差注意力模型,实现对抗攻击方法的有效识别。多因子误差注意力模型主要包含压缩模块、解压缩模块、误差分支注意力模块、注意力特征统计层以及卷积层、全连接层等,如图 5 所示。

颜色空间转换是线性变换,变换过程为

$$\begin{cases} \boldsymbol{X}_{\mathrm{Y}} = 0.299\boldsymbol{X}_{\mathrm{R}} + 0.587\boldsymbol{X}_{\mathrm{G}} + 0.114\boldsymbol{X}_{\mathrm{B}} \\ \boldsymbol{X}_{\mathrm{Cb}} = -0.168\,7\boldsymbol{X}_{\mathrm{R}} - 0.331\,3\boldsymbol{X}_{\mathrm{G}} + 0.5\boldsymbol{X}_{\mathrm{B}} + 128 \\ \boldsymbol{X}_{\mathrm{Cr}} = 0.5\boldsymbol{X}_{\mathrm{R}} - 0.418\,7\boldsymbol{X}_{\mathrm{G}} - 0.081\,3\boldsymbol{X}_{\mathrm{B}} + 128 \end{cases} \tag{6}$$

式中,$\boldsymbol{X}_{\mathrm{Y}}$,$\boldsymbol{X}_{\mathrm{Cb}}$,$\boldsymbol{X}_{\mathrm{Cr}}$,$\boldsymbol{X}_{\mathrm{R}}$,$\boldsymbol{X}_{\mathrm{G}}$ 和 $\boldsymbol{X}_{\mathrm{B}}$ 分别表示输入图像 $\boldsymbol{X}$ 的 Y,Cb,Cr,R,G 和 B 颜色通道。$\boldsymbol{X}_{\mathrm{Y}}$ 的计算过程相当于使用三通道的 1 × 1 卷积核(权重初始化为 0.299、0.587、0.114)对 RGB 颜色空间的图像 $\boldsymbol{X}$ 进行卷积。只需要使用 3 个 1 × 1 的卷积核,就可以实现 $\boldsymbol{X}$ 从 RGB 颜色空间到 YCbCr 颜色空间的转换。同理,颜色空间逆变换也可以使用 1 × 1 的卷积实现。

图 5 中的压缩模块不包含量化步骤,解压缩模块也不包含反量化步骤。对抗样本首先输入到解压缩模块,实现颜色空间转换和 DCT 变换。使用不同质量因子获取误差时,量化—反量化系数也相应改变。多个反量化后的数据分别通过解压缩模块,并与原始输入样本作差获取压缩误差。接着,多分支压缩误差经由误差分支注意力进入卷积模块,再将最上层卷积特征图送入注意力特征统计层获取带权

高维统计特征。最后,将特征值输入全连接层进行分类,从而实现攻击方法的识别。

在图 4 所示的 JPEG 误差提取过程中,颜色空间转换、DCT 变换的过程与质量因子无关。这意味着即使提取同一样本的关于多个质量因子的压缩误差,也只需要经过一次压缩模块。虽然解压缩模块可以共用,但是多个分支反量化后的数据存在差异,每个分支的输出都需要经过解压缩模块。

压缩、解压缩模块的使用是为了使用深度神经网络组件并行加速 JPEG 误差的提取,这两个模块并不需要参与识别网络的训练过程。此外,压缩—解压缩过程中使用的量化、取整等函数均不可导,影响了梯度的反向传播。压缩、解压缩模块内的参数无法随识别网络的训练过程优化。

