南京大学 NANJING UNIVERSITY

RINC Robotic Intelligence & Neural Computing Group

从神经元出发的卷积神经网络可解释性研究

Research on the Interpretability of Convolutional Neural Networks from the Perspective of Neurons

答辩人：窦 慧

指导老师：申富饶 教授

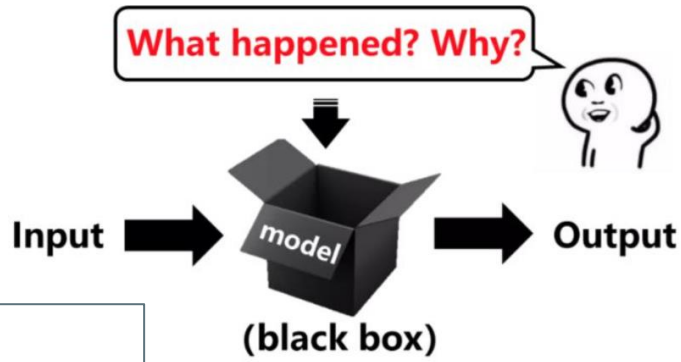
RINC
Robotic Intelligence & Neural Computing Group

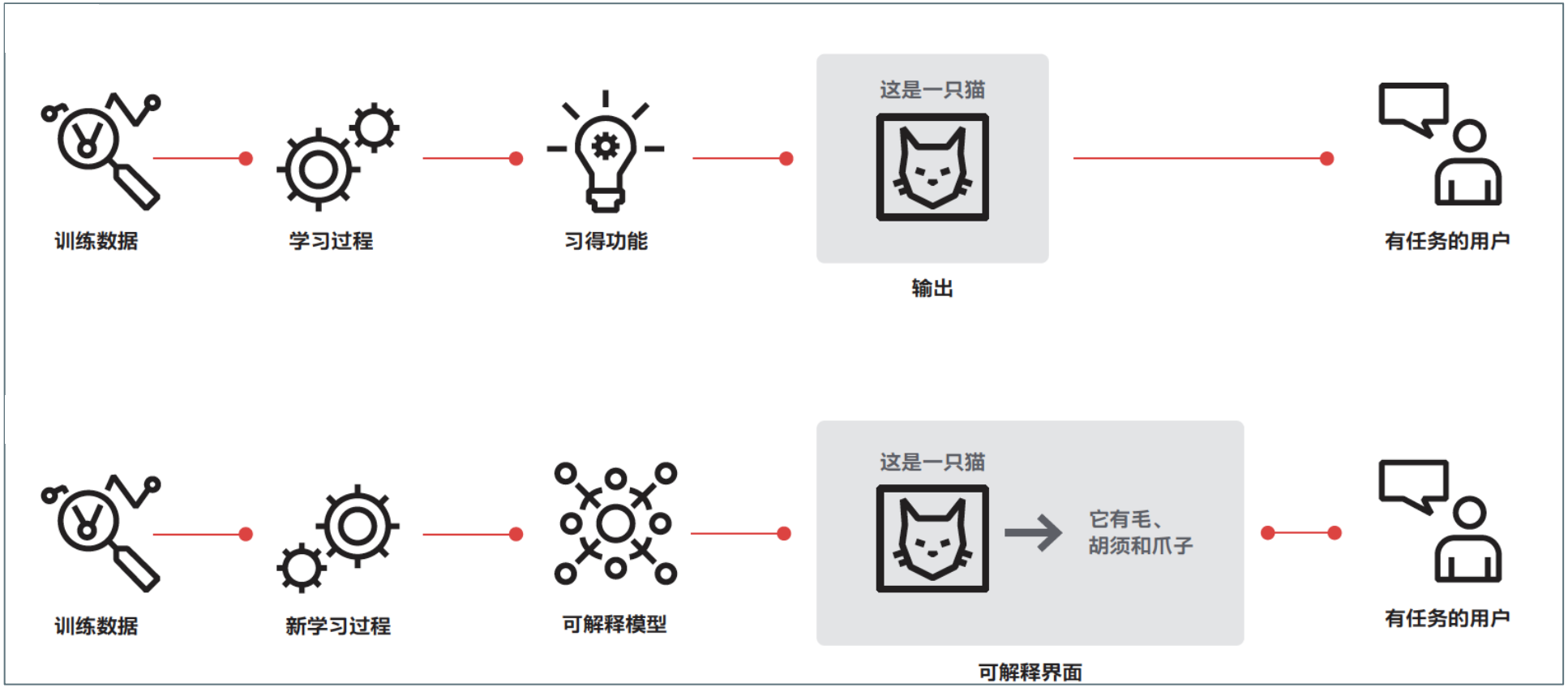
目录

誠樸雄偉 勵學敦行

壹 研究背景及现状

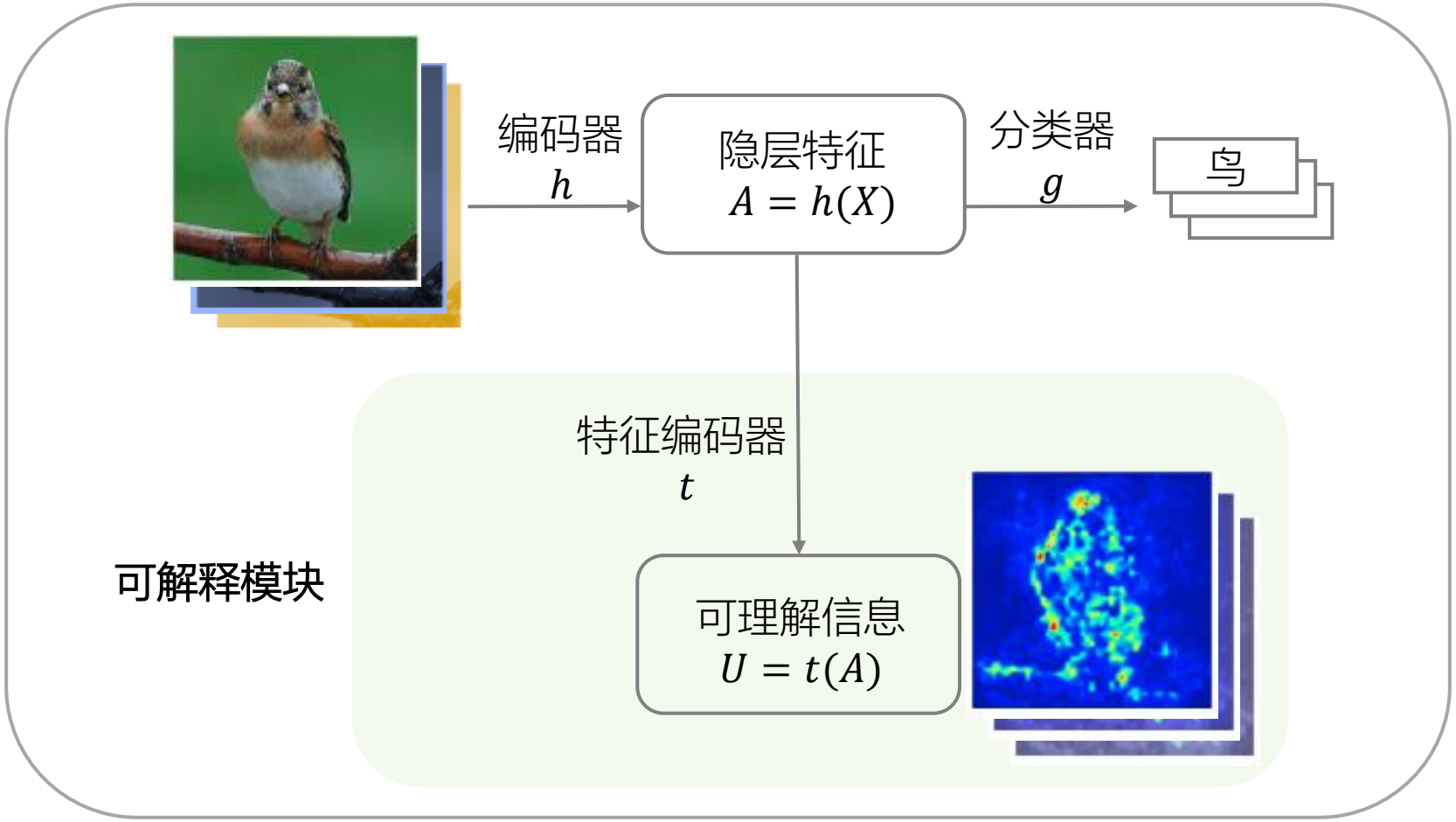
神经网络是**黑盒模型**，使用的过程中会产生一个疑问：为什么它是可行的？





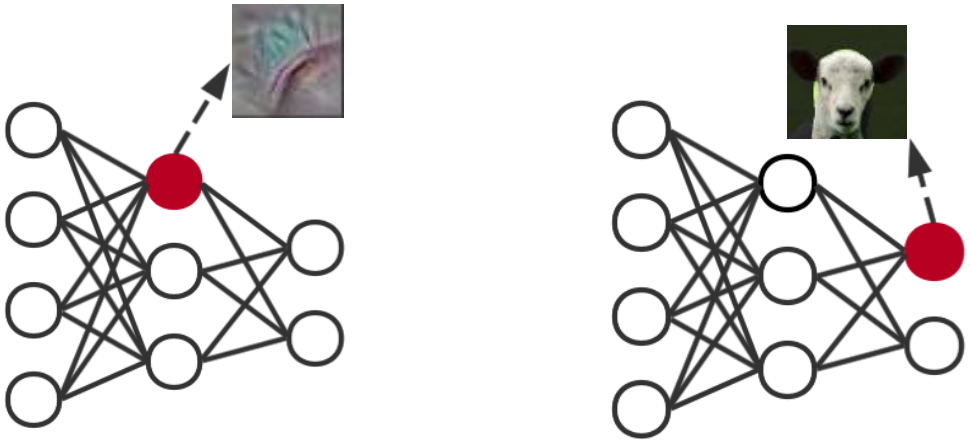
可解释性（Interpretability）：以人类可以理解的方式呈现神经网络模型的属性和结果。

