

基础模型的可解释性

汇报人：陈若愚

日期：2023.10.20

提纲

一、传统可解释AI与基础模型

- 传统可解释AI方法简介
- 基础模型的特点及其新挑战

二、基础模型的可解释性研究现状

- 大语言模型的可解释性
- 多模态编码式基础模型的可解释性
- 多模态问答式基础模型的可解释性

三、我们思考的一些方法

一、传统可解释AI与基础模型

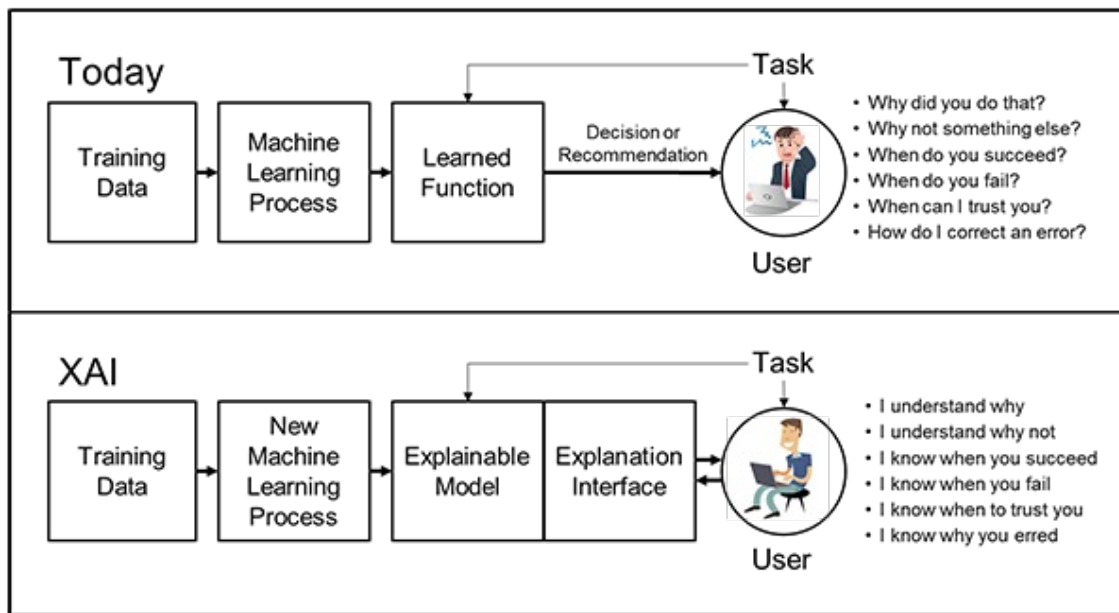
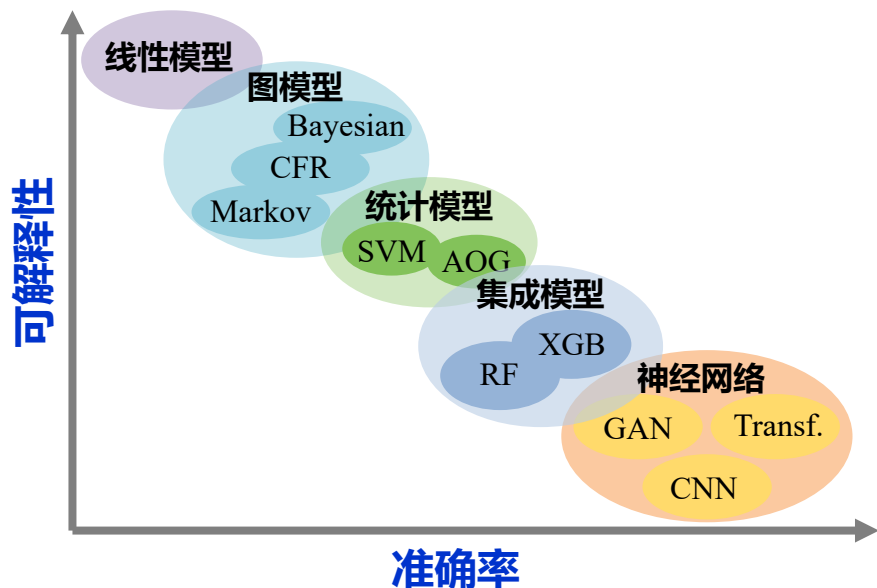
- 传统可解释AI方法简介
- 基础模型的特点及其新挑战

一、传统可解释AI与基础模型

- 传统可解释AI方法简介
- 基础模型的特点及其新挑战

可解释AI的重要性

可解释AI



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。



可解释的人工智能 (XAI) 计划^[1]:

- 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。

[1] Explainable Artificial Intelligence, <https://www.darpa.mil/program/explainable-artificial-intelligence>

可解释AI的相关术语

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)

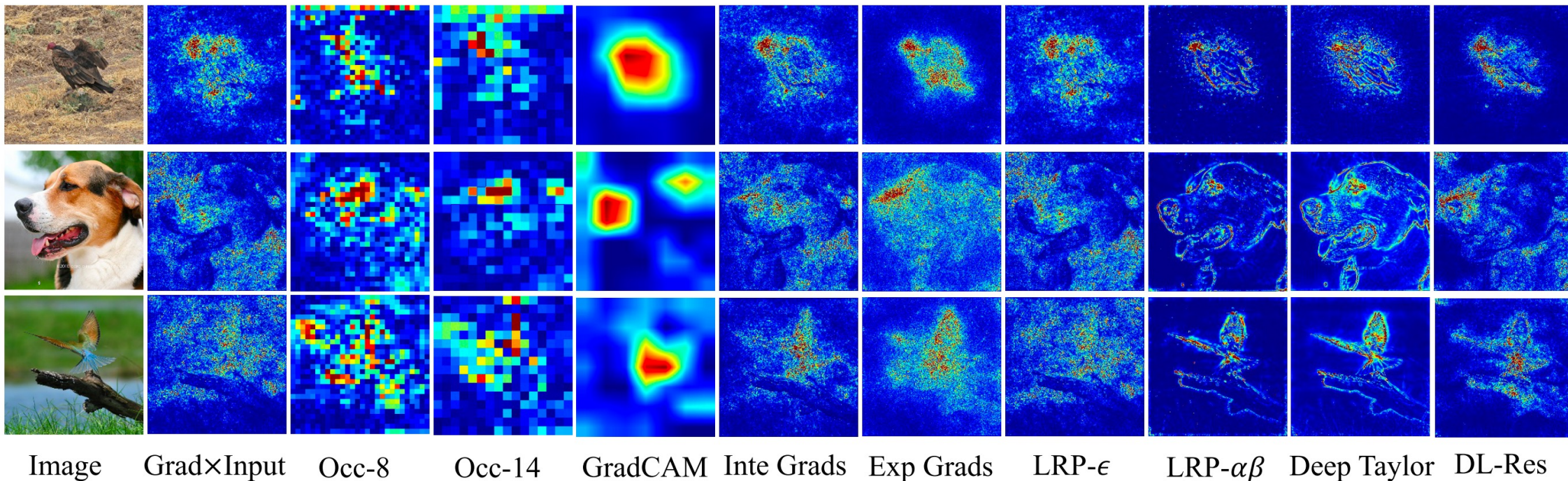
- 解释一个预训练模型或其决策的结果；
- 在模型做完决策后提供的解释。

传统可解释AI方法简介

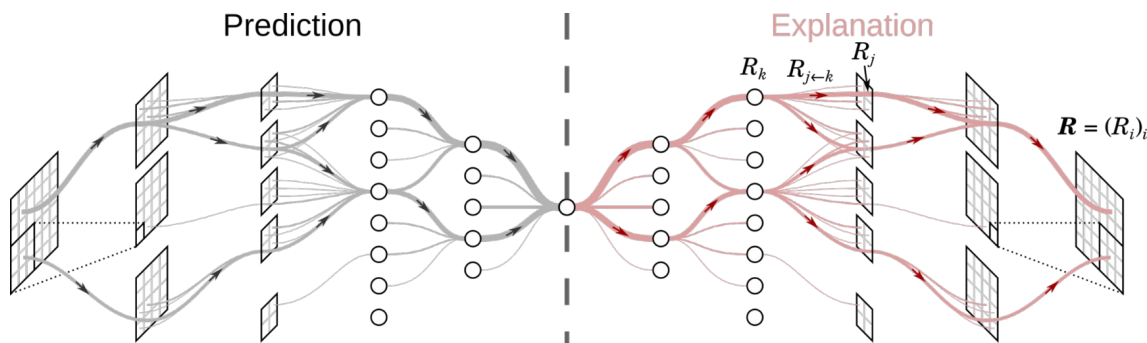
- 基于归因的方法
- 基于特征可视化的方法
- 基于语义概念的方法
- 基于关联模型的方法
- 基于其他模态模型的方法
- 基于设计的方法
- 基于因果的方法