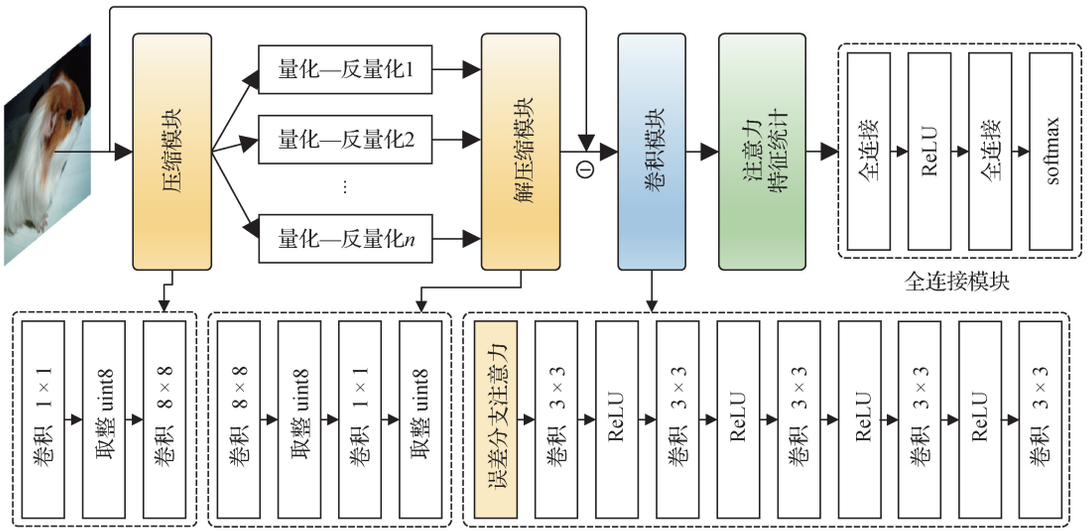

图 5　多因子误差注意力模型结构

Fig. 5　Structure of multi-factor error attention model

2.2　多误差分支注意力

在多质量因子压缩误差的提取过程中,图 4 中的颜色空间转换和 DCT 变换过程不涉及质量因子,多个质量因子压缩可以公用压缩模块。虽然多因子误差注意力模型仅包含一个解压缩模块,但是不同量化—反量化分支输出的数据需要依次通过这个模块。逻辑上的压缩解压缩模块如图 6 所示,各个解压缩模块的参数完全一致。

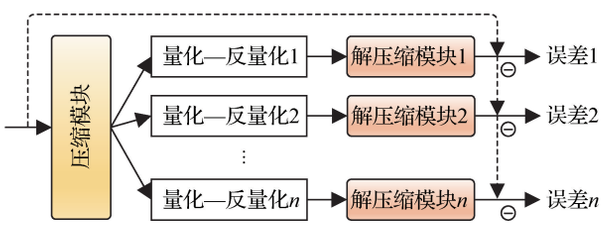

图 6　误差提取逻辑结构

Fig. 6　Logical structure of the error extraction module

为分析不同质量因子压缩误差对于识别任务的重要程度,本文在多因子误差注意力模型中设计了误差分支注意力机制,如图 7 所示。各分支压缩—解压缩误差经过一个卷积层后,在通道维度上连接形成图 7 中的特征图 1,同时进行全局最大池化和全局均值池化。池化后的数据拉伸形成 1 维向量,输入全连接层计算特征图 1 各通道的权重。这些权重分组输入后续全连接层,计算各个误差分支的权重。图 7 中 n 个误差分支分别输入包含 k 个卷积核的卷积层,拼接形成包含 $n \times k$ 个通道的特征图 1。全局最大池化和全局平均池化后摊平的向量各包含 $n \times k$ 个特征值,输入包含 $n \times k/2$ 个神经元的全连接层,再分别连接到 $n \times k$ 个神经元的全连接层输出通道权重。通道权重以 k 个为一组进行顺序分组,分别输入到包含 n 个神经元的全连接层,最终输出 n 个误差分支对应的权重。误差分支注意力机制以分支为单位评估误差特征重要性,在限定关注范围的同时,避免了训练过程中可能出现的颜色通道整体丢失,为后续攻击方式的识别充分保留了完整的误差信息。

2.3　注意力特征统计层

特征统计层以全局池化层(Lin 等,2014)为基

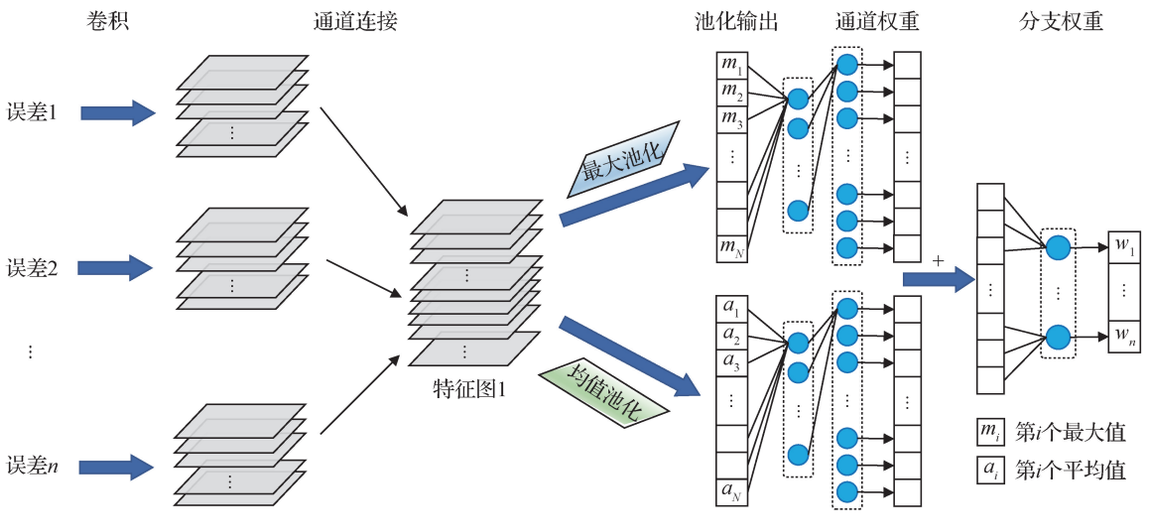

图 7　误差分支注意力

Fig. 7　Attention mechanism of compression factors

础，在全局特征最大池化层和全局均值池化层的基础上同时获取特征图各通道的最大值、均值、最小值和方差等统计量，组合形成高维特征向量（Wang 等，2022）。使用全局池化层替代卷积层与全连接层之间的摊平变换，大幅减少了网络参数量，减轻了过拟合现象。特征统计层在不大幅增加参数量的同时，丰富了特征值类型。