- 解释模型输入：模型特征、模型的输入输出
- 解释模型输出：可理解信息（图片、文本）



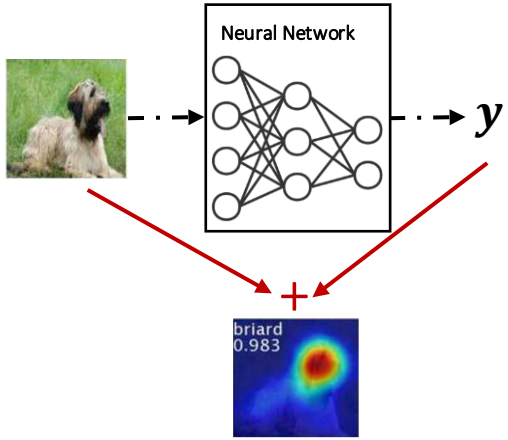
基于网络的解释方法

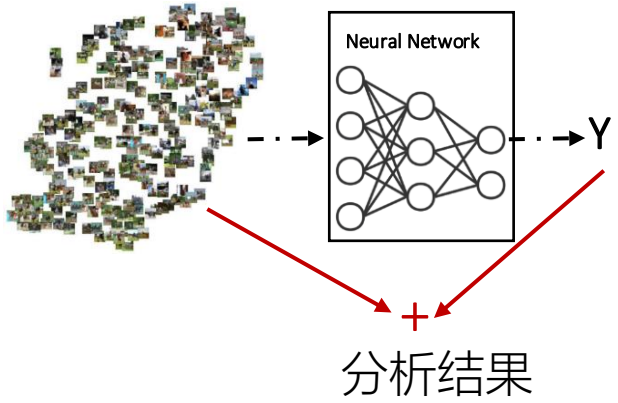
- 理想样本
- 真实样例



基于输入的解释方法

- 单一输入的解释
- 多个输入的解释





缺乏具有普适性的可解释性方法

- 现有方法与特定的神经网络结构紧密耦合，依赖于模型内部特定模块的行为特征进行解释。
- 限制了可解释性方法在不同模型间的可迁移性和可比较性。

解释结果的主观性与缺乏量化标准

- 现有主流方法依赖特征图、热力图等形式，高度依赖人工进一步解释。
- 这种主观性导致可解释性研究在实践中缺乏严谨性与一致性。

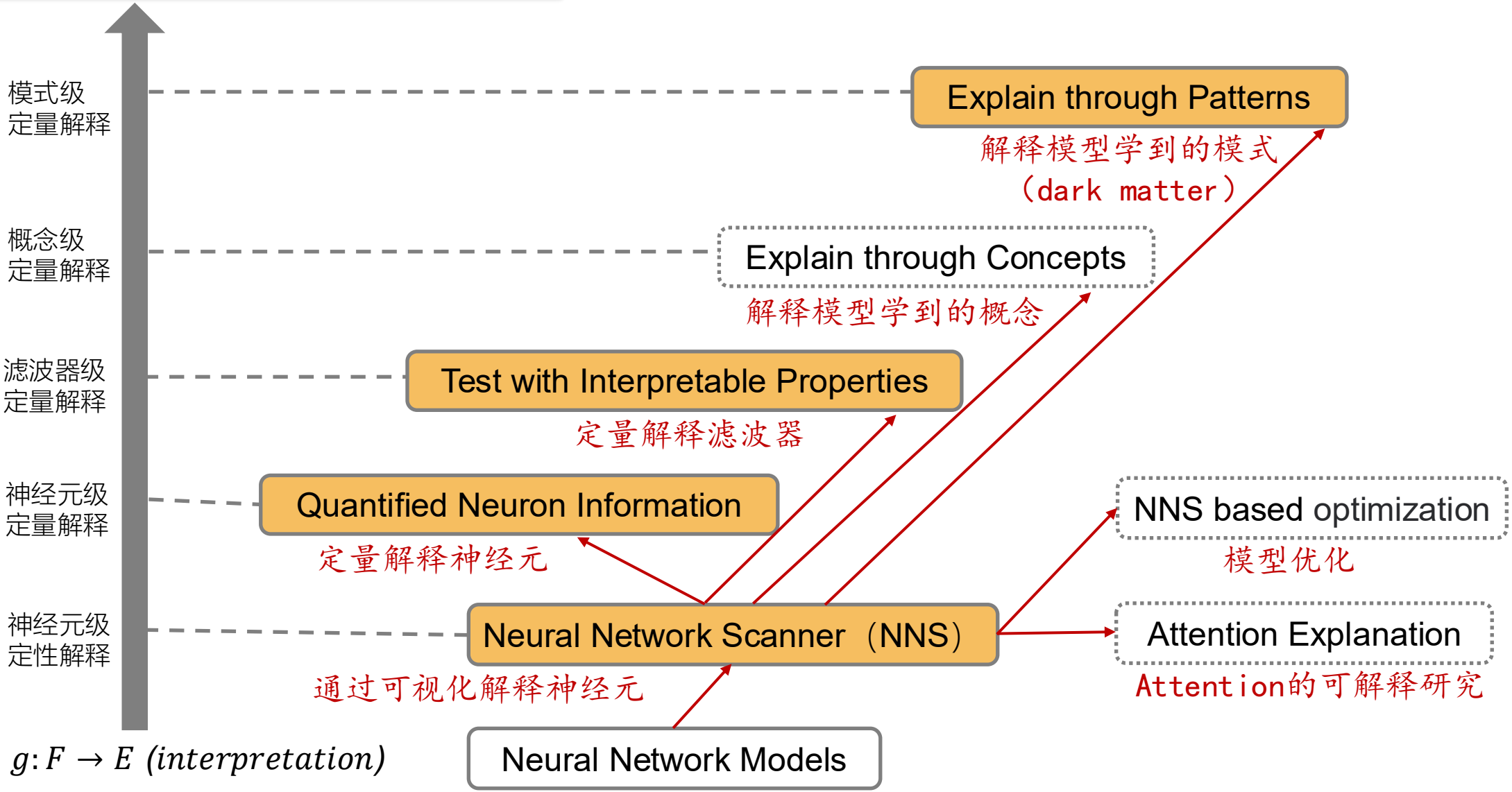
缺乏可扩展的可解释性评估框架

- 解释指标缺乏可扩展性已成为制约方法评估一致性与通用性的核心问题之一。
- 当前研究缺乏具备可扩展性的解释性评价指标体系，使得模型能够在统一框架下全面解释。

异常行为识别能力不足

- 神经网络在训练过程中可能引入大量冗余或功能异常的单元，导致模型在性能下降。
- 当前可解释性研究在识别和理解模型的异常行为方面仍显薄弱。

基于NNS的神经网络可解释性研究脉络



本论文的主要工作

科学问题：神经网络模型的解释性研究

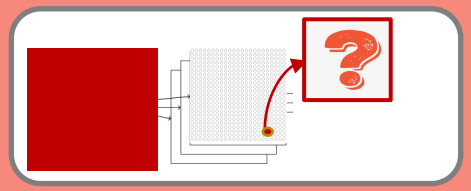
提出一种新型解释算法分类方法

（对应第二章内容）

通用神经网络解释算法

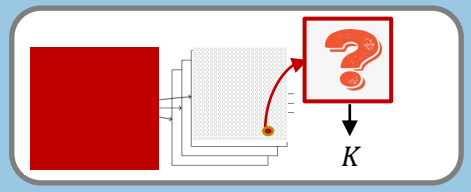
神经网络扫描仪（NNS）

基于神经元的神经网络定性解释算法



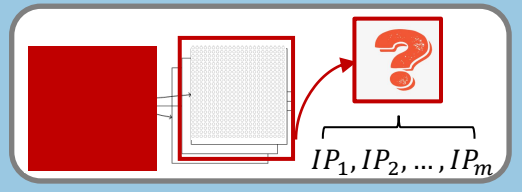
（对应第三章内容）

NNS+特征量



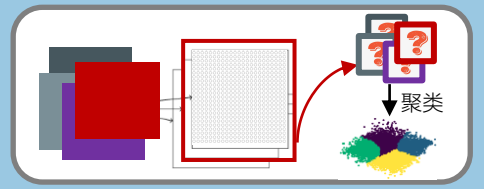
（对应第四章内容）

NNS+可解释属性



（对应第五章内容）

NNS+聚类



（对应第六章内容）

神经元级定量解释 滤波器级定量解释 模式级定量解释

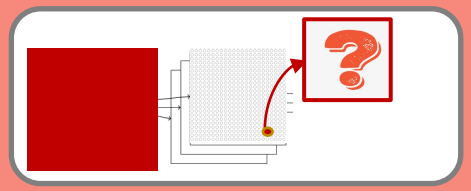
科学问题：神经网络模型的解释性研究

提出一种新型解释算法分类方法
（对应第二章内容）

通用神经网络解释算法

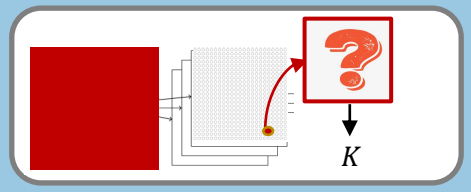
神经网络扫描仪（NNS）

基于神经元的神经网络定性解释算法



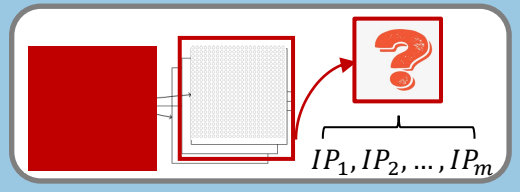
（对应第三章内容）

NNS+特征量



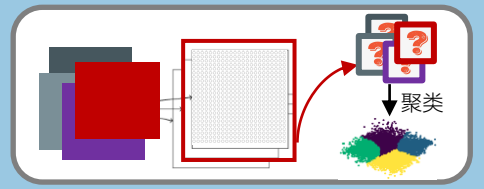
（对应第四章内容）

NNS+可解释属性



（对应第五章内容）

NNS+聚类



（对应第六章内容）

神经元级定量解释　　滤波器级定量解释　　模式级定量解释

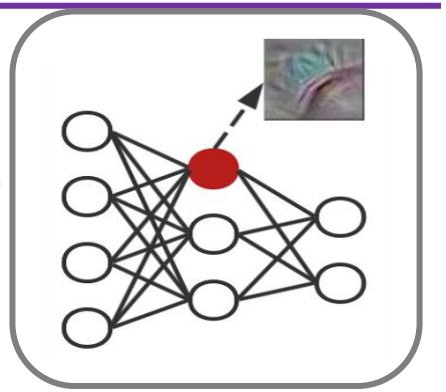
RINC
Robotic Intelligence & Neural Computing Group

贰 基于神经元的可解释性方法：神经网络扫描仪NNS

神经网络可视化

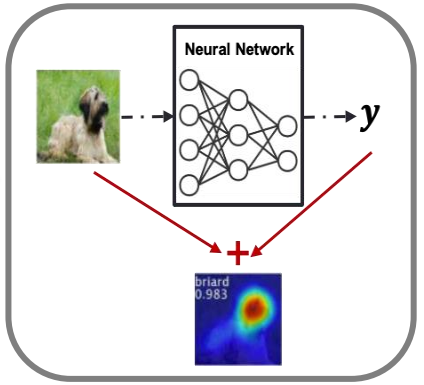
激活最大化

专注于**模型单元**关注的模式。
不适合解释单一输入样本的结果。

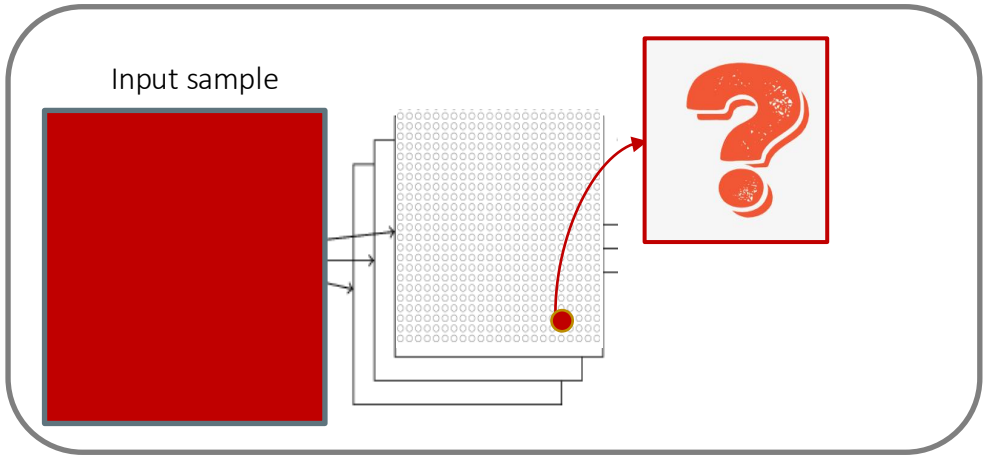


显著图

显示了输入样本的哪些区域对网络而言是重要的。
与**固定的**神经网络模型密切相关，很难比较不同神经网络模块的工作机制。



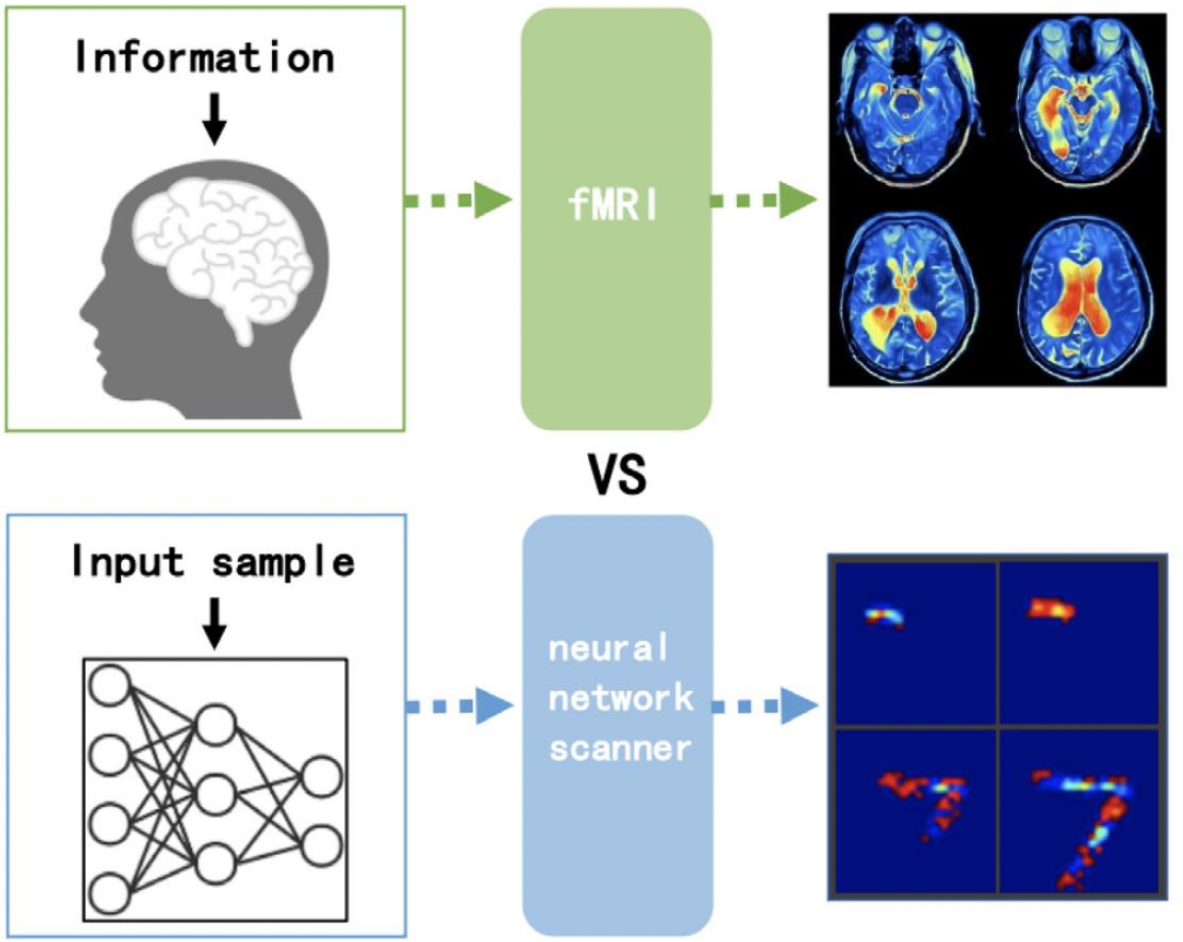
如何以统一的方式实现模型解释?



主要工作

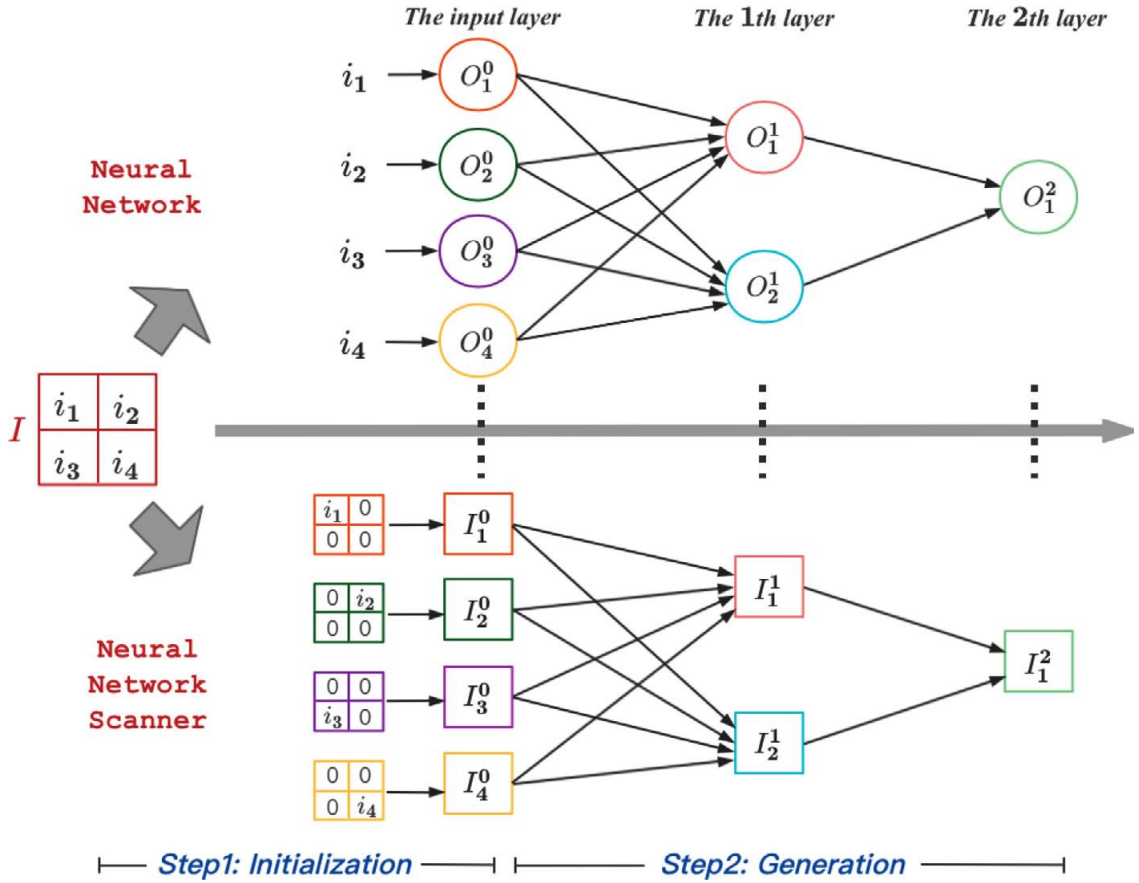
- 提出的NNS 可以可视化神经元的学习过程，展示每个神经元学习到的特征。
- 通过灵活地结合单个神经元学习到的特征，分析神经网络不同模块的工作机制。
- 多角度实验评估CNN 的可解释性。

Neural Network Scanner：
基于单个神经元的可视化解释



神经网络扫描仪NNS

对于一个输入样本，图像中物体的位置信息与数值信息同样重要。在全连接层中，仅使用数值信息，而位置信息被丢弃，这使得全连接层缺乏定位物体的能力。而在NNS 中，为神经元生成学习图像，位置信息得以保留。



学习图像初始化

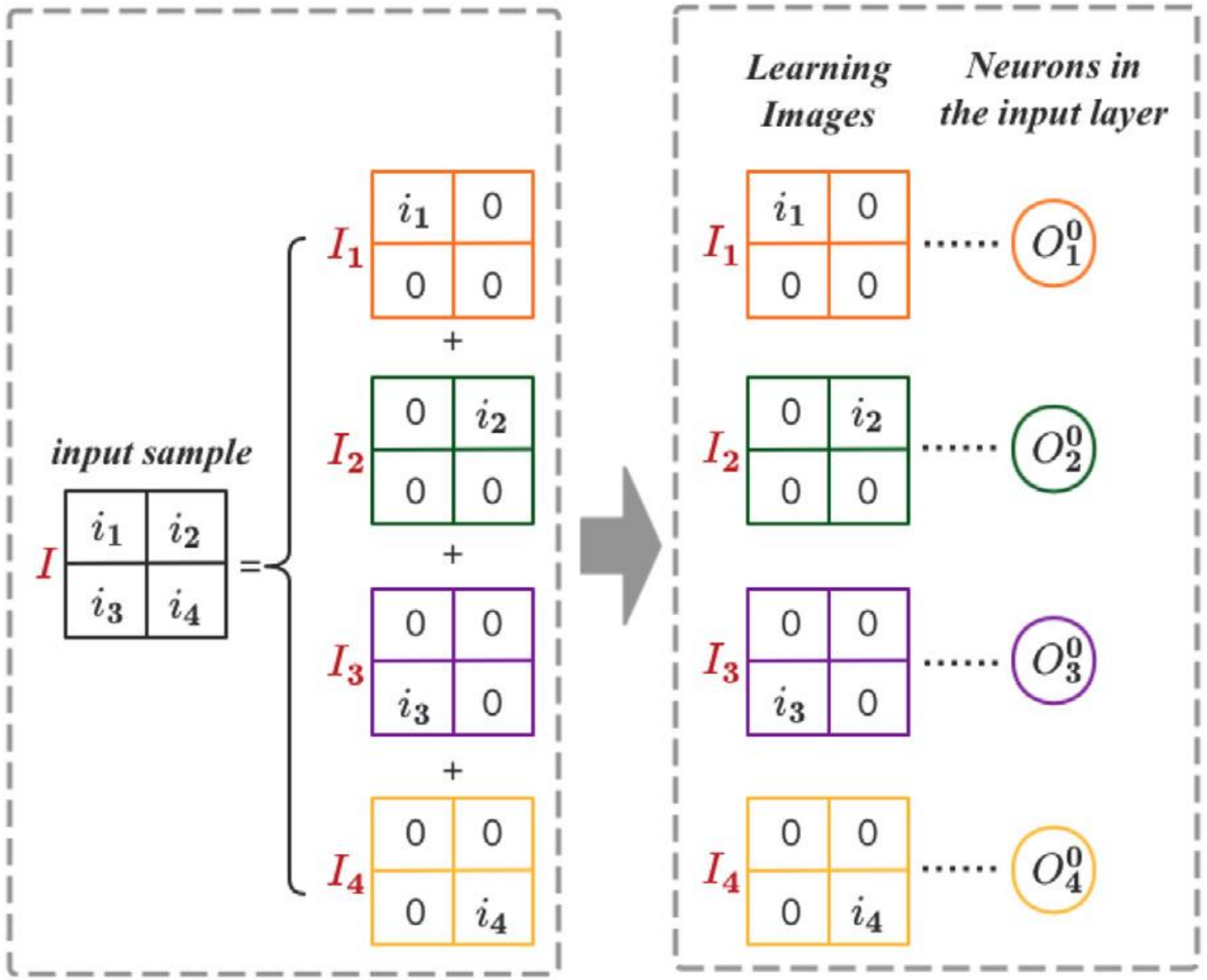
- 输入样本分割：

$$I = \sum_{p\in[1,N]} I_p$$

$$I_p(x) = \begin{cases} i_p, & \text{if } x == p \\ 0, & \text{otherwise} \end{cases}$$

