兰州理工大学

硕士论文

学校代号 10731　　学　号 202083500010

分 类 号 TP391　　密　级 公　开

蘭州理工大学
LANZHOU UNIVERSITY OF TECHNOLOGY

硕士学位论文

初级视皮层启发的鲁棒深度卷积神经网络研究

学位申请人姓名 董占国

培 养 单 位 计算机与通信学院

导师姓名及职称 柯铭 教授 王路斌 副研究员

王刚 副研究员

学 科 专 业 软件工程

研 究 方 向 生物信息处理

论文提交日期 2023 年 3 月 20 日

学校代号：10731

学　　号：202083500010

密　　级：公开

兰州理工大学硕士学位论文

初级视皮层启发的鲁棒深度卷积神经网络研究

学位申请人姓名：　董占国

导师姓名及职称：　柯铭 教授　王路斌 副研究员

王刚 副研究员

培　养　单　位：　计算机与通信学院

专　业　名　称：　软件工程

论文提交日期：　2023 年 3 月 20 日

论文答辩日期：　2023 年 5 月 26 日

答辩委员会主席：　火久元　教授

Research on Robust Deep Convolution Neural Network Inspired by the Primary Visual Cortex

by

DONG Zhanguo

B.E.(Beijing Information Science and Technology University)2018

A thesis submitted in partial satisfaction of the

Requirements for the degree of

Master of Engineering

in

Software Engineering

in the

School of Computer and Communication

of

Lanzhou University of Technology

Supervisor

Professor KE Ming

Associate Professor WANG Lubin

Associate Professor WANG Gang

目　录

摘 要

深度卷积神经网络（Deep Convolutional Neural Network，DCNN）作为深度学习的核心，在许多视觉任务中取得了令人瞩目的成绩。但是，目前高级的目标识别 DCNN 模型却非常容易被人为设计的微小扰动欺骗，并且难以识别损坏图像中的目标。这种现象称为对抗攻击。然而，人类视觉系统可以轻易识别这些损坏图片，说明人类视觉系统可以有效排除对抗扰动带来的影响。很多研究指出将生物视觉机制与 DCNN 融合是提高 DCNN 模型鲁棒性很有前景的研究方向。初级视皮层（V1 区）是脑皮层信息处理的关键脑区，包含多种简单细胞方位选择感受野，对应的简单细胞能够对不同类型低层特征产生特异性响应。神经科学领域的研究发现很多高斯卷积核与视觉系统的神经元感受野特性比较吻合。其中，椭圆形的各向异性高斯核更符合 V1 区长条形状感受野，且能够提取高信噪比图像低层特征。于是，此研究借助多尺度各向异性高斯核，将 V1 区简单细胞方位选择感受野引入到 DCNN 的前端低层，提出了初级视皮层启发的鲁棒 DCNN 模型。研究的主要工作和创新点如下：

（1）此研究构建了一个初级视皮层启发的混合架构 DCNN 模型，能够显著提高模型的对抗鲁棒性。模型包含一个方位选择感受野启发的前端和一个标准 CNN 层搭建的神经网络后端。前端由三层组成：包含多种生物可信卷积的卷积层、包含简单细胞和复杂细胞非线性的非线性层和包含 V1 神经噪声生成器的 V1 神经随机层。前端是模型的主要优势，也是对抗鲁棒性的主要来源。前端对图像的处理近似 V1 区对图像的处理。而且，前端不会为模型带来额外的参数量，能够以较小的额外训练成本实现较大的对抗鲁棒性收益。

（2）在 CIFAR-10、CIFAR-100、Mini-ImageNet 和 ImageNet 上的对抗鲁棒性分析实验中，研究提出模型的性能远超基线模型，部分实验超过最先进的 V1 启发模型 VOneNet。此外，与基于训练过程的数据增强结合也能够进一步提高模型的对抗鲁棒性。

（3）在 Mini-ImageNet 上的消融研究验证了研究提出的前端对模型鲁棒性的贡献。生物可信卷积与 V1 神经随机性组合，共同提高模型的对抗鲁棒性。而且，在模型可解释性分析中，研究提出的模型对分类关键特征的定位更准确，同时也表现出对边缘特征和线性特征的偏向。

关键词：鲁棒深度卷积神经网络；对抗鲁棒性；初级视皮层；目标识别

Abstract

Deep Convolutional Neural Networks (DCNNs) are the core of deep learning, which have shown successful results for vision tasks. However, current advanced DCNN models for object recognition can be easily bewildered by imperceptibly small and explicitly crafted perturbations, and can hardly recognize objects in corrupted images. This phenomenon is called adversarial attack. Although, humans have no trouble with such corrupted images, indicating that human visual systems can effectively suppress the inference of adversarial perturbations. Many studies pointed out that the fusion of biological vision mechanisms and DCNN is a promising way to improve the robustness of the model. The primary visual cortex (V1) is a key brain region for cortical visual information processing and contains a variety of simple cell orientation-selective receptive fields. The corresponding simple cells can respond specifically to different types of low-level features. Research in the neuroscience field found that many Gaussian convolution kernels are more consistent with the receptive field characteristics of the visual system. Among them, the elliptical anisotropic Gaussian kernels are more similar with the long-strip receptive fields in V1 and can extract the low-level features of the image with a high signal-to-noise ratio. Thus, with the help of multi-scale anisotropic Gaussian kernels, this thesis introduces simple cell orientation-selective receptive fields into the front layer of the DCNN and proposes a robust DCNN model inspired by the Primary visual cortex. The main works and innovations of this thesis are listed as followed:

(1) This thesis develops a hybrid DCNN model inspired by the primary visual cortex, which can significantly improve the adversarial robustness of the model. The model consists of a front-end inspired by orientation-selective receptive fields and a neural network back-end built from standard CNN layers. The front-end contains three layers: a convolution layer with multiple bio-credible convolutions, a nonlinear layer with simple cell and complex cell nonlinearity, and a V1 neuronal stochasticity layer with a V1 neural noise generator. The front-end is the advantage of the model and the source of

adversarial robustness. The image processing of the front-end approximates the V1. Moreover, the front-end does not increase the parameter of the model and can obtain larger gains of adversarial robustness with small additional training costs.

(2) In the adversarial robustness analysis experiments on CIFAR-10, CIFAR-100, Mini-ImageNet, and ImageNet, the performance of the model in this thesis far exceeds that of the baseline model, and outperforms the state-of-the-art V1-inspired model VOneNet in some experiments. Additionally, the combination with the training-based data augmentation can further improve the adversarial robustness of the model.

(3) Ablation studies on Mini-ImageNet validate the contribution of the front-end proposed in this thesis to model robustness. The combination of bio-credible convolutions and V1 neural stochasticity improves the adversarial robustness jointly. In the model interpretability analysis, the model in this thesis locates the key features for classification with more accuracy and shows a bias toward edge and line features.

Keywords： Robust Deep Convolutional Neural Networks; Adversarial Robustness; Primary Visual Cortex; Object Recognition

插图索引

附表索引

第1章 绪 论

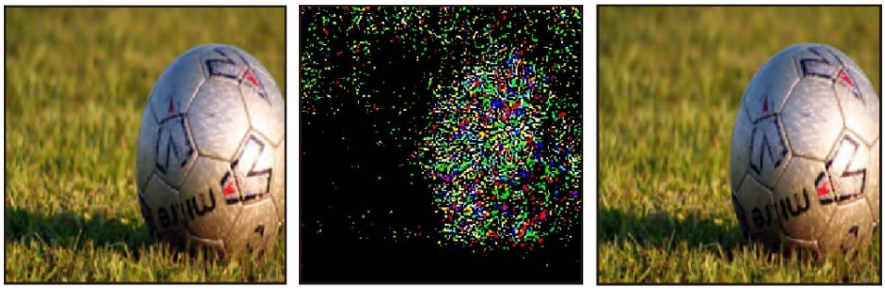
1.1 研究背景

近年来，以深度卷积神经网络（Deep Convolutional Neural Networks，DCNNs）为基础的深度学习（Deep Learning， DL）成为人工智能（Artificial Intelligence， AI）研究的热点，被广泛应用于各种视觉任务，例如基于视觉的汽车自动驾驶[1]、基于人脸识别的人脸支付[2]、基于视频目标检测的行人识别[3]等。在过去的十几年中，深度学习一直主导着目标识别领域[4]，很多经典的目标识别 DCNN 模型也相继被提出，如 AlexNet[5]，VGG[6]，ResNet[7]等。在一些特定的目标识别任务中，深度学习模型取得了令人瞩目的成绩，表现甚至超过人类[8]。但是，这些精巧的 DCNN 模型却非常容易被精心设计的微小对抗扰动欺骗从而输出错误的模式判别结果。这种现象被称为对抗攻击（Adversarial Attack）[9]。这种精心设计的微小扰动被称为对抗扰动（Adversarial Perturbations），针对对抗攻击进行的防御被称为对抗防御（Adversarial Defense）。对抗攻击会严重降低 DCNN 模型的性能，甚至使其完全失效，且一些高级的DCNN模型（如ResNet，VGG等）也无法抵御对抗攻击。此外，对抗攻击现象普遍存在，不仅仅是对象识别，语音识别、自然语言处理[10]等领域也存在对应的对抗攻击现象。对抗攻击会极大威胁基于深度学习的计算机视觉系统的安全性，提高 DCNN 对对抗扰动的鲁棒性已成为深度学习领域的重点研究方向[11-13]。

自 DCNN 模型的对抗攻击现象被发现后，涌现出许多对抗攻击算法，如快速梯度符号法（Fast Gradient Sign Method，FGSM）[14]、投影梯度下降（Projected Gradient Descent，PGD）[13]等。相应的，许多对抗防御算法也被提出，如对抗训练、防御蒸馏、随机梯度裁剪、附加防御网络等[15]。在应对诸如 ImageNet[16]等大尺寸图像分类任务时，虽然没有神经生物学先验知识的 DCNN 在对抗鲁棒性方面有很大的提升，但融合神经科学先验知识使 DCNN 的表现更加接近灵长类视觉依然是深度学习领域研究的热点[17-20]。

如图 1.1 所示，原始样本从 ImageNet 数据集中随机选取，其真实标签和在ImageNet上预训练的ResNet模型的分类结果均为足球。对原始样本添加对抗扰动后生成对抗样本，预训练的 Resnet 模型将其误分类为橄榄球。相较而言，人类视觉系统却几乎察觉不到原始样本与对抗样本的区别，说明人类视觉系统能够

很好的排除对抗扰动带来的影响。因此可将人类视觉或者灵长类视觉的某些视觉机制引入 DCNN 模型，以提高 DCNN 模型对对抗扰动的鲁棒性。



(a) 原始样本　(b) 对抗扰动　(c) 对抗样本

图 1.1 对抗样本示例

将灵长类视觉机制引入 DCNN 模型是使其获得更加接近灵长类视觉系统对对抗扰动的鲁棒性最具前景的研究方向之一[21,22]。对比原始样本和对抗样本，二者没有视觉感知上的差异，说明人类视觉系统能够有效过滤对抗扰动。同样，对抗扰动和目标类别也没有视觉感知上的一致性，表明基于任务的 DCNN 模式判别所依赖的视觉特征与人类视觉系统所使用的视觉特征不完全相同[23]。但是，研究表明一些 DCNN 模型已经成功预测灵长类腹侧流多个阶段的神经响应[20,24]。此外，Dapello 等人[17]指出每个 DCNN 模型解释初级视皮层神经响应的能力与其对对抗扰动的鲁棒性密切相关，即 DCNN 模型越具“生物性”，其对图像损坏和对抗扰动噪声越鲁棒。

初级视皮层（Primary Visual Cortex，V1 区）是脑皮层信息处理的关键脑区，是连接皮层下通路（包括视网膜和外膝体）和更高级视觉皮层（如 V4 区）的中枢[25]。V1 区有多种类型的细胞感受野，包括椭圆形的简单细胞方位选择感受野，其对应的简单细胞能够对边缘、线、点等低层特征产生特异性响应[26]。在计算视觉领域，学者们提出了很多图像特征低层提取的卷积核，包括各向同性和各向异性高斯核。从形态上来看，椭圆形的各向异性高斯核更加符合长条状的 V1 区简单细胞方位选择感受野。此外，神经科学领域的很多研究发现，很多高斯卷积核与视觉系统的神经元感受野特性比较吻合。于是，此研究中通过多尺度各项异性高斯核，将 V1 区简单细胞方位选择感受野引入 DCNN 模型，使其更具“生物性”，进而获得更接近灵长类视觉的对抗鲁棒性。

1.2 对抗攻击与防御研究现状

1.2.1 对抗攻击背景知识

对抗攻击是发生在 DCNN 推理阶段的攻击行为[27]。在图像分类任务中，对于给定的 DCNN 模型$f(x)=y$，$x \in R^m$为模型输入，$y \in Y$为针对当前输入x的输

出。对抗攻击即在目标 DCNN 模型$f(\cdot)$的输入x上添加一个小的噪声ϵ，使得$f(x+\epsilon)\neq f(x)$；为了保证噪声数据对图像的改动足够小，大多数对抗攻击算法会使用$\|\epsilon\|_0$、$\|\epsilon\|_2$或$\|\epsilon\|_\infty$来约束对抗扰动噪声ϵ。用$\|\epsilon\|_\infty$来约束对抗扰动大小时，无目标攻击下的对抗样本生成问题描述如下：

$$\begin{aligned} &f(x+\epsilon)\neq y; x+\epsilon\in R^m \\ &Minimize\ \|\epsilon\|_\infty \end{aligned} \tag{1.1}$$

其中：y是原始类别；x为原始输入样本；ϵ为对抗扰动噪声；$x+\epsilon$表示对原始样本添加对抗扰动噪声后生成的对抗样本。目标攻击下的对抗样本生成问题描述如下：

$$\begin{aligned} &f(x+\epsilon)=y^t; x+\epsilon\in R^m \\ &Minimize\ \|\epsilon\|_\infty \end{aligned} \tag{1.2}$$

其中：y^t是目标类别，其他参数含义与公式 1.1 相同。

下面介绍一下对抗攻击分类。按照攻击者所能获得的模型信息，对抗攻击可划分为白盒对抗攻击、灰盒对抗攻击和黑盒对抗攻击。在白盒攻击中，攻击者能够获得目标模型的所有参数，比如数据集、训练方式、训练参数、模型结构、模型权重等。在灰盒攻击中，攻击者仅能获取目标模型的部分信息。而对于黑盒模型，攻击者则无法得到目标模型的任何参数。根据攻击目的的不同，可将对抗攻击分为目标攻击和非目标攻击。目标攻击希望对抗样本被目标模型错分至特定类别，非目标攻击则希望对抗样本被目标模型分类至除原始类别外的其他任意类别。根据对抗扰动作用的对象，可将对抗攻击分为针对特定图像的特定攻击和针对整个数据集中所有图像的通用攻击。从模型输入对抗样本的方式，可将对抗攻击分为数字攻击和物理攻击。

1.2.2 对抗攻击研究现状

2013 年，Szegedy 等人发现 DCNN 的对抗攻击现象，并提出了 Box-constrained L-BFGS（Box-Constrained Limited-memory BFGS）的对抗攻击算法[27]。Box-constrained L-BFGS 是第一个对抗攻击算法。该算法首次将计算对抗样本的过程抽象为一个凸优化问题，是重要的基于优化的对抗攻击算法。另一个基于优化的对抗攻击算法是 Carlini 和 Wagner 提出的 C&W（Carlini&Wagner）算法[9]，由 Box-constrained L-BFGS 算法改进而来，是目前比较强大的对抗攻击算法。该算法成功攻破 Papernot 等人提出的防御蒸馏（Defensive Distillation）网络[28]。

2014 年，Goodfellow 等人[14]在其研究中指出，对抗样本的产生是由于 DCNN 模型局部空间的线性性质导致，并提出了快速梯度符号法（Fast Gradient

Sign Method，FGSM）。FGSM 算法通过将对抗扰动变化量沿着模型损失梯度上升的方向移动，从而使损失函数增大，达到让模型误分类的目的。FGSM 是经典的基于梯度的对抗攻击算法，在其基础上发展出了很多基于梯度的对抗攻击算法，例如 BIM（Basic Iterative Method）算法[29]，PGD（Projected Gradient Descent）算法[13]，动量迭代的 FGSM（Momentum Iterative FGSM）算法[30]、多样性攻击算法[31]等。FGSM 是单步的对抗攻击算法，其优点是计算速度快，迁移性好，但攻击成功率较低。

为解决 FGSM 攻击成功率低的痛点，Kurakin 等人[29]提出迭代的 FGSM 算法 BIM，该方法以较小的步长执行 FGSM，能够生成攻击能力更强的对抗样本。2017 年，Madry 等人[13]提出 PGD 攻击算法，是目前公认的最强的白盒攻击算法，也是用于评估模型鲁棒性测试的基准算法之一。PGD 可以认为是 BIM 的广义形式，本质上也是迭代的 FGSM 算法。

Moosavi-Dezfooli 等人[32]对模型的决策边界分析后提出一种精确计算对抗扰动的 DeepFool 方法，该方法从几何学角度出发，考虑分类问题的超平面，将样本点不断向距离最近的分类超平面移动，直到跨过分类超平面，使模型决策错误。

Baluja 等人利用生成式神经网络生成对抗样本，并设计了 ATN（Adversarial Transformation Network）网络来生成对抗样本[27]。ATN 方法生成对抗样本速度快，生成的对抗样本攻击力强，但迁移攻击能力弱。Hayes 等人[33]则是提出了基于神经网络生成通用对抗扰动的 UAN（Universal Adversarial Network）攻击算法。UAN 攻击通过训练一个简单的反卷积神经网络将一个自然分布$N(0,1)^{100}$上采样的随机噪声转换为通用对抗扰动。此外，Xiao 等人[34]在神经网络生成攻击算法的基础之上，首次引入了生成式对抗网络（Generative Adversarial Network，GAN）的思想，提出了包含生成器、鉴别器和攻击目标模型的 AdvGAN。经过训练的 AdvGAN 网络可以将随机噪声转换为有效的对抗样本。

白盒对抗攻击算法虽然攻击性强，但在现实中 DCNN 模型相关的信息很难获取，所以研究者们也提出了很多黑盒对抗攻击算法。常见的黑盒攻击算法主要包括两类：一类是基于查询的算法，对某模型构造特定的输入，然后根据模型的反馈不断迭代修改输入，比较经典算法的是单像素攻击（Single Pixel Attack）算法、本地搜索攻击（Local Search Attack）算法、Square Attack 等[35]；另一类是基于迁移学习的方法，即用白盒算法攻击类似的开源模型生成对抗样本，再用生成的对抗样本进行相同任务的黑盒攻击。

除了数字空间的对抗样本外，还有物理空间的对抗样本。Sharif 等人[36]针

对部署在真实场景下的人脸识别系统进行了对抗攻击测试，并提出一种对抗贴纸的攻击方式，将贴纸粘贴在眼镜框来达到对抗攻击的目的。Eykholt 等人[37]设计的针对复杂物理场景下的深度学习系统进行攻击的 RP2（Robust Physical Perturbations）算法，在针对自动驾驶车辆的道路交通标志识别攻击测试中，将生成的对抗扰动打印并粘贴在被攻击的道路交通标志上，成功使自动驾驶系统错误识别该标志。

1.2.3 对抗防御研究现状

随着对抗攻击算法的不断更新和发展，越来越来越多的研究者将目光聚焦于如何提高模型对对抗扰动的鲁棒性，一些优秀的对抗防御算法也随之提出。对抗防御是提高 DCNN 模型安全性的一种手段。对抗防御大致可分为数据层面的防御和模型层面的防御。

数据层面的防御即数据处理，在图像输入网络之前对图像进行处理，从而减小甚至消除对抗扰动带来的影响。基于数据处理的对抗防御优点是计算开销小，不需要修改网络结构，但基于数据处理的防御策略所带来的防御力非常有限。在图像输入模型之前，对图像进行随机裁剪、翻转、缩放等图像变换操作，能够减少样本中的对抗扰动[38]。对抗样本通过JPEG压缩方法，在一定程度上降低对抗扰动带来的影响。Liu 等人[39]通过重新设计标准的 JPEG 压缩算法来抵御对抗样本，该方法又被叫做特征蒸馏（Feature Distillation）。Jia 等人[40]提出了一种使用图像压缩网络的方法，进一步消除对抗扰动。该方法通过预先压缩对抗样本，以实现对抗扰动的消除，然后通过重构方法恢复干净样本。