传统可解释AI方法简介

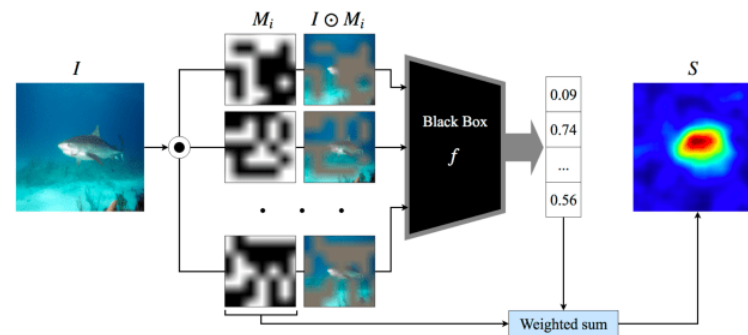
基于归因的方法



基于模型内部机理（白盒）

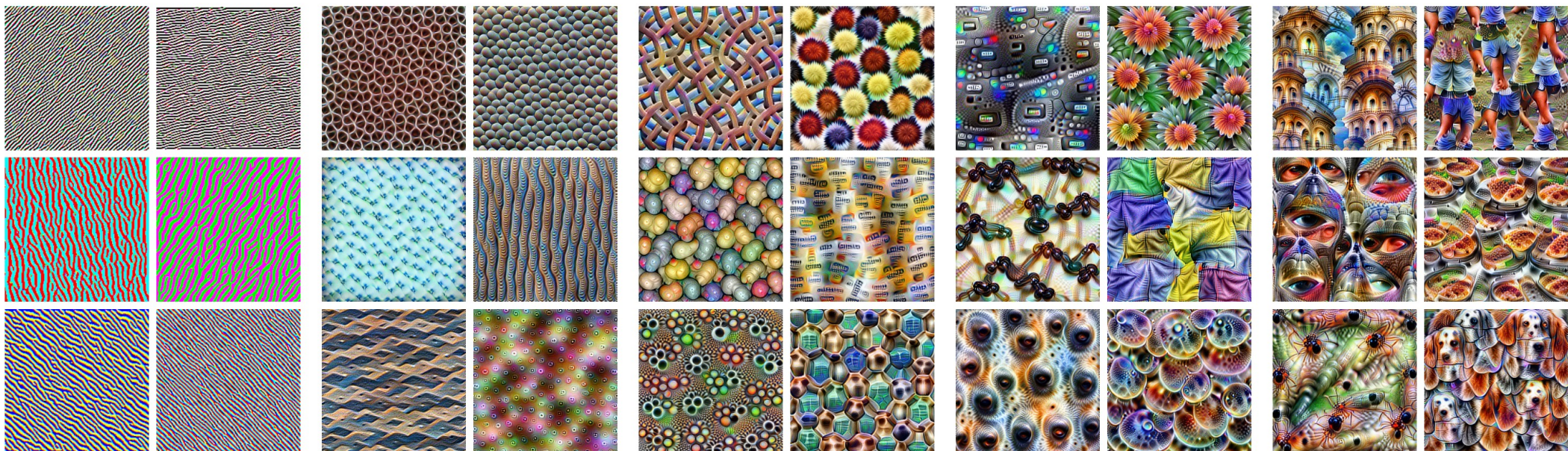


基于扰动（黑盒）



传统可解释AI方法简介

基于特征可视化的方法



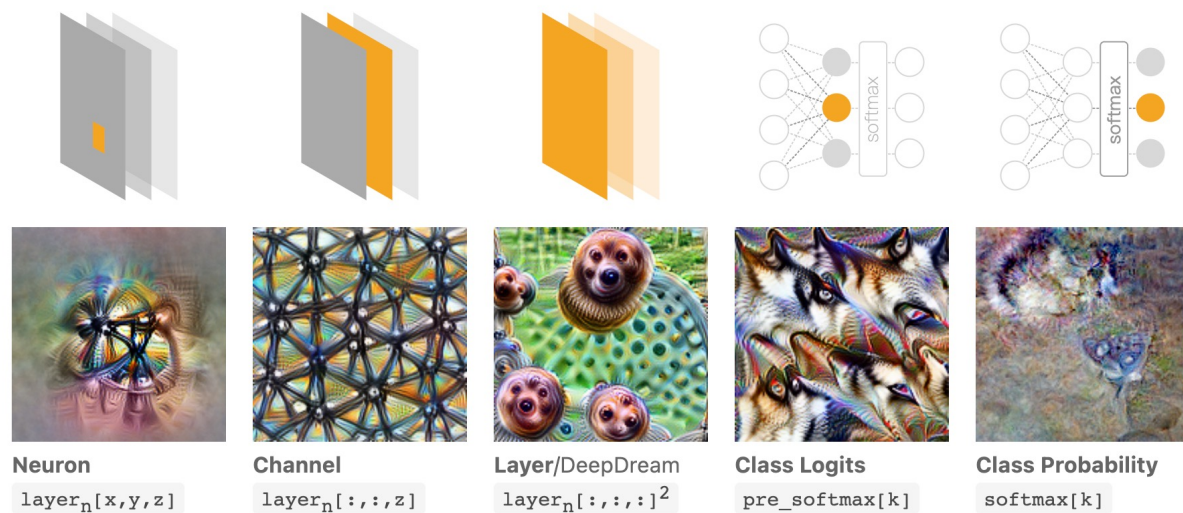
Edges (layer conv2d0)

Textures (layer mixed3a)

Patterns (layer mixed4a)

Parts (layers mixed4b & mixed4c)

Objects (layers mixed4d & mixed4e)



Neuron

$layer_n[x, y, z]$

Channel

$layer_n[:, :, z]$

Layer/DeepDream

$layer_n[:, :, :]^2$

Class Logits

$pre_softmax[k]$

Class Probability

$softmax[k]$

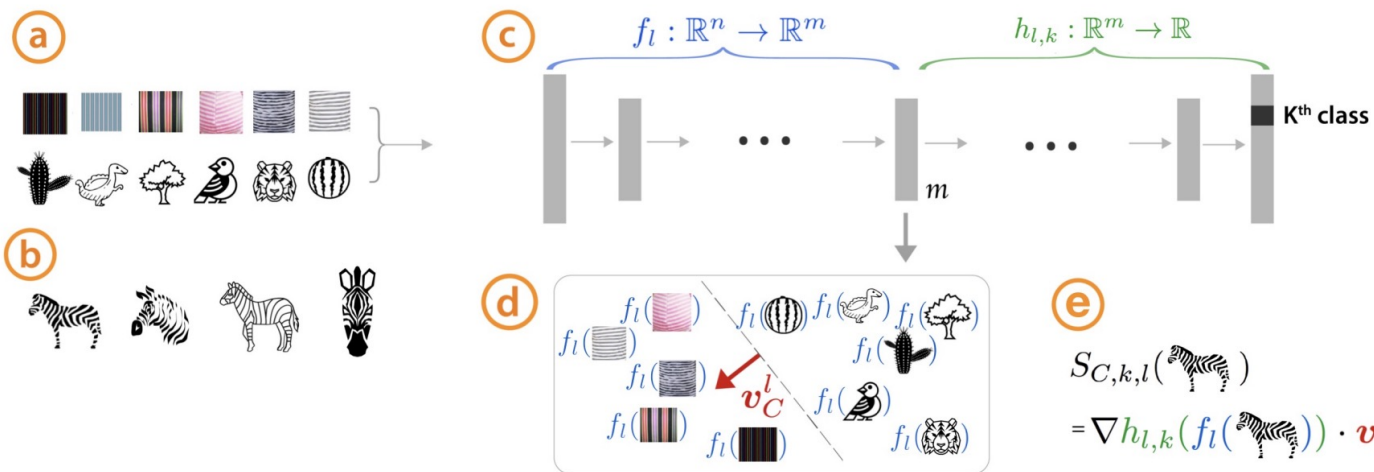
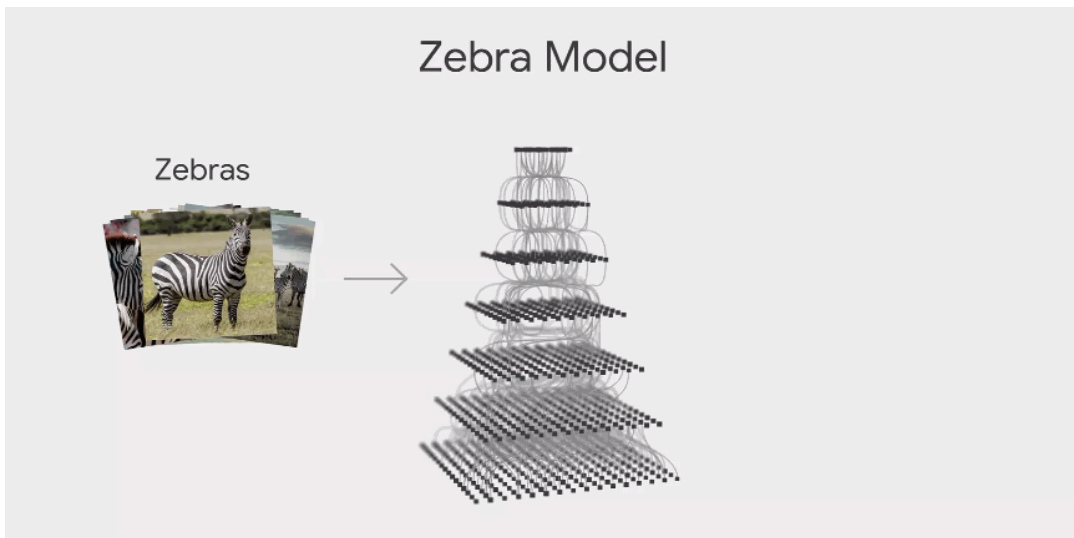
Feature Visualization^[2]:

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。

[2] Feature Visualization, <https://distill.pub/2017/feature-visualization/>

传统可解释AI方法简介

基于语义概念的方法



TCAV^[3]: 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：

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

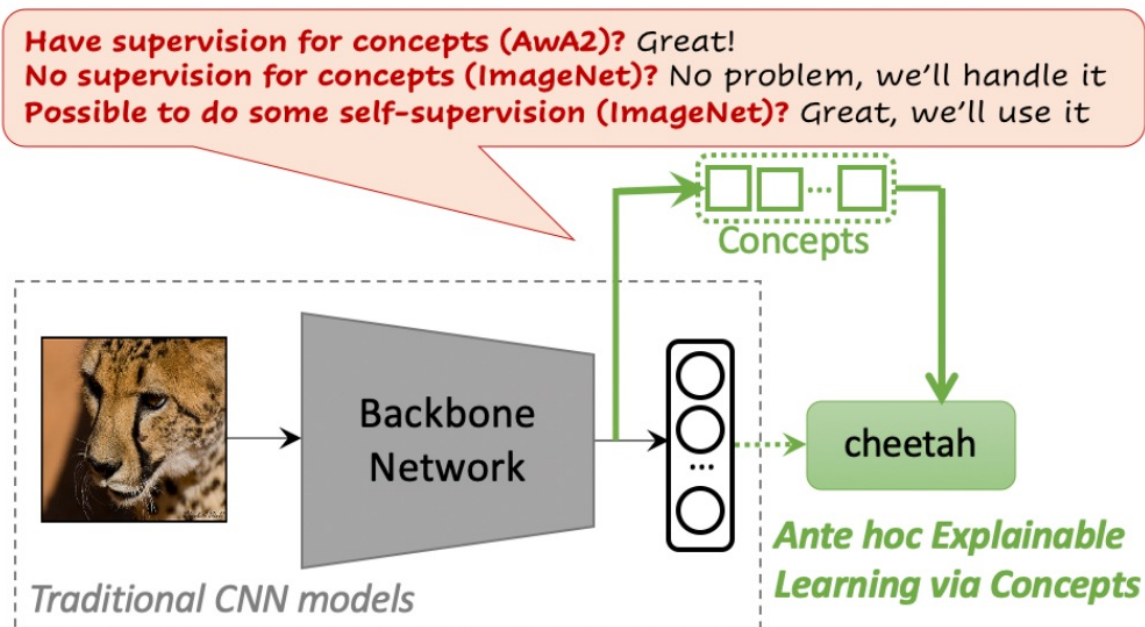
Ramaswamy *et al.*^[4]: 数据集中的概念信息通常不那么突出，也比它们声称要解释的类信息**更难学习**。

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

[4] Ramaswamy, Vikram V., et al. "Overlooked Factors in Concept-Based Explanations: Dataset Choice, Concept Learnability, and Human Capability." *CVPR*. 2023.

