SYSU 2023
PRESENTATION SLIDE

中山大學
SUN YAT-SEN UNIVERSITY

中国科学院信息工程研究
INSTITUTE OF INFORMATION ENGINEERING

中山大學

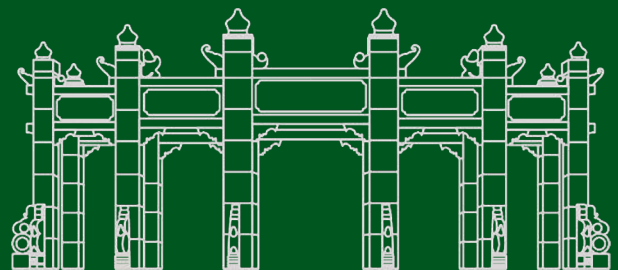
可解释人工智能

Explainable Artificial Intelligence

讲者：陈若愚 (chenruoyu@iie.ac.cn)

导师：操晓春 教授

院系：网络空间安全学院

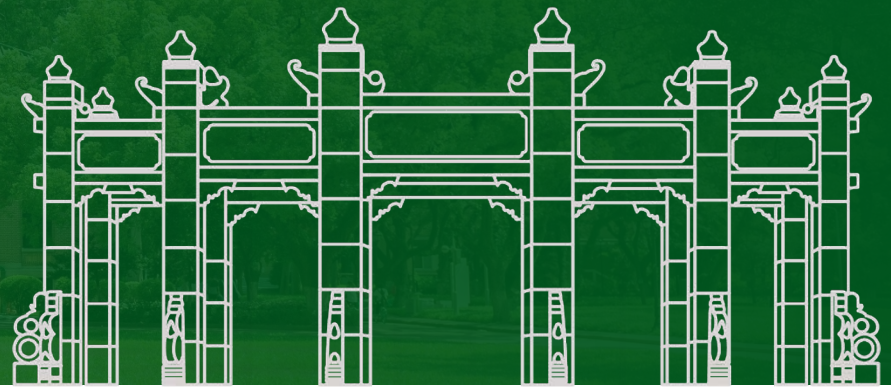


中山大學
SUN YAT-SEN UNIVERSITY

中国科学院信息工程研究
INSTITUTE OF INFORMATION ENGINEERING

目录 Contents

在此输入内容摘要或关键词，不宜过长，若无内容也可删除
也可添加适当辅助信息 建议不要超过50字



SYSU 2023
PRESENTATION SLIDE

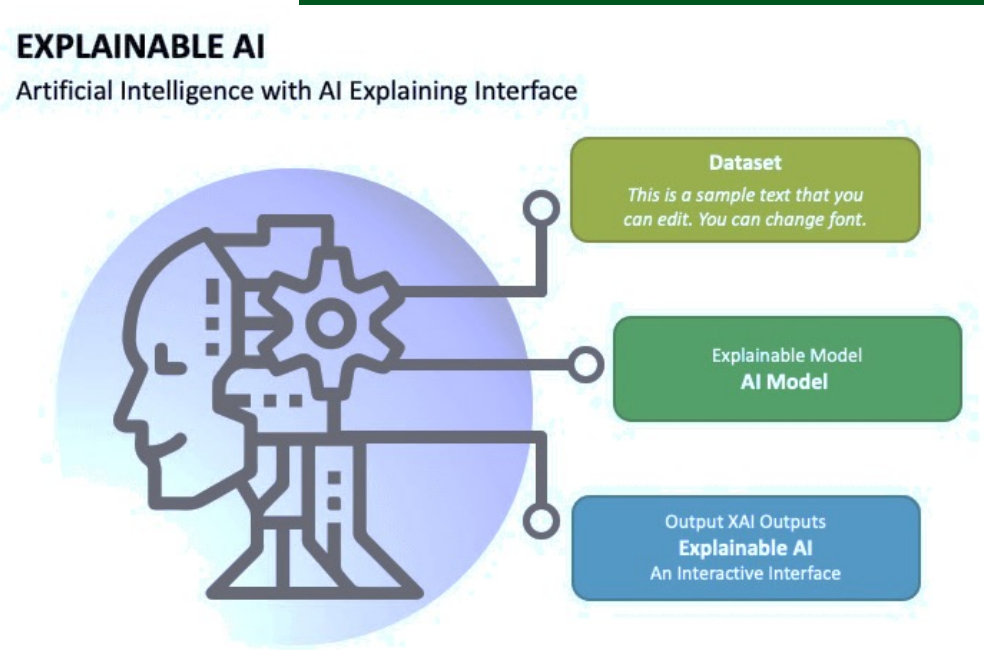
中山大學
SUN YAT-SEN UNIVERSITY

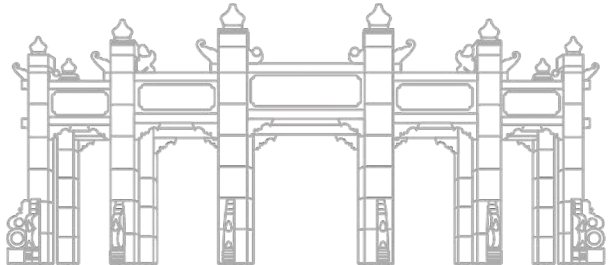
中国科学院信息工程研究
INSTITUTE OF INFORMATION ENGINEERING

01

可解释基本概述

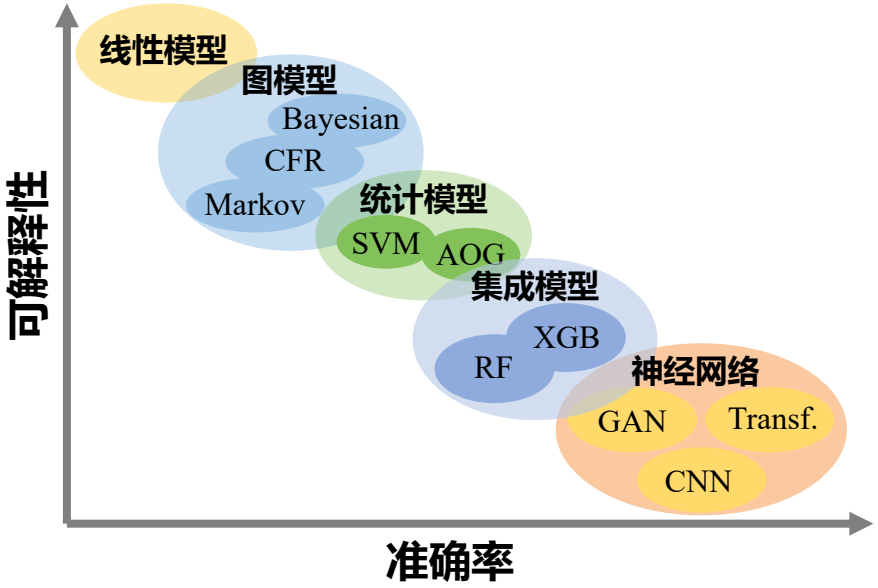
Basic Overview of Interpretation

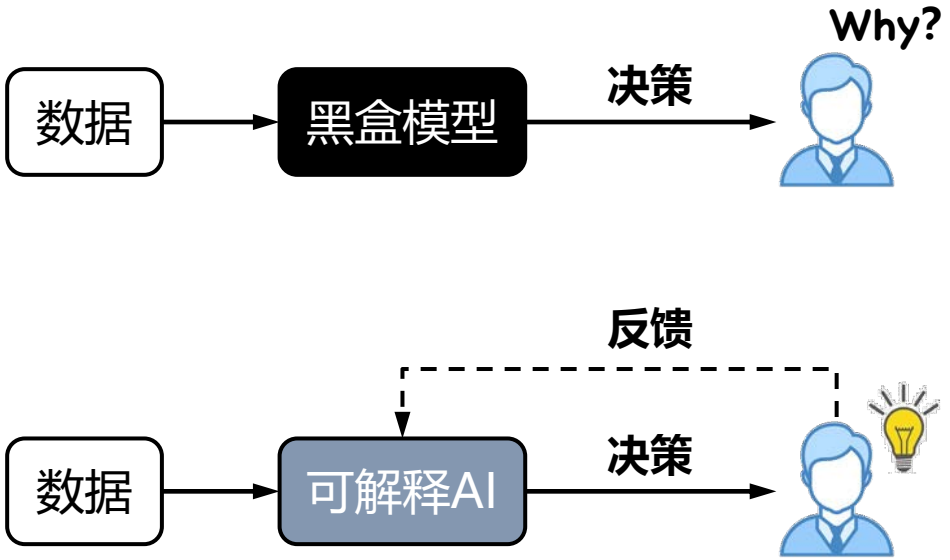




我们为什么需要可解释?

Why We Need XAI?





AI黑盒模型存在**决策不合理**，**不合法**或者**无详细解释**的风险。

可解释的AI有利于人类**理解模型决策**，更加**信任模型**，并根据不断**反馈提升AI模型**。

ML的巨大成功使AI的能力爆炸式增长，但其有效性将受到机器**无法向人类用户解释其决策和行动**的限制。**XAI**对于用户理解、适当信任和有效管理**新一代人工智能**至关重要。

自动驾驶

教育

金融风险

医疗健康

国家需求与国际前沿

National Needs and International Frontiers

国家需求

《AI框架发展白皮书》

- AI 框架需要具备三个层面的能力支 持可解释人工智能。
- 可解释性的需求增加对AI 框架提出进阶性要求。

《可解释、可通用的下一代人工智能方法重大研究计划2023年度项目指南》

- 数据驱动与知识驱动融合的人工智能方法。
- 深度学习的表示理论和泛化理论来指导深度学习模型和算法设计。
- 可通用的专业领域人机交互方法。

可解释、可通用的下一代人工智能方法重大研究计划面向人工智能发展**国家重大战略需求**

国际前沿

DARPA 美国

可解释的人工智能 (XAI) 计划：

- 产生更可解释的模型，同时保持高水平的学习性能（预测准确性）；
- 使人类用户能够理解、适当信任并有效管理新一代人工智能合作伙伴。

欧盟

基于以用户为中心的XAI设计方法：

- 对人工智能提供商的风险分析和管理系统的本地可解释性提供具体指导；
- 以用户为中心的设计方法，作为衡量给定解释对于其受众和目的是否“有意义”的方法。

可解释AI相关术语

Related Terms for XAI

可解释AI

Interpretation

- 模型背后实际的**运行机理**；
- 准确将模型的原因与结果联系起来；
- 确定模型实际学习了什么；
- 在一定条件下是正确的。

Explanation

- 以**人类可理解**的方式表示决策过程或者结果；
- 关联各种反馈的模态，以及控制语义表达程度；
- 不一定是正确的。

Ante-hoc (拉丁语)

- 直接解释**白盒模型**；
- 在模型的决策过程中已产生可解释。

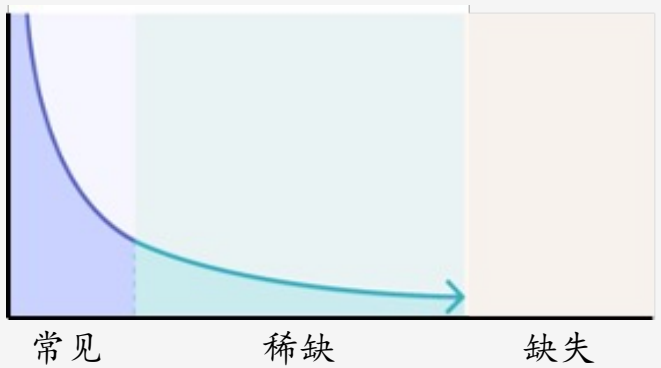
Post-hoc (拉丁语)

- 解释一个预训练模型或其决策的结果；
- 在模型做完决策后提供的解释。

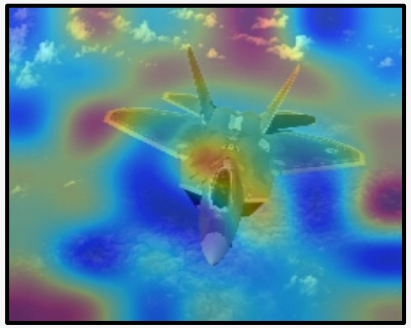
现实挑战

Realistic Challenges

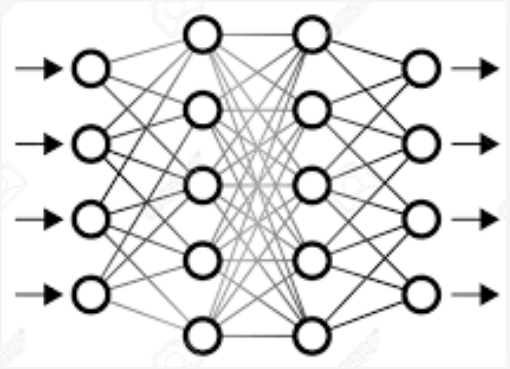
为什么AI模型仍存在错误？



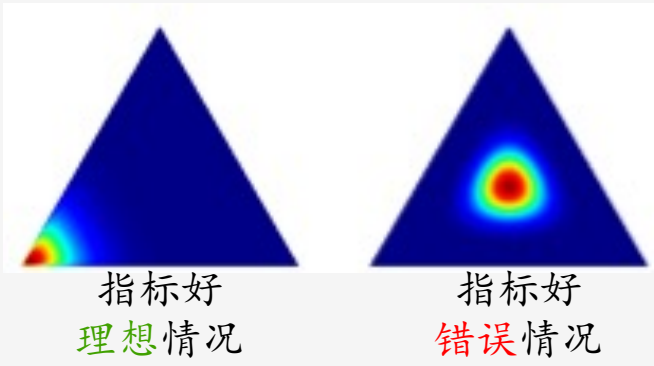
数据分布不全面



监督信息少



模型自身的缺陷

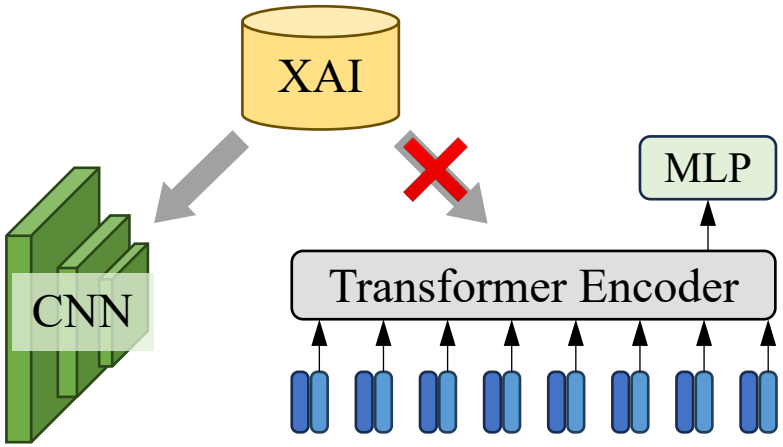


评价指标缺陷

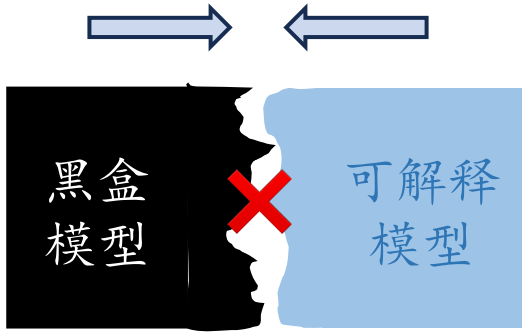
所以我们需要可解释性！

可解释AI涉及部分科学问题

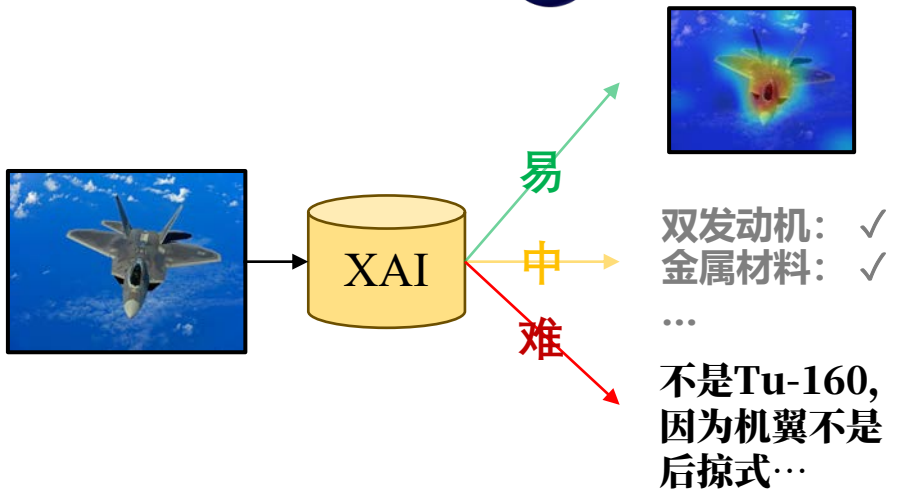
Scientific issues



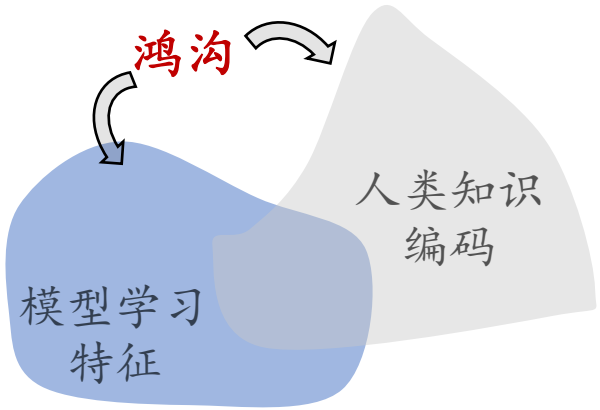
1. 解释范式不通用



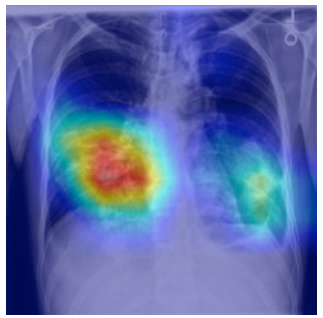
2. 可解释模型难设计



3. 高程度语义反馈难解释

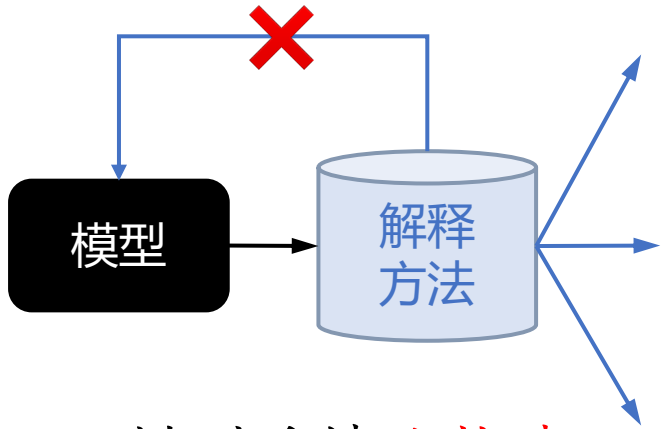


4. 人类知识难融合



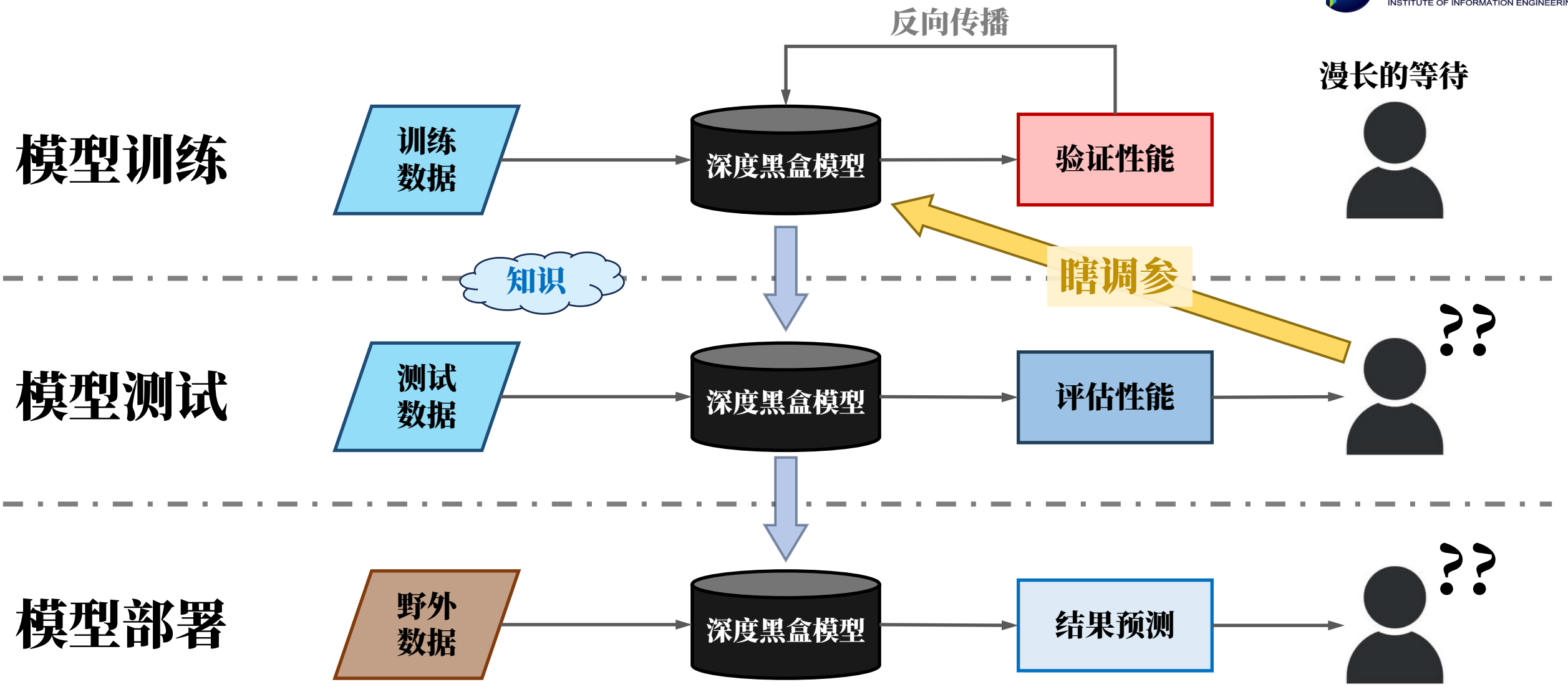
？？

5. 解释结果难评估

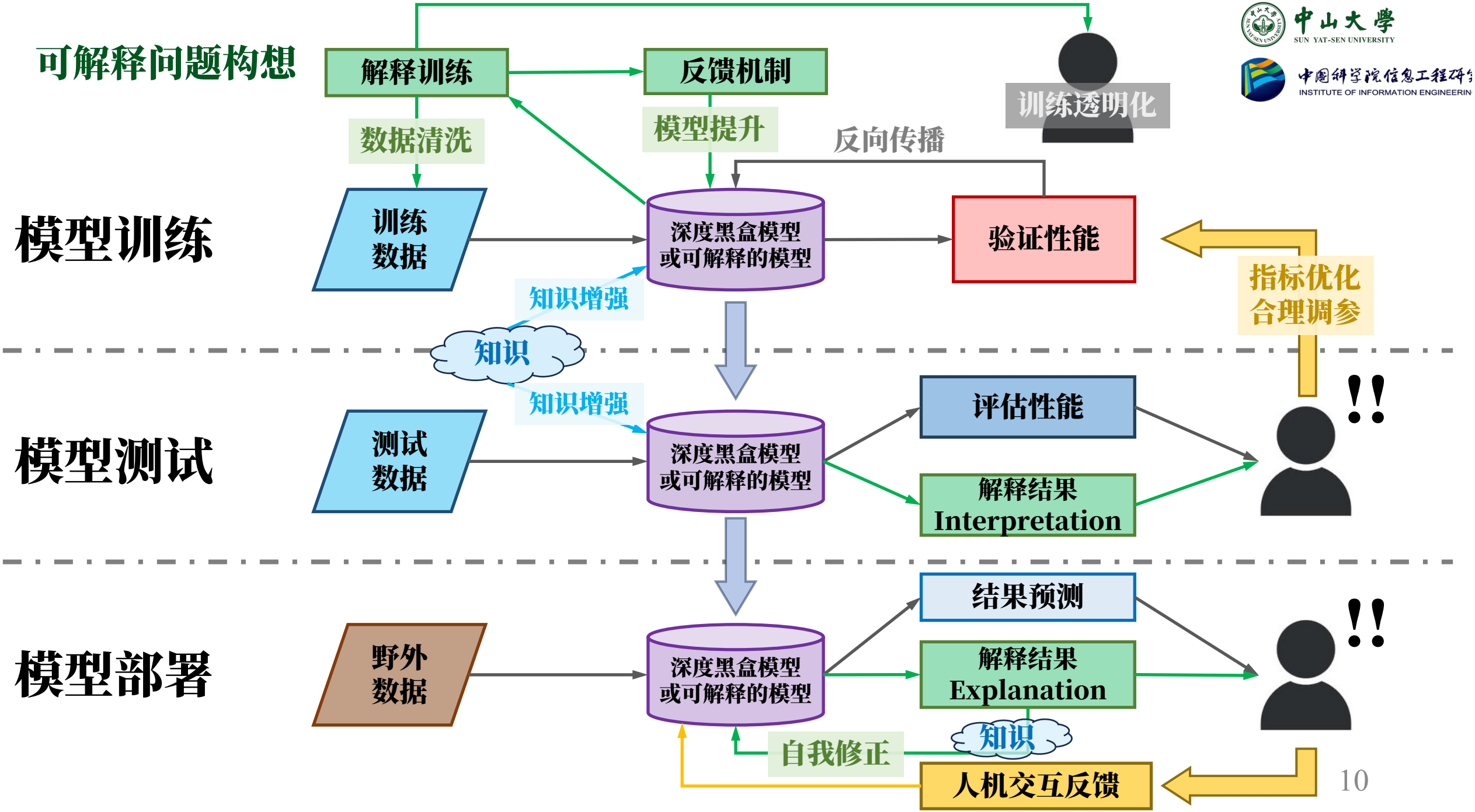


6. 模型反馈难构建

可解释问题构想


反向传播
漫长的等待
模型训练
训练数据
深度黑盒模型
验证性能
知识
瞎调参
模型测试
测试数据
深度黑盒模型
评估性能
??
模型部署
野外数据
深度黑盒模型
结果预测
??

可解释问题构想


解释训练
反馈机制
训练透明化
数据清洗
模型提升
反向传播
模型训练
训练
数据
深度黑盒模型
或可解释的模型
验证性能
指标优化
合理调参
知识增强
知识
知识增强
模型测试
测试
数据
深度黑盒模型
或可解释的模型
评估性能
解释结果
Interpretation
!!
模型部署
野外
数据
深度黑盒模型
或可解释的模型
结果预测
解释结果
Explanation
!!
知识
自我修正
人机交互反馈

SYSU 2023
PRESENTATION SLIDE

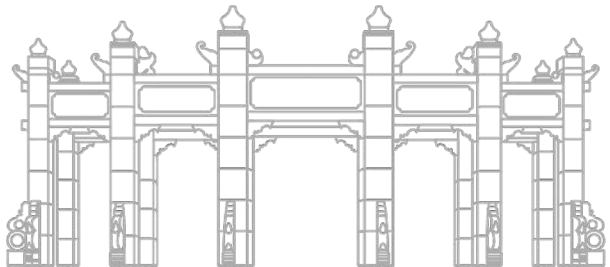
中山大學
SUN YAT-SEN UNIVERSITY

中国科学院信息工程研究
INSTITUTE OF INFORMATION ENGINEERING

02

基于归因的可解释

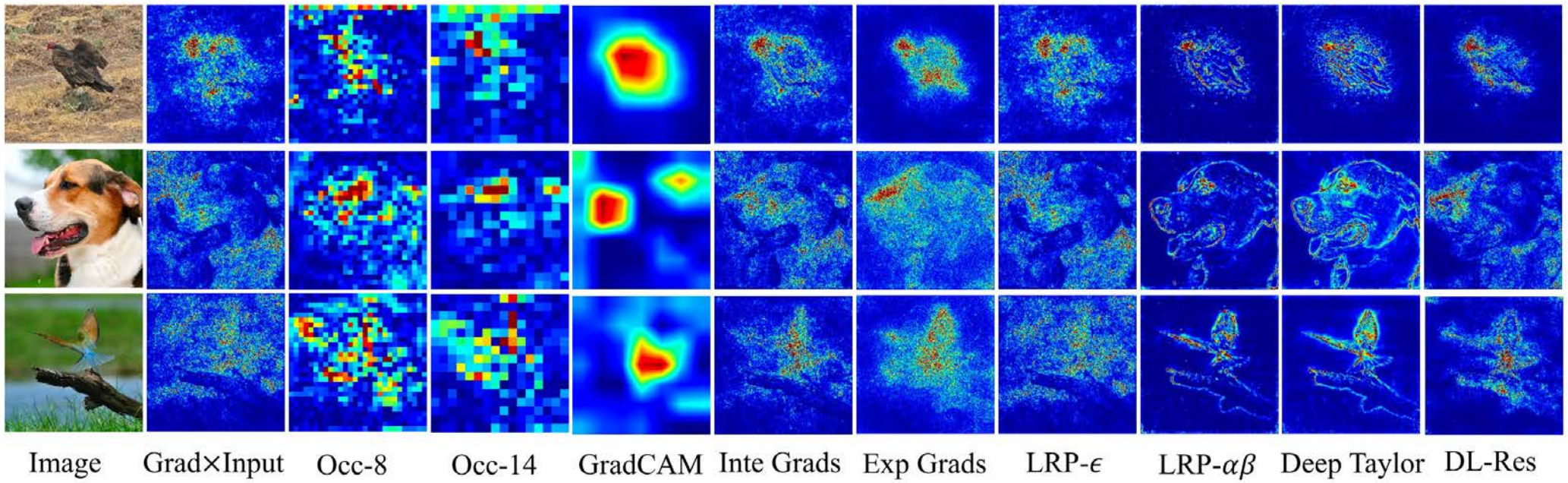
Attribution-based interpretable algorithms



什么是归因?

