第 58 卷 第 22 期/2021 年 11 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于信息熵的卷积神经网络图像特征度量

陈文俊，丛超*，黄丽雯

重庆理工大学电气与电子工程学院，重庆 400054

摘要 在对卷积神经网络(CNN)的可解释性研究中，针对特征信息的定量分析是研究的重点。提出一种基于信息熵的特征度量方法，用于定量分析 CNN 的特征提取性能。该方法首先针对不同类别的图像计算特征层的激活直方图，然后统计其归一化熵，将其定义为特征纯度。在 CIFAR10 和 ImageNet 数据集上对不同 CNN 模型及其内部结构的特征纯度进行量化评估；在评估特征纯度的同时，结合类激活图进行可视化解释，以验证特定 CNN 模型的特征提取能力和特征纯度之间的关系。实验结果表明，特征纯度与模型的特征提取能力、模型的分类性能有统计意义上的显著相关性；同时，特征纯度的计算不依赖于分类标签，也不局限于具体网络结构，具有较强的鲁棒性与实用性。所提出的量化评估方法可以有效地评估 CNN 的特征提取性能。

关键词 成像系统；卷积神经网络；特征度量；图像特征；网络可解释性；信息熵

中图分类号 TP391 **文献标志码** A **doi**：10.3788/LOP202158.2211004

Convolutional Neural Network Image Feature Measurement Based on Information Entropy

Chen Wenjun, Cong Chao*, Huang Liwen

College of Electrical and Electronic Engineering, Chongqing University of Technology, Chongqing 400054, *China*

Abstract In the interpretability research of convolutional neural network (CNN), the quantitative analysis of feature information is the focus of the research. In this paper, we proposed a feature measurement method based on information entropy to quantitatively analyze the performance of feature extraction in CNN. Firstly, aiming to the different types of images, the activation histogram of the feature layer is calculated, and then the normalized entropy is calculated to define the characteristic purity. Different CNN models and their feature purity of internal structures were quantitatively evaluated on CIFAR10 and ImageNet datasets. At the same time, visual interpretation was performed by combining class activation maps to verify the relationship between feature extraction ability and feature purity of a specific CNN model. The experimental results show that the feature purity is significantly correlated with the feature extraction ability and classification performance of the model. At the same time, the calculation of feature purity is not dependent on the classification label, is not limited to the specific network structure, and has strong robustness and practical value. The proposed quantitative evaluation method can effectively evaluate the feature extraction performance of CNN.

Key words imaging systems; convolutional neural networks; feature metrics; image feature; network interpretability; information entropy

OCIS codes 110.4155; 100.4996; 100.2960

收稿日期：2020-12-16；**修回日期**：2021-01-11；**录用日期**：2021-02-12

基金项目：重庆市教委科学技术研究项目(KJQN202001131)、重庆理工大学研究生创新项目资助(clgycx 20202040)

通信作者：*chenwj@2019.cqut.edu.cn

1 引 言

近几年，深度学习技术在计算机视觉领域快速发展，应用于基础的图像分类、目标检测[1-3]、语义分割[4-6]等，其性能优于基于人工设计的视觉特征的传统方法。然而，在深度学习领域，神经网络一直是“黑盒”，其较差的算法可解释性与复杂内部运行机制是其缺陷。神经网络可解释性的提高，不但可使工程师对其技术原理有深入的理解，而且成为深度学习在业务上落地并被人们普遍接受的重要推动力。因此，近年来神经网络的可解释性得到高度的关注，许多研究者开始着手研究“黑盒”内部的工作机理。

神经网络的可解释性研究方法已经经历了多个阶段[7-8]，经典的方法包括反卷积[9]和导向反向传播[10]等，它们对神经网络浅层和深层的特征进行了阐述。更进一步地，Zhou 等[11]利用 CAM（Class Activation Mapping）将全连接层替换为全局池化层，经重新训练得到权重，并得到图像中每个部分对类别的重要程度，进而对神经网络的分类结果进行可视化解释。Selvaraju 等[12]提出 Grad-CAM（Gradient-weighted Class Activation Mapping），其无需替换全连接层，采用梯度的全局平均来计算权重，得到图像像素对类别的重要程度。最近，Sturmfels 等[13]提出积分梯度法，通过对梯度沿不同路径进行积分，期望得到非饱和区的非零梯度对决策重要性的贡献，以解决梯度饱和问题并提高模型的可见性。

另一方面有研究者利用通用手段或工具来对神经网络进行解释，LIME（Local Interpretable Model-Agnostic Explanation）[14]通过在复杂神经网络分类模型的局部拟合出一个简单的可解释模型（线性分类）来解释神经网络。Influence function[15]通过研究训练样本变化（如减少一个样本、对样本进行扰动）对模型预测的影响，对神经网络进行解释。Wang 等[16]分析了卷积神经网络的泛化行为与图像数据集频谱之间的关系。Liang 等[17]通过知识一致性来判别卷积神经网络中间层的表征能力。Ma 等[18]通过使用基于信息熵的信息丢失和重建的不确定度来对网络进行解释。