传统可解释AI方法简介

基于语义概念的方法

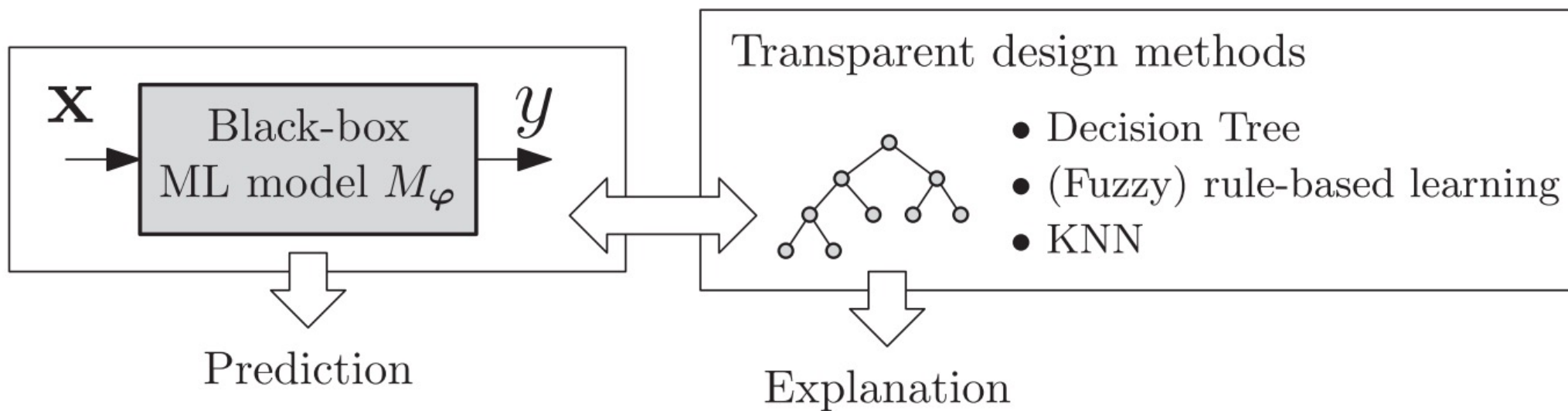


Self-Explaining Neural Networks

自解释网络：在模型学习过程中显式地学习有标注的语义概念，并在推理类别时联合类别特征与概念信息。其可解释性在于模型决策时产生的语义概念。

传统可解释AI方法简介

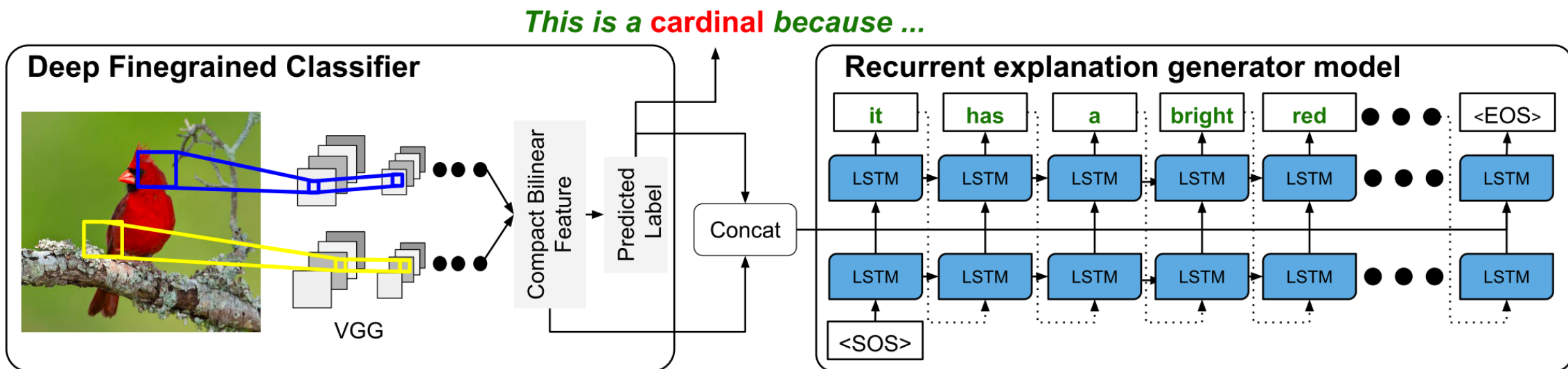
基于关联模型的方法



映射一个不可解释的黑盒系统变成了白盒孪生体，更易于解释。但通常会**影响最终模型的性能**[6]。

传统可解释AI方法简介

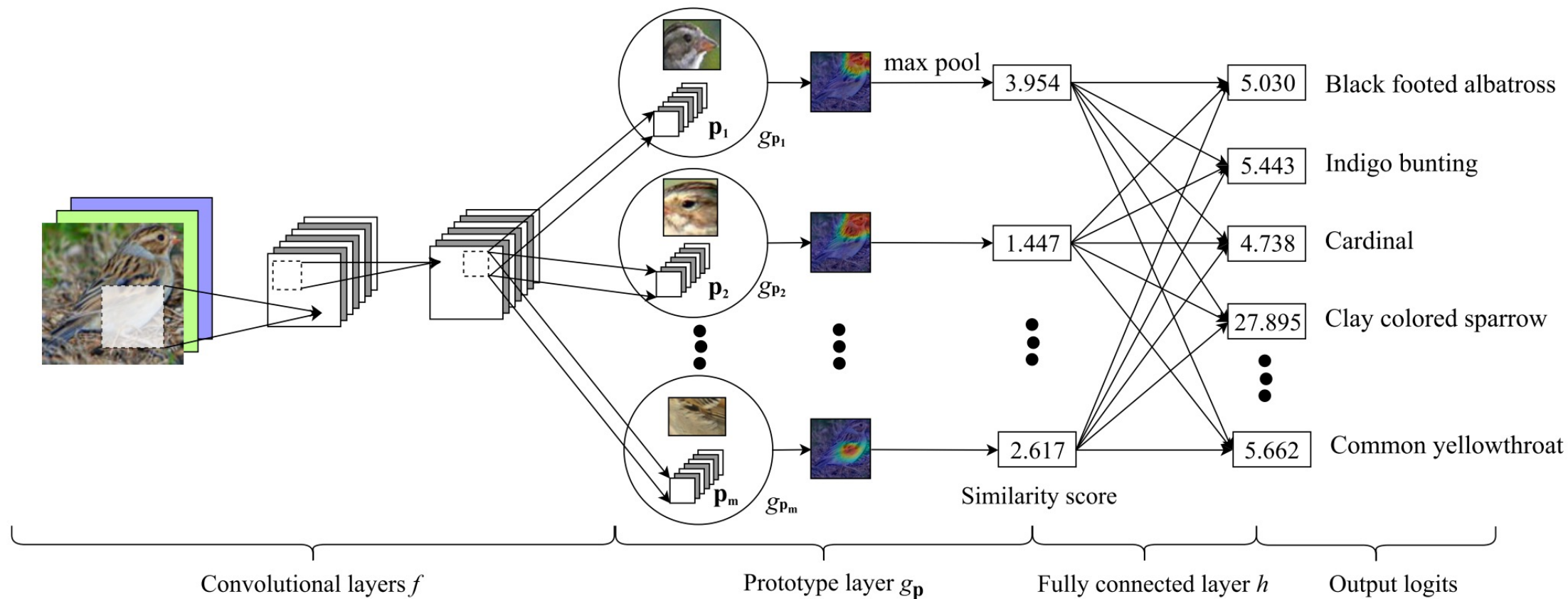
基于其他模态模型的方法



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。

传统可解释AI方法简介

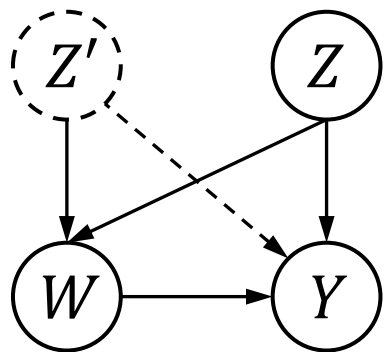
基于设计的方法



需要给定指定的概念原型特征，通用性、可扩展性差。

传统可解释AI方法简介

基于因果的方法



反事实推理

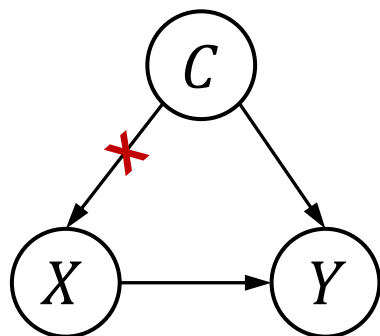
Query (Cardinal)

Why is the prediction a Cardinal? Why is the prediction a Summer Tanager? Why is the prediction confident?

Attributive Explanations

Why is it a Cardinal not a Summer Tanager?

Discriminant Explanations



因果干预

Prediction: Bird

Prediction: Bird

$$P(Y|do(X)) = \sum_{t \in \mathcal{T}} P(Y|X, t)P(t)$$

先验的可能导致模型产生偏见的因素

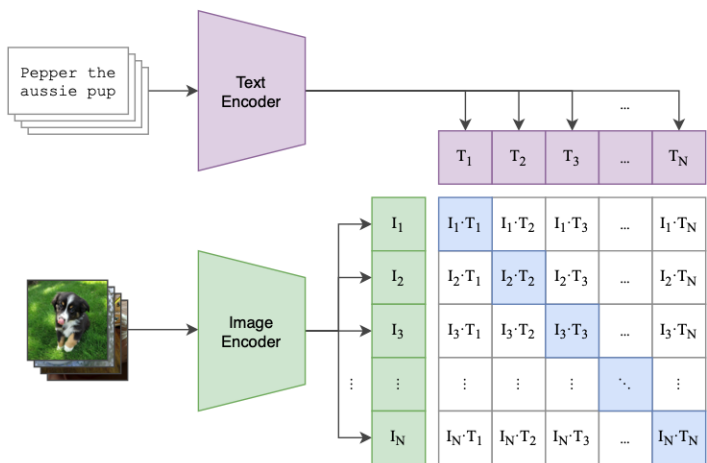
[9] Wang, Pei, and Nuno Vasconcelos. "Scout: Self-aware discriminant counterfactual explanations." *CVPR*. 2020.

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

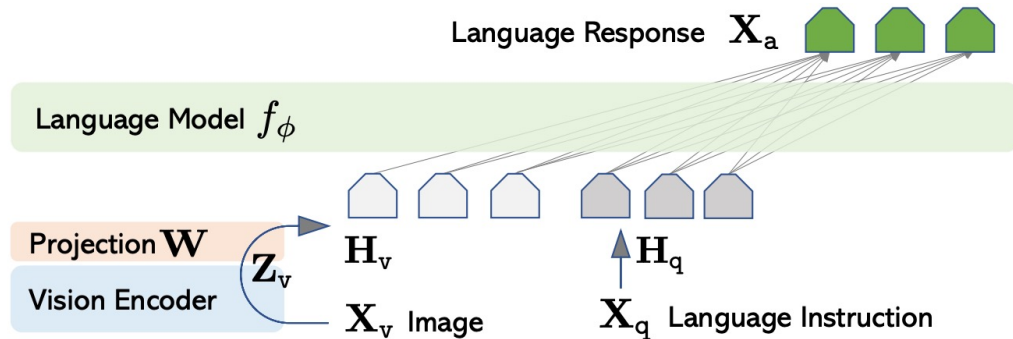
一、传统可解释AI与基础模型

- 传统可解释AI方法简介
- 基础模型的特点及其新挑战

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。

二、基础模型的可解释性研究现状

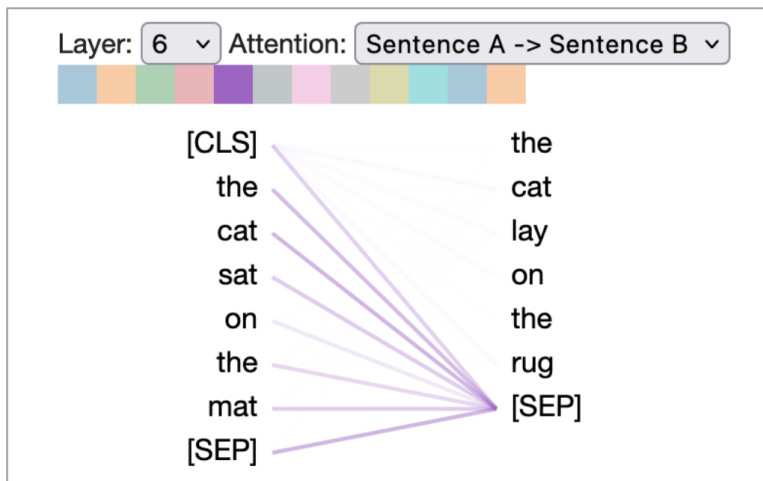
- 大语言模型的可解释性
- 多模态编码式基础模型的可解释性
- 多模态问答式基础模型的可解释性