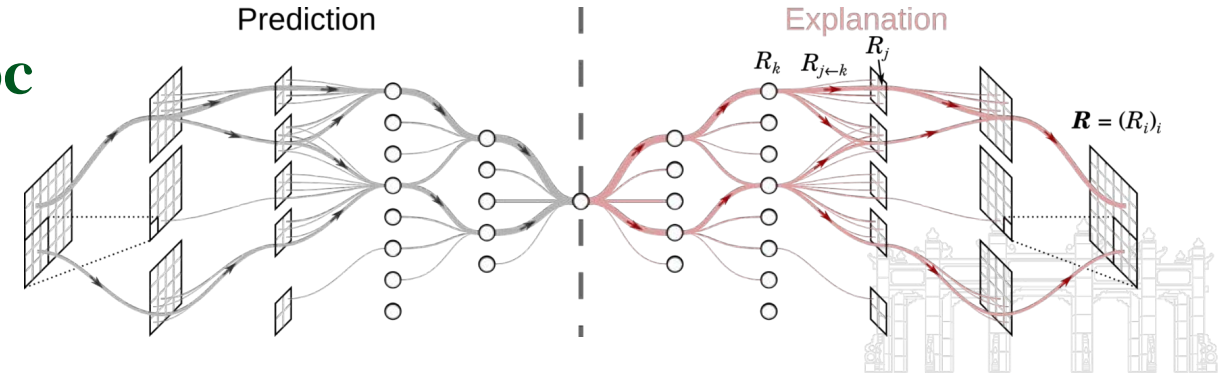
What's attribution

基于归因的方法

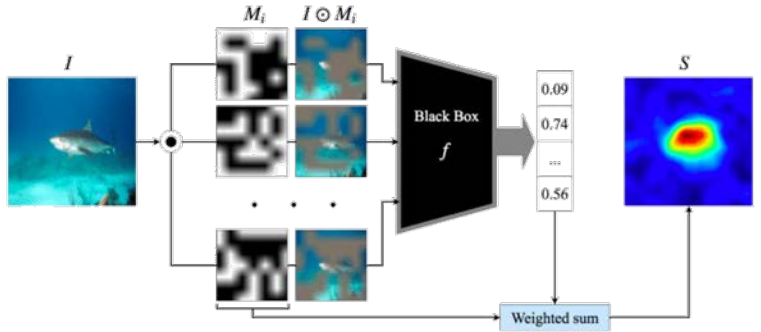


Post-hoc

基于模型内部机理（白盒）



基于扰动（黑盒）



类激活图

Class Activation Map

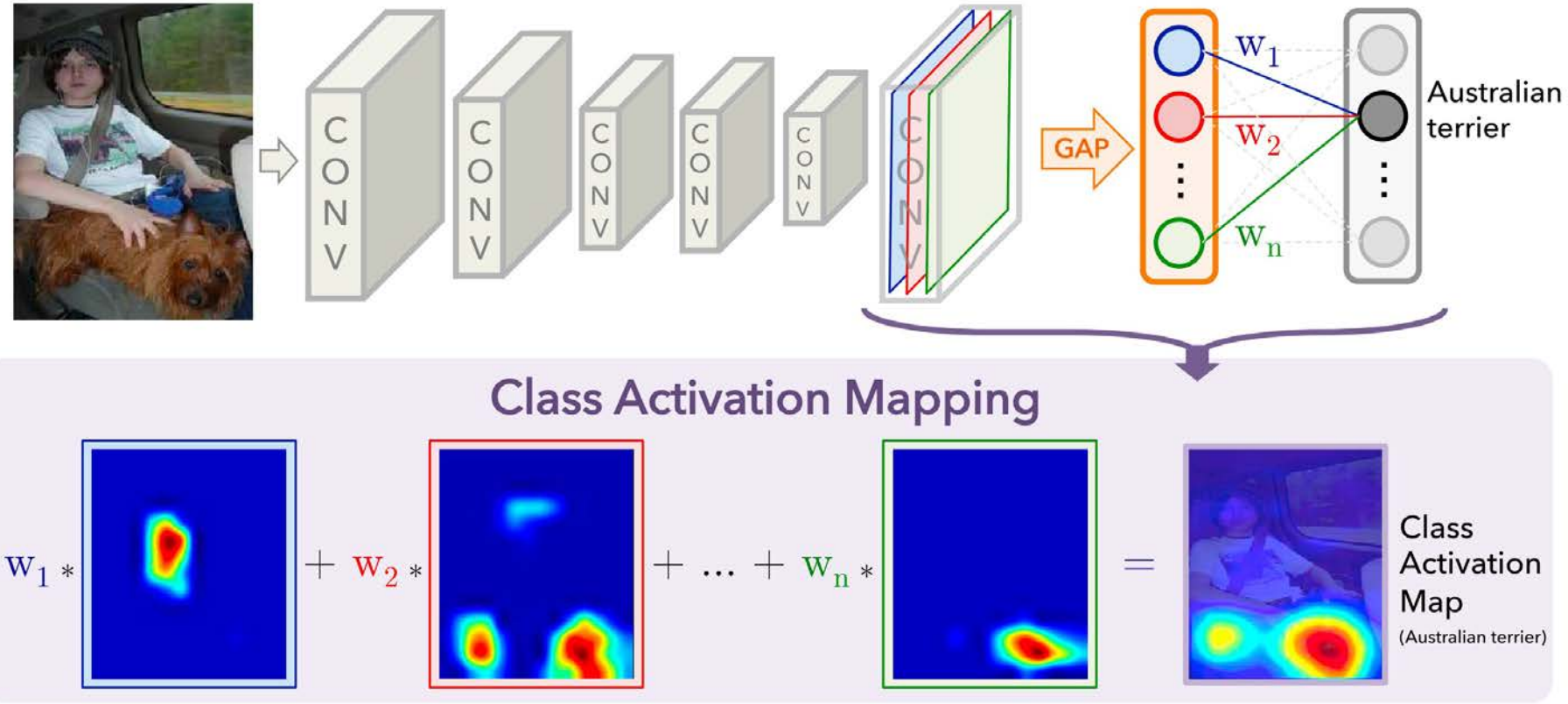


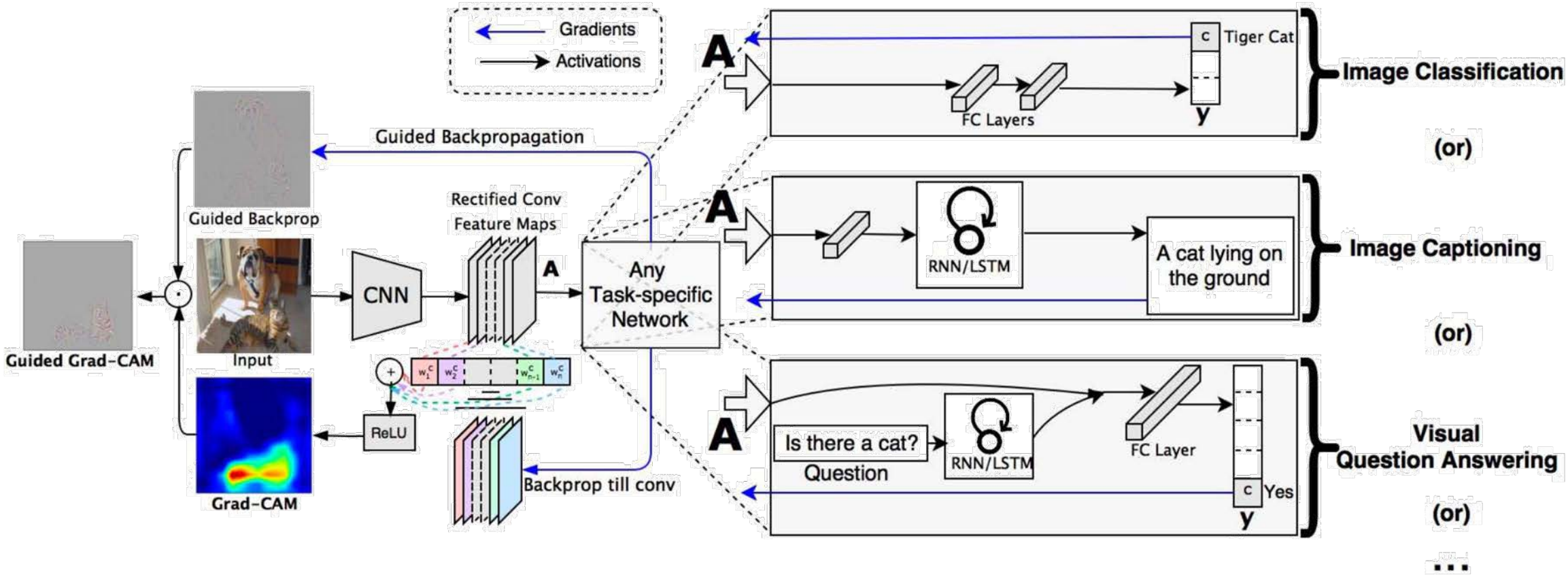
Figure 2. Class Activation Mapping: the predicted class score is mapped back to the previous convolutional layer to generate the class activation maps (CAMs). The CAM highlights the class-specific discriminative regions.

$$M_c(x,y) = \sum_k w_k^c f_k(x,y)$$

Zhou, Bolei, et al. "Learning deep features for discriminative localization." *CVPR*. 2016.

Grad-CAM

Gradient-based Localization

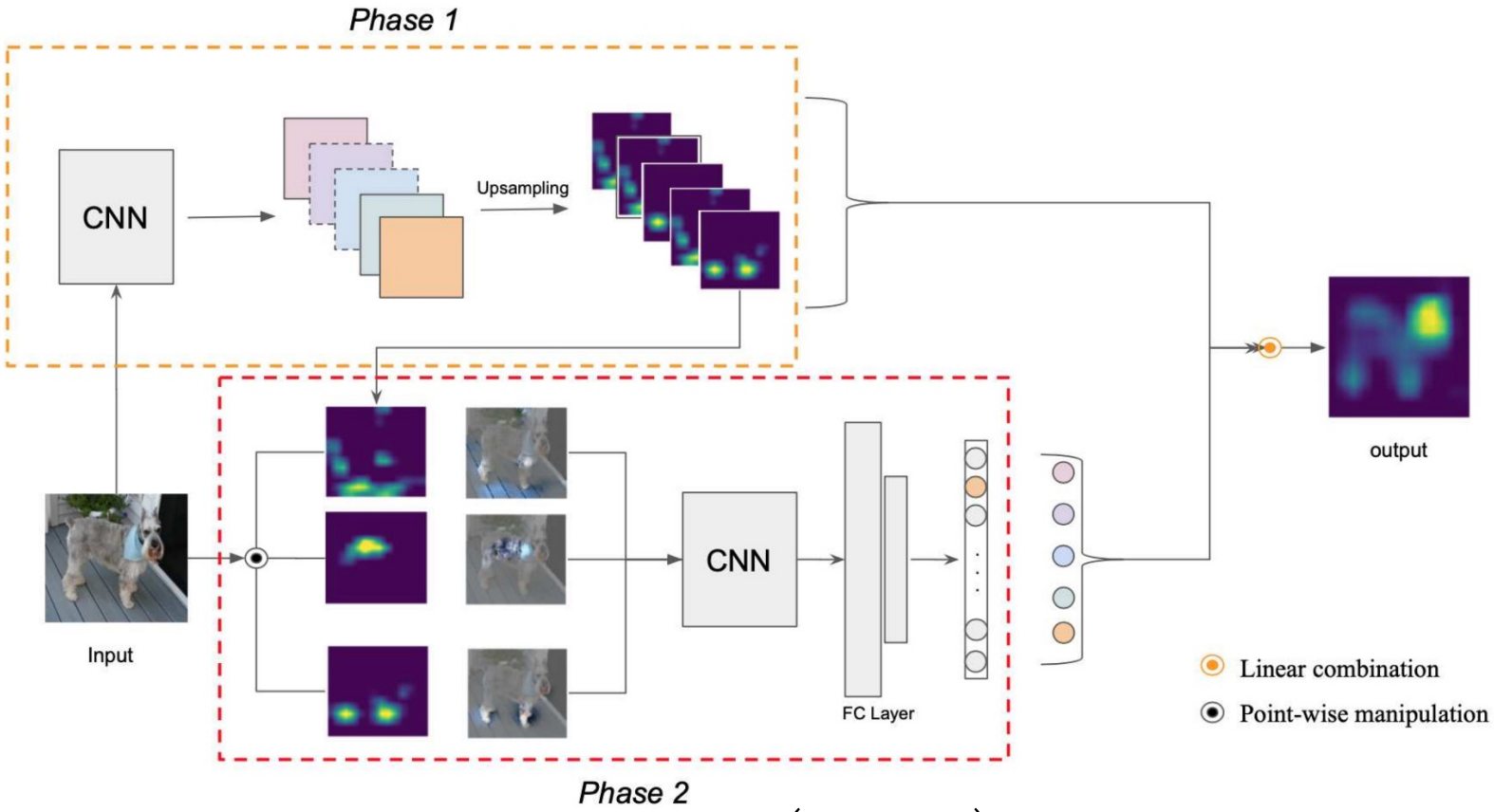


$$\boldsymbol{g} = \max\left(0, \sum_k \alpha_k^c \boldsymbol{A}^k\right)$$

Selvaraju, Ramprasaath R., et al. "Grad-CAM: Visual explanations from deep networks via gradient-based localization." *ICCV*. 2017.

Score-CAM

Score-weighted visual explanations



$$L^c_{Score} = ReLU\left(\sum_k \alpha^c_k A^k_l\right)$$

Wang, Haofan, et al. "Score-CAM: Score-weighted visual explanations for convolutional neural networks." *WACV*. 2020.

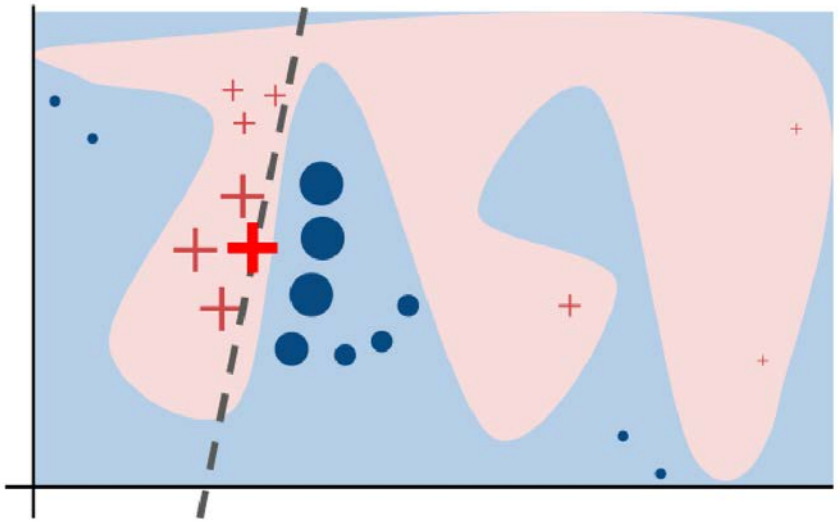
CAM类推荐阅读论文

Gradient-based Localization

1. Chattopadhay, Aditya, et al. "**Grad-CAM++**: Generalized gradient-based visual explanations for deep convolutional networks." *WACV*, 2018.
2. Fu, Ruigang, et al. "**Axiom-based grad-cam**: Towards accurate visualization and explanation of cnns." *WACV*, 2018.
3. Omeiza, Daniel, et al. "**Smooth Grad-CAM++**: An enhanced inference level visualization technique for deep convolutional neural network models." *arXiv preprint arXiv:1908.01224* (2019).
4. Wang, Haofan, et al. "**SS-CAM**: Smoothed Score-CAM for sharper visual feature localization." *arXiv preprint arXiv:2006.14255* (2020).
5. Muhammad, Mohammed Bany, and Mohammed Yeasin. "**Eigen-CAM**: Class activation map using principal components." *IJCNN*, 2020.
6. Ramaswamy, Harish Guruprasad. "**Ablation-CAM**: Visual explanations for deep convolutional network via gradient-free localization." *WACV*. 2020.
7. Zhang, Qinglong, Lu Rao, and Yubin Yang. "**Group-CAM**: Group score-weighted visual explanations for deep convolutional networks." *arXiv preprint arXiv:2103.13859* (2021).
8. Jiang, Peng-Tao, et al. "**LayerCAM**: Exploring hierarchical class activation maps for localization." *IEEE Transactions on Image Processing* 30 (2021): 5875-5888.

LIME

Black-Box Explanation Method



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

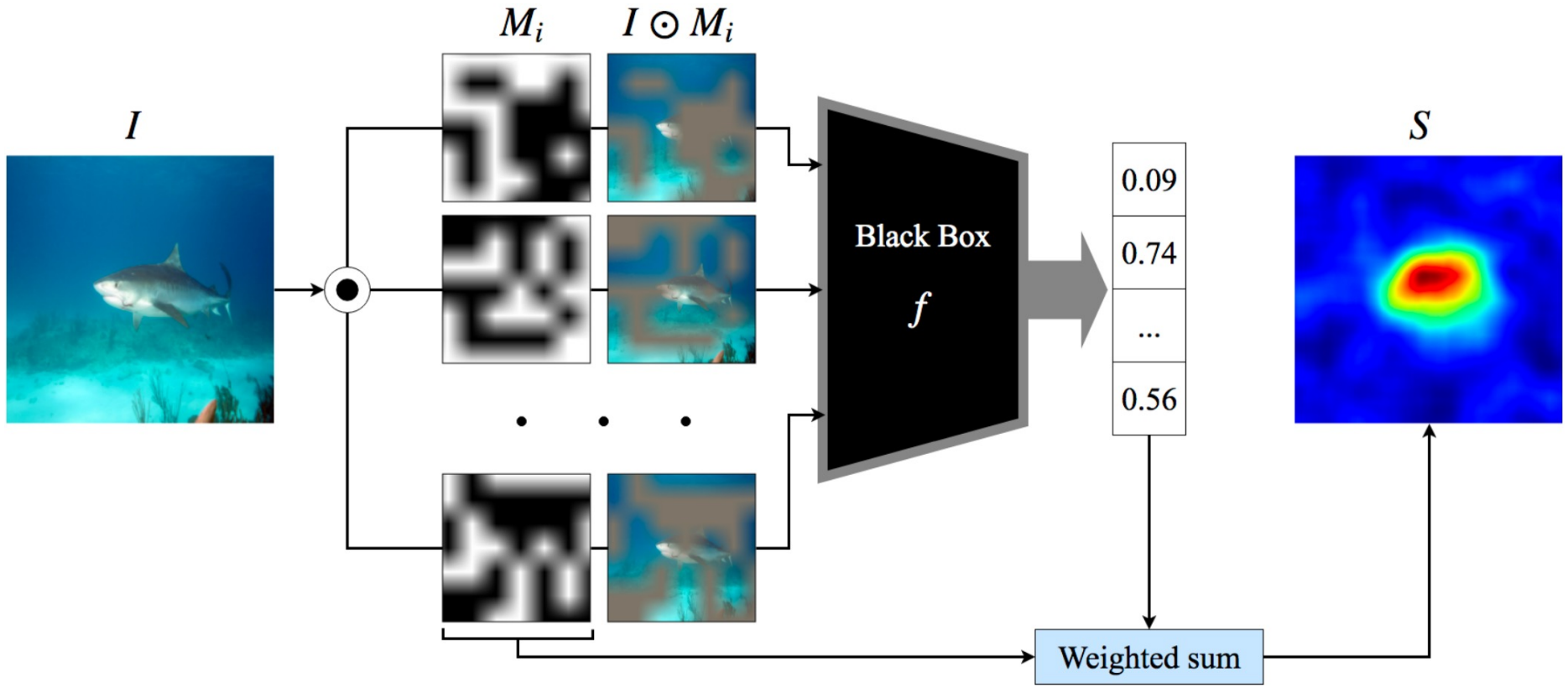
Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

$$\xi = \arg\min_{g \in \mathcal{G}} \; L(f, g, \pi_{x'}) + \Omega(g).$$

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. ""Why should i trust you?" Explaining the predictions of any classifier." *ACM SIGKDD*. 2016.

RISE

Black-Box Explanation Method

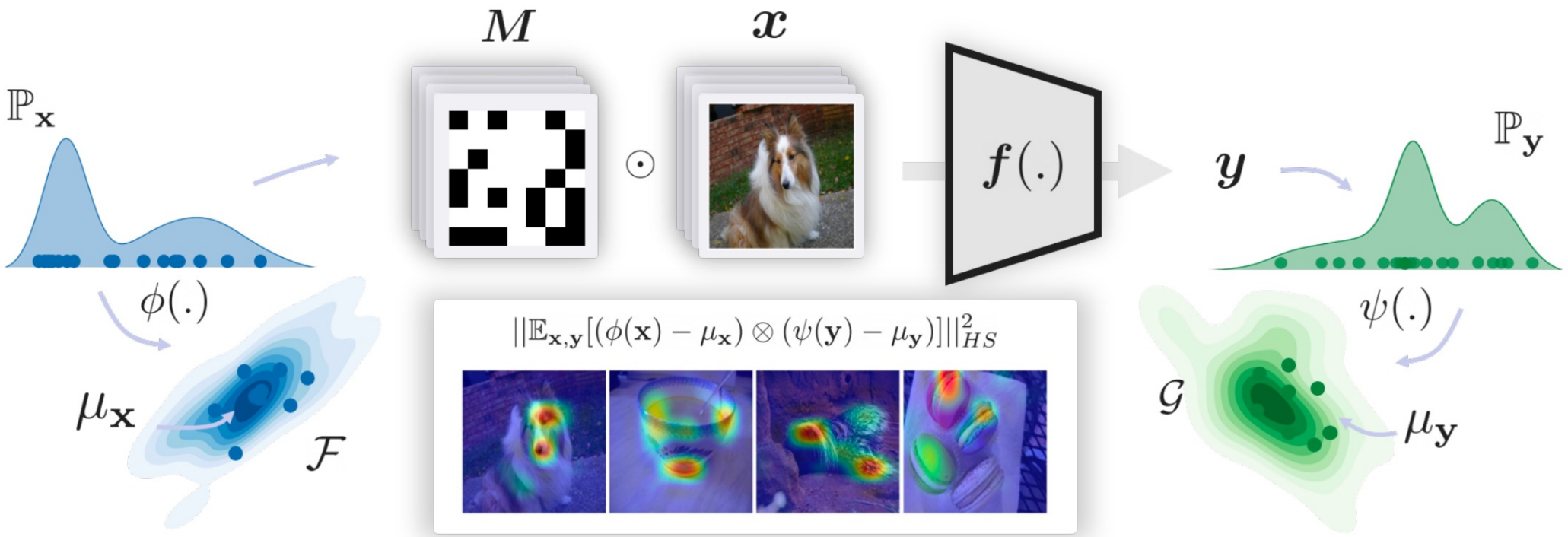


$$\boldsymbol{g}(x) = \frac{1}{\mathbb{E}(\mathcal{M})N}\sum_{i=0}^{N}\boldsymbol{f}(x \odot \boldsymbol{m}_i)\boldsymbol{m}_i$$

Petsiuk, Vitali, Abir Das, and Kate Saenko. "Rise: Randomized input sampling for explanation of black-box models." *BMVC*. 2018.

HSIC-Attribution

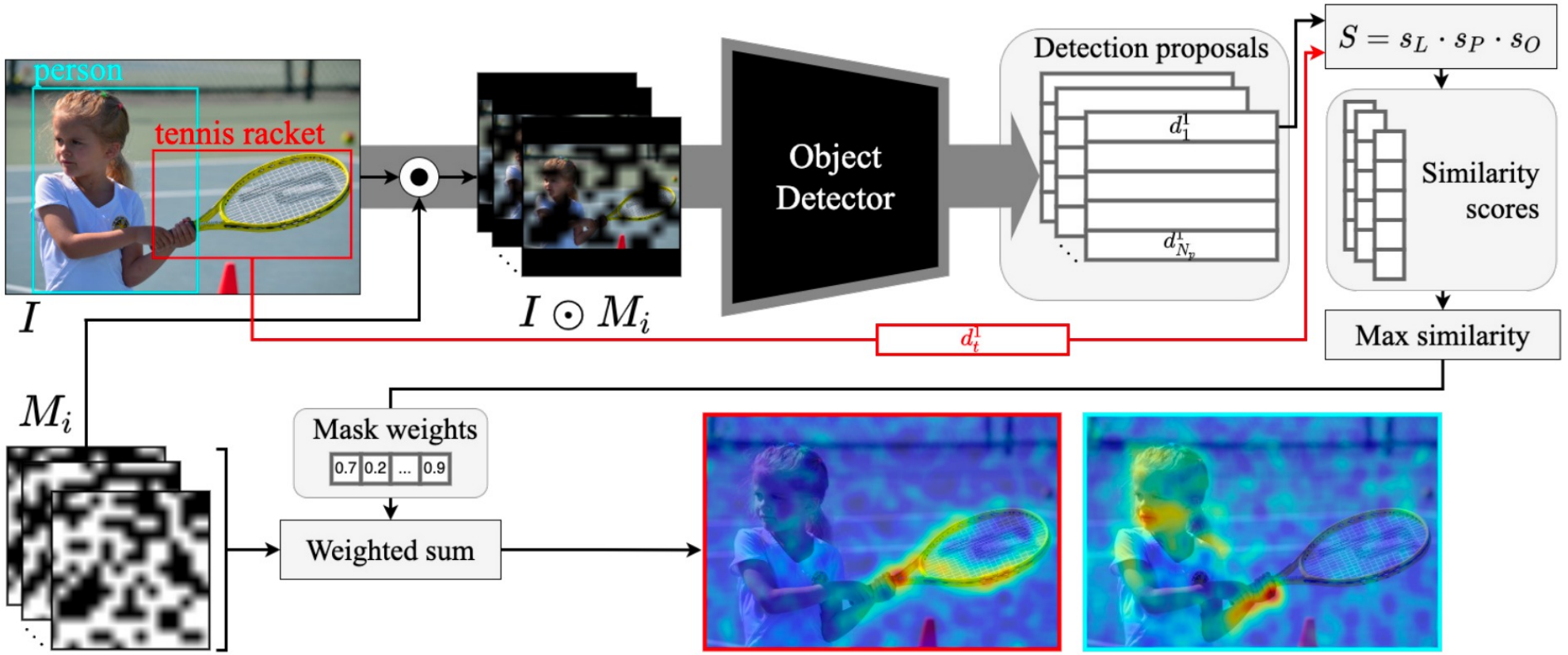
Black-Box Explanation Method



Novello, Paul, Thomas Fel, and David Vigouroux. "Making sense of dependence: Efficient black-box explanations using dependence measure." *NeurIPS* 35 (2022): 4344-4357.

D-RISE 目标检测的可解释

Object Detection Saliency Map



目标检测可解释复现：https://github.com/RuoyuChen10/objectdetection-saliency-maps
（请顺便Star一下该仓库）

Petsiuk, Vitali, et al. "Black-box explanation of object detectors via saliency maps." *CVPR*. 2021.