- 学习图像分配：

$$I_p^0 = I_p, \quad p \in [1, N]$$



全连接层学习图像生成

通过获取输入神经元对指定神经元的贡献，基于输入神经元的学习图像生成该神经元的学习图像。

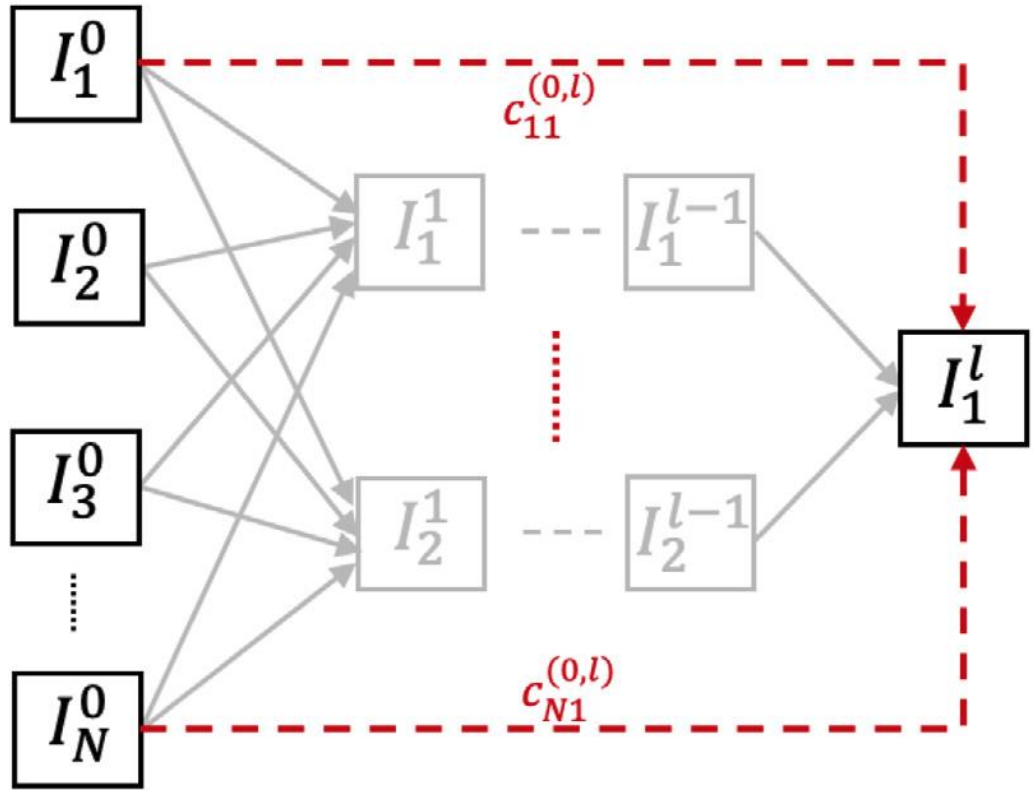
- 指定神经元的学习图像生成：

$$I' = \sum_{p\in[1,N]} c_{pq}^{(0,1)} \times I_p^0 \qquad I_q^1 = \begin{cases} I', & \text{if } x' > 0 \\ \mathbf{0}, & \text{otherwise} \end{cases}$$

- 一般情况下的学习图像生成：

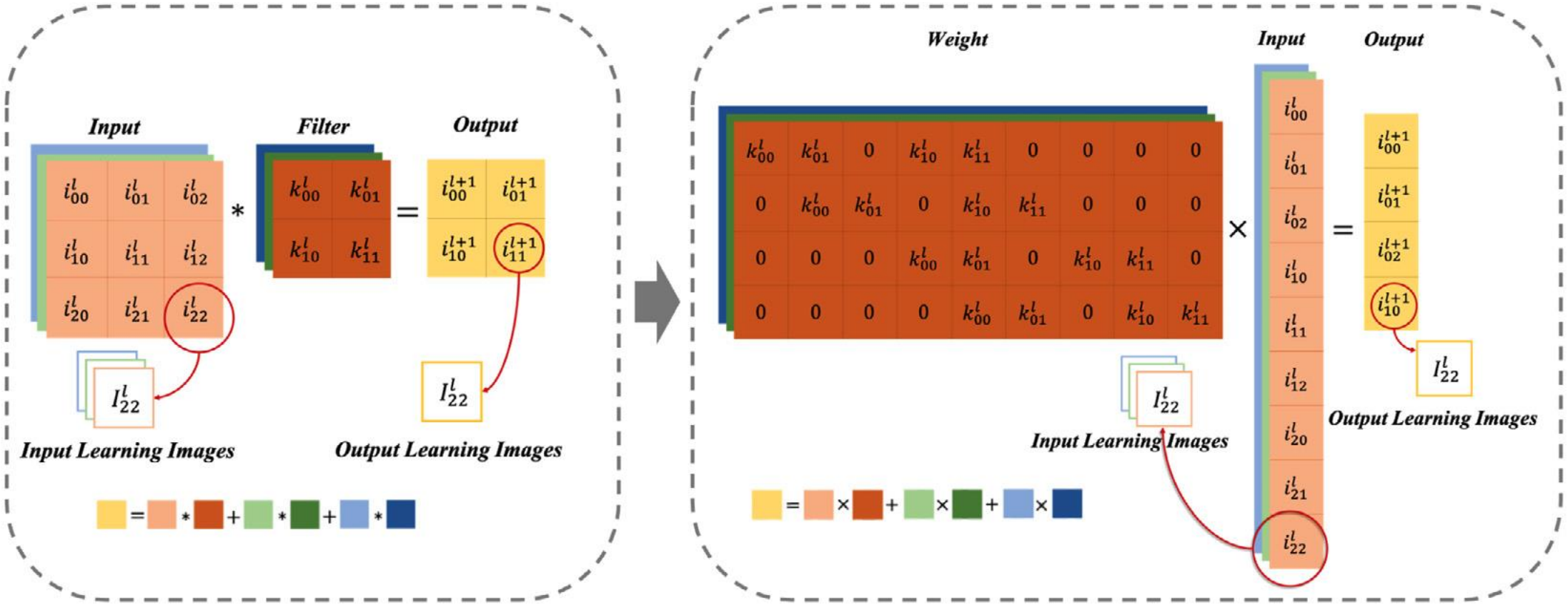
$$C_q^{(0,l)} = W_q^l \cdot \prod_{i=1}^{l-1} C^{(i-1,i)}$$

$$I_q^l = I_{sum}^0 \cdot C_q^{(0,l)}$$



卷积层学习图像生成

将卷积和池化操作转化为线性操作，然后通过与全连接层计算学习图像相同的方式，获得卷积层和池化层中的学习图像。



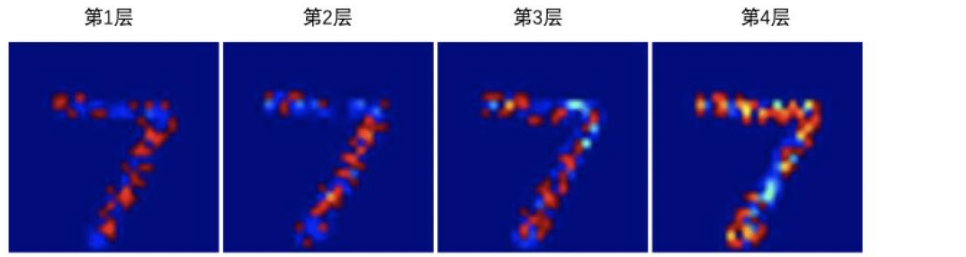
不同模块运行机制分析

- 在FCN 中，学习特征通过**激活强度**变化实现。

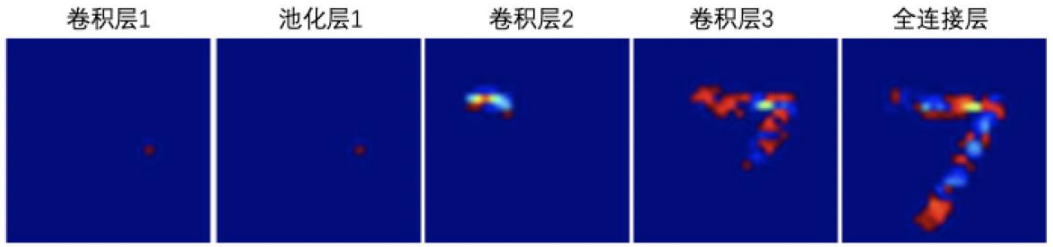
FCN 在图像相同位置上学习到的特征复杂性随着层数的增加而增加。

- 在CNN 中，学习特征通过**激活数量**的变化实现。

低层神经元学习简单的局部特征，而高层神经元学习复杂的全局特征。

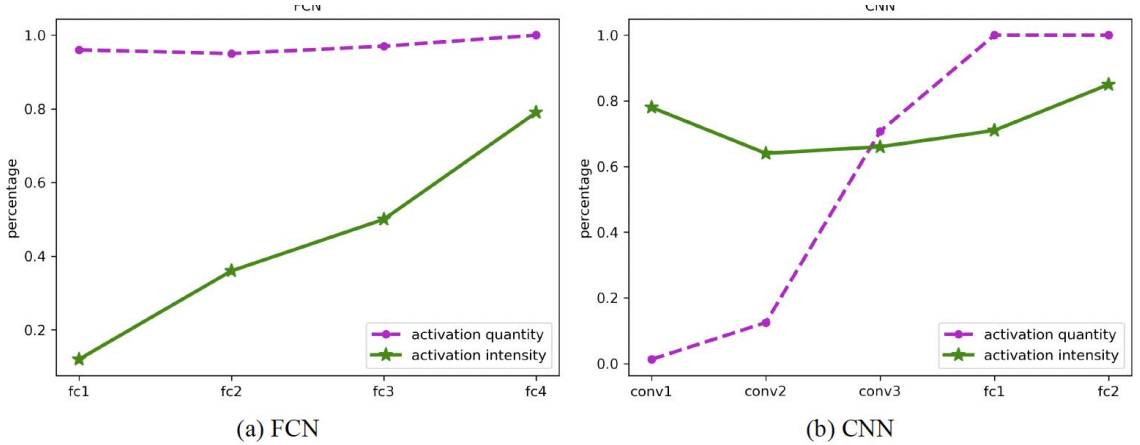


(a) FCN



(b) CNN

图 3-7 神经元学习过程可视化



(a) FCN (b) CNN

图 3-8 FCN 和 CNN 每层知识量的比较

不同模块神经元学习规则

$$dist_o^I = \|I_o - I\|$$

- 全连接层中的神经元没有固定的学习特征。神经元学习的特征与输入样本始终有很大的关系。
- 卷积层中的神经元倾向于学习固定特征，但这种特性随着层数的上升而减弱。

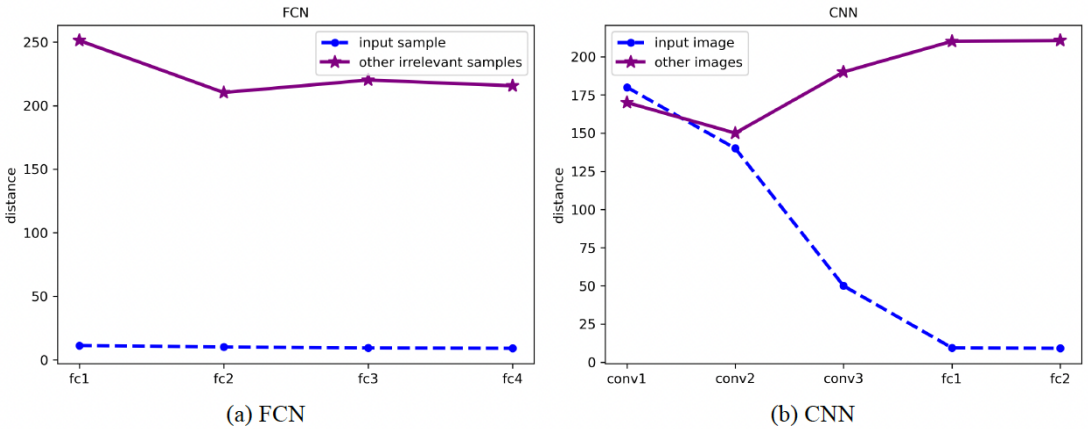


图 3-9 学习图像与样本之间的相似性

获胜神经元分析

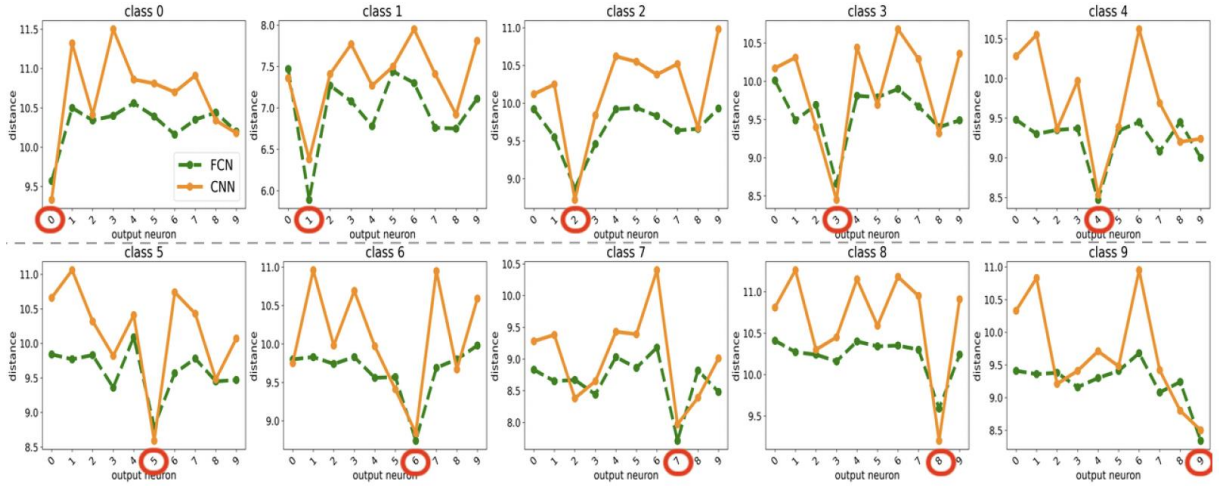


图 3-10 输入样本与输出神经元学习图像之间的距离

和输出层中的其他神经元相比，获胜神经元能够更好地重建输入样本。

CNN层级解释评估

学习图像与特征图之间的相关性。 $C^{k,l} = \frac{S^{k,l} \cap I^{k,l}}{I^{k,l}}$

表 3-3 特征图与学习图像的相关性

	第 1 层	第 2 层	第 3 层	第 4 层	第 5 层
AlexNet	0.998	0.872	0.641	0.634	0.623
VGG-16	0.952	0.923	0.688	0.620	0.538

- 学习图像与其对应的特征图之间高度相关。
- 学习图像中的信息比对应特征图中的信息更细粒度。

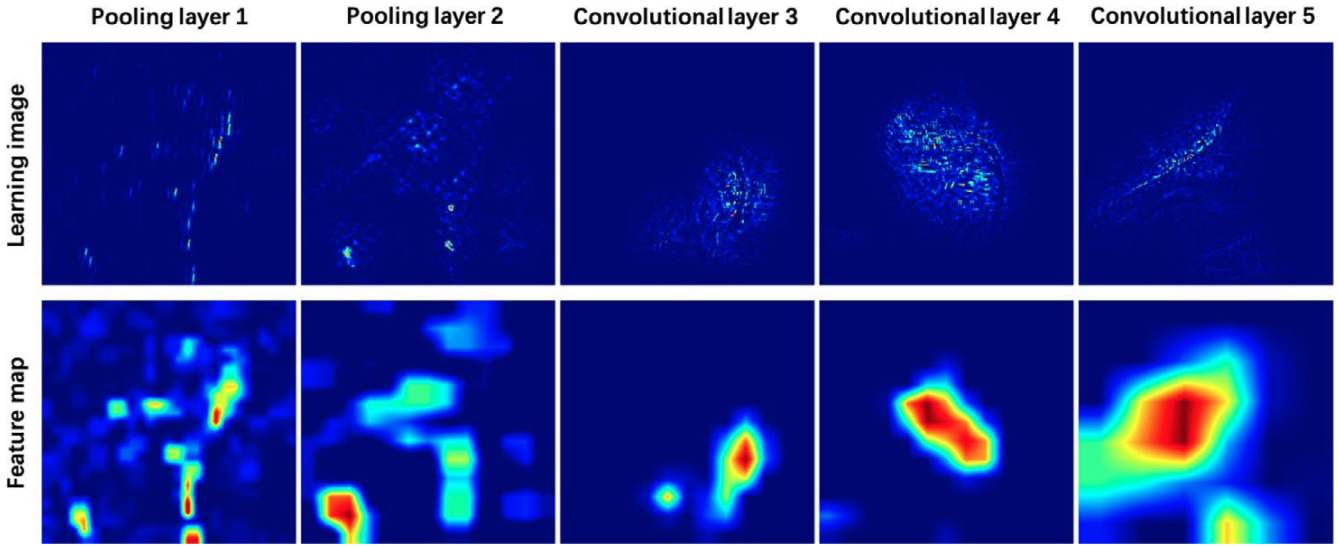


图 3-13 AlexNet 中学习图像及其对应的特征图

CNN神经元级解释评估

$$IoU^{k,l} = \frac{I_1^{k,l} \cap I_2^{k,l}}{I_1^{k,l} \cup I_2^{k,l}}$$

表 3-4 高度激活神经元间的学习图像相关性

	第 1 层	第 2 层	第 3 层	第 4 层	第 5 层
AlexNet	0.999	0.953	0.761	0.676	0.615
VGG-16	0.997	0.990	0.873	0.835	0.804