Raffle 等人[41]提出了一种综合防御方法，通过集成多种单一的对抗防御方法来抵御对抗扰动。他们在图像输入网络之前，通过应用一系列变换，如 JPEG 压缩、小波去噪、非局部均值滤波、位深度缩减等方法来消除对抗样本中的扰动。Prakash 等人[42]提出了像素偏转法，该方法通过随机选择少量像素并替换为领域内随机选取的像素来抵御对抗攻击，并使用小波去噪消除替换后的像素值产生的噪声。Mustatf 等人[43]提出一种基于超分辨率的防御方法，能够将流形边缘的样本重新映射到自然图像流形上。该方法不需要进行任何模型训练即可增强图像质量，并保持模型在原始图像上的分类正确率。Osadchy 等人[44]认为对抗扰动是一种噪声，可以利用滤波器将其消除，于是设计了以像素为目标的去噪器（Pixel-guided Denoiser）来消除抗扰动带来的影响。Liao 等人[45]在此基础上进一步提出了以高阶表征为目标的去噪器（High-guided denoiser）。

GoodFellow 等人[14]最早提出对抗训练的概念。对抗训练是一种直观的防御方法，其核心思想是将对抗攻击算法生成的对抗样本添加到训练集并用于模型的

训练。对抗训练可以提高模型针对特定对抗扰动的鲁棒性，且够大幅度提高模型对特定攻击类型的防御力，但对其他对抗攻击算法失效。Kurakin等人[29]通过新的训练策略将模型扩大到更大的训练集，并通过批量标准化来提高对抗训练效率。对抗训练存在标签泄露的问题，因此生成对抗样本时需要多次计算输入图像的梯度，训练成本非常高。Moosavi-Dezfooli 等人[32]指出，对抗训练只能使模型对训练集中的对抗样本有很好的鲁棒性，但对抗样本无法穷尽，因此无论添加多少对抗样本，依然存在新的对抗样本使模型出错。

一些随机化的方法也能提高 DCNN 模型的对抗鲁棒性。Xie 等人[46]提出了两种随机变换：随机调整大小和随机填充，以减轻模型推理时对抗扰动的影响。这些变换可以在对抗样本和原始样本上使用，从而增加模型对对抗扰动的鲁棒性。Guo 等人[47]的防御方法则是在将图像输入 DCNN 模型之前使用随机性的图像变换，包括位深度减小、JPEG 压缩、总方差最小化和图像缝合等，以减轻对抗攻击的影响。上述两种方法能够很好的抵御一些黑盒攻击算法，但均无法抵御转换期望（Expectation over Transformation，EoT）算法[48]。Dhillon 等人[49]提出了一种随机激活修剪（Stochastic Activation Pruning，SAP）的方法进行对抗防御。除了随机化，降噪也是简单有效的方法。Xu 等人[12]首先利用两种压缩（去噪）方法：位减少和图像模糊，以减少自由度并消除对抗扰动。

相较于数据层面的防御方法，模型层面的防御策略能带来更强的防御，但其迁移性能通常较差。模型层面的防御主要通过修改模型结构、基于模型梯度的防御、使用附加网络等方法来实现。

修改模型结构即优化模型网络架构、参数等，使模型本身更具鲁棒性，以抵御对抗攻击。DCNN 输出对输入的梯度幅度过大是造成其过于敏感的原因。Ross 等人[50]提出梯度正则化的方法来提升网络的对抗鲁棒性。在网络的训练过程中，惩罚输出相对于输入的变化幅度，使得输出对于输入的敏感性降低，从而达到隐藏梯度的效果。混淆梯度（Obfuscated Gradients）使基于梯度的攻击失效。混淆梯度有三种类型，分别是破碎梯度、随机梯度、消失/爆炸梯度。破碎梯度添加了一个不可微的预处理，使攻击者无法找到用于攻击的梯度；随机梯度训练一组分类器，在预测时随机选取一个分类器预测，由于攻击者不知道网络模型选择那个分类器，因此攻击的成功率很低；消失和爆炸梯度将一层计算的输出作为下一层计算的输入，因此来自每一层的偏导数累计乘积使得偏导数极大或者极小，攻击者无法将其用于对抗样本的生成。然而，基于混淆梯度的防御能够被规避。Athalya 等人[51]利用梯度近似方法，成功打破 ICLR2018 中 9 种依赖混淆梯度防御的 7 种防御方法。

防御蒸馏（Defensive Distillation）基于对模型做平滑输出的思想实现。蒸

馏是由 Hinton 等人[52]提出的一种知识迁移方法，将复杂模型的知识迁移到较小的模型上，在压缩神经网络的同时保持其预测准确性。Papernote 等人[28]提出的防御蒸馏，通过将原始训练集的标签替换为教师模型的预测概率分布来完成对学生模型的训练。通过训练的软标签，模型的分类边界更加平滑，模型的敏感性降低，从而提高模型对对抗扰动的鲁棒性。

Lee 等人通过生成对抗网络（Generative Adversarial Networks，GAN）中的生成器和判别器的相互博弈来构造分类网络，生成网络用来生成对抗样本，判别网络用来分类，从而使模型鲁棒性提高[38]。GAN 网络并不会对原网络产生影响。在充分训练的前提下，基于 GAN 的防御方法可以有效防御对抗攻击。Samangouei 等人[53]提出的 Defense-GAN 防御策略，使用生成模型将可能存在的对抗样本投影到干净的数据流形上，然后再对其分类。生成模型可认为是将对抗样本转化为干净样本的转化器。

1.3 融合神经计算的鲁棒深度卷积神经网络研究现状

DCNN 已经被广泛应用于物体分类等视觉任务中[54]，但是 DCNN 模型本身存在“大数据-小任务”的问题，有时甚至面临“少数据-小任务”的矛盾，在可解释性、泛化能力和抗干扰能力方面存在局限性[55]。而且，DCNN 模型在面一些常见的图像损坏和针对 DCNN 专门设计的对抗扰动噪声时，其性能急剧下降。

许多学者在对抗攻击与对抗防御的研究中，提出了很多提高 DCNN 模型对抗鲁棒性的方法，如噪声数据增广、对抗训练、随机特征修剪、特征去噪等[15]。由于灵长类视觉系统几乎不受噪声（包括随机噪声、固定模式噪声等）或对抗扰动（包括白盒和黑盒攻击）的影响，因此将灵长类视觉机制引入 DCNN 模型使其获取更加接近灵长类视觉的性能越来越受到深度学习领域的关注[56]。

生物约束是使 DCNN 模型获得更加接近灵长类视觉性能的最具前景的方法。Kubilius 等人[57]认为虽然 DCNN 模型一定程度上模拟了视觉腹侧通路的物体识别机制，但是在架构和连接机制方面与生物视觉系统还有较大差距，因此该团队借助基本循环网络模块搭建了一个浅层（四层）网络，分别模拟 V1、V2、V4 和 IT 脑区，在 ImageNet 数据集上获得了较好的性能，而且与腹侧通路的响应规律更为接近。尽管该工作对生物启发的神经网络进行了积极探索，但是对于视网膜、外膝体特别是 V1 区的模拟还比较粗糙，在 DCNN 鲁棒性方面的工作还比较有限。Vuyyuru 等人[18]从视网膜非均匀采样和 V1 多尺度感受野机制出发，对输入图像进行了非均匀采样和多尺度高斯下采样，并分别使用 ResNet[7]模块进行特征提取，证明了非均匀变换和高斯多尺度变换能够提高网络的抗攻击能力。Lindeberg[58]发现主流 DCNN 模型不具有尺度协变性，因为通过重复使用固定尺

寸的卷积核滤波以及邻域池化操作，使得深层网络的特征产生固定间隔的非线性而不是尺度空间协变性，因此提出了级联多尺度各项异性高斯卷积核的多层网络，在理论上实现了多层网络的尺度协变性，同时又具有可解释性和抗噪性，在纹理分类任务上取得了较好性能。Li 等人[59]指出，正常的正则化训练不能使 DCNN 模型学习到有限样本数据之外的数据分布，并通过小鼠视觉皮层神经元对复杂自然场景的神经响应数据来正则化模型，从而使卷积特征接近神经活动的特性。实验表明由生物神经元响应正则化的 DCNN 模型可以有效消除对抗性扰动带来的影响。

Safarani 等人[60]联合训练了一个 DCNN 模型来执行图像分类任务，在预测猕猴 V1 对相同自然刺激下的神经活动的同时，成功利用了这些诱导偏差。与猕猴 V1 数据的联合训练可以提高 DCNN 模型对常见图像失真的鲁棒性，虽然 DCNN 模型在训练期间没有见过这些图像这些失真。Dapello 等人[17]发现，DCNN 的白盒对抗鲁棒性与灵长类 V1 的解释方差密切相关。基于这一发现，他们提出了一种混合架构的 DCNN 模型——VOneNet。VOneNet 由灵长类 V1 约束的前端和可训练的基础 DCNN 后端组成。VOneNet 可以显著提高模型对图像损坏和对抗性扰动的鲁棒性，同时保持原始图像分类准确率。

1.4 此研究的研究内容和组织结构

1.4.1 此研究的研究内容及意义

针对严重威胁 DCNN 安全性的对抗攻击问题，此研究在前人的研究基础之上，借鉴 V1 区视觉信息加工处理机制，通过多尺度各向异性高斯核将 V1 区简单细胞方位选择感受野引入 DCNN 模型，提出 V1 区启发的 DCNN 模型，使 DCNN 获得更加接近灵长类视觉的鲁棒性。最后，再借助消融实验和模型可解释性技术，分析并验证此研究提出的方位选择感受野启发的前端的有效性。此研究的主要研究内容如下：

（1）此研究借助多尺度各项异性高斯核，将 V1 区简单细胞方位选择感受野引入到 DCNN 的前端低层，提出了初级视皮层启发的鲁棒 DCNN 模型，在扩增 DCNN 低层特征层的同时，提高 DCNN 模型的对抗鲁棒性。

（2）在 CIFAR-10、CIFAR-100、Mini-ImageNet 和 ImageNet 数据集上的模型对抗鲁棒性分析实验中，此研究提出模型性能显著高于基线模型，在某些测试中甚至超过 VOneNet。为进一步提高模型的性能，研究中将基于训练过程的数据增强与研究提出的方法结合。

（3）此研究设计的消融实验验证了研究提出的前端能够提高模型的对抗鲁棒性，同时也证明了一系列生物可信卷积滤波器对模型的贡献。此外，研究中通

过模型可解释技术对模型进行可视化分析，也证明了研究提出模型对边缘、线条等低层特征的偏向。

此研究的研究成果在图像去噪、生物医学图像分析以及对象分类等计算机视觉任务中具有良好的应用前景，而且对改善 DCNN 视觉系统的安全性和稳定性有重要理论指导价值和实际应用意义。

此研究来自于国家自然科学青年基金《初级视皮层启发的各向异性高斯核极其在鲁棒深度卷积神经网络中的应用》和国家自然科学基金地区基金《基于磁共振成像的原发性癫痫脑网络相机故障模型研究》。

1.4.2 此研究的组织结构

此研究主要由五个章节构成：

第一章：绪论。包括此研究的背景，对抗攻击的相关基础知识，对抗攻击、对抗防御的研究现状，融合神经计算的鲁棒深度卷积神经网络研究现状，以及此研究的研究内容与研究意义。

第二章：初级视皮层方位选择感受野的模型。介绍了灵长类视觉及 V1 区细胞的方位选择性，同时介绍了作为 V1 区方位选择感受野模型的多尺度各向异性高斯核及其在图像低层特征提取方面的优势。

第三章：初级视皮层启发的鲁棒 DCNN 模型。介绍了研究提出的 DCNN 模型架构、优势及其各个组件，同时也详细介绍了如何通过数学模型将生物可信机制融合到 DCNN 前端低层。

第四章：对抗鲁棒性分析与模型优化。介绍了实验数据集、训练方式、训练超参数等模型相关的详细信息，介绍对抗鲁棒性评估指标的同时进行了一系列鲁棒性分析实验。此外，还将研究提出模型与基于训练过程的数据增强方法结合，以进一步优化模型性能。

第五章：消融实验与基于模型可解释性的模型分析。通过消融实验，验证了研究提出的包含 V1 区方位选择感受野的前端的有效性与生物可信卷积对模型对抗鲁棒性的贡献，并通过模型可解释性技术对模型进行了可视化分析，验证了此研究提出的模型对边缘、线条等低层特征的偏向。

总结与展望：分析了此研究的不足之处和未来的发展方向。

第2章 初级视皮层方位选择感受野模型

V1 区作为人类视觉系统中承上启下的脑区，其地位至关重要。在上世纪五六十年代，就开始了关于 V1 区的研究。V1 区的研究在源源不断提供新知识的同时，还启发了人类对大脑工作机制的理解。“方位选择性”是 V1 区细胞的一个重要特性。计算视觉领域的研究发现，椭圆形各向异性高斯核更加符合部分 V1 区简单细胞方位选择感受野的表观特性，且高斯卷积核具备方位调谐和尺度适应性，能实现鲁棒的图像低层特征提取。本章将在简要介绍人类视觉系统的同时，详细阐述 V1 区细胞的“方位选择性”机制，并进一步介绍 V1 区方位选择感受野与各向异性高斯核，同时阐述作为 V1 区方位选择感受野模型的各向异性高斯核在图像低层特征提取方面的优势。

2.1 人类视觉系统与初级视皮层

视觉系统是人类最为重要的感觉系统，人类大脑皮层三分之一的区域与视觉有关。人类接收的外部信息，视觉信息占绝大部分。且视觉系统强烈的影响人类的认知、决策、情绪乃至于潜意识的活动。

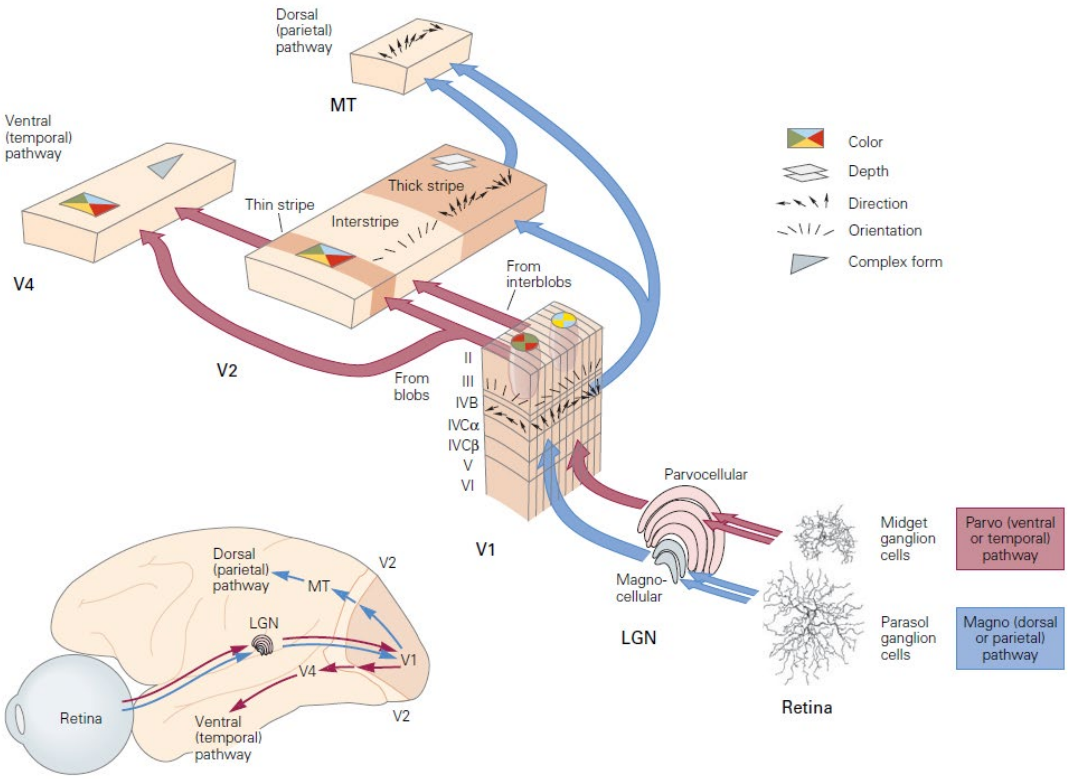


图 2.1 人类视觉系统及初级视皮层位置示意图

初级视皮层在人类视觉系统中的位置，如图 2.1 所示。人类视觉系统主要由眼球、视神经、外侧膝状体、初级视皮层，高级视皮层（腹侧视觉皮层、背侧视觉皮层）等组成。人类视觉系统的起点是眼球。光包含着外部世界的各种结构信息，经过视觉系统的光学折射系统（如晶状体、玻璃体等），投射到眼球底部的视网膜上。视网膜上的光感受器细胞将光信号转化为电信号，然后传递给其他细胞（如双极细胞、水平细胞、无长突细胞等），进行初步的信息整合和加工。多种细胞在视网膜上协同工作，将融合后的视觉信息传递到视网膜神经节细胞，由其将视觉信息通过视神经传递到大脑。

视觉信息进入大脑后，首先被传递到丘脑中的外侧膝状体。在这个小核团中，视觉信息会被进一步处理和整合，提取出其中的关键信息并滤除无用信息。视觉信息经过外侧膝状体的整合和加工后，关键信息被提取并进一步传递到 V1 区。在 V1 区，这些信息被进一步加工和处理，以提取更高级的特征信息，例如方向、运动和颜色等。随后，这些信息被传递到高级视觉皮层。高级视觉皮层包含多个处理特定信息的高级区域，例如运动检测脑区、形状识别脑区和人脸识别脑区等。这些高级区域进一步整合和分析信息，以产生对视觉场景的更深刻理解。视觉系统中的各个单元各司其职，通过从上到下的协同分工，共同完成对视觉信号的处理，并将最终结果传递到其他脑区，从而影响人的行为和思想。

自视觉研究的奠基人 Hubel 和 Wissel[61]上世纪 50-60 年代系统的研究了 V1 区的信息处理机制以来，V1 区就一直为全世界的神经科学家们的关注，并提供源源不断的新知识。同时，V1 区的早期研究也大大促进了人类对大脑工作机制的理解，如皮层功能柱、视觉特征提取、信号的分级处理等概念和思想，在 V1 区的研究中不断被提及，也影响到了神经科学领域的其他研究。

下面介绍一些关键概念。首先是感受野（Receptive Field）[62]。通俗来说，感受野是指视觉细胞所能感觉到原始刺激输入的面积总和。一个视觉细胞的感受野，就是它负责的视野区域总和。形象来讲，视觉细胞的感受野就是该细胞能够“看到”视野区域。视觉细胞仅对感受野内部视觉刺激有明显响应，而感受野以外的区域则无法引起该视觉细胞的明显响应。大部分视觉细胞的感受野都有其固定的结构，感受野内不同的区域，对光的反应有所区别。对于比较低层级的视觉细胞来说，细胞感受野的形状和细胞检测视觉特征的能力息息相关。

然后是方位选择性（Orientation Selectivity）。方位即“空间朝向”（Orientation）。V1 区细胞的一种性质就是“方位选择性”。如果给 V1 区细胞一个不同方位的视觉刺激，当视觉刺激的方位与这个细胞的最优方位重合时，细胞响应达到最强。如果视觉刺激的方位逐渐偏离细胞最优方位，细胞响应就越来越弱。如果视觉刺激的方位和最优方位垂直，此时细胞响应达到最低点。过程如

图 2.2 所示。因此，可以通过观察细胞响应的强弱变化，来得知这个细胞的最优方位。V1 区细胞的上述性质，被称为“方位选择性”。类似的，还存在“运动方向选择性”、“空间频率/时间频率选择性”、“颜色选择性”等等。这些选择性，本质上使得 V1 区具备了检测并编码各种视觉特征的能力。

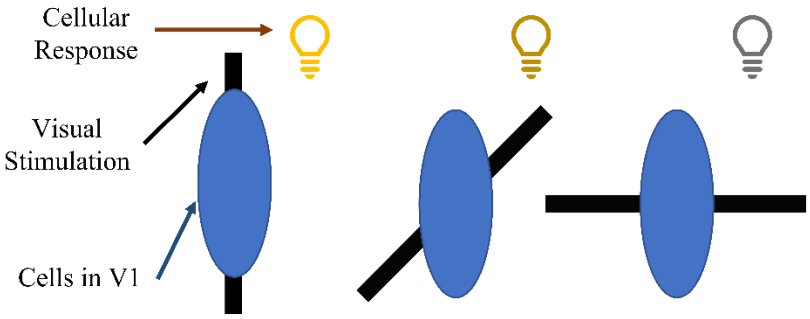


图 2.2 V1 区细胞方位选择性示意图

2.2 初级视皮层的方位选择感受野

V1 区是脑皮层处理视觉信息的关键脑区，是连接皮层下通路（包括视网膜和外膝体）和更高级视皮层（如 V4 区）的中枢[25]，视觉神经科学领域围绕 V1 区的结构和功能进行大量的研究[63]。V1 区存在多种类型细胞感受野，可以提取不同类型的低层特征，加工局部模式、颜色和方位等信息[64]。这些细胞感受野主要分为简单细胞感受野和复杂细胞感受野。1981 年诺贝尔奖获得者 Hubel 和 Wiesel[61]的工作揭示 V1 区典型简单细胞感受野代表包括“奇对称梯度敏感型”方位选择性感受野、“偶对称双侧对比度敏感型”方位选择性感受野和“非对称单侧对比度敏感型”方位选择性感受野等，如图 2.3 所示。相应地，V1 区简单细胞可以分别对边缘、线和点等低层特征产生特异性响应。