Explain Any Concept

Shapley Value with SAM

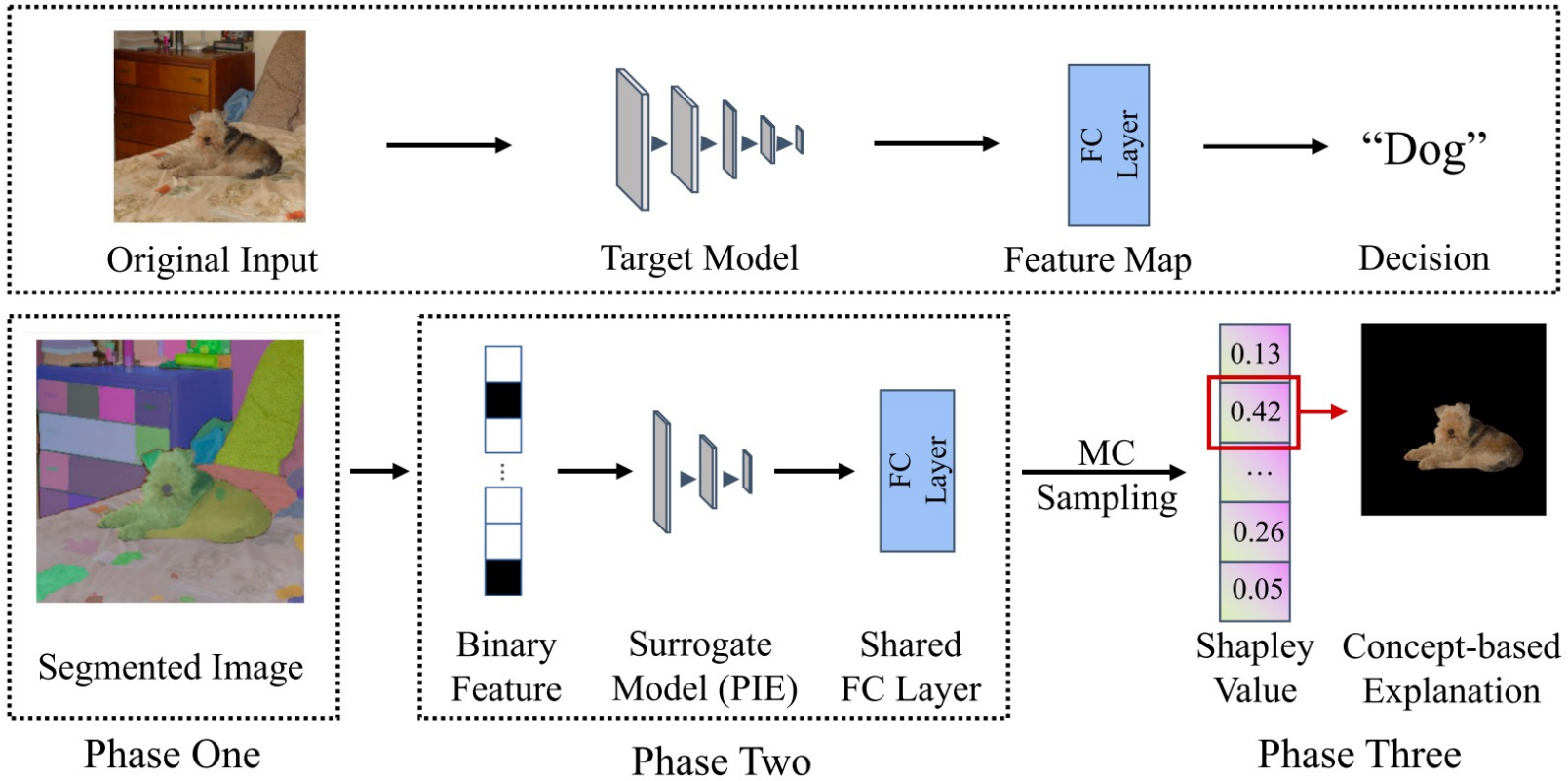


Figure 1: The technical pipeline of EAC in a three-phase form.

评价指标

Evaluation Metric

忠实度 (Faithfulness)

给定一个预测器f，一个可解释函数g，一个数据点x，一个子集大小$|S|$，定义g对$f(x)$的解释方程。则g的特征重要性得分应对应于f的x的重要特征。

复杂度

一个复杂的解释是使用所有d个特征来解释x的哪些特征对f是重要的。尽管这种解释可能忠实于模型(如上所定义)，但对于用户来说可能太难理解(特别是如果d很大)。

敏感度 (Sensitivity)

如果输入彼此接近并且它们的模型输出相似，那么它们的解释应该彼此接近。假设f是可微的，我们希望一个解释函数g在兴趣点x附近的区域具有低灵敏度，这意味着g的局部平滑。

可理解 (Understandability)

XAI过程必须提供一个清晰的、人类可以理解的解释，使用户能够轻松地掌握模型决策过程背后的推理。

Bhatt, Umang, Adrian Weller, and José MF Moura. "Evaluating and aggregating feature-based model explanations." *IJCAI*. 2021.

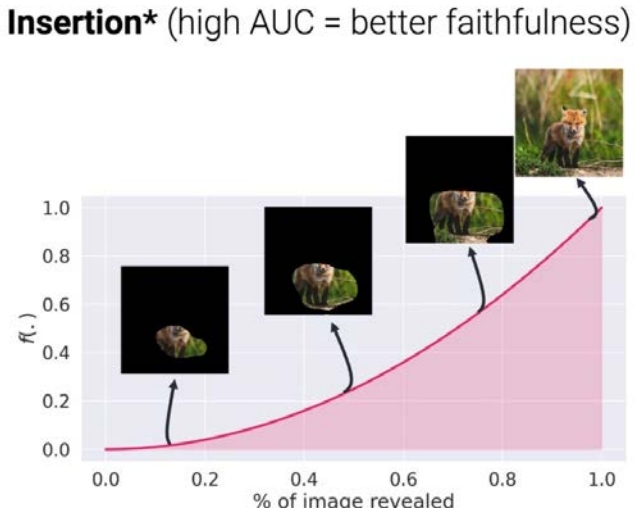
评价指标 - Faithfulness

Evaluation Metric

定义 (Insertion AUC score)

Insertion AUC score执行Deletion的逆过程，从基线状态的图像开始，然后逐步添加最重要的变量。

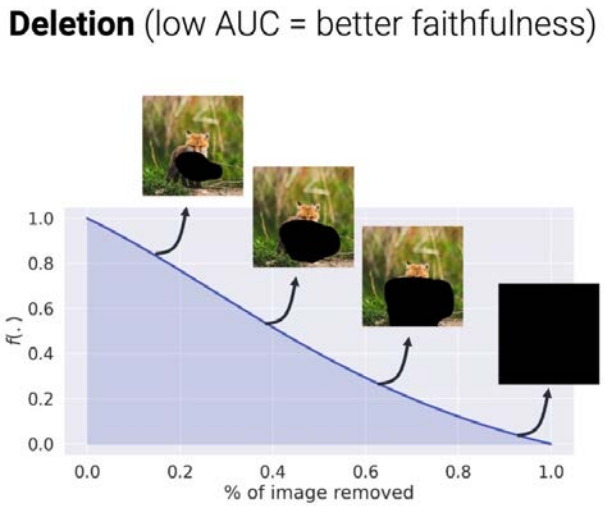
$$\text{Insertion}^{(k)} = f\left(x_{[x_{\bar{u}}=x_0]}\right)$$



定义 (Deletion AUC score)

Deletion AUC score测量当重要变量设置为基线状态时分数的下降。直观上，急剧下降表明解释方法已经很好地识别了决策的重要变量。

$$\text{Deletion}^{(k)} = f\left(x_{[x_u=x_0]}\right)$$



Petsiuk, Vitali, Abir Das, and Kate Saenko. "Rise: Randomized input sampling for explanation of black-box models." *BMVC*. 2018.

评价指标 - Faithfulness

Evaluation Metric

定义 (μFidelity)

给定一个预测器f，一个可解释函数g，一个数据点x，一个子集大小$|S|$，定义g对$f(x)$的忠实度为：

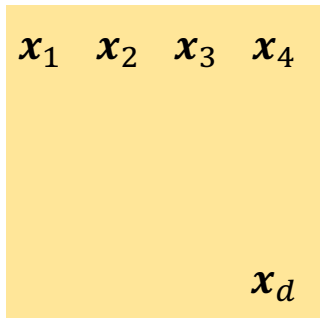
$$\mu_F(f;g;x) = \underset{S\in\binom{[d]}{|S|}}{\text{corr}} \left(\sum_{i\in S} g(f,x)_i , f(x) - f\left(x_{[x_S=\bar{x}_S]}\right) \right)$$

物理意义： g 的特征重要性得分应对应于x的重要特征。

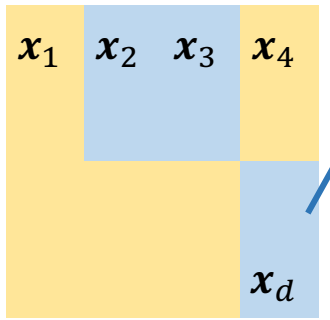
方法思路： 采样多个子区域，计算子区域显著图分数，与删除该子区域扰动的关联性（原始图像预测分数-删除重要区域的图像预测分数）。

指标性质： 越高越好。

原始图像



$g(f,x)$
显著图

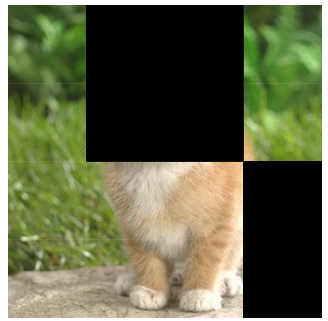


采样区域 $S \in \binom{[d]}{|S|}$

$\sum_{i\in S} g(f,x)_i$
蓝色区域显著图分数和

f 

$f(x)$
原始图像预测置信度

$-$ f 

$f\left(x_{[x_S=\bar{x}_S]}\right)$
Mask图像预测置信度

Bhatt, Umang, Adrian Weller, and José MF Moura. "Evaluating and aggregating feature-based model explanations." *IJCAI*. 2021.

Less is More

Fewer Interpretable Region via Submodular Subset Selection

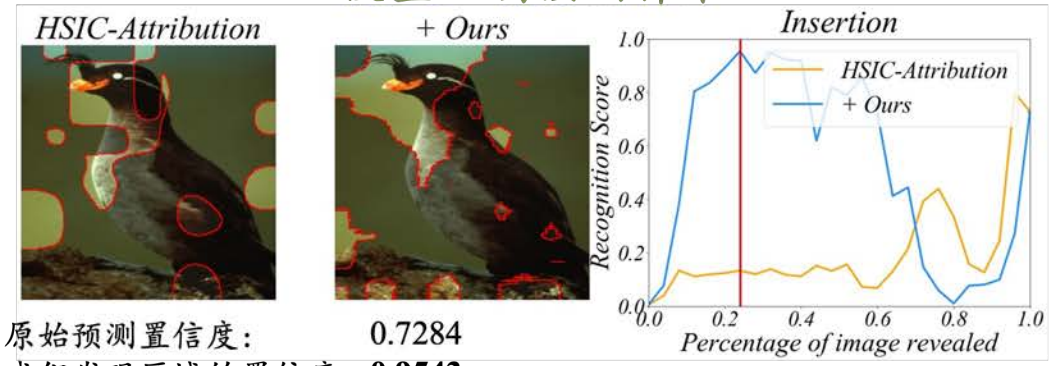
定义 (子模态子集选择理论)

给定一个集合V，一个子模方程$\mathcal{F}(\cdot)$，给定需要搜索的元素数量k，我们的目标是发现一个子集$S \subseteq V$，使子模方程数值最大化：

$$\max_{S \subseteq V, |S| < k} \mathcal{F}(S).$$

转变为图像归因问题

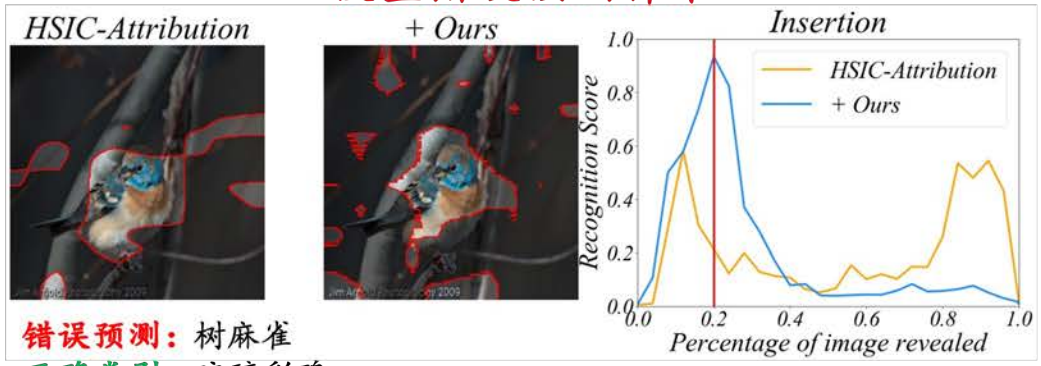
模型正确预测样本



原始预测置信度： 0.7284

我们发现区域的置信度： **0.9543**

模型错误预测样本



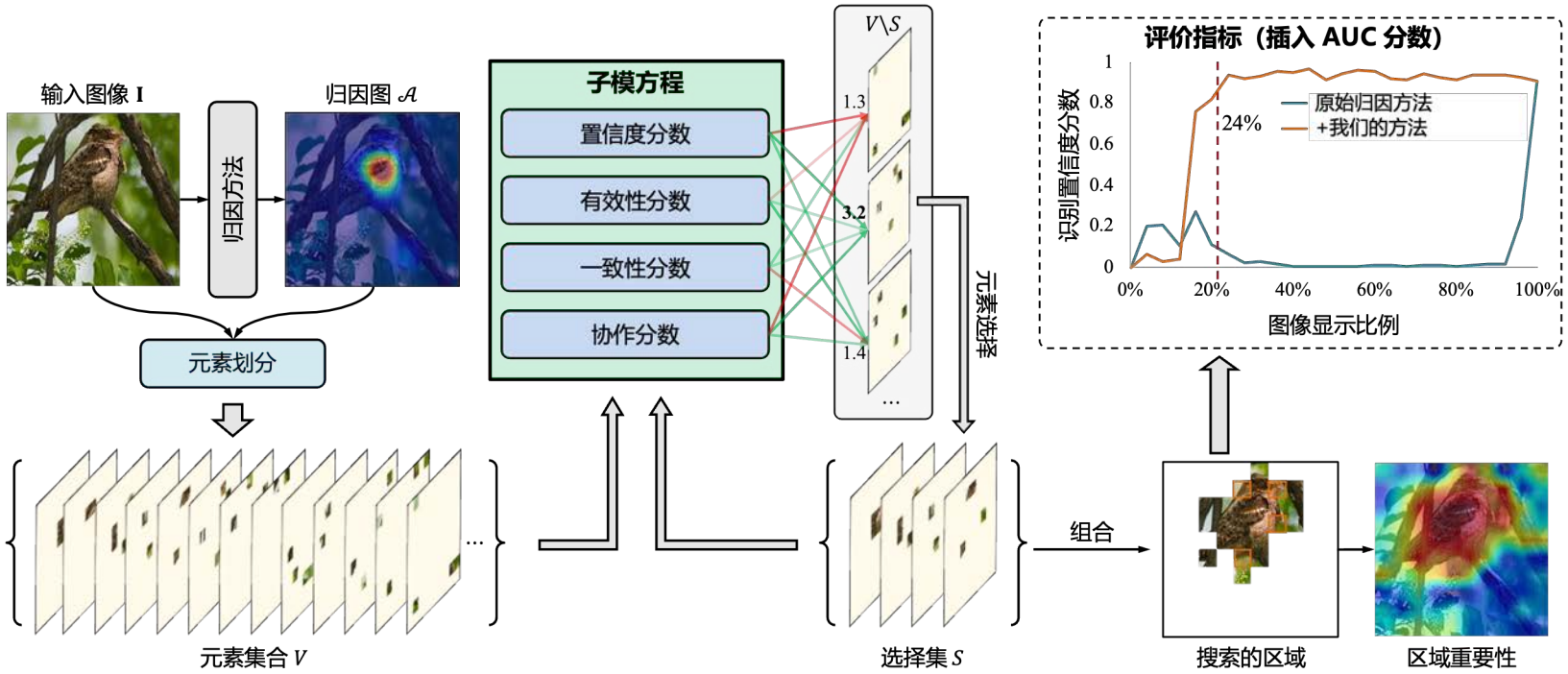
错误预测： 树麻雀

正确类别： 琉璃彩鸦

Chen, Ruoyu, et al. "Less is More: Fewer Interpretable Region via Submodular Subset Selection" *Submitted to ICLR 2024*.

Less is More

Fewer Interpretable Region via Submodular Subset Selection



设计子模方程 $\mathcal{F}(\cdot)$

置信度分数：

$$s_{\text{conf.}}(\mathbf{x}) = 1 - u = 1 - \frac{K}{\sum_{k_c=1}^{K}\left(\mathbf{e}_{k_c} + 1\right)},$$

有效性分数：

$$s_{\text{eff.}}(S) = \sum_{s_i \in S} \min_{s_j \in S, s_i \neq s_j} \text{dist}\left(F\left(s_i\right), F\left(s_j\right)\right),$$

一致性分数：

$$s_{\text{cons.}}\left(S, \boldsymbol{f}_s\right) = \frac{F\left(\sum_{\mathbf{I}^M \in S} \mathbf{I}^M\right) \cdot \boldsymbol{f}_s}{\left\|F\left(\sum_{\mathbf{I}^M \in S} \mathbf{I}^M\right)\right\| \left\|\boldsymbol{f}_s\right\|},$$

协作分数：

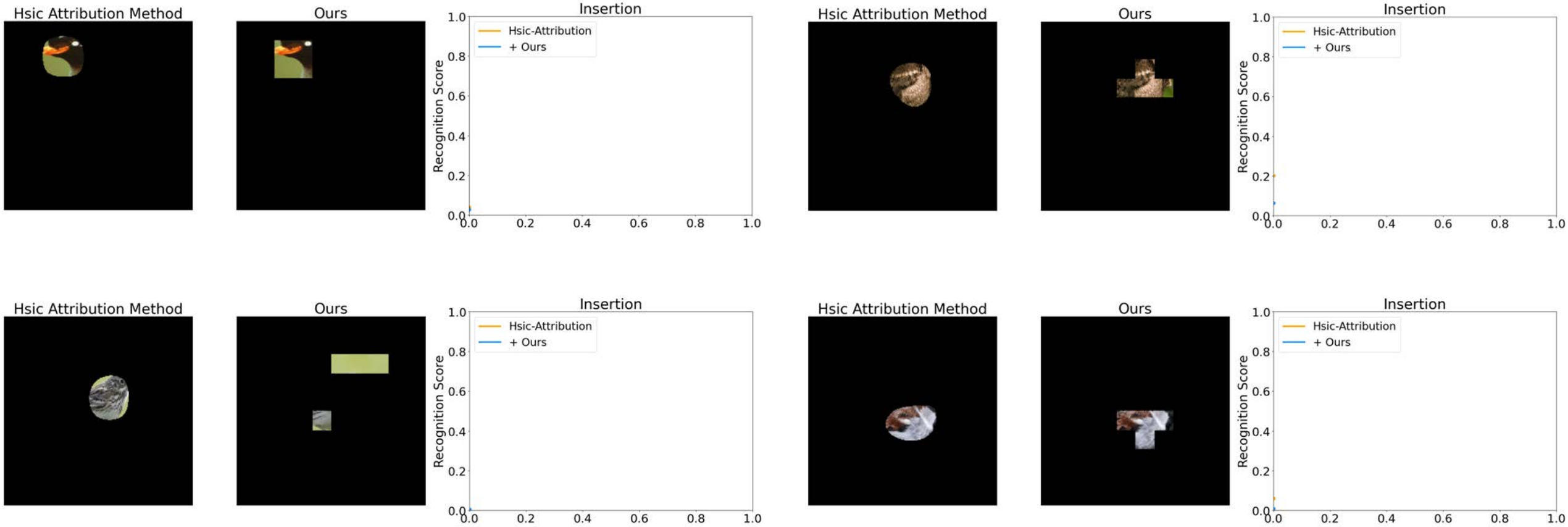
$$s_{\text{colla.}}\left(S, \mathbf{I}, \boldsymbol{f}_s\right) = 1 - \frac{F\left(\mathbf{I} - \sum_{\mathbf{I}^M \in S} \mathbf{I}^M\right) \cdot \boldsymbol{f}_s}{\left\|F\left(\mathbf{I} - \sum_{\mathbf{I}^M \in S} \mathbf{I}^M\right)\right\| \left\|\boldsymbol{f}_s\right\|},$$

Chen, Ruoyu, et al. "Less is More: Fewer Interpretable Region via Submodular Subset Selection" *Submitted to ICLR 2024*.

Less is More

Fewer Interpretable Region via Submodular Subset Selection

针对预测正确的样本，我们的方法有更好的归因效果

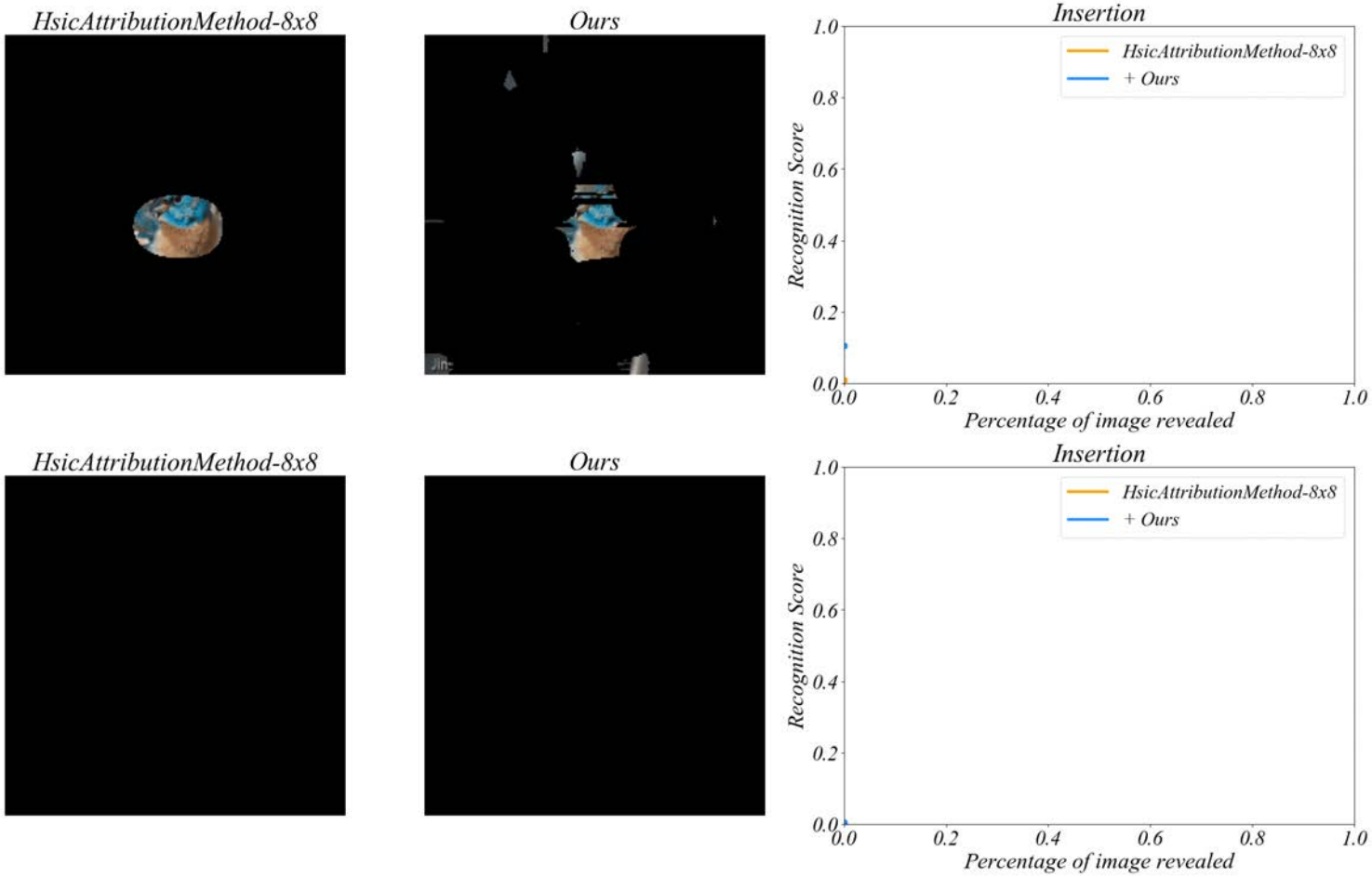


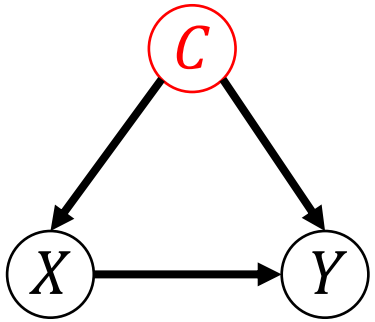
Chen, Ruoyu, et al. "Less is More: Fewer Interpretable Region via Submodular Subset Selection" *Submitted to ICLR 2024*.

Less is More

Fewer Interpretable Region via Submodular Subset Selection

寻找令模型预测错误的原因





针对置信度较差或者识别错误的方法，通过我们的方法是否能寻找到**导致错误的因**。

Chen, Ruoyu, et al. "Less is More: Fewer Interpretable Region via Submodular Subset Selection" *Submitted to ICLR 2024*.