目前针对神经网络可解释性的研究工作还存在一些亟待解决的问题。现有的研究主要通过对卷积层提取的单一类别的特征进行可视化解释，或从模型的局部和通过研究训练样本对神经网络进行解释。对卷积层提取的全局特征信息进行评估以及制定相应的评估准则有助于更进一步理解和评估神经网络模型。因此，本文提出一种基于信息熵的卷积神经网络图像特征度量方法，主要通过评估卷积层特征提取性能来度量全局特征信息。在模型内、模型间以及不同训练程度模型的卷积层评估了特征纯度的有效性，并使用 Grad-CAM 进行简单的可视化解释。实验结果表明，卷积层的特征纯度越高，其提取性能越优，对各类的定位能力越好；同时，随着训练的进行，卷积层特征提取性能逐渐提升，特征纯度也逐渐提高。

2 基本原理

2.1 特征度量

熵是衡量信息理论中的无序性或不确定性的常用度量。同时，信息熵常被用作系统信息内容的定量指标。较大的熵值意味着系统包含更多的信息。基于此，在卷积神经网络中，如果卷积层特征提取性能较好，那么其包含的特征信息就越多。

因此，本研究提出基于特征纯度对特征层的全局特征信息进行评估，特征纯度定义为某一特征层各神经元对各类图像激活概率的归一化信息熵。在神经网络中，如果神经元尽可能地被特定类别的图像所激活，不同的神经元所表达的跨类信息重叠越小，那么其对各类图像的激活就越均匀，信息熵就越大，即这一层所提取的特征信息就越丰富，特征纯度越高，如图 1(a)所示。若神经元未充分激活各类图像，特征提取性能较差，那么其特征纯度就相对较低，如图 1(b)所示。在神经元相对较多的深层，相对应地，可以认为某几个神经元学习到某一类图像的不同特征，其特征提取性能更好，学习到的特征更加丰富，更有助于模型决策，其特征纯度也相应较高，如图 1(c)所示，其中不同形状代表被不同类图像激活，虚线圈代表未激活。

2.2 特征度量算法实现

特征纯度计算的具体步骤如图 2 所示，其中 c、h、w 分别代表特征层的通道数、长度和宽度，n 分别代表图片数量，y 代表卷积神经网络的输出，m 代表预测的类；⊙代表特征图与对应权重的点乘。

对于卷积层 $B\in\mathbf{R}^{c\times h\times w}$，定义其中第 c 张特征图对类别 m 的权重为 $\beta_c^{(m)}$，其计算式为

$$\beta_c^{(m)}=\frac{1}{Z}\sum_i\sum_j\frac{\partial y^{(m)}}{\partial B_{ij}^{(c)}},\tag{1}$$

式中：Z 为特征图的像素个数；$y^{(m)}$ 是对应类别 m 的得分；$B_{ij}^{(c)}$ 表示第 c 张特征图中 (i,j) 位置处的像素值。

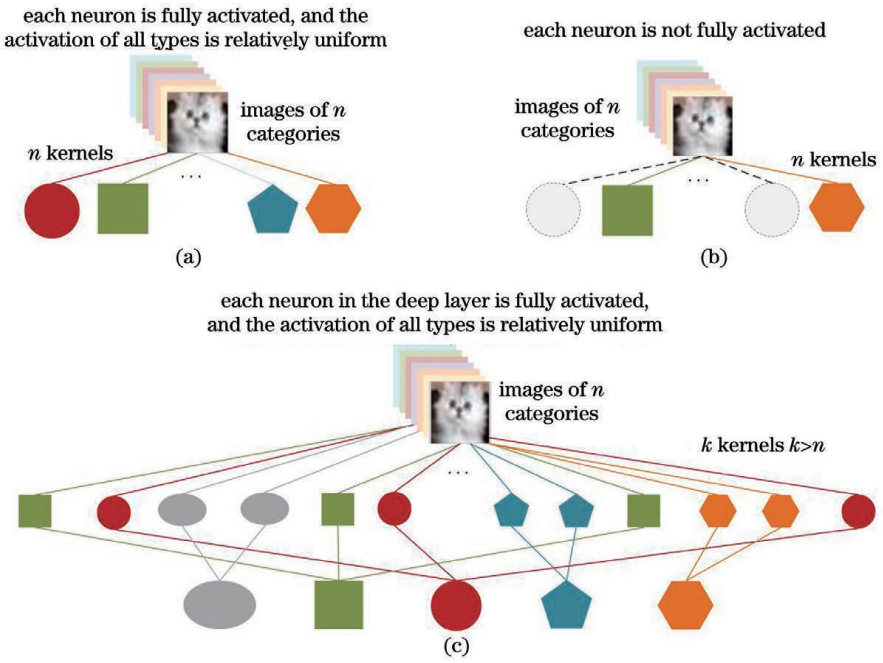

图 1　特征层不同激活情况示例。(a)充分激活神经元示例；(b)未充分激活神经元示例；(c)充分激活深层神经元示例

Fig. 1　Examples of different activations of feature layers. (a) Example of fully activated neurons; (b) example of insufficient activation of neurons; (c) example of full activation of deep layer neurons