二、基础模型的可解释性研究现状

- 大语言模型的可解释性
- 多模态编码式基础模型的可解释性
- 多模态问答式基础模型的可解释性

大语言模型的可解释性

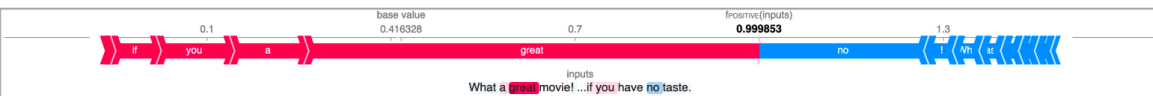
(a) Attention Visualization



(b) Question Answering

Context: In 1899, John Jacob Astor IV invested \$100,000 for Tesla to further develop and produce a new lighting system. Instead, Tesla used the money to fund his **Colorado Springs experiments**.
Question: What did Tesla spend Astor's money on?
Confidence: 0.78 -> 0.91

(c) Sentiment Analysis



(d) Commonsense Reasoning

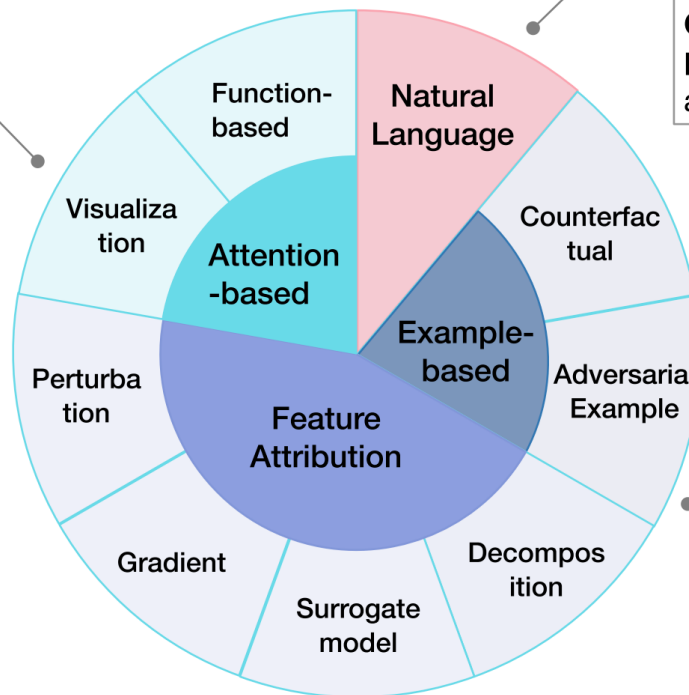
Question: While eating a **hamburger with friends**, what are people trying to do?
Choices: have fun, tasty, or indigestion
Explanation: Usually a hamburger with friends indicates a good time.

(e) Sentiment Analysis

Original text: It is great for kids (**positive**).
Negation examples: It is not great for kids (**negative**)

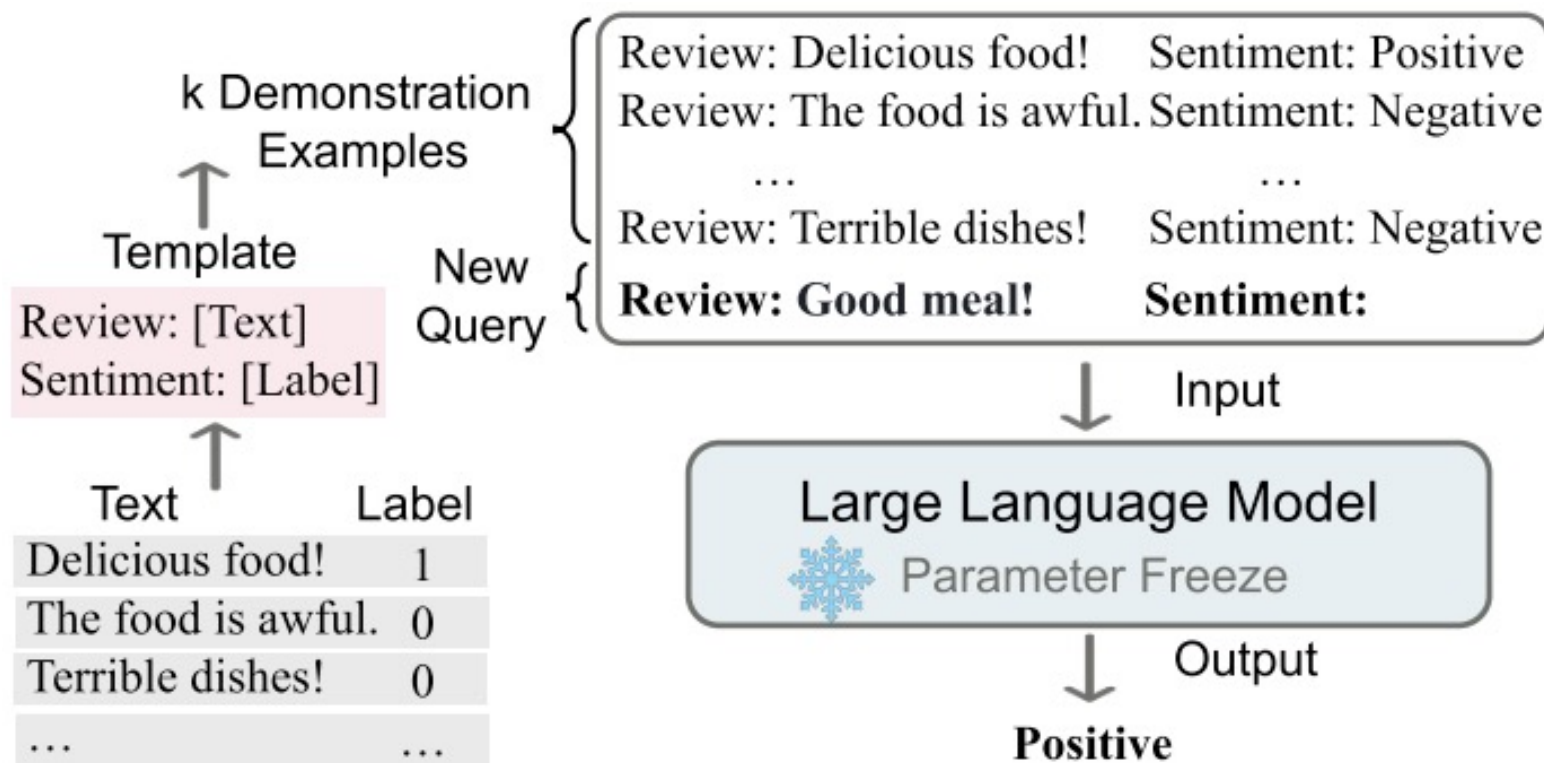
(f) Classification

Original text: The characters, cast in impossibly contrived situations, are totally estranged from reality (**Negative**).
Perturbed text: The characters, cast in impossibly engineered circumstances, are fully estranged from reality (**Positive**)



传统语言模型的局部可解释方法汇总

大语言模型的可解释性



上下文学习的释例。

大语言模型的可解释性

思维链 Chain-of-thought (CoT) 谷歌Brain团队

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

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。

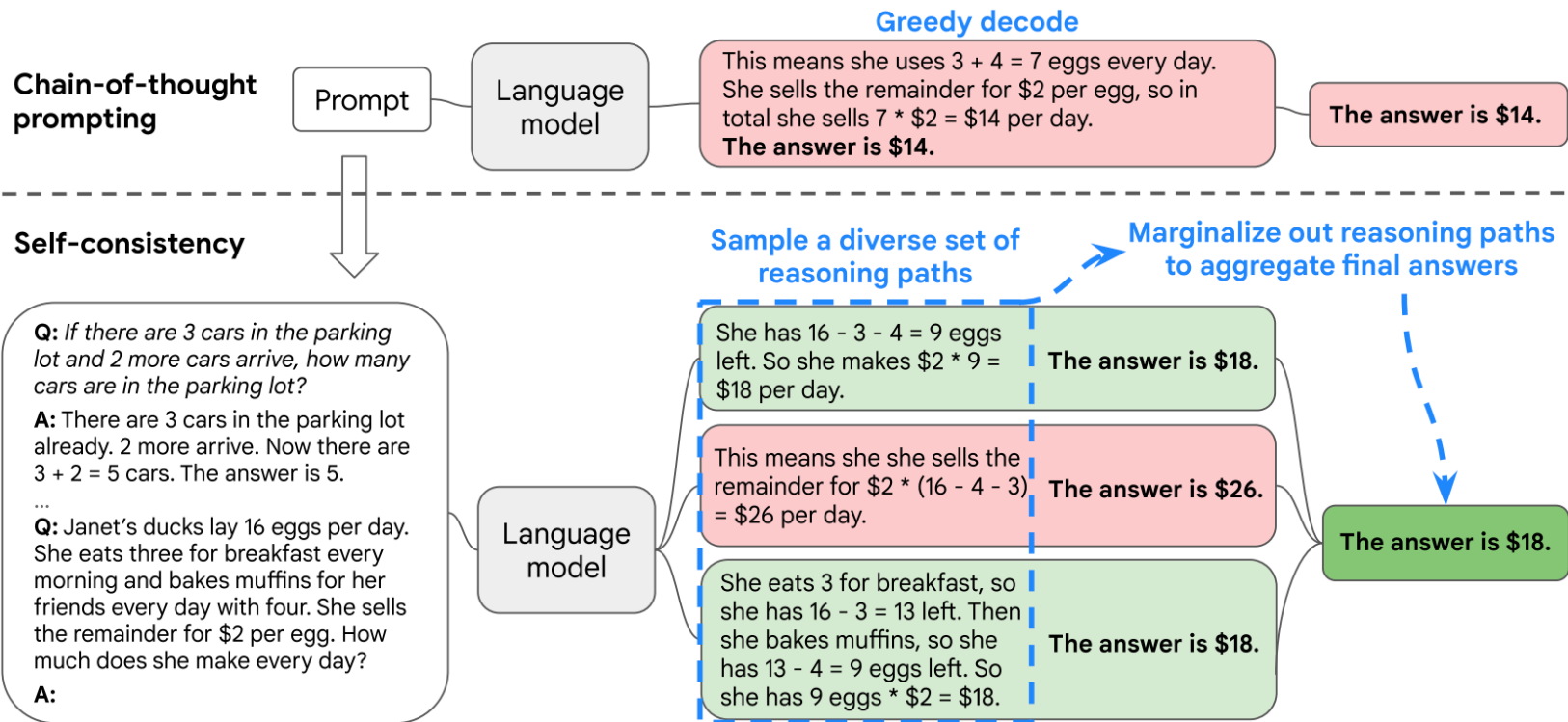
思维链提示使大型语言模型能够处理复杂的算术、常识和符号推理任务。强调了思维链推理过程。

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。

大语言模型的可解释性

Self-Consistency with CoT (CoT-SC)

谷歌Brain团队



步骤:

1. 使用**思维链 (CoT)** 来提示语言模型;
2. 通过从语言模型的解码器中采样来**生成一组不同的推理路径**来替换 CoT 提示中的“贪婪解码”;
3. 通过在最终答案集中**选择最一致的答案**来边缘化推理路径并进行聚合。