特征统计层的计算过程如图 8 所示。卷积层输出的特征图分层进入特征统计层，在全局尺度上统计最大值、均值、最小值和方差等特征值，并重新排列形成 1 维特征向量。这里统计的特征不局限于上述 4 种，是一个可选的组合。

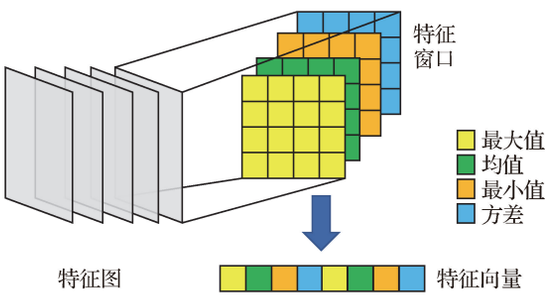

图 8　特征统计层

Fig. 8　Feature statistical layer

本文在特征统计层的基础上设计了注意力特征统计层，如图 9 所示。注意力特征统计层采用了最大值、均值、最小值、方差和偏态 5 种特征。图 9 的左半部分描述了通道注意力的计算过程，特征图分通道进行最大池化和均值池化，分别形成 1 维向量，送入多层感知机，将结果输出后相加形成通道权重。图 9 的右半部分描述了特征注意力的计算过程，特征图的多种统计特征在特征类型维度上分别进行最大池化和均值池化，并分别送入多层感知机，输出特征权重。最后，通道权重与特征权重相乘，并经过 sigmoid 激活形成关于每个特征值的权重。输出特征向量 $\boldsymbol{X}_f$ 的计算过程为

$$\boldsymbol{X}_f = \boldsymbol{W}_f \times (\max(\boldsymbol{X}_t) \cup \text{mean}(\boldsymbol{X}_t) \cup \min(\boldsymbol{X}_t) \cup \text{var}(\boldsymbol{X}_t) \cup \text{skew}(\boldsymbol{X}_t)) \tag{7}$$

式中，max，mean，min，var 和 skew 分别表示最大值、均值、最小值、方差和偏态的计算，∪ 表示取并集，用 $\boldsymbol{X}_t$ 表示注意力特征统计层的输入，$\boldsymbol{W}_f$ 表示图 9 中形成的权重向量。

3　实验与分析

3.1　实验说明

3.1.1　实验环境

本文使用的多因子误差注意力模型基于 PyTorch 深度学习框架实现。具体的软件环境为 windows 10，python3.7 以及 PyTorch 1.11。硬件环境为英特尔 I5 11400 CPU，32 GB 内存，英伟达 RTX 3080Ti GPU。

3.1.2　数据集说明

本文以 ImageNet 图像分类数据集为基础，使用 FGSM（Goodfellow 等，2015）、BIM（Kurakin 等，2017）、Deepfool（Moosavi-Dezfooli 等，2016）、C&W