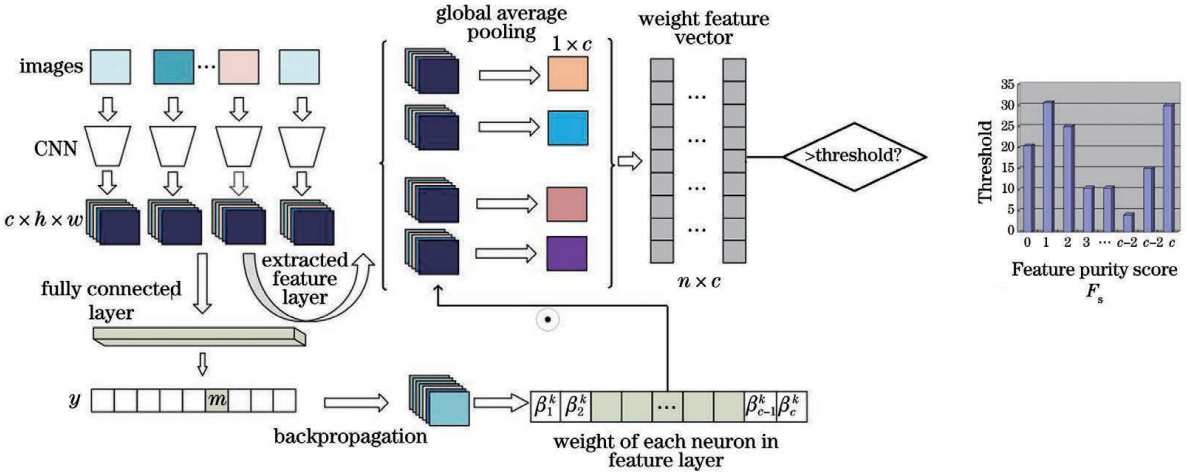

图 2　基于信息熵的图像特征度量算法流程图

Fig. 2　Flow chart of image feature measurement algorithm based on information entropy

对于所有特征图，可以得到权重矩阵 $\boldsymbol{\beta}^{(m)}$，$\boldsymbol{\beta}^{(m)}\in\mathbf{R}^{1\times c}$，然后将权重一一乘上对应的特征图，得到权重特征图 $\boldsymbol{L}\in\mathbf{R}^{c\times h\times w}$：

$$\boldsymbol{L}=\boldsymbol{\beta}^{(m)}B。\tag{2}$$

将权重特征图进行全局平均池化，得到特征激活矩阵 $\boldsymbol{M}\in\mathbf{R}^{1\times c}$，可以表示为

$$\boldsymbol{M}=\frac{1}{Z}\sum_{i=0}^{h}\sum_{j=0}^{w}\boldsymbol{L},\tag{3}$$

n 张图像对应的特征激活矩阵 $\boldsymbol{M}_n=\mathbf{R}^{n\times c}$，然后设定阈值以统计 n 张图像每个通道的激活次数，得到关于 c 个特征图的激活直方图。

通过激活直方图得到每个神经元的激活概率 p_c 并计算该层的信息熵 H，将其进行归一化，理想情况下，c 个通道激活的信息熵最大值 $H_{\max}=-1/c\cdot\log_2(1/c)$，将计算出的值除以最大值，得到对应特征层的归一化特征纯度得分 F_s：

$$H=-\sum_{c}p_c\cdot\log_2 p_c,\tag{4}$$

$$F_s=H/H_{\max},\tag{5}$$

式中：F_s 的取值范围为 0～1，其值越大，说明特征层内部的神经元分工越精细，对各类图像的激活相对均匀，卷积层提取特征的性能越好，所提取的特征

包含更多的特征信息。理想情况下，当每个通道的激活概率相等：$p_c=1/c$，F_s 达到最大值 1。若其值相对较小，说明各个通道没有被充分地激活，特征提取性能较差。

3　实验方案设计

3.1　数据集及实验设定

实验选择在数据集 CIFAR10[19] 和 ImageNet（ILSVRC-12[20]）上进行，CIFAR10 数据集包含 10 类图像，图像分为训练集和测试集，训练集包含 50000 张照片，测试集包含 10000 张照片。ImageNet 数据集包含 1000 类图像，图像分为训练集、验证集和测试集，训练集包含 1281167 张图片，验证集包含 50000 张图片，测试集包含 100000 张图片。

在 CIFAR10 数据集上，选择 Tiny VGG[21] 和 ResNet18[22] 作为基准模型，并分别对其进行训练。训练时对初始图像进行数据增强，首先在图像外边进行补 0，其次进行随机水平翻转，最后进行随机剪裁，batch_size 设置为 64，学习率设置为 0.001，使用 Adam 优化网络参数。

在 ImageNet（ILSVRC-12）数据集上，本文选择 AlexNet[23]、VGG[24]、DenseNet121[25]、ResNet 及 SENet154[26] 作为实验模型，权重选择 pytorch 官方公布的预训练模型，以保证在每个模型都达到最优性能时进行对比。