- 高度激活神经元趋向于学习相似的特征。

与其他可解释方法比较分析

相较于其他方法，NNS更侧重于描述了输入样本的轮廓和重要细节。

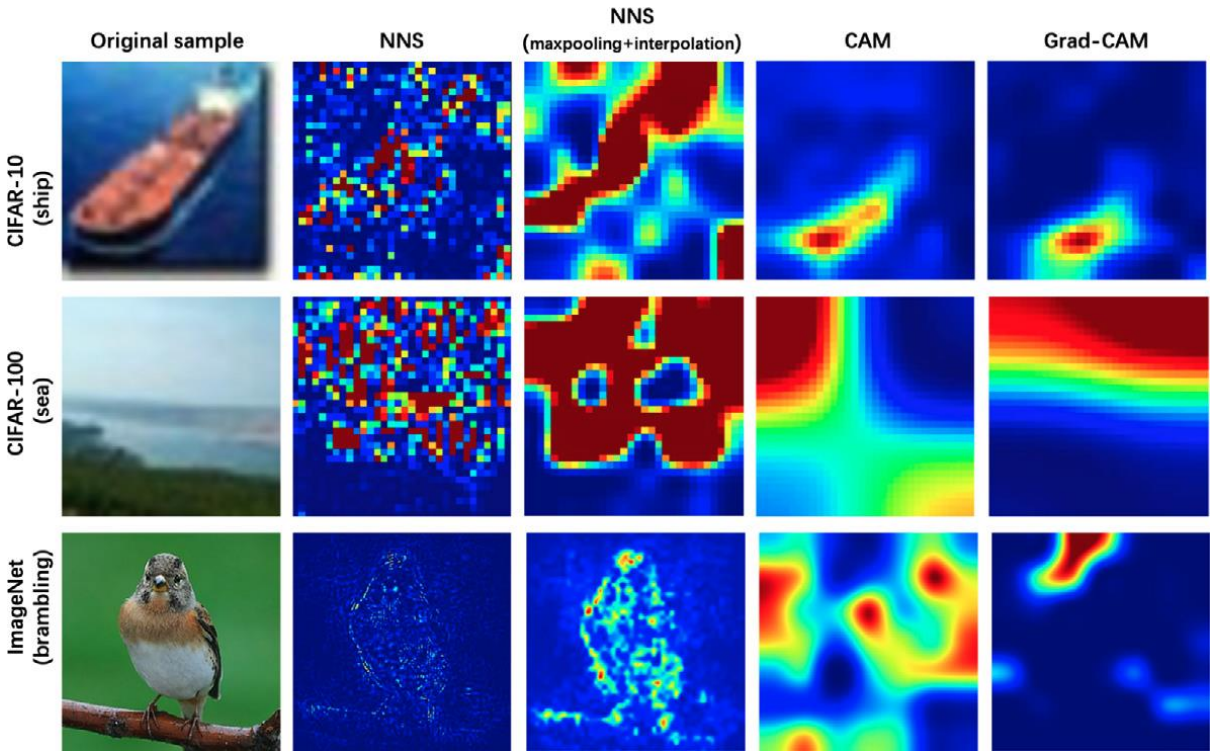


图 3-14 不同解释方法的可视化结果比较

神经网络扫描仪（NNS）

一 可视化神经元的学习过程，展示每个神经元学习到的特征，保持定位能力。

二 通过灵活地结合单个神经元学习到的特征，统一分析神经网络不同模块的工作机制。

三 多角度实验解释CNN模型。

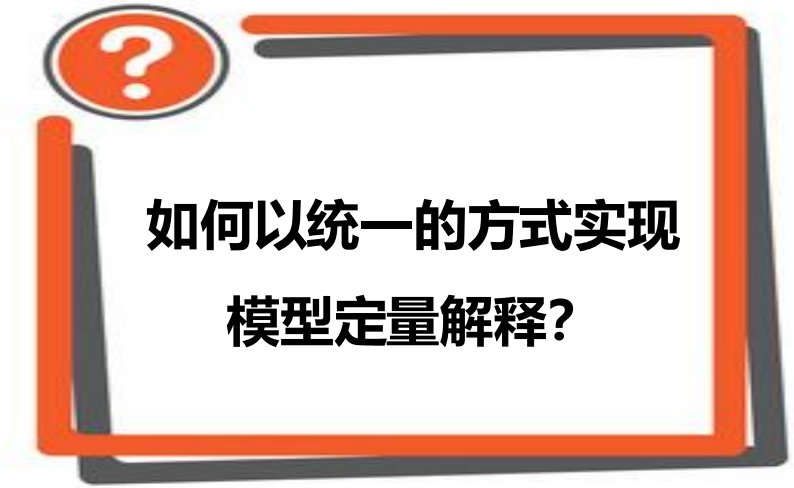
叁 基于NNS的神经元层面解释

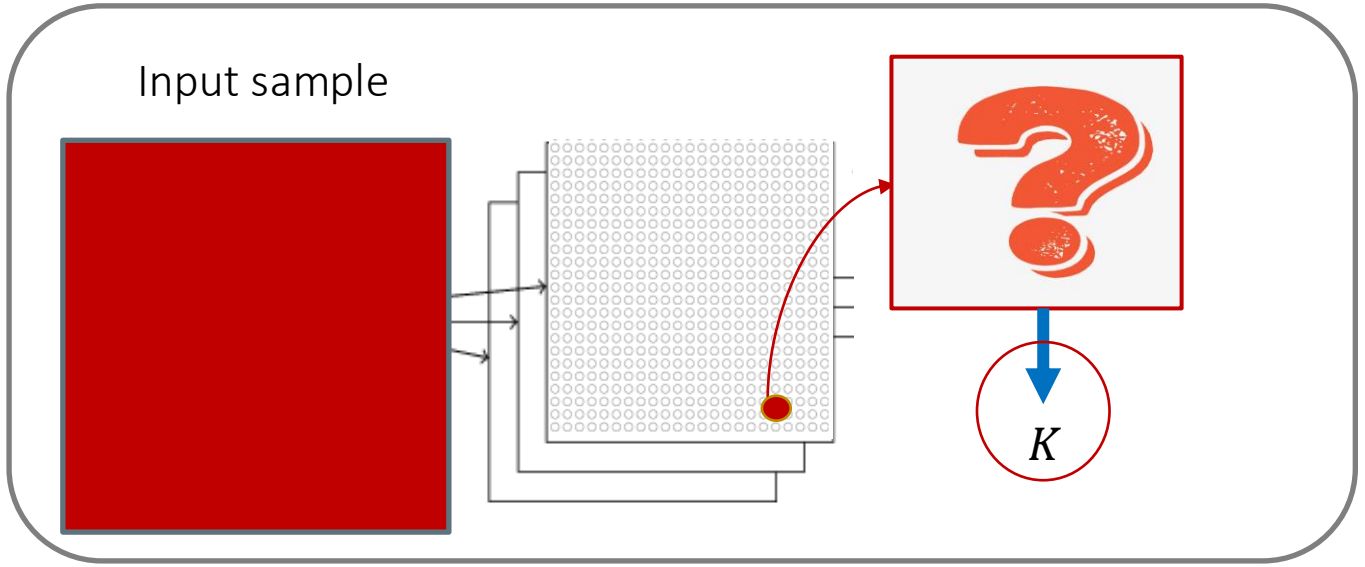
现有解释方法：

- 大多研究集中在特征图或整个模型，通常与**特定的架构**联系紧密。

NNS在神经元定量解释中的瓶颈：

- NNS能够有效地可视化神经元学习到的特征，但在**定量衡量**这些特征方面存在不足。由于缺乏标准评估准则，NNS的结果往往具有主观性，通常需要额外的人工解释。

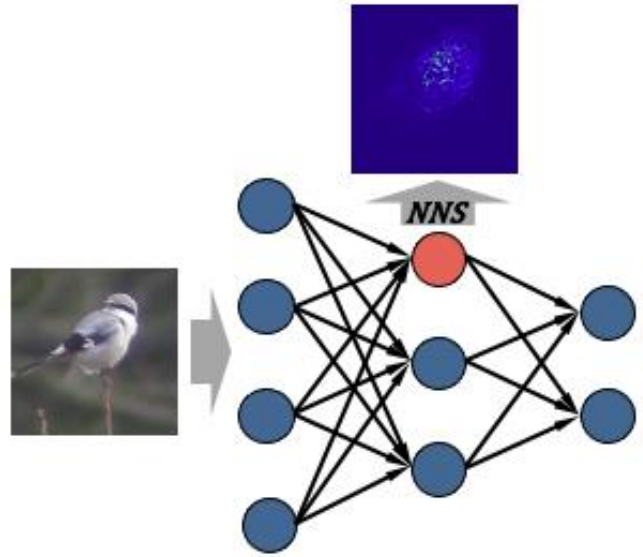


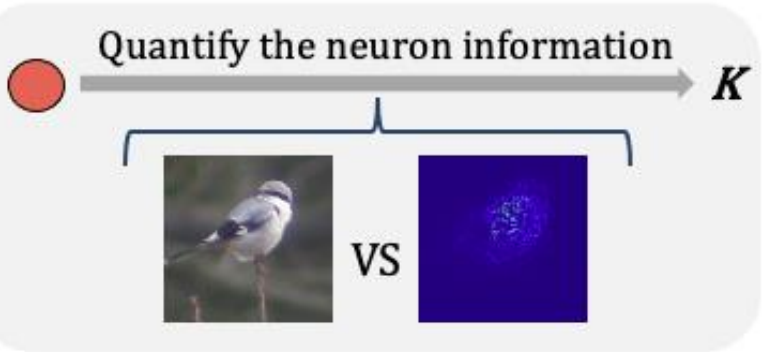


主要工作

- 通过**NNS**的学习图像量化每个神经元的编码信息。
- 定量获取神经元学习结果，不需要人对解释结果进行进一步的加工。
- 基于特征量从三个方面分析模型。

Quantified Neuron Information：
度量神经元的可解释特征量





神经元特征量的计算

引入了“特征量”这一概念来量化了神经元学习的特征。

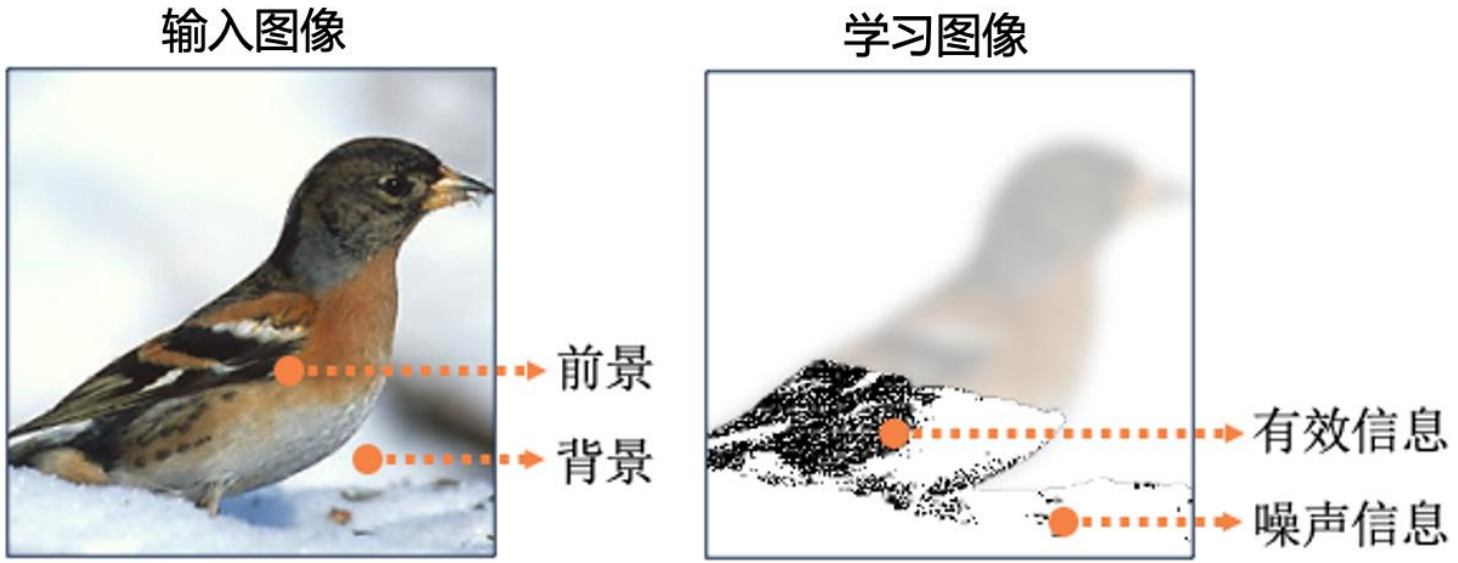
通过量化和比较前景和背景信息中的编码特征，探讨这些神经元学习到的有效信息和噪声信息。

$$I_x^{fe} = I^f \cap M_x$$

前景信息

$$I_x^e = I_x \cap M^f, I_x^n = I_x \cap M^b$$

有效信息　　噪声信息



神经元特征量的计算

亮度对比为：$$l_x = \frac{2\mu^{fe}\mu^{e}}{\mu^{fe2} + \mu^{e2} + C1}$$

对比度对比为：$$c_x = \frac{2\sigma^{fe}\sigma^{e}}{\sigma^{fe2} + \sigma^{e2} + C2}$$

结构对比为：$$s_x = \frac{\sigma^{fee}}{\sigma^{fe}\sigma^{e} + C3}$$

神经元 x 对指定输入的特征量通过比较 I 和 I_x 之间的差异来量化：

$$F(I,x) = K\left(\frac{N_x^n}{N}\right)\left(\frac{N_x^e}{N^f}\right)(l_x \cdot c_x \cdot s_x)$$

从数量角度量化特征　　从质量角度量化特征

神经元x的特征量：

$$FQ(x) = F(I^*, x), s.t. I^* := \arg\max_{i\in\{1,\cdots,M\}} F(I^i, x)$$

特殊状态下的神经元特征量分析

神经元完全学习到输入样本的结构特征：可视作学习图像为输入图像的线性缩放版本。

$$I_x = pI$$

$$F(I,x) = \frac{2p\mu^{fe^2}}{(1+p^2)\mu^{fe^2}+C1} \times \frac{2p\sigma^{fe^2}}{(1+p^2)\sigma^{fe^2}+C2} \times \frac{p^2\sigma^{fe^2}}{\sigma^{fe^2}+C3}$$

当 $p=1$ 且 $C_1, C_2 \to 0$ 时，，有：$F(I,x) \to 1$ 。

若学习图像与输入图像在结构上完全一致，则神经元特征量可达最大值1，反映神经元完全学习到了输入特征。

特殊状态下的神经元特征量分析

神经元完全未学习任何结构特征：若神经元未被激活，即对图像无响应，其学习图像为全零。

$$I_x = 0$$

$$F(I, x) = 0$$

当学习图像为空白时，与原图结构相似性为0，意味着神经元未能学习到任何输入特征。

神经元学习状态与特征量间的关系

学习图像情况	学习图像形式	特征量	学习图像与原图相似性
完全学习特征	$I_x \propto I$	接近 1	高
完全未学习特征	$I_x = 0$	接近 0	低

不同度量标准的比较

- 低层的神经元学习的特征很少，学习图像接近空白；随着层数的增加，神经元学习的特征变得更加复杂。
- 特征量有效地考虑了学习图像中空白区域的影响。

表 4-2 不同度量标准下的学习图像的特征

		conv1	conv2	conv3	conv4	conv5	conv6	conv7	conv8
AlexNet	ILSVRC-2013DET dataset								
	SSIM	0.95	0.52	0.24	0.76	0.42	-	-	-
	FQ	**0.0012**	**0.02**	**0.18**	**0.61**	**0.34**	-	-	-
	CUB200-2011 dataset								
	SSIM	0.92	0.48	0.35	0.74	0.41	-	-	-
	FQ	**0.0003**	**0.015**	**0.13**	**0.52**	**0.29**	-	-	-
VGG-11	ILSVRC-2013DET dataset								
	SSIM	0.96	0.92	0.84	0.67	0.71	0.74	0.68	0.53
	FQ	**0.0013**	**0.026**	**0.34**	**0.47**	**0.54**	**0.69**	**0.63**	**0.45**
	CUB200-2011 dataset								
	SSIM	0.95	0.94	0.81	0.51	0.62	0.67	0.87	0.79
	FQ	**0.0003**	**0.0013**	**0.23**	**0.37**	**0.59**	**0.64**	**0.78**	**0.54**
VGG-16	ILSVRC-2013DET dataset								
	SSIM	0.94	0.94	0.92	0.87	0.84	0.88	0.79	0.74
	FQ	**0.0011**	**0.018**	**0.23**	**0.36**	**0.48**	**0.51**	**0.57**	**0.60**
	CUB200-2011 dataset								
	SSIM	0.95	0.92	0.90	0.87	0.86	0.89	0.88	0.90
	FQ	**0.0010**	**0.013**	**0.18**	**0.27**	**0.33**	**0.39**	**0.48**	**0.51**
ResNet-18	ILSVRC-2013DET dataset								
	SSIM	0.97	0.95	0.90	0.88	0.79	0.75	0.70	0.66
	FQ	**0.0022**	**0.031**	**0.14**	**0.22**	**0.48**	**0.51**	**0.55**	**0.61**
	CUB200-2011 dataset								
	SSIM	0.96	0.91	0.89	0.83	0.77	0.74	0.71	0.59
	FQ	**0.0015**	**0.0027**	**0.18**	**0.27**	**0.32**	**0.44**	**0.47**	**0.51**