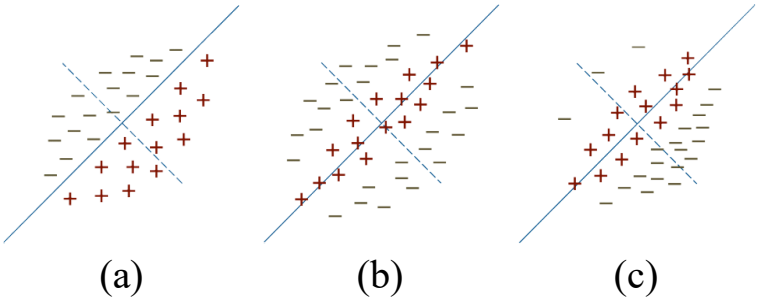


图 2.3 V1 区简单细胞方位选择感受野示意图

（a）“奇对称梯度敏感型”方位选择性感受野特性示意图（“+”表示给光区，“-”表示撤光区）；（b）“偶对称双侧对比度敏感型”方位选择性感受野特性示意图；（c）“非对称单侧对比度敏感型”方位选择感受野特性示意图。

V1 区的基本功能模块是超柱，每个超柱模块存在功能类型相似的感受野，感受野中心基本保持一致，在尺寸和最优方位等方面具有明显规律[65]。如图 2.4（a）所示，在纵向方向，感受野的最优朝向大致相同，但感受野尺寸存在差异；如图 2.4（b）所示，在横向方向，感受野的最优朝向具有连续性。V1 区存在特征聚合机制，能够以非线性方式整合多个超柱的信息加工结果，进而提取特征的方位、尺度和位置等信息。

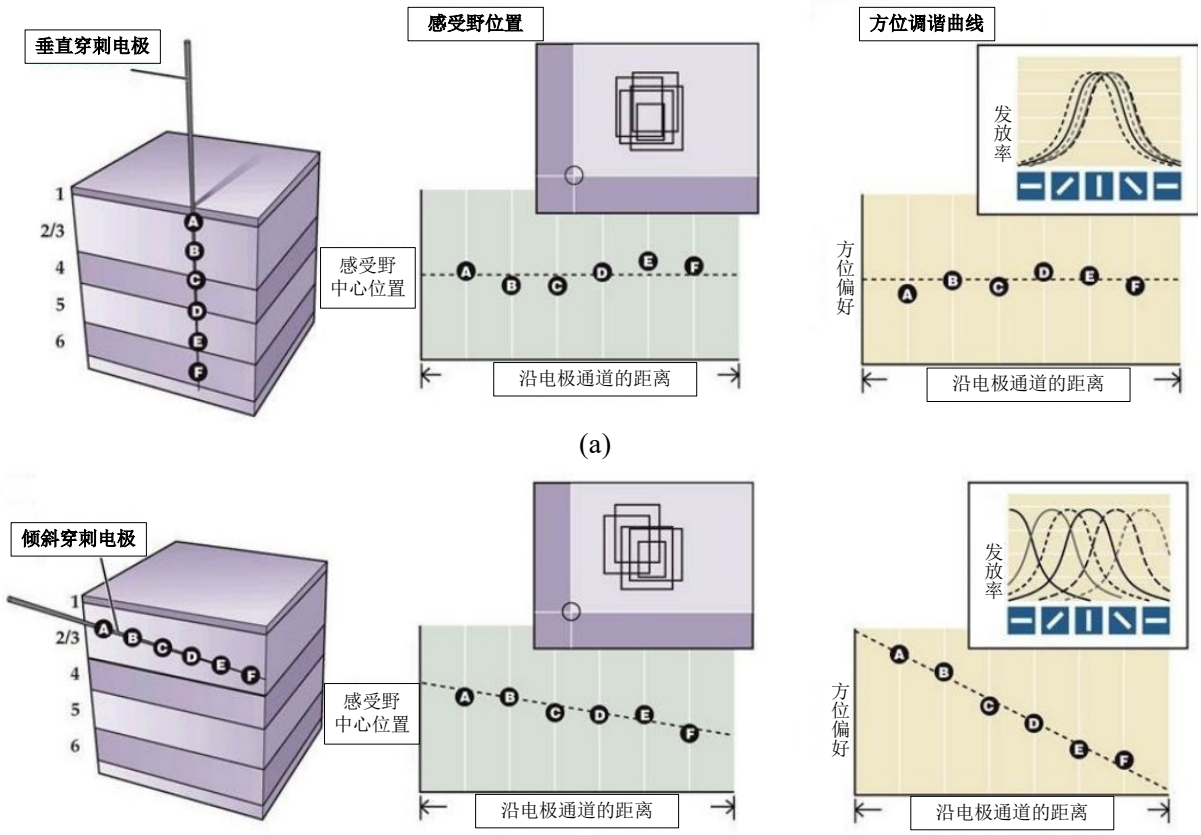


图 2.4 V1 区部分简单细胞的感受野尺寸和方位的空间分布特性

（a）V1 区皮质纵向细胞柱的感受野尺寸和方位偏好规律；（b）V1 区皮质横向细胞群的感受野尺寸和方位偏好变化规律。

2.3 方位选择感受野的数学模型

V1 区的简单细胞方位选择感受野对应的简单细胞能够对边缘、线、点等低层特征产生特异性响应。图像低层特征提取时，图像中的边缘、轮廓、线、点等低层特征具有多变的朝向和尺度，因此要求特征提取卷积核需具备方位调谐和尺度适应性[66]。在计算视觉领域，学者开发了多种面向特征提取的卷积核，其中高斯卷积核已经被证明是尺度空间完备卷积核[67]。此外，神经科学领域研究也

发现很多高斯卷积核与视觉系统的神经元感受野特性比较吻合[68]。高斯卷积核分为各向同性高斯核和各向异性高斯（Anisotropic Gaussian，AG）核。各向同性高斯核的形状通常为圆形（图 2.5 a 和 c），各向异性高斯核的形状一般为椭圆形（图 2.5 b 和 d）。各向异性高斯核的椭圆形状更接近 V1 区长条形状的感受野。

Canny[69]和 Geusebroek[70]指出，与各向同性高斯核相比，各向异性高斯核更符合边缘和线条表观特性。Shui[71]通过引入各向异性因子将一阶导数各向同性高斯核发展为各向异性高斯一阶导数（First-order derivative of Anisotropic Gaussian，FAG）核（图 2.5 b）。在形态方面，FAG 核更符合 V1“奇对称梯度敏感型”方位选择感受野。Wang 等人[72]成功解决了基于 FAG 核彩色图像边缘检测，而 Zhang 等人[73]的研究解决了各项异性因子的自适应计算问题。此外，Wang 等人[74,75]进一步证明了 FAG 卷积核的边缘响应和各向异性因子之间的无关性。受 FAG 的启发，Lopez-Molina 等人[76] 提出了基于各向异性高斯二阶导数（Second-order derivative of Anisotropic Gaussian，SAG）卷积核（图 2.5 d）的线特征提取方法，在形态上符合 V1 区“偶对称双侧对比度敏感型”方位选择感受野。Wang 等人[77]证明了各向异性因子能够显著提高线特征提取的信噪比。

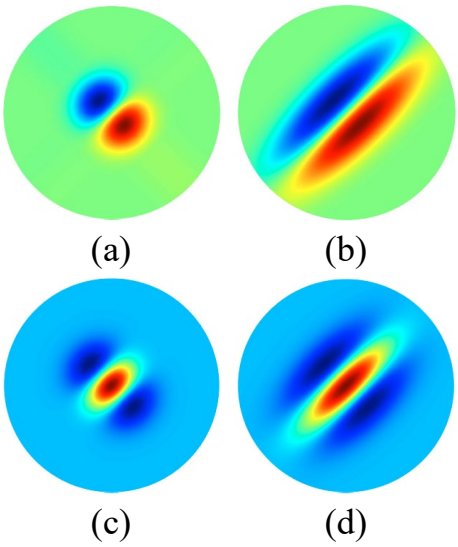


图 2.5 模拟 V1 区方位选择感受野的各向同性、各向异性高斯核

（a）各向同性高斯一阶导数核；（b）各向异性高斯一阶导数核（FAG），其中（a）和（b）采用相同的卷积核尺度；（c）各向同性高斯二阶导数核；（d）各向异性高斯二阶导数核（SAG），其中（c）和（d）采用相同的卷积核尺度。

在此研究中，借助多尺度各向异性高斯核，将 V1 区简单细胞方位选择感受野引入 DCNN 的前端低层，模拟 V1 区简单细胞方位选择感受野的特征提取过

程，在实现鲁棒的低层特征提取的同时丰富 DCNN 的低层特征层，使 DCNN 获得更加接近灵长类视觉的鲁棒性，提高 DCNN 对对抗扰动噪声的鲁棒性。

2.4 本章小结

本章主要介绍人类视觉系统以及视觉系统中的重要一环——初级视皮层。同时，介绍了 V1 区的“方位选择性”机制和 V1 区三种典型的简单细胞方位选择感受野。因为 FAG 核、SAG 核与“奇对称梯度敏感型”、“偶对称双侧对比度敏感型”方位选择感受野具有相似的表观特性，且多尺度的 FAG 核、SAG 核能够提取鲁棒的图像低层特征。因此，在此研究中通过两种各向异性高斯卷积核来模拟 V1 区的简单细胞方位选择感受野，并将其融合到 DCNN 的前端低层，在丰富 DCNN 低层特征层的同时，使 DCNN 更具“生物性”，从而使得 DCNN 获得更加接近灵长类视觉的鲁棒性。

第3章 初级视皮层启发的鲁棒深度卷积神经网络模型

初级视皮层是脑皮层视觉信息处理的关键脑区，针对 V1 区的大量研究不仅解析了 V1 区的视觉机制，同时大大促进了人类对大脑工作机制的理解。DCNN 受大脑的宽松启发，在很多视觉任务中取得了令人瞩目的成绩。针对严重威胁 DCNN 安全性的对抗攻击问题，很多研究者将生物视觉机制引入 DCNN 以增强其对抗鲁棒性。Dapello 等人[17]提出的 VOneNet，是目前最先进的 V1 区启发的鲁棒 DCNN 模型。VoneNet 由灵长类 V1 区宽松约束的前端 VOneBlock 和卷积神经网络后端组成，能够显著提高 DCNN 的对抗鲁棒性。但 VOneNet 的前端只使用 Gabor 滤波器来模拟 V1 区简单细胞响应，对 V1 区的模拟不够精确，尤其是 V1 区的方位选择感受野。此研究通过多尺度各向异性高斯核将部分 V1 区简单细胞方位选择感受野引入到 DCNN 的前端低层，并引入其他生物视觉机制，因此对 V1 区的模拟更精确。理论上，研究构建的模型将比 VOneNet 更能抵御对抗攻击。本章将对此研究构建的初级视皮层启发的鲁棒 DCNN 模型进行详细介绍。

3.1 初级视皮层启发的鲁棒 DCNN 模型

此研究提出了一个混合架构的 DCNN 模型，模型架构如图 3.1 所示。模型由一个 V1 区方位选择感受野启发的前端和一个标准 CNN 层（一般包括卷积层、池化层、归一化层和全连接层）搭建的神经网络后端组成，前端与后端之间通过一个瓶颈层连接。前端包含多种生物视觉机制，在丰富 DCNN 的低层特征层的同时实现图像鲁棒的低层特征提取。前端对图像的处理近似 V1 区对图像的处理。后端则进一步加工图像低层特征，提取高级语义特征，并进行最后的分类决策。

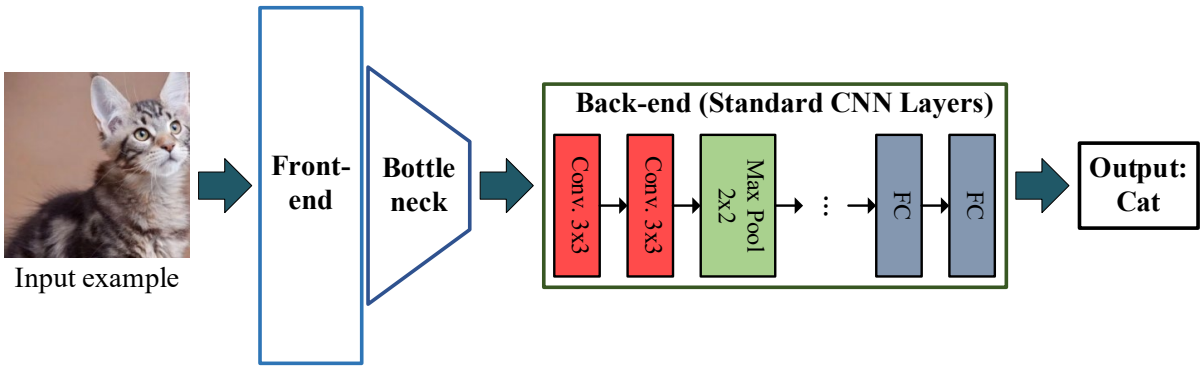


图 3.1 初级视皮层启发的鲁棒 DCNN 模型架构

研究提出的 DCNN 模型的主要优势是 V1 区方位选择感受野启发的前端。前端是模型对抗鲁棒性的主要来源。前端融合了多种生物视觉机制，包括 V1 区简单细胞方位选择感受野、视网膜和 V1 区环绕调制机制（Surround Modulation）、Gabor 滤波器（Gabor 滤波器是模拟初级视皮层简单细胞响应的一种卷积滤波器）。前端由三层构成：参数固定的生物可信卷积层（包括模拟方位选择感受野的各向异性卷积滤波器组，模拟视网膜和 V1 区环绕调制机制的 LoG（Laplace of Gaussian）卷积滤波器组和 Gabor 卷积滤波器组）、非线性层（包括简单细胞非线性和复杂细胞非线性）和 V1 神经随机层（包含一个 V1 神经噪声生成器）。预处理后的 RGB 图像首先到达前端，经过前端的生物可信卷积、非线性处理后，会得到包含多种低层特征的卷积特征图。这些卷积特征图会被堆叠到较高的维度，而 V1 区神经噪声生成器则向每个卷积通道添加独立的高斯噪声，以模拟 V1 区神经细胞的随机性。

在构建此研究提出的模型时，只需用研究中提出的前端替换标准 DCNN 模型的第一层。一般来说，研究设计的前端拥有比标准 DCNN 模型第一层更高的维度，因此前端无法直接与后端网络连接。于是，在研究中，借鉴 VOneNet 的设计，在前端和后端之间增加了一个由1×1卷积构成的瓶颈层。通过瓶颈层，可将前端高维的卷积特征图压缩到与标准 DCNN 模型第一层相同的维度，从而实现前端与后端的连接。此外，由于加入了瓶颈层，研究提出的前端可轻松嵌入到不同的后端网络中，将来自前端的对抗鲁棒性迁移到其他 DCNN 模型。

此研究构建的 DCNN 模型，对模型结构进行了改动，属于模型层面的提高模型对抗鲁棒性的方法。此研究提出的 DCNN 模型与标准 DCNN 模型相比，主要区别在于前端。由于前端的卷积滤波器组是参数固定的卷积滤波器组，因此不会为模型带来额外的参数量。相较于标准的 DCNN 模型，研究提出的模型可训练参数量更少，都来自可训练的卷积神经网络后端。因为前端的维度较标准 DCNN 第一层高，因此研究中提出的 DCNN 模型需要额外的计算成本，但所需的额外计算开销很小。

对比公认的防御力较强的对抗训练策略，此研究提出的 DCNN 模型能够以较小的训练开销实现较大的对抗鲁棒性收益。对抗训练的思想是将对抗样本加入训练集，通过 DCNN 模型的自我学习过程使模型更鲁棒，但训练的模型只能防御来自参与对抗训练的攻击算法的攻击，对其他攻击则失效。且对抗训练需要在训练过程中同步计算对抗样本，需要很高的计算与训练开销。而此研究提出的模型则通过对模型结构的修改，在提高模型的泛化识别能力的同时，也获得了比对抗训练更加广谱的对抗鲁棒性。

3.2 基于多尺度各向异性高斯核的图像低层特征提取

研究中，借助多尺度各向异性卷积核，在 DCNN 模型的低层模拟 V1 区简单细胞的方位选择感受野，增加 DCNN 模型感受野的异质性，在扩增 DCNN 模型的低层特征层的同时实现鲁棒低层特征提取，提高 DCNN 模型对对抗扰动的鲁棒性。此研究采用多尺度、可变各向异性因子和方向调谐的 FAG、SAG 卷积核分别作为边缘和线条特征提取的卷积算子。

FAG、SAG 卷积核来源于高斯函数。高斯函数描述如下：

$$g(\boldsymbol{x};\sigma)=\frac{1}{2\pi\sigma^2}exp\left(-\frac{\boldsymbol{x}^T\boldsymbol{x}}{2\sigma^2}\right) \tag{3.1}$$

其中，$\boldsymbol{x}=[x,y]^T$表示平面坐标系，$\sigma\in\mathbb{R}_+$表示空间尺度的标准差，即尺度。公式 3.1 为零阶高斯函数。对零阶高斯函数求关于$\boldsymbol{x}$的一阶偏导数，即可得到各向同性高斯一阶导数，其描述如下：

$$g'(\boldsymbol{x};\sigma)=-\frac{x}{\sigma^2}\cdot g(\boldsymbol{x};\sigma) \tag{3.2}$$

将公式 3.2 中的各项同性高斯一阶导数旋转一个角度θ，从而得到角度版本的高斯一阶导数：

$$g'(\boldsymbol{x};\sigma,\theta)=-\frac{[\cos\theta,\sin\theta]\boldsymbol{x}}{\sigma^2}\cdot\frac{1}{2\pi\sigma^2}exp\left(-\frac{1}{2\sigma^2}\boldsymbol{x}^T\boldsymbol{R}_\theta^T\boldsymbol{R}_\theta\boldsymbol{x}\right) \tag{3.3}$$

其中，

$$\boldsymbol{R}_\theta=\begin{bmatrix}cos\,\theta & sin\,\theta\\ -sin\,\theta & cos\,\theta\end{bmatrix} \tag{3.4}$$

表示旋转矩阵，$\theta\in[0,2\pi]$表示朝向。

各向异性高斯函数的方向版本定义为[74]：

$$g(\boldsymbol{x};\sigma,\varphi,\theta)=\frac{1}{2\pi\varphi\sigma^2}exp\left(-\frac{1}{2\sigma^2}\boldsymbol{x}^T\boldsymbol{R}_\theta^T\begin{bmatrix}1 & 0\\ 0 & \varphi^{-2}\end{bmatrix}\boldsymbol{R}_\theta\boldsymbol{x}\right) \tag{3.5}$$

其中，$\varphi\geq1$表示各向异性因子。值得注意的是，当$\varphi=1$时，公式 3.5 中的各向异性高斯核会还原为各向同性高斯核。由公式 3.5 和公式 3.3，可得到 FAG 核：

$$g'(\boldsymbol{x};\sigma,\varphi,\theta)=-\frac{[cos\,\theta,sin\,\theta]\boldsymbol{x}}{\sigma^2}\cdot g(\boldsymbol{x};\sigma,\varphi,\theta) \tag{3.6}$$

于是，面向边缘特征提取的 FAG 卷积核描述如下：

$$K_{fag}(\boldsymbol{x};\sigma,\varphi,\theta)=-\frac{[cos\,\theta,sin\,\theta]\boldsymbol{x}}{2\pi\varphi\sigma^4}.exp\left(-\frac{1}{2\sigma^2}\boldsymbol{x}^T\boldsymbol{R}_\theta^T\begin{bmatrix}1 & 0\\ 0 & \varphi^{-2}\end{bmatrix}\boldsymbol{R}_\theta\boldsymbol{x}\right) \tag{3.7}$$

FAG 卷积核的三维空间形态如图 3.2（a）所示。

面向线条特征提取的 SAG 卷积核是各向异性高斯函数对$\boldsymbol{x}$的二阶偏导数，其

描述如下：

$$K_{sag}(\boldsymbol{x};\sigma,\varphi,\theta)=\left(\frac{(x\cos\theta+y\sin\theta)^2}{2\pi\varphi\sigma^6}-\frac{1}{2\pi\varphi\sigma^4}\right)$$

$$\cdot exp\left(-\frac{1}{2\sigma^2}\boldsymbol{x}^T\boldsymbol{R}_\theta^T\begin{bmatrix}1 & 0\\ 0 & \varphi^{-2}\end{bmatrix}\boldsymbol{R}_\theta\boldsymbol{x}\right) \tag{3.8}$$