特征纯度均选择在各类分布均匀的测试集上计算，并通过 Grad-CAM 对相应特征层提取的特征信息进行可视化。

3.2　阈值的选择

首先在 CIFAR10 和 ImageNet 上分别选择 ResNet18 和 VGG19 模型计算各特征层在不同阈值（0.1～0.9）下的特征纯度得分，观察特征度量的有效性。图 3 展示了 ResNet18 以及 VGG19 各特征层在不同阈值下的特征纯度得分折线图，其中 L 代表 layer，C 代表 conv，F 代表 features。可以看出：当阈值为 0.4～0.5 时，各特征层的特征纯度有明显的区别，可用于较好地对比不同特征层提取特征的性能。同时发现：随着阈值的增加，浅层的纯度降低得较多，深层纯度降低得相对较少。当阈值为 0.4～0.5 时，浅层纯度降低很多，深层纯度几乎没有降低，说明浅层神经元的激活相对较弱，深层神经元的激活相对较强，特征提取性能较好。当阈值大于 0.5 时，特征纯度开始大幅降低（特别是浅层特征层），导致特征度量有效性变差。故在保证特征度量有效性及特征信息尽量不丢失的情况下，选择阈值为 0.5 进行实验。

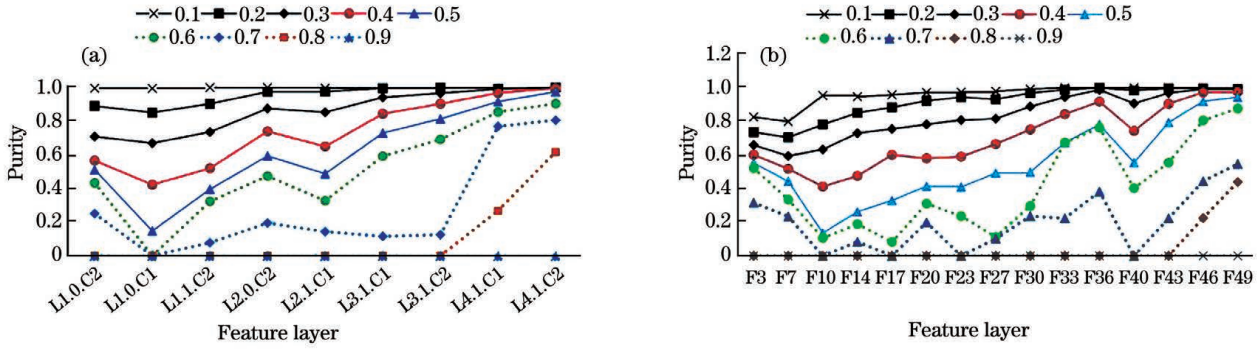

图 3　ResNet18 以及 VGG19 各个特征层在不同阈值下的特征纯度。(a)ResNet18；(b)VGG19

Fig. 3　Feature purity of each feature layer of ResNet18 and VGG19 under different thresholds. (a) ResNet18; (b) VGG19

3.3　模型内各特征层的比较

相应地，在 ResNet18 模型上，选择每个类的图像，通过 Grad-CAM 可视化观察同一模型不同特征层的特征提取性能。图 4 展示了 CIFAR10 各个类对应层的 Grad-CAM 激活图，其中 layer 代表不同的特征层，右侧加粗的数字是阈值在 0.5 时对应特征层的纯度得分。

通过对比特征纯度得分与相应层的 Grad-CAM 激活图，可以发现特征纯度得分比较大的特征层提取的特征信息更丰富，同时对各类图像的激活响应比较好。如图 4 所示，layer4.1.conv2 的纯度得分为 0.971，从 Grad-CAM 类激活图可以看出，Grad-CAM 对各类图像的激活较好，定位能力更好。而对于 layer1.1.conv1、layer1.1.conv2，其纯度得分分别为 0.147 和 0.394，从相应层各类图像的 Grad-CAM 可以看出，其激活性能相对较差，激活区域比较散乱，对类别的定位效果相对较差，这说明神经元没有被充分激活。其中针对 layer4.1.conv1，Grad-CAM 对鸟几乎没有激活，但从总体来看，Grad-CAM 对其他类图像的激活是相对较好的，这总体