SYSU 2023
PRESENTATION SLIDE

中山大學
SUN YAT-SEN UNIVERSITY

中国科学院信息工程研究
INSTITUTE OF INFORMATION ENGINEERING

03

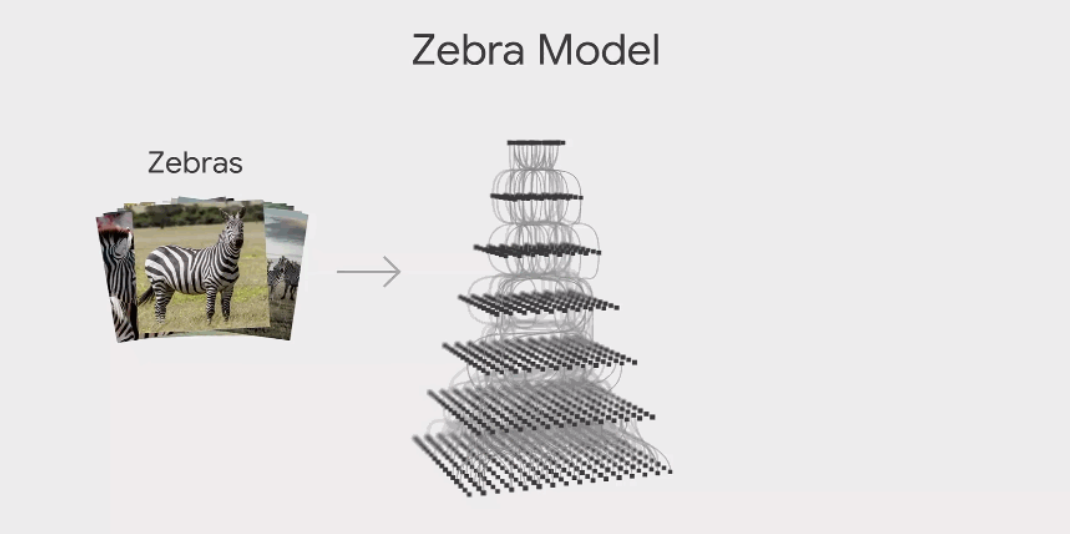
基于概念的可解释

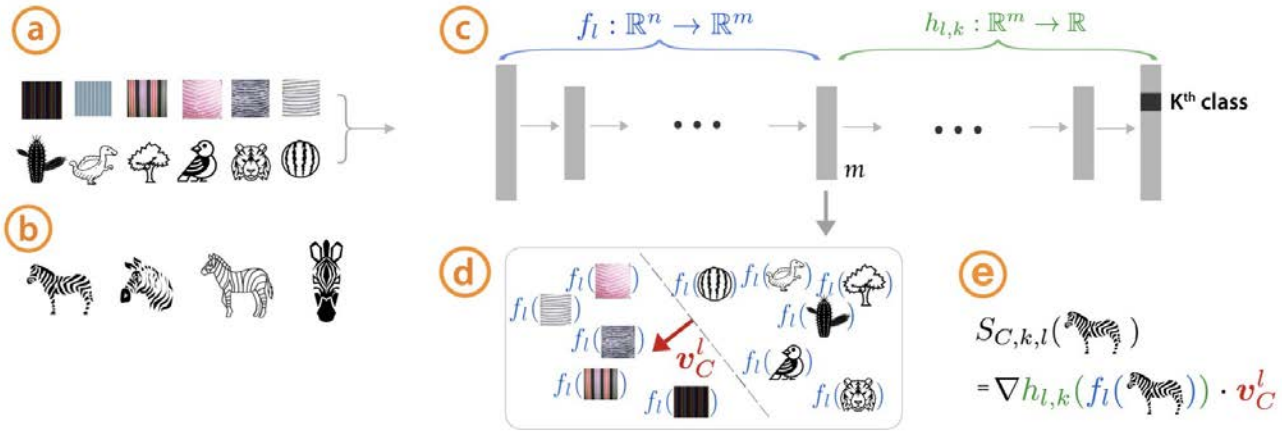
Concept-based interpretable algorithms

TCAV

Post-hoc Method

Concept Activation Vectors





TCAV: 对于一个在模型第f_l层的概念激活向量v_l，其中类别为c，预测分数为f_c。则：

$$S_c(x) = v_l \cdot \frac{\partial f_c(x)}{\partial f_l(x)},$$

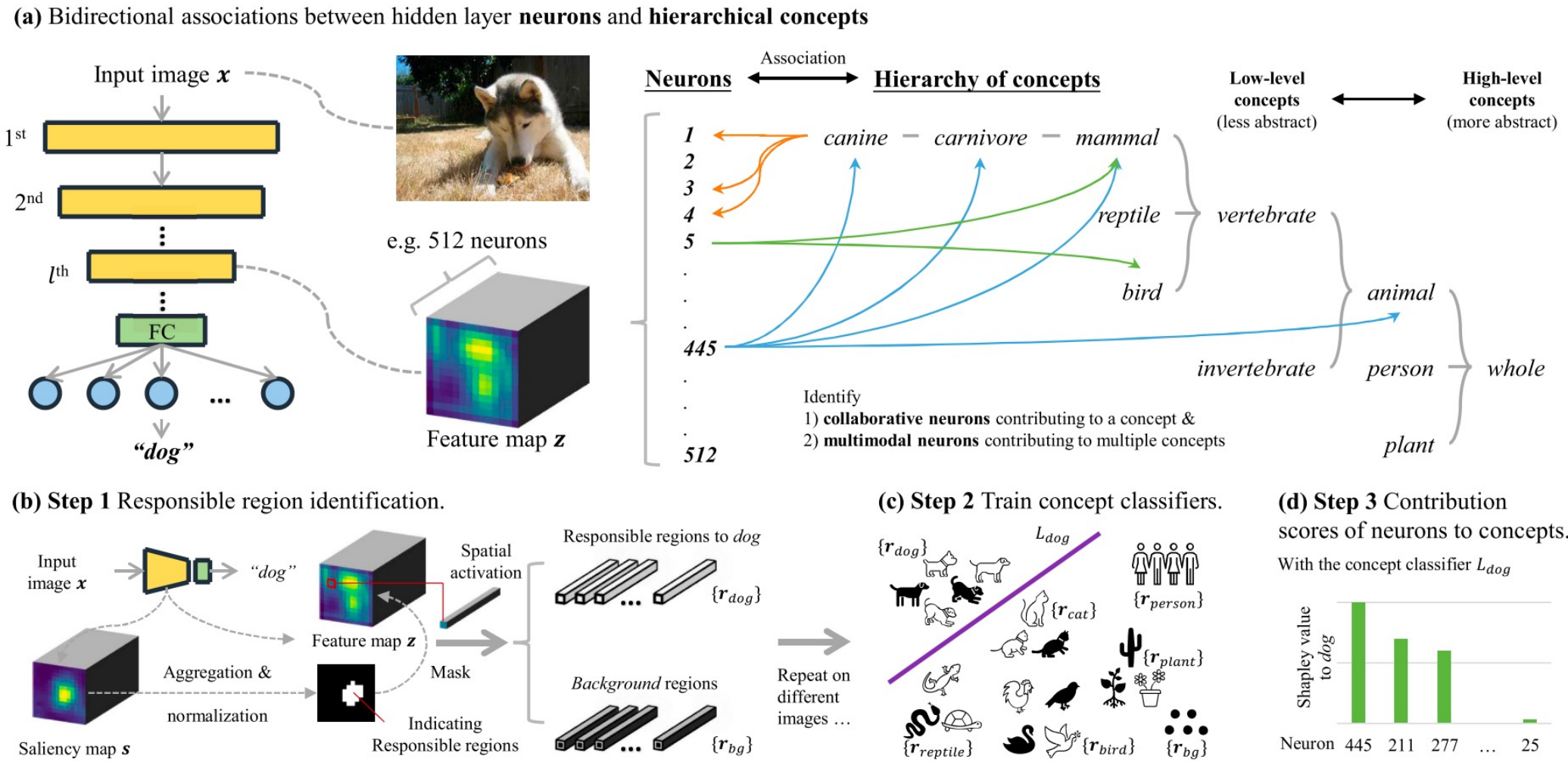
TCAV分数是指类别c中得分S_c为正的元素所占的百分比：

$$TCAV_c = \frac{|x \in \chi^c : S_c(x) > 0|}{|\chi^c|}.$$

Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)." *ICML*. 2018.

HINT Post-hoc Method

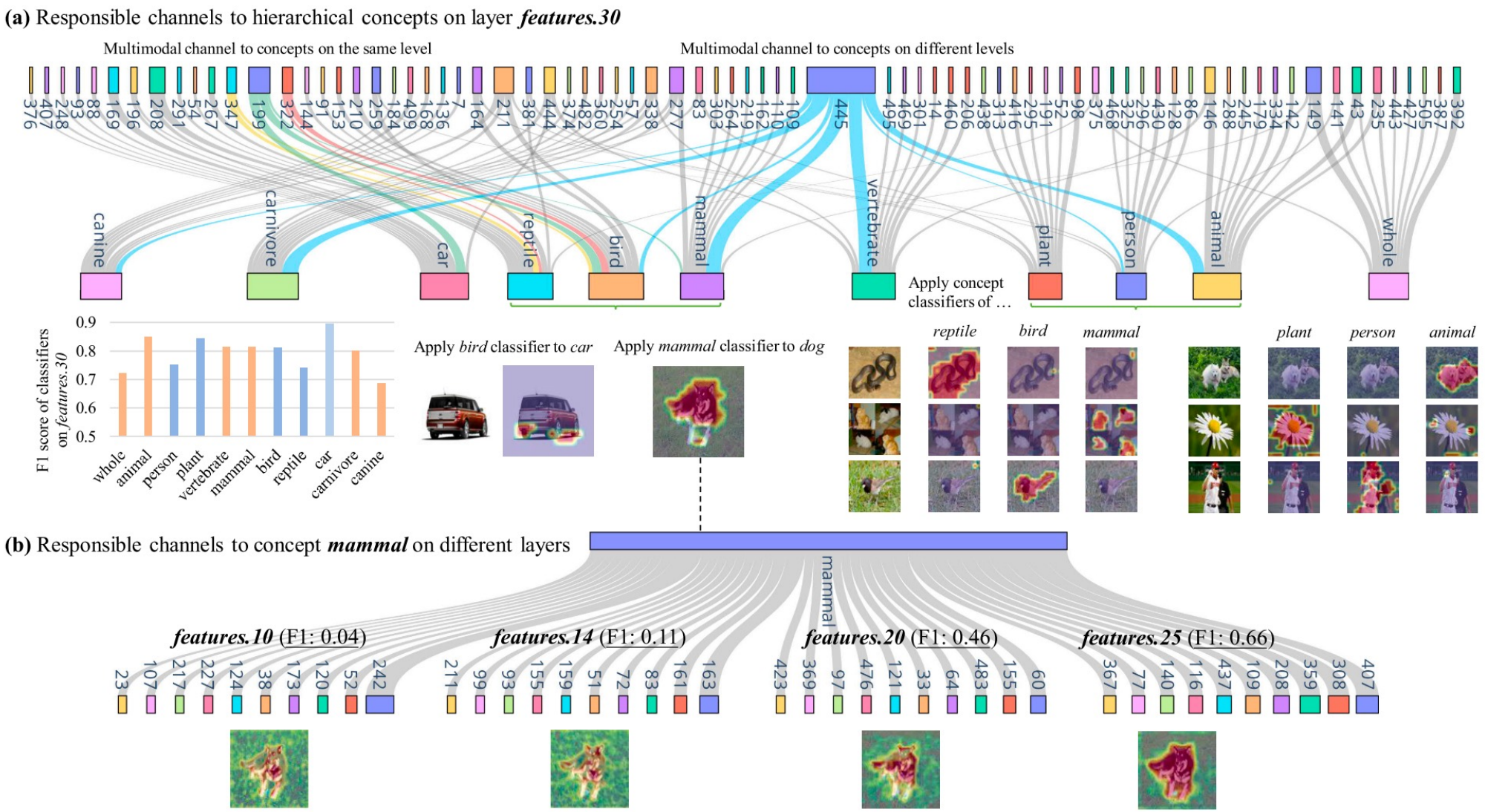
Hierarchical Neuron Concept Explainer



Wang, Andong, Wei-Ning Lee, and Xiaojuan Qi. "HINT: Hierarchical neuron concept explainer." *CVPR*. 2022.

HINT Post-hoc Method

Hierarchical Neuron Concept Explainer

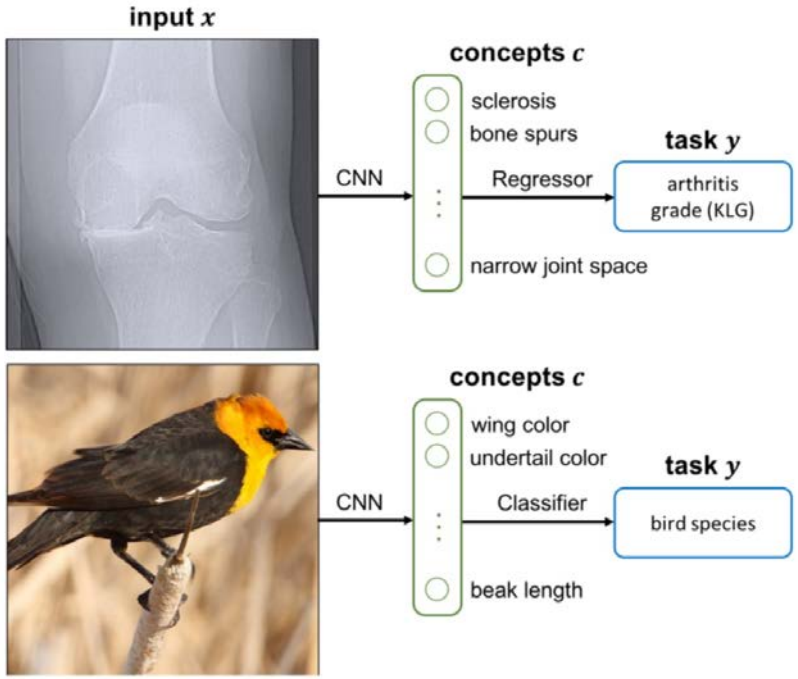


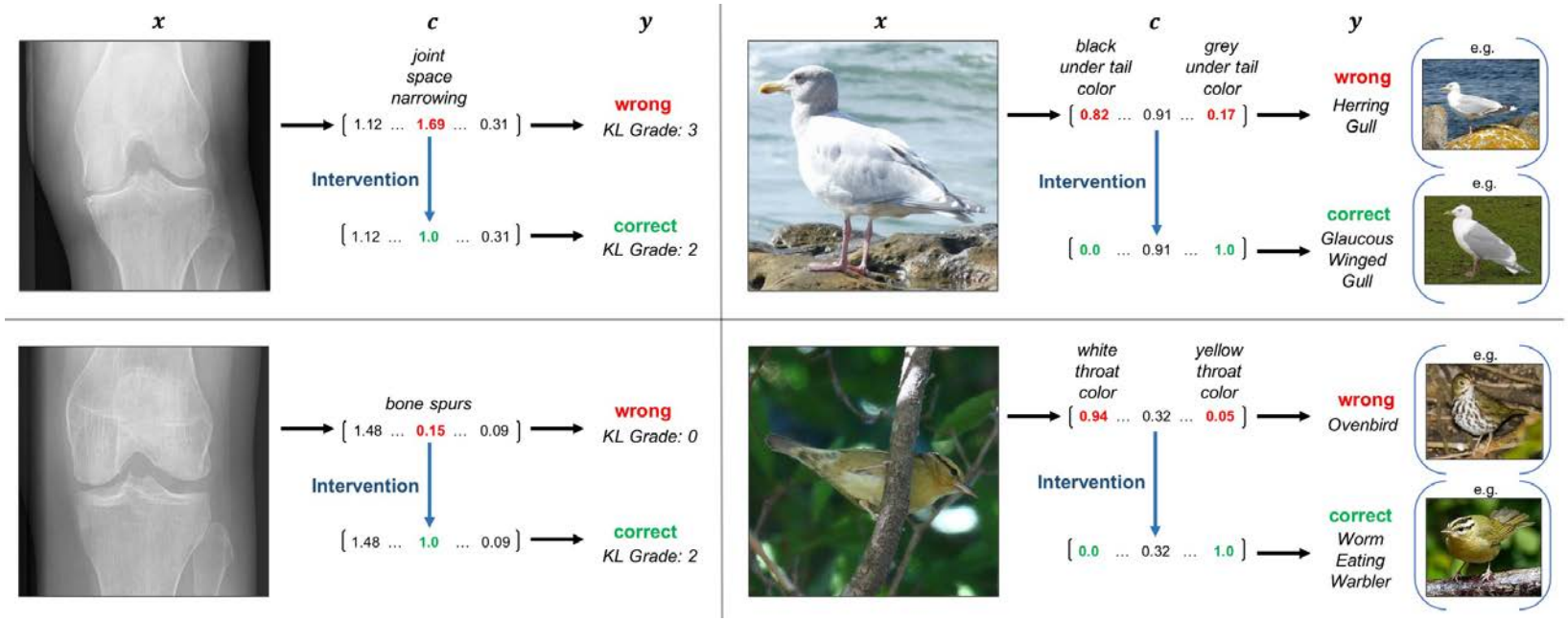
Wang, Andong, Wei-Ning Lee, and Xiaojuan Qi. "HINT: Hierarchical neuron concept explainer." *CVPR*. 2022.

CBM

Ante-hoc Method

Concept Bottleneck Models





Koh, Pang Wei, et al. "Concept bottleneck models." *ICML*. 2020.

SYSU 2023
PRESENTATION SLIDE

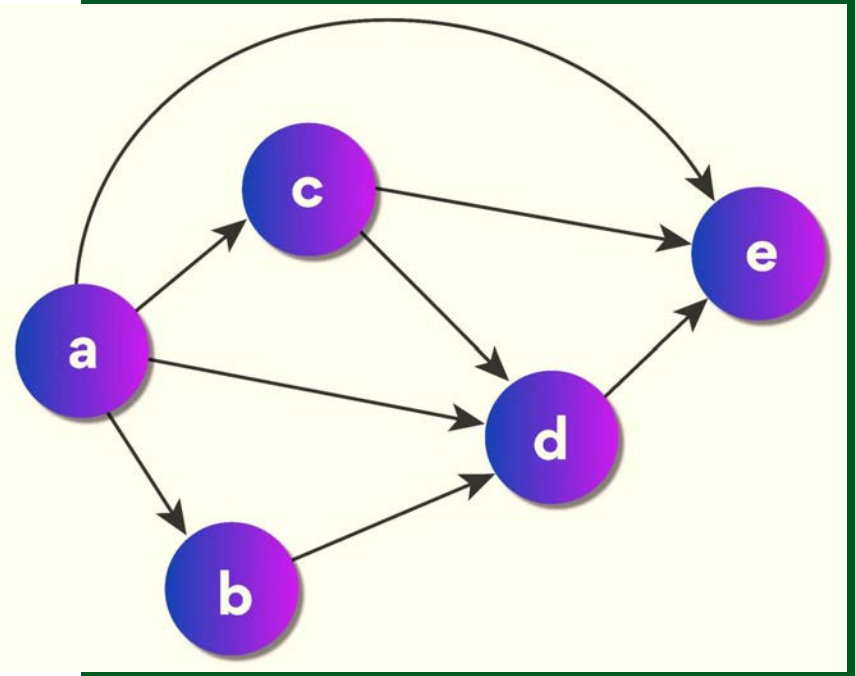
中山大學
SUN YAT-SEN UNIVERSITY

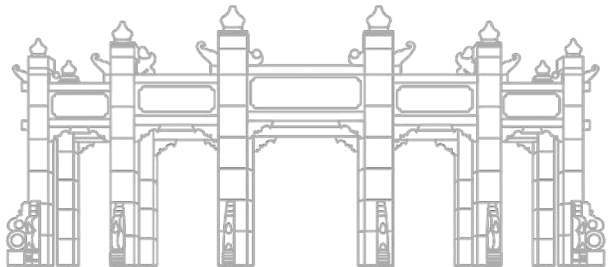
中国科学院信息工程研究
INSTITUTE OF INFORMATION ENGINEERING

04

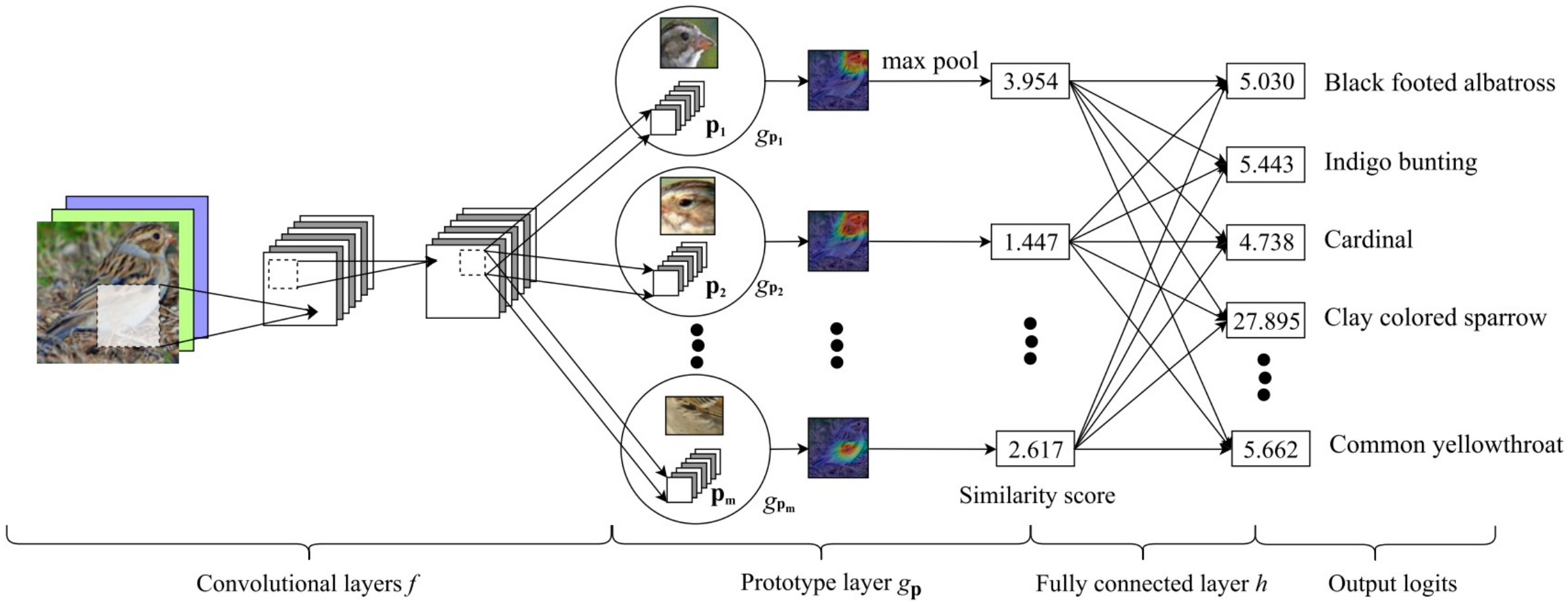
基于设计的可解释

Design-based interpretable algorithms





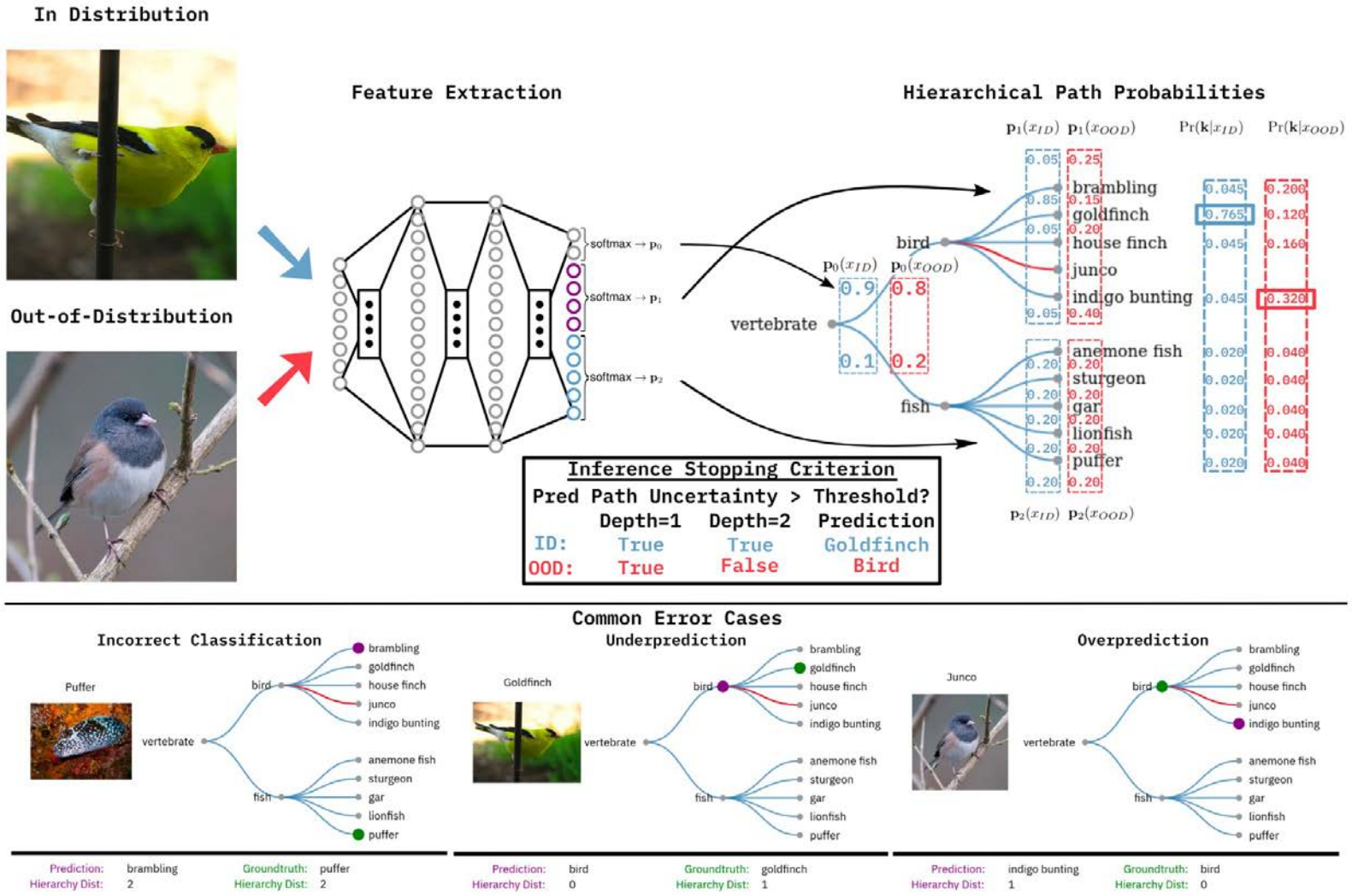
ProtoPNet Ante-hoc Method



Chen, Chaofan, et al. "This looks like that: deep learning for interpretable image recognition." NeurIPS 32 (2019).

Hierarchical Uncertainty Estimation

Ante-hoc Method



Linderman, Randolph, et al. "Fine-grain inference on out-of-distribution data with hierarchical classification." *Conference on Lifelong Learning Agents*. PMLR, 2023.

其它从数学角度设计的方法 Ante-hoc Methods

xNNs	Vaughan, Joel, et al. "Explainable neural networks based on additive index models." *arXiv preprint arXiv:1806.01933* (2018).
EBMs	Nori, Harsha, et al. "Interpretml: A unified framework for machine learning interpretability." *arXiv preprint arXiv:1909.09223* (2019).
SLIMs	Ustun, Berk, and Cynthia Rudin. "Supersparse linear integer models for optimized medical scoring systems." *Machine Learning* 102 (2016): 349-391.
RETAIN	Choi, Edward, et al. "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism." *NeurIPS* 29 (2016).
Bayesian Deep Learning	Wilson, Andrew Gordon. "The case for Bayesian deep learning." *arXiv preprint arXiv:2001.10995* (2020).

SYSU 2023
PRESENTATION SLIDE

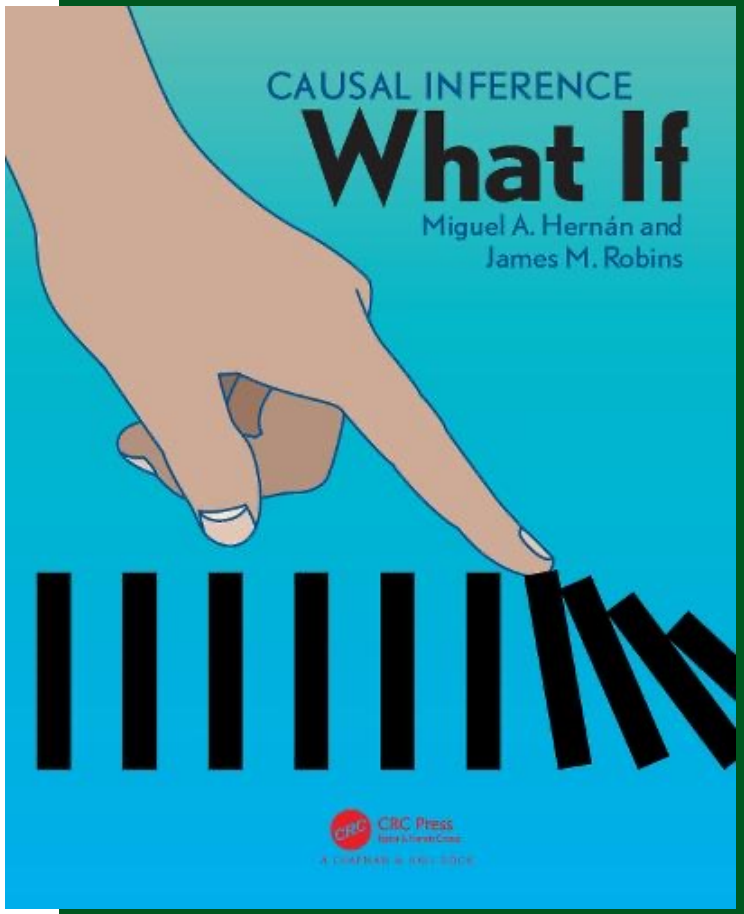
中山大學
SUN YAT-SEN UNIVERSITY

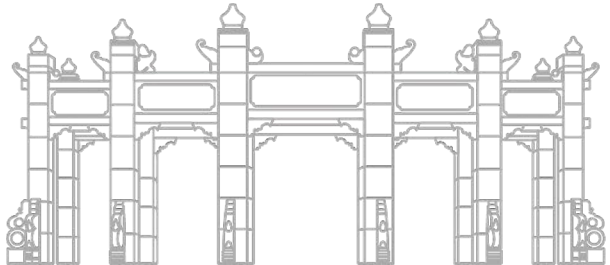
中国科学院信息工程研究
INSTITUTE OF INFORMATION ENGINEERING