基于特征量的模型分析

（1）神经元的特征量与激活值的关系

度量：

通过测量皮尔逊相关系数来分析激活值集 A 与特征量集 K 之间的关系：

$$r(A,K)=\frac{\sigma_{AK}}{\sigma_A\sigma_K}$$

各模型中绝大多数卷积层的神经元特征量与激活值之间表现出显著的正相关性。

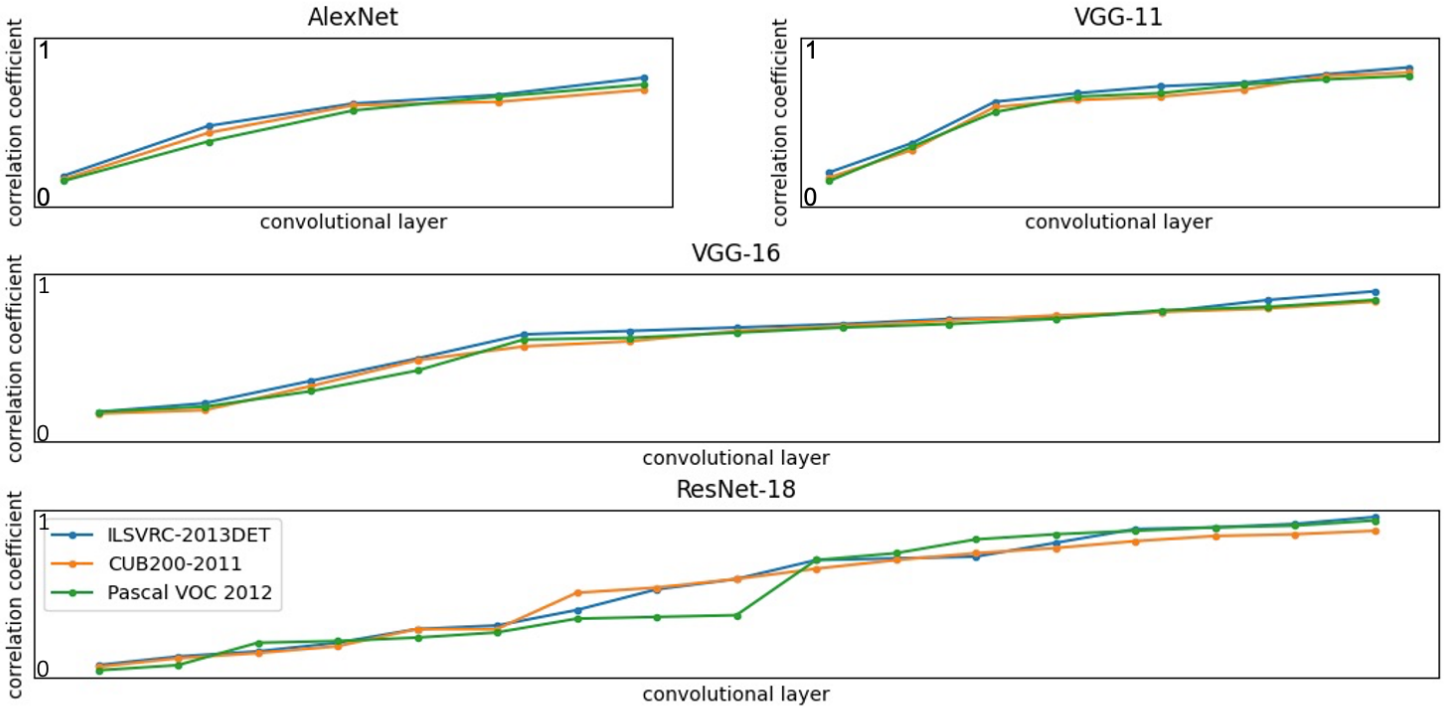


图 4-3 特征量与激活值的 Pearson 相关系数

基于特征量的模型分析

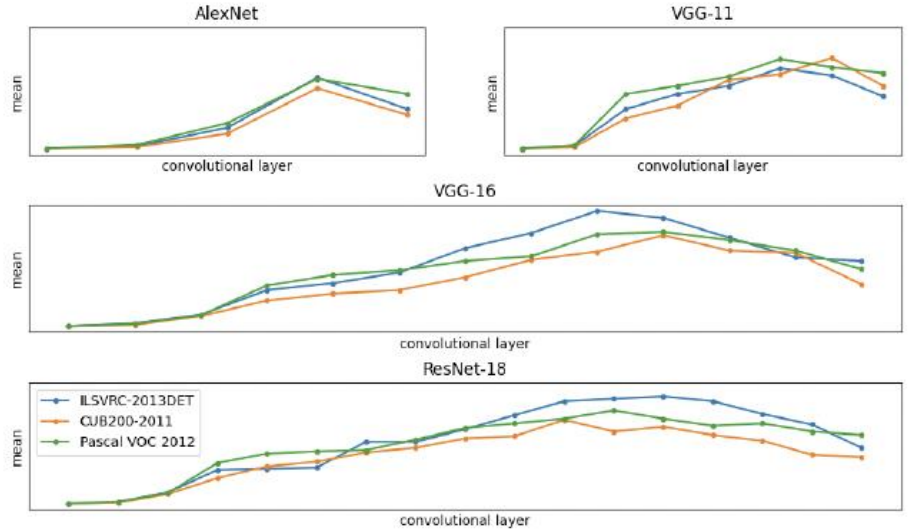
（2）不同层级的特征量的变化情况

度量：

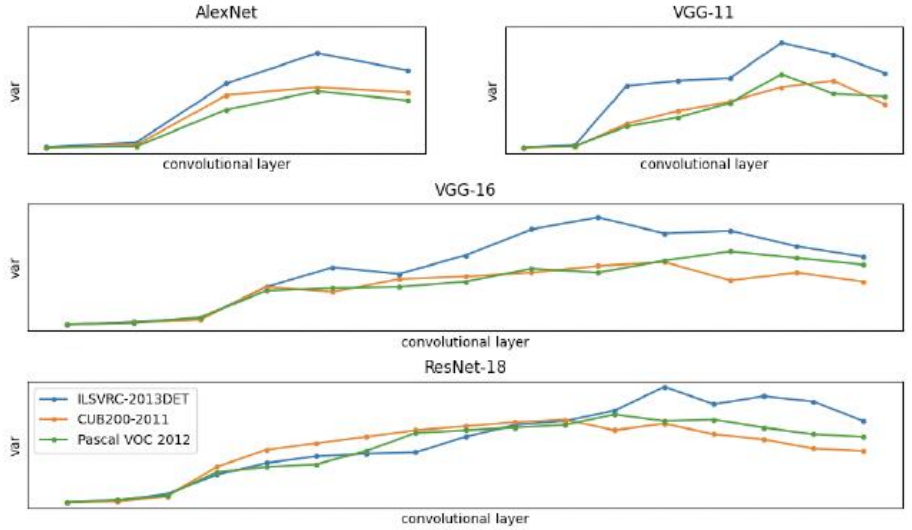
分别从“响应强度的集中趋势”与“表示的多样性”两个维度分析神经元特征量的层级特征。

卷积神经网络在特征学习过程中表现出由“多样性”向“一致性”的特征表示过渡。

- 低层：通用的、低级的
- 中间层：学习到更为抽象且多样的表示
- 高层： 统一的更具判别性的模式



(a) Mean



(b) Variance

图 4-4 不同层神经元特征量的均值和方差

基于特征量的模型分析

（3）不同滤波器的特征学习能力

度量：

采用DBSCAN聚类，分别对聚类数排名前20（高多样性）和后20（低多样性）的滤波器进行掩码实验，观察性能变化。

表 4-3 滤波器掩码后模型损失的百分比变化

模型	AlexNet		VGG-11		VGG-16		ResNet-18	
掩码滤波器	前 20	后 20	前 20	后 20	前 20	后 20	前 20	后 20
数据集 1	+108.3%	-12.8%	+93.4%	-1.9%	+95.2%	-0.01%	+78.2%	-2.6%
数据集 2	+84.9%	-8.7%	+90.5%	-10.3%	+89.6%	+0.01%	+88.3%	-0.4%
数据集 3	+112.5%	+0.03%	+92.8%	-2.7%	+88.7%	-3.1%	+82.2%	+1.3%

卷积层内部滤波器间存在显著的学习能力差异：

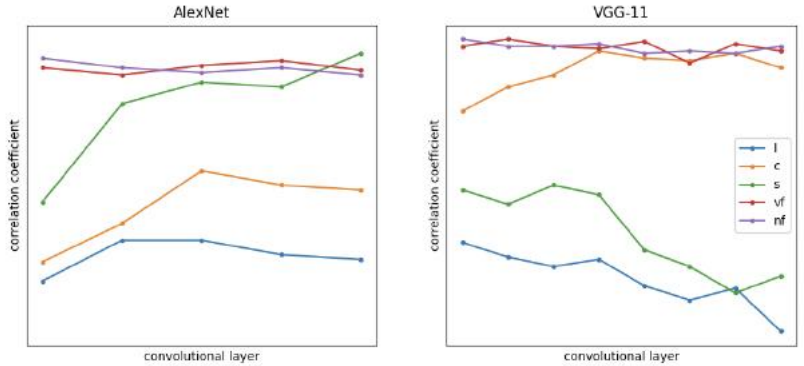
- 高多样性滤波器专注于多样化的特征，对模型性能影响大
- 低多样性滤波器集中在单一或冗余的特征表示，对模型性能影响小

消融实验

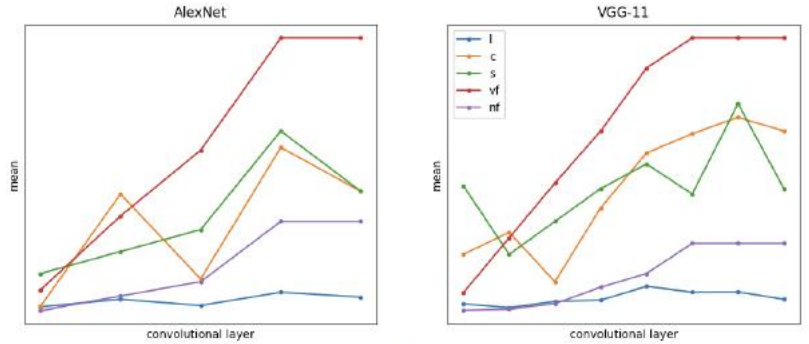
表 4-4 掩码特征量中不同元素后模型损失的百分比变化

模型	AlexNet		VGG-11	
掩码滤波器	前 20	后 20	前 20	后 20
l	-32.7%	+23.4%	-0.18%	+67.3%
c	+1.4%	+3.5%	-19.6%	-8.7%
s	+34.1%	-10.4%	-4.2%	-0.12%
vf	+56.8%	+58.4%	+70.6%	+65.8%
nf	+56.8%	+58.4%	+70.6%	+65.8%

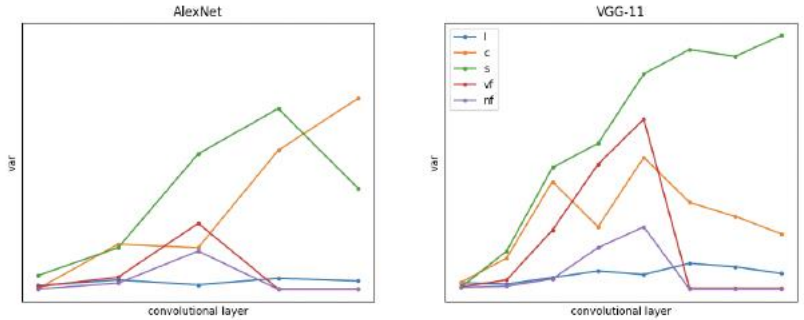
特征量各组成部分在整体特征表达中均产生一定影响，但其单独作用呈现出不一致性。



(a) Pearson correlation coefficient



(b) Mean



(c) Variance

图 4-5 特征量中不同元素对方向 1 和 2 结果的影响

与其他可解释方法比较分析

度量:

采用IOU作为衡量显著图与原图中目标区域一致性的度量指标。

$$\mathrm{IOU} = \frac{|M_{\mathrm{sal}} \cap M_{\mathrm{gt}}|}{|M_{\mathrm{sal}} \cup M_{\mathrm{gt}}|}$$

表 4-5 不同解释方法的 IOU

	数据集 1			数据集 2			数据集 3		
	CAM	Grad-CAM	**FQ**	CAM	Grad-CAM	**FQ**	CAM	Grad-CAM	**FQ**
AlexNet	0.08	0.17	**0.32**	0.09	0.23	**0.46**	0.08	0.22	**0.34**
VGG-11	0.11	0.24	**0.52**	0.14	0.19	**0.45**	0.09	0.21	**0.35**
VGG-16	0.17	0.31	**0.49**	0.16	0.36	**0.53**	0.07	0.15	**0.28**
ResNet-18	0.19	0.29	**0.58**	0.21	0.28	**0.34**	0.12	0.32	**0.44**