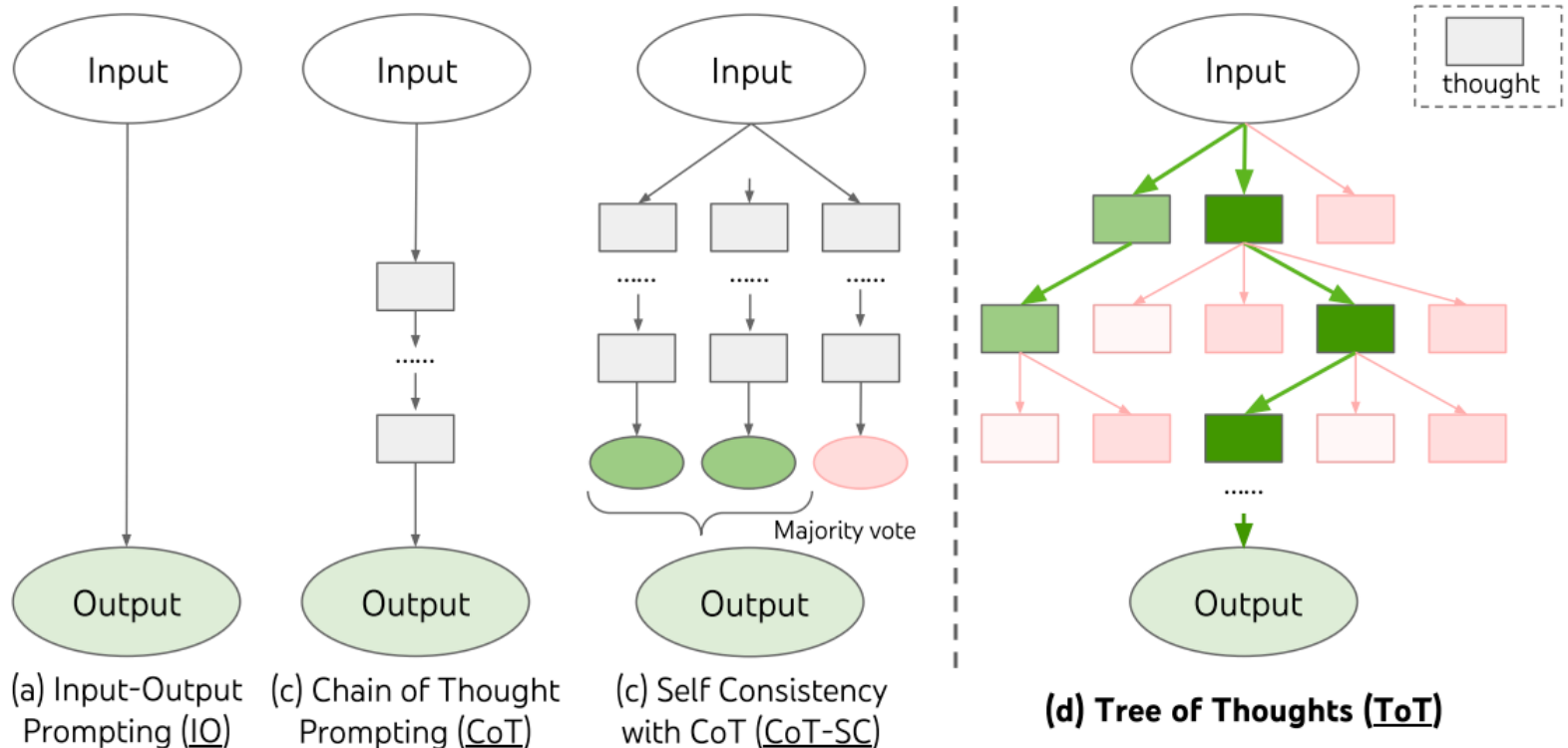
大语言模型的可解释性

思维树 Tree-of-thought (ToT) 普林斯顿大学、谷歌DeepMind团队



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.*



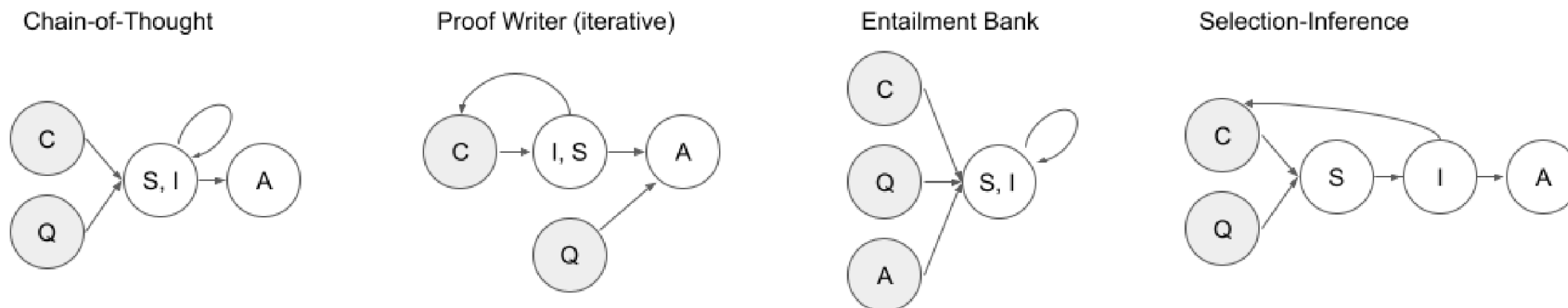
ToT步骤:

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

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

大语言模型的可解释性

Selection-Inference (SI) 谷歌DeepMind团队



C - context, Q - question, A - answer, S - selection, I - inference.灰色圆圈-给定内容, 白色圆圈-模型输出。循环表示多步推理。单个圆圈中字母的顺序表示模型输出相应步骤的顺序。

大语言模型的可解释性

Selection-Inference (SI) 谷歌DeepMind团队

Algorithm 1 Selection-Inference

Require: An n-shot selection prompt, p_{select} .

Require: An n-shot inference prompt, p_{infer} .

Require: Initial Context, C^0 , made up of statements e.g. facts and rules.

Require: The question, q .

Require: Language model, LLM.

Require: The number of reasoning steps, H .

$t=0$

while $t < H$ **do**

$s^t \leftarrow \text{Selection_Module}(p_{select}, C^t, q, \text{LLM})$

$i^t \leftarrow \text{Inference_Module}(p_{infer}, s^t)$

$C^{t+1} \leftarrow C^t \cup i^t$

$t \leftarrow t+1$

end while

return s^t

▷ Start at step 0.

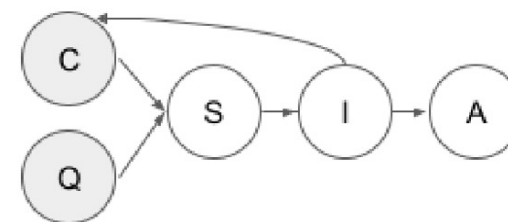
▷ Do selection.

▷ Do inference.

▷ Add the newly inferred fact to the context.

▷ Move onto the next step of reasoning

Selection-Inference



Selection module

```
# n-shot prompt
# First example.
<context 1> \n <question 1>
# Example selection
<fact>. We know that <fact>[ and <fact>]*. Therefore,
...
# Problem to solve.
<context> \n <question>
```

Inference module

```
#n-shot inference prompt
# First example.
<fact>. We know that <fact>[ and <fact>]*. Therefore, <new fact>.
...
# Problem to solve.
<output of the Selection module>. Therefore,
```


大语言模型的可解释性

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。

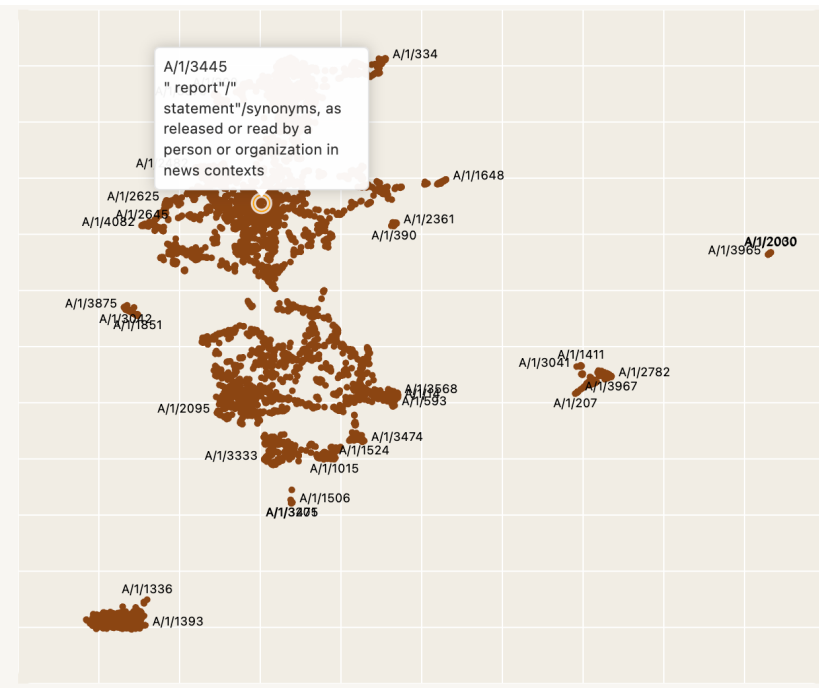
大语言模型的可解释性

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	<ul style="list-style-type: none">● 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	<ul style="list-style-type: none">● 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	<ul style="list-style-type: none">● 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"



大语言模型的可解释性

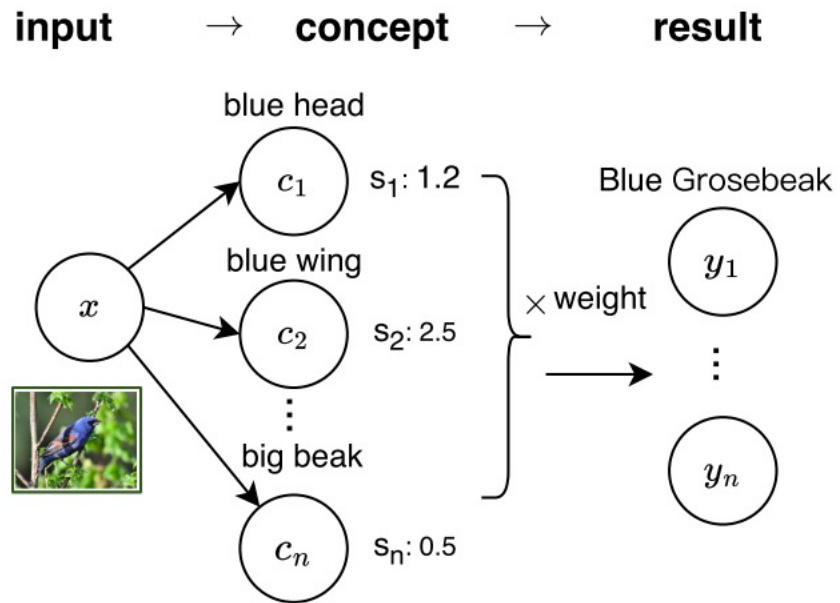
总结

- 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？

二、基础模型的可解释性研究现状

- 大语言模型的可解释性
- 多模态编码式基础模型的可解释性
- 多模态问答式基础模型的可解释性

多模态编码式基础模型的可解释性

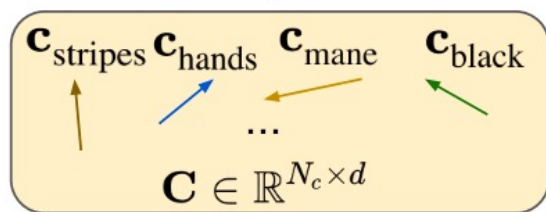
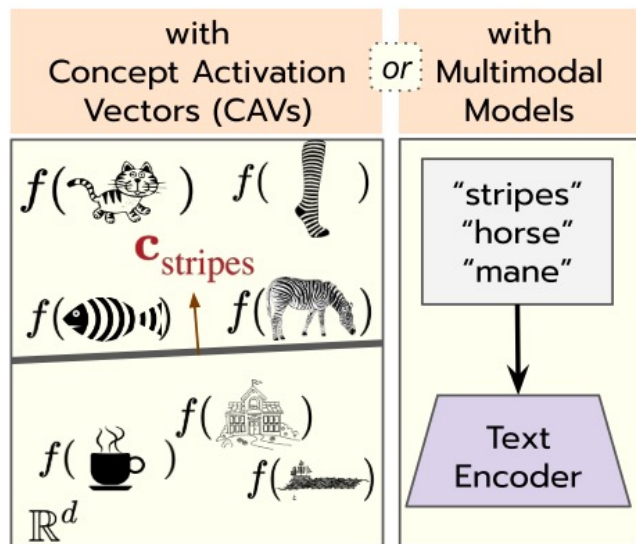