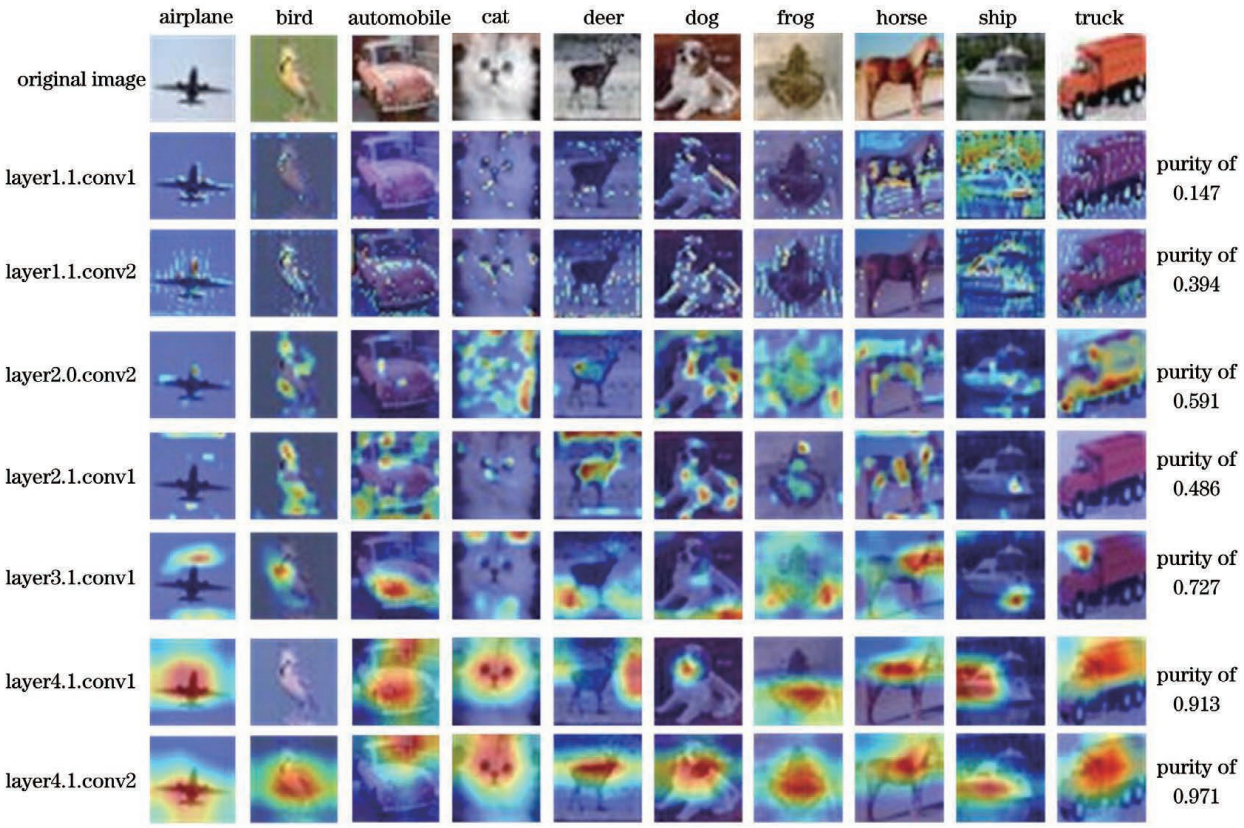

图 4　CIFAR10 数据集中各类图像在 ResNet18 特征层的 Grad-CAM 与特征纯度对比

Fig. 4　Comparison of Grad-CAM and feature purity of ResNet18 feature layer of different types of images in CIFAR10 dataset

上符合反卷积中所述内容，即浅层非线性能力较弱，学习能力相对较弱，提取的是物体的边缘、纹理、轮廓等低级特征，故特征提取性能相对较弱，特征的纯度相对较低，随着层数的增加，特征层非线性能力增强，提取的特征是关于类别的高级语义特征，特征提取性能较好，提取的特征包含更多的特征信息。

3.4　模型间特征层的度量

对于模型间特征提取的度量，本文主要考虑两个方面进行对比：1）选择每个模型下采样前包含低、中、高级语义特征的每个特征层，通过计算其特征纯度的平均来对模型特征提取性能进行评估；2）考虑到深层的特征层包含丰富的高级语义特征，对分类的最终结果影响较大，故对模型的最后一个特征层进行对比。

在 ImageNet（ILSVRC-12）数据集上，选择 VGG、ResNet 等模型进行实验，统计每个模型下采样前以及最后一层特征层的特征纯度并计算它们的平均值，观察其各特征层的特征纯度的得分。表 1 中 C1～C5 代表模型每次下采样前的特征层，average代表5个特征层的纯度平均值。可以看到

表 1　VGG、ResNet 各模型不同特征层的特征纯度对比

Table 1　Comparison of feature purity of different feature layers of VGG and ResNet models

Layer	VGG 13_bn	VGG 16_bn	VGG 19_bn	ResNet34	ResNet50	ResNet101
C1	0.407	0.559	0.550	0.629	0.692	0.644
C2	0.404	0.308	0.135	0.506	0.641	0.739
C3	0.409	0.283	0.407	0.226	0.491	0.575
C4	0.763	0.820	0.777	0.443	0.337	0.515
C5	0.930	0.934	0.939	0.931	0.948	0.952
Average	0.583	0.581	0.562	0.547	0.622	0.639

最后一层的特征纯度普遍较高。随着模型深度的增加，模型性能提升，特征纯度有一定的提升。模型最后一层的特征纯度与性能之间呈正相关，包含的语义信息更加丰富，故选择模型的最后一层来对不同模型的特征纯度进行对比。