SAG 卷积核的三维空间形态如图 3.2（b）所示。

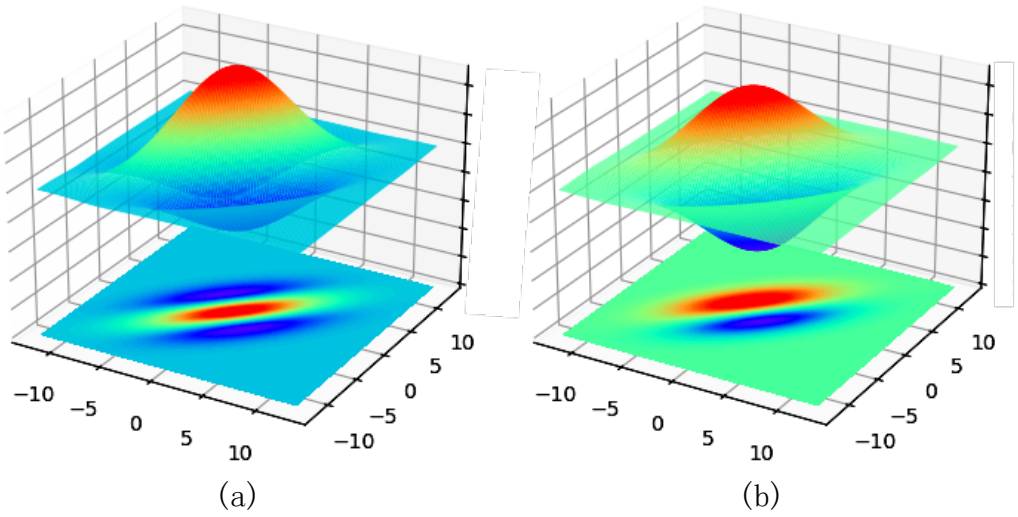


图 3.2 FAG 和 SAG 卷积核的三维空间形态

（a）FAG 的三维空间形态；（b）SAG 的三维空间形态。

尺度不变归一化（Scale-Invariant Normalization）是一种用于将数据进行归一化的技术，其目的是确保归一化后的值不受数据的尺度或大小的影响。当处理具有不同尺度或度量单位的特征或变量时，尺度不变归一化特别有用。尺度不变归一化的目的是确保不同变量的归一化值具有类似的范围或分布，而不考虑它们的原始尺度。这种归一化技术可以帮助提高依赖于特征相对大小的机器学习模型的性能。

在此研究中，为提高图像低层特征的鲁棒性和可重复性，对 FAG 和 SAG 卷积核进行尺度不变归一化。尺度不变归一化的主要思想是将图像块的像素值减去均值并除以标准差，使得块的像素值分布在一个相对较小的范围内。然后，对块进行尺度变换，使其能够适应不同的图像尺度。在这个过程中，均值和标准差也会随着尺度的变化而变化，因此需要对它们进行归一化，以保证图像特征的尺度不变性。尺度不变归一化的各向异性高斯卷积核的描述如下：

$$\widetilde{K}_{fag}(\boldsymbol{x};\sigma,\varphi,\theta)=\beta\sigma^{2\gamma}\cdot K_{fag}(\boldsymbol{x};\sigma,\varphi,\theta) \tag{3.9}$$

$$\widetilde{K}_{sag}(\boldsymbol{x};\sigma,\varphi,\theta)=\beta\sigma^{2\gamma}\cdot K_{sag}(\boldsymbol{x};\sigma,\varphi,\theta) \tag{3.10}$$

其中，β是控制卷积响应幅度的参数，γ是尺度归一化因子。

Wang 等人[75,78]的研究证明，面向边缘特征提取的 FAG 卷积核尺度不变归一化的最佳参数为$\beta = 2\sqrt{\pi}$，$\gamma = 1/4$，面向线性特征提取的SAG卷积核尺度不变归一化的最佳参数为$\beta = -\frac{3\sqrt{3}}{2}$，$\gamma = 1$。在研究中，FAG 和 SAG 卷积核在计算时采用上述的尺度不变归一化最佳参数。

为了使各向异性高斯卷积核适用于数字图像滤波，应该给出其离散版本。通过在二维整数坐标$\mathbb{Z}^2$中对公式 3.5、3.9 和 3.10 进行采样，得到离散的 FAG、SAG 卷积核，描述如下：

$$\widetilde{K}_{fag}(\boldsymbol{m};\sigma_i,\varphi_j,\theta_k) = -2\sqrt{\pi}{\sigma_i}^{\frac{1}{2}} \cdot \frac{[cos\,\theta_k, sin\,\theta_k]\boldsymbol{m}}{{\sigma_i}^2} \cdot g(\boldsymbol{x};\sigma_i,\varphi_j,\theta_k) \tag{3.11}$$

$$\widetilde{K}_{sag}(\boldsymbol{m};\sigma_i,\varphi_j,\theta_k) = -\frac{3\sqrt{3}}{2}{\sigma_i}^2 \cdot \left(\frac{([cos\,\theta_k + sin\,\theta_k]\boldsymbol{m})^2}{{\sigma_i}^4} - \frac{1}{{\sigma_i}^2}\right) \cdot g(\boldsymbol{x};\sigma_i,\varphi_j,\theta_k) \tag{3.12}$$

其中，

$$g(\boldsymbol{m};\sigma_i,\varphi_j,\theta_k) = \frac{1}{2\pi\varphi_j{\sigma_i}^2} exp\left(-\frac{1}{2{\sigma_i}^2}\boldsymbol{m}^T\boldsymbol{R}_k^T\begin{bmatrix}1 & 0\\ 0 & {\varphi_j}^{-2}\end{bmatrix}\boldsymbol{R}_k\boldsymbol{m}\right) \tag{3.13}$$

$$\boldsymbol{R}_k = \begin{bmatrix} cos\,\theta_k & sin\,\theta_k \\ -sin\,\theta_k & cos\,\theta_k \end{bmatrix} \tag{3.14}$$

$\boldsymbol{m} = [m_x, m_y]^T \in \mathbb{Z}^2$表示图像的坐标，$\sigma_i \in \mathbb{S}$表示尺度，$\theta_k \in \mathbb{D}$表示朝向，$\varphi_j \in \mathbb{A}$表示各向异性因子，而$\mathbb{S}$，$\mathbb{D}$，$\mathbb{A}$分别代表尺度集、朝向集和各向异性因子集。一个各向异性高斯卷积核由尺度、朝向和各向异性因子三个参数共同控制。

离散的各项异性高斯卷积核与图像的卷积运算，可实现图像低层特征的提取。离散的 FAG、SAG 卷积核与图像I进行卷积运算后，得到一系列包含图像低层特征的卷积特征图L_{fag}和L_{sag}：

$$L_{fag}(\boldsymbol{m};\sigma_i,\varphi_j,\theta_k) = I * \widetilde{K}_{fag}(\boldsymbol{m};\sigma_i,\varphi_j,\theta_k) \tag{3.15}$$

$$L_{sag}(\boldsymbol{m};\sigma_i,\varphi_j,\theta_k) = I * \widetilde{K}_{sag}(\boldsymbol{m};\sigma_i,\varphi_j,\theta_k) \tag{3.16}$$

其中，$*$表示图像卷积运算。

为使本章构建的DCNN模型更具“生物性”，此研究中的尺度集$\mathbb{S}$、朝向集$\mathbb{D}$和各向异性因子集$\mathbb{A}$来自于麻省理工学院 DiCarlo 团队[17]在其提出的 VOneNet 模型中使用的近似灵长类 V1 区神经响应的数据，这些数据能够帮助 DCNN 的低层更好的近似灵长类 V1 区。

3.3 方位选择感受野启发的卷积神经网络前端

V1 区简单细胞方位选择感受野启发的前端的架构如图 3.3 所示。前端由三

层组成：卷积层、非线性层和 V1 区神经随机层。卷积层由一系列生物可信的卷积滤波器组构成，包括模拟 V1 区“奇对称梯度敏感型”方位选择感受野的 FAG 卷积滤波器组，模拟 V1 区“偶对称双侧对比度敏感型”方位选择感受野的 SAG 卷积滤波器组，模拟视网膜和 V1 区环绕调制机制的 LoG 卷积滤波器组和模简单 V1 区简单细胞神经响应的 Gabor 卷积滤波器组，此外还有对模型鲁棒性有帮助的零阶高斯卷积滤波器组。非线性层由两种非线性构成：简单细胞非线性和复杂细胞非线性。V1 区神经随机层包括 V1 区神经噪声生成器，对卷积特征图添加独立的高斯噪声，以模拟 V1 区神经细胞的随机性。前端融合的多种视觉机制，使得前端对图像的处理近似 V1 区对图像的处理过程。因此，前端也会为模型带来更加接近灵长类视觉的对抗鲁棒性。

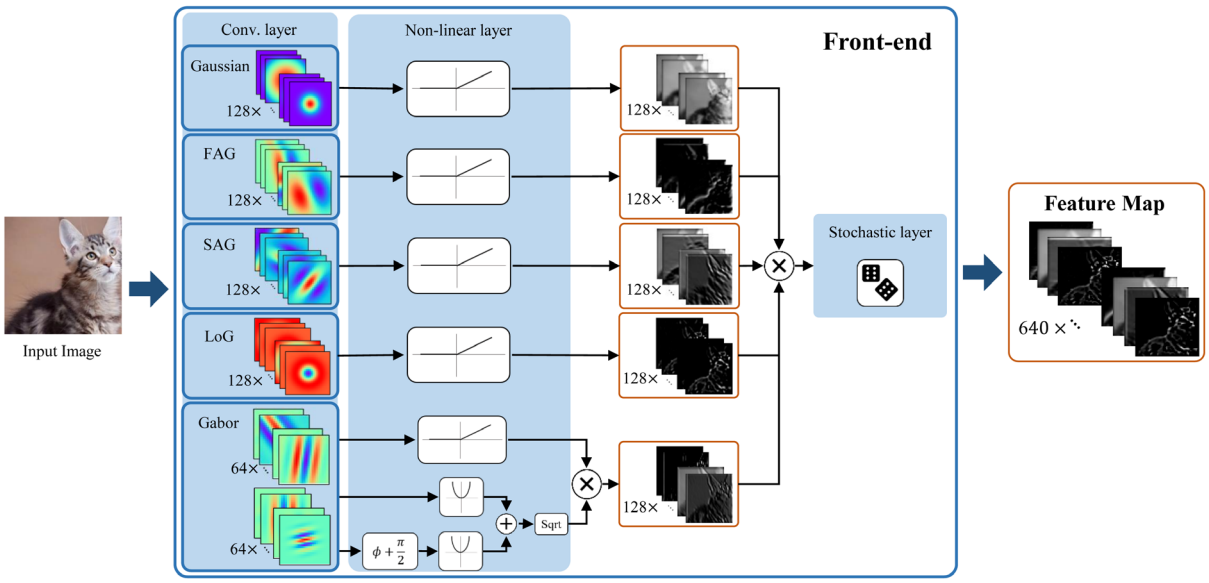


图 3.3 V1 方位选择感受野启发的前端

卷积层包含五种数学参数化的卷积滤波器组，这些滤波器组的卷积核权重获得过程如本章第二小节所描述。这些生物可信卷积滤波器组的权重不会随着模型的训练而更新，这区别于与可学习的卷积滤波器组。这些参数固定的卷积滤波器组通过图像卷积滤波运算，实现图像低层特征提取。V1 区方位选择感受野的数学模型和模拟方位选择感受野的各项异性高斯卷积核在本章第二小节中详细介绍。

在视网膜和 V1 区中，神经元的响应不仅受到刺激物体的直接影响，还受到周围区域的影响，这种周围区域的影响可以通过环绕调制机制来解释。环绕调制机制是指周围区域的神经元响应与刺激物体的神经元响应产生相互作用的现象。在视网膜和 V1 区中，周围区域的神经元响应通常具有抑制作用，即当周围区域被刺激时，它们的响应会抑制刺激物体周围的神经元响应。这种抑制作用可以帮助神经元更好地检测边缘和轮廓等视觉特征，提高视觉系统对复杂场景的适应性

和鲁棒性。

LoG（LapLace of Gaussian）核的响应函数在空间上呈现环状分布，与环绕调制相符合，被认为是视网膜和 V1 区环绕调制机制的模型[79,80]。其描述如下：

$$LoG(\boldsymbol{x};\sigma)=\frac{\boldsymbol{x}^T\boldsymbol{x}-2\sigma^2}{\sigma^4}exp\left(-\frac{\boldsymbol{x}^T\boldsymbol{x}}{2\sigma^2}\right) \tag{3.17}$$

其中，$\boldsymbol{x}=[x,y]^T$是平面坐标系，σ是高斯函数标准差。LoG 滤波器的三维空间形态如图 3.4（a）所示。同样需要给出 LoG 卷积核离散版本，使其可用于图像卷积：

$$LoG(x,y;\sigma_i)=\frac{x^2+y^2-2{\sigma_i}^2}{{\sigma_i}^4}exp\left(-\frac{x^2+y^2}{2{\sigma_i}^2}\right) \tag{3.18}$$

其中，$\sigma_i\in\mathbb{S}$。离散版本的 LoG 卷积核与图像I进行卷积运算，得到 LoG 的卷积特征图L_{LoG}。

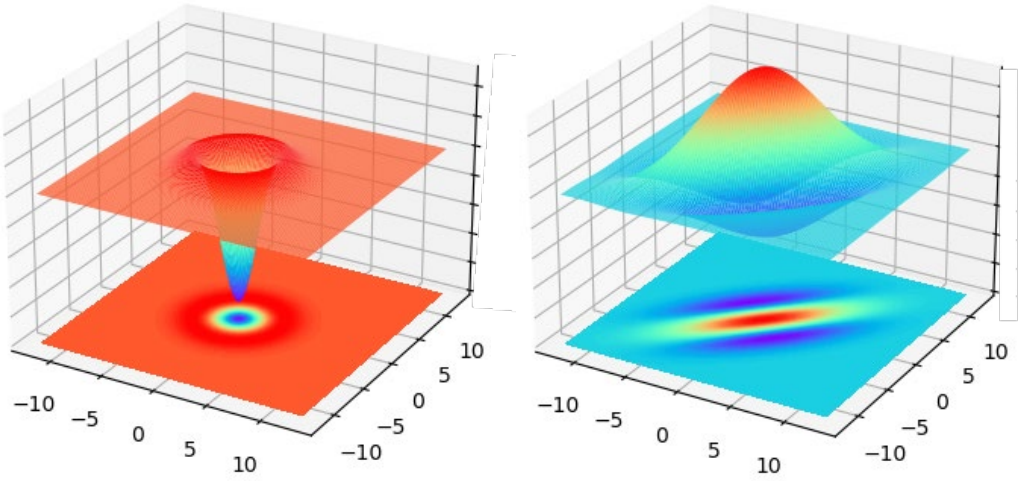


图 3.4 LoG 卷积核和 Gabor 滤波器的三维空间形态

（a）LoG 卷积核的三维空间形态；（b）Gabor 滤波器的三维空间形态。

Gabor 滤波器是一种常用的图像处理滤波器，它模拟了生物视觉系统中 V1 区简单细胞的响应。在 V1 区，简单细胞的响应类似于 Gabor 滤波器的响应，它可以通过在不同方向和不同空间频率上对视觉信息进行滤波来实现对视觉特征的选择性响应。Gabor 滤波器的基本原理就是在不同方向和不同频率上对图像进行滤波，从而模拟简单细胞的响应。Gabor 滤波器的形状类似于二维正弦波，它可以在空间域和频率域中同时表示。在空间域中，Gabor 滤波器通常是一个高斯函数与一个正弦函数的乘积，这个乘积可以产生一个局部的正弦波。Gabor 滤波器的描述如下：

$$g(\boldsymbol{x};\sigma,\varphi,\theta,\lambda,\psi)=\frac{1}{2\pi\sigma^2}exp\left(-\frac{1}{2\sigma^2}\boldsymbol{x}^T\boldsymbol{R}_\theta^T\begin{bmatrix}1 & 0\\ 0 & \varphi^{-2}\end{bmatrix}\boldsymbol{R}_\theta\boldsymbol{x}\right)$$
$$\cdot\cos\left(2\pi\frac{[\cos\theta,\sin\theta]\boldsymbol{x}}{\lambda}+\psi\right) \quad (3.19)$$

其中，$\boldsymbol{x}=[x,y]^T$；σ表示高斯函数标准差，即尺度；φ表示各向异性因子；θ表示旋转角度；λ表示滤波器的波长；ψ表示相位偏移量。Gabor 滤波器的三维空间形态如图 3.4（b）所示。离散版本的 Gabor 滤波器与图像I进行卷积，得到 Gabor 滤波器卷积特征图L_{gabor}。

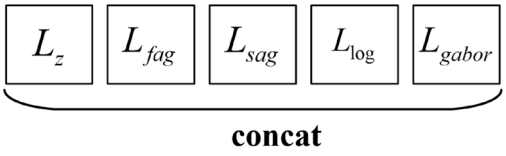


图 3.5 低层特征级联叠加示意图

如图 3.5 所示，将零阶高斯，FAG，SAG，LoG，Gabor 卷积滤波器提取的特征图级联叠加到 DCNN 前端底层来扩增 DCNN 鲁棒低层特征的种类和数量：

$$L=\mathcal{F}_{concat}(L_z,L_{fag},L_{sag},L_{LoG},L_{gabor}) \quad (3.20)$$

其中，$\mathcal{F}_{concat}$表示级联叠加操作。在实际操作中，将每个卷积滤波器与输入图像卷积后得到的包含低层特征的 4 维卷积特征图在第二个维度拼接。由此产生的滤波器组比标准 DCNN 的第一层中的滤波器异质性要高得多，并且能更好地近似 V1 区感受野的多样性。研究提出的模型中，每种卷积滤波器组 128 通道，5 种卷积滤波器组共有 640 通道。卷积核大小为25×25，较标准 DCNN 第一层的卷积核尺寸大，这是因为大尺寸的卷积核更利于大尺度的图像低层特征提取。经过卷积层的图像卷积运算，大小为224×224的 3 通道的 RGB 图像变成了大小为56×56，640 通道的卷积特征图。此外，卷层独立处理每一个颜色通道，即每个卷积通道与输入 RGB 图像中的单一颜色通道进行卷积，这个单一的颜色通道在每轮的训练过程中随机选取。

非线性层基于线性-非线性-泊松（Linear-Nonlinear-Poisson，LNP）模型的第二阶段[81]。LNP 模型是一个经典的神经科学模型，由三个连续的处理阶段组成：卷积、非线性和随机性发生器。在此研究中，非线性层有两种非线性：简单细胞非线性和复杂细胞非线性。简单细胞曾被认为是计算复杂细胞反应的一个中间步骤，但已有的研究发现它们形成了对 V2 区的大部分下游投影[82]。因此，研究采用了两种非线性单元，分别应用于不同的通道。其中简单细胞非线性是一个线性整流变换：

$$S_{nl} = \begin{cases} S_l, & if\ S_l \geq 0 \\ 0, & otherwise \end{cases} \tag{3.21}$$

其中，S_l和S_{nl}表示简单神经元的线性和非线性响应。复杂细胞非线性是一个正交相位对的频谱功率：

$$C_{\psi}^{nl} = \frac{1}{\sqrt{2}} \sqrt{\left(C_{\psi}^{l}\right)^2 + \left(C_{\frac{\pi}{2}+\psi}^{l}\right)^2} \tag{3.22}$$

其中，C^l和C^{nl}表示复杂神经元的线性和非线性响应。简单细胞非线性分别应用于零阶高斯、FAG、SAG 和 LoG 的每个卷积通道。对于 Gabor 滤波器组，此研究参考 VOneNet 设计，将 Gabor 滤波器组分为两组，分别是简单通道和复杂通道。简单通道和复杂通道的相位ψ相差$\pi/2$。其中，简单通道应用简单细胞非线性，复杂通道应用复杂细胞非线性。

神经元响应的一个决定性属性是它们的随机性。在清醒的猕猴中，V1 区神经元的脉冲序列是近似泊松的，即在一个给定的时间窗内以相同的视觉输入对一个神经元进行重复测量，会产生不同的脉冲序列，这些脉冲序列脉冲计数的方差和均值大致相同。平均的脉冲计算取决于呈现的图像，且每次实验的脉冲序列近似泊松分布[83]。为了近似神经元响应的这一特性，此研究在 V1 神经随机层向卷积层的每个卷积通道添加独立的高斯噪声。由于泊松分布不是连续的，它打破了白盒对抗攻击中的梯度，会造成虚假的对抗鲁棒性[51]。为了避免这种情况，评估模型的真实鲁棒性，研究中采用的 V1 随机噪声生成器使用连续的二阶近似泊松噪声，即加入方差等于均值的高斯噪声。和大脑中一样，V1 神经随机性在训练和推理阶段均开启。