传统的概念瓶颈模型

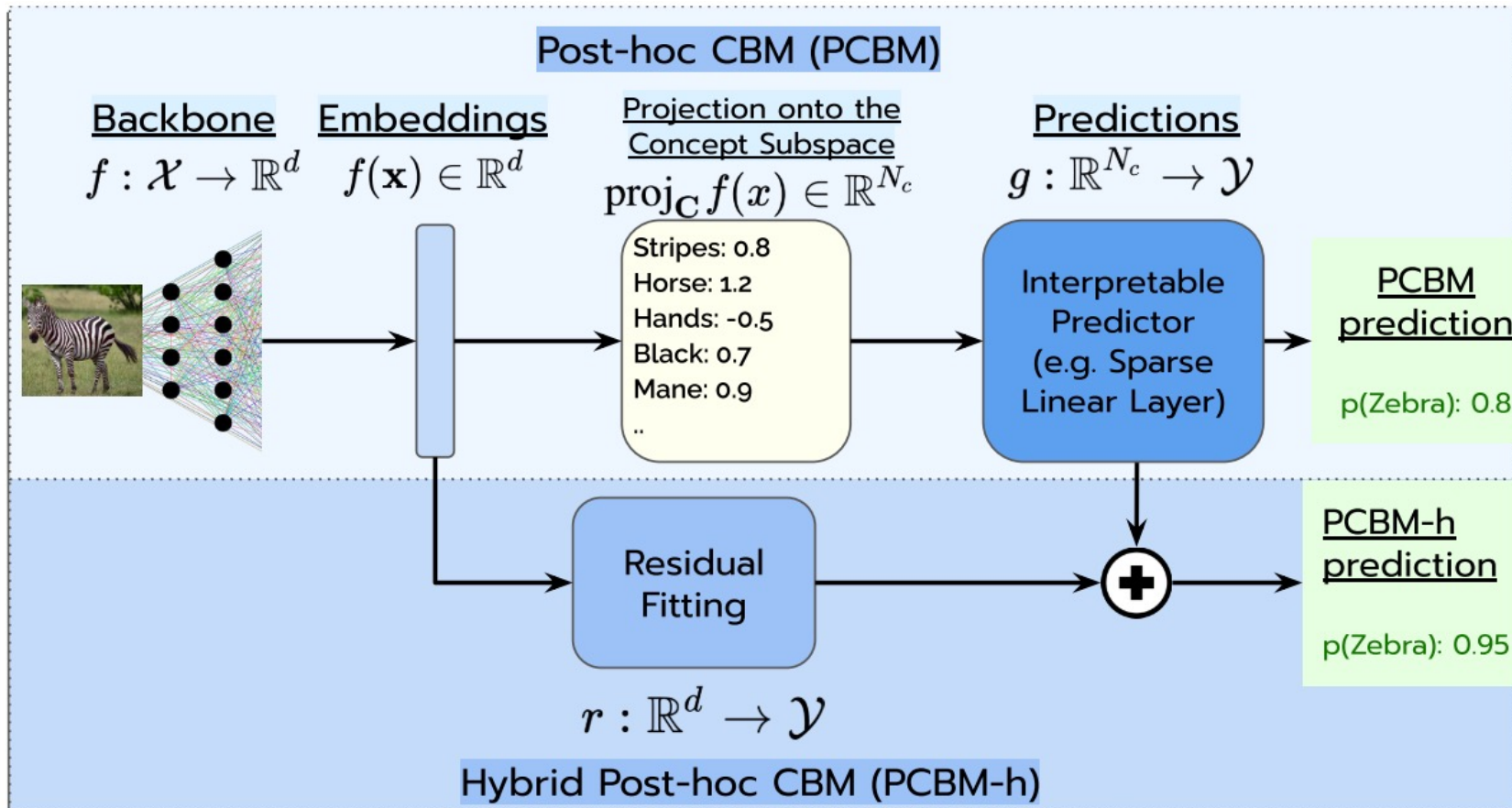
缺点:

- 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



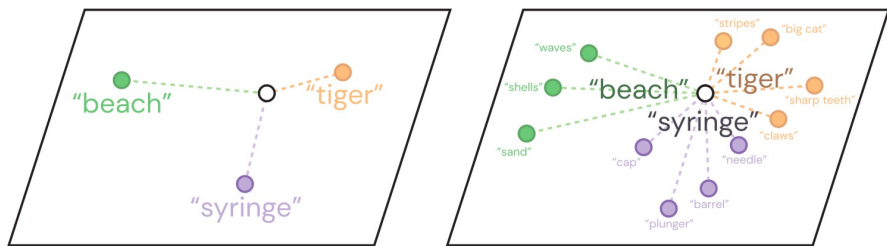
Learning the Concept Bank



Fitting the Concept Bottleneck

Post-hoc Concept Bottleneck Models.

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

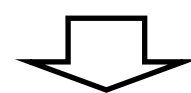
Violin

- 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

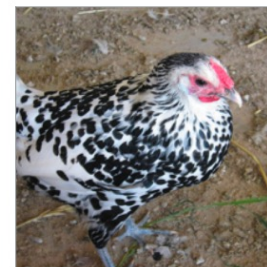
- a large, sailing vessel
- a flag with a skull and crossbones
- cannons on the deck
- a wooden hull
- portholes
- rigging
- a crow's nest

GPT-3生成的描述符模式示例。



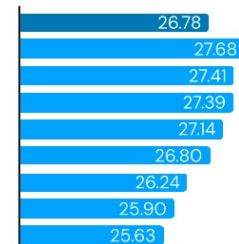
$$s(c, x) = \frac{1}{|D(c)|} \sum_{d \in D(c)} \phi(d, x)$$

Architecture for ϕ		ImageNet			ImageNetV2			CUB		
		Ours	CLIP	Δ	Ours	CLIP	Δ	Ours	CLIP	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



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

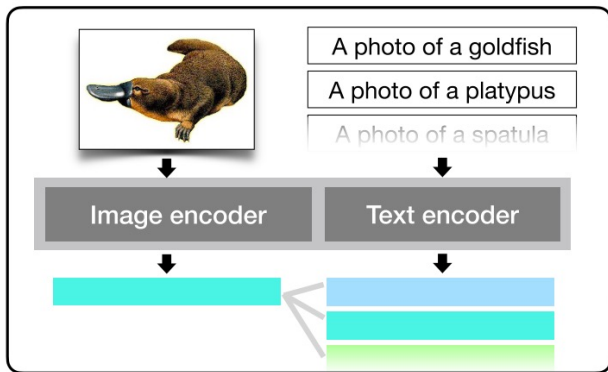


CLIP通过描述符进行决策。

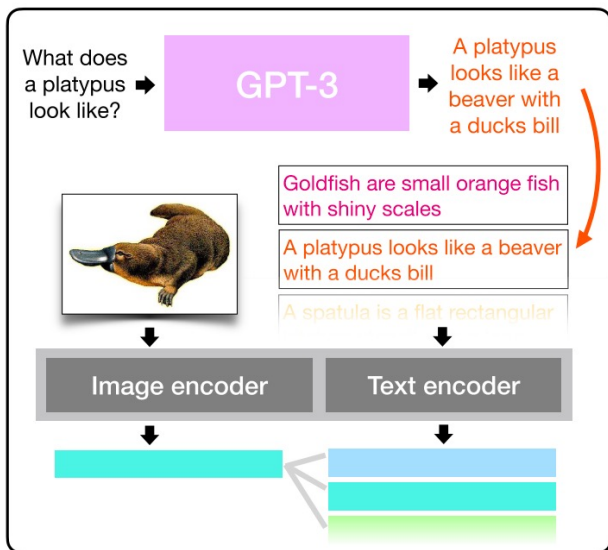
ImageNet和ImageNetV2的模型有一致的~ 3-5%的改进, CUB有~ 1%的改进。

多模态编码式基础模型的可解释性

Standard
Zero-shot



Customized Prompts via
Language models (CuPL)



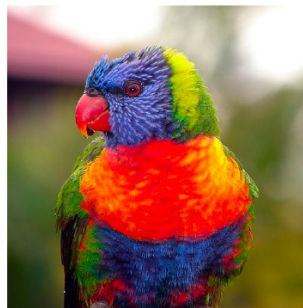
LLM-prompts:

"What does a
{lorikeet, marimba,
viaduct, papillon}
look like?"

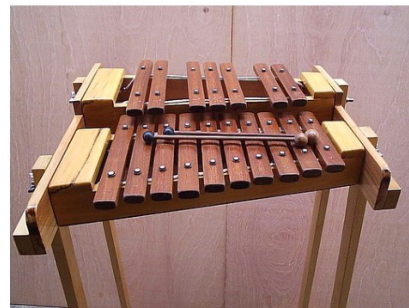
GPT-3

Image-prompts:

"A **lorikeet** is a small to medium-sized parrot with a brightly colored plumage."
"A **marimba** is a large wooden percussion instrument that looks like a xylophone."
"A **viaduct** is a bridge composed of several spans supported by piers or pillars."
"A **papillon** is a small, spaniel-type dog with a long, silky coat and fringed ears."