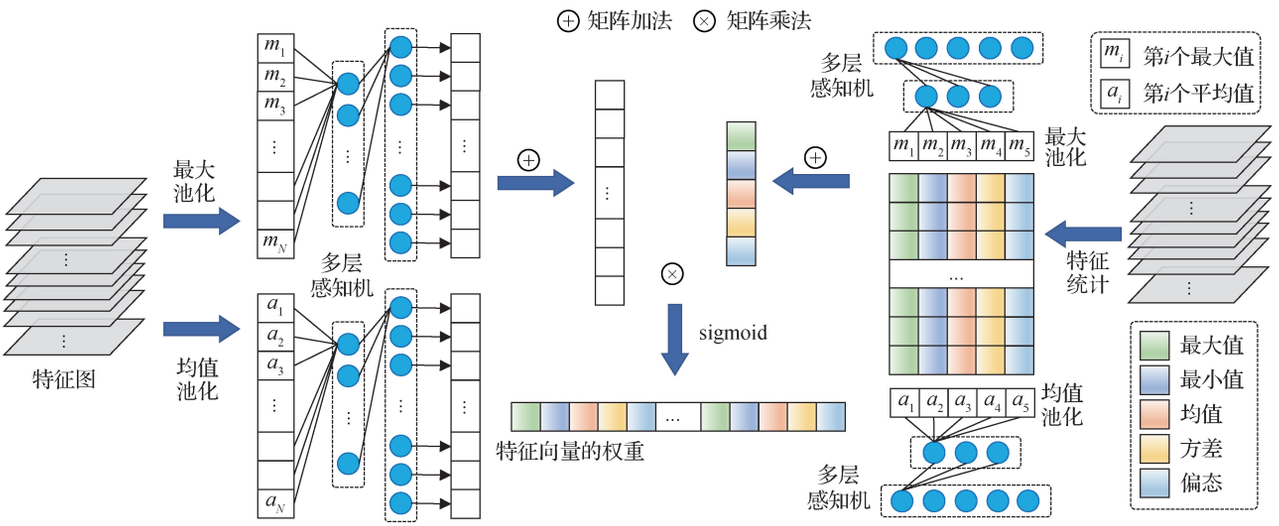

图 9　注意力特征统计层

Fig. 9　Feature statistical layer with attention

(Carlini 等,2017)、DDN(Rony 等,2019)、Boundary-Attack (Brendel 等, 2018)、BrendelBethgeAttack (Brendel 等,2019)和 BanditsAttack(Ilyas 等,2019)等 8 种攻击方法生成对抗样本。为方便描述,本文将这 8 种攻击方法分别称为 FGSM、BIM、Deepfool、C&W、DDN、Boundary、Brendel 和 Bandits。其中,FGSM 和 BIM 都分别使用了 2、4、6、8 等 4 个扰动系数,Bandits 包含 L_2 和 L_∞ 两个版本。因此,本文一共生成了 15 个子数据集。在每个子数据集的生成过程中,本文从 ImageNet 数据集中随机挑选 30 000 幅图像尝试进行攻击。由于不同攻击方法成功率的差异,每个子数据集最终生成的对抗样本数量不同。在训练、测试时,随机从每个子数据集中挑选 12 000 个对抗样本。其中 10 000 个用于训练,2 000 个用于测试。

3.1.3　实验参数

实验过程中,误差分支数量为 4。为充分保留样本原有信息,第 4 个分支使用归一化样本代替 JPEG 误差。其余 3 个通道采用的质量因子分别为 90、95、100。模型中卷积层的具体参数如表 1 所示。此外,模型还包含 3 个全连接层。表 1 中卷积模块的第 2 个与第 3 个卷积层之间出现了通道数量的缩减,这是通过误差分支注意力机制实现通道合并的结果。

表 1　模型参数

Table 1　Model parameters

卷积核尺寸	步长	卷积核数量	激活函数
1×1	1×1	3	uint8
8×8	8×8	64	–
8×8	8×8	64	uint8
1×1	1×1	3	uint8
3×3	1×1	16	ReLU
3×3	1×1	32	ReLU
3×3	1×1	128	ReLU
3×3	1×2	256	ReLU
3×3	1×1	256	ReLU
3×3	2×1	512	ReLU
3×3	1×1	512	ReLU

注:“–”表示不使用激活函数。

3.2　性能评估

3.2.1　攻击方法识别

对抗样本攻击方法识别是本文的主要任务。实验在由 15 个子集组成的数据集上和在由 8 个子集组成的数据集上进行多因子误差注意力模型的训练和测试。8 个子数据集分别为扰动系数为 2 的 FGSM 和 BIM,L_2 版本的 Bandits 和其他攻击方法生成的子数据集。在训练过程中,训练集和测试集上的检测准确率如图 10 所示。

图 10(a)(b)分别展示了 15 分类和 8 分类的训练效果。在图 10(a)所示的训练过程中,20 个周期之后的测试准确率趋于稳定,测试准确率达到了 91.19%。然而在图 10(b)所示的 8 类别训练过程中,测试准确率仅能达到 84% 左右,且 20 个周期之

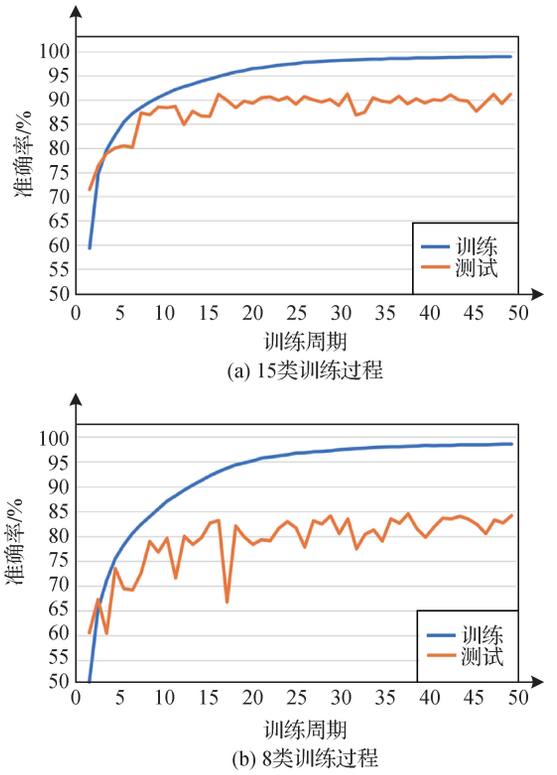

图 10　多因子误差注意力模型训练效果

Fig. 10　Training effect of multi-factor error attention model

((a) training process for 15 categories;

(b) training process for 8 categories)