05

基于因果的可解释

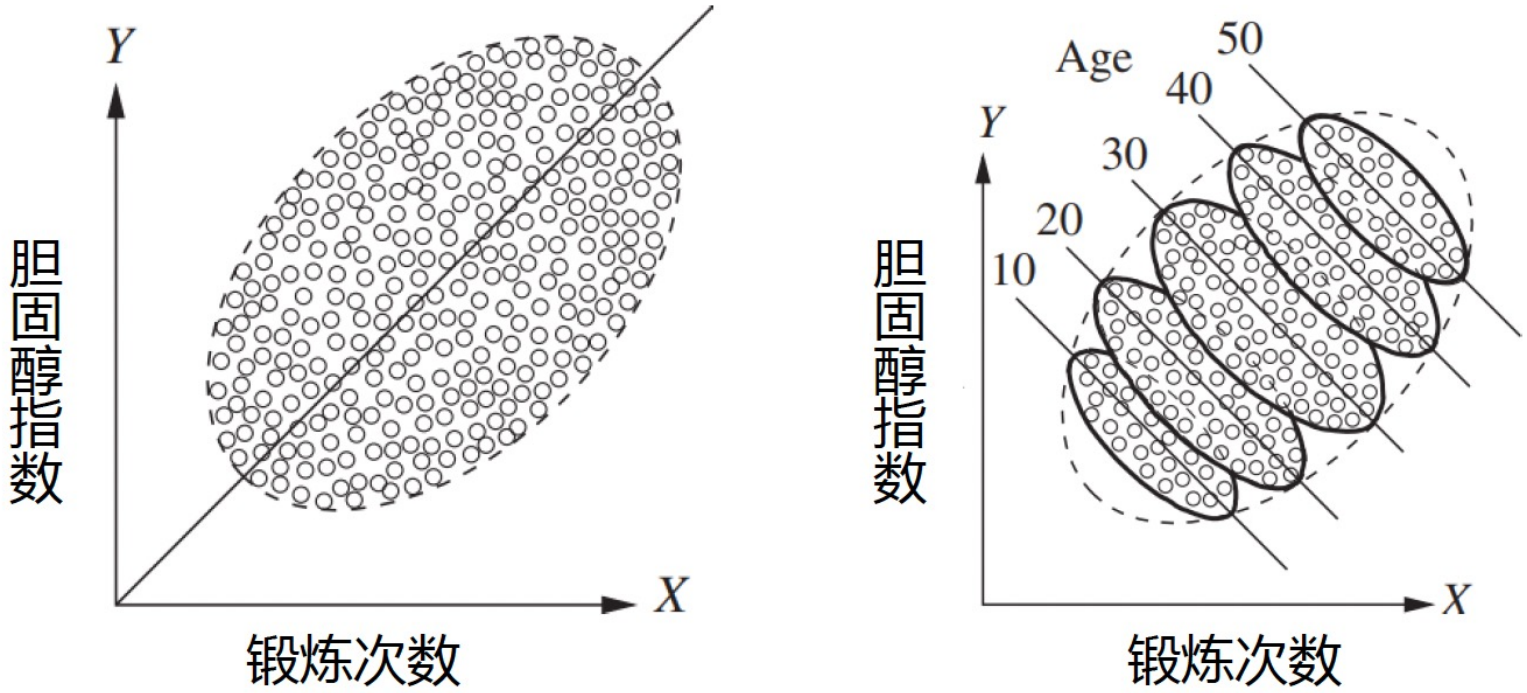
Causal-based interpretable algorithms





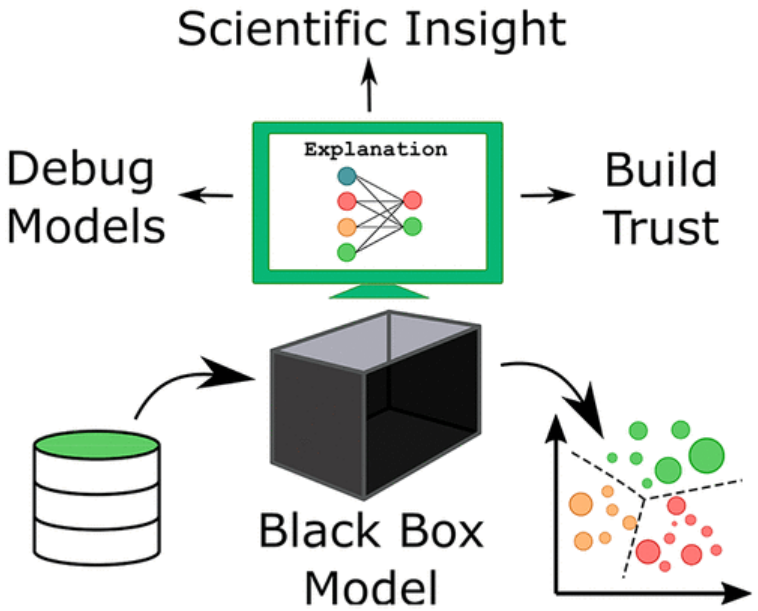
辛普森悖论

总体数据上得出的统计结论和分组数据上的统计结论相反

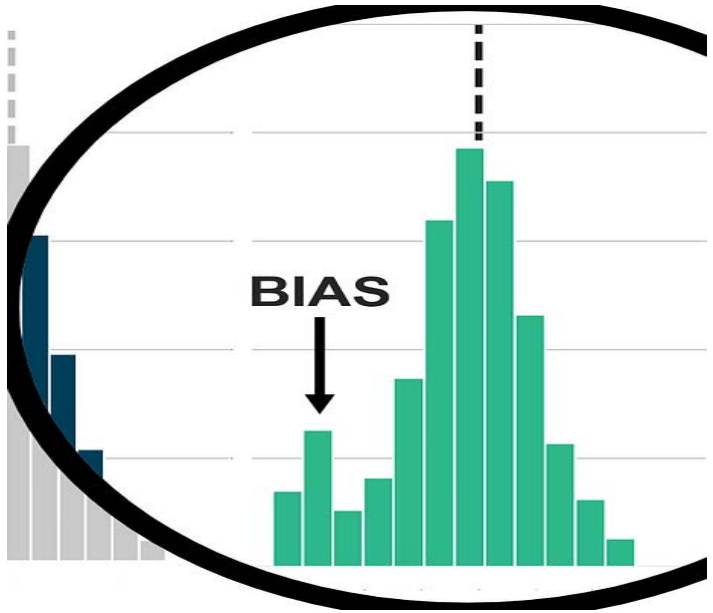


年龄高是锻炼次数高和胆固醇指数高的的共同原因

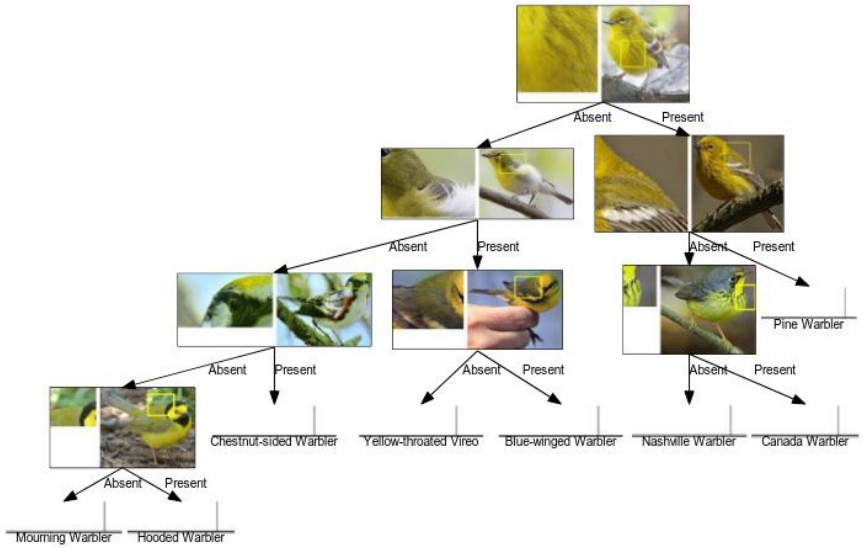
因果可以帮助我们什么?



调试和增强人工智能模型



检测潜在的Bias，提升模型的鲁棒性



理解模型背后的决策逻辑

结构因果模型

SCM: Structural Causal Model

因果模型图以图的方式更直观地表示结构因果模型

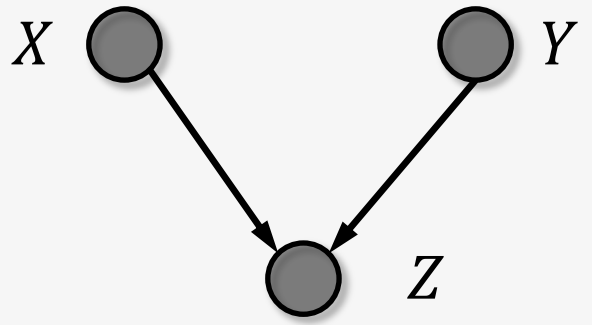
- 外生变量集合： U
  - 外生变量不依赖于其他变量
- 内生变量集合：V
  - 内生变量至少依赖一个变量
- 确定内生变量取值的函数集合： F

结构因果模型：

$$U = \{X, Y\},$$
$$V = \{Z\},$$
$$F = \{f_Z\},$$

$$f_Z: Z = 2X + 3Y$$

因果模型图：

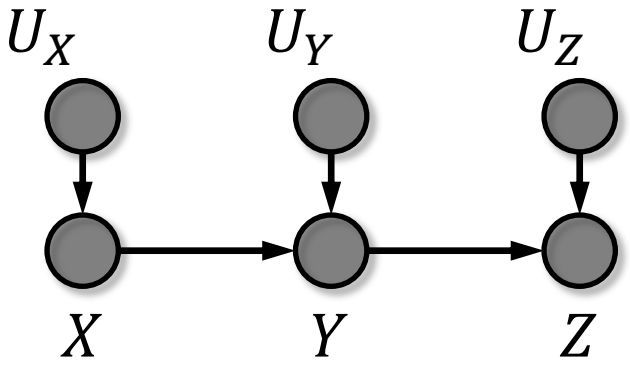


✓ 因果模型图中，节点表示变量，边表示变量间的依赖关系

✓ 因果模型图是一个有向无环图（DAG：Directed Acyclic Graphs）

因果模型图与独立性

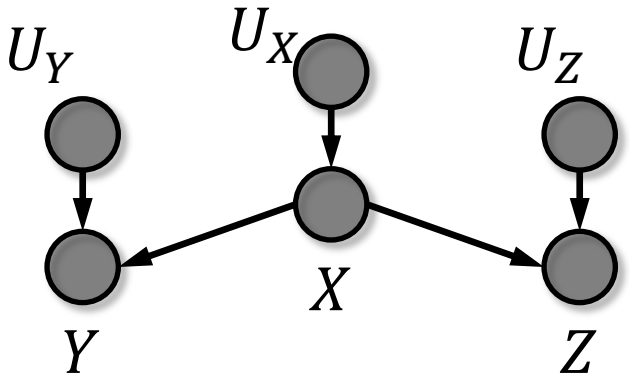
链结构



链结构中的条件独立性：$X \perp Z|Y$

已知外生变量U_X和U_Z独立，给定Y时，Z只受U_Z影响，X只受U_X影响，故X、Z独立。

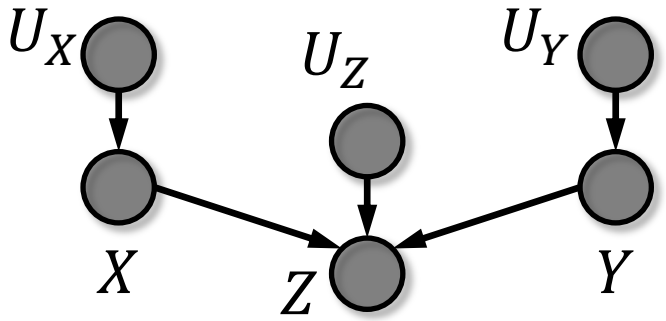
分叉结构



分叉结构中的条件独立性：$Y \perp Z|X$

已知外生变量U_Y和U_Z独立，给定X时，Z只受U_Z影响，Y只受U_Y影响，故Y、Z独立。

对撞结构

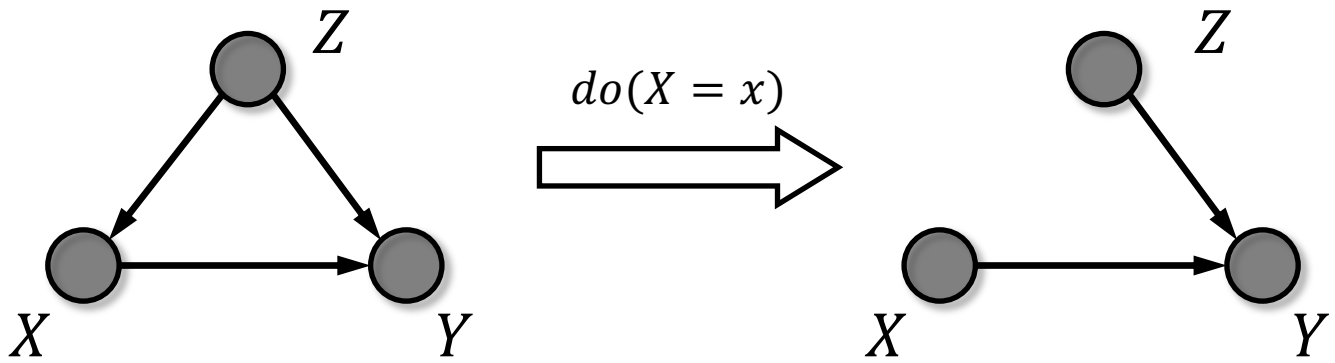


对撞结构中的条件独立性：$X \perp Y$，但给定Z或者Z的子孙时，X与Y不相互独立。

干预 (Intervention) 的定义：

✓ 将变量其固定为某个值，限制了该变量随其他变量而变化的自然趋势

✓ 记为$do(X = x)$，在因果模型图中即为去掉所有指向X的边

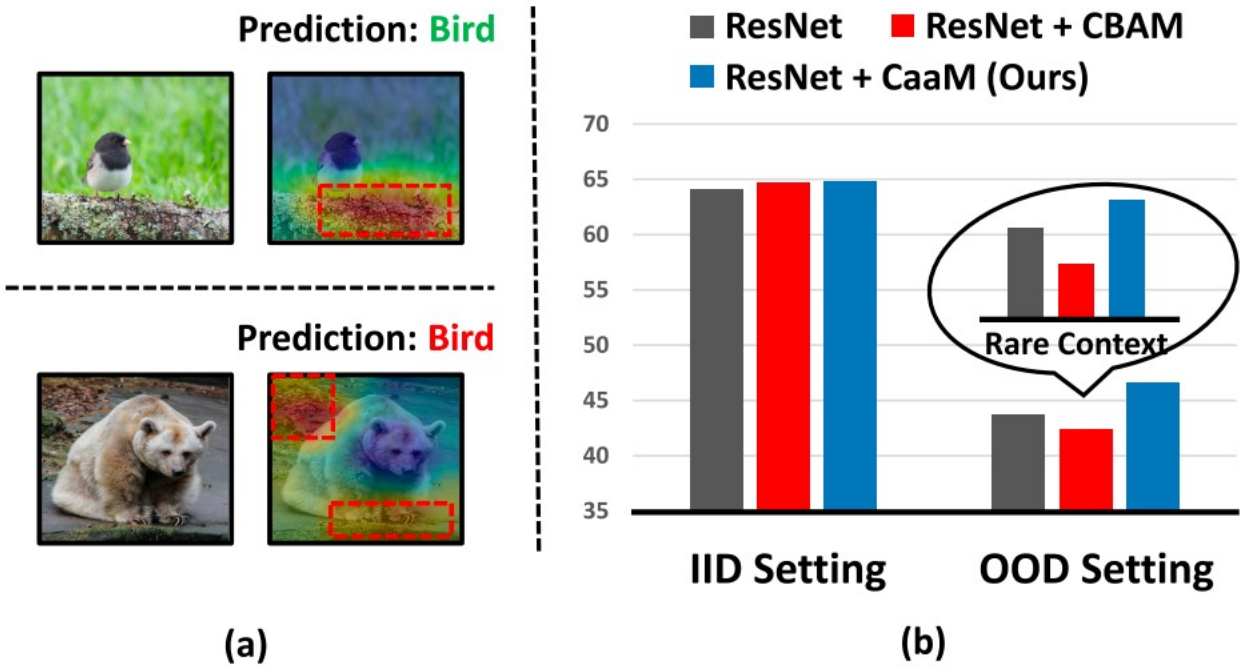


常用方法-以变量为条件：$\boldsymbol{P(Y = y|X = x)}$

- 在$\boldsymbol{X = x}$的条件下$\boldsymbol{Y = y}$的概率
- 在变量$\boldsymbol{X}$的取值都为$\boldsymbol{x}$的这些个体上，$\boldsymbol{Y}$的总体分布
- 关注问题的子集，改变的仅是我们对世界的看法，而不是改变了世界。

CAAM

SCM: Structural Causal Model



根据Human的先验经验构建结构因果图模型：

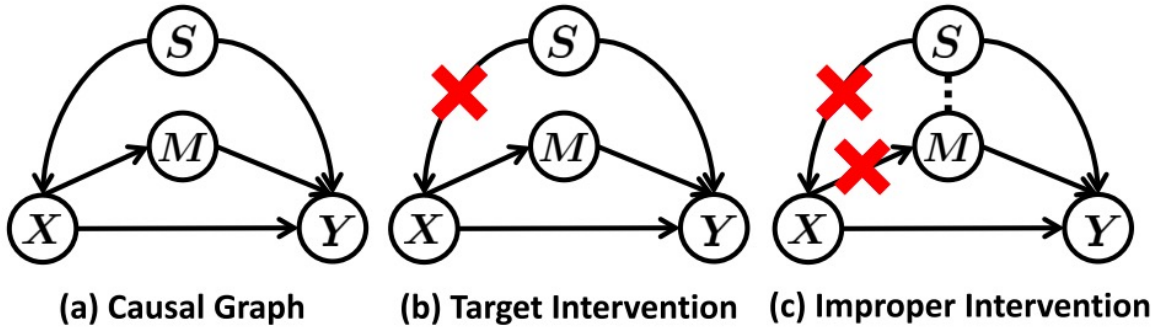


Figure 3. The causal graphs of visual recognition.

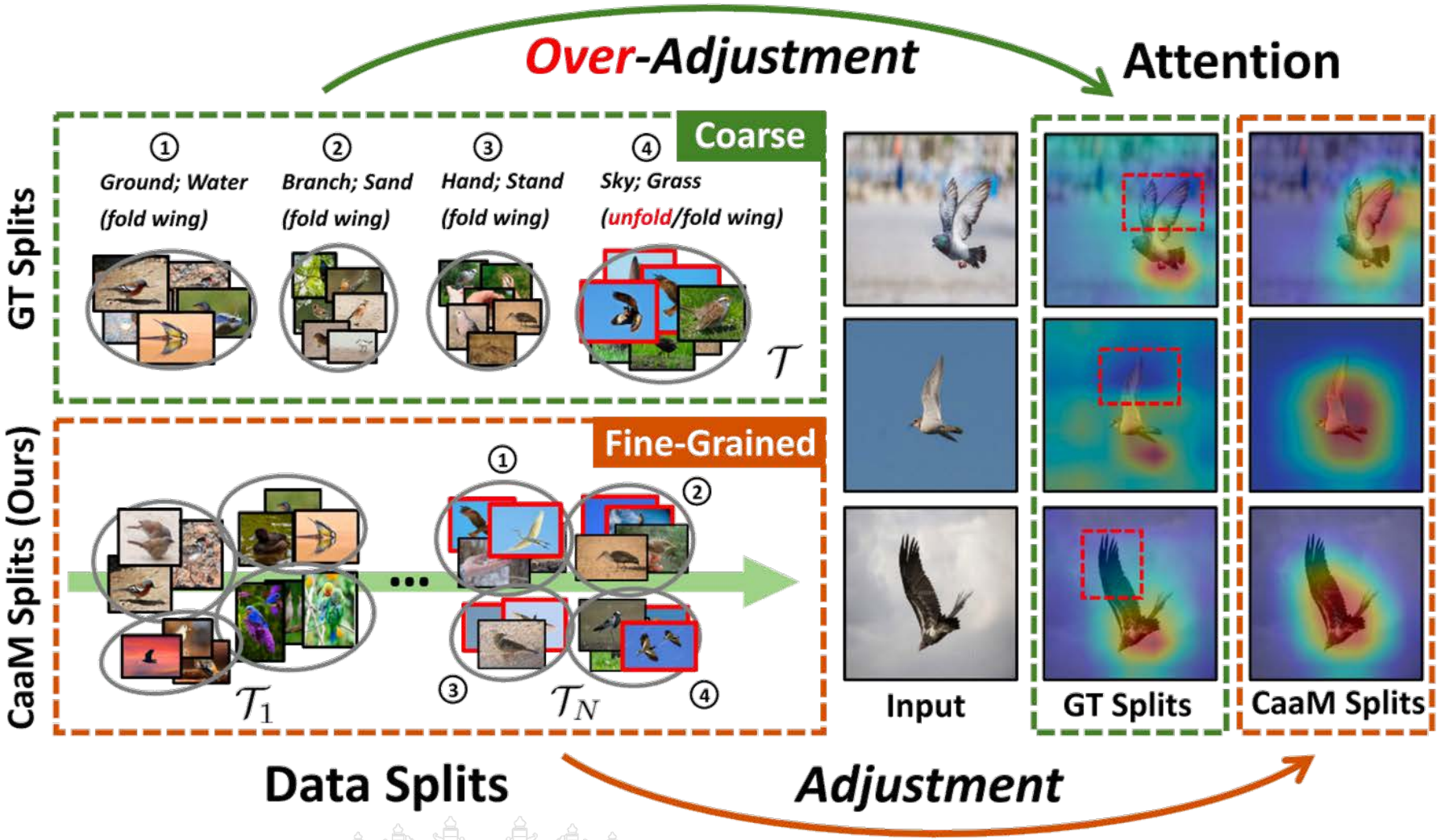
S: confounder (混淆变量，不利于模型学习)
M: mediator (有利于模型学习的中间变量)；
X: image;
Y: label

Wang, Tan, et al. "Causal attention for unbiased visual recognition." *ICCV*. 2021.

CAAM

SCM: Structural Causal Model

通过训练时控制数据变量划分实现干预



Wang, Tan, et al. "Causal attention for unbiased visual recognition." *ICCV*. 2021.

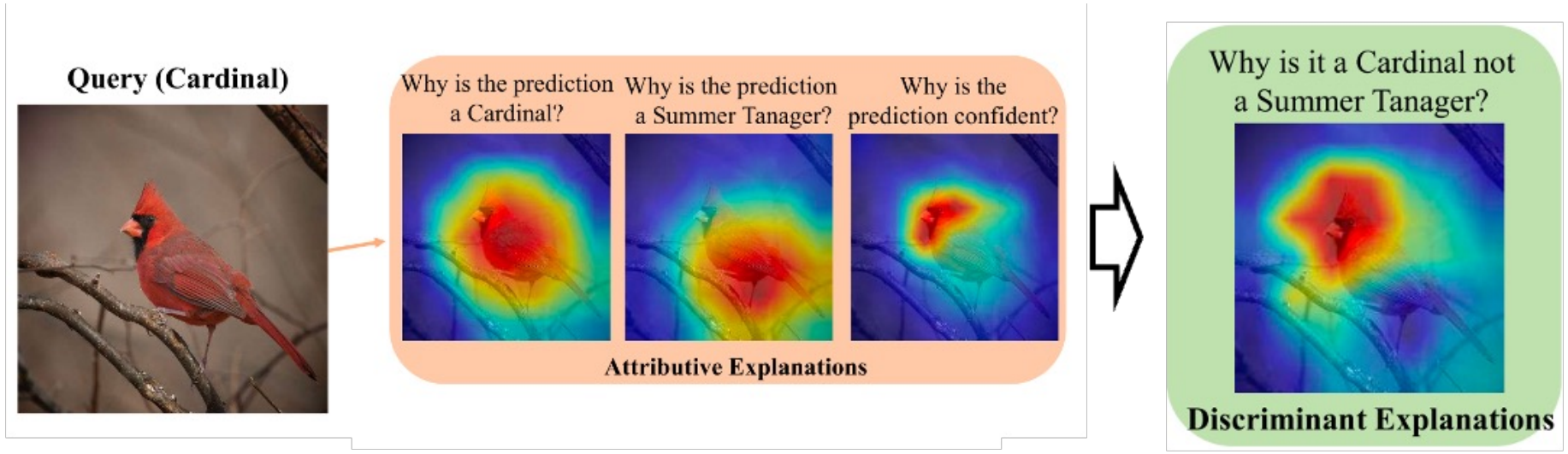
反事实

Counterfactual

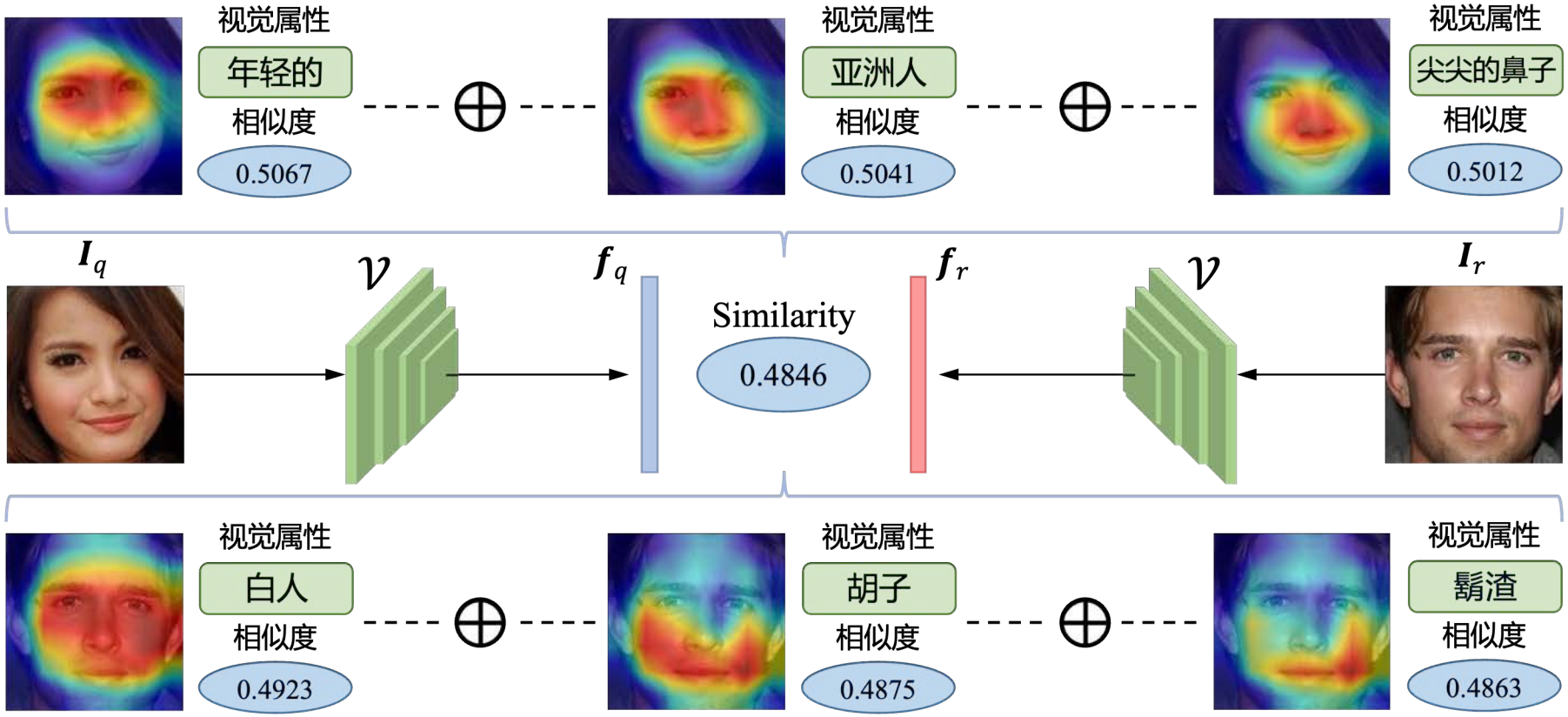
反事实(counterfactual)的定义：

✓ 本质意思是指在实际生活中,某些情况并未发生,即与“事实”相反。

✓ 分类任务中，反事实可以是解释模型为什么决策为A而不是B?