Lorikeet



Marimba

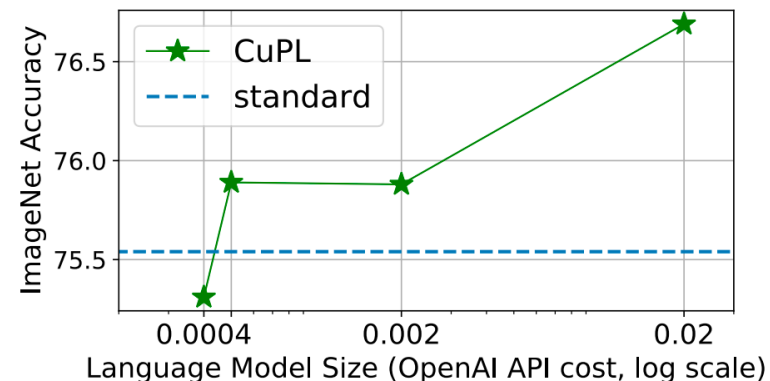
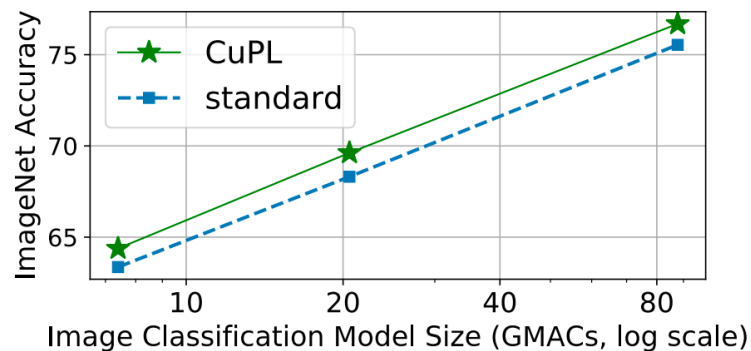


Viaduct

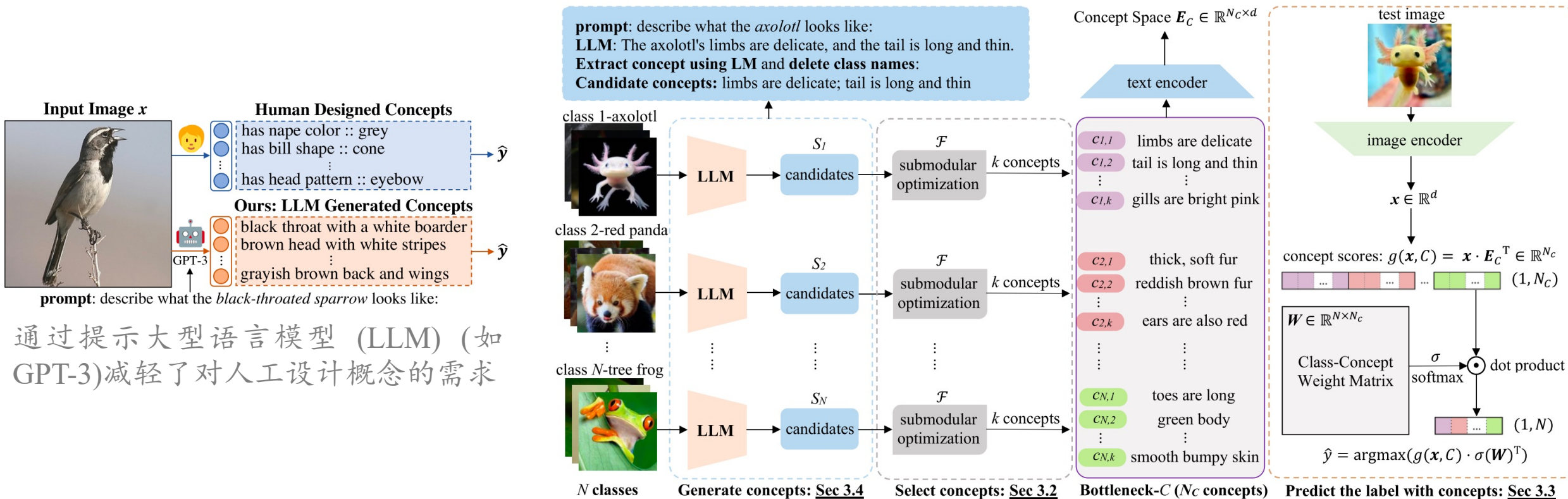


Papillon

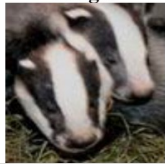



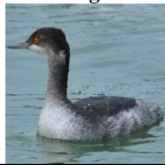












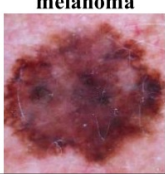


显示了 LLM 生成的图像提示和来自 ImageNet 的相关图像的示例。仅图像提示用于下游图像分类。



多模态编码式基础模型的可解释性



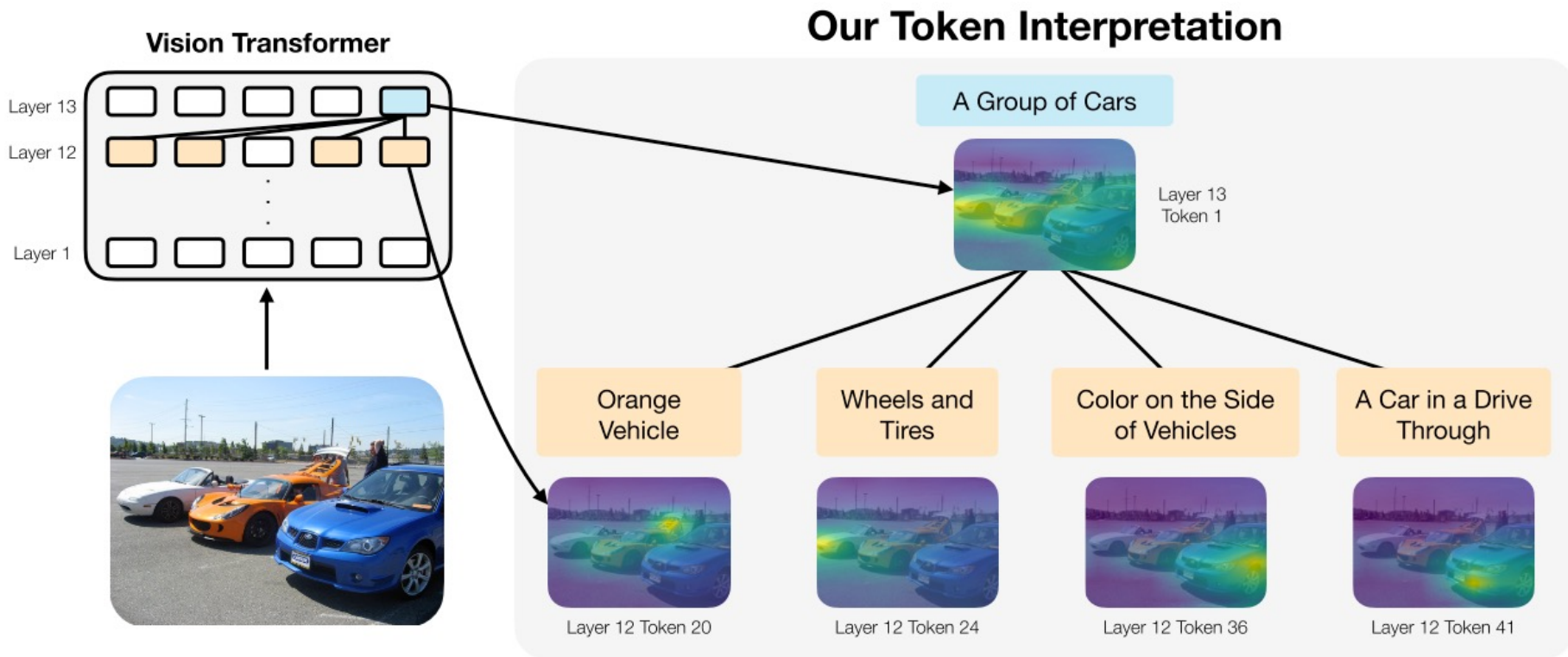
多模态编码式基础模型的可解释性

	Class Name	Top-3 Concepts	Class Name	Top-3 Concepts	Class Name	Top-3 Concepts	Class Name	Top-3 Concepts
ImageNet	badger 	<ol style="list-style-type: none"> 1. short legs and long body make it an excellent digger 2. black-and-white striped fur 3. coat is very shaggy 	ant 	<ol style="list-style-type: none"> 1. black and red stinger 2. small, black insect with six legs 3. long, slender antennae that it uses to smell and touch 	hammer 	<ol style="list-style-type: none"> 1. long, thin tool with a wooden handle 2. great tool for pounding object 3. used to pound on surfaces 	water buffalo 	<ol style="list-style-type: none"> 1. large head with short, curved horns 2. heaviest living species of bovid 3. huge, dark-colored animal
	CUB	eared grebe 	<ol style="list-style-type: none"> 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 	<ol style="list-style-type: none"> 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 	<ol style="list-style-type: none"> 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 
Flower		water lily 	<ol style="list-style-type: none"> 1. depicted in artworks of ponds and waterfall 2. member of the nymphaeaceae family 3. lily pads float 	barbeton daisy 	<ol style="list-style-type: none"> 1. scientific name for the flower is taraxacum officinal 2. named after the city of barberton 3. member of the daisy family 	marigold 	<ol style="list-style-type: none"> 1. central disc with smaller florets 2. have a slightly furry texture 3. bold and vibrant color palette 	tiger lily 
	UCF-101	archery 	<ol style="list-style-type: none"> 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 	<ol style="list-style-type: none"> 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 	<ol style="list-style-type: none"> 1. deep blue color 2. tans contrast with the white of their boards 3. sending a spray of water into the air 	long jump 
HAM10000		dermatofibroma 	<ol style="list-style-type: none"> 1. generally not painful 2. red, brown, or purple in color 3. thin white halo around them 	melanoma 	<ol style="list-style-type: none"> 1. dark brown or black in color 2. large and dark 3. flesh-colored, brown, or black 	melanocytic nevi 	<ol style="list-style-type: none"> 1. color is tan 2. dark brown or black color 3. small, round, and slightly raised 	benign lesions 

生成的概念示意图

多模态编码式基础模型的可解释性

解释CLIP中潜在token的表征



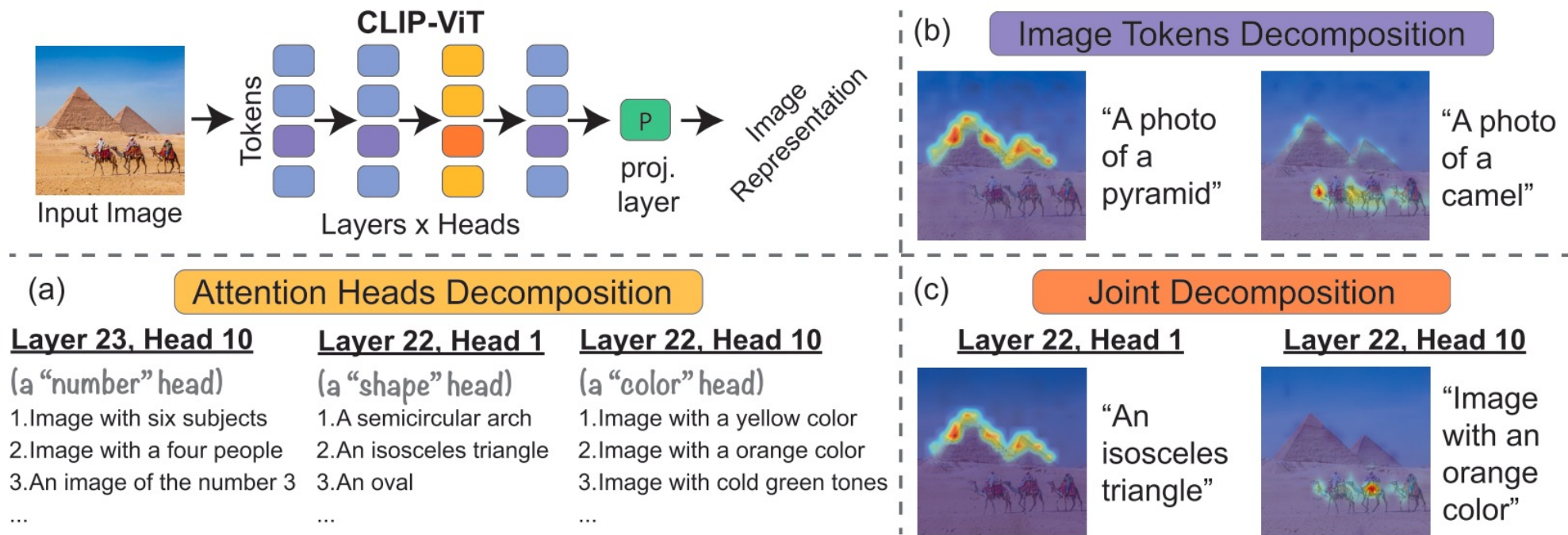
Transformer视觉推理的解释，所提出的方法允许对潜在token进行文本解释，而无需任何训练或数据收集。