随后在 ImageNet（ILSVRC-12）数据集上，选择 AlexNet、VGG16、DenseNet121、ResNet50 以及 SENet154 这几个具有不同性能的模型进行实验。如表 2 所示，性能越好的模型，其最后一个特征层的特征纯度越高，图 5 展示了各模型对于 ImageNet 各类图像的纯度与 Grad-CAM 对比图。可以看出，纯度越高，模型性能越好，模型卷积层内部神经元分工越精细，特征提取性能越好，包含更少的背景信息，定位相对准确。

表 2　跨模型特征纯度在 ImageNet1000 数据集上的对比

Table 2　Comparison of cross-model feature purity on ImageNet1000 dataset

Dataset	AlexNet	VGG16	DenseNet121	ResNet50	SENet154
Accuracy /%	56.43	71.64	74.67	76.00	81.30
Purity	0.512	0.793	0.870	0.948	0.963

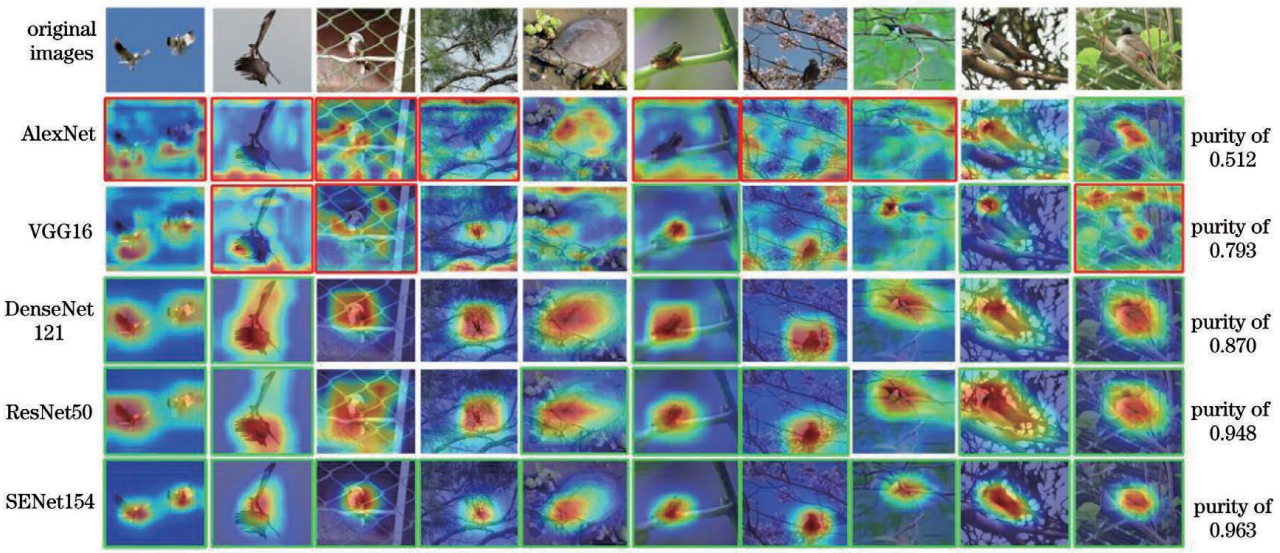

图 5　不同性能模型在 ImageNet1000 上特征层的特征激活图与特征纯度对比

Fig. 5　Comparison of feature activation maps and feature purity of feature layers on ImageNet1000 for models with different performance

3.5　不同训练程度模型的特征纯度对比

更进一步地，为了研究训练过程中不同训练程度模型的纯度变化，在保证训练参数一致的情况下，本研究在 CIFAR10 上用 Tiny VGG 和 ResNet18 对比同一模型在不同训练程度下的特征纯度。分别选择训练 10，20，50，100 轮的模型对比最后一层特征层的特征纯度，如表 3 所示，其中 Tiny VGG_10 及 ResNet18_10 代表训练 10 个 epoch 之后的模型，Tiny VGG _ 20、ResNet18 _ 20、Tiny VGG _ 50、ResNet18_50、Tiny VGG_100、ResNet18_100 代表的含义以此类推。随着训练轮次的增加，模型逐渐优化，准确率逐渐上升，特征纯度也不断增大。如图 6、7 所示，本文使用 Grad-CAM 对 Tiny VGG 和 ResNet18训练不同轮次后的最后一层特征层进行

表 3　ResNet18 模型和 Tiny VGG 模型在不同训练 epoch 下的特征纯度得分对比

Table 3　Comparison of feature purity scores of ResNet18 model and Tiny VGG model under different training epochs

Model	Accuracy	Purity	Model	Accuracy	Purity
Tiny VGG_10	81.25	0.713	ResNet18_10	87.50	0.810
Tiny VGG_20	81.25	0.756	ResNet18_20	87.50	0.826
Tiny VGG_50	85.50	0.823	ResNet18_50	93.75	0.887
Tiny VGG_100	87.50	0.835	ResNet18_100	95.50	0.946