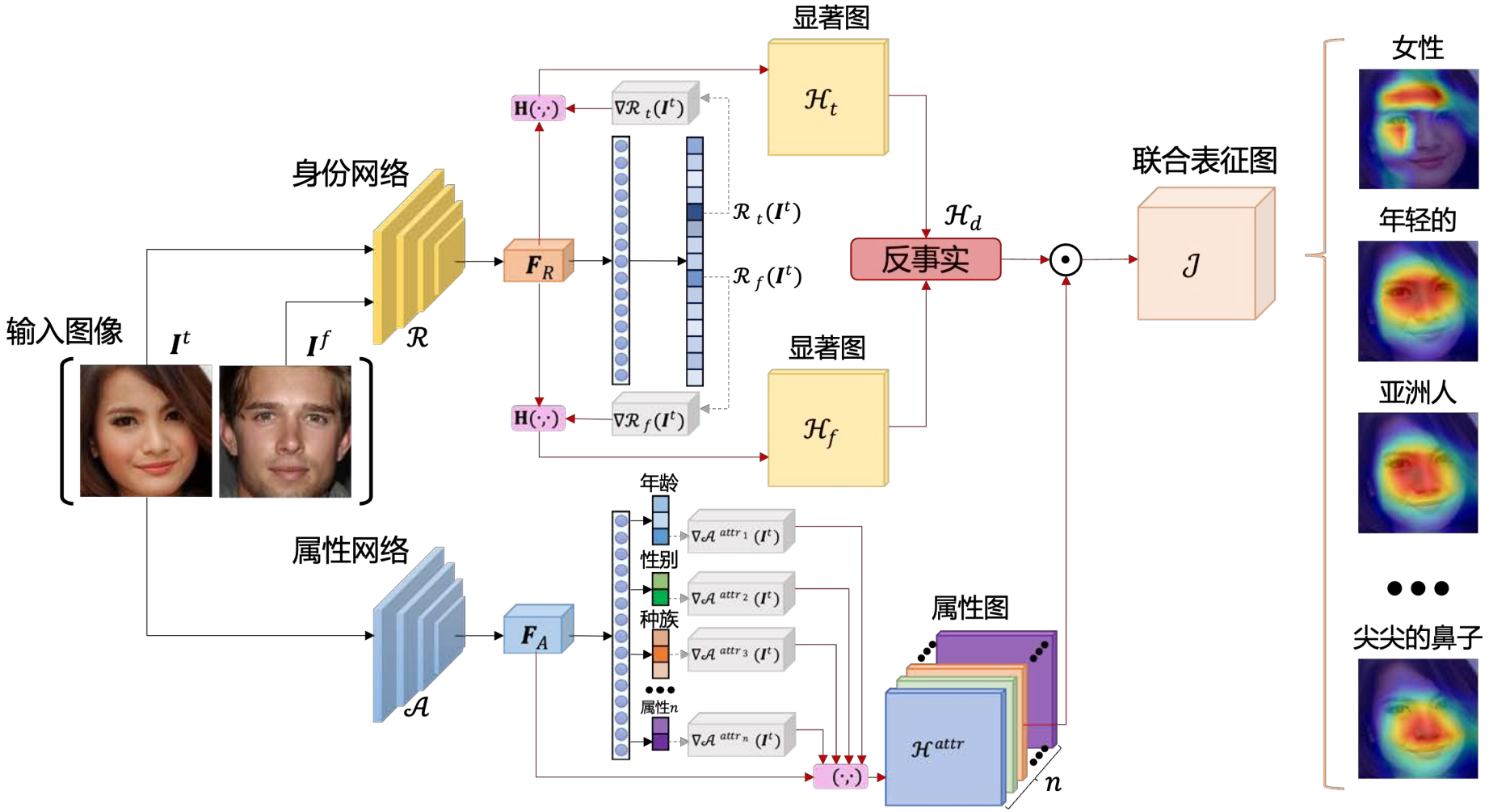
Sim2Word



提出一种新颖的**视觉相似度任务**解释模型，旨在用视觉属性集解释相似度得分，并提供视觉证据和语义描述。具体来说，我们将相似的分数分解为**视觉属性的组合**。

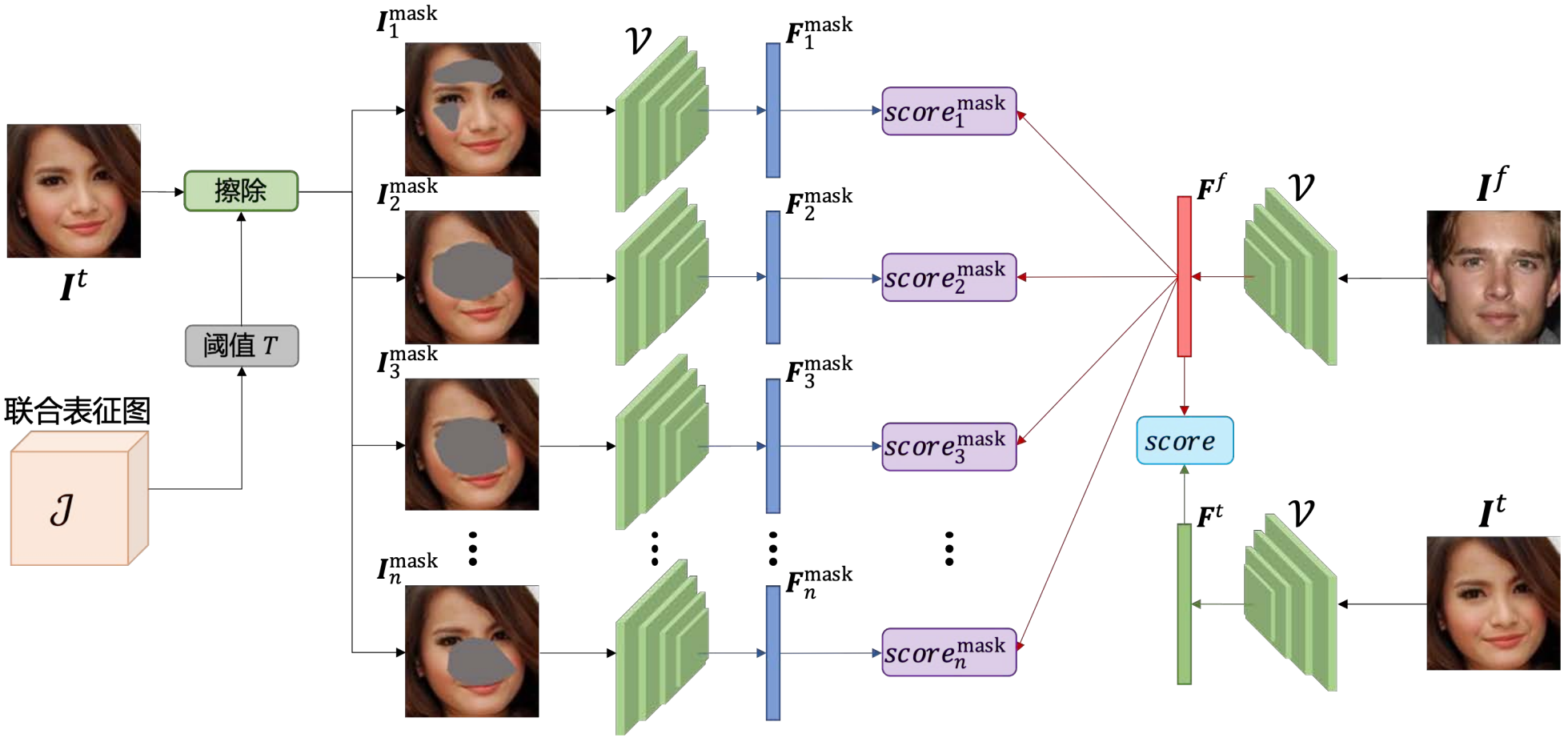
Chen, Ruoyu, et al. "Sim2Word: Explaining similarity with representative attribute words via counterfactual explanations." *ACM Transactions on Multimedia Computing, Communications and Applications* 19.6 (2023): 1-22.

Sim2Word



Chen, Ruoyu, et al. "Sim2Word: Explaining similarity with representative attribute words via counterfactual explanations." *ACM Transactions on Multimedia Computing, Communications and Applications* 19.6 (2023): 1-22.

Sim2Word



Chen, Ruoyu, et al. "Sim2Word: Explaining similarity with representative attribute words via counterfactual explanations." *ACM Transactions on Multimedia Computing, Communications and Applications* 19.6 (2023): 1-22.

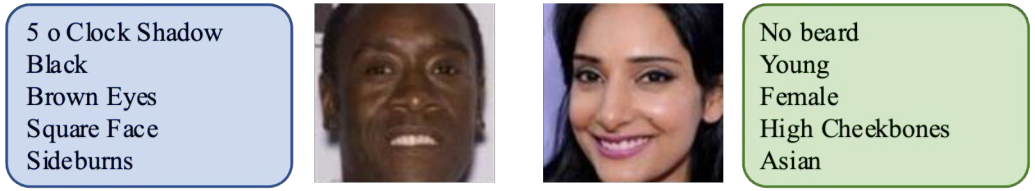
Sim2Word

Sim2Word 模型提供的解释包括：

- 显著图

What regions does the model focus on?

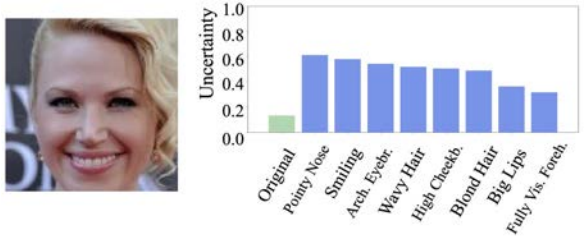
前五个最具代表性的面部属性



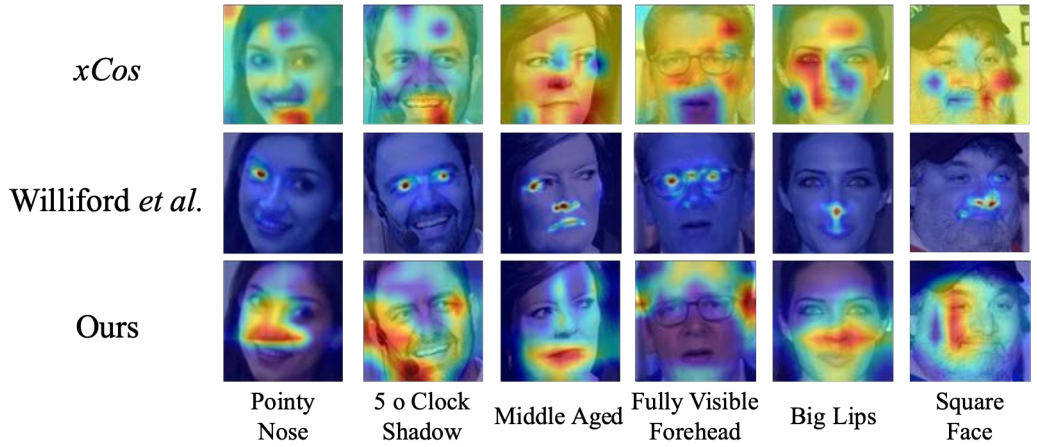
- 文本描述

The most characteristic attribute is the **pointy nose**

- 数值分数



最先进方法比较



Chen, Ruoyu, et al. "Sim2Word: Explaining similarity with representative attribute words via counterfactual explanations." *ACM Transactions on Multimedia Computing, Communications and Applications* 19.6 (2023): 1-22.

SYSU 2023
PRESENTATION SLIDE

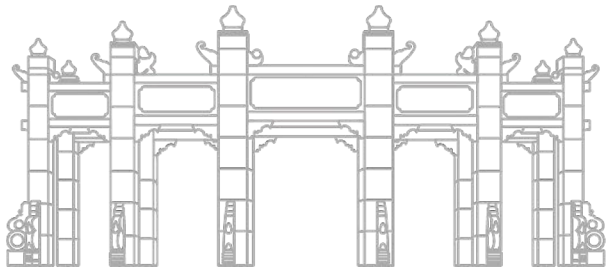
中山大學
SUN YAT-SEN UNIVERSITY

中国科学院信息工程研究
INSTITUTE OF INFORMATION ENGINEERING

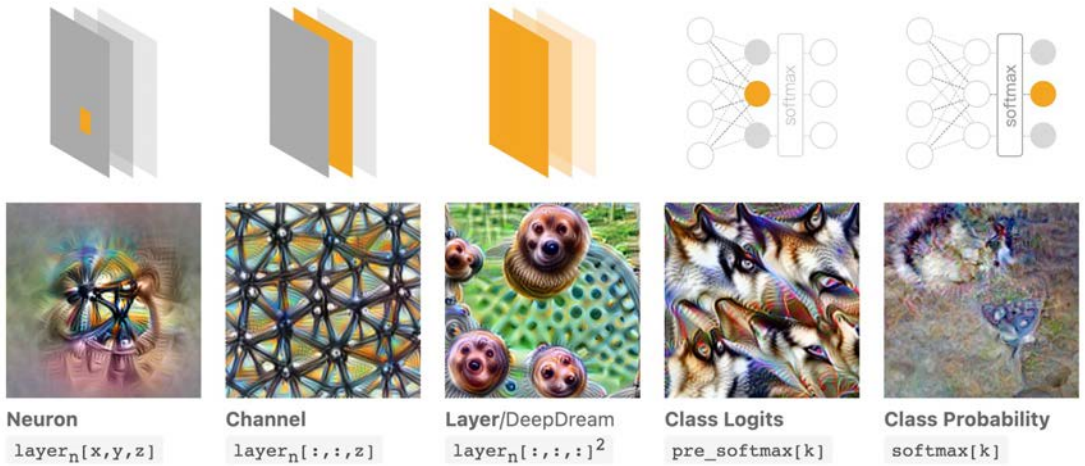
06

其他可解释方法

Other interpretable algorithms



Feature Visualization

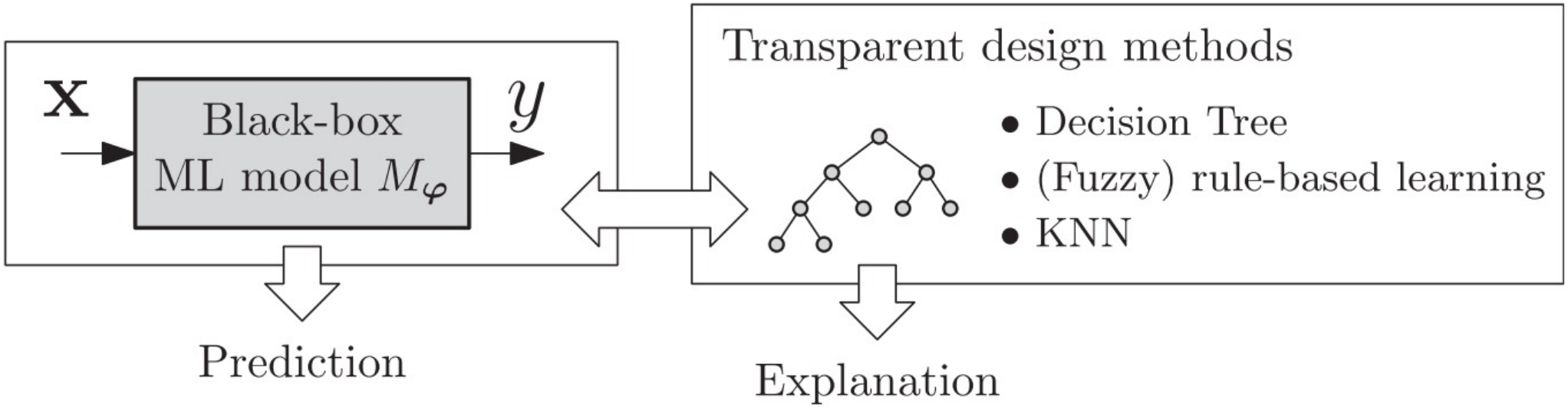


Feature Visulization[2]:
指定中间单元，优化输入，使目标单元有最大激活响应，观察优化的输入图像。

Feature Visualization, https://distill.pub/2017/feature-visualization/

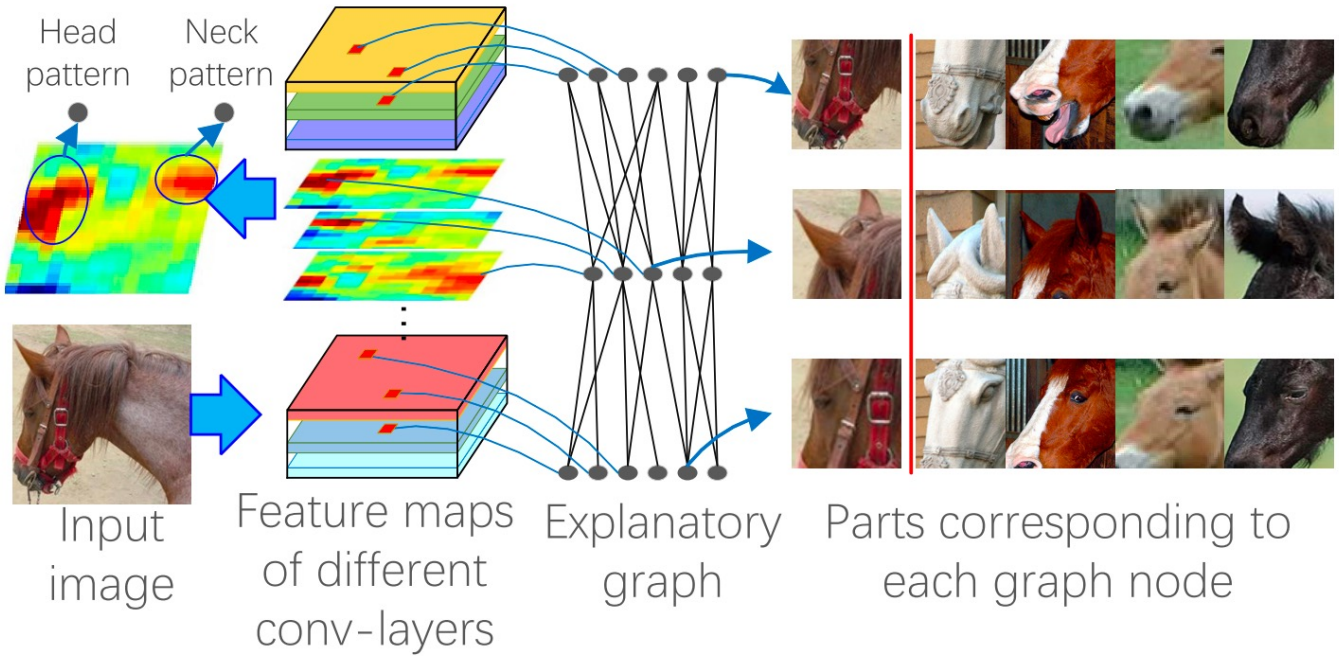
代理模型进行解释

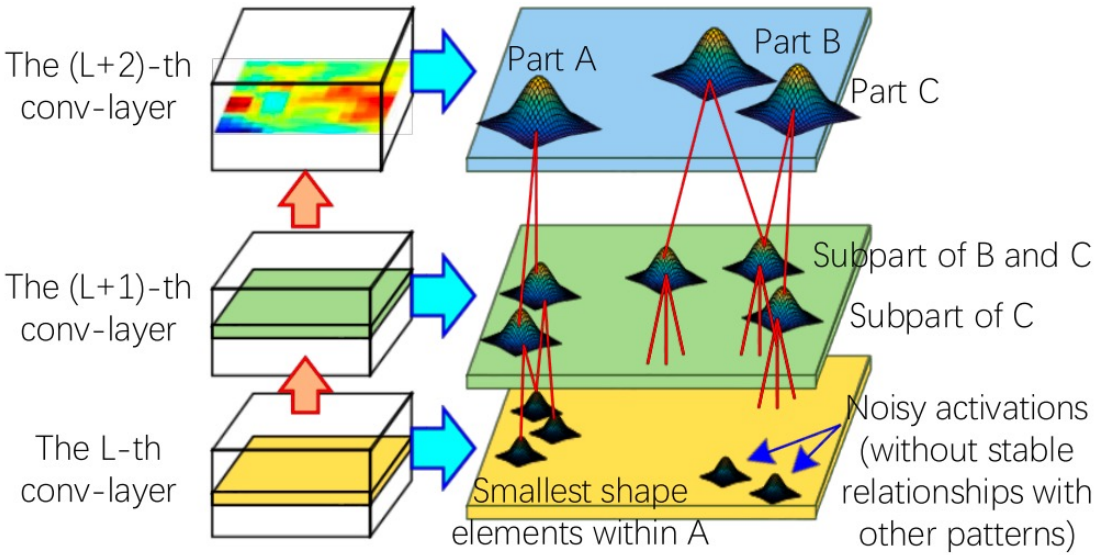
Evaluation Metric



Arrieta, Alejandro Barredo, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." *Information Fusion* 58 (2020): 82-115.

Explanatory Graph

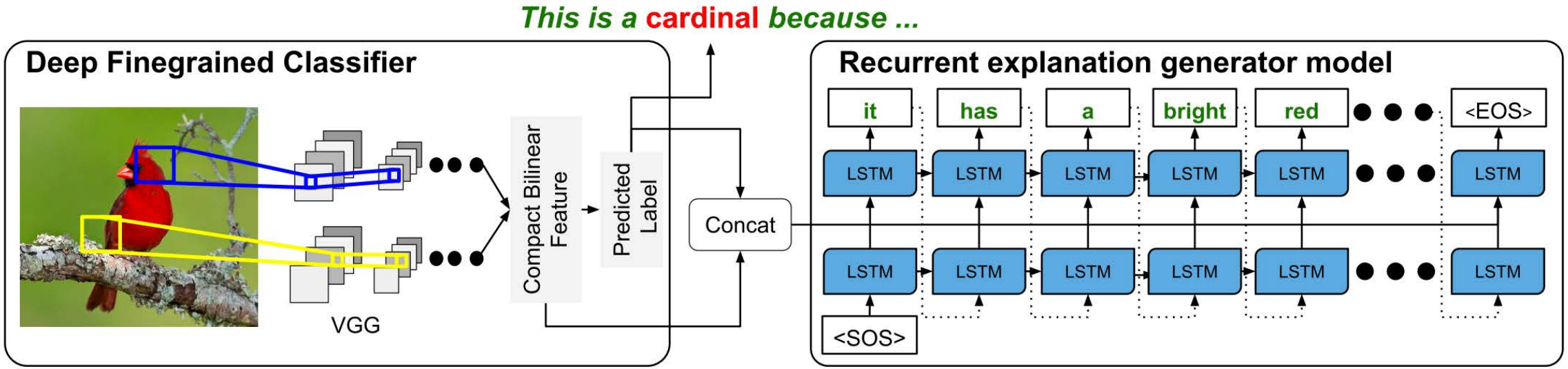




卷积核对应的概念仍需要人工标注，可能存在偏见！

Zhang, Quanshi, et al. "Interpreting CNN knowledge via an explanatory graph." *AAAI*. Vol. 32. No. 1. 2018.

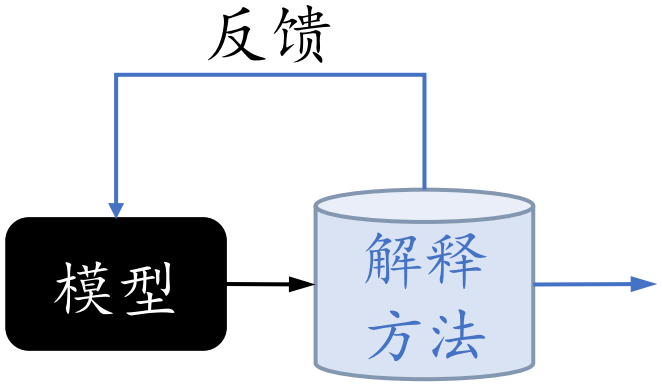
通过其他模型辅助解释



用一个不可解释的模型来解释一个黑盒模型是令人担忧的。

Hendricks, Lisa Anne, et al. "Generating visual explanations." *ECCV*, 2016.

通过可解释方法Debug并Enhance模型



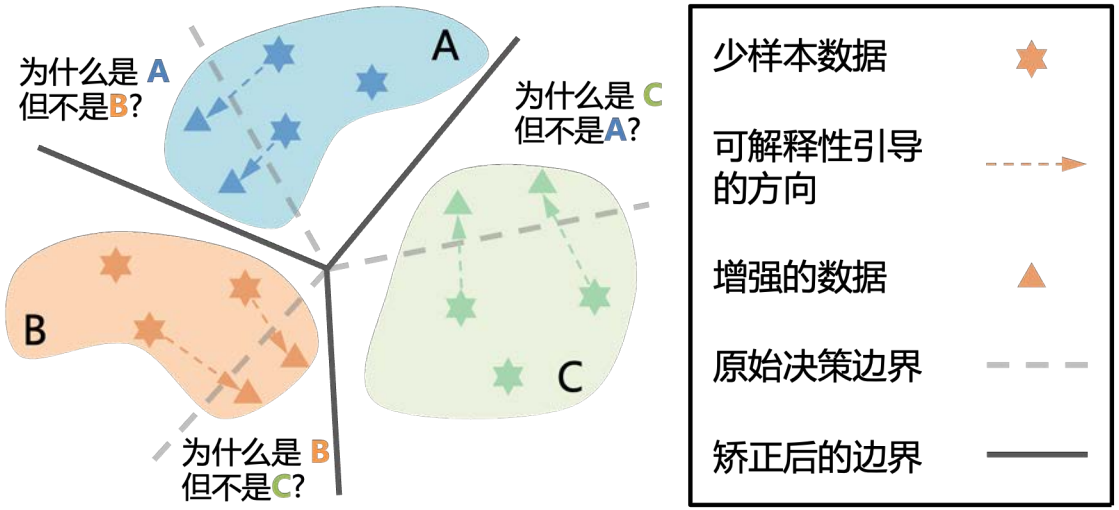
- 因果干预可以对模型去偏，但是需要人工先验辅助模型，如果通过可解释方法在无人干预的情况下自动修正模型？
- 根据特定任务，构建特定可解释方法解释特定的内容，并将内容反馈给模型。

通过可解释方法Debug并Enhance模型

Evaluation Metric

以小样本目标检测为例：

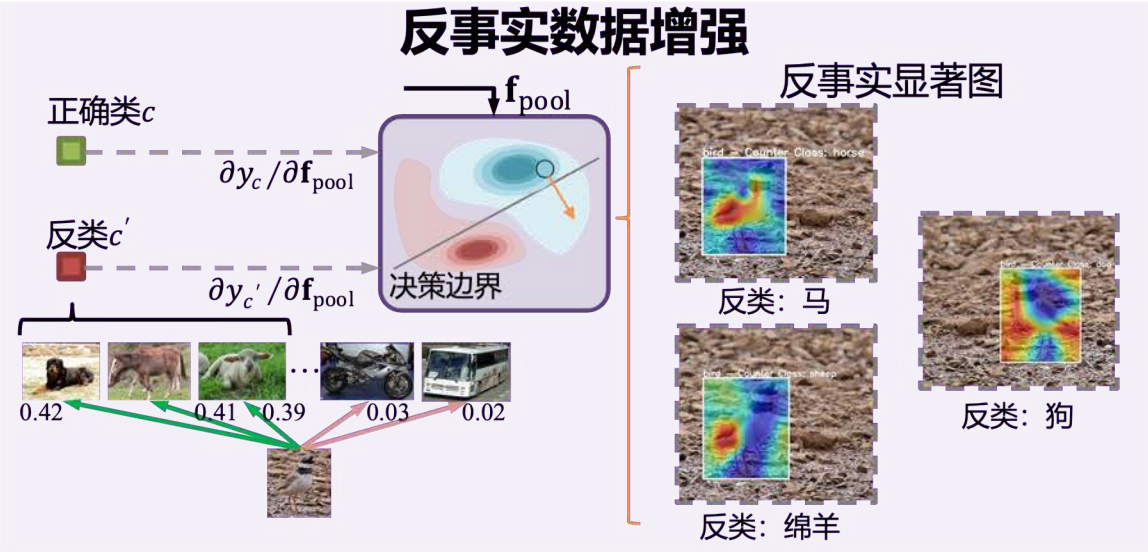
- **分析任务存在的问题：**小样本学习的新类数据量很少，相似类别之间的可分性很差；样本量不足以覆盖标准分布，可能存在过拟合问题。
- **设计何种解释方法：**针对相似类别间可分性差，是因为模型学习时候存在偏见，需要设计可解释方法，解释训练时模型为什么判别A而不是B，其中A和B是相似类别。
- **解释结果如何反馈：**假设可解释方法发现的内容是模型可能存在的偏见，如果是，则去除偏见。即使不是偏见，反馈也不应该对模型造成影响。

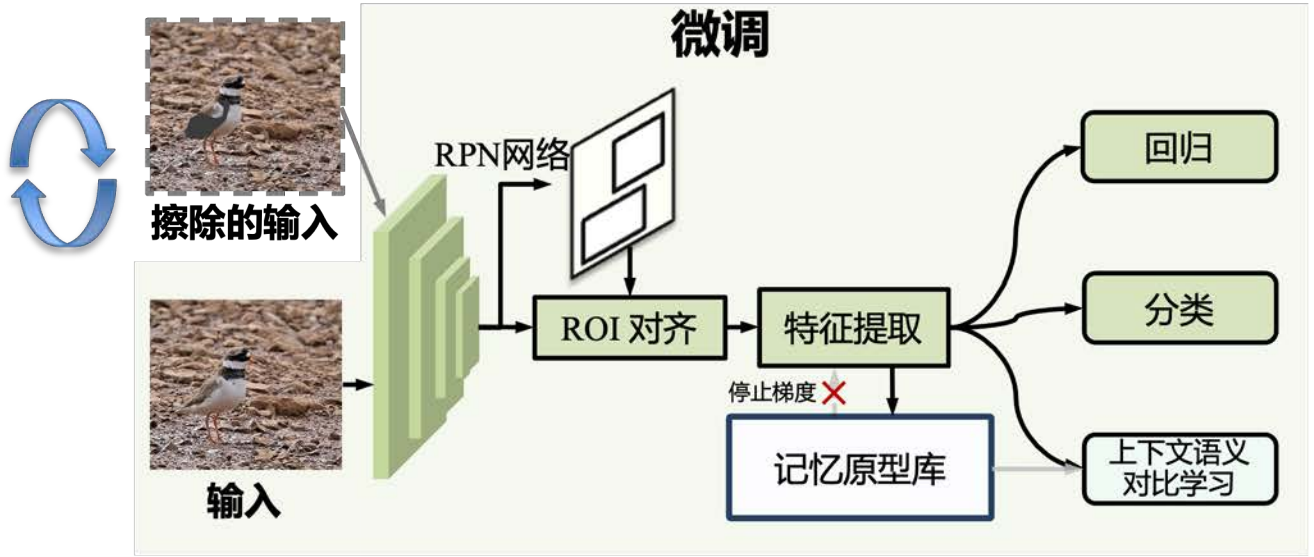


Ruoyu Chen, Hua Zhang, Jingzhi Li, Li Liu, and Xiaochu Cao. "Generalized Semantic Contrastive Learning via Embedding Side Information for Few-Shot Object Detection." Submitted to *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