对比 SOTA 的 VOneNet 模型中 V1 区宽松启发的前端 VOneBlock，此研究提出的方位选择感受野启发的前端不仅仅包括 Gabor 卷积滤波器组，还增加了 FAG、SAG、LoG 三种生物可信的卷积滤波器组和对模型对抗鲁棒性有帮助的零阶高斯卷积滤波器组。在滤波器的种类和数量上，此研究提出的前端比 VOneBlock 更丰富，对图像的处理也更加接近 V1 区对图像的处理。理论上研究文提出的前端将带来比 VOneBlock 更大的对抗鲁棒性收益，构建的模型也将会有超越 VOneNet 的对抗鲁棒性。

3.4 标准 CNN 层搭建的卷积神经网络后端

此研究提出的模型由方位选择感受野启发的卷积神经网络前端和标准 CNN 层搭建的卷积神经网络后端组成，二者通过瓶颈层连接。后端来自一些标准的 DCNN 模型，例如 AlexNet、ResNet、CORNet-S[57]等。在构建研究中提出的模

型时，使用提出的前端替换标准 DCNN 的第一层。例如，以 ResNet50 为基准模型构建提出的模型时，将 ResNet50 的第一层（包含一个7×7的卷积层、一个批标准化（Batch Normalization）层、一个 ReLU 层和一个2×2的最大池化层）替换为研究中提出的前端即可。通过 ResNet50 的第一层的运算，224×224的 3 通道 RGB 输入图像成为了 64 通道56×56的特征图，而相同的输入图像经过前端的处理后，成为 640 通道56×56的特征图。瓶颈层将 640 通道高维的前端卷积特征图压缩到 64 通道，与 ResNet50 模型的第一层相同。通过瓶颈层对前端高维特征的压缩，实现前后端之间的连接。

后端来自标准的 DCNN 模型，通常由卷积层（Convolution Layer）、池化层（Pooling Layer）、归一化层（Normalization Layer）、激活函数层（Activation Function Layer）、全连接层（Full Connection Layer）等 CNN 标准层通过不同的方式搭建。卷积层是卷积神经网络的核心层。每个卷积层由多个卷积单元组成，这些卷积单元的参数会随着反向传播算法的优化过程而不断更新。卷积层通过对输入图像进行卷积运算来提取图像的各种特征。通常，网络低层的卷积层能够提取一些低级特征，例如边缘、线条和角等。而在深层的卷积层中，网络可以从低级特征中迭代提取更加复杂的特征，这些特征服务于更高级别的语义特征和的最后的分类决策。

每个卷积层由许多卷积单元组成，每个卷积单元的参数随着训练过程由反向传播算法优化。卷积层通过图像卷积运算提取输入图像的不同特征，第一个卷积层可能只能提取一些低级特征，例如边缘、线条和角等，更深层的卷积层能从低级特征中迭代提取更复杂的特征。二维图像卷积运算描述如下：

$$s(i,j)=\sum_{u=0}^{M-1}\sum_{v=0}^{N-1}x(i-u,j-v)w(u,v) \tag{3.23}$$

其中，M是卷积核宽度，N是卷积核高度，x是输入图像数据，w是卷积核权重。对于不同的卷积核，提取的输出图像数据的特征不同。二维图像卷积的示意图如图 3.6 所示。

池化层仿照人类视觉系统对视觉特征进行降维和抽象。在 DCNN 中，某个特征与其他特征的相对位置关系比该特征的精确位置重要的多，而池化层可以在保持特征相对位置的同时缩小数据空间，从而降低参数量和计算量，这有助于加速模型的训练，避免过拟合。实际上，池化层是某种形式的下采样，它将输入图像分割成多个矩形区域，每个区域计算出一个特定的值。常用的池化层包括最大池化、平均池化、全局最大池化、全局平均池化和重叠池化等。图 3.7 为2×2的最大池化示意图，最大池化即在划分的矩形区域中计算该区域的最大值。

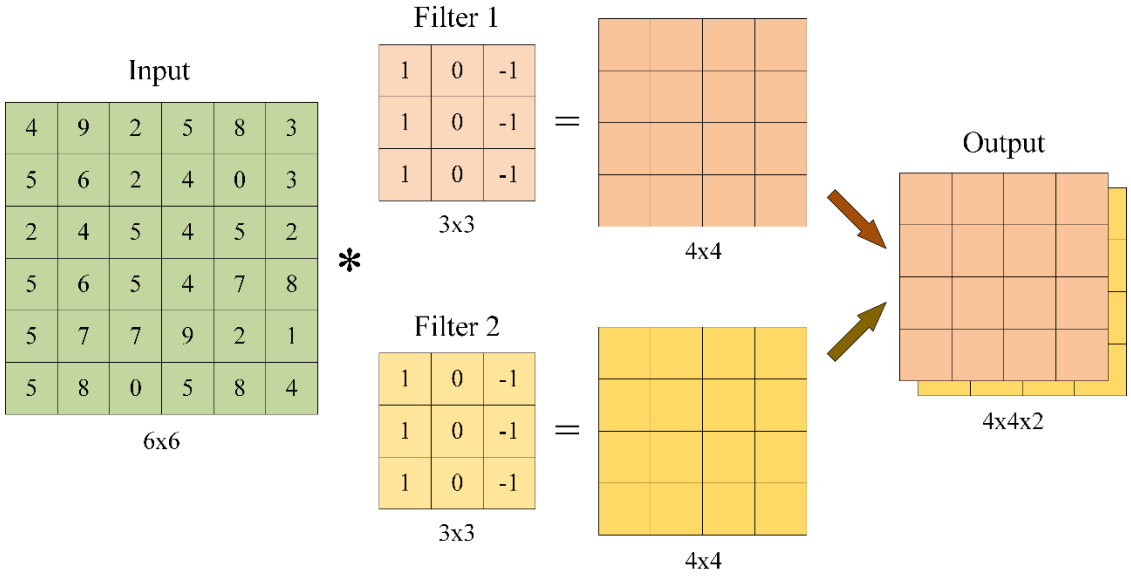


图 3.6 二维图像卷积示意图

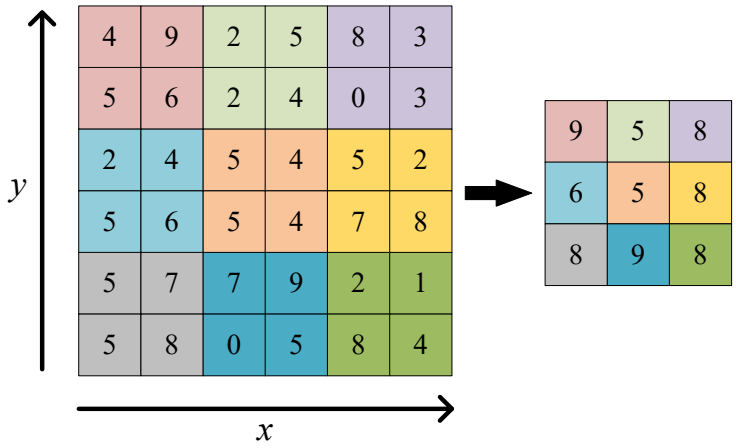


图 3.7 最大池化示意图

归一化层的主要作用是平滑模型中间层的输入，使其稳定分布在合适范围，缓解梯度消失和梯度爆炸问题，加速模型的收敛，加快模型的训练过程，并提高模型对输入扰动的鲁棒性。归一化的数学描述如下：

$$y = \frac{x - E(x)}{\sqrt{Var(x) + \epsilon}}\gamma + \beta \tag{3.24}$$

其中，γ和β是可学习参数。

如图 3.8 所示，在神经元中，输入信号经过加权求和后，通过一个被称为激活函数的单元进行处理。激活函数在神经网络模型的学习起着至关重要的作用，因为它能够增强整个网络的表达能力。通过使用非线性的激活函数，可以实现输入和输出之间的非线性映射，这有助于更好地处理一些线性不可分问题。同时它使得神经网络能够处理更加复杂的任务，因为激活函数提供了对输入数据的非线

性变换，从而使网络更具灵活性并提高网络的表达能力。常用的激活函数有 Sigmoid 激活函数、Tanh（双曲正切）激活函数、ReLU 激活函数、SoftMax 激活函数等。

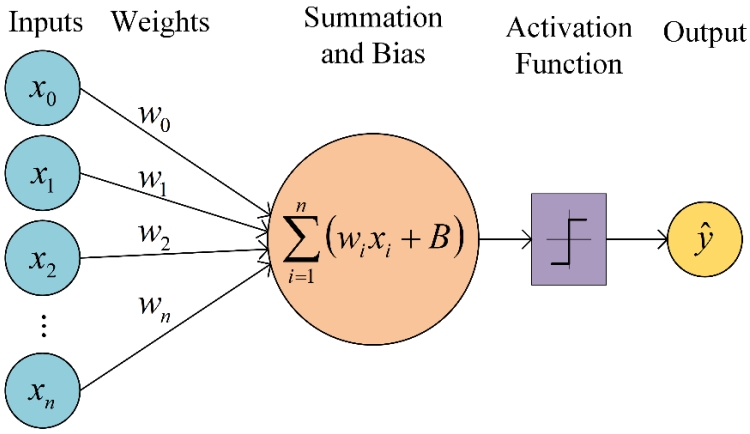


图 3.8 激活函数示意图

全连接层通常位于卷积神经网络的末尾，扮演着“分类器”的角色。前面的卷积层、池化层和激活函数层等对原始数据进行处理，将其映射到隐层特征空间中，提取数据的特征表示。而全连接层则将学习到的特征表示映射到样本标签空间，用于进行分类、回归等任务。全连接层建立网络的抽象特征与样本的真实标签之间联系，使得网络能够对未“见过”的数据进行准确的预测。

3.5 本章小结

本章详细介绍了初级视皮层启发的鲁棒 DCNN 模型。模型由方位选择感受野启发的前端和卷积神经网络后端组成，前后端之间通过瓶颈层连接。前端融合了多种生物视觉机制，由多种生物可信卷积滤波器组组构成的卷积层、非线性层和 V1 区神经随机层组成。卷积层包含一系列生物可信卷积，其中零阶高斯被证明有助于提高模型的鲁棒性；FAG 和 SAG 模拟 V1 区简单细胞方位选择感受野，实现边缘和线条等低层特征的提取；LoG 是视网膜和 V1 区环绕调制机制的模型，能够很好检测边缘和轮廓；Gabor 滤波器模拟 V1 区简单细胞的神经响应，在纹理特征描述中起着重要作用。非线性层由简单细胞非线性和复杂细胞非线性组成，分别应用于不同的卷积通道。V1 区神经随机层近似 V1 区神经元的随机性。理论上，融合多种视觉机制的前端会为模型带来额外的对抗鲁棒性收益，且对抗鲁棒性超过只包含 Gabor 滤波器的 VOneNet。

第4章 对抗鲁棒性分析与模型优化

此研究提出的初级视皮层启发的鲁棒 DCNN 模型，由方向选择感受野启发的前端和标准神经网络后端组成。对比 VOneNet 模型的前端 VOneBlock，研究构建前端融合了多种生物视觉机制。理论上，此研究构建的模型更具“生物性”，也会更具鲁棒性。本章在 CIFAR-10、CIFAR-100、Mini-ImageNet 和 ImageNet 四个分类数据集上进行模型对抗鲁棒性评估，以验证提出的模型。为进一步提高模型性能，研究中还将此研究的方法与基于训练过程的数据增强策略结合。

4.1 数据集、训练超参数与对比模型

此研究提出的模型基于四个图像分类数据集进行评估：CIFAR-10、CIFAR-100、Mini-ImageNet[84]和 ImageNet[16]。CIFAR 系列数据集由加拿大高级研究所提出，包括 CIFAR-10 和 CIFAR-100。CIFAR-10 数据集包含 10 个类别的 60000 个样本，每个类别包含 6000 个样本。这些类别包括：飞机、汽车、鸟、猫、鹿、狗、青蛙、马、船和卡车。CIFAR-100 数据集是 CIFAR-10 的扩展版本，包含 100 个类别的 60000 个样本，每个类别包含 600 个图像。这些类别包括 20 个小类别和 80 个粗略类别，小类别包括不同的动物和交通工具，而粗略类别包括植物、家居用品等。CIFAR 数据集是计算机视觉中经典的数据集之一，被广泛用于研究和比较不同算法的性能。Mini-ImageNet 最初为小样本学习而提出，包含 60000 个样本，共 100 个类别，每个类有 600 个样本。Mini-ImageNet 可以帮助研究人员更好地理解在小数据集上训练模型的挑战，并开发更具有泛化能力的学习算法。在实验中，使用 ImageNet 的子集 ImageNet-2012。ImageNet-2012 包含超过 120 万张高分辨率的彩色图像，涵盖 1000 个类别。ImageNet-2012 数据集的标签由人类标注，每个图像都被分配到一个精细的类别中，这使得 ImageNet-2012 成为计算机视觉研究中广泛使用的数据集之一，尤其是用于评估深度学习算法在大规模图像分类问题上的表现。

在实验前，对四个数据集进行了划分。CIFAR-10 和 CIFAR-100 已划分好训练集和测试集。60000 样本量中，50000 作为训练集，10000 作为测试集。Mini-ImageNet 有 60000 的样本量，共 100 个类别，每个类别 600 个样本。在实验中，从每个类别中随机选取 100 个样本，共 10000 个样本作为测试集，剩余的 50000 样本作为训练集。对于 ImageNet-2012 数据集，官方已提前划分好测试集和训练集。训练集包含大约 130 万个样本，测试集 50000 个样本。为平衡对抗鲁棒性

测试的准确率和时间成本，从 50000 个测试集样本中随机选取 10000 个样本（1000 个类别，每个类别随机选取 10 个样本），构成新的测试集，进行对抗鲁棒性测试。

所有的模型均在本构建的训练测试框架下进行训练和对抗鲁棒性测试。所有的模型使用相同的训练超参数。优化器为随机梯度下降（Stochastic Gradient Descent，SGD）。SGD 是一种常用的优化算法，主要用于训练深度学习模型。与传统的批量梯度下降（Batch Gradient Descent，BGD）相比，SGD 每次只使用小部分样本来计算损失函数的梯度，并利用这个梯度来更新模型参数。由于每次迭代只使用小部分样本，因此 SGD 具有较低的计算复杂度和存储需求，可以更快地收敛到最优解。此外，SGD 还具有较好的泛化能力，因为在每次迭代时，它都会随机选择小部分样本，避免了模型对特定样本的过度拟合。SGD 更新规则为：

$$\theta_{t+1} = \theta_t - \alpha \Delta J\big(\theta; x(i), y(i)\big) \tag{4.1}$$

其中,θ是模型参数，α是学习率，J是损失函数，ΔJ是损失函数的梯度，$x(i), y(i)$分别代表第i个样本的特征和标签。

损失函数为交叉熵损失（Cross-Entropy Loss）。交叉熵损失是深度学习中常用的一种损失函数，具有以下优点：首先，交叉熵损失在模型预测与真实标签之间的差异较大时，损失函数值会相对较大，可以促使模型更快地收敛；其次，交叉熵损失的梯度计算较为简单，可以使用反向传播算法（Backpropagation）高效地求解。交叉熵损失常用于监督学习中。多分类情况下的交叉熵损失的描述如下：

$$L = \frac{1}{N}\sum_{i} L_i = -\frac{1}{N}\sum_{i}\sum_{c=1}^{M} y_{ic}\, log(p_{ic}) \tag{4.2}$$

其中，M是类别的数量；y_{ic}是符号函数，如果样本i的真实类别等于c，则取 1，否则取 0；p_{ic}表示模型预测的样本i属于类别c的概率。

所有实验使用相同的训练环境和实验超参数。实验在 Ubuntu20.04 系统中进行。实验使用的编程环境为 Python-3.8，深度学习框架为 Pytroch-1.9，使用两块 Nvidia Tesla A100 GPU，配合 CUDA 完成模型的训练与测试。此外，在训练时，使用 Pytorch 的分布训练 DDP（Distribute Data Parallel）来加速模型的训练。为使模型的训练过程更加平滑，模型训练中使用的梯度裁剪。经过实验前期的查询与探索，模型训练的最优参数为：batch size 为 256，以 0.1 的初始学习率训练 120 轮，学习率在第$[35, 85, 105]$下降至原来的 0.1 倍。

在对抗鲁棒性分析实验中，设置 3 个对比模型，分别是 ResNet50、Mixup 和 VOneNet。ResNet 由 He 等人[7]提出，在 VGG 和 GoogleNet 等经典网络的基

础上进一步改进和优化，解决了深度神经网络中的梯度消失和退化问题，使得网络可以训练到更深的层数，从而获得更好的性能。ResNet 被广泛应用于计算机视觉领域的各种任务。在实验中，选取 ResNet50 作为基线模型。在 Dapello 等人关于 VOneNet 的研究中，选取了 AlexNet，Resnet50 以及 CORNet-S 三种模型作为后端网络，而三种不同的后端所构建的模型对抗鲁棒性均明显提高。此研究提出模型的构建参照 VOneNet 模型，因此也能够适应不同的后端网络。在实验中，选取较为普遍的 ResNet50 模型，将其去掉第一层作后作为此研究提出模型及 VOneNet 模型的后端网络。。

其次是 Mixup。Zhang 等人[85]提出的 Mixup 是一种基于数据增强的提高模型鲁棒性的方法。Mixup 是一种图像混类的数据增强方法，它一次对两个或多个样本进行操作，在输入空间或特征空间中对样本进行插值，同时也对图像分类的目标标签进行插值。Mixup 的基本思想是将不同的样本进行线性混合，生成新的训练样本。具体来说，对于两个不同的输入样本，可以按照一定的权重将它们进行混合。假设有两个输入样本x_i和x_j，对应的标签分别为y_i和y_j，则 Mixup 生成的新样本为：

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j \tag{4.3}$$

其中λ是一个从β分布中采样得到的随机权重，用于控制两个样本之间的混合程度。相应的，标签也可以按照相同的比例进行混合：

$$\tilde{y} = \lambda y_i + (1-\lambda)y_j \tag{4.4}$$

Mixup 在训练数据中引入更多的样本变化，平滑了远离训练数据的决策边界，使网络对相似类别的识别更加鲁棒。此外，Mixup 还可以降低训练样本之间的相关性，减少过拟合的风险。

最后是 VOneNet。Dapello 等人[17]提出的 VOneNet，是最先进的初级视皮层启发的鲁棒 DCNN 模型。VOneNet 模型与此研究中的模型具有相似的结构，包含一个灵长类 V1 约束的前端 VOneBlock，和一个卷积神经网络后端。其中，VOneBlock 替换标准 DCNN 的第一层，以模拟灵长类视觉系统中的低层特征处理过程。VOneNet 能够在保持原始图像分类准确率基本不变的同时，显著提高模型对图像损坏和对抗扰动的鲁棒性。同时，VOneNet 为 DCNN 的设计提供了新的思路，即通过模拟灵长类视觉系统来提高 DCNN 模型在实际应用中的综合性能。

4.2 对抗鲁棒性实验设计与评估指标

4.2.1 对抗鲁棒性实验设计

此研究设计了四个实验来全面的、准确的评估研究提出的 DCNN 模型，分

别是：无目标白盒对抗鲁棒性实验、无目标黑盒对抗鲁棒性实验、目标白盒对抗鲁棒性实验和无界攻击实验：

（1）无目标白盒对抗鲁棒性实验：采用了四种白盒攻击算法：$\|\epsilon\|_\infty$约束下的 FGSM 算法[14]，$\|\epsilon\|_\infty$约束下的 PGD 算法[13]，$\|\epsilon\|_2$约束下的 PGD 算法和$\|\epsilon\|_\infty$约束下的 FAB（Fast Adaptive Boundary Attack）算法[86]。其中，$\|\epsilon\|_\infty$约束下的 FGSM、PGD 和 FAB 算法的对抗扰动ϵ大小为$1/255$，$\|\epsilon\|_2$约束下的 PGD 算法的对抗扰动ϵ大小为 0.9。对于多步的 PGD 算法，以步长为$\epsilon/5$迭代 10 次；多步的 FAB 算法迭代 10 次。

（2）无目标黑盒对抗鲁棒性实验：采用$\|\epsilon\|_\infty$约束下的黑盒攻击算法 Square Attack。对抗扰动ϵ大小为$1/255$，随机查询 5000 次。

（3）目标白盒对抗鲁棒性实验：采用目标 PGD 算法来生成对抗样本，$\|\epsilon\|_\infty$约束下的目标 PGD 的对抗扰动ϵ大小为$1/255$，并以步长为$\epsilon/5$迭代 10 次。

（4）在无界攻击鲁棒性实验：测试了三种对比模型和此研究提出模型，在白盒算法 FGSM 和 PGD 攻击下，对抗扰动ϵ不断增大的情况下模型的性能。此外还测试了模型在 PGD 算法攻击下，迭代次数不断增加的情况下模型的性能。