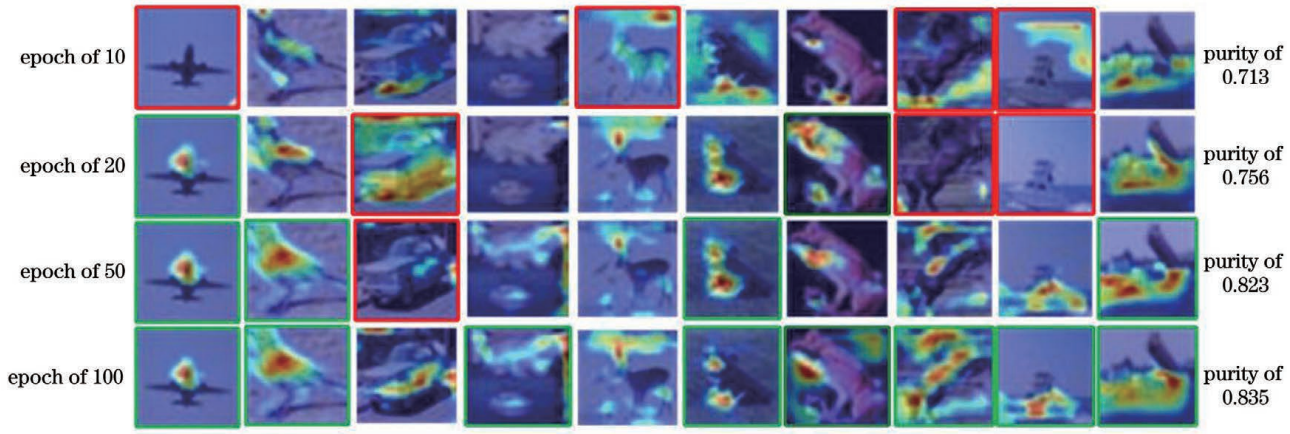

图 6　Tiny VGG 在不同 epoch 下最后一层特征层的 Grad-CAM 与特征纯度对比

Fig. 6　Comparison of Grad-CAM and feature purity of last feature layer under different epochs of Tiny VGG

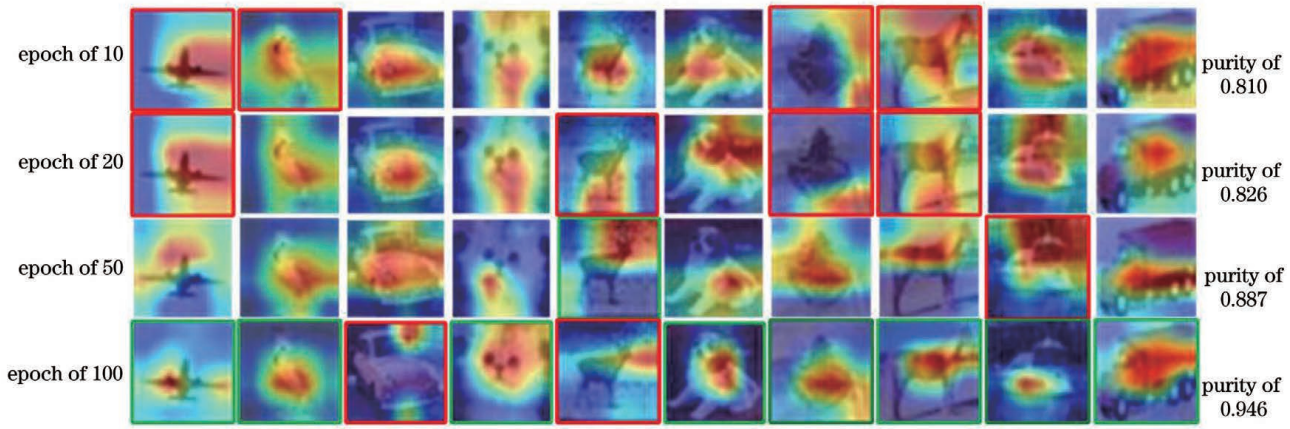

图 7　ResNet18 模型在不同 epoch 下最后一层特征层的 Grad-CAM 与特征纯度对比

Fig. 7　Comparison of Grad-CAM and feature purity of last feature layer of ResNet18 model under different epochs

可视化，观察相应特征层纯度的变化。随着训练程度的加深，Tiny VGG 和 ResNet18 提取的特征逐渐优化，提取的语义信息更好，逐渐排除了背景信息的干扰，对类别的定位能力更好。同时，随着训练轮次的增加，模型存在过拟合的情况，虽然特征纯度有所提升，但特征层提取的某些图像的特征逐渐变差，特征被定位在图像的边缘或者背景。例如，在图 7 中，汽车和鹿在 epoch 为 50 的激活比 epoch 为 100 的激活更好，在 epoch 为 100 时，模型出现过拟合现象，将汽车和鹿定位在了车的边缘和鹿的背景上面。

4　结　　论

提出一种基于信息熵的卷积神经网络图像特征度量方法。在无需标签的情况下，对 CNN 特征提取性能进行了量化评估，这有助于进一步理解 CNN 内部运行机制。在 CIFAR10 和 ImageNet 数据集上进行训练与评估测试。实验结果表明，在针对不同结构的模型以及同一模型的不同训练程度的性能评估上，特征纯度的量化指标与模型的分类性能在统计意义上具有一定的相关性。另一方面，针对 CNN 模型内部结构的特征纯度评估实验也表明了不同深度的特征层在特征提取方面的差异性，层次较深的特征层的特征特化与整体的特征提取能力要明显优于层次较浅的特征层。更进一步地，实验结合了类激活图方法，在进行量化评估的同时，从可视化的角度印证了特征纯度的准确性和有效性。

参　考　文　献

[1]　Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE Press, 2014: 580-587.