通过可解释方法Debug并Enhance模型

Evaluation Metric





Ruoyu Chen, Hua Zhang, Jingzhi Li, Li Liu, and Xiaochu Cao.
"Generalized Semantic Contrastive Learning via Embedding Side Information for Few-Shot Object Detection." Submitted to *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

SYSU 2023
PRESENTATION SLIDE

中山大學
SUN YAT-SEN UNIVERSITY

中国科学院信息工程研究
INSTITUTE OF INFORMATION ENGINEERING

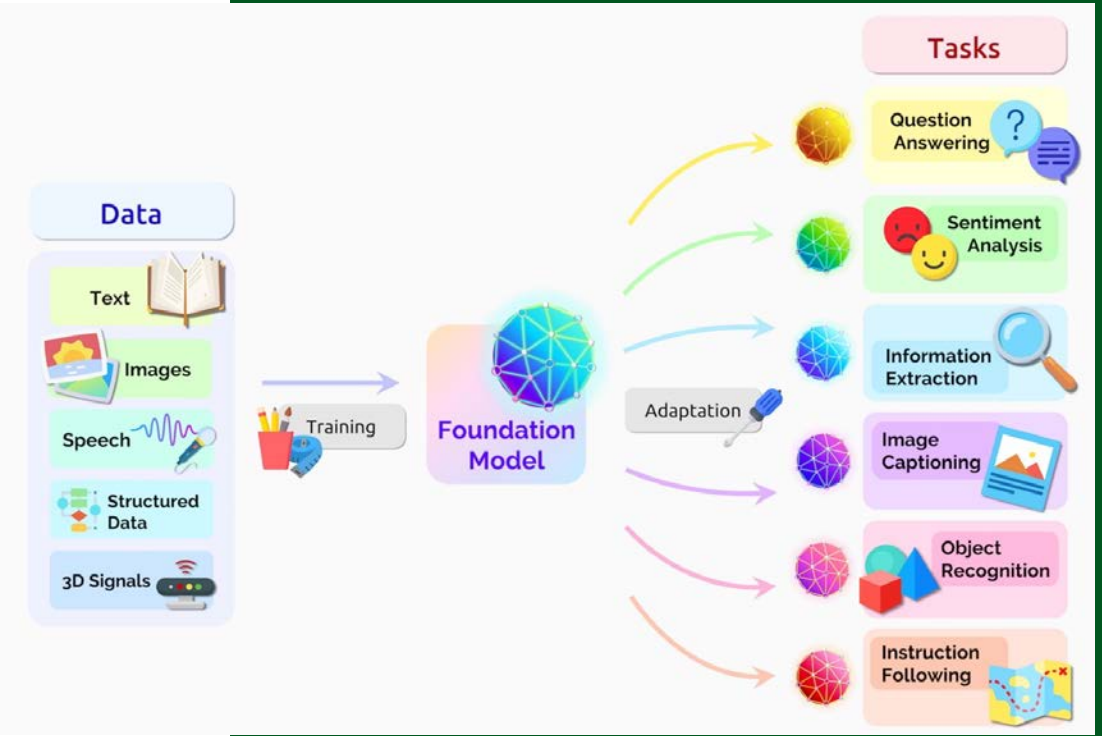
07

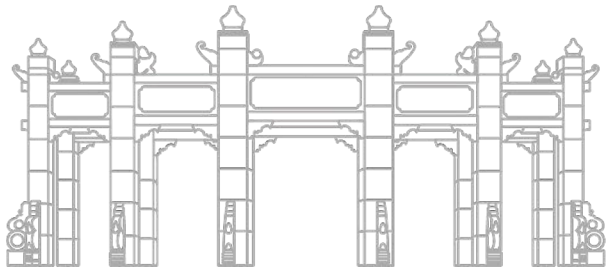
基础模型的可解释

Interpretability of the foundation model

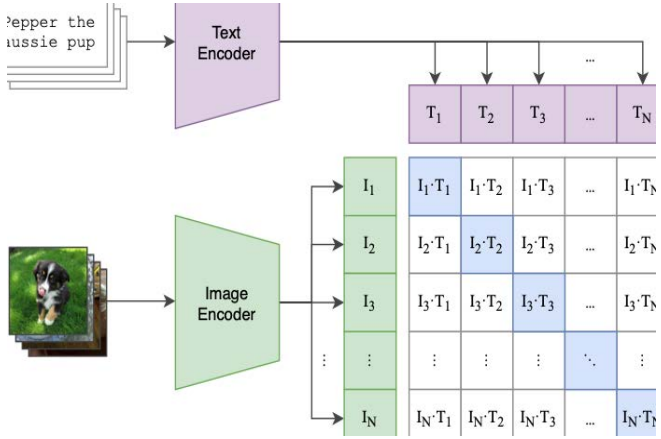
完整版PowerPoint请见：

https://ruoyuchen10.github.io/talk/Ruoyu_Chen-Interpretation_of_foundation_model.pdf

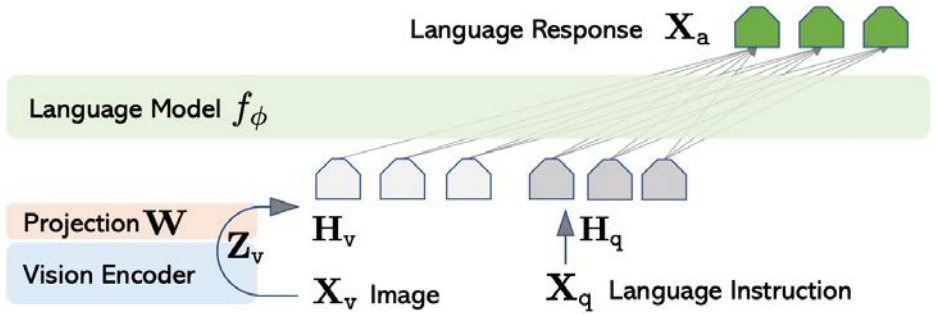




基础模型的特点及其新挑战



多模态编码式基础模型



多模态问答式基础模型

大语言模型

特点

- ✓ 双流结构
- ✓ 编码式模型
- ✓ Zero-Shot能力

- ✓ 参数量较大
- ✓ 编码+生成式模型
- ✓ 双模态输入
- ✓ 提示学习能力

- ✓ 生成式模型
- ✓ 参数量非常大
- ✓ 内部结构非常复杂

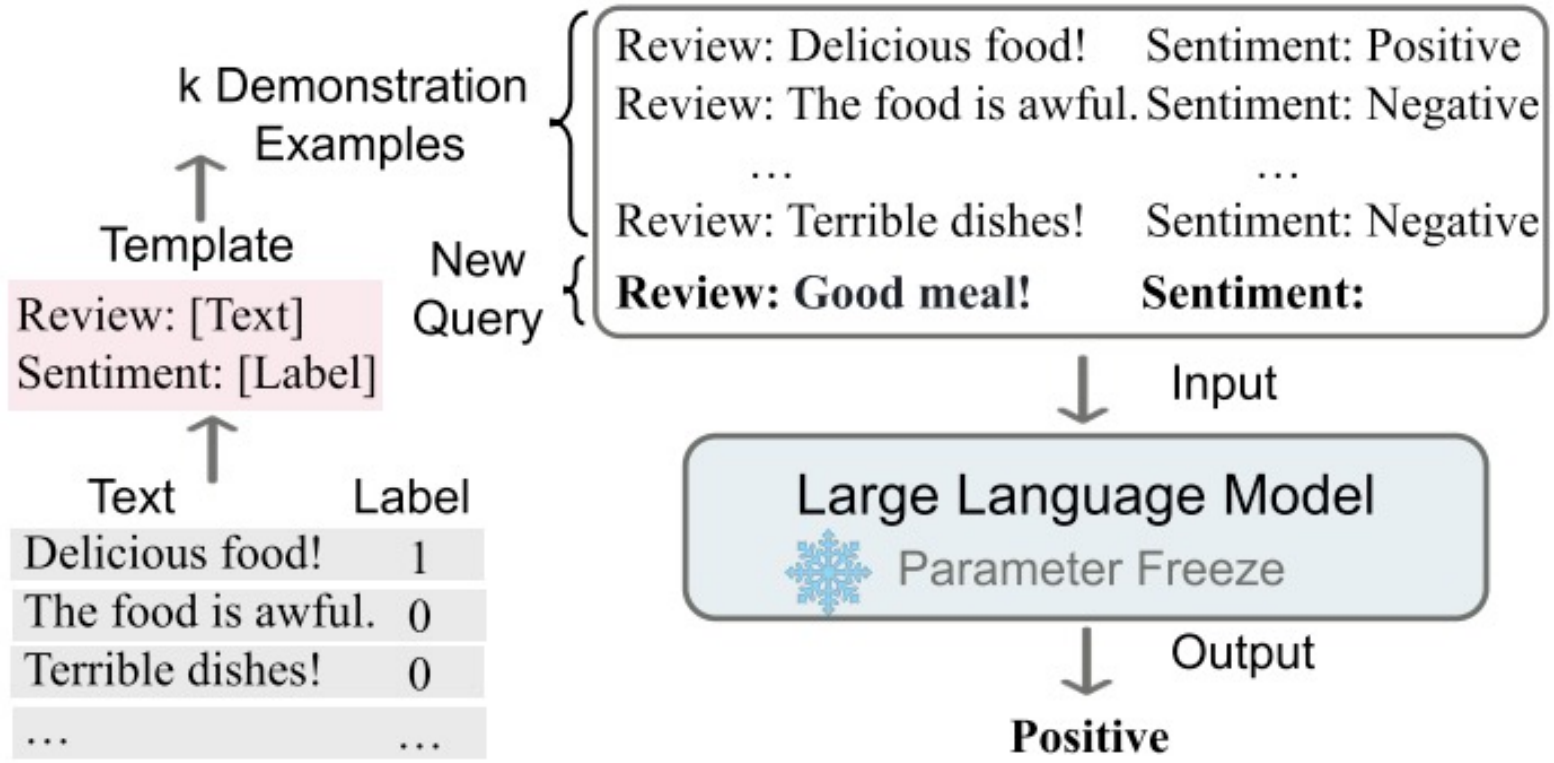
传统方法缺陷

- 传统的解释方法大多只针对单一模态的模型，他们可能并不适合解释处理多模态输入的模型。
- 传统方法缺少考虑多模态模型特有的性质，例如多模态模型与文本高度关联，增强人类理解。
- ViT和CNN的解释方法不通用！

- 模型依赖对话历史和当前的多模态信息，这种依赖性为解释模型行为添加了额外的复杂性。
- 模型可能会忽略某些信息，如何选择、提示信息对解释很重要。
- 以何种方式解释？不同推理框架？不同模态信息组成？新的设计方法？

- 由于参数的复杂交互，传统的可视化和解释工具无法提供关于这些交互和内部处理的清晰视图。
- 训练数据量大，不易理解数据中提取模型的领域知识与偏好。

大语言模型的可解释性

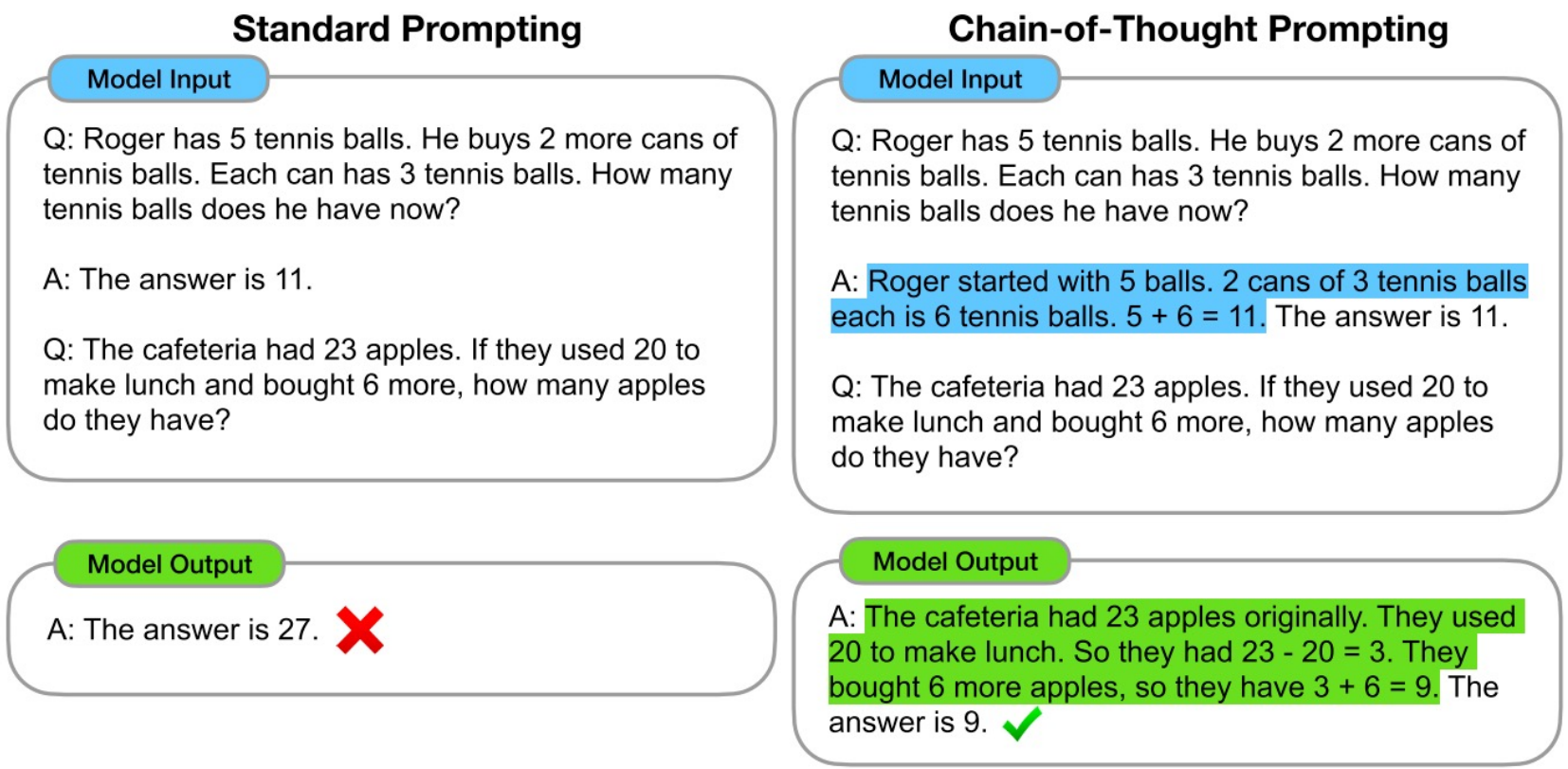


上下文学习的释例。

Dong, Qingxiu, et al. "A survey for in-context learning." *arXiv preprint arXiv:2301.00234* (2022).

思维链 Chain-of-thought (CoT)

谷歌Brain团队



思维链提示使大型语言模型能够处理复杂的算术、常识和符号推理任务。强调了思维链推理过程。

In-context few-shot learning via *prompting*:

<input, *chain-of-thought*, output>

特性：

1. 思维链原则上允许模型将多步骤问题分解为中间步骤；
2. 为模型的行为提供了一个**可解释的窗口**，表明如何得出特定的答案，并提供调试推理路径出错的地方的机会；
3. 可能适用于（至少原则上）人类可通过语言解决的任何任务。
4. 在**足够大**的语言模型中，只要将思维链序列的示例包含到少数提示的示例中，就可以很容易地推导出思维链推理。

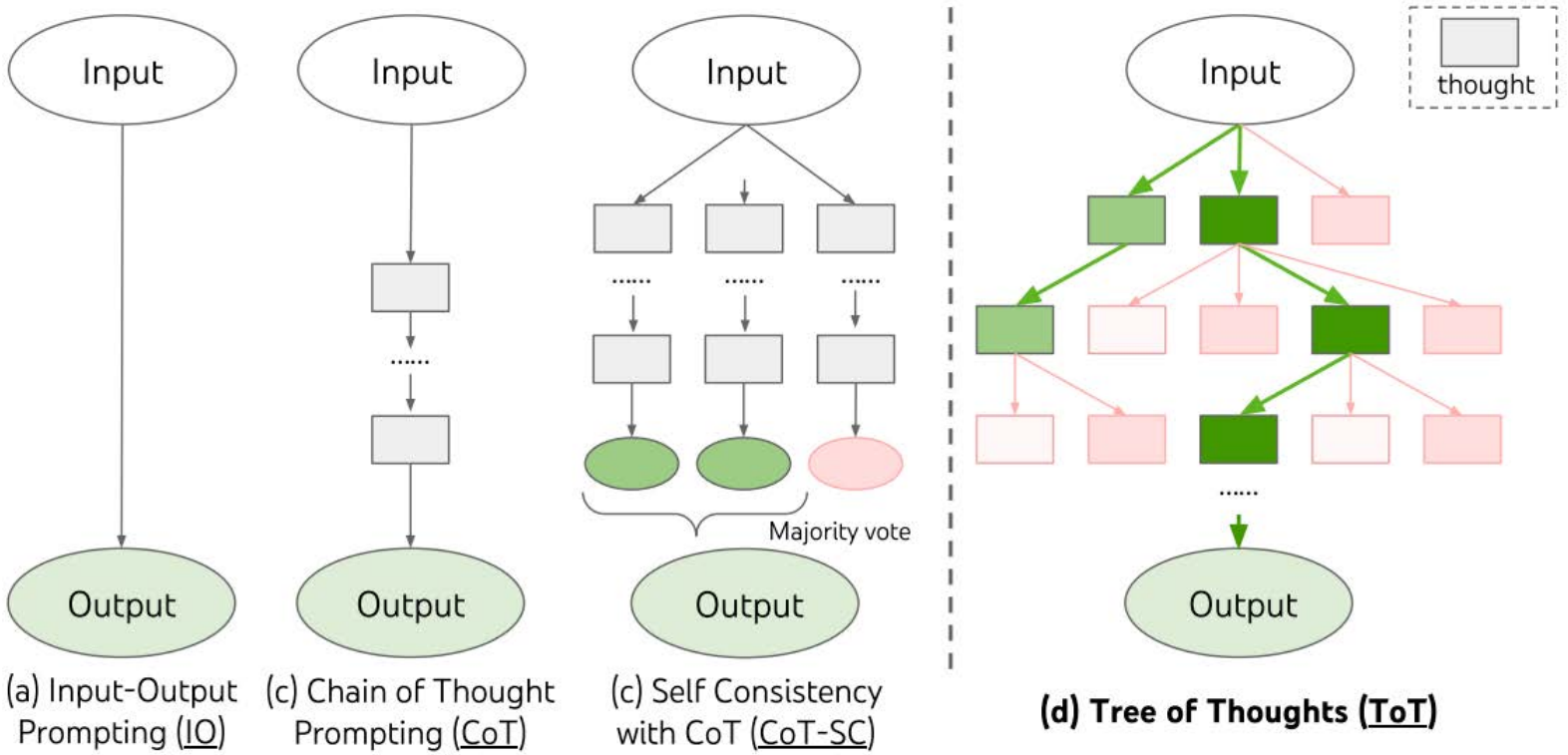
注：尽管完全表征一个模型的计算支持一个答案仍然是一个悬而未决的问题。

Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." *NeurIPS* 35 (2022): 24824-24837.

Tree-of-thought (ToT)

真正的解决问题的过程涉及重复使用可用信息来启动探索，进而揭示更多信息，直到最终找到解决方案的方法。

——Allen Newell *et al.*



ToT步骤：

1. Thought decomposition 思维分解
2. Thought generator 思维生成
   a) Sample 采样
   b) Propose 建议
3. State evaluator 状态评估
   a) Value 价值
   b) Vote 投票
4. Search algorithm 搜索算法
   a) 广度优先
   b) 深度优先

Yao, Shunyu, et al. "Tree of thoughts: Deliberate problem solving with large language models." *arXiv preprint arXiv:2305.10601* (2023).

分解大语言模型

2023.10.4 Claude背后公司Anthropic发布Poster：
Towards Monosemanticity: Decomposing Language Models With Dictionary Learning
使用稀疏自编码器，从一个单层Transformer中提取了大量的可解释特征。

问题：对语言模型来说，它的不可解释性主要体现在网络中的大多数神经元都是**“多语义的”**。

一个潜在的因素是“叠加” (superposition)，指的是模型将许多不相关的概念全部压缩到一个少量神经元中的操作。

团队又采用了一种称为**稀疏自动编码器的弱字典学习算法**。在神经网络激活上使用字典学习的相关方法，以解耦 (disentanglement) 相关的内容。

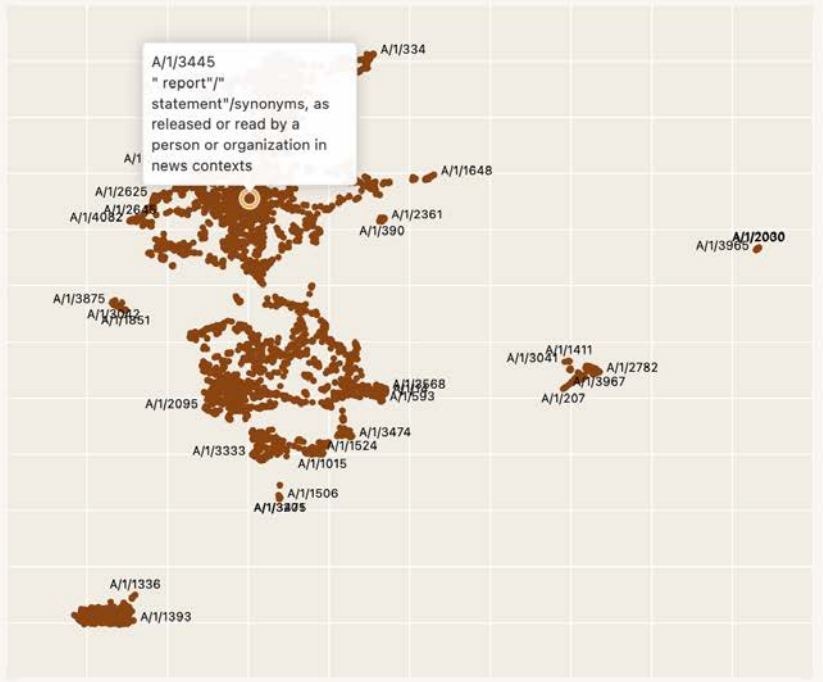
Trenton Bricken, *et al.*, "Towards Monosemanticity: Decomposing Language Models With Dictionary Learning." https://transformer-circuits.pub/2023/monosemantic-features. 2023.

分解大语言模型

2023.10.4 Claude背后公司Anthropic发布Poster：
Towards Monosemanticity: Decomposing Language Models With Dictionary Learning
使用稀疏自编码器，从一个单层Transformer中提取了大量的可解释特征。

Anthropic采用一个具有512个神经元的MLP单层Transformer，通过在具有80亿个数据点的MLP激活上训练稀疏自动编码器，最终将MLP激活分解为**相对可解释的特征**，扩展因子范围可以从1x（512个特征）增长到256x（131072个特征）。

Cluster #49	A/0/307	This feature fires for references to citations in scientific pa...
	A/0/311	This feature fires for reference citations in academic paper...
	A/1/776	Years in some citation notation
	A/1/1538	Citations in a [@author] or [@authoryear] format
	A/1/1875	Markdown Citation (Predict year)
	A/1/2252	" [@"
	A/1/2237	[Ultralow density cluster]
Cluster #42	A/0/126	This feature seems to fire on section headings, specifically ...
	A/1/357	"ref" in [context]
	A/1/1469	"s"/"sec" after "{#", section reference in some markup
	A/1/3841	"Sec"
	A/1/3898	Section number in {#SecX}
	A/1/4083	" {#"
	A/1/2129	"." in [context]
	A/1/553	"]{#" in [context]
Cluster #43	A/0/8	This feature attends to text formatting markups such as ref...
	A/0/398	This feature attends to references to figures and tables.
	A/0/454	This feature fires on reference/bibliographic citations in LaT...
	A/1/35	"){"
	A/1/366	"type"
	A/1/945	"ref" in [context]
	A/1/1895	"-" in [context]
	A/1/2176	"fig"

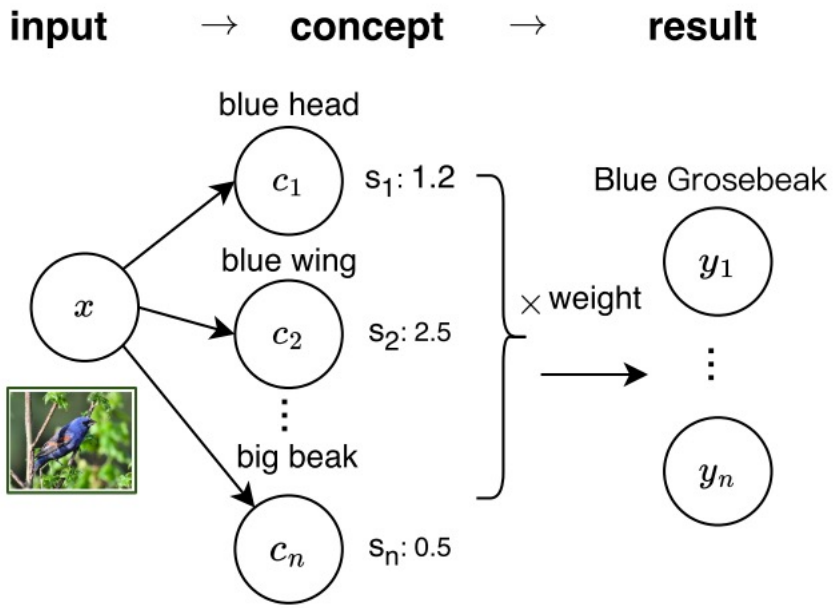


Trenton Bricken, *et al.*, "Towards Monosemanticity: Decomposing Language Models With Dictionary Learning." https://transformer-circuits.pub/2023/monosemantic-features. 2023.

总结

- 如何利用大语言模型LLM的In-Context Learning的特性，设计更合理的推理框架？
- 如何设计更合理的因果图以解释LLM的决策？
- 如何解释大语言模型内部逻辑？解释什么？解释的结果有什么作用？

多模态编码式基础模型的可解释性

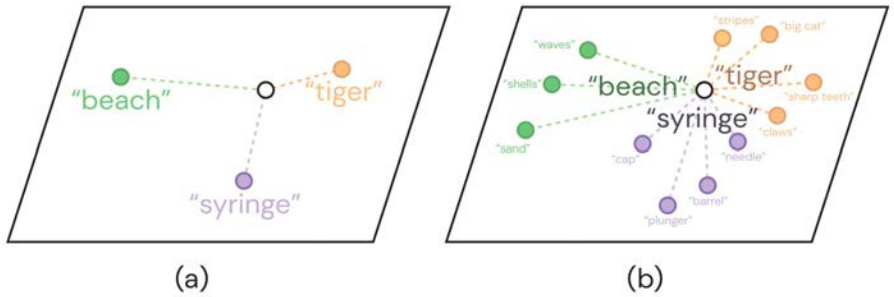


传统的概念瓶颈模型

缺点：

- 如何确定语义概念集？
- 需要人工密集的语义标注。

多模态编码式基础模型的可解释性



(a) (b)

挖掘大型语言模型来自动构建描述符

School bus
- large, yellow vehicle
- the words "school bus" written on the side
- a stop sign that deploys from the side of the bus
- flashing lights on the top of the bus
- large windows

Shoe store
- a building with a sign that says "shoe store"
- a large selection of shoes in the window
- shoes on display racks inside the store
- a cash register
- a salesperson or customer

Volcano
- a large, cone-shaped mountain
- a crater at the top of the mountain
- lava or ash flowing from the crater
- a plume of smoke or ash rising from the crater

Barber shop
- a building with a large, open storefront
- a barber pole or sign outside the shop
- barber chairs inside the shop
- mirrors on the walls
- shelves or cabinets for storing supplies
- a cash register
- a waiting area for customers

Cheeseburger
- a burger patty
- cheese
- a bun
- lettuce
- tomato
- onion
- pickles
- ketchup
- mustard

Violin
- a stringed instrument
- typically has four strings
- a wooden body
- a neck and fingerboard
- tuning pegs
- a bridge
- a soundpost
- f-holes
- a bow

Pirate ship
- a large, sailing vessel
- a flag with a skull and crossbones
- cannons on the deck
- a wooden hull
- portholes
- rigging
- a crow's nest

GPT-3生成的描述符模式示例。

Architecture for ϕ		ImageNet Ours	ImageNet CLIP	ImageNet Δ	ImageNetV2 Ours	ImageNetV2 CLIP	ImageNetV2 Δ	CUB Ours	CUB CLIP	CUB Δ
Vision Transformers	ViT-B/32	**62.97**	58.46	4.51	**55.52**	51.90	3.62	**52.57**	51.95	0.62
	ViT-B/16	**68.03**	64.05	3.98	**61.54**	57.88	3.66	**57.75**	56.35	1.40
	ViT-L/14	**75.00**	71.58	3.42	**69.3**	65.33	3.97	**63.46**	63.08	0.38
	ViT-L/14@336px	**76.16**	72.97	3.19	**70.32**	66.58	3.74	**65.257**	63.41	1.847
ResNets	RN50	**59.44**	54.81	4.63	**52.98**	49.43	3.55	**48.91**	47.79	1.12
	RN101	**61.88**	57.65	4.23	**55.43**	51.13	4.30	**51.59**	49.46	2.13
	RN50x4	**66.05**	61.48	4.27	**59.23**	54.85	4.38	**55.97**	54.99	0.98
	RN50x16	**69.45**	66.28	3.17	**62.68**	58.8	3.88	**59.03**	57.59	1.44
	RN50x64	**73.19**	69.63	3.56	**66.82**	63.02	3.80	**64.62**	64.24	0.38

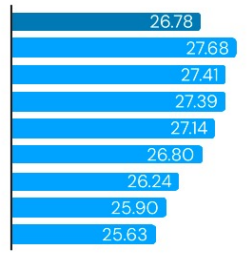
ImageNet和ImageNetV2的模型有一致的~ 3-5%的改进，CUB有~ 1%的改进。

$$s(c,x) = \frac{1}{|D(c)|} \sum_{d \in D(c)} \phi(d,x)$$

Our top prediction: Hen

and we say that because...