后依然存在较大幅度的波动。两个数据集之间的差异主要体现在 FGSM 和 BIM 攻击方法扰动系数上。因此,本文计算了 15 类模型关于每个类别的召回率,分析扰动系数对模型的影响。召回率(recall) R 计算为

$$R = \frac{TP}{TP + FN} \tag{8}$$

式中, TP (true positive)表示正样本中预测为正样本的数量, FN (false negative)表示正样本中预测为负样本的数量。然而对于多分类任务,不能简单区分正负类样本。第 k 类召回率 R_k 计算为

$$R_k = \frac{(D_y \mid T_k)}{(D_y \mid T_k) + (D_n \mid T_k)} \tag{9}$$

式中,$(D_y \mid T_k)$ 表示第 k 个类别中检测为该类别的样本数量,$(D_n \mid T_k)$ 表示第 k 个类别中未检测为该类别的样本数量。图 10(a)中第 50 个周期各类别的召回率如表 2 所示。

比较表 2 中 BIM 和 FGSM 各扰动系数与其他类别的召回率,可以发现各扰动系数召回率均偏高。

表 2　15 分类各类别召回率

Table 2　Recall rate of 15 categories

类别	召回率/%
BIM(ε = 2)	96.05
BIM(ε = 4)	95.04
BIM(ε = 6)	96.95
BIM(ε = 8)	92.04
FGSM(ε = 2)	93.85
FGSM(ε = 4)	98.35
FGSM(ε = 5)	97.20
FGSM(ε = 8)	99.39
Deepfool	72.75
C&W	72.34
DDN	85.58
Boundary	89.63
Brendel	80.64
Bandits(L_2)	98.55
Bandits(L_∞)	99.90

这表明所提多因子误差注意力模型对于扰动系数的识别任务具有良好效果。

本文随机选择了 4 个子数据集对所提模型进行训练,测试准确率如表 3 所示。表 3 中 BIM 和 FGSM 的扰动系数均为 2。对比表 3 中的最后一行数据与表 2 中的数据,可以发现表 3 中最后一行的数据为表 2 中召回率最低的 4 个类别的组合。由于各个子数据集中测试集的样本数量一致,这 4 个类别召回率的均值即为在 4 个子数据集上的测试准确率,具体为

$$Acc_j^{j+3} = \frac{1}{4}\sum_{k=j}^{j+3} R_k \tag{10}$$

式中,Acc_j^{j+3} 表示从 j 类开始连续 4 个类别的召回率,R_k 使用式(9)计算得到。15 类模型在 Deepfool、C&W、Brendel 和 DDN 等 4 个子数据集上的测试准确率为 77.83%,低于表 3 中的准确率。表 3 中最后一行数据对应的召回率如表 4 所示。C&W 和 Deepfool 数据集上的召回率比表 3 有明显提升。这表明多因子误差注意力模型在特定数据集上进行针对性训练后,可以得到更好的测试效果。

表 3　4 类别识别准确率

Table 3　Accuracy of 4 categories

/%

数据集	准确率
BIM、DDN、Boundary、Bandits	95.68
FGSM、DDN、C&W、Bandits(L_∞)	94.93
Deepfool、C&W、Brendel、DDN	81.85

表 4　C&W、Deepfool、Brendel 和 DDN 的召回率

Table 4　Recall rate of C&W, Deepfool, Brendel and DDN

数据集	召回率/%
C&W	77.60
Deepfool	97.35
Brendel	75.95
DDN	76.50

3.2.2　攻击强度检测

所提模型在攻击方法识别任务中的表现表明对扰动强度检测同样有着良好表现。实验分别使用 4 种扰动系数的 FGSM 和 BIM 子数据集对多因子误差注意力模型进行训练，准确率和召回率如表 5 所示。表 5 的结果表明，所提多因子误差注意力模型在单独的对抗攻击扰动强度检测任务中同样具有优异的性能。

表 5　攻击强度检测准确率

Table 5　Accuracy of attack coefficients

/%

数据集	准确率	召回率			
		$\varepsilon=2$	$\varepsilon=4$	$\varepsilon=6$	$\varepsilon=8$
FGSM	99.67	99.75	99.60	99.50	99.85
BIM	96.85	97.65	97.90	94.25	97.60

3.2.3　质量因子分析

为评估所提模型各误差分支 JPEG 压缩质量因子对检测效果的影响，本文使用多种质量因子的组合对模型进行了训练，测试结果如表 6 所示。表 6 中的训练和测试过程均是在 15 类数据集上进行的。实验结果显示，80 以上质量因子的组合均可取得良好的分类效果。

3.3　模块性能分析

模块性能分析实验均在 15 类数据集上训练及测试。对照所提多因子误差注意力模型各项参数，本文构建了普通的卷积神经网络。4 个分支输入均为归一化样本，注意力特征统计层使用 5 个完全相同的全局池化层代替。全局池化分为最大池化和均值池化两种，训练过程如图 11 所示。从图 11 中两种池化方式的测试效果可以看出，池化方式对于模型的对抗攻击方法检测性能没有明显影响。多因子误差注意力模型不仅在测试效果上优于普通神经网络，且收敛速度更快。本文模型在 15 个训练周期之后已经趋于收敛，而最大池化、均值池化模型均在 20 个周期之后。