关键假设是，当Transformer中的潜在token不关注其他token时，它们在后续层中保持相同的语义信息。所提出的方法利用这个属性来解释潜在token，通过前向传播将潜在嵌入映射到最后一层，而无需自注意力操作。

然后利用开放世界词汇表来解释潜在的token嵌入。通过返回每个潜在token的语言描述，提出的方法直接阐明了在Transformer中学习的概念。

该假设未在论文中理论性地证实，有待考证！

多模态编码式基础模型的可解释性



通过将 CLIP 的图像表示分解为各个图像块、模型层和注意力头的总和，我们可以 (a) 通过自动查找跨越其输出空间的文本可解释方向来表征每个头的角色，(b) 突出显示有助于图像和文本之间相似性得分的图像区域，以及 (c) 呈现哪些区域有助于在特定头部找到文本方向。

多模态编码式基础模型的可解释性

Causality Inspired Model Interpreter (CIMI) 中国科学技术大学、微软亚洲研究院

为了深入了解大模型的科学原理并确保其安全，可解释变得日益重要。解释大模型带来了许多独特挑战：（1）大模型参数特别多，怎么尽可能确保解释速度？（2）大模型涉及的样本特别多，如何让用户尽可能少看一些样本的解释也能了解大模型的全貌？这两个问题都指向了对大模型解释效率的要求，而本文希望通过新的范式，为构建大模型高效解释之路提供一个思路。

目前仍然缺乏可解释性的正式且统一的因果视角，一些关键的研究问题仍然难以回答，例如：

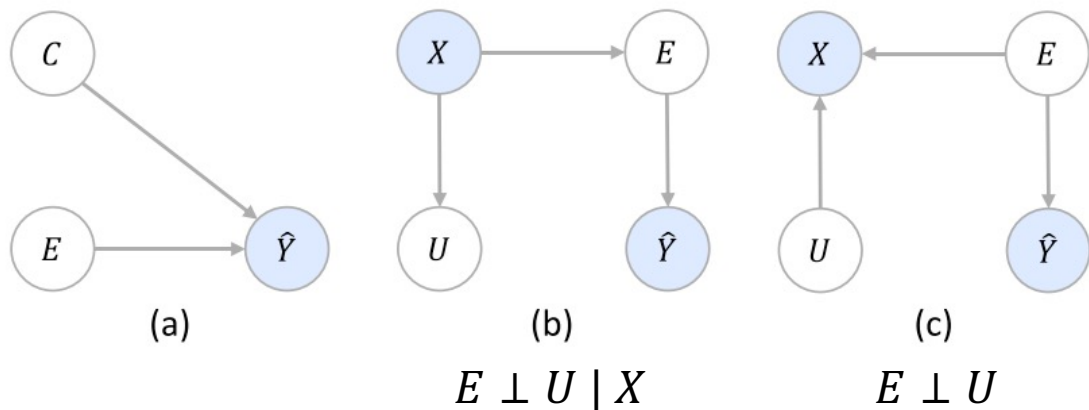
- 现有的解释方法能否在因果理论框架内构建？
如果是这样，采用的因果模型是什么，它们之间有何区别？
- 利用因果推理进行模型解释的主要挑战是什么？
通过解决这些挑战我们可以实现什么好处？
- 如何改进因果模型来克服这些挑战？

多模态编码式基础模型的可解释性

Causality Inspired Model Interpreter (CIMI) 中国科学技术大学、微软亚洲研究院

X : 输入变量
 \hat{Y} : 模型预测

E : 解释的未知随机变量
 U : 非解释的未知随机变量



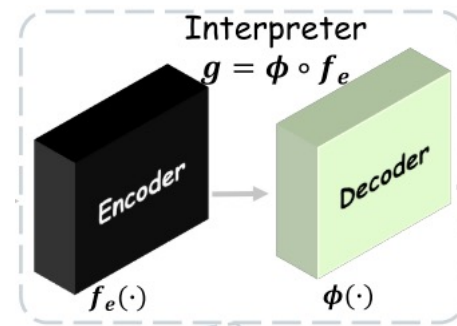
(a)现有的可解释方法的因果图，(b)另一种因果图，其中的解释在因果关系上对预测是充分的，但不能泛化， X 改变时， E 和 U 非相互独立，以及(c)本文提出的模型，其中解释 E 是可泛化的（ X 改变时， E 和 U 相互独立），并建模为 \hat{Y} 的唯一原因。观察到的变量用蓝色阴影表示。

假设 E 是 X 中影响 \hat{Y} 的特征，而 U 是 X 中另外的特征，与 E 不相交。

$$E = M \odot X$$

$$U = (1 - M) \odot X$$

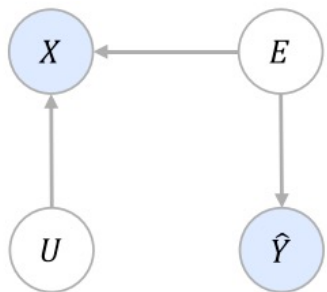
目标即为学习一个方程 $g: X \rightarrow M$ 。



多模态编码式基础模型的可解释性

Causality Inspired Model Interpreter (CIMI)

中国科学技术大学、微软亚洲研究院

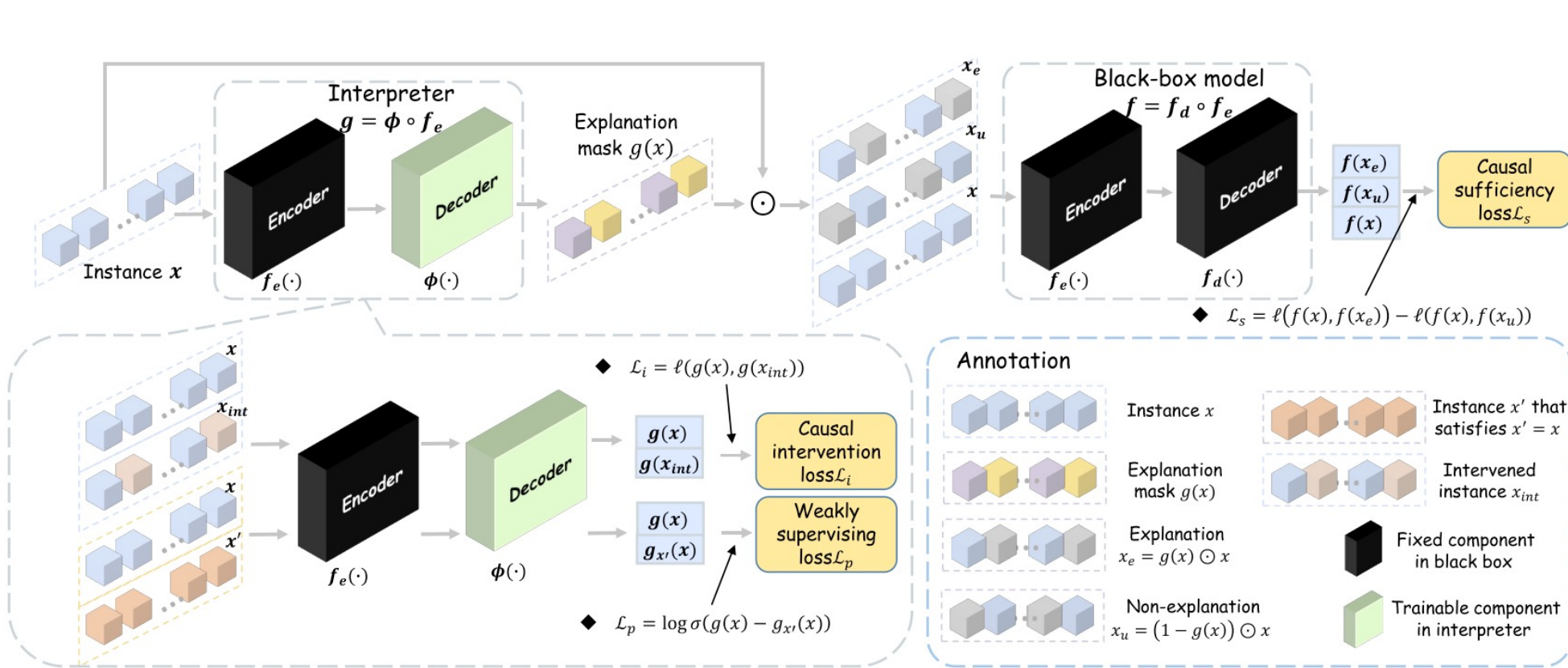


假设 E 是 X 中影响 \hat{Y} 的特征，而 U 是 X 中另外的特征，与 E 不相交。

$$E = M \odot X$$

$$U = (1 - M) \odot X$$

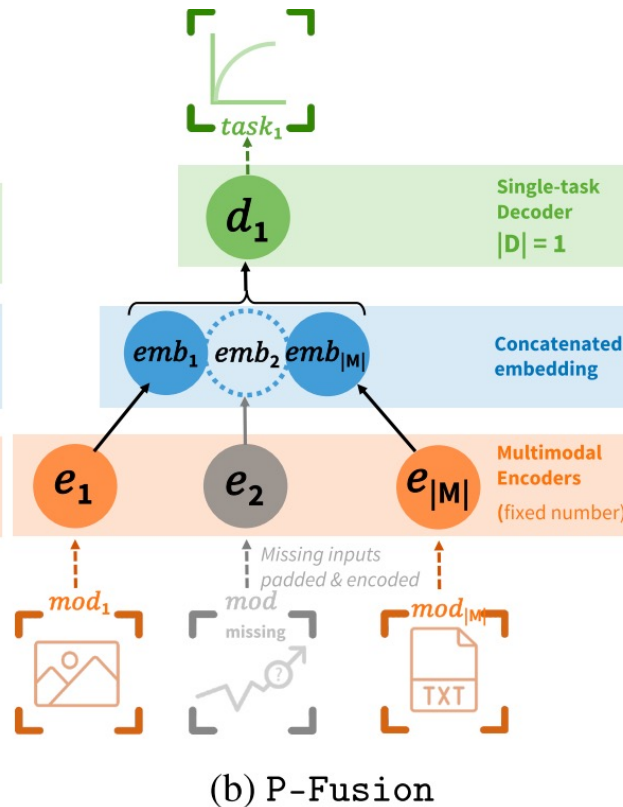