提出的方法目标定位能力优于其他显著图方法，验证了神经元特征量是作为度量模型学习能力的有效指标。

度量神经元的可解释特征量

一 引入了神经元特征量的概念，用于量化CNN 中神经元学习的特征。

二 基于这一度量分析了神经元的功能，并探索了卷积层的工作机制。

三 从三个方面分析模型，并讨论了不同模型之间的异同。

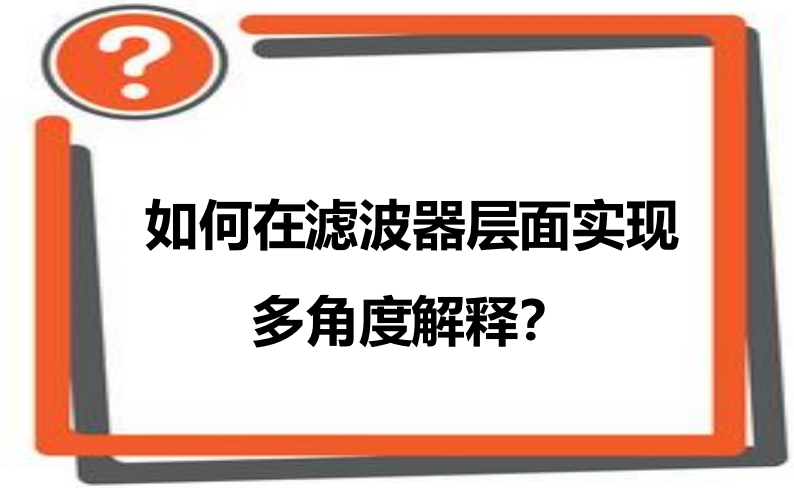
肆 基于NNS的滤波器层面解释

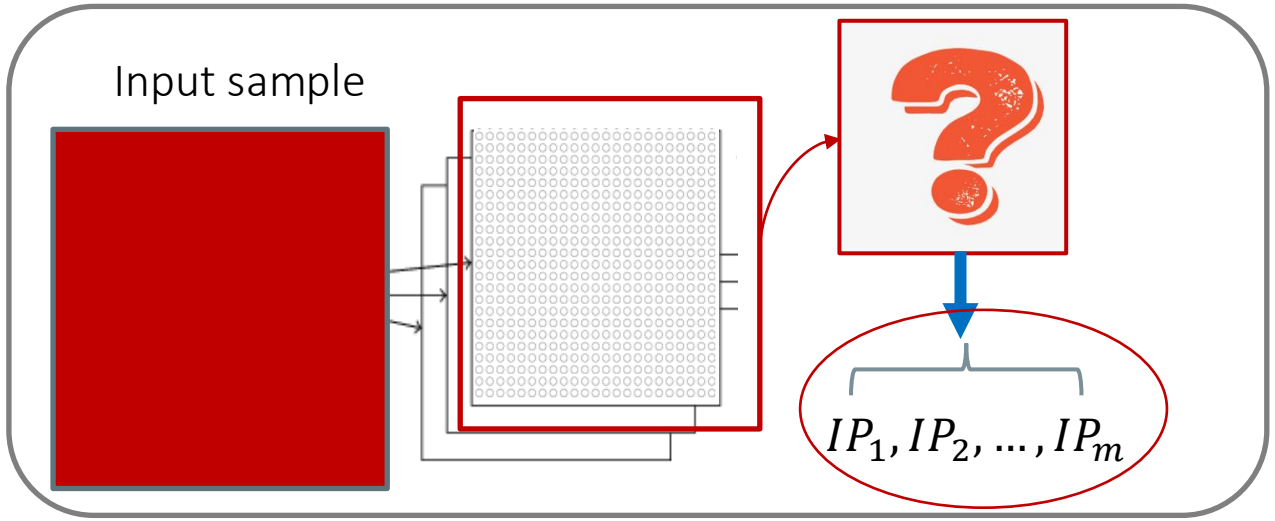
现有解释方法：

- 通常基于**单一标准**进行解释，过于简化模型内部复杂的关系，进而导致**部分解释**。

NNS在滤波器定量解释中的瓶颈：

- 缺乏针对神经网络模型更高阶单元—**滤波器**的解释，导致解释结果不够全面。

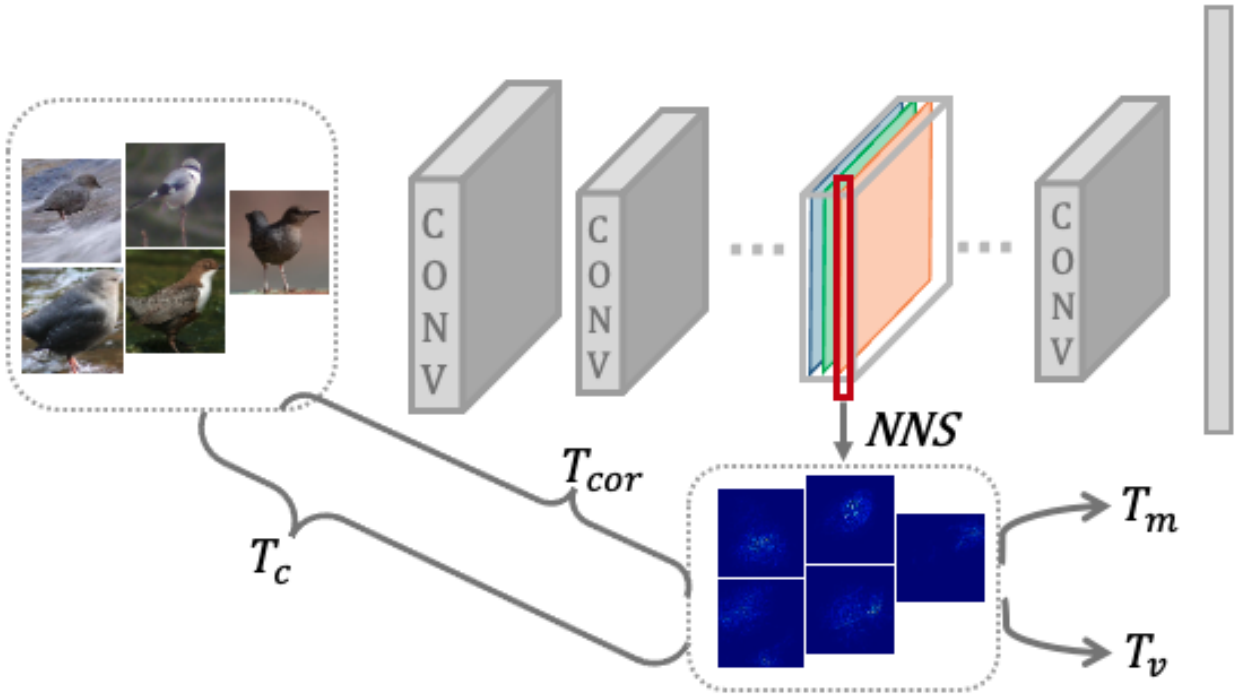




主要工作

- 提出一个多角度可扩展的可解释性框架TIP
- 基于TIP，提出三种类型的度量来解释模型
- 通过实验分析三个关键问题

Test with Interpretable Properties：
度量滤波器学习到的可解释特性



基于NNS的模型可解释特性测量框架

滤波器 k 的学习图像：

$$L^k(x) = \sum_{i,j\in[1,M]} I^k_{(i,j)}(x) \times f^k_{(i,j)}(x)$$

可解释特性：

特征强度：

$$P^k_m(x) = \frac{\sum_{i,j\in[1,M]} L^k_{(i,j)}(x)}{M \times M}$$

特征多样性：

$$P^k_v(x) = \frac{\sum_{i,j\in[1,M]} (L^k_{(i,j)}(x) - \mu_L)^2}{M \times M}$$

输入依赖性：

$$P^k_c(x) = \frac{\sum_{i,j\in[1,M]} (L^k_{(i,j)}(x) - \mu_L)(x_{(i,j)} - \mu_x)}{M \times M}$$

标准化输入相关性：

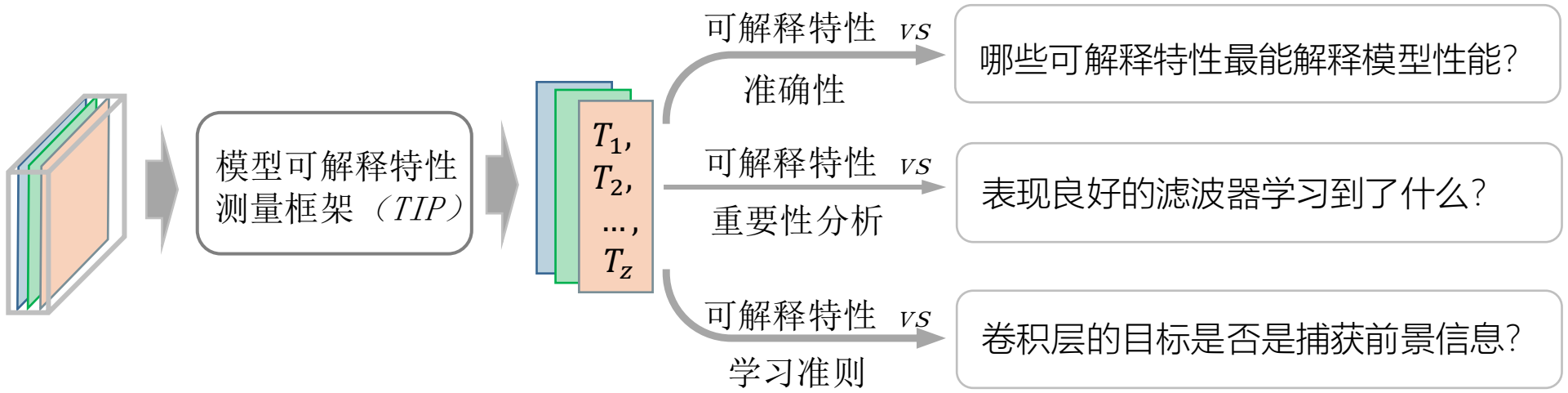
$$P^k_{cor}(x) = \frac{\sum_{i,j\in[1,M]} (L^k_{(i,j)}(x) - \mu_L)(x_{(i,j)} - \mu_x)}{\sigma_L \times \sigma_x}$$

基于可解释特性的模型分析

可解释特性测量框架（Test with Interpretable Properties，TIP）：使用模型的固有特性（可解释特性）来解释模型行为。

第k 滤波器的可解释特性定义如下：

$$T^k = \frac{\sum_{x \in X} P^k(x)}{S}$$



基于可解释特性的模型分析

➢ 模型性能分析

准确率可解释特性曲线（AIPC）：根据指定的可解释特性对滤波器进行排序，然后按排序结果从高到低依次对滤波器进行掩码。 x 轴表示掩码滤波器占比，y轴表示掩码后准确率的变化。

➢ 滤波器重要性分析

滤波器的激活值（AVoU）：对给定输入样本的响应强度

$$AVoU^k = \frac{\sum_{x \in X} \sum_{i,j \in [1,M]} f^k_{(i,j)}(x)}{M \times M \times S}$$

滤波器的敏感性（SoU）：对损失函数的影响

$$\triangle loss_w(x) = loss_w(\mathbf{1} \odot \omega) - loss_w((\mathbf{1} - \mathbf{e_w}) \odot \omega)$$

$$SoU^k = \frac{\sum_{x \in X} \sum_{w \in [1,W]} ||\Delta loss_w(x)||}{W \times S}$$

基于可解释特性的模型分析

➢ 模型学习准则

根据滤波器捕捉的前景和背景信息进行分类。rf 和 rb 是排序后满足指定条件的元素索引集合：

$$rf = \{i \in i, 2, .., K : Tf^i satisfies\ conditions\}$$

$$rb = \{i \in i, 2, .., K : Tb^i satisfies\ conditions\}$$

前景中 T 测量值

背景中 T 测量值

评估：

1. 滤波器类型的分布分析。
2. 不同滤波器类型对模型准确率的影响。

基于可解释特性的可视化结果分析

可解释特性排序与可视化：

每种可解释特性代表一种单一的解释维度，通过对比不同属性下的滤波器学习图像，从**不同维度**分析模型的学习机制。

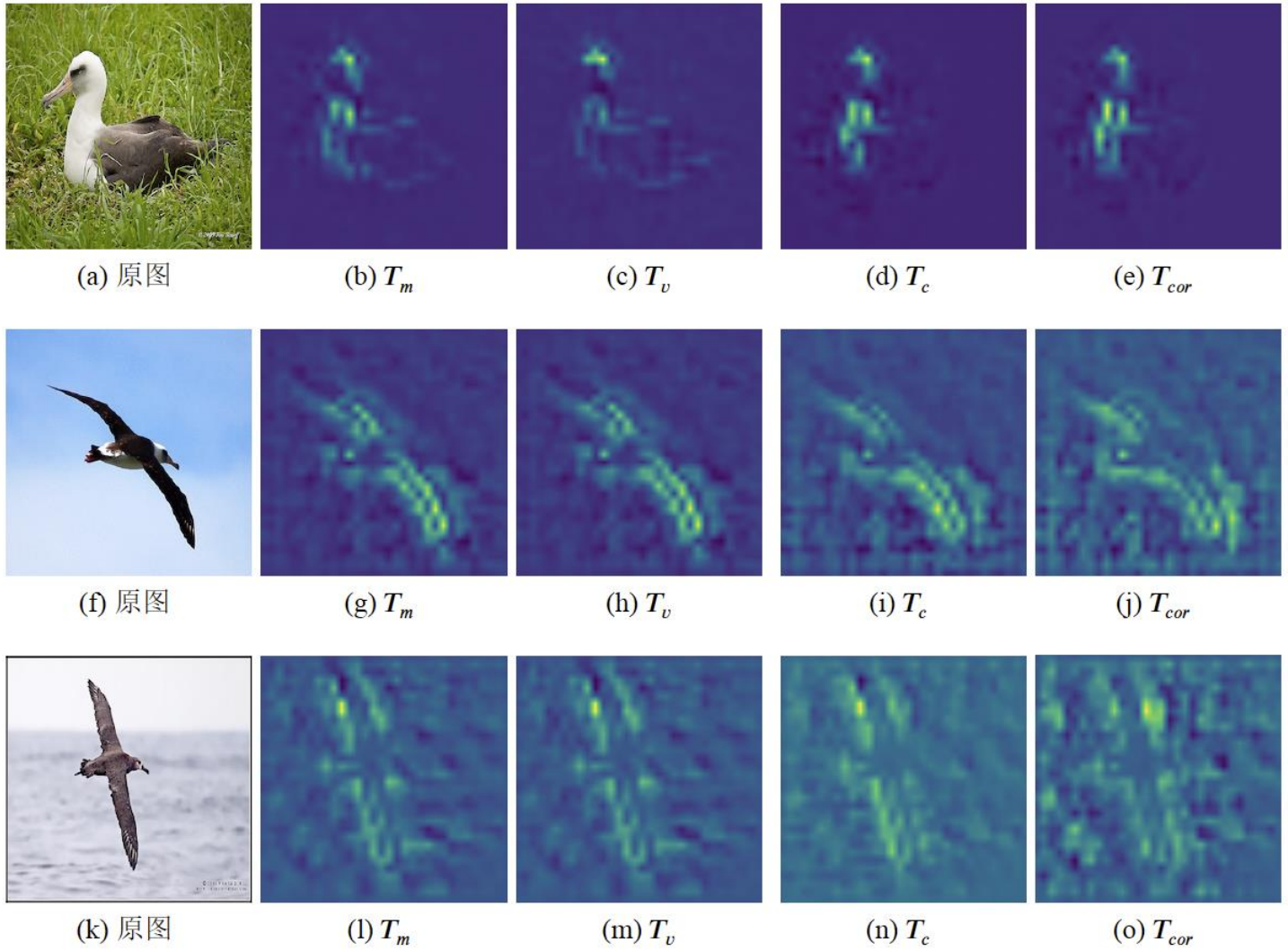


图 5-3 不同可解释特性排序下，排名最高滤波器对应的学习图像

基于可解释特性的模型分析

1. 哪些可解释特性最能解释模型性能?

不同模型的差异:

- AlexNet：T_c（分布广泛）高的滤波器更影响模型性能。
- VGG：T_m（高均值）是与模型性能相关性更高的可解释特性。

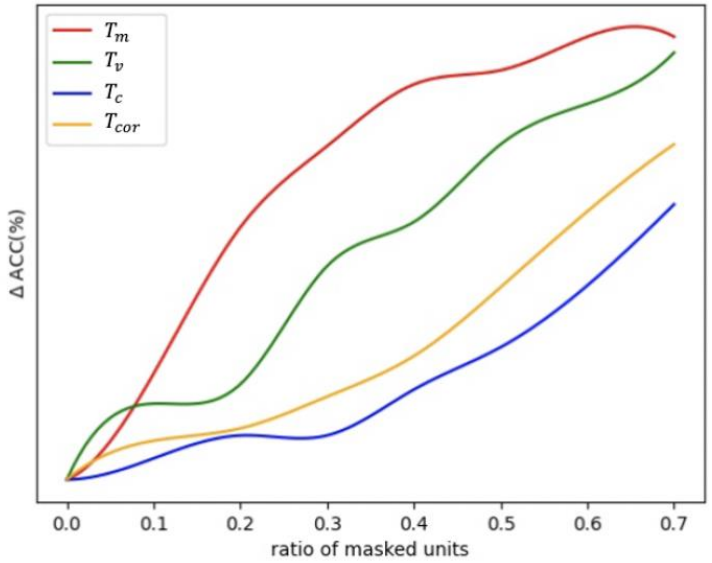


图 5-4 在 ILSVRC-2013 DET 数据集上训练的 VGG-16 模型的 APIC 值

表 5-1 APIC 曲线下的面积

	AlexNet				VGG-11				VGG-16			
	T_m	T_v	T_c	T_{cor}	T_m	T_v	T_c	T_{cor}	T_m	T_v	T_c	T_{cor}
数据 1	41.20	41.40	**52.40**	24.40	**40.60**	19.40	7.40	10.20	**55.20**	53.00	36.40	29.40
数据 2	31.51	26.05	**36.72**	28.26	**35.67**	32.94	11.20	12.24	**52.61**	52.34	21.23	14.32
数据 3	18.46	22.31	**36.92**	24.61	**25.38**	17.70	8.46	4.62	56.15	**66.15**	17.69	15.38

基于可解释特性的模型分析

2. 表现良好的滤波器学习到了什么?

采用秩相关系数，衡量显著性度量与可解释特性之间的关联程度：

$$\rho = 1 - \frac{6\sum d_i^2}{K(K^2-1)}$$

表 5-2 显著性度量与可解释性属性之间的相关系数

	AlexNet				VGG-11				VGG-16			
	T_m	T_v	T_c	T_{cor}	T_m	T_v	T_c	T_{cor}	T_m	T_v	T_c	T_{cor}
AVoU												
数据 1	**0.99**	0.97	-0.46	-0.50	**0.99**	0.98	-0.40	-0.32	**0.98**	0.84	-0.54	-0.37
数据 2	**1.00**	1.00	-0.09	-0.51	**1.00**	0.98	-0.16	-0.12	**0.98**	0.88	-0.47	-0.32
数据 3	**0.99**	0.95	0.22	0.06	**1.00**	0.99	-0.54	-0.43	**0.97**	0.83	-0.42	-0.22
SoU												
数据 1	0.74	**0.77**	-0.21	-0.26	0.83	**0.84**	-0.29	-0.22	**0.57**	0.56	-0.29	-0.16
数据 2	**0.92**	**0.92**	-0.09	-0.49	0.77	**0.78**	-0.23	-0.22	0.58	**0.61**	-0.23	-0.14
数据 3	0.75	**0.79**	0.18	0.10	**0.83**	**0.83**	-0.42	-0.33	**0.57**	0.47	-0.18	-0.06

- AVoU：高平均激活值的滤波器更关注高均值特性。
- SoU：敏感度高的滤波器通常学习到具有高均值特性或分布广泛特性。

基于可解释特性的模型分析

3. 卷积层的目标是否是捕获前景信息?

（1）卷积层的滤波器是否主要专注于学习图像中的前景信息，而较少关注背景?

滤波器在学习过程中倾向于捕捉多样化的图像特征，而非单纯聚焦于前景。

（2）专注于前景特征的滤波器是否在模型的整体性能中扮演更关键的角色?