表 6　各压缩质量因子准确率对比

Table 6　Accuracy of different compression quality factors

/%

压缩因子				准确率
分支 1	分支 2	分支 3	分支 4	
90	95	100	–	**91.19**
85	90	95	100	90.16
80	85	90	95	90.15

注：加粗字体表示最优结果。“–”表示输入数据未压缩。

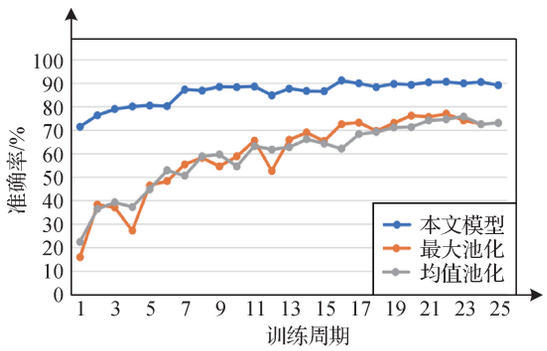

图 11　本文模型与同规模普通神经网络对比

Fig. 11　Performance of the proposed model versus neural networks with the same amount of parameters

多误差分支模块并行计算样本关于多个压缩质量因子的误差，避免了在特定任务中人工筛选适合质量因子的过程。当多误差分支采用的质量因子相同时，测试效果如图 12 所示。质量因子选择不当时，如图 12 中的 100 或 80，测试过程中的准确率波动较大，无法保证模型的稳定识别效果。图中的曲线也表明，多个质量因子的使用也提升了模型的识别准确率。

在保留多个压缩因子的前提下，本文模型中的

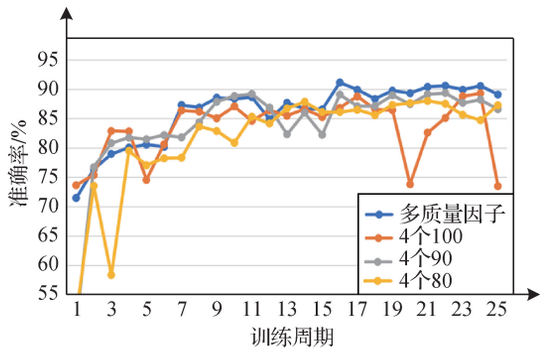

图 12　多误差分支模块性能

Fig. 12　Performance of the multi-brunch error

注意力特征统计层分别替换为特征统计层(Wang 等,2022)、全局最大池化层和全局平均池化层,测试结果如图 13 所示。使用特征统计层的准确率略高于全局池化层,且波动更小。注意力特征统计层识别效果明显好于特征统计层和全局池化层。

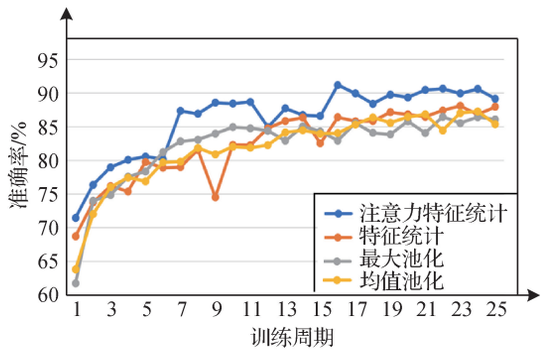

图 13　注意力特征统计模块性能

Fig. 13　Performance of the feature statistical layer with attention

3.4　对抗样本检测

对抗样本检测任务需要使用单个子数据集对模型进行训练及测试,测试集仅包含 2 000 个对抗样本。为避免数据量过小带来的测试偏差以及训练过程中出现过拟合现象,重新使用 FGSM、Deepfool 和 C&W 攻击方法针对 VGG16 网络各生成了 30 000 个对抗样本。其中 25 000 个用于训练,5 000 个用于测试。每个对抗样本均与一个自然样本对应。因此,每个数据集包含 50 000 个训练样本和 10 000 个测试数据。

在 6 个数据集上,本文模型分别与 ESRM(Liu 等,2019)、SmsNet(stochastic multifilter statistical network)(Wang 等,2022)和 DCT-Like(Zhao 和 Wang,2021)3 种对抗样本检测方案进行准确率对比,结果如表 7 所示。多因子误差注意力模型各项参数设置与对抗攻击方法识别中一致,未针对对抗样本检测任务调整。所提模型检测准确率与 SmsNet 及 DCT-Like 模型相比,准确率差距小于 1%。同时,在较难检测的 C&W 数据集上,所提模型表现优于现有模型。实验表明,多因子误差注意力模型在对抗样本检测任务中同样具有良好表现。

表 7　对抗样本检测准确率

Table 7　Accuracy of adversarial example detection

/%

模型	FGSM				Deepfool	C&W
	$\varepsilon=2$	$\varepsilon=4$	$\varepsilon=6$	$\varepsilon=8$		
ESRM	98.10	98.53	99.03	99.35	95.13	92.87
SmsNet	98.48	99.46	99.74	99.78	98.26	93.83
DCT-Like	**99.64**	**99.66**	99.03	**99.98**	98.54	95.05
本文	99.18	99.65	**99.75**	99.86	**98.71**	**96.27**