实验中使用了四种对抗攻击算法，包括白盒攻击算法 FGSM、PGD 和 FAB，以及黑盒攻击算法 Square Attack。FGSM 是对抗攻击领域较为经典的单步攻击算法。PGD 是公认的攻击力较强的多步对抗攻击算法，常用于防御算法鲁棒性的评估。Corece 等人[86]提出的 FAB 白盒对抗攻击算法，基于边界攻击（Boundary Attack）方法改进，采用自适应策略来减少攻击所需的迭代次数。边界攻击是一种基于距离度量的对抗攻击方法，它通过对原始输入样本的边界进行搜索，寻找使模型的预测结果偏离原始标签的对抗性扰动。但边界攻击需要大量的迭代来寻找有效的对抗样本，这限制了其实际应用。为解决这一问题，FAB 提出了一种自适应策略，可以根据目标模型的不同特点来调整攻击的速度和强度。该方法利用当前攻击成功的对抗性样和边界信息，自适应地确定下一步搜索的方向和步长。FAB 还使用了一种双线性插值方法来在每次迭代中更新对抗样本，从而进一步加快攻击速度。FAB 与部分专门针对P范数的攻击表现相似或更好，并且对混淆梯度[51]具有鲁棒性。

在无目标黑盒对抗鲁棒性实验中，采用黑盒攻击算法 Square Attack 来评估模型的鲁棒性。Andriushchenko 等人[35]提出的 Square Attack 是一种基于得分的黑盒攻击算法。Square Attack 的基本思想是在每个搜索步骤中，将输入空间分为若干个网格，对于每个网格中的输入点，通过计算它们与原始样本的距离来评估它们的攻击能力，并选择最具攻击性的点进行下一步搜索。然后，在每个网格内进行更精细的搜索，以找到最终的对抗样本。Square Attack 不受模型梯度

的限制，可以在较少的搜索步骤中生成有效的对抗样本。

4.2.2 对抗鲁棒性评估指标

在对抗鲁棒性实验中，通过评估模型对原始样本、对抗样本的 Top-1 分类准确率来评估模型的性能：

$$Acc = \frac{N_t}{N_t + N_f} \tag{4.5}$$

其中，Acc代表分类准确率，N_t表示分类正确的样本数量，N_f表示分类错误的样本数量。对于无目标白盒对抗鲁棒性实验，因为有四种白盒评估算法，此研究取模型对原始样本的分类准确率和在四种攻击算法攻击下的分类准确率的均值，作为模型无目标白盒对抗鲁棒性的评价指标：

$$Acc_{white}^{no-target} = \frac{Acc_{clean} + Acc_{fgsm} + Acc_{pgd-l\infty} + Acc_{pgd-l2} + Acc_{fab}}{5} \tag{4.6}$$

其中，Acc_{clean}表示模型对原始样本的分类准确率；Acc_{fgsm}表示模型在 FGSM 算法攻击下的分类准确率；$Acc_{pgd-l\infty}$表示模型在$\|\epsilon\|_\infty$约束 PGD 算法攻击下的分类准确率；Acc_{pgd-l2}表示模型在$\|\epsilon\|_\infty$约束 PGD 算法攻击下的分类准确率；Acc_{fab}表示模型在 FAB 算法攻击下的分类准确率。

对于目标白盒对抗鲁棒性实验，评价指标也是对抗样本的 Top-1 分类准确率，但与无目标白盒对抗鲁棒性不同的是，目标攻击下的准确率越高，表示模型越容易被攻击算法攻破。在目标白盒对抗鲁棒性实验中，模型对对抗样本的分类准确率越低，表示模型的防御力越好。这与无目标白盒对抗鲁棒性的指标相反。在无目标的白盒、黑盒对抗鲁棒性实验中，模型对对抗样本的分类准确率越高，表示模型的性能越好。

4.3 对抗鲁棒性分析

4.3.1 无目标白盒对抗鲁棒性分析

表 4.1、表 4.2、表 4.3 和表 4.4 分别报告了我们的模型和对比模型在 CIFAR-10，CIFAR-100，Mini-ImageNet 和 ImageNet 四个数据集上的无目标白盒对抗鲁棒性。图 4.1 总结了四个模型在四个数据集上的表现。在无目标白盒对抗鲁棒性测试中，研究提出模型在四个数据集 CIFAR-10、CIAFR-100、Mini-ImageNet 和 ImageNet 上对抗鲁棒性的表现远高于基线模型，相比基线模型分别提高了 34.95%、30.12%、27.87%和 31.37%。对比基线模型和增加 Mixup 数据增强的基线模型，Mixup 会提高模型原始图像的分类准确率，但是对模型对抗鲁棒性的贡献较小，这与前人的研究一致。此外，此研究的模型在四个数据集上的

表现比目前最先进的神经科学启发鲁棒 DCNN 模型 VOneNet 更好，在 CIFAR-10、CIAFR-100、Mini-ImageNet 和 ImageNet 上分别提高 1.32%，3.05%，6.40%和 2.58%。

与基线模型相比，研究提出模型的无目标白盒对抗性鲁棒性提升明显。例如在 Mini-ImageNet 数据集上的 PGD-L_∞算法攻击下，基线模型对对抗样本的分类准确率仅 5.56%，基线模型几乎失效。相同条件下，VOneNet 模型对对抗样本的分类准确率为 27.03%，而此研究的模型对对抗样本的分类准确率为 38.52%，研究提出模型的性能明显高于基线模型和 VOneNet。

与数据增强方法 Mixup 相比，此研究的模型在对抗鲁棒性方面的收益更加明显。此研究的方法与 Mixup 均需要较小的额外训练成本，但 Mixup 仅对攻击力较弱的算法有效，无法抵御来自攻击力较强的算法的攻击。例如 CIFAR-10 数据集上，FGSM 算法攻击下 Mixup 模型分类准确率比基线模型高 5.73%，但在 PGD-L_∞攻击下 Mixup 模型分类准确率仅比基线模型高 0.26%。在相同数据集中，FGSM 攻击下此研究提出的模型分类准确率比基线模型高 34.14%，在 PGD-L_∞攻击下此研究提出的模型分类准确率比基线模型高 46.07%。显然，此研究的方法所带来的对抗鲁棒性收益远高于 Mixup。

表 4.1 CIFAR-10 数据集上无目标白盒算法攻击下的 Top-1 分类准确率

Attacks	Clean (%)	FGSM (%)	PGD-L_∞ (%)	PGD-L_2 (%)	FAB (%)	Mean (%)
Baseline	90.89	43.64	28.88	28.79	34.80	45.40
Mixup	92.23	49.37	29.14	29.59	35.04	47.07
VOneNet	90.55	74.23	70.53	71.57	88.30	79.04
Ours	89.26	77.78	74.95	71.57	88.21	80.35

表 4.2 CIFAR-100 数据集上无目标白盒算法攻击下的 Top-1 分类准确率

Attacks	Clean (%)	FGSM (%)	PGD-L_∞ (%)	PGD-L_2 (%)	FAB (%)	Mean (%)
Baseline	70.62	21.41	10.59	10.31	20.97	26.78
Mixup	71.91	23.94	12.91	12.12	20.60	28.30
VOneNet	68.96	46.96	42.84	44.98	65.53	53.85
Ours	67.14	52.16	48.95	50.55	65.70	56.90

表 4.3 Mini-ImageNet 数据集上无目标白盒算法攻击下的 Top-1 分类准确率

Attacks	Clean (%)	FGSM (%)	PGD-L_∞ (%)	PGD-L_2 (%)	FAB (%)	Mean (%)
Baseline	72.10	15.91	5.56	6.35	17.27	23.44
Mixup	73.53	18.01	5.94	7.01	18.52	24.60
VOneNet	70.04	35.41	27.03	27.49	64.56	44.91
Ours	68.23	45.66	38.52	38.40	65.74	51.31

表 4.4 ImageNet 数据集上无目标白盒算法攻击下的 Top-1 分类准确率

Attacks	Clean (%)	FGSM (%)	PGD-L_∞ (%)	PGD-L_2 (%)	FAB (%)	Mean (%)
Baseline	73.21	21.41	10.59	10.31	13.20	25.74
Mixup	74.84	23.94	12.91	12.12	13.44	27.45
VOneNet	70.37	46.96	42.84	44.98	67.52	54.53
Ours	67.94	52.16	48.95	50.55	65.96	57.11

研究中提出方法的对抗鲁棒性高于最先进的神经科学启发模型 VOneNet。例如，在 Mini-ImageNet 数据集中，在 PGD-L_∞攻击下，此研究提出模型的对抗鲁棒性比 VOneNet 高 11.49%，白盒对抗鲁棒性提高了 6.40%。在其他数据集上也有不同程度的提高，但是效果没有 VOneNet 明显，是因为模型前期的参数优化在 Mini-ImageNet 数据集上进行。实验结果正好印证了第三章中此研究提出模型与 VOneNet 模型的分析对比，进一步证实了使用生物可信卷积滤波器组来增强模型对抗鲁棒性的可行性，同时也证明了研究中提出方法的有效性。此研究的模型比 VOneNet 模型在对抗鲁棒性方面有一定的优势。相较于与 VOneNet 模型，此研究的模型融合了多种生物视觉机制，包括模拟方位选择感受野的 FAG 和 SAG 卷积核，模拟视网膜和 V1 区环绕调制机制的 LoG 卷积核以及零阶高斯。二者之间对抗鲁棒性的差异，也有力的证明了研究中引入的其他生物可信卷积滤波器组的有效性。

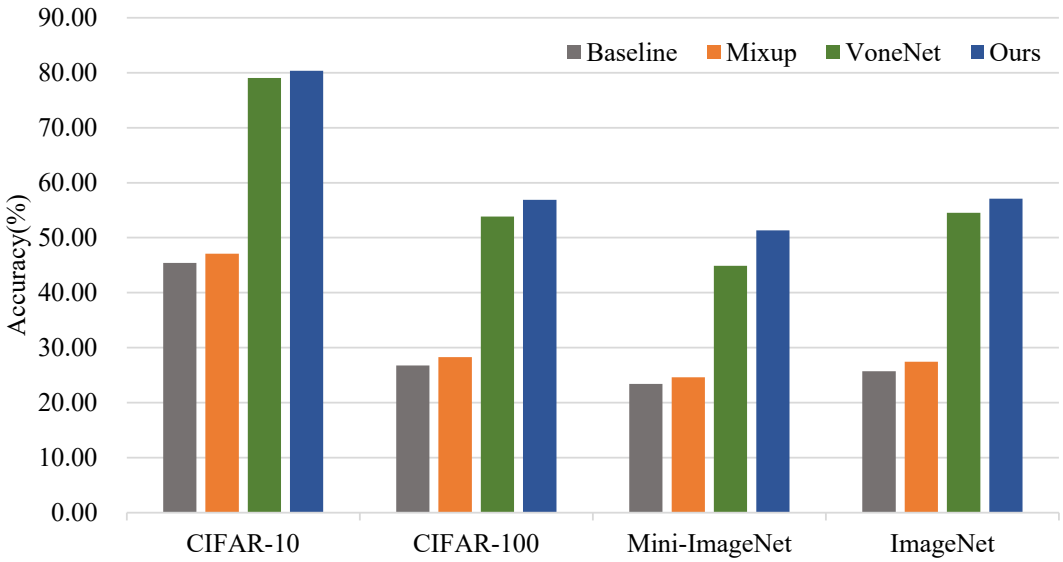


图 4.1 无目标白盒对抗鲁棒性实验分析

4.3.2 无目标黑盒对抗鲁棒性分析

黑盒扰动由 Square Attack 算法生成。表 4.5 报告了四个数据集上此研究提出模型和对比模型的黑盒分类准确率。此研究提出模型的无目标黑盒鲁棒性也远优于基线模型和 Mixup 模型。例如表 4.5 中，CIFAR-100 数据集上的无目标黑盒对抗鲁棒性对比中，基线模型和 Mixup 模型的黑盒分类准确率分别为 30.31%，29.18%，而此研究提出的模型和 VOneNet 模型的黑盒分类准确率分别为 66.69%和 68.72%，远高于基线模型和 Mixup 模型。

表 4.5 CIFAR-10、CIFAR-100、Mini-ImageNet 和 ImageNet 数据集上无目标黑盒算法攻击下的 Top-1 分类准确率

Datasets	CIFAR-10 (%)	CIFAR-100 (%)	Mini-ImageNet (%)	ImageNet (%)
baseline	54.64	30.31	46.41	41.54
Mixup	58.12	29.18	49.41	42.51
VOneNet	90.71	68.72	69.72	70.55
Ours	88.93	66.69	67.8	67.73

此研究提出的模型和 VOneNet 模型的黑盒分类准确率与表 4.1、表 4.2、表 4.3 和表 4.4 中的原始图像分类准确率相比，发现研究提出的模型和 VOneNet 模型的黑盒分类准确率和原始图像分类准确率几乎相同，说明此研究中的模型和 VOneNet 模型几乎对 Square Attack 黑盒攻击免疫。需要说明的是，此研究提出的模型在黑盒测试中的表现比 VOneNet 模型略差，是因为研究提出模型的原始

图像分类准确率略低于 VOneNet 模型。

4.3.3 目标白盒对抗鲁棒性分析

在本小节，进行了目标白盒对抗鲁棒性实验。对抗扰动由目标 PGD 算法生成，目标类别在除真实类别以外的其他类别中随机选取。表 4.6 报告了四个数据集上此研究提出的模型和对比模型在 PGD 目标攻击下的性能。目标攻击在基线模型和 Mixup 模型上取得了很好效果。例如，在 ImageNet 数据集上，目标攻击下的基线模型和 mixup 模型对对抗样本的分类正确率为 69.30%和 89.28%，这说明这两种模型很容易被目标攻击误导，即两种模型目标白盒鲁棒性较差。对比此研究提出的模型和 VOneNet 模型，在 ImageNet 数据集上对对抗样本的分类准确率为 0.51%和 1.67%，即目标攻击对研究提出模型和 VOneNet 模型几乎失效。在四数据集上分别对比此研究提出模型和 VOneNet 表现，发现此研究提出模型的目标白盒对抗鲁棒性均比 VOneNet 更好。

表 4.6 CIFAR-10、CIFAR-100、Mini-ImageNet 和 ImageNet 数据集上目标白盒算法攻击下的 Top-1 分类准确率

Datasets	CIFAR-10 (%)	CIFAR-100 (%)	Mini-ImageNet (%)	ImageNet (%)
baseline	52.03	22.47	46.94	69.30
Mixup	52.65	21.87	47.86	89.28
VOneNet	12.78	2.26	7.29	1.67
Ours	9.55	1.48	2.91	0.51

4.3.4 无界攻击分析

为了验证模型的有效性，此研究进行了无界攻击实验，实验结果如图 4.2 所示。如图 4.2（a）所示，对于攻击力较弱的 FGSM 算法，随着对抗扰动大小ϵ从 1/255 增加到 36/255，模型的无目标白盒对抗鲁棒性逐渐收敛到 0%或者随机水平。如图 4.2（b）所示，对于攻击力较强的 PGD 算法，所有模型的无目标白盒对抗鲁棒性收敛到 0%。任何有效的防御模型都无法抵御这种无界攻击，因为较大的对抗扰动会修改整个图像并破坏图像的关键特征。此外，研究还测试了 PGD 算法无限迭代的对抗鲁棒性，实验结果如图 4.2（c）所示，所有测试模型在 100 次迭代后收敛。

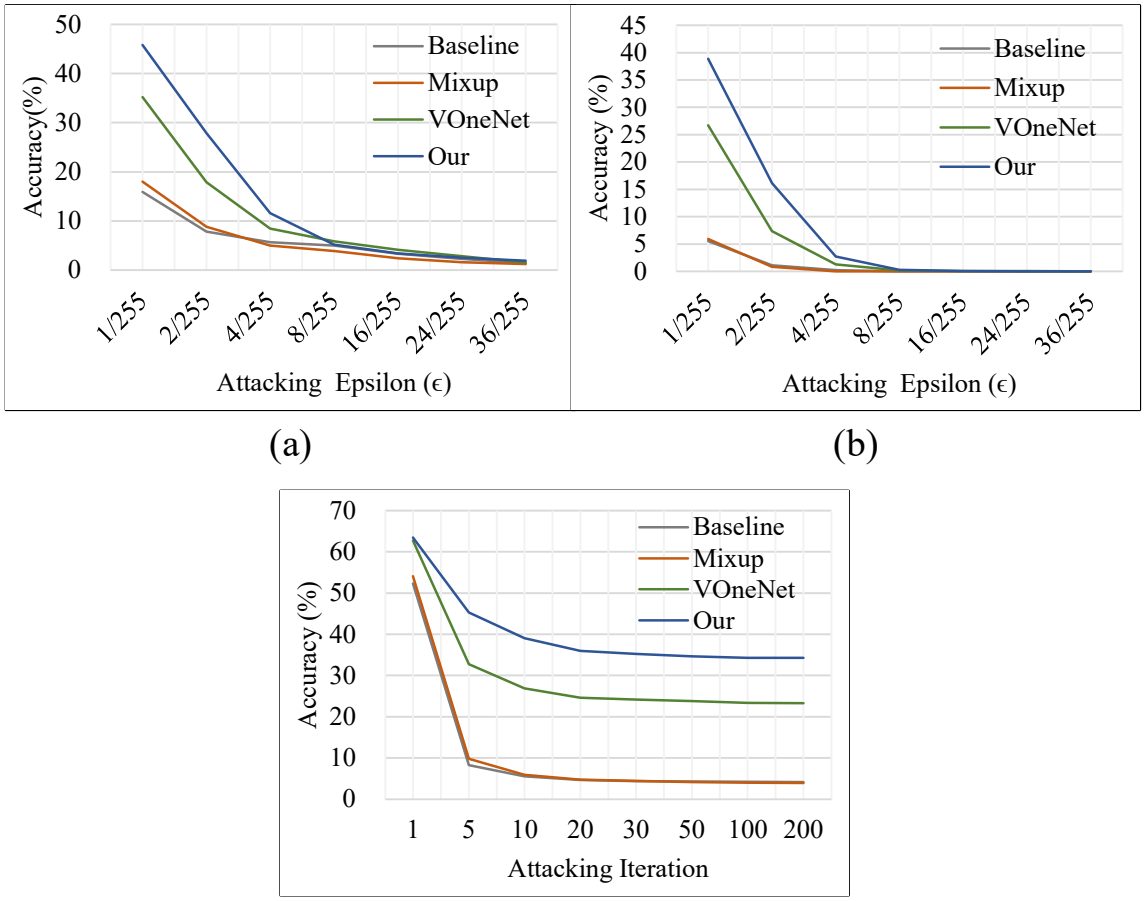


图 4.2 在 Mini-ImageNet 数据集上的无界攻击

（a）模型在 FGSM 算法无限扰动强度攻击下的对抗鲁棒性；（b）模型在 PGD 算法无限扰动强度攻击下的对抗鲁棒性；（c）模型在 PGD 算法无限迭代次数攻击下的对抗鲁棒性。

4.3.5 对抗鲁棒性实验完整性检验

在第三章的模型介绍中，此研究提出模型的前端融合了多种视觉机制，且生物可信卷积滤波器组的种类比 VOneNet 丰富，因此研究提出模型会有比基础 DCNN 和 VOneNet 更好的对抗鲁棒性。在第三小结的三个对抗鲁棒性分析实验中，此研究提出的模型表现均超过 VOneNet 模型，与第三章中研究提出模型和 VOneNet 模型的分析对比重合，验证了此研究中引入的 FAG 和 SAG 卷积滤波器组对模型对抗鲁棒性的贡献，验证了此研究工作的有效性，验证了通过数学模型将生物视觉机制引入 DCNN 的可行性，验证了使用生物视觉机制来增强 DCNN 模型对抗鲁棒性的可行性。在研究中，从理论层面和实验层面分别分析检验了此研究提出模型相对基线模型 ResNet50 和 VOneNet 的优势，增加了此研究的真实性，为后续的研究以及此研究在现实世界的应用提供了牢靠的基础。

此外，为了验证此研究提出的模型不会受到有缺陷或者不完整的对抗鲁棒性

评估，研究还遵循一些 DCNN 模型对抗鲁棒检验准则，来检验此研究提出的模型和对抗鲁棒性分析实验：

（1）迭代攻击优于单步攻击，例如表 4.1、表 4.2、表 4.3、表 4.4 以及图 4.2 中的（a）和（b）中的 FGSM 算法和 PGD 算法。

（2）无界攻击测试中，模型对抗鲁棒性降低到 0%或者随机水平，例如图 4.2 中的（a）和（b）。

（3）模型的对抗鲁棒性随着攻击步数的增加而逐渐收敛，例如图 4.2 中的（c）。

（4）对抗鲁棒性实验涵盖无目标攻击和目标攻击，例如表 4.6。

（5）使用黑盒攻击算法和绕过混淆梯度的方法来避免的潜在的错误评估，例如表 4.5 中的 Square attack 算法和和表 4.1、表 4.2、表 4.3 和表 4.4 中 FAB 算法的攻击结果。