滤波器的有效性不仅取决于是否关注前景信息，还与其是否能在前景和背景之间建立起有效的特征联系相关。

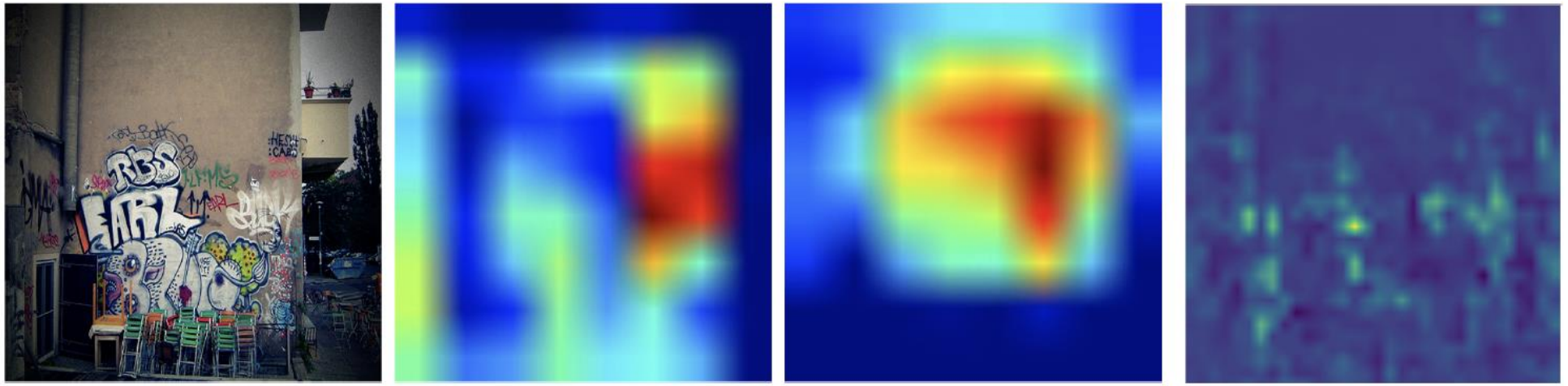
表 5-3 滤波器的分布和重要性

	Proportion of units			Change of accuracy		
	Phase1	Phase2	Phase3	Phase1	Phase2	Phase3
H-H	**0.28**	**0.28**	**0.28**	**17.4**	**22.00**	**26.00**
H-noH	0.03	0.03	0.03	0.40	0.40	0.80
noH-H	0.03	0.03	0.03	1.60	1.40	1.80
L-L	**0.29**	**0.29**	**0.29**	**0.00**	**0.40**	**0.00**
L-noL	0.02	0.02	0.02	0.20	0.20	0.80
noL-L	0.02	0.02	0.02	0.00	0.20	0.00

注释：此表按顺序展示了具有高前景-高背景、高前景-非高背景、非高前景-高背景、低前景-低背景、低前景-非低背景和非低前景-低背景的滤波器。

与其他可解释方法比较分析

与CAM 和Grad-CAM 相比，TIP 能够更清晰地定位图像中关键但区域上较小的细节特征，例如小物体、边缘轮廓等。



(a) 原图 (b) CAM (c) Grad-CAM (d) TIP

图 5-6 不同解释方法的可视化比较

度量滤波器学习到的可解释特征

一 提出了一个统一的可解释性框架TIP，采用具有数学意义的可解释特性，阐明模型的运行机制。

二 提出了三种类型的度量，分别从三个方面解释模型。

三 通过实验验证了所提出的TIP 的有效性，并分析了三个关键问题。

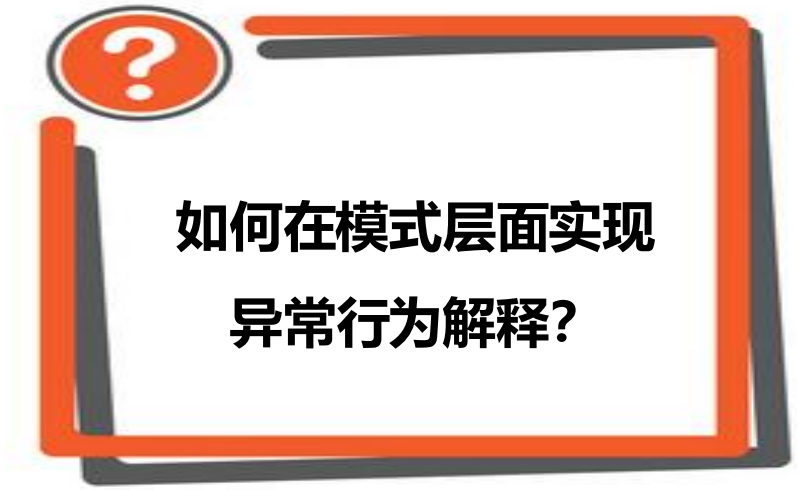
伍 基于NNS的模式层面解释

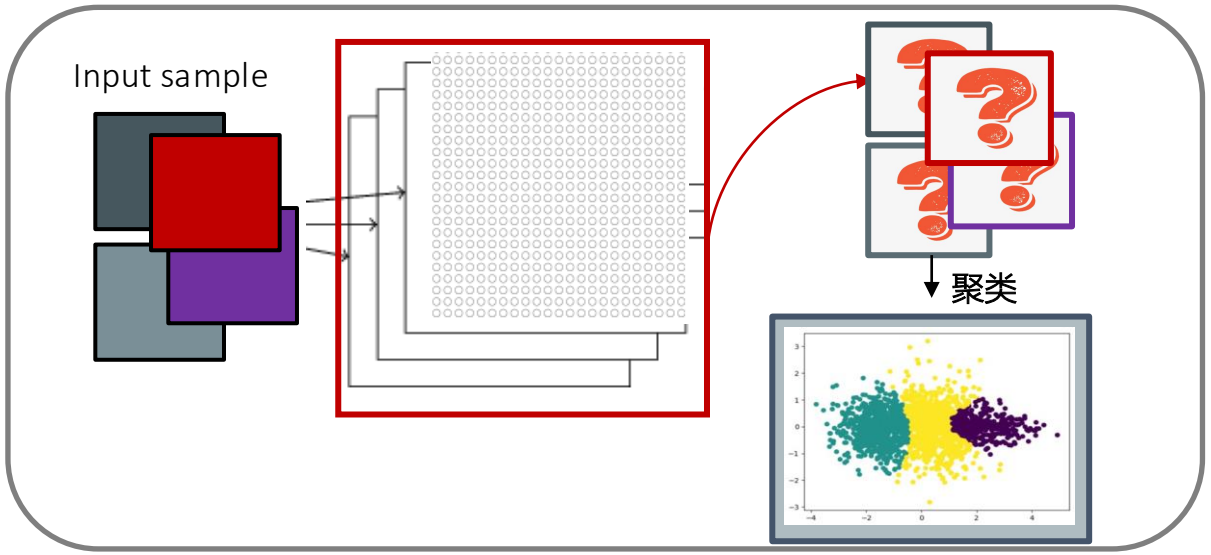
现有解释方法：

- 在识别和理解模型的**异常行为**方面仍显薄弱。

NNS在模式层面解释中的瓶颈：

- 在局部或特征层面提供了一定程度的模型解释，但这类解释缺乏**全局**结构性，难以揭示神经网络模型在更高层次上所学习到的模式。

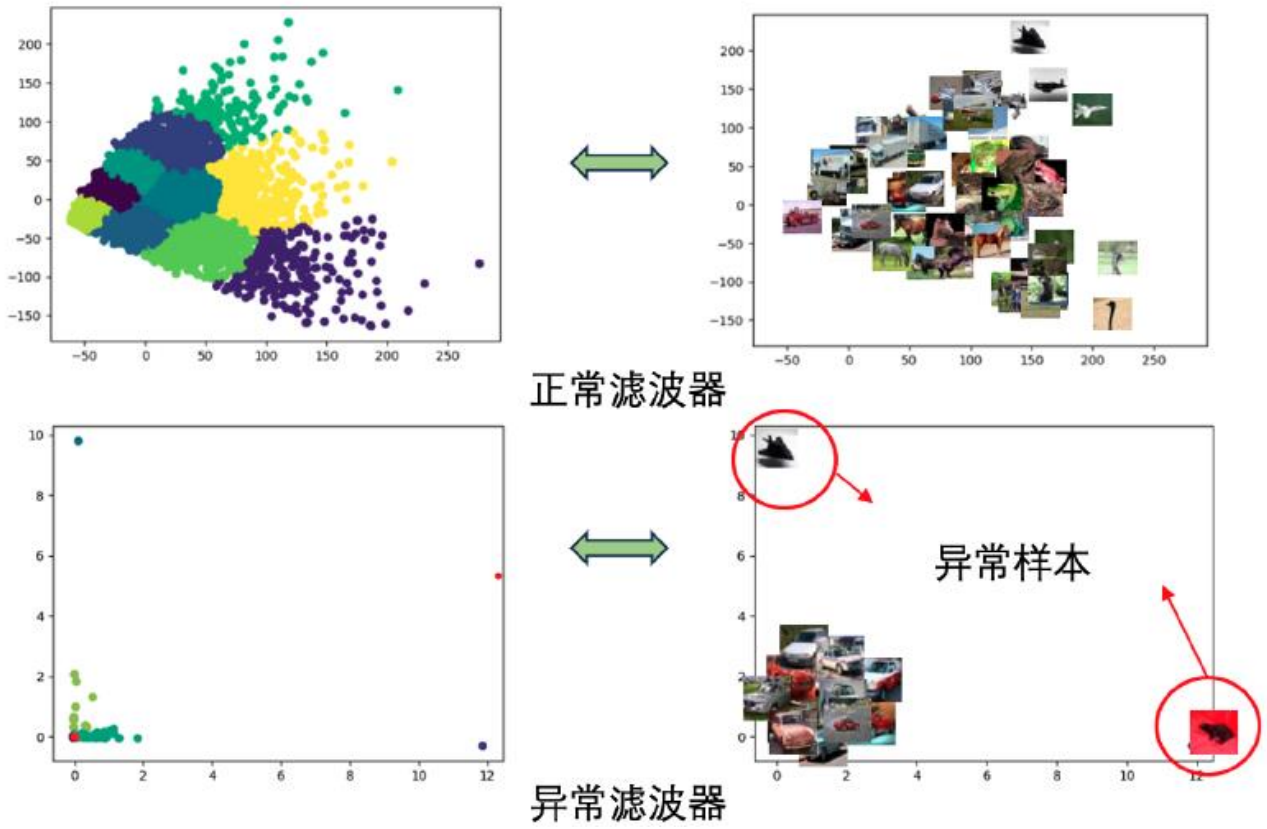




主要工作

- 提出了一种基于聚类的解释方法
- 提出了定量指标来评估异常滤波器
- 探究异常滤波器与过拟合的关系

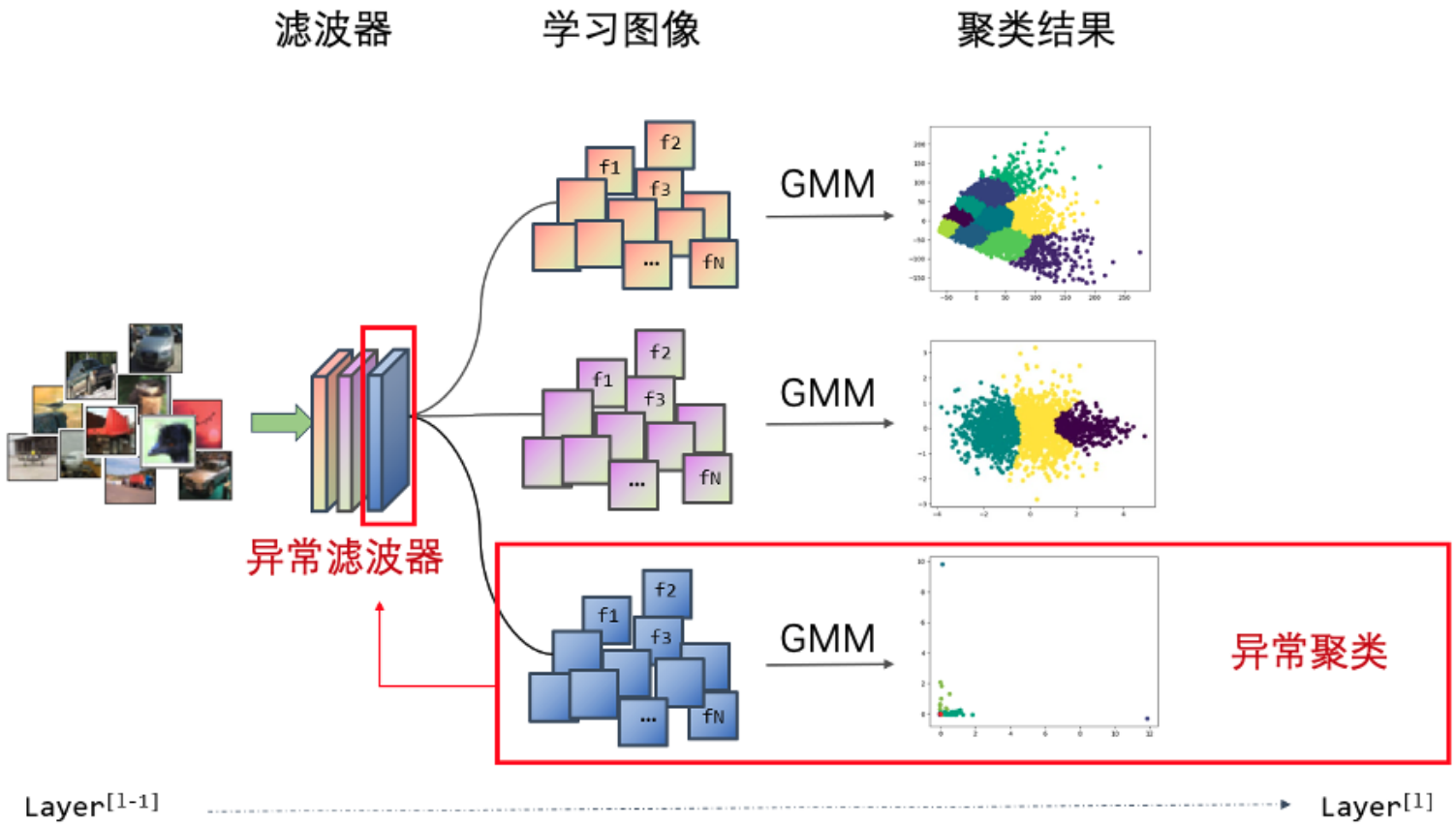
Search for Anomaly Filters:
寻找模型的异常滤波器



基于NNS的异常滤波器识别

使用高斯混合模型（Gaussian Mixture Model, GMM）对每个滤波器的学习图像进行聚类。

异常滤波器是指滤波器的聚类结果中存在离群点，讨论异常滤波器与模型过拟合的关系。



滤波器层级聚类评估

使用CH 指数作为评估指标：

SSB 和SSW 分别表示类间和类内的散度矩阵。

$$CH = \frac{SSB/(K-1)}{SSW/(N-K)},$$

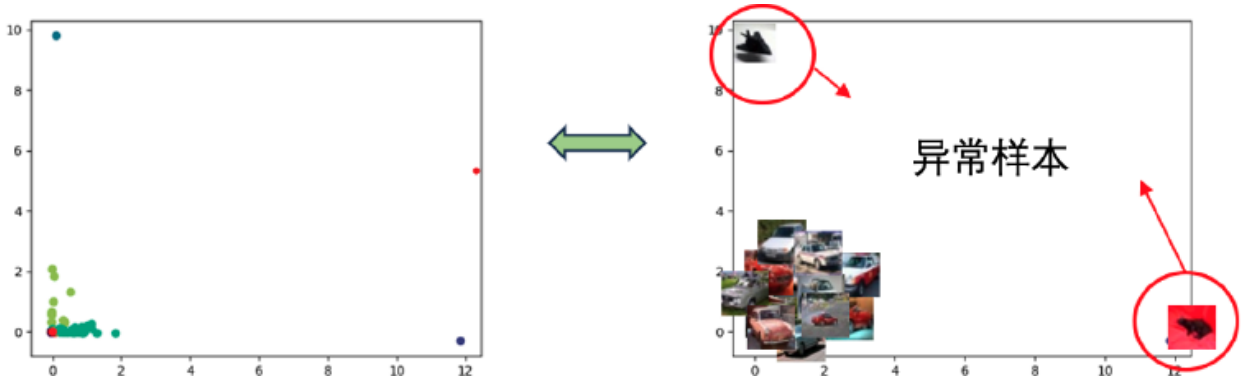
$$SSB = \sum_{i=1}^{K} Z_i \cdot ||\boldsymbol{m_i} - \boldsymbol{m}||_2.$$

$$SSW = \sum_{i=1}^{K} \sum_{i=1}^{Z_i} ||\boldsymbol{x_{ij}} - \boldsymbol{m_i}||_2,$$

聚类结果评估

异常滤波器的聚类结果的三个关键特征：

1. 类别分布不平衡
2. 异常高的CH 指数
3. 高激活值



异常滤波器分布分析

根据异常滤波器数量的变化，分析其在神经网络不同阶段的变化趋势。

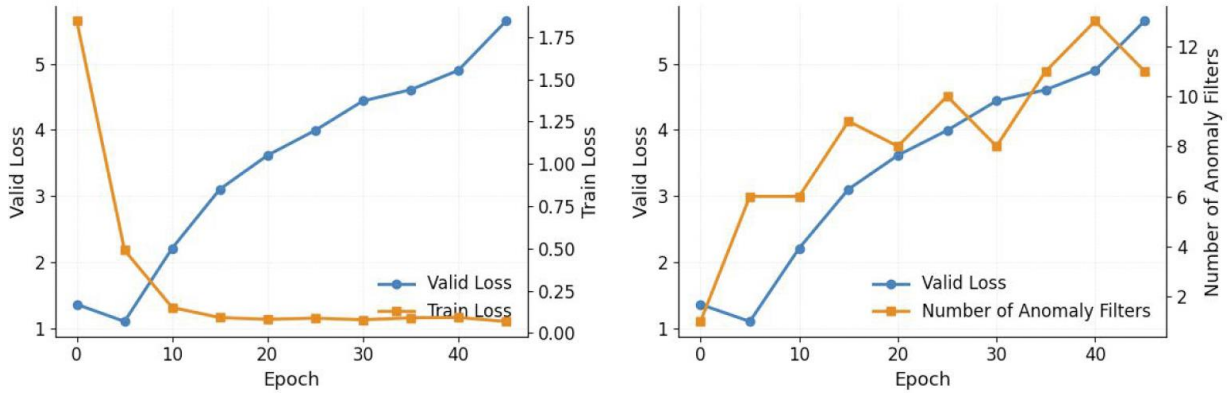


图 6-4 CNN 模型在 CIFAR-10 数据集:（a）模型损失；（b）异常滤波器数量

表 6-2 不同模型中的异常滤波器数量

模型 \ 数据集	CIFAR-10		CIFAR-100		ILSVRC-2013 DET	
	WT	OF	WT	OF	WT	OF
LeNet-5	2	**12**	8	**13**	-	-
Simple CNN	5	**12**	3	**20**	-	-
AlexNet	10	**53**	30	**37**	6	4
ResNet-18	123	**374**	104	**419**	207	**301**
VGG-16	11	10	16	**29**	2	3