注:加粗字体表示各列最优结果。

所提模型的跨库检测性能如表 8 所示。其中 FGSM 攻击的扰动系数 ε 为 2。从表 8 可以看出,训练集为 C&W 数据集时的跨库测试效果最好,训练集为 FGSM 数据集时的效果最差。这个趋势与 Wang 等人(2022)的实验结果一致。

表 8　跨库检测准确率

Table 8　Detection accuracy of cross dataset

/%

训练集	测试集	准确率
FGSM	Deepfool	76.93
FGSM	C&W	64.41
Deepfool	FGSM	98.41
Deepfool	C&W	79.59
C&W	FGSM	98.34
C&W	Deepfool	97.87

4　结　论

本文从 JPEG 误差对于对抗噪声的放大作用出发,提出多因子误差注意力模型。由于对抗攻击方法的多样性,使用 JPEG 误差放大对抗噪声时压缩因子难以调节。本文利用卷积、取整等方式构建了

压缩模块和解压缩模块,并行获取关于多个质量因子的压缩误差,进而设计了误差分支注意力机制评估各误差分支的重要性,合并通道后输入卷积模块。最后一个卷积层输出的特征图送入注意力特征统计层计算带权高维特征,并将这些特征值输入全连接层。

所提多因子误差注意力模型不仅在攻击方法识别任务中达到了91.19%的准确率,在对抗样本扰动强度检测、对抗样本检测等任务中均有良好表现。在对抗样本检测任务中,本文模型在较难检测的C&W数据集上超过了现有方案的准确率。在其他数据集上的检测准确率与现有方法的差距也小于1%。实验结果表明了本文模型的优越性。

由于JPEG压缩—解压缩过程存在不可导的步骤,本文模型未能实现压缩质量因子的自适应优化。在未来的工作中,将探索质量因子的自动调整方案。但正是由于不可导的步骤阻碍了梯度的传播,无法针对本文模型生成基于梯度的对抗样本,未来将测试非梯度攻击对于本文模型的攻击效果,并尝试设计相应的抵御方案。

参考文献(References)

Akhtar N and Mian A. 2018. Threat of adversarial attacks on deep learning in computer vision: a survey. IEEE Access, 6: 14410-14430 [DOI: 10.1109/ACCESS.2018.2807385]

Brendel W, Rauber J and Bethge M. 2018. Decision-based adversarial attacks: reliable attacks against black-box machine learning models//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: OpenReview.net

Brendel W, Rauber J, Kümmerer M, Ustyuzhaninov I and Bethge M. 2019. Accurate, reliable and fast robustness evaluation//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: [s.n.]: 1152

Carlini N and Wagner D. 2017. Towards evaluating the robustness of neural networks//2017 IEEE Symposium on Security and Privacy (SP). San Jose, USA: IEEE: 39-57 [DOI: 10.1109/SP.2017.49]

Carrara B, Becarelli R, Caldelli R, Falchi F and Amato G. 2019. Adversarial examples detection in features distance spaces//Proceedings of 2019 European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 313-327 [DOI: 10.1007/978-3-030-11012-3_26]

Chen H B, Zhao L, Zhang H M, Wang Z Z, Zuo Z W, Li A L, Xing W and Lu D M. 2021. Diverse image style transfer via invertible cross-space mapping//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 6363-6372 [DOI: 10.1109/ICCV48922.2021.01461]

Feinman R, Curtin R R, Shintre S and Gardner A B. 2017. Detecting adversarial samples from artifacts [EB/OL]. [2022-05-30]. https://arxiv.org/pdf/1703.00410.pdf

Fridrich J and Kodovsky J. 2012. Rich models for steganalysis of digital images. IEEE Transactions on Information Forensics and Security, 7(3): 868-882 [DOI: 10.1109/TIFS.2012.2190402]

Gao S H, Cheng M M, Zhao K, Zhang X Y, Yang M H and Torr P. 2021. Res2 Net: a new multi-scale backbone architecture. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(2): 652-662 [DOI: 10.1109/TPAMI.2019.2938758]

Goodfellow I J, Shlens J and Szegedy C. 2015. Explaining and harnessing adversarial examples//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: [s.n.]