4.4 基于数据增强的模型优化

在第一章对抗防御的介绍中，介绍了基于数据的提高模型对抗鲁棒性的方法，其中包括基于训练过程的数据增强方法。为进一步提高此研究提出模型的综合性能，本节将一些基于训练过程的数据增强方法与此研究提出的模型相结合。研究中，选择了两种基于训练过程的数据增强方法：RandAugment[87]和 Mixup[85]。

RandAugment 由 Google Brain 团队提出，是一种数据增强方法，旨在提高图像的分类准确率。其基本思想是通过随机应用一组固定的图像增强来扩增数据集，从而提高模型的泛化能力。RandAugment 包含两个主要部分：增强策略和超参数。增强策略是一组图像增强操作，它们以随机的方式应用于每个输入图像。RandAugment 包含 15 种不同的图像增强，包括：随机裁剪、随机翻转、随机颜色扭曲、随机旋转等。每个操作都有不同的参数范围，例如旋转角度、剪裁大小等。对于每个输入图像，从增强策略中等概率抽取随N个图像增强方式，并以随机顺序应用于输入图像。每个数据增强以 M 的强度应用于每个图像。RandAugment 是一种简单而有效的数据增强方法，被广泛应用于图像分类任务。RandAugment 是为了提高 DCNN 模型对原始样本的分类准确率，并不是专门为提高对抗鲁棒性而设计。第二种基于训练过程的数据增强方法是 Mixup，已在第四章对比模型中详细介绍。

将基于训练的数据增强方法应用此研究提出的模型，即在图像进入 DCNN 模型之前，对其应用数据增强。RandAugment 是在图像预处理时对图像进行随机数据增强。Mixup 在图像预处理后，送入模型之前，对图像和标签进行混类增强。此外，Mixup 会处理模型输出，以正确计算损失。

在本节的实验中，评估了此研究提出的模型和两个变体模型：此研究提出模型和 RandAugment 组合，此研究提出的模型和 Mixup 组合。评估指标为模型无目标白盒对抗鲁棒性。实验环境与超参数和第四章模型对抗鲁棒性分析实验一致。

实验结果如图 4.3 所示。此研究提出的模型与数据增强能进一步提高模型的对抗鲁棒性。在 CIFAR-100、Mini-ImageNet 和 ImageNet 三个数据集上，RandAugment 与此研究提出模型结合的变体模型和 Mixup 与此研究提出模型结合的变体模型，鲁棒性均提高。其中，此研究提出模型与 RandAugment 结合，在 ImageNet 数据集上表现出较好的性能，此研究提出模型与 Mixup 结合在 Mini-ImageNet 上表现出较好的性能。

值得指出的是，在 CIFAR-10 上，此研究提出模型与数据增结合的变体模型表现不佳，甚至出现性能下降，或许是因为 CIFAR-10 数据集类别太少，分类精度已经达到卷积神经网络后端的上限。

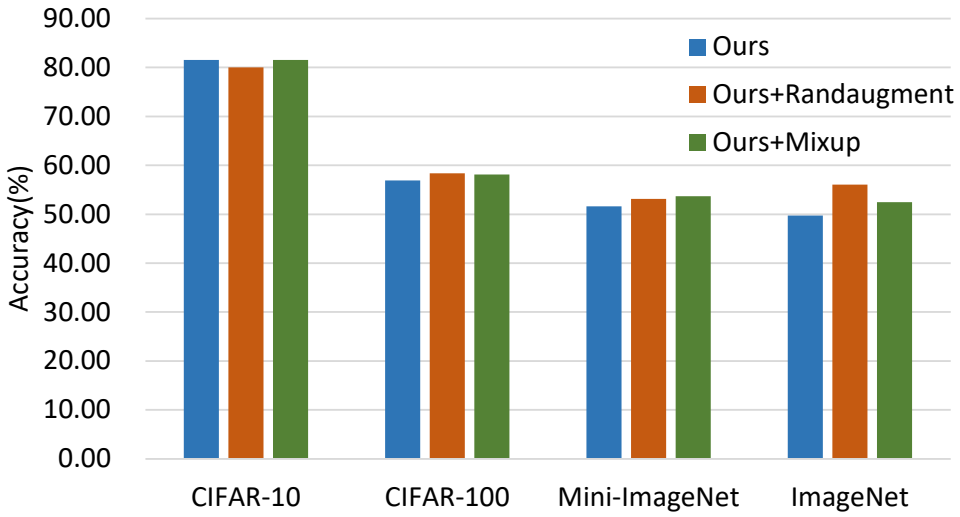


图 4.3 此研究提出模型与数据增强方法结合对比图

4.5 本章小结

本章进行了对模型的一系列对抗鲁棒性分析实验。首先，在无目标白盒对抗鲁棒性实验中，此研究提出模型的性能远超基线模型，且表现超过目前最先进的 VOneNet 模型。其次，在无目标黑盒对抗鲁棒性实验中，此研究提出模型的表现同样优异，黑盒对抗鲁棒性远超基线模型，甚至直接免疫无目标黑盒攻击。然后，在目标白盒对抗鲁棒性实验中，此研究提出的模型性能依然优于 VOneNet 模型。最后的无界攻击实验也验证了此研究提出模型的有效性。除了对抗鲁棒性评估实验，此研究还进行了模型优化。本章测试了此研究提出模型与基于训练过程的数据增强方法结合的模型优化策略。实验结果表明，此研究提出的模型与数据增强结合，能进一步提高模型的对抗鲁棒性。

第5章 消融研究与基于模型可解释技术的模型分析

在对此研究提出模型的对抗鲁棒性分析实验中，此研究提出模型的对抗鲁棒性明显提高。为探究前端生物可信卷积滤波器组对模型对抗鲁棒性的作用，此研究设计了消融实验来分析生物可信卷积滤波器的作用。同时，本章还通过模型可解释性技术，来探索基线模型与此研究提出的模型在分类时决策过程，为模型的进一步优化提供思路。

5.1 消融研究

5.1.1 消融实验设计

消融研究对深度学习研究至关重要。了解系统中的因果关系是产生可靠知识的最直接方法，不管出于何种研究目的。消融是研究因果关系的一种非常低成本的方法，在机器学习，特别是复杂的深度卷积神经网络的背景下，已经广泛采用“消融研究”来描述去除网络的某些部分的过程，以便更好地理解网络的行为。此研究提出的方位选择感受野启发的前端，是此研究的主要创新点，也是模型对抗鲁棒性的主要来源。前端包含三层：卷积层、非线性层和 V1 神经随机层。在卷积层，包含了模拟 V1 区方位选择感受野的 FAG、SAG 卷积核，模拟视网膜和 V1 区环绕调制机制的 LoG 卷积核，模拟 V1 区神经细胞响应的 Gabor 滤波器，还有包括零阶高斯核。这些卷积核共同负责图像低层特征的提取。为了探索这些生物可信卷积滤波器组对模型对抗鲁棒性的贡献，将这些生物可信卷积滤波器组作为一个整体。在变体模型中，则将这些生物可信卷积滤波器组替换为可学习的卷积滤波器组，然后通过对比分析，来验证生物可信卷积滤波器组的作用。变体模型示意图如图 5.1 所示。

消融实验在Mini-ImageNet数据集上进行，模型的训练与测试环境以及模型训练超参数与第四章实验一致。在消融实验中，测试了三个模型的无目标白盒对抗鲁棒：基线模型 ResNet50，变体模型和此研究提出的模型。基线模型与变体模型的主要区别是 V1 神经随机性，通过二者之间的对比可以直观的分析 V1 神经随机性的作用。基线模型和此研究提出模型的对比，可以分析整个前端的贡献。而变体模型和此研究提出模型的比较，可以分析生物可信卷积滤波器组的作用。

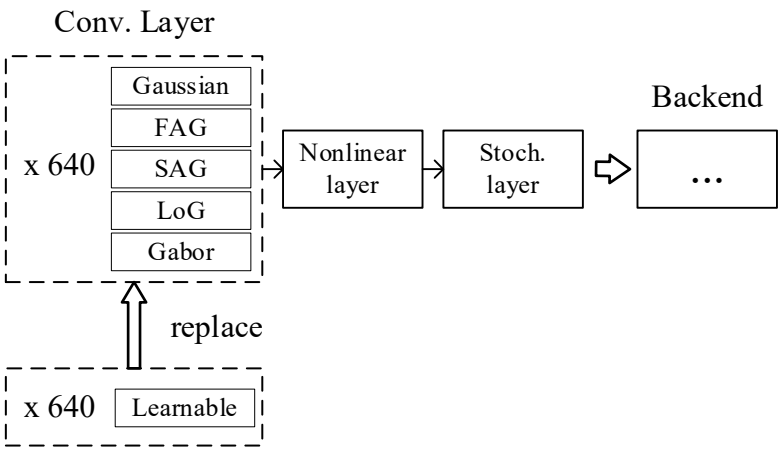


图 5.1 变体模型示意图

5.1.2 消融实验结果与分析

实验结果如表 5.1 所示。对表中数据总结后，结果如图 5.1 所示。基线模型与模型之间的区别是方位选择感受野启发的前端，两者之间的比较说明融合多种视觉机制的前端为模型提供了额外的对抗鲁棒性，且模型对抗鲁棒性明显提高。对比基线模型与变体模型，二者的区别是变体模型第一层的卷通道数高于基线模型，且变体模型包含了 V1 神经随机性。更高的卷积通道数会为模型带来更高的性能，但已有的研究证明只增加通道数带来的增益十分有限。因此，基线模型和变体模型之间对抗鲁棒性之间的差异，大部分来自于 V1 神经随机性，说明 V1 神经随机性对模型对抗鲁棒性的提高有帮助。

表 5.1 基线模型、变体模型和此研究提出模型在 Mini-ImageNet 上的 Top-1 分类准确率

Attacks	Clean (%)	FGSM (%)	PGD-L_∞ (%)	PGD-L_2 (%)	FAB (%)	Mean (%)
Baseline	72.10	15.91	5.56	6.35	17.27	23.44
Variant	66.18	26.47	19.44	22.77	58.59	38.69
Ours	68.23	45.66	38.52	38.40	65.74	51.31

变体模型和此研究提出的模型相比，变体模型将生物可信卷积滤波器组替换成了相同通道数的可学习卷积滤波器组。二者对抗鲁棒性之间的差异，说明生物可信卷积滤波器组对模型有助于提高模型的抗鲁棒性。此外，二者对抗鲁棒性之间的差异也说明来自参数固定的生物可信卷积滤波器组的收益无法通过模型的自我学习过程获得。消融实验表明，前端生物可信卷积滤波器组和 V1 神经随机性有机组合，共同提高模型的对抗鲁棒性。

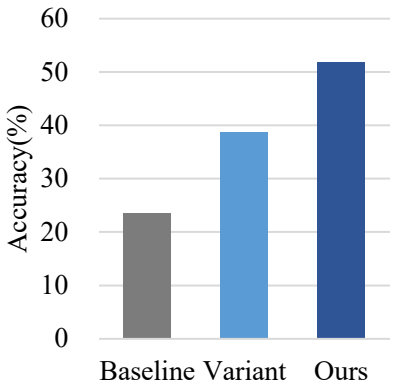


图 5.2 消融实验结果分析

5.2 基于模型可解释技术的模型可视化分析

5.2.1 模型可解释技术

深度学习模型在许多领域的成功表现得到广泛认可，但这些模型缺乏可解释性，这对于其在现实任务中的应用，特别是安全敏感任务中的应用造成了严重限制。深度学习模型通常由卷积层、池化层、归一化层、非线性层和全连接层等组合堆叠而成，这些层之间的复杂关系使得深度学习模型的决策过程非常难以理解和解释。因此，开发深度学习模型的可解释性技术是非常重要的。模型可解释性技术可以帮助用户更好地理解机器学习模型的决策过程，并提高模型的可靠性和可信度。同时，模型可解释技术也可以帮助研究者进行模型验证、模型诊断、辅助分析、发现新知识等[88]。

深度学习模型可解释性技术可分为anti-hoc可解释性和post-hoc可解释性。anti-hoc 可解释性指模型本身具备可解释性，不需要额外的信息就可以理解模型的决策过程或决策依据，这种可解释性在模型训练之前就被设计，因此也被称为事前可解释性。anti-hoc 可解释性通常使用结构简单、易于理解的自解释模型来实现，例如朴素贝叶斯、线性回归、决策树、基于规则的模型等。此外，也可以通过构建将可解释性直接结合到具体的模型结构中的学习模型来实现模型的内置可解释性[89]。post-hoc 可解释性也被称事后可解释性，是在模型训练完成后进行的。它旨在利用解释方法或构建解释模型，来解释一个已经训练好的学习模型的工作机制、决策行为和决策依据。因此，post-hoc 可解释性的关注点在于设计高保真度的解释方法或构建高精度的解释模型。根据解释目的和解释对象的不同，post-hoc 可解释性可分为全局可解释性和局部可解释性，对应的方法则称为全局可解释方法和局部可解释方法。

深度学习模型的全局可解释性旨在帮助人们从整体上理解模型内部的复杂逻辑和工作机制，例如模型的学习过程、从训练数据中学到的知识、决策依据等。

这需要可解释方法能够以一种人类可理解的方式来表达一个训练好的复杂学习模型。典型的全局解释方法包括解释模型／规则提取[90]、模型蒸馏[91]、激活最大化解释[92]等。深度学习模型的局部可解释性旨在帮助人们理解模型在特定输入样本上做出决策的过程和依据。与全局可解释性不同，局部可解释性以输入样本为导向，通常可以通过分析每个输入样本特征对模型做出最终决策的贡献来实现。在实际应用中，由于深度学习模型算法的不透明性、模型结构的复杂性和应用场景的多样性，全局可解释性比局部可解释性更加困难。因此，局部解释方法比全局解释方法更常见，也更受到广泛关注和研究。经典的局部解释方法包括敏感性分析解释、局部近似解释、梯度反向传播解释、特征反演解释以及类激活映射解释等[88]。

研究表明，DCNN 不同层的卷积滤波器包含大量的位置信息，使其具备良好的定位能力。基于卷积滤波器的定位能力，可以定位输入样本中用于 DCNN 决策的核心区域，如分类任务中的核心决策特征。但传统的 DCNN 通常采用全连接层对卷积层提取的特征进行组合用于最终的决策，从而导致模型的定位能力丧失。Zhou 等人[93]提出的类激活映射（Class Activation Mapping，CAM）解释方法解决了这一问题。CAM 方法利用全局平均池化（Global Average Pooling）层来替代传统 DCNN 模型中除 Softmax 层以外的所有全连接层，并通过将输出层的权重投影到卷积特征图来识别图像中的重要区域，但是 CAM 方法需要修改网络结果并重新训练模型，在实际应用中并不实用。Selvaraju 等人[94]对 CAM 方法进行了改进，提出了一种将梯度信息与特征映射相结合的梯度加权类激活映射方法——Grad-CAM。Grad-CAM 是一种用于定位图像中类别判别关键区域的可解释性方法。给定一个输入样本，Grad-CAM 首先计算目标类别相对于最后一个卷积层中每个特征图的梯度，然后对梯度进行全局平均池化以获得每个特征图的重要性权重。然后，根据这些权重对特征图进行加权，得到一个粗粒度的梯度加权类激活图，从而定位输入样本中与类别判别相关的重要区域。与 CAM 方向相比，Grad-CAM 无需修改网络结构或重新训练模型，避免了解释性和准确性之间的平衡，并且适用于不同的任务和任何 DCNN 模型架构。尽管 Grad-CAM 具有良好的类别判别能力并能够准确地定位相关图像区域，但是它无法提供如 DeconvNet[95]和 Guided BP[96]等算法像素级别梯度可视化解释方法所能提供的对细粒度特征的精细定位。为了获得更精细的特征定位，研究者将 Grad-CAM 与 Guided BP 方法相结合，提出了导向梯度加权类激活映射方法——Guided Grad-CAM[94]。该方法结合了 Grad-CAM 和 Guided BP 的优点，不仅能够进行类别判别和区域定位，还能提供更细粒度的特征重要性分析，从而使得模型的解释能力更加完整。Guided Grad-CAM 首先将梯度加权类激活图上采样到输入图

片大小，然后与 Guided BP 方法的输出结果进行点乘，从而得到更细粒度的类别判别特征定位图。研究表明，Guided Grad-CAM 方法的解释效果优于 Guided BP 和 Grad-CAM。

5.2.2 基于模型可解释技术的模型分析

在此研究中，通过 Grad-CAM 模型可解释性方法和 Guided Grad-CAM 模型可解释方法，来探索基线模型和此研究提出的模型对原始样本和对抗样本分类决策过程，寻找决定图像类别的关键特征，同时分析比较两个模型用于决策的关键特征的差异。对可视化的结果进行筛选，得到了两种模型可解释方法可视化图，如图 5.3 所示。

图中左边三列为 Grad-CAM 方法对基线模型和此研究提出模型的原始样本和对抗样本决策关键特征的可视化结果，右边三列为 Guided Grad-CAM 方法。从上到下，依次是原始图像、基线模型对原始样本分类的关键特征可视化图、基线模型对对抗样本分类的关键特征可视化图、此研究提出的模型对原始样本分类的关键特征可视化图、此研究提出的模型对对抗样本关键特征可视化图。

在两种可解释方法下，比较基线模型和此研究提出的模型，发现此研究提出的模型对图像分类关键特征的定位更为准确。例如图 5.3 中第三列，Grad-CAM 方法中，基线模型对原始图像分类的关键特征定位到了狗的头部，但是对对抗样本分类的关键特征却定位到了架子，完全避开了狗的头部；而此研究提出的模型不管是对原始图像还是对抗样本，均能正确定位到分类的关键特征。图 5.3 中第五列，Guided Grad-CAM 方法下，基线模型对原始图像的分类关键特征定位到了书架，但是对对抗样本分类的关键特征定位到了人，而原始样本的标签是“书橱”。

在更为精细的 Guide Grad-CAM 方法中，也表现出对边缘、线性特征的偏向。例如图中第四列，Guide Grad-CAM 方法中，对比此研究提出模型与基线模型，此研究提出的模型无论是原始图像还是对抗样本，垃圾桶的边缘特征都要更清晰。在图像低层特征提取的介绍中，SAG 卷积核主要面向图像边缘特征提取，SAG 卷积核主要面向图像线条特征提取。此研究提出的模型对边缘特征和线条特征的偏向进有力证明了融合到前端的模拟 V1 区方位选择感受野的 SAG 和 FAG 卷积核有效性。

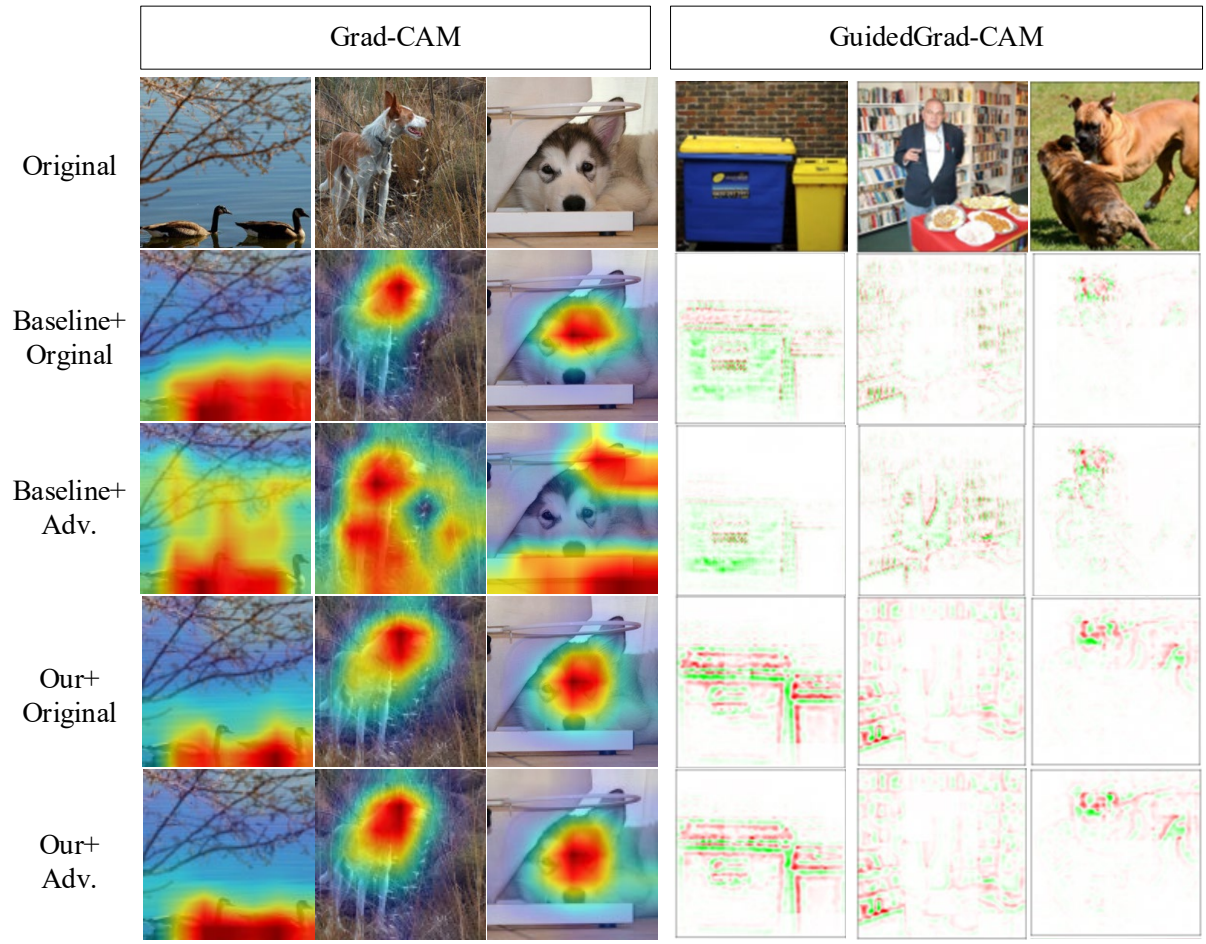


图 5.3 模型对原始样本和对抗样本分类关键特征的可视化图

5.3 本章小结

此研究的消融实验证明了研究中提出的生物可信卷积滤波器组对模型对抗鲁棒性的贡献，同时也证明了 DCNN 模型的自我学过程无法学习到参数固定的生物可信卷积滤波器组为模型带来的对抗鲁棒性。初级视皮层启发前端的三个组成部分：卷积层、非线性层、V1 神经随机层有机组合，共同提高模型的对抗鲁棒性。本节还选择了两种基于梯度的局部可解释性技术：Grad-CAM 和 Guided Grad-Cam，来对模型进行可视化分析。两种可解释方法中，不管是原始样本还是对抗样本，此研究提出的模型对分类决策的关键特征定位均更准确。在更为精细的 Guided Grad-CAM 方法中，此研究提出的模型也表现出对边缘、线条特征的偏向。