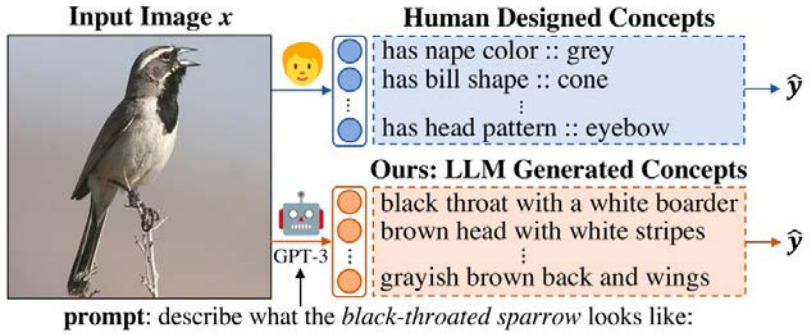
Average
- two legs
- red, brown, or white feathers
- a small body
- a small head
- two wings
- a tail
- a beak
- a chicken



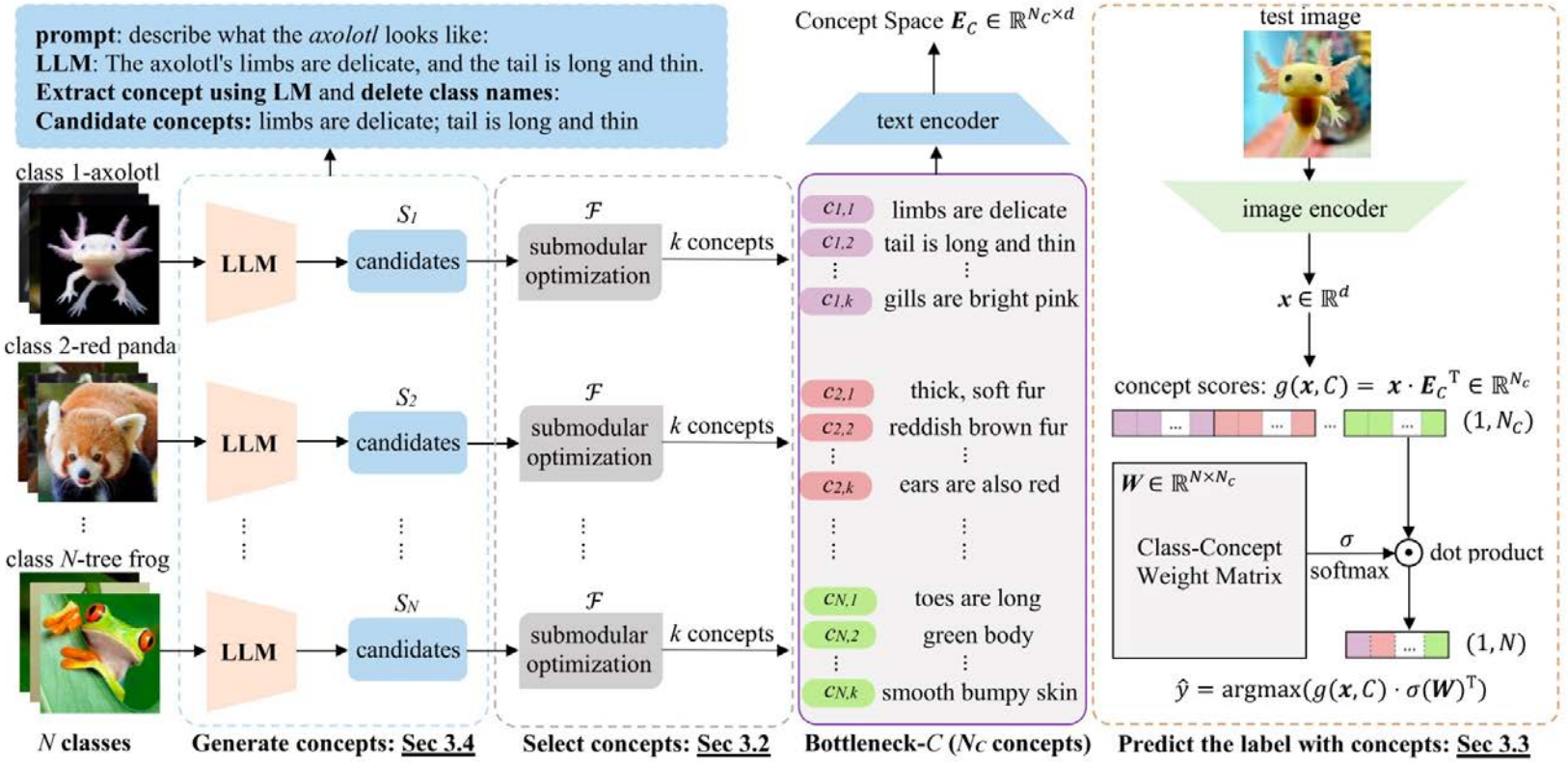
CLIP通过描述符进行决策。

[19] Menon, Sachit, and Carl Vondrick. "Visual Classification via Description from Large Language Models." *ICLR*. 2023.

多模态编码式基础模型的可解释性

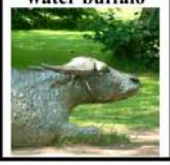


通过提示大型语言模型（LLM）（如GPT-3)减轻了对人工设计概念的需求



Yang, Yue, et al. "Language in a bottle: Language model guided concept bottlenecks for interpretable image classification." *CVPR*. 2023.

多模态编码式基础模型的可解释性

	Class Name	Top-3 Concepts	Class Name	Top-3 Concepts	Class Name	Top-3 Concepts	Class Name	Top-3 Concepts
ImageNet	badger 	1.short legs and long body make it an excellent digger 2. black-and-white striped fur 3. coat is very shaggy	ant 	1. black and red stinger 2. small, black insect with six legs 3. long, slender antennae that it uses to smell and touch	hammer 	1.long, thin tool with a wooden handle 2. great tool for pounding object 3. used to pound on surfaces	water buffalo 	1. large head with short, curved horns 2. heaviest living species of bovid 3. huge, dark-colored animal
CUB	eared grebe 	1. black and white plumage that is striking in the sunlight 2. black body with a long, slender neck 3. red and black bill	horned lark 	1. black line running through yellow face 2. head is black with a white horn on each side 3. black horn on each side of their head	white pelican 	1. long neck and bill make it look like a giant swan 2. large, white bird with black wingtips 3. bill is huge and yellow	arctic tern 	1. breeds in greenland, iceland, and northern russia 2. black and white markings 3. small white bird with a black cap
Flower	water lily 	1. depicted in artworks of ponds and waterfall 2. member of the nymphaeaceae family 3. lily pads float	barbeton daisy 	1. scientific name for the flower is taraxacum officinal 2. named after the city of barberton 3. member of the daisy family	marigold 	1. central disc with smaller florets 2. have a slightly furry texture 3. bold and vibrant color palette	tiger lily 	1. long, protruding stamen 2. orange with black spots and stripes 3. scientific name is lilium columbianum
UCF-101	archery 	1. grip bow tightly in their left hand 2. focused and concentrated on their task 3. keep bow and arrows in safe and dry place when not in use	drumming 	1. blur as they fly over the drums 2. sitting on a stool in front of a drum set 3. position the drumstick so it is resting on your index finger	surfing 	1.deep blue color 2. tans contrast with the white of their boards 3. sending a spray of water into the air	long jump 	1. get out of sandpit 2. series of movements starting from a standing position 3. person tucks their knees up to chest
HAM10000	dermatofibroma 	1. generally not painful 2. red, brown, or purple in color 3. thin white halo around them	melanoma 	1. dark brown or black in color 2. large and dark 3. flesh-colored, brown, or black	melanocytic nevi 	1. color is tan 2. dark brown or black color 3. small, round, and slightly raised	benign lesions 	1. color ranges from light brown to black 2. rough or scaly texture 3. darker in color, such as brown or black

生成的概念示意图

Yang, Yue, et al. "Language in a bottle: Language model guided concept bottlenecks for interpretable image classification." *CVPR*. 2023.

总结

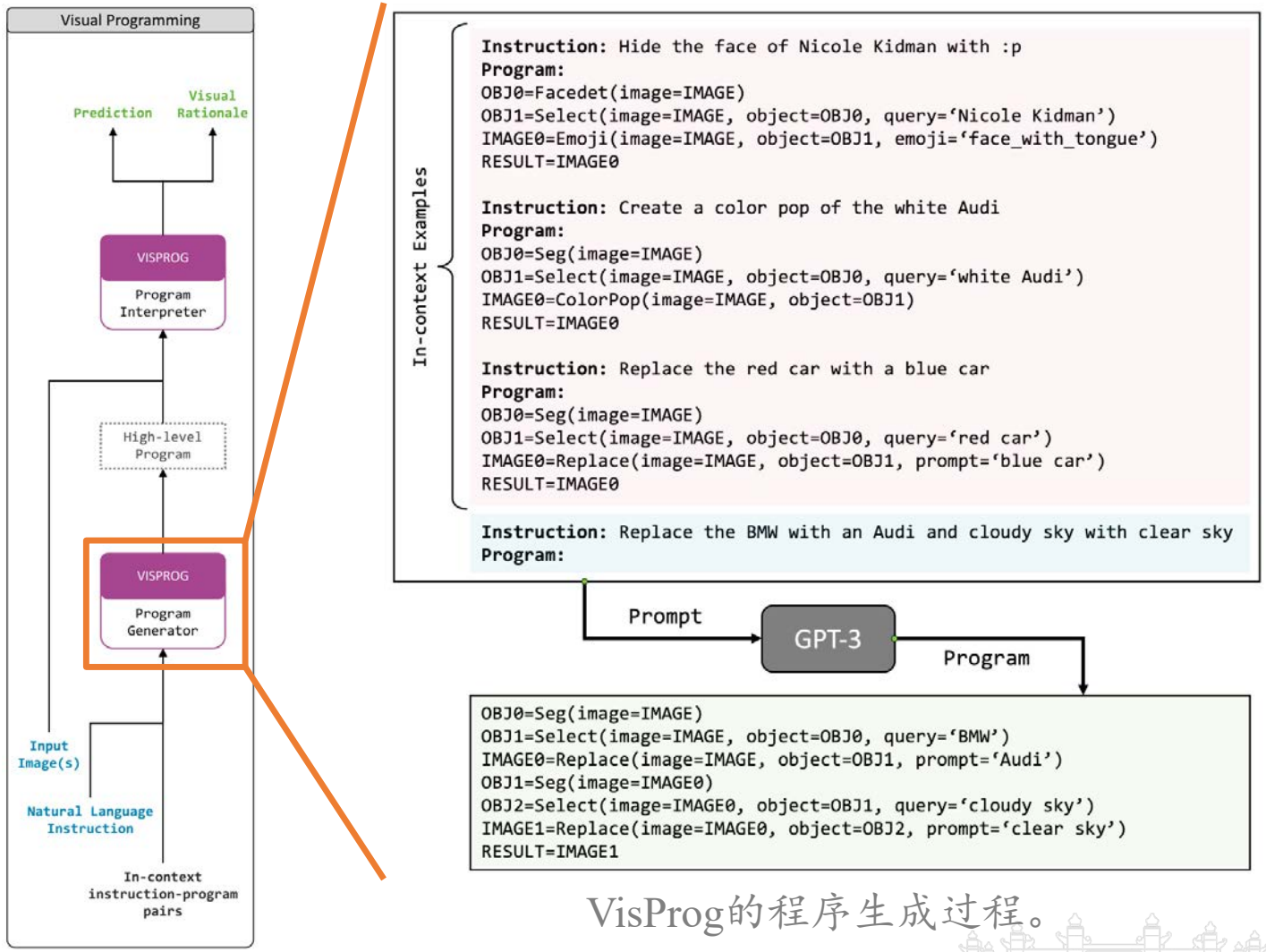
- 如何利用多模态编码基础模型的特性，利用人类容易理解的文本描述来辅助解释？
- 如何理解多模态基础模型内部的运行机理？一些假设是否是正确的？或者是不是应该这样被理解？
- 如何构建统一的因果图模型，以应对大模型中巨大的参数量与消耗的参数推理的挑战？
- 如何分解特征，以帮助人类的理解？
- 如何设计一个更方便可解释的模型结构，同时适应在大模型训练数据量巨大的情况。

多模态问答式基础模型的可解释性-VisProg

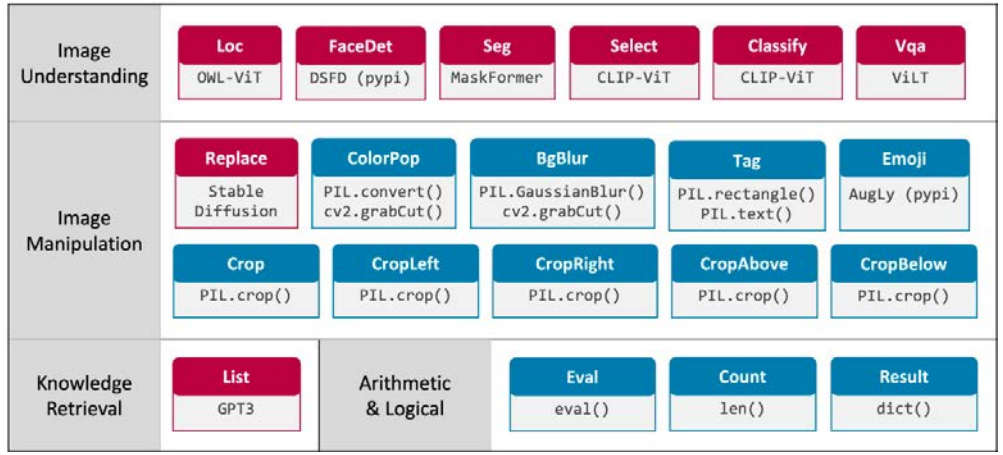
VisProg是一个模块化和可解释的神经符号系统，用于组合视觉推理。

[28] Gupta, Tanmay, and Aniruddha Kembhavi. "Visual programming: Compositional visual reasoning without training." *CVPR*. 2023. (Best Paper)

多模态问答式基础模型的可解释性-VisProg



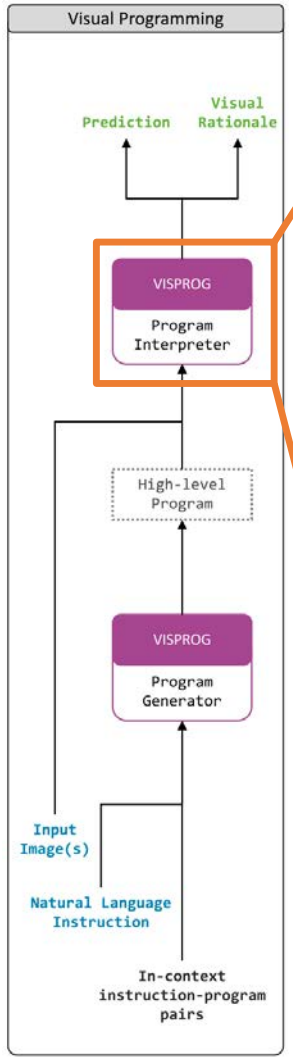
VisProg的程序生成过程。

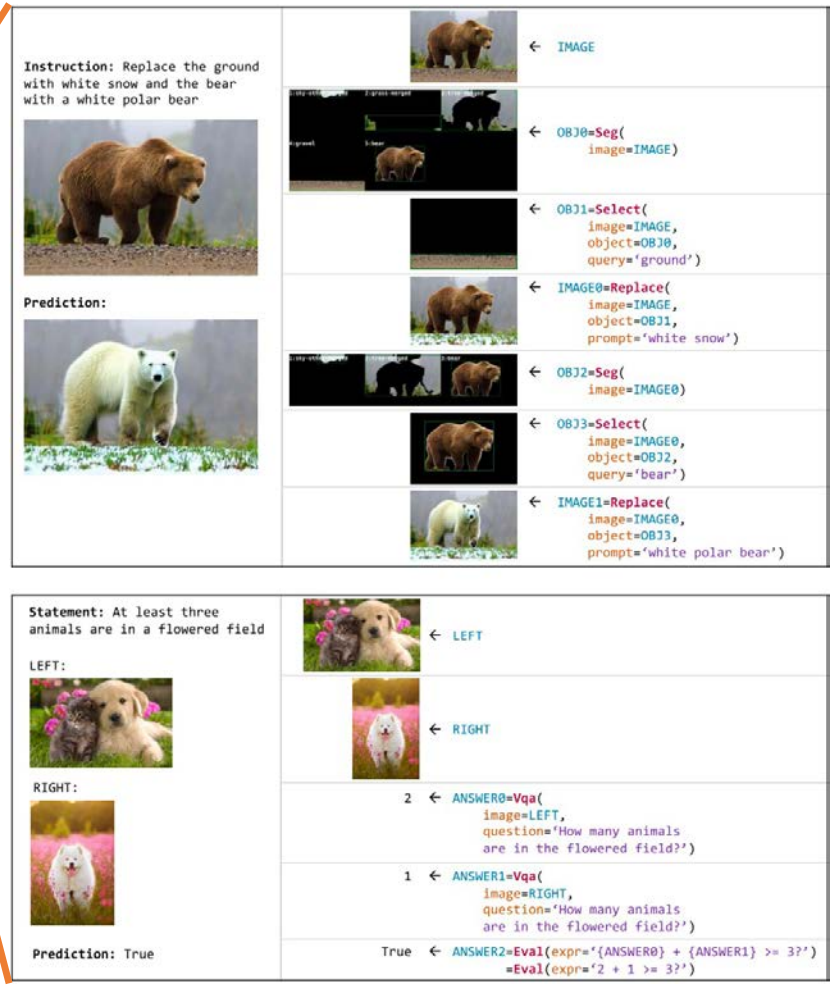


VisProg已有的所支持的功能模块。

[28] Gupta, Tanmay, and Aniruddha Kembhavi. "Visual programming: Compositional visual reasoning without training." *CVPR*. 2023. (Best Paper)

多模态问答式基础模型的可解释性-VisProg





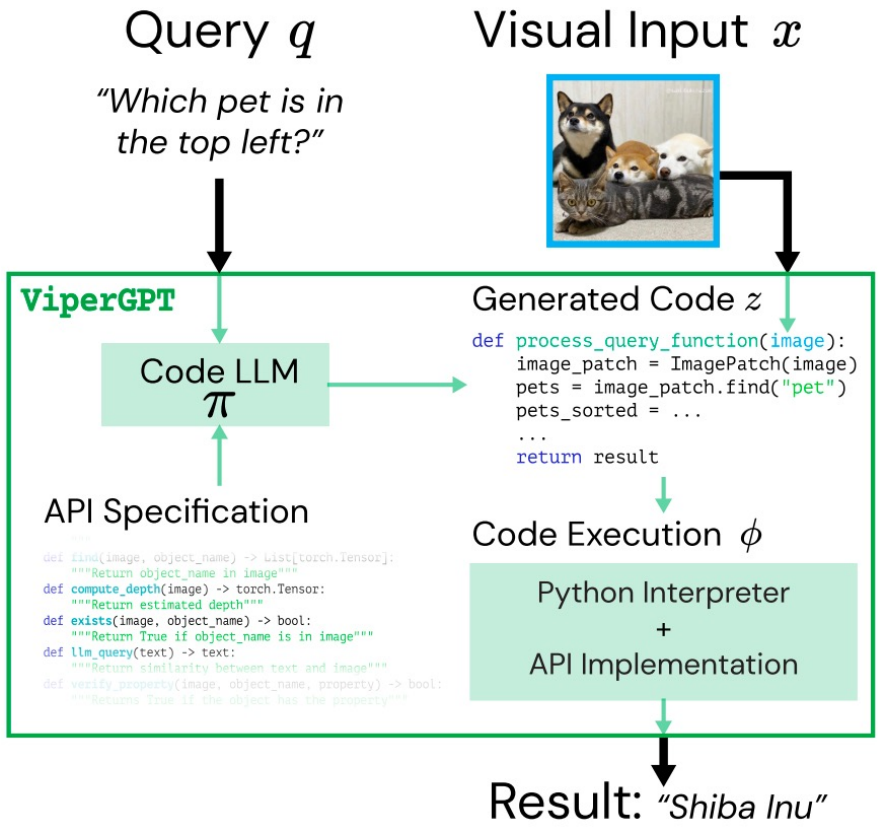
VisProg生成的视觉原理。

Task	Input	Output	Modules
Compositional Visual QA (GQA)	Image + Question	Text	Loc, Vqa, Eval, Count, Crop, CropLeft, CropRight, CropAbove, CropBelow
Reasoning on Image Pairs (NLVR)	Image Pair + Statement	True/False	Vqa, Eval
Factual Knowledge Object Tagging	Image + Instruction	Image	FaceDet, List, Classify, Loc, Tag
Image Editing with Natural Language	Image + Instruction	Image	FaceDet, Seg, Select, Replace, ColorPop, BgBlur, Emoji

在一系列不同的任务上评估VisProg。

[28] Gupta, Tanmay, and Aniruddha Kembhavi. "Visual programming: Compositional visual reasoning without training." *CVPR*. 2023. (Best Paper)

多模态问答式基础模型的可解释性-ViperGPT



ViperGPT是一个以编程方式解决复杂可视化查询的框架。

Surís, Dídac, Sachit Menon, and Carl Vondrick. "ViperGPT: Visual inference via python execution for reasoning." *ICCV* (2023).

多模态问答式基础模型的可解释性-ViperGPT

Surís, Dídac, Sachit Menon, and Carl Vondrick. "ViperGPT: Visual inference via python execution for reasoning." *ICCV* (2023).

总结

- 如何利用大语言模型（LLM）的特点辅助模型推理？
- 如何针对特定任务构建专家知识，以帮助模型更好的适应下游任务？
- 需要何种解释？以模型直接反馈推理过程？

SYSU 2023
PRESENTATION SLIDE

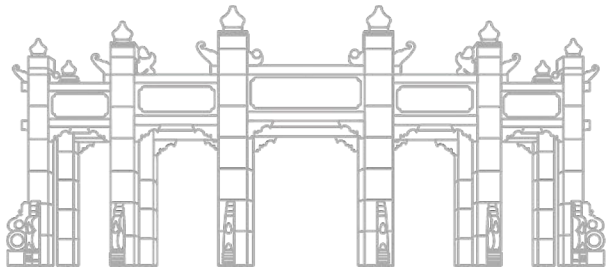
中山大學
SUN YAT-SEN UNIVERSITY

中国科学院信息工程研究
INSTITUTE OF INFORMATION ENGINEERING

08

总结与展望

Summary and Outlook



推荐可解释方法Survey

1 Arrieta, Alejandro Barredo, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." *Information fusion* 58 (2020): 82-115.

2 Minh, Dang, et al. "Explainable artificial intelligence: a comprehensive review." *Artificial Intelligence Review* (2022): 1-66.

3 Ahmed, Imran, Gwanggil Jeon, and Francesco Piccialli. "From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where." *IEEE Transactions on Industrial Informatics* 18.8 (2022): 5031-5042.

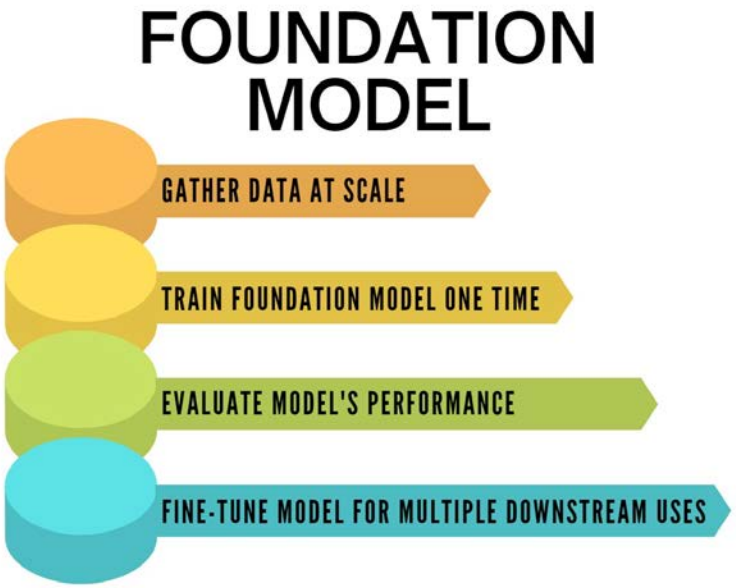
4 Deng, Huiqi, et al. "Understanding and Unifying Fourteen Attribution Methods with Taylor Interactions." *arXiv preprint arXiv:2303.01506* (2023).

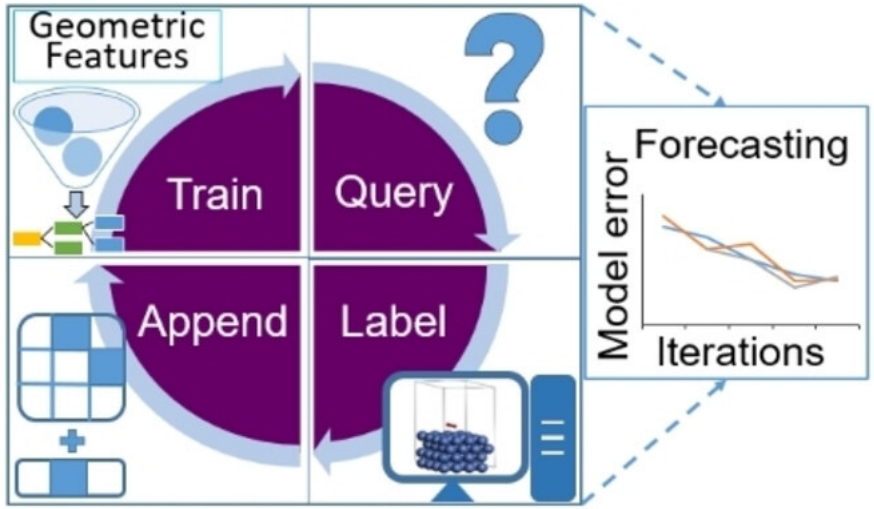
5 Nauta, Meike, et al. "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai." *ACM Computing Surveys* 55.13s (2023): 1-42.

6 Dwivedi, Rudresh, et al. "Explainable AI (XAI): Core ideas, techniques, and solutions." *ACM Computing Surveys* 55.9 (2023): 1-33.

7 Rong, Yao, et al. "Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

一些令人兴奋的方向





探索基础模型的可解释

- 设计Ante-Hoc可解释模型
- 怎么解释海量参数模型
- 解释数据集，哪些是脏数据
- 怎么融合人类的知识？

可解释如何增强模型性能

- 针对什么任务做解释？
- 如何设计合理的反馈机制？
- 下游任务怎么验证可解释合理？
- 训练阶段怎么引入？
- 测试阶段怎么引入？

Human-Centered可解释

- 如何研究人机交互？
- 如何人机对齐？
- 怎么验证合理性？
- 实验怎么做？用大语言模型模仿人类？

仍有很多未知的可解释方法！

可解释性仍是一个有争议的话题！

更多的方法值得我们去探索发现！

欢迎大家加入可解释人工智能的研究！

中山大學
SUN YAT-SEN UNIVERSITY

谢谢观看

SUN YAT-SEN UNIVERSITY 2023

陈若愚