Grosse K, Manoharan P, Papernot N, Backes M and McDaniel P. 2017. On the (statistical) detection of adversarial examples [EB/OL]. [2022-05-30]. https://arxiv.org/pdf/1702.06280v2.pdf

Hu J, Shen L and Sun G. 2018. Squeeze-and-excitation networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7132-7141 [DOI: 10.1109/CVPR.2018.00745]

Hu L Q, Kan M N, Shan S G and Chen X L. 2022. Large pose face recognition with morphing field learning. Journal of Image and Graphics, 27(7): 2171-2184 (胡蓝青, 阚美娜, 山世光, 陈熙霖. 2022. 面向大姿态人脸识别的正面化形变场学习. 中国图象图形学报, 27(7): 2171-2184) [DOI: 10.11834/jig.210011]

Ilyas A, Engstrom L and Madry A. 2019. Prior convictions: black-box adversarial attacks with bandits and priors//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: OpenReview.net

Jaderberg M, Simonyan K, Zisserman A and Kavukcuoglu K. 2015. Spatial transformer networks//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 2017-2025

Kurakin A, Goodfellow I J and Bengio S. 2017. Adversarial examples in the physical world//Proceedings of the 5th International Conference on Learning Representations. Toulon, France: OpenReview.net

Li X and Li F X. 2017. Adversarial examples detection in deep networks with convolutional filter statistics//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 5775-5783 [DOI: 10.1109/ICCV.2017.615]

Lin M, Chen Q and Yan S C. 2014. Network in network [EB/OL]. [2022-05-30]. https://arxiv.org/pdf/1312.4400v3.pdf

Liu J Y, Zhang W M, Zhang Y W, Hou D D, Liu Y J, Zha H Y and Yu N H. 2019. Detection based defense against adversarial examples from the steganalysis point of view//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Long Beach, USA: IEEE: 4820-4829 [DOI: 10.1109/CVPR.2019.00496]

Madry A, Makelov A, Schmidt L, Tsipras D and Vladu A. 2018. Towards deep learning models resistant to adversarial attacks//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: OpenReview.net

Moosavi-Dezfooli S M, Fawzi A and Frossard P. 2016. DeepFool: a simple and accurate method to fool deep neural networks//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 2574-2582 [DOI: 10.1109/CVPR.2016.282]

Papernot N, McDaniel P, Wu X, Jha S and Swami A. 2016. Distillation as a defense to adversarial perturbations against deep neural networks//Proceedings of 2016 IEEE Symposium on Security and Privacy (SP). San Jose, USA: IEEE: 582-597 [DOI: 10.1109/SP.2016.41]

Rony J, Hafemann L G, Oliveira L S, Ayed I B, Sabourin R and Granger E. 2019. Decoupling direction and norm for efficient gradient-based L_2 adversarial attacks and defenses//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 4317-4325 [DOI: 10.1109/CVPR.2019.00445]

Schöttle P, Schlögl A, Pasquini C and Böhme R. 2018. Detecting adversarial examples —— a lesson from multimedia security//Proceedings of the 26th European Signal Processing Conference (EUSIPCO). Rome, Italy: IEEE: 947-951 [DOI: 10.23919/EUSIPCO.2018.8553164]

Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: [s.n.]

Wang J W, Wang H, Li J, Luo X Y, Shi Y Q and Jha S K. 2020. Detecting double JPEG compressed color images with the same quantization matrix in spherical coordinates. IEEE Transactions on Circuits and Systems for Video Technology, 30(8): 2736-2749 [DOI: 10.1109/TCSVT.2019.2922309]

Wang J W, Zhao J J, Yin Q L, Luo X Y, Zheng Y H, Shi Y Q and Jha S K. 2022. SmsNet: a new deep convolutional neural network model for adversarial example detection. IEEE Transactions on Multimedia, 24: 230-244 [DOI: 10.1109/TMM.2021.3050057]

Woo S, Park J, Lee J Y and Kweon I S. 2018. CBAM: convolutional block attention module//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 3-19 [DOI: 10.1007/978-3-030-01234-2_1]

Zhang Y, Luo X Y, Wang J W, Lu W, Yang C F and Liu F L. 2022. Research progress on digital image robust steganography. Journal of Image and Graphics, 27(1): 3-26 (张祎, 罗向阳, 王金伟, 卢伟, 杨春芳, 刘粉林. 2022. 数字图像鲁棒隐写综述. 中国图象图形学报, 27(1): 3-26) [DOI: 10.11834/jig.210449]

Zhao J J and Wang J W. 2021. Lightweight DCT-like domain forensics model for adversarial example//Proceedings of the 19th International Workshop on Digital Watermarking. Melbourne, Australia: Springer: 265-279 [DOI: 10.1007/978-3-030-69449-4_20]

Zou J H, Duan Y X, Ren C L, Qiu J Y, Zhou X Y and Pan Z S. 2022. Perturbation initialization, adam-nesterov and quasi-hyperbolic momentum for adversarial examples. Acta Electronica Sinica, 50(1): 207-216 (邹军华, 段晔鑫, 任传伦, 邱俊洋, 周星宇, 潘志松. 2022. 基于噪声初始化、Adam-Nesterov 方法和准双曲动量方法的对抗样本生成方法. 电子学报, 50(1): 207-216) [DOI: 10.12263/DZXB.20200839]

作者简介

赵俊杰,男,博士研究生,主要研究方向为信息安全和对抗样本取证。E-mail: gino1912@163.com

王金伟,通信作者,男,教授,主要研究方向为信息安全和数字多媒体取证。E-mail: wjwei_2004@163.com

吴俊凤,女,硕士研究生,主要研究方向为对抗样本应用。E-mail: wjf1916219959@outlook.com