总结与展望

总结

面对严重威胁 DCNN 模型安全性的对抗攻击问题，此研究受初级视皮层简单细胞方位选择感受野的启发，将 V1 区简单细胞方位选择感受野、视网膜和 V1 区环绕调制机制、模拟 V1 区简单细胞神经响应的 Gabor 滤波器融合到了 DCNN 的前端低层，增加 DCNN 感受野的异质性，丰富 DCNN 低层特征层，并提高末模型对对抗扰动噪声的鲁棒性。此研究的主要工作和创新点如下：

（1）在此研究中，通过多尺度各向异性高斯核，将初级视皮层简单细胞方位选择感受野引入到 DCNN 的第一层，提出了初级视皮层启发的鲁棒 DCNN 模型。模型由生物可信卷积神经网络前端和标准 CNN 层搭建的卷积神经网络后端组成，前后端之间通过一个由1×1卷积构成的瓶颈层连接。借助瓶颈层，来自前端的对抗鲁棒性可以很容易嵌入到不同的后端网络中，从而实现对抗鲁棒性的迁移。模型的主要优势是 V1 区方位选择感受野启发的前端。前端由三层组成：卷积层、非线性层和 V1 神经随机层。卷积层包含一系列生物视觉模型约束的数学参数化的卷积滤波器组，这些卷积滤波器组与预处理后的输入图像进行图像卷积运算后得到包含图像低层特征的卷积特征图。非线性层包含两种非线性：线性整流变换的简单细胞非线性和正交相位对频谱功率的复杂细胞非线性。两种非线性分别应用于不同的卷积通道。V1 神经随机层包含一个 V1 神经噪声生成器，将独立的高斯噪声添加到每个卷积通道，以模拟 V1 区神经元的随机性。

（2）在 CIFAR-10、CIFAR-100、Mini-ImageNet 和 ImageNet 数据集上的对抗鲁棒性测试实验中，此研究提出的模型表现优异。在无目标白盒对抗鲁棒性实验中，此研究提出的模型远超基线模型，性能超过最先进的神经科学启发模型 VOneNet。在无目标黑盒对抗鲁棒性实验中，此研究提出的模型几乎免疫黑盒攻击算法。在目标白盒对抗鲁棒性实验中，此研究提出的模型性能也显著高于基线模型，在四个数据集上均超过 VOneNet。此外，研究进行的无界攻击测试和一系列对抗鲁棒性评估，也保证了模型没有遭受错误或者不完整的对抗鲁棒性测试。为进一步提高模型的性能，研究中将基于训练过程的数据增强与此研究提出的模型结合。实验结果表明，与数据增强的结合也能够进一步提高模型的对抗鲁棒性。

（3）消融研究也验证了此研究提出的前端对模型对抗鲁棒性的贡献。生物可信卷积滤波器组和 V1 神经随机性有机组合，共同提高模型对抗鲁棒性，且来自前端生物可信卷积滤波器组的对抗鲁棒性无法通过模型训练过程中的自我学习

获得。在基于模型可解释性技术的模型可视化分析中，不论是原始样本还是对抗样本，此研究提出的模型均能够更准确的定位到分类关键特征，且此研究提出的模型表现出对边缘、线条等低层特征的偏向。

展望

虽然此研究构建的方法能够显著提高 DCNN 模型对对抗扰动噪声的鲁棒性，但是研究中还存在一些不足：

（1）此研究只加入了两种方位选择感受野模型，种类较少。对于“非对称单侧对比度敏感型”方位选择感受野，因为缺乏相关的建模计算，无法将其用于增强 DCNN 的感受野。

（2）在无目标白盒对抗鲁棒性实验中，此研究提出模型的原始图像分类准确率下降，或许是因为低层特征融合方式。后续的工作将探索更为合理的低层特征融合方式，使模型对抗鲁棒性的提高不以牺牲原始图像分类精度为代价。

（3）此研究的消融实验虽然证明了生物可信卷对模型对抗鲁棒性的贡献，却无法细致分析每种生物可信卷积的作用。在后续的研究中，将设计更合理的消融实验，详细分析每种生物可信卷积的作用，为模型的进一步优化提供策略。

参考文献

[1] Xu H, Gao Y, Yu F, et al. End-to-end learning of driving models from large-scale video datasets[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2174-2182.

[2] Xu Z, Zhang T, Zeng Y, et al. A Secure Mobile Payment Framework Based On Face Authentication[C]// Proceedings of the International MultiConference of Engineers and Computer Scientists. 2015, 1.

[3] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[4] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

[5] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[6] Simonyan K , Zisserman A . Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014.

[7] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[8] He K, Zhang X, Ren S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification[C]// Proceedings of the IEEE international conference on computer vision. 2015: 1026-1034.

[9] Carlini N, Wagner D. Towards evaluating the robustness of neural networks[C]// Proceedings of the IEEE symposium on security and privacy (sp). IEEE, 2017: 39-57.

[10] Yakura H, Sakuma J. Robust audio adversarial example for a physical attack [C]// Proceedings of the 28th International Joint Conference on Artificial Intelligence. 2019.

[11] Liu X, Cheng M, Zhang H, et al. Towards robust neural networks via random self-ensemble[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018: 369-385.

[12] Xu W, Evans D, Qi Y. Feature squeezing: Detecting adversarial examples in deep neural networks[J]. Proceedings of the 25th Annual Network and Distributed System Security Symposium (NDSS). 2018.

[13] Madry A, Makelov A, Schmidt L, et al. Towards Deep Learning Models Resistant to Adversarial Attacks[J]. stat, 2019, 1050: 4.

[14] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. stat, 2015, 1050: 20.

[15] Ren K, Zheng T, Qin Z, et al. Adversarial attacks and defenses in deep learning[J]. Engineering, 2020, 6(3): 346-360.

[16] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, 2009: 248-255.

[17] Dapello J, Marques T, Schrimpf M, et al. Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations[J]. Advances in Neural Information Processing Systems, 2020, 33: 13073-13087.

[18] Vuyyuru M R, Banburski A, Pant N, et al. Biologically inspired mechanisms for adversarial robustness[J]. Advances in Neural Information Processing Systems, 2020, 33: 2135-2146.

[19] Strisciuglio N, Lopez-Antequera M, Petkov N. Enhanced robustness of convolutional networks with a push–pull inhibition layer[J]. Neural Computing and Applications, 2020, 32: 17957-17971.

[20] Zhuang C, Yan S, Nayebi A, et al. Unsupervised neural network models of the ventral visual stream[J]. Proceedings of the National Academy of Sciences, 2021, 118(3): e2014196118.

[21] Marblestone A H, Wayne G, Kording K P. Toward an integration of deep learning and neuroscience[J]. Frontiers in computational neuroscience, 2016: 94.

[22] Hassabis D, Kumaran D, Summerfield C, et al. Neuroscience-inspired artificial intelligence[J]. Neuron, 2017, 95(2): 245-258.

[23] Geirhos R, Rubisch P, Michaelis C, et al. ImageNet-trained CNNs are

biased towards texture; increasing shape bias improves accuracy and robustness[C]// International Conference on Learning Representations. 2018.

[24] Cadena S A, Denfield G H, Walker E Y, et al. Deep convolutional models improve predictions of macaque V1 responses to natural images[J]. PLoS computational biology, 2019, 15(4): e1006897.

[25] Goldstein E B, Cacciamani L. Sensation and perception[M]. Cengage Learning, 2021.

[26] Kremkow J, Jin J, Wang Y, et al. Principles underlying sensory map topography in primary visual cortex[J]. Nature, 2016, 533(7601): 52-57.

[27] 魏佳璇, 杜世康, 于志轩, 等. 图像分类中的白盒对抗攻击技术综述[J]. 计算机应用, 2022, 42(9): 2732.

[28] Papernot N, McDaniel P, Wu X, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]// Proceedings of the IEEE symposium on security and privacy (SP). IEEE, 2016: 582-597.

[29] Kurakin A, Goodfellow I J, Bengio S. Adversarial examples in the physical world[M]// Artificial intelligence safety and security. Chapman and Hall/CRC, 2018: 99-112.

[30] Dong Y, Liao F, Pang T, et al. Boosting adversarial attacks with momentum[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 9185-9193.

[31] Xie C, Zhang Z, Zhou Y, et al. Improving transferability of adversarial examples with input diversity[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 2730-2739.

[32] Moosavi-Dezfooli S M, Fawzi A, Frossard P. Deepfool: a simple and accurate method to fool deep neural networks[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2574-2582.

[33] Hayes J, Danezis G. Learning universal adversarial perturbations with generative models[C]// Proceedings of the IEEE Security and Privacy Workshops (SPW). IEEE, 2018: 43-49.

[34] Xiao C, Li B, Zhu J Y, et al. Generating adversarial examples with

adversarial networks[C]// Proceedings of the 27th International Joint Conference on Artificial Intelligence. 2018: 3905-3911.

[35] Andriushchenko M, Croce F, Flammarion N, et al. Square attack: a query-efficient black-box adversarial attack via random search[C]// Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII. Cham: Springer International Publishing, 2020: 484-501.

[36] Sharif M, Bhagavatula S, Bauer L, et al. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition[C]// Proceedings of the 2016 acm sigsac conference on computer and communications security. 2016: 1528-1540.

[37] Eykholt K, Evtimov I, Fernandes E, et al. Robust physical-world attacks on deep learning visual classification[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1625-1634.

[38] 徐金才, 任民, 李琦, 等. 图像对抗样本的安全性研究概述[J]. 信息安全研究, 2021, 7(4): 294.

[39] Liu Z, Liu Q, Liu T, et al. Feature distillation: Dnn-oriented jpeg compression against adversarial examples[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019: 860-868.

[40] Jia X, Wei X, Cao X, et al. Comdefend: An efficient image compression model to defend adversarial examples[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 6084-6092.

[41] Raff E, Sylvester J, Forsyth S, et al. Barrage of random transforms for adversarially robust defense[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 6528-6537.

[42] Prakash A, Moran N, Garber S, et al. Deflecting adversarial attacks with pixel deflection[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8571-8580.

[43] Mustafa A, Khan S H, Hayat M, et al. Image super-resolution as a defense against adversarial attacks[J]. IEEE Transactions on Image Processing, 2019, 29: 1711-1724.

[44] Osadchy M, Hernandez-Castro J, Gibson S, et al. No bot expects the

DeepCAPTCHA! Introducing immutable adversarial examples, with applications to CAPTCHA generation[J]. IEEE Transactions on Information Forensics and Security, 2017, 12(11): 2640-2653.

[45] Liao F, Liang M, Dong Y, et al. Defense against adversarial attacks using high-level representation guided denoiser[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1778-1787.

[46] Xie C, Wang J, Zhang Z, et al. Mitigating adversarial effects through randomization[C]// International Conference on Learning Representations.2017.

[47] Guo C, Rana M, Cisse M, et al. Countering adversarial images using input transformations[C]// International Conference on Learning Representations.2017.

[48] Athalye A, Engstrom L, Ilyas A, et al. Synthesizing robust adversarial examples[C]// Proceedings of the International conference on machine learning. PMLR, 2018: 284-293.

[49] Dhillon G S, Azizzadenesheli K, Lipton Z C, et al. Stochastic activation pruning for robust adversarial defense[C]// International Conference on Learning Representations.2018.

[50] Ross A, Doshi-Velez F. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).

[51] Athalye A, Carlini N, Wagner D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples[C]// Proceedings of the International conference on machine learning. PMLR, 2018: 274-283.

[52] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. stat, 2015, 1050: 9.

[53] Samangouei P, Kabkab M, Chellappa R. Defense-gan: Protecting classifiers against adversarial attacks using generative models[C]// International Conference on Learning Representations.2018.

[54] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. nature, 2015, 521(7553): 436-444.

[55] Hendrycks D, Dietterich T. Benchmarking neural network robustness to

common corruptions and perturbations[C]// International Conference on Learning Representations. 2019.

[56] 曾毅, 刘成林, 谭铁牛. 类脑智能研究的回顾与展望[J]. 计算机学报, 2016, 39(1): 212-222.

[57] Kubilius J, Schrimpf M, Kar K, et al. Brain-like object recognition with high-performing shallow recurrent ANNs[J]. Advances in neural information processing systems, 2019, 32.

[58] Lindeberg T. Provably scale-covariant continuous hierarchical networks based on scale-normalized differential expressions coupled in cascade[J]. Journal of Mathematical Imaging and Vision, 2020, 62(1): 120-148.

[59] Li Z, Brendel W, Walker E, et al. Learning from brains how to regularize machines[J]. Advances in neural information processing systems, 2019, 32.

[60] Safarani S, Nix A, Willeke K, et al. Towards robust vision by multi-task learning on monkey visual cortex[J]. Advances in Neural Information Processing Systems, 2021, 34: 739-751.

[61] Hubel D H, Wiesel T N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex[J]. The Journal of physiology, 1962, 160(1): 106.

[62] Livingstone M, Hubel D. Segregation of form, color, movement, and depth: anatomy, physiology, and perception[J]. Science, 1988, 240(4853): 740-749.

[63] Herreras E B. Cognitive neuroscience; The biology of the mind[J]. Cuadernos de Neuropsicología/Panamerican Journal of Neuropsychology, 2010, 4(1): 87-90.

[64] Kremkow J, Jin J, Wang Y, et al. Principles underlying sensory map topography in primary visual cortex[J]. Nature, 2016, 533(7601): 52-57.

[65] 寿天德, 杨雄里. 视觉信息处理的脑机制[M]. 上海科技教育出版社, 1997.

[66] Sofka M, Stewart C V. Retinal vessel centerline extraction using multiscale matched filters, confidence and edge measures[J]. IEEE transactions on medical imaging, 2006, 25(12): 1531-1546.

[67] Lindeberg T. Feature detection with automatic scale selection[J]. International journal of computer vision, 1998, 30(2): 79-116.

[68] Chen R, Wang F, Liang H, et al. Synergistic processing of visual contours

across cortical layers in V1 and V2[J]. Neuron, 2017, 96(6): 1388-1402. e4.

[69] Canny J. A computational approach to edge detection[J]. IEEE Transactions on pattern analysis and machine intelligence, 1986 (6): 679-698.

[70] Geusebroek J M, Smeulders A W M, Van De Weijer J. Fast anisotropic gauss filtering[J]. IEEE transactions on image processing, 2003, 12(8): 938-943.

[71] Shui P L, Zhang W C. Noise-robust edge detector combining isotropic and anisotropic Gaussian kernels[J]. Pattern Recognition, 2012, 45(2): 806-820.

[72] Wang F P, Shui P L. Noise-robust color edge detector using gradient matrix and anisotropic Gaussian directional derivative matrix[J]. Pattern Recognition, 2016, 52: 346-357.

[73] Zhang W C, Zhao Y L, Breckon T P, et al. Noise robust image edge detection based upon the automatic anisotropic Gaussian kernels[J]. Pattern Recognition, 2017, 63: 193-205.

[74] WANG G, De Baets B. Edge detection based on the fusion of multiscale anisotropic edge strength measurements[C]// 10th Conference of the European-Society-for-Fuzzy-Logic-and-Technology (EUSFLAT)/16th International Workshop on Intuitionistic Fuzzy Sets and Generalized Nets (IWIFSGN). Springer, 2018, 643: 530-536.

[75] Wang G, Lopez-Molina C, De Baets B. Multiscale edge detection using first-order derivative of anisotropic Gaussian kernels[J]. Journal of Mathematical Imaging and Vision, 2019, 61(8): 1096-1111.

[76] Lopez-Molina C, De Ulzurrun G V D, Baetens J M, et al. Unsupervised ridge detection using second order anisotropic Gaussian kernels[J]. Signal Processing, 2015, 116: 55-67.

[77] Wang G, Lopez-Molina C, De Baets B. Blob reconstruction using unilateral second order Gaussian kernels with application to high-ISO long-exposure image denoising[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 4817-4825.

[78] Wang G, Lopez-Molina C, de Ulzurrun G V D, et al. Noise-robust line detection using normalized and adaptive second-order anisotropic

Gaussian kernels[J]. Signal Processing, 2019, 160: 252-262.

[79] Babaiee Z, Hasani R, Lechner M, et al. On-off center-surround receptive fields for accurate and robust image classification[C]// International Conference on Machine Learning. PMLR, 2021: 478-489.

[80] Hasani H, Soleymani M, Aghajan H. Surround modulation: A bio-inspired connectivity structure for convolutional neural networks[J]. Advances in neural information processing systems, 2019, 32.

[81] Rust N C, Schwartz O, Movshon J A, et al. Spatiotemporal elements of macaque v1 receptive fields[J]. Neuron, 2005, 46(6): 945-956.

[82] El-Shamayleh Y, Kumbhani R D, Dhruv N T, et al. Visual response properties of V1 neurons projecting to V2 in macaque[J]. Journal of Neuroscience, 2013, 33(42): 16594-16605.

[83] Softky W R, Koch C. The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs[J]. Journal of neuroscience, 1993, 13(1): 334-350.

[84] Vinyals O, Blundell C, Lillicrap T, et al. Matching networks for one shot learning[J]. Advances in neural information processing systems, 2016, 29.

[85] Zhang H, Cisse M, Dauphin Y N, et al. mixup: Beyond empirical risk minimization[C]// International Conference on Learning Representations. 2017.

[86] Croce F, Hein M. Minimally distorted adversarial examples with a fast adaptive boundary attack[C]// International Conference on Machine Learning. PMLR, 2020: 2196-2205.

[87] Cubuk E D, Zoph B, Shlens J, et al. Randaugment: Practical automated data augmentation with a reduced search space[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020: 702-703.

[88] 纪守领, 李进锋, 杜天宇, 等. 机器学习模型可解释性方法, 应用与安全研究综述[J]. 计算机研究与发展, 2019, 56(10): 2071-2096.

[89] Alvarez Melis D, Jaakkola T. Towards robust interpretability with self-explaining neural networks[J]. Advances in neural information processing systems, 2018, 31.

[90] Tickle A B, Orlowski M, Diederich J. DEDEC: A methodology for extracting rules from trained artificial neural networks[M].

Neurocomputing Research Centre, Queensland University of Technology, 1996.

[91] Liu X, Wang X, Matwin S. Improving the interpretability of deep neural networks with knowledge distillation[C]// Proceedings of the IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2018: 905-912.

[92] Zhang Q, Wang W, Zhu S C. Examining CNN representations with respect to dataset bias[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).

[93] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2921-2929.

[94] Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]// Proceedings of the IEEE international conference on computer vision. 2017: 618-626.

[95] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]// Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13. Springer International Publishing, 2014: 818-833.

[96] Springenberg J T, Dosovitskiy A, Brox T, et al. Striving for simplicity: The all convolutional net[J]. International Conference on Learning Representations. 2015.