[2]　Feng X Y, Mei W, Hu D S. Aerial target detection

based on improved Faster R-CNN[J]. Acta Optica Sinica, 2018, 38(6): 0615004.
冯小雨, 梅卫, 胡大帅. 基于改进 Faster R-CNN 的空中目标检测[J]. 光学学报, 2018, 38(6): 0615004.

[3] Ju M R, Luo H B, Wang Z B, et al. Improved YOLO V3 algorithm and its application in small target detection[J]. Acta Optica Sinica, 2019, 39(7): 0715004.
鞠默然, 罗海波, 王仲博, 等. 改进的 YOLO V3 算法及其在小目标检测中的应用[J]. 光学学报, 2019, 39(7): 0715004.

[4] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 3431-3440.

[5] Wang J, Zhang X Y, Cai Y F, et al. CT image segmentation method combining wavelet transform and RSF model[J]. Acta Optica Sinica, 2020, 40(21): 2110003.
王珏, 张秀英, 蔡玉芳, 等. 联合小波变换和 RSF 模型的 CT 图像分割方法[J]. 光学学报, 2020, 40(21): 2110003.

[6] Yang G L, Lai Z D, Wang Y. Skin lesion image segmentation algorithm based on multi-scale DenseNet[J]. Laser & Optoelectronics Progress, 2020, 57(18): 181020.
杨国亮, 赖振东, 王杨. 基于多尺度密集块网络的皮肤病变图像分割算法[J]. 激光与光电子学进展, 2020, 57(18): 181020.

[7] Cheng K Y, Wang N, Shi W X, et al. Research advances in the interpretability of deep learning[J]. Journal of Computer Research and Development, 2020, 57(6): 1208-1217.
成科扬, 王宁, 师文喜, 等. 深度学习可解释性研究进展[J]. 计算机研究与发展, 2020, 57(6): 1208-1217.

[8] Hua Y Y, Zhang D C, Ge S M. Research progress in the interpretability of deep learning models[J]. Journal of Cyber Security, 2020, 5(3): 1-12.
化盈盈, 张岱墀, 葛仕明. 深度学习模型可解释性的研究进展[J]. 信息安全学报, 2020, 5(3): 1-12.

[9] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[M]//Fleet D, Pajdla T, Schiele B, et al. Computer vision-ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8689: 818-833.

[10] Springenberg J T, Dosovitskiy A, Brox T, et al. Striving for simplicity: the all convolutional net[EB/OL]. (2014-12-21)[2020-12-10]. https://arxiv.org/abs/1412.6806v3.

[11] Zhou B L, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 2921-2929.

[12] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[J]. International Journal of Computer Vision, 2020, 128(2): 336-359.

[13] Sturmfels P, Lundberg S, Lee S I. Visualizing the impact of feature attribution baselines[J]. Distill, 2020, 5(1): 22.

[14] Ribeiro M T, Singh S, Guestrin C. "Why should I trust You?": explaining the predictions of any classifier[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 13-17, 2016, San Francisco, California, USA. New York: ACM, 2016: 1135-1144.

[15] Koh P W, Liang P. Understanding black-box predictions via influence functions[EB/OL]. (2017-03-14)[2020-12-10]. https://arxiv.org/abs/1703.04730.

[16] Wang H H, Wu X D, Huang Z Y, et al. High-frequency component helps explain the generalization of convolutional neural networks[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 8681-8691.

[17] Liang R F, Li T L, Li L F, et al. Knowledge consistency between neural networks and beyond[C]//8th International Conference on Learning Representations (ICLR), April 26-30, 2020, Addis Ababa, Ethiopia. Trier: DBLP, 2020.

[18] Ma H T, Zhang Y Q, Zhou F, et al. Quantifying layerwise information discarding of neural networks[EB/OL]. (2019-06-10)[2020-12-10]. https://arxiv.org/abs/1906.04109.

[19] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images[J]. Handbook of Systemic Autoimmune Diseases, 2009, 1(4): 1-60.

[20] Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[21] Wang Z J, Turko R, Shaikh O, et al. CNN explainer: learning convolutional neural networks

with interactive visualization[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(2): 1396-1406.

[22] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

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

[24] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04)[2020-12-10]. https://arxiv.org/abs/1409.1556.

[25] Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 2261-2269.

[26] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7132-7141.