在大多数配置中，过拟合模型（OF）中的异常滤波器数量显著高于良好训练模型（WT）。表明异常滤波器在过拟合模型中更为常见。

异常样本贡献评估

异常样本与模型过拟合之间的关系

表 6-3 不同样本的平均梯度值

模型 \ 数据集	CIFAR-10		CIFAR-100		ILSVRC-2013 DET	
	Outlier	Normal	Outlier	Normal	Outlier	Normal
LeNet-5 (WT)	**0.461**	0.052	**0.163**	0.032	-	-
LeNet-5 (OF)	**2.811**	0.201	**0.264**	0.162	-	-
Simple CNN (WT)	**0.079**	0.024	**0.083**	0.015	-	-
Simple CNN (OF)	**0.281**	0.066	**0.312**	0.178	-	-
AlexNet (WT)	**0.005**	0.001	**0.017**	0.002	**0.012**	0.016
AlexNet (OF)	**0.014**	0.005	**0.063**	0.005	**0.045**	0.014
ResNet-18 (WT)	0.003	0.003	**0.003**	0.002	0.006	**0.007**
ResNet-18 (OF)	0.006	**0.008**	**0.009**	0.008	0.010	0.010
VGG-16 (WT)	**0.006**	0.004	0.004	0.004	0.018	**0.020**
VGG-16 (OF)	0.003	**0.004**	**0.017**	0.010	**0.070**	0.032

在多数情况下，过拟合模型在异常样本处的平均梯度值显著高于正常模型。

表明与异常滤波器相关的异常样本促使了模型的过拟合。

异常滤波器重要性评估

异常滤波器对模型性能的影响

表 6-5 掩码不同滤波器后验证准确率变化

模型 \ 数据集	CIFAR-10		CIFAR-100		ILSVRC-2013 DET	
	Anomaly	Random	Anomaly	Random	Anomaly	Random
LeNet-5	**-1.21%**	-3.94%	**-1.27%**	-6.92%	-	-
Simple CNN	**+0.10%**	-1.51%	**-0.18%**	-0.74%	-	-
AlexNet	**+0.06%**	-0.15%	**+0.01%**	-0.43%	**+0.02%**	-0.10%
ResNet-18	**+0.15%**	-0.04%	**+0.02%**	-0.30%	**+0.01%**	-0.09%
VGG-16	**+0.01%**	-0.48%	**+0.00%**	-0.61%	**+0.02%**	-0.04%

表 6-4 掩码后发生准确率变化的滤波器数量

模型 \ 数据集	CIFAR-10		CIFAR-100		ILSVRC-2013 DET	
	训练集 ↓	验证集 ↑	训练集 ↓	验证集 ↑	训练集 ↓	验证集 ↑
LeNet-5	3/3	1/3	2/2	0/2	-	-
Simple CNN	3/3	1/3	8/8	6/8	-	-
AlexNet	6/6	5/6	13/15	13/15	4/4	2/4
ResNet-18	218/234	102/234	316/356	247/356	3/3	2/3
VGG-16	7/8	5/8	14/16	12/16	209/237	114/237

相比随机掩码方式，掩码异常滤波器对验证准确率的影响更小。表明异常滤波器与过拟合强相关。移除异常滤波器能缓解模型的过拟合。

多数情形下，掩码后的模型在训练准确率下降，而验证准确率上升。表明异常滤波器导致模型对训练集的过度学习。

寻找模型隐藏的异常滤波器

一 提出了一种细粒度的解释方法，通过无监督聚类，探索CNN 本质特征。

二 提出定量指标来检测与模型过拟合相关的异常滤波器。

三 发现一种指示过拟合与异常滤波器关系的模式，并从三个方面实验进行分析。

陆 总结与展望

神经网络扫描仪NNS

- 本章提出了一种通用的基于神经元的定性解释方法NNS
- 对神经元学习到的特征可视化。通过灵活地组合，分析神经网络的不同模块。

基于NNS的神经元层面解释

- 本章提出一种基于NNS的神经元级定量解释方法
- 通过引入特征量，解决NNS的结果具有主观性缺乏标准评估准则的问题。

基于NNS的滤波器层面解释

- 本章提出一种基于NNS的滤波器级定量解释方法
- 提出一个统一的可解释框架，解决NNS缺乏对滤波器的分析及从单一维度进行解释的问题。

基于NNS的模式层面解释

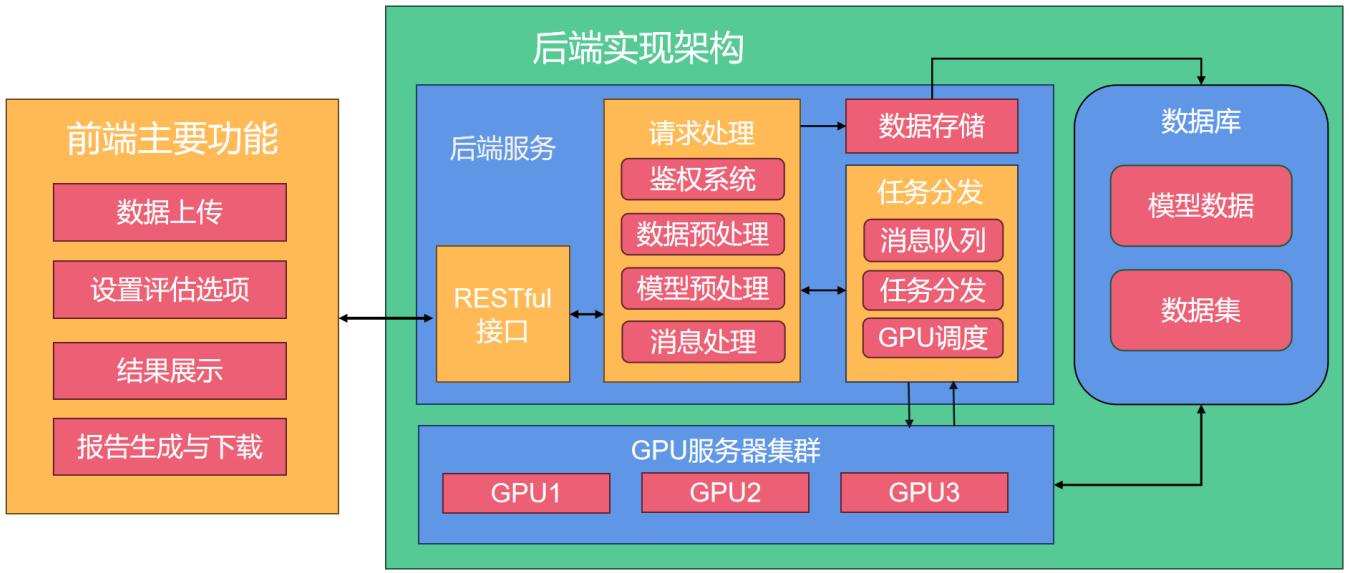
- 本章提出一种基于NNS的模式级定量解释方法
- 通过对学习图像进行聚类，探究异常滤波器与模型过拟合之间的关系。

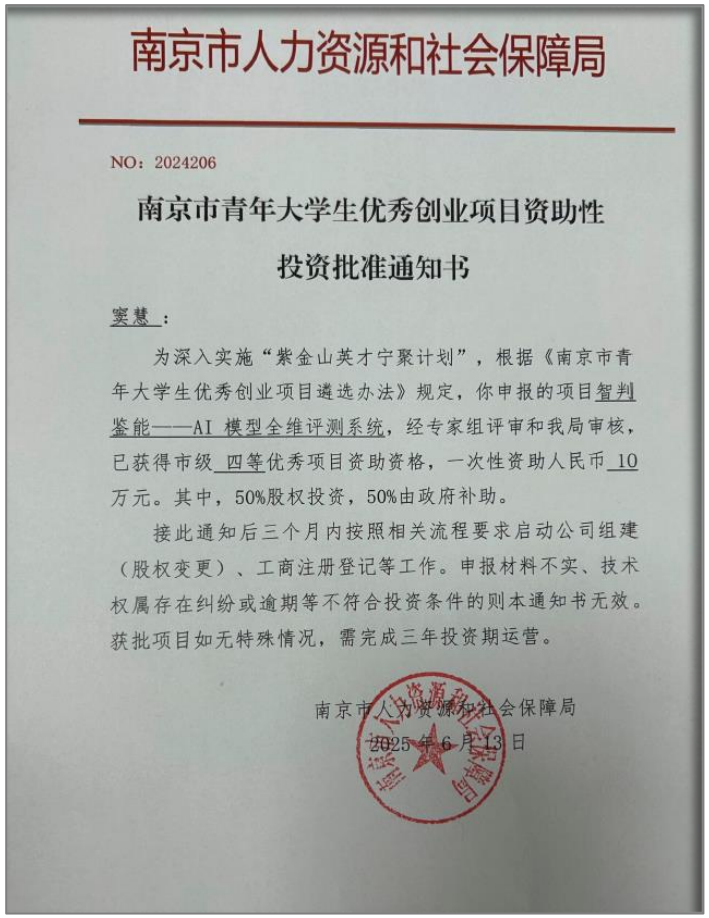
神经网络模型评测系统——进行多维度的评估，包括基础性能监控、可解释性分析、鲁棒性评估和潜力分析等，形成模型评估报告。

待评估模型 →

指标	结果
准确率	0.5
错误率	0.5
运行时间	0.00874
宏查全率（宏召回率）	0.001
宏查准率	0.001
宏F1分数	0.001
微查全率（微召回率）	0.5
微查准率	0.5
微F1分数	0.5
模型性能得分	3.34

...





南京市人力资源和社会保障局

NO：2024206

南京市青年大学生优秀创业项目资助性投资批准通知书

窦慧：

为深入实施“紫金山英才宁聚计划”，根据《南京市青年大学生优秀创业项目遴选办法》规定，你申报的项目智判鉴能——AI 模型全维评测系统，经专家组评审和我局审核，已获得市级 四等优秀项目资助资格，一次性资助人民币 10 万元。其中，50%股权投资，50%由政府补助。

接此通知后三个月内按照相关流程要求启动公司组建（股权变更）、工商注册登记等工作。申报材料不实、技术权属存在纠纷或逾期等不符合投资条件的则本通知书无效。获批项目如无特殊情况，需完成三年投资期运营。

南京市人力资源和社会保障局

2025 年 6 月 13 日

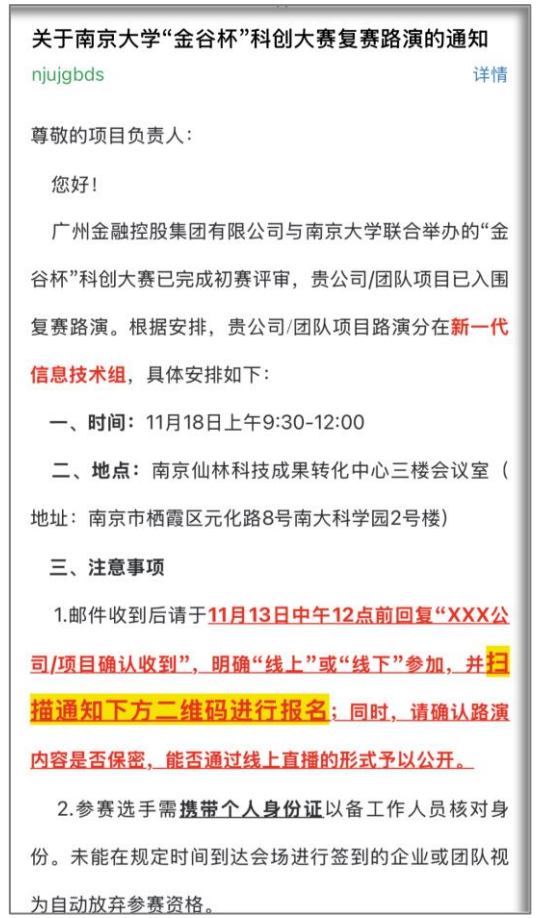
巾帼创新绽芳华｜南大博士团队勇闯 AI“无人区”：从南大科技园启程，打造模型安全评估新标杆

南大科技园NJUSP 2025年10月15日 17:24

江苏 听全文 星标

南京大学 国家大学科技园 NANJING UNIVERSITY NATIONAL SCIENCE PARK

在人工智能快速发展的浪潮中，依托南京大学国家大学科技园（以下简称“南大科技园”）的创新孵化平台，南京大学计算机专业博士生窦慧带领团队在 AI 安全领域开辟新赛道，成功研发“模型安全与性能评估系统”。这套系统不仅为 AI 模型的可靠性与可解释性提供了全新解决方案，更填补了国内模型全维度评估的技术空白。团队以科技创新诠释了新时代女性创业者的价值，在自动驾驶、医疗诊断等关键场景中展现出广阔应用前景，成为南大科技园培育高校前沿科研成果、推动科技成果转化的生动实践。



关于南京大学“金谷杯”科创大赛复赛路演的通知

njujgbds 详情

尊敬的项目负责人：

您好！

广州金融控股集团有限公司与南京大学联合举办的“金谷杯”科创大赛已完成初赛评审，贵公司/团队项目已入围复赛路演。根据安排，贵公司/团队项目路演分在新一代信息技术组，具体安排如下：

一、时间：11月18日上午9:30-12:00

二、地点：南京仙林科技成果转化中心三楼会议室（地址：南京市栖霞区元化路8号南大科学园2号楼）

三、注意事项

1.邮件收到后请于11月13日中午12点前回复“XXX公司/项目确认收到”，明确“线上”或“线下”参加，并扫描通知下方二维码进行报名；同时，请确认路演内容是否保密，能否通过线上直播的形式予以公开。

2.参赛选手需携带个人身份证以备工作人员核对身份。未能在规定时间到达会场进行签到的企业或团队视为自动放弃参赛资格。

2025年南京大学大学生创业训练计划立项项目名单公示

发布者：院办公室 发布时间：2025-04-15 浏览次数： 4522

28	智判鉴能——AI模型全维评测系统	窦慧	博士生	计算机学院	穆欣雨，韦奕，叶子琪，郭梦遥	申富饶	重点项目

模型单元学习模式分析

分析单元的表现模式
- 层次化特征抽取
- 概念链

分析单元的学习模式
- 学习行为演化
- 架构功能区分

应用验证 ⇅ 理论支持

基于可解释性的模型优化

核心问题	解决思路
可解释+剪枝技术	识别在不同任务中无关或低贡献的单元

基于可解释性的模型安全

核心问题	解决思路
可解释性+对抗攻击检测	揭示网络易受攻击的特征模式和高风险模型单元

已发表成果：

1. **Hui Dou**, Furao Shen, and Jian Zhao. Understanding neural network through neuron level visualization, *Neural Networks, 2023*. (CCF-B 类期刊, 中科院分区1 区, IF: 7.8)
2. **Hui Dou**, Baile Xu, Furao Shen, and Jian Zhao. V-SOINN: A topology preserving visualization method for multidimensional data, *Neurocomputing, 2021*. (CCF-C 类期刊, 中科院分区2 区, top, IF: 6)
3. **窦慧**, 张凌茗, 韩峰, 申富饶, 赵健. 卷积神经网络的可解释性研究综述, *软件学报, 2024*.(CCF-A 类中文期刊, 中文核心期刊, 影响因子: 2.138)
4. Yang, Hongchao, Suorong Yang, **Hui Dou**, Furao Shen. CS-QCFS: Bridging the performance gap in ultra-low latency spiking neural networks. *Neural Networks,2025*. (CCF-B 类期刊, 中科院分区1 区, IF: 7.8)
5. **窦慧**, 徐百乐, 申富饶. 一种支持拓扑结构保持的高维数据可视化方法, ZL201911179884.0,2023. (已授权)

目前在投论文：

1. **Hui Dou**, Xinyu Mu, Furao Shen, and Jian Zhao. Explaining Model Overfitting in CNNs via GMM Clustering.
2. **Hui Dou**, Furao Shen, and Jian Zhao. A Unified Approach to Explaining CNNs through Interpretable Properties.
3. **Hui Dou**, Xinyu Mu, Furao Shen, and Jian Zhao. Explaining the Convolutional Layer from the Neuron-Level Perspective.

◆ 科研课题

- 面向增量式无监督学习的新型神经网络研究，**国家自然基金面上项目**，2023.01至2026.12. 参与。
- 基于神经可塑性的脉冲网络高效学习机制与类脑智能系统，**科技部科技创新2030 重大项目**，2022.01至2026.12. 参与。
- 基于深度感知增量式联想记忆神经网络的信息融合系统研究，**国家自然基金面上项目**，2019.01至2022.12。
- 神经网络模型评测系统科研课题，**自研项目**，2021.10至2025.12，主要负责。

◆ 专著教材

- 申富饶， **窦慧**， 郭苏涵， 易梦军等。 简明神经网络[M]。 机械工业出版社， 2025。
  (该教材为"十四五"规划省级重点教材)
- 申富饶， **窦慧**， 郭苏涵， 易梦军等。 理解深度学习[M]。 (初稿完成)
- 申富饶， **窦慧**， 郭苏涵， 易梦军等。 剖析大模型[M]。 (编写中)
- 申富饶， 徐百乐， **窦慧**等。 自组织增量学习神经网络[M]。 电子工业出版社，2024。
  (该专著为2023 年度工信学术出版基金资助项目)

感谢各位老师

敬请批评指正