南京大学 NANJING UNIVERSITY

RINC Robotic Intelligence & Neural Computing Group

基于前向传播的神经网络可解释性算法研究

答辩人：张凌茗 MG21330071

导师：申富饶 教授

日期：2024年5月15日

誠樸雄偉 勵學敦行

RINC
Robotic Intelligence & Neural Computing Group

目录

壹 **研究背景**

贰 **研究内容**

- 基于前向传播的特征可视化可解释算法
- 基于前向传播的特征归因可解释算法

叁 **实际应用**

肆 **研究生期间工作成果**

伍 **总结**

誠樸雄偉 勵學敦行

RINC
Robotic Intelligence & Neural Computing Group

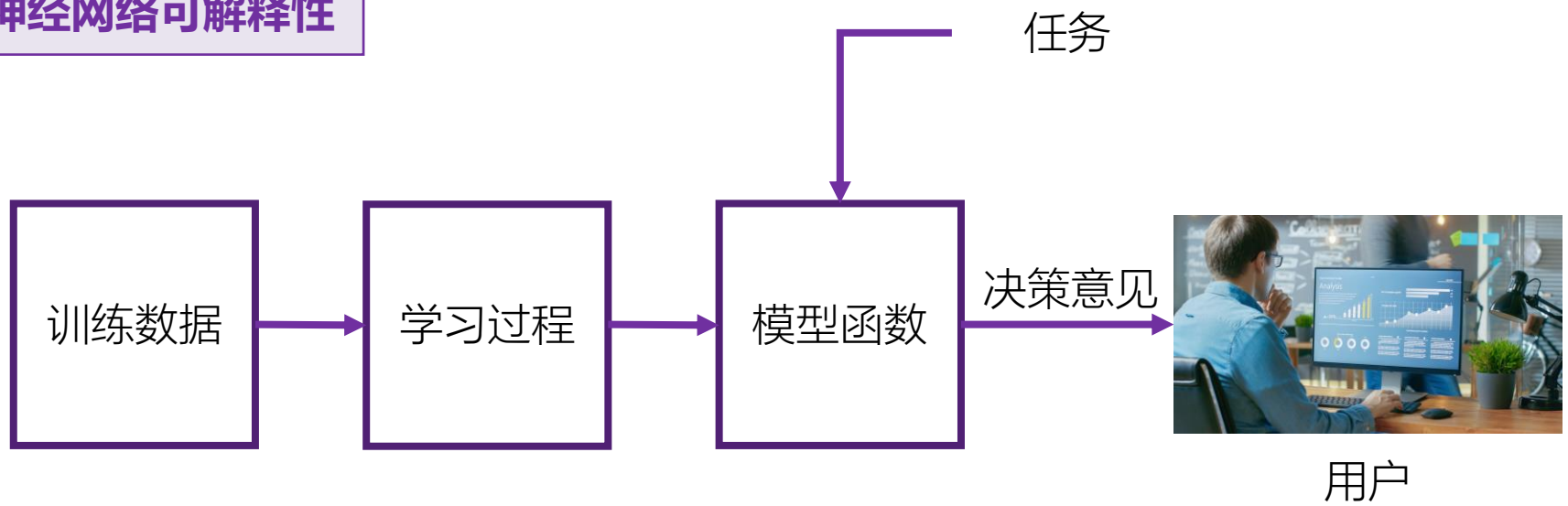
第一部分

研究背景

Research Background

神经网络可解释性 | 研究意义 | 困难与挑战

1.1 神经网络可解释性



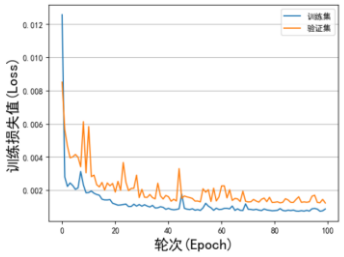
- 为什么是这个结论
- 为什么不是其他的结论
- 什么时候模型会成功
- 什么时候模型会失败
- 什么时候可以信任模型
- 如何去得知模型的错误并纠正错误

南京大学 NANJING UNIVERSITY

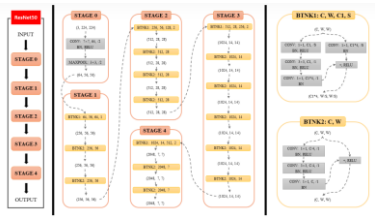
1.2 研究意义

意义1

了解神经网络决策过程，揭示内部**工作原理**以及机制。



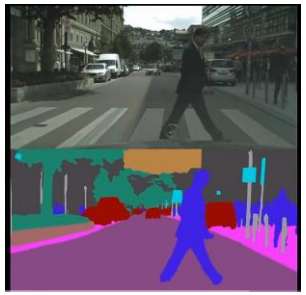
意义2



通过对模型学习过程的了解，进一步对模型进行**优化**，并有助于设计更好的模型架构。

意义3

涉及安全与健康的关键领域，需要通过可解释性提高神经网络的**信任度以及可靠性**。



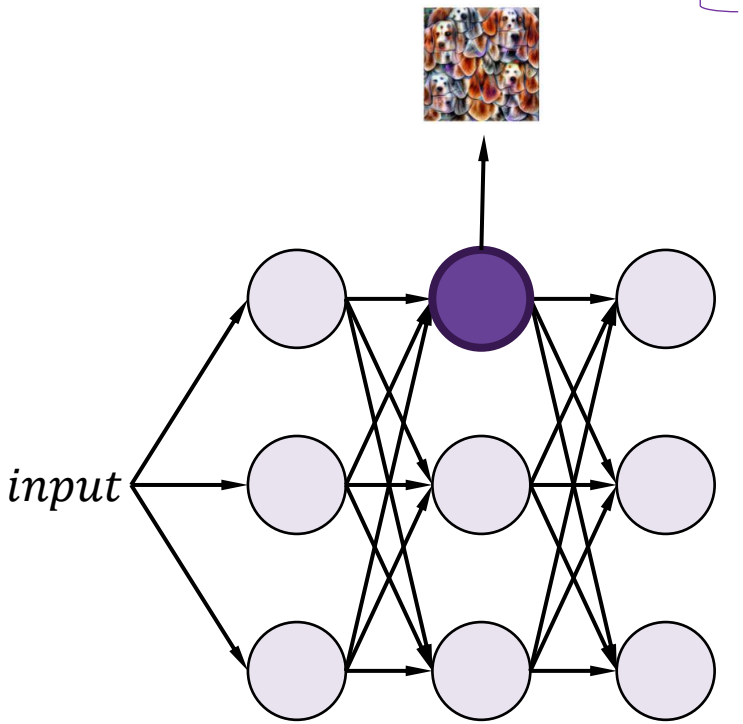
意义4

从**法律和伦理**角度，人类对模型算法的透明度和可解释性提出了要求

南京大学 NANJING UNIVERSITY

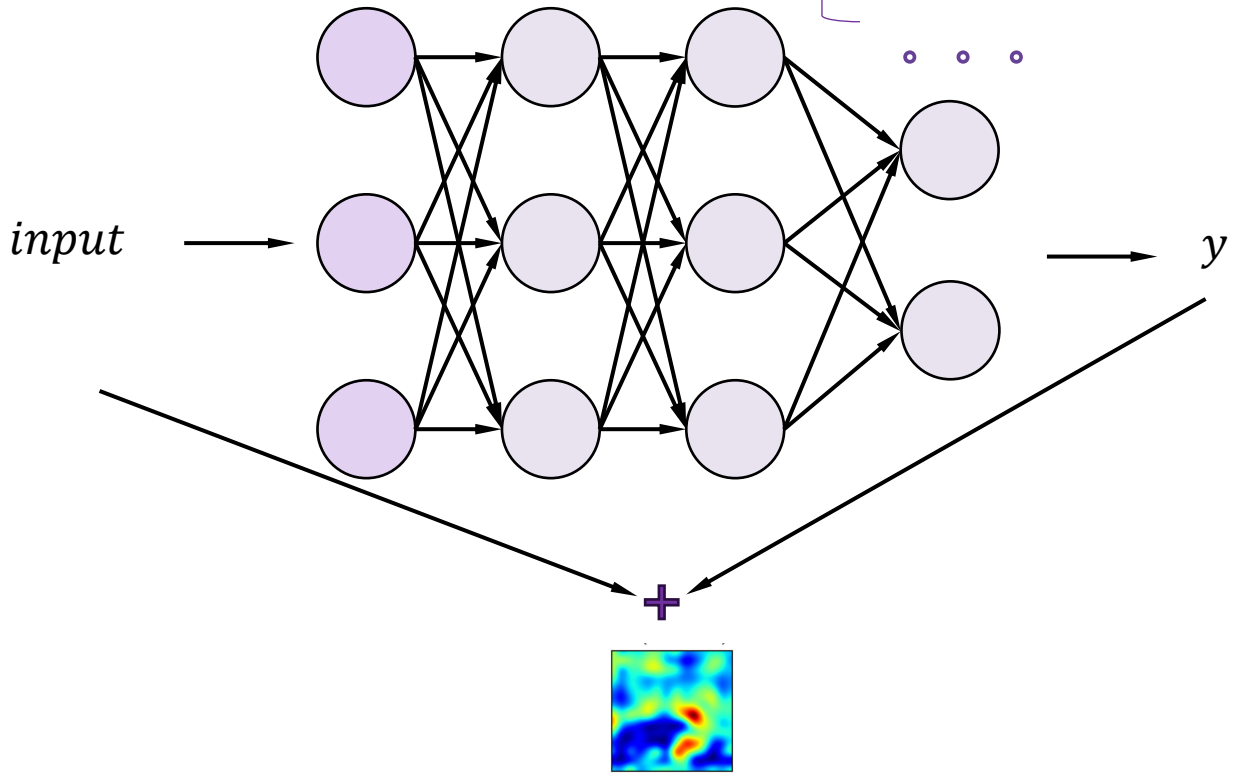
1.3 研究现状

特征可视化
- 理想样本
- 激活最大化
- 。。。



- 特征可视化挖掘神经网络内部机制
- 解释结果多数为高维特征变量，难以具有明确的语义信息。

特征归因图
- 基于梯度
- 基于扰动
- 。。。



- 特征归因图关注输入与输出之间的联系
- 忽略了对神经网络内部的运作机制的探索，与白盒解释方法接近。

1.4 困难与挑战

- 需要对内部特征变量进行合适的维度控制；
- 解释结果需要与人类主观判断一致。

- 解释方法对整个网络中的任意目标有效；
- 解释方法适用于任意的网络结构。

第二部分

研究内容

Research Content

基于前向传播的特征可视化可解释算法 | 基于前向传播的特征归因可解释算法

引言

两项工作的关系

相关性分数传播

以神经元为目标解释单元
(注重神经网络运算过程)

以输出为目标解释单元
(注重神经网络输出结果)

工作1

基于前向传播的特征可视化

特征可视化

工作2

基于前向传播的特征归因

特征归因

南京大学 NANJING UNIVERSITY

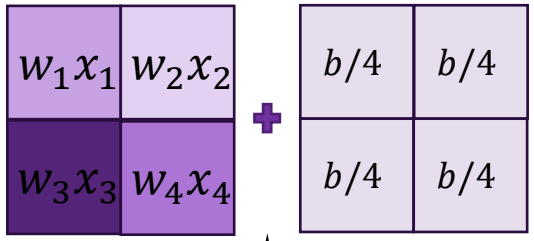
2.1 基于前向传播的特征可视化可解释算法：算法步骤

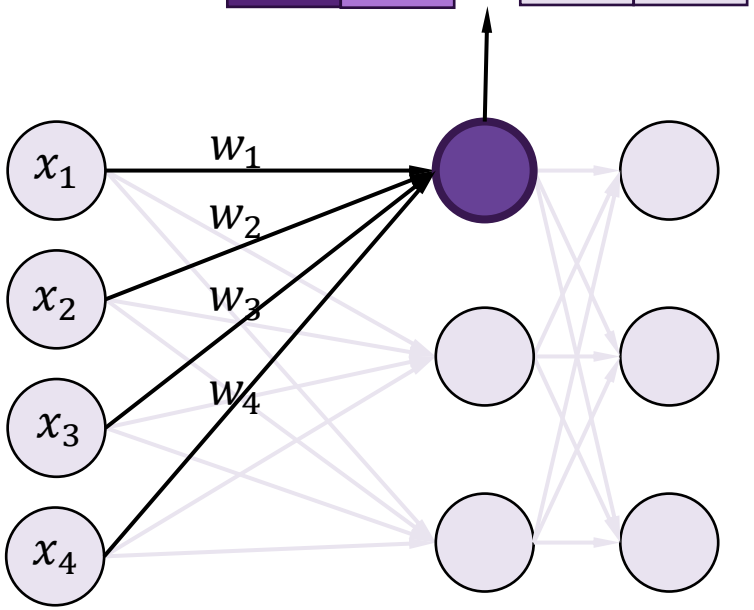
※ 以2x2的输入样本为示例，不同的颜色表示不同的权重

神经网络第一层的解释结果：

$$V_1^1 = \left(w_1^1 \odot X + W_b\right) \times \left|\operatorname{sgn}\left(Out_1^1\right)\right|$$

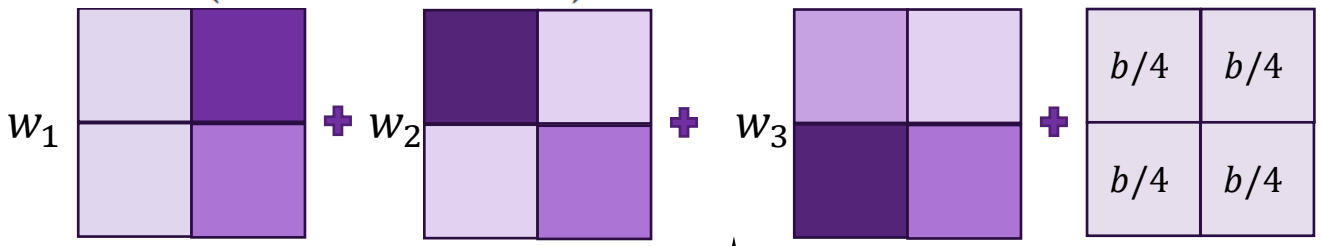
$$\left(W_b\right)_{x,y} = \frac{b}{N}$$

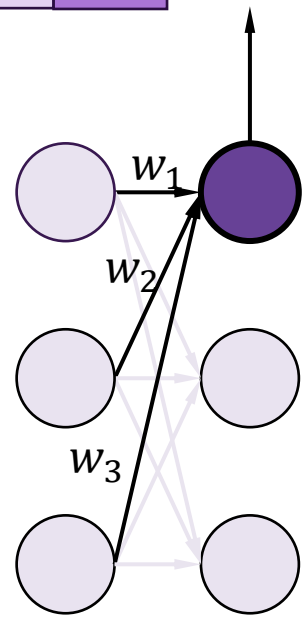




神经网络中间层的解释结果：

$$V_i^l = \left(\sum_{j=1}^{(l-1)} w_{ij}^{(l)} V_j^{l-1} + W_{b_i^l}\right) \times \left|\operatorname{sgn}\left(Out_i^l\right)\right| \text{ where } l \neq 1$$





RINC Robotic Intelligence & Neural Computing Group

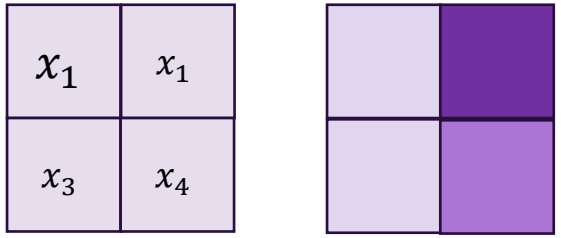
南京大学 NANJING UNIVERSITY

2.1 基于前向传播的特征可视化可解释算法：特点分析

※ 以2x2的输入样本为示例，不同的颜色表示不同的权重

神经网络第一层的解释结果

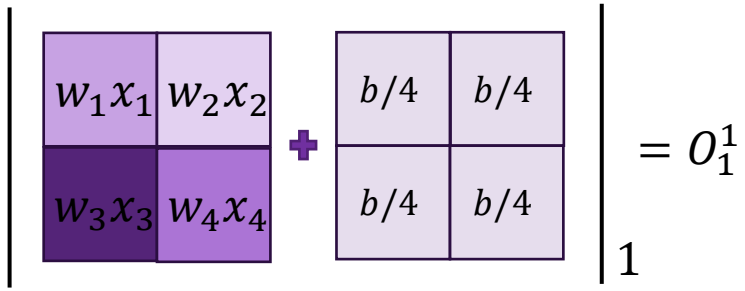
解释结果与原输入尺度一致



- 保证视觉上的尺寸一致；
- 便于与原输入图像进行对比。

通过前向传播得到其他神经元的解释结果

解释结果L1范数等于目标解释单元的特征值



- 与原神经网络前向传播方法保持一致；
- 不带入额外的噪声。

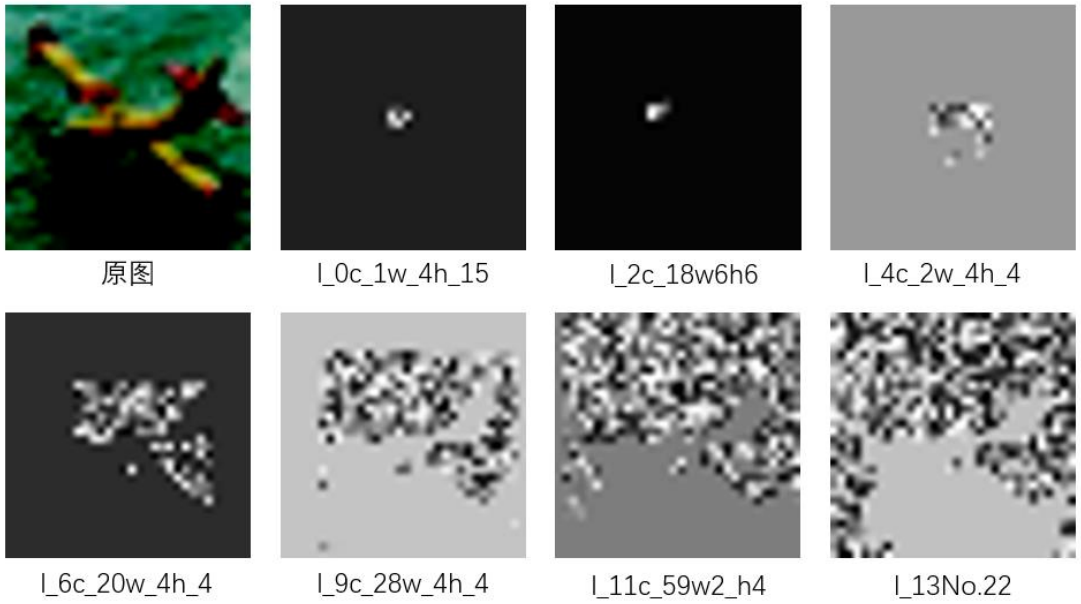
RINC Robotic Intelligence & Neural Computing Group

南京大学 NANJING UNIVERSITY

2.1 基于前向传播的特征可视化可解释算法：可视化实验

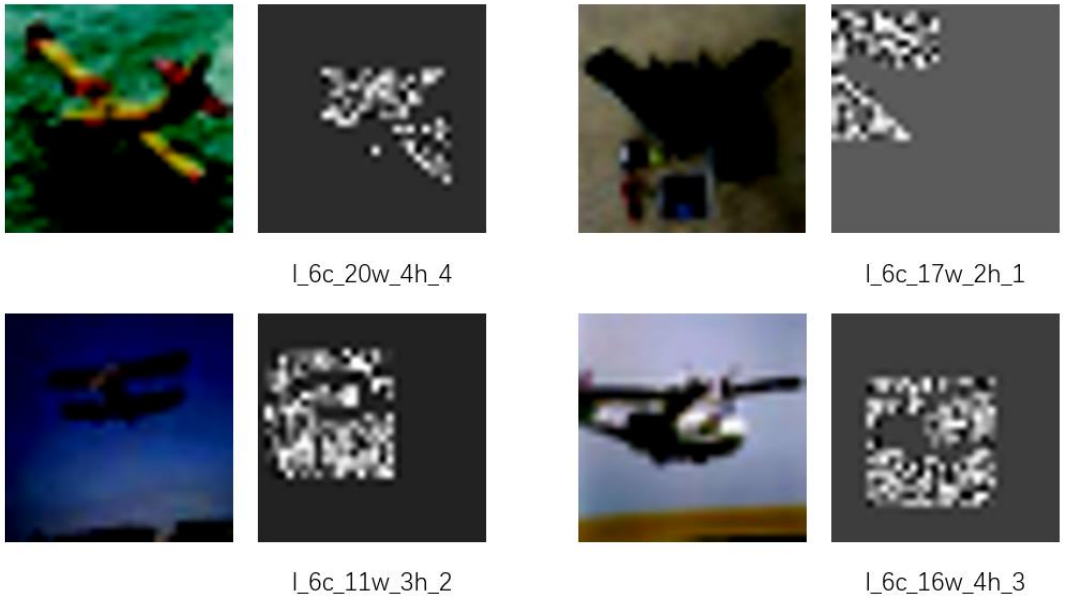
※ l_0c_1w_4h_15表示第0层第1个通道第4列第15个神经元

对同一输入不同目标单元的解释结果



解释结果展示了卷积神经网络中神经元的感受野变化；
深层的神经元具有更大的感受野。

对不同一输入不同目标单元的解释结果



解释结果展示了神经元对飞机目标的轮廓感知；
对不同的输入，神经元在对特定目标和模式进行感知。

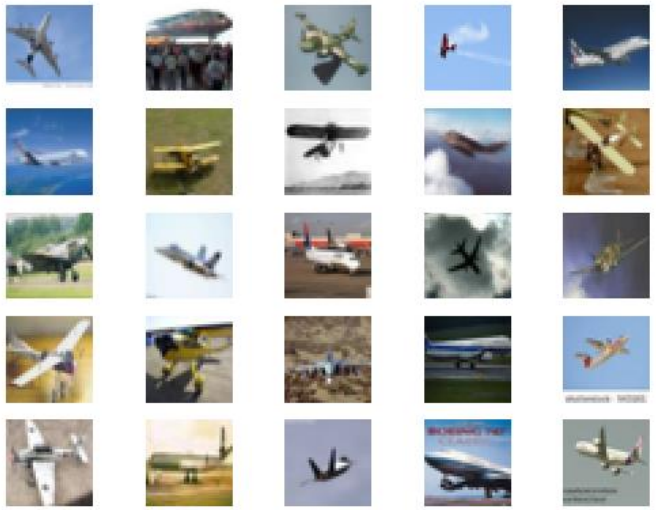
- 解释结果依据网络结构，能充分体现神经网络的特征；√
- 解释结果具有一定的语义信息。

RINC Robotic Intelligence & Neural Computing Group

南京大学 NANJING UNIVERSITY

2.1 基于前向传播的特征可视化可解释算法：可视化实验

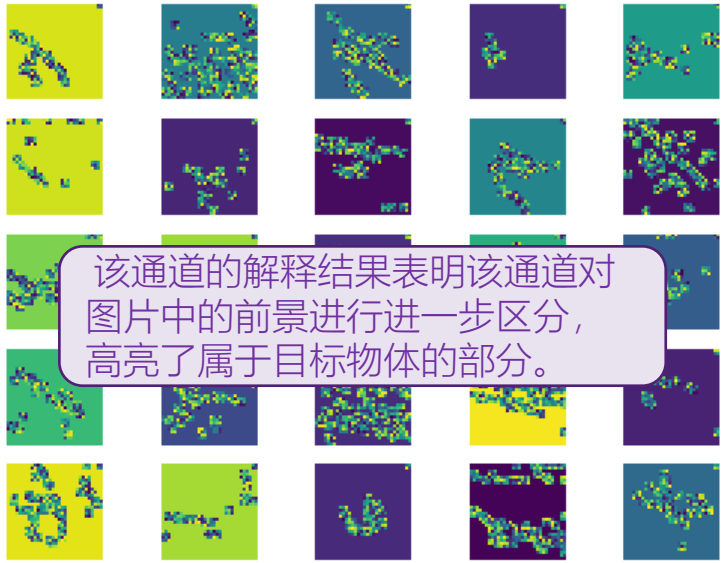
解释结果在通道上的聚焦具有语义信息



第2层第20通道

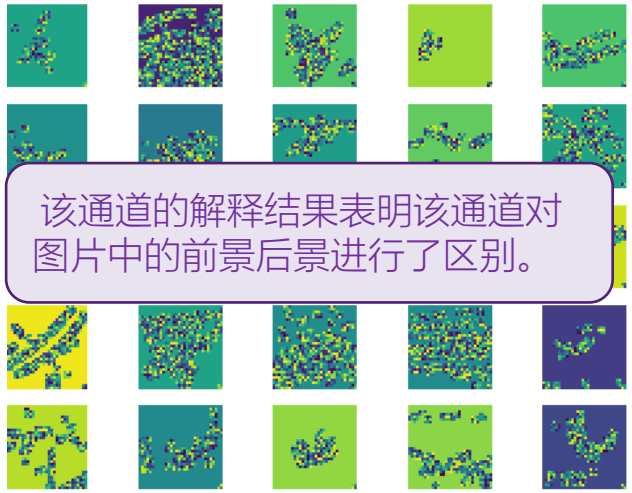
该通道的解释结果表明该通道对目标物体的一些关键点进行了检测。

第2层第62通道



该通道的解释结果表明该通道对图片中的前景进行进一步区分，高亮了属于目标物体的部分。

第1层第30通道



该通道的解释结果表明该通道对图片中的前景后景进行了区别。

- 解释结果依据网络结构，能充分体现神经网络的特征；
- 解释结果具有一定的语义信息。√

RINC Robotic Intelligence & Neural Computing Group

南京大学 NANJING UNIVERSITY

2.1 基于前向传播的特征可视化可解释算法：有效性实验

根据语义信息主观挑选通道

根据分类的结果进行聚类与剪枝分析

结构化剪枝准确率检测

通道内特征向量聚类分析

方法	训练集准确率	测试集准确率
baseline	99.87%	73.97%
random	91.35%	69.23%
good_channel	**87.47%**	**67.24%**
bad_channel	95.40%	70.56%

channel	ARI(avg)
all	0.079 57
good_channel	**0.085 09**
bad_channel	0.072 20

channel	silhouette(avg)
all	0.240 044 73
good_channel	**0.264 953 74**
bad_channel	0.129 083 13

- 具有语义信息的通道对网络分类能力更为重要；
- 具有语义信息的通道对输入样本的抽象能力更强；
- 该方法产生的解释结果不仅能体现通道的语义信息，也能体现通道对整个网络的重要性。

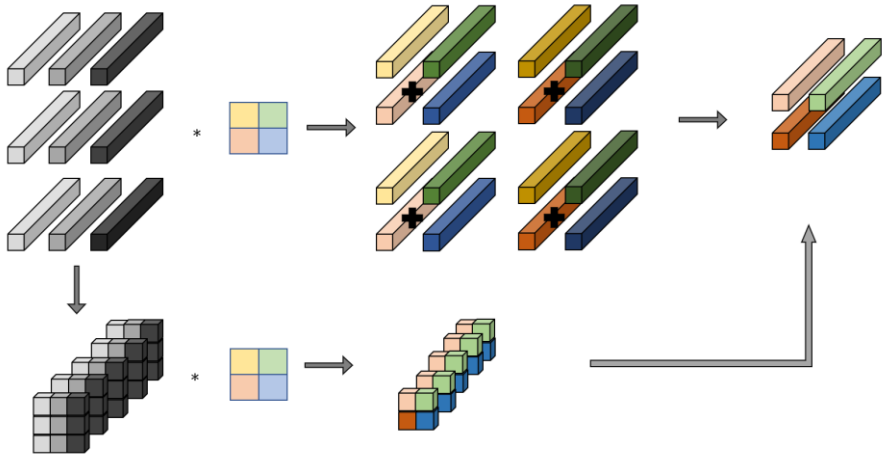
RINC Robotic Intelligence & Neural Computing Group

南京大学 NANJING UNIVERSITY

2.2 基于前向传播的特征归因可解释算法：算法原理

该方法可以为不同的网络结构生成显著性图

卷积神经网络特征归因：



算法 1 卷积层相关性分数计算算法

输入： 当前层所有的神经元节点的相关性分数矩阵 $V_{\text{in}}^0, V_{\text{in}}^1, \cdots, V_{\text{in}}^{W_{\text{in}} \times H_{\text{in}}}$，卷积核参数矩阵 K_{conv}

输出： 经过卷积层后输出节点的相关性分数矩阵 $V_{\text{out}}^0, V_{\text{out}}^1, \cdots, V_{\text{out}}^{W_{\text{out}} \times H_{\text{out}}}$

1: 所有输入节点的相关性分数矩阵拼接成 $V_{\text{in}} = cat\left(V_{\text{in}}^0, V_{\text{in}}^1, \cdots, V_{\text{in}}^{W_{\text{in}} \times H_{\text{in}}}\right)$

2: 将 V_{in} 进行转置，将维度 $\text{dim}_{\text{in}}^0, \text{dim}_{\text{in}}^1, \text{dim}_{\text{in}}^2$ 转置为 $\text{dim}_{\text{in}}^2, \text{dim}_{\text{in}}^0, \text{dim}_{\text{in}}^1$

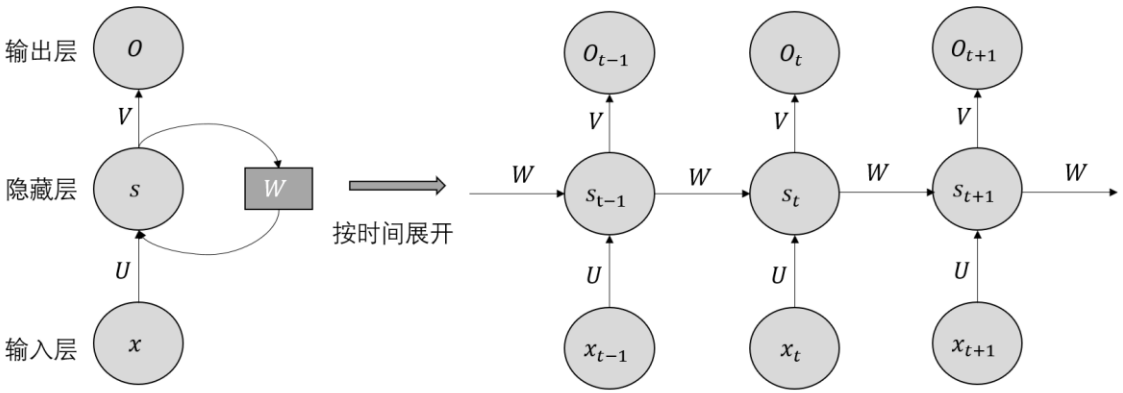
3: **for** $i = 0$ **to** N **do**

4: $\quad V_{\text{out}}^i = V_{\text{in}}^i * K_{conv}$

5: **end for**

6: 将 V_{out} 进行转置，将维度 $\text{dim}_{\text{out}}^0, \text{dim}_{\text{out}}^1, \text{dim}_{\text{out}}^2$ 转置为 $\text{dim}_{\text{out}}^1, \text{dim}_{\text{out}}^2, \text{dim}_{\text{out}}^0$

循环神经网络特征归因：

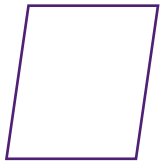


对于 x_t，以及任意的 k，$k \geqslant t$ 可以递推出 S_k 以及 O_k 的显著性分数为：

$$
\begin{aligned}
Score_{x_t}^{o_t} &= g\left(V \times f\left(U \odot x_t\right)\right) \times |\text{sgn}\left(o_t\right)|, \\
Score_{x_t}^{s_t} &= f\left(U \odot x_t\right) \times |\text{sgn}\left(s_t\right)|, \\
Score_{x_t}^{o_k} &= g\left(V \times f\left(W \times Score_{x_t}^{s_k}\right)\right) \times |\text{sgn}\left(o_t\right)|, \text{where } t < k, \\
Score_{x_t}^{s_k} &= f\left(W \times Score_{x_t}^{s_{k-1}}\right) \times |\text{sgn}\left(s_t\right)|, \text{where } t < k.
\end{aligned}
\tag{4}
$$

南京大学 NANJING UNIVERSITY

2.2 基于前向传播的特征归因可解释算法：对比实验

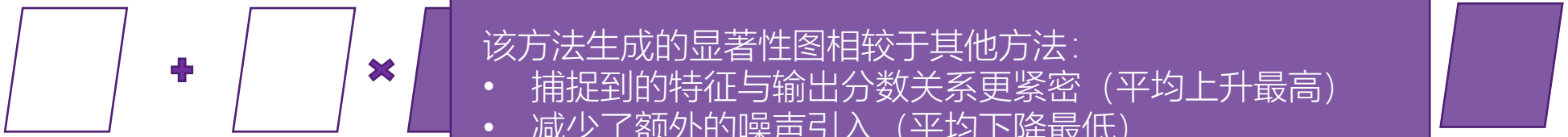

原图


显著性图

敏感度分数：

平均上升指标 $\sum_{i=1}^{N} \frac{\text{Sign}\left(Y_i^c < O_i^c\right)}{N}$

平均下降指标 $\sum_{i=1}^{N} \frac{\text{Max}\left(0, Y_i^c - O_i^c\right)}{Y_i^c}$



该方法生成的显著性图相较于其他方法：
- 捕捉到的特征与输出分数关系更紧密（平均上升最高）
- 减少了额外的噪声引入（平均下降最低）

方法	平均下降 (%)	平均上升 (%)
Mask	86.9	54.9
GradCAM	76.7	56.7
GradCAM++	86.3	47.7
FAFRP(Ours)	**67.9**	**62.3**

输出为原始分数

方法	平均下降 (%)	平均上升 (%)
Mask	80.3	38.4
GradCAM	68.3	35.1
GradCAM++	75.2	32.7
FAFRP(Ours)	**67.5**	**48.1**

输出为Softmax分数

RINC Robotic Intelligence & Neural Computing Group

2.2 基于前向传播的特征归因可解释算法：对比实验

基于能量的指向游戏：

$$S_{energy} = \frac{\sum V^c_{(i,j)\in bbox}}{\sum V^c_{(i,j)\in bbox} + \sum V^c_{(i,j)\notin bbox}}$$

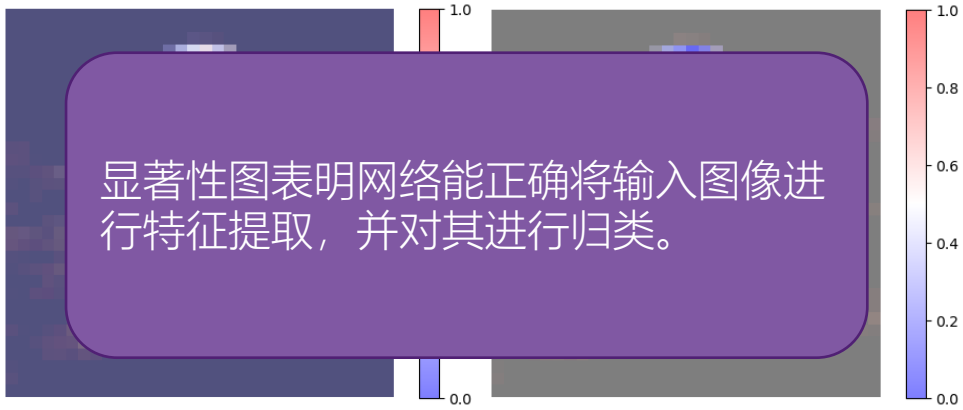
该方法生成的显著性图相较于其他方法：

- 对目标物体的定位能力更强（框内能量更高）
- 产生的噪声更低（框外能量更低）

	Mask	GradCAM	GradCAM++	FAFRP(Ours)
$S_{energy}(\%)$	56.1	48.1	49.3	**57.2**

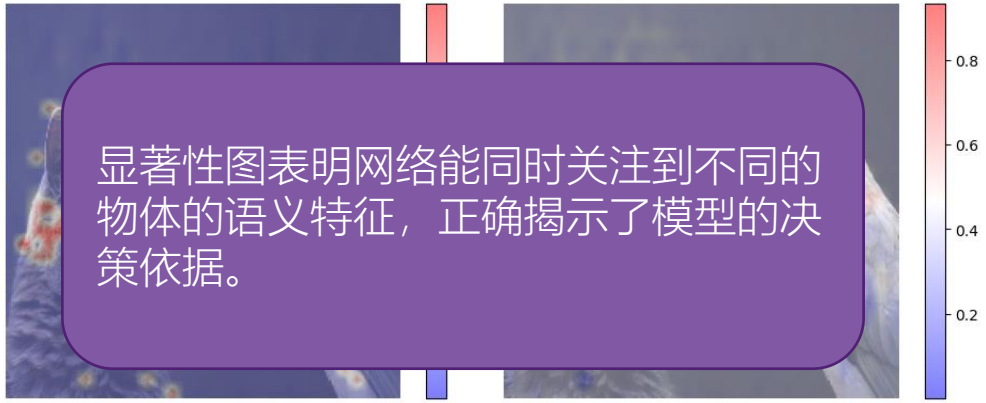
南京大学 NANJING UNIVERSITY

2.2 基于前向传播的特征归因可解释算法：可视化实验



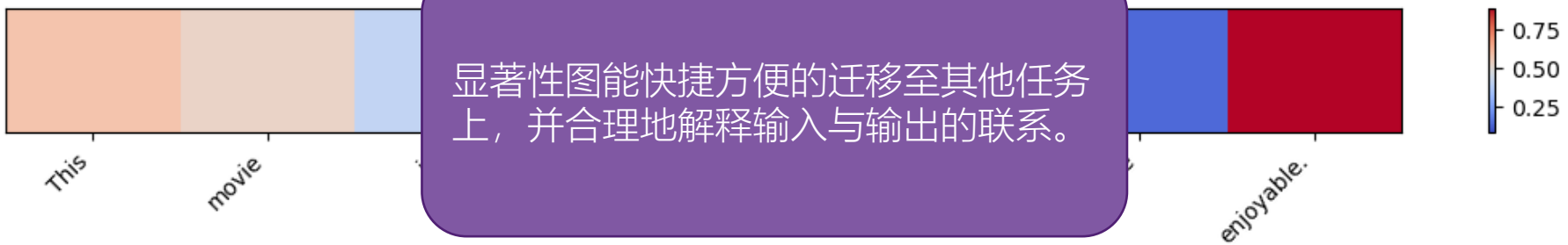
显著性图表明网络能正确将输入图像进行特征提取，并对其进行归类。

输入图像在不同输出分数上的显著性图



显著性图表明网络能同时关注到不同的物体的语义特征，正确揭示了模型的决策依据。

含有不同物体的输入图像在不同输出分数上的显著性图



显著性图能快捷方便的迁移至其他任务上，并合理地解释输入与输出的联系。

在情感分类任务上的显著性图可视化

RINC Robotic Intelligence & Neural Computing Group

RINC
Robotic Intelligence & Neural Computing Group

第三部分

实际应用

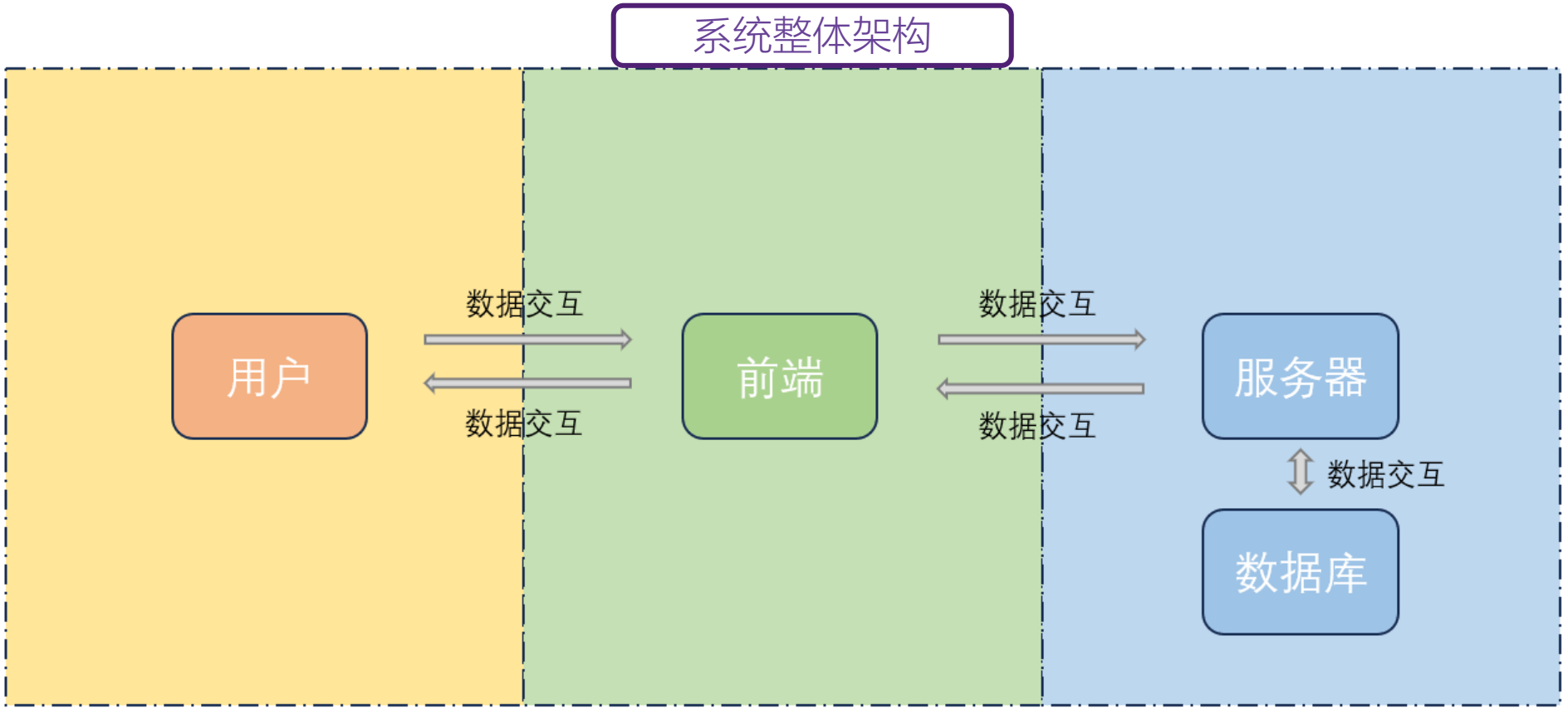
Applications

神经网络可解释性系统

3.1 系统设计

系统需求：

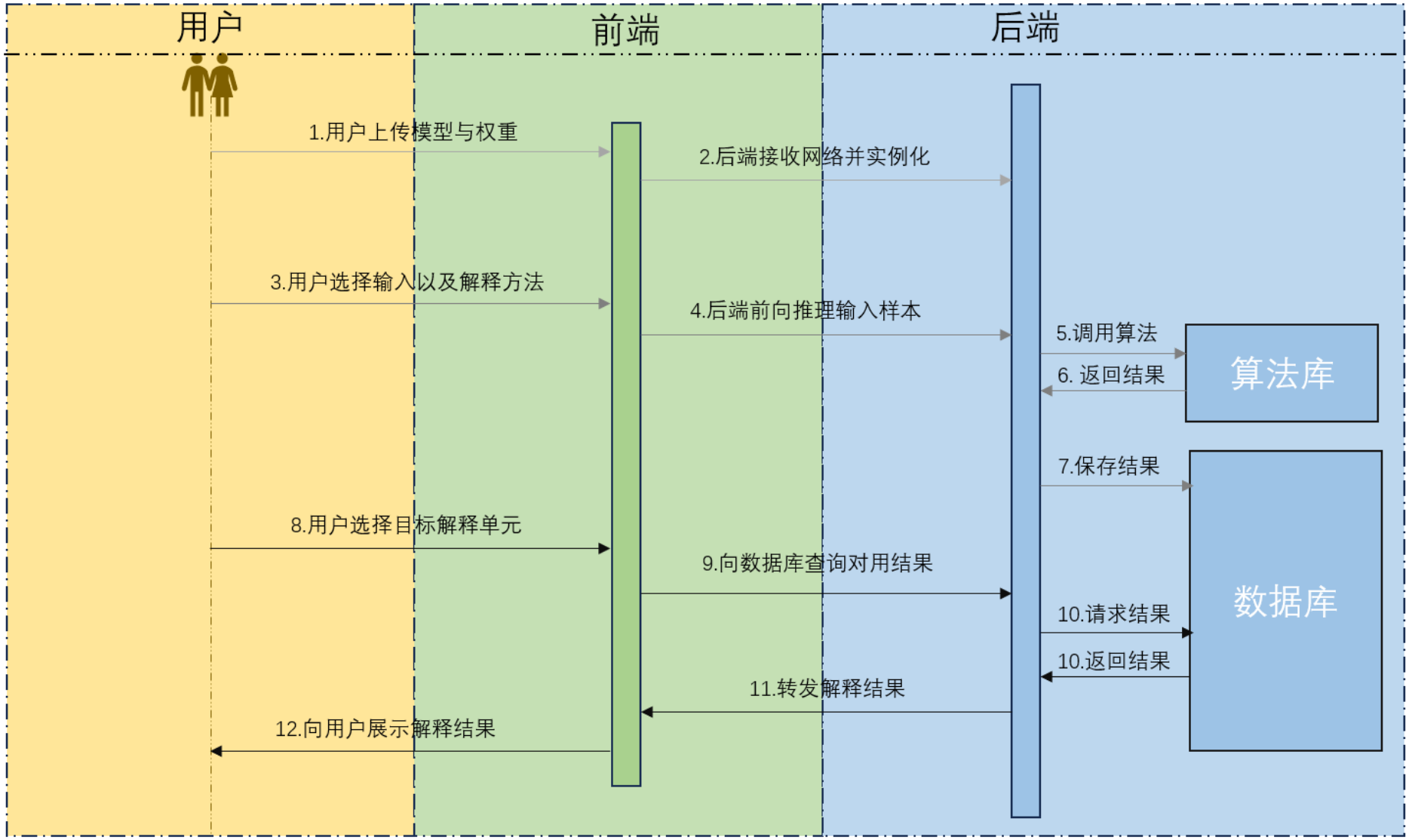
- 用户上传网络
- 用户上传数据
- 用户选择可解释方法
- 解析网络
- 用户查看可解释结果



南京大学 NANJING UNIVERSITY

3.1 系统设计

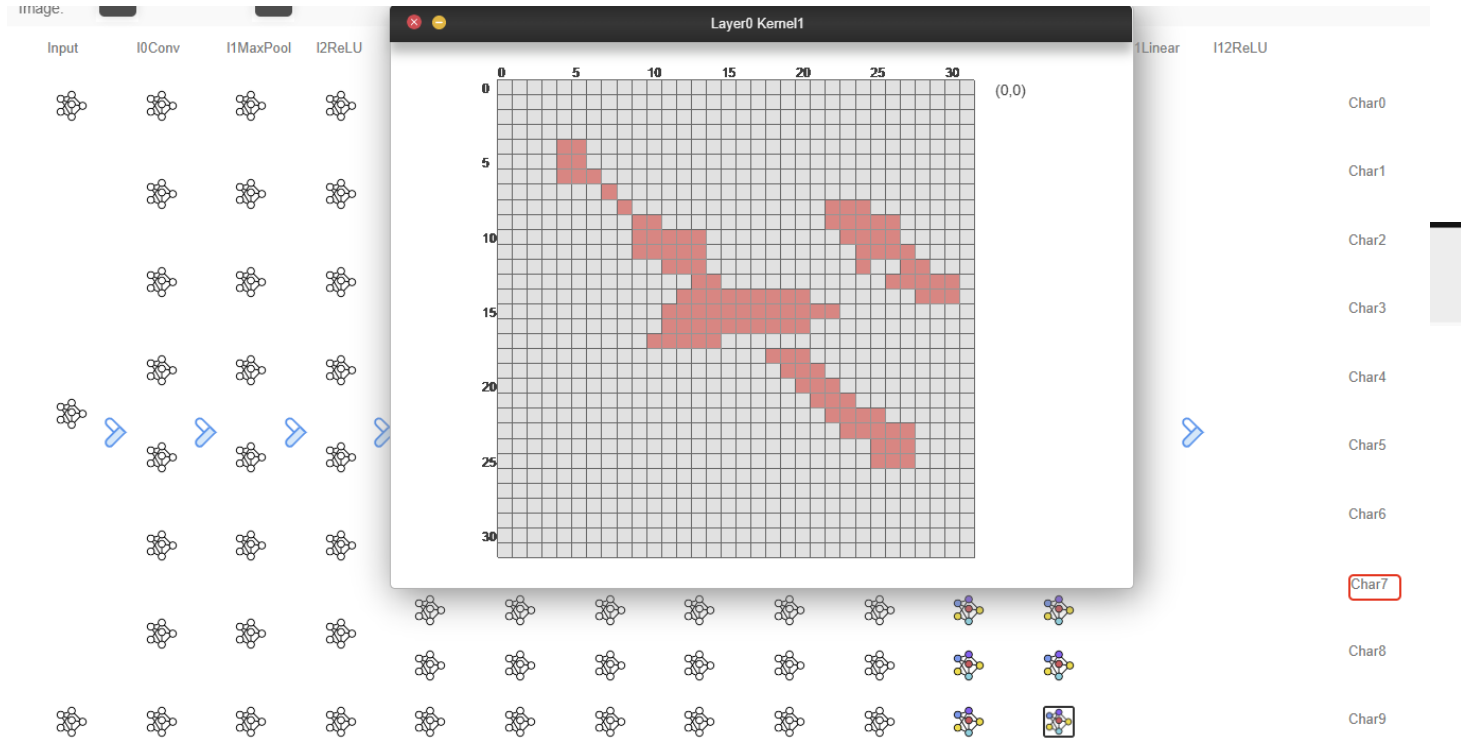
系统运行流程



南京大学 NANJING UNIVERSITY

3.2 系统展示

查看网络每层输出特征向量

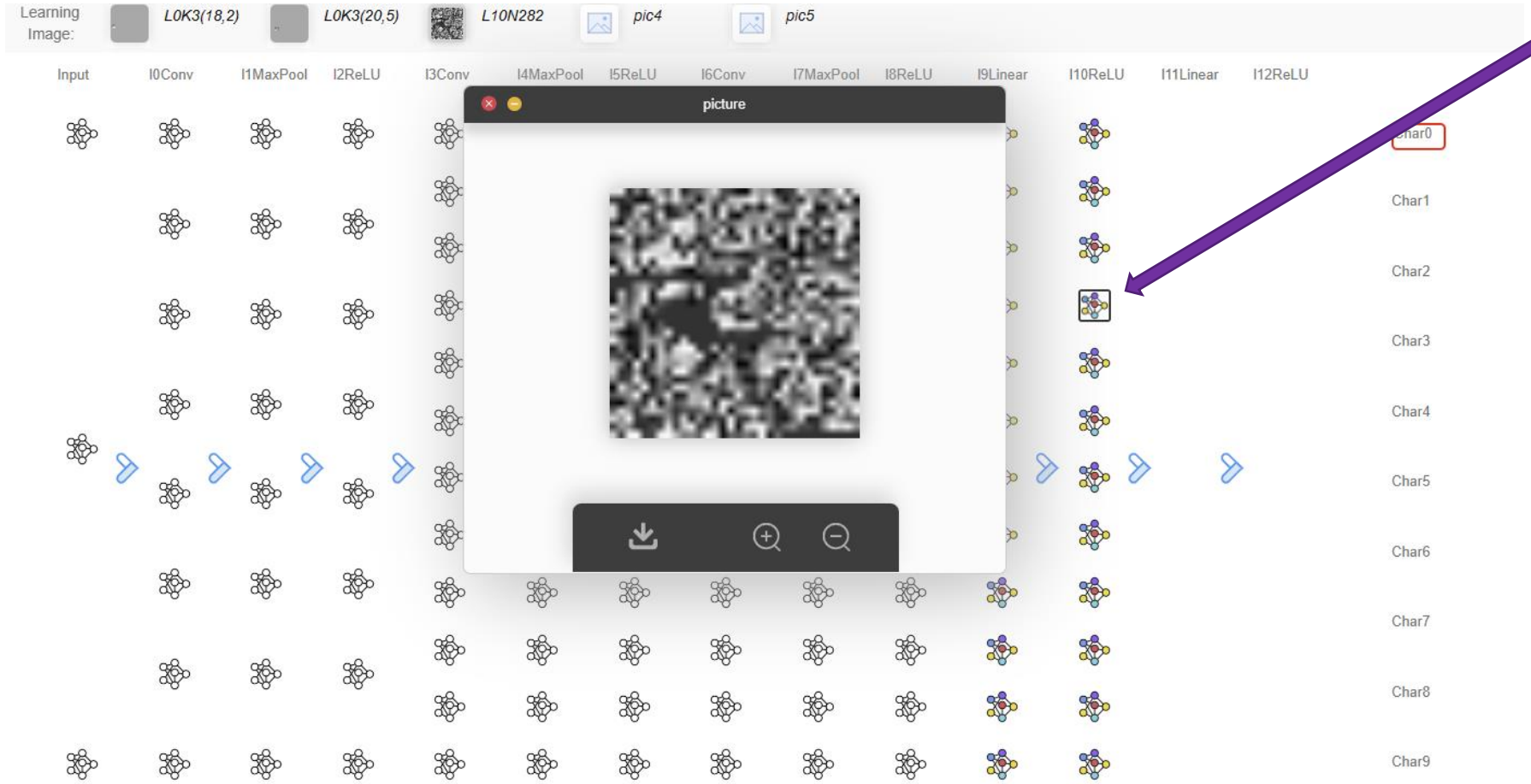


南京大学 NANJING UNIVERSITY

3.2 系统展示

解释结果查看

点击对应的解释目标单元，即可得到对应的特征可视化图像



RINC
Robotic Intelligence & Neural Computing Group

第四部分

研究生期间工作成果

Work Product

相关成果列举

论文

- 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[J].软件学报,2022

项目

- 科技部重大项目 “基于神经可塑性的脉冲神经网络高效学习机制与类脑智能系统”（参与课题年限2021年9月——2024年6月），负责神经网络模型相关研究。
- 国家电网项目 “基于多维巡检影像匹配和对比技术的变电设备缺陷分析技术研究”（参与课题年限2021年9月——2022年12月），负责图像对比与目标检测相关研究。

第五部分

总结

Summary

全文总结

基于前向传播的特征可视化可解释算法

- 基于前向传播设计相关性分数传播方法，适用于**多种神经网络结构**；
- 在通道级别聚合解释结果，得到对应的**语义信息**；

基于前向传播的特征归因可解释算法

- 基于前向传播设计特征归因算法，得到输入样本对结果的**显著性图**。
- 得到的显著性图具有更好的**解释效果**，以及**物体定位能力**。

神经网络可解释性系统

- 将算法应用于实际任务中
- 提供了上传模型、上传数据、解释算法以及查看结果等功能，构建了一个**成熟并且完整**的系统

南京大学 NANJING UNIVERSITY

RINC Robotic Intelligence & Neural Computing Group

感谢各位老师批评指正

答辩人：张凌茗 MG21330071

导师：申富饶 教授

日期：2024年5月15日

誠樸雄偉 勵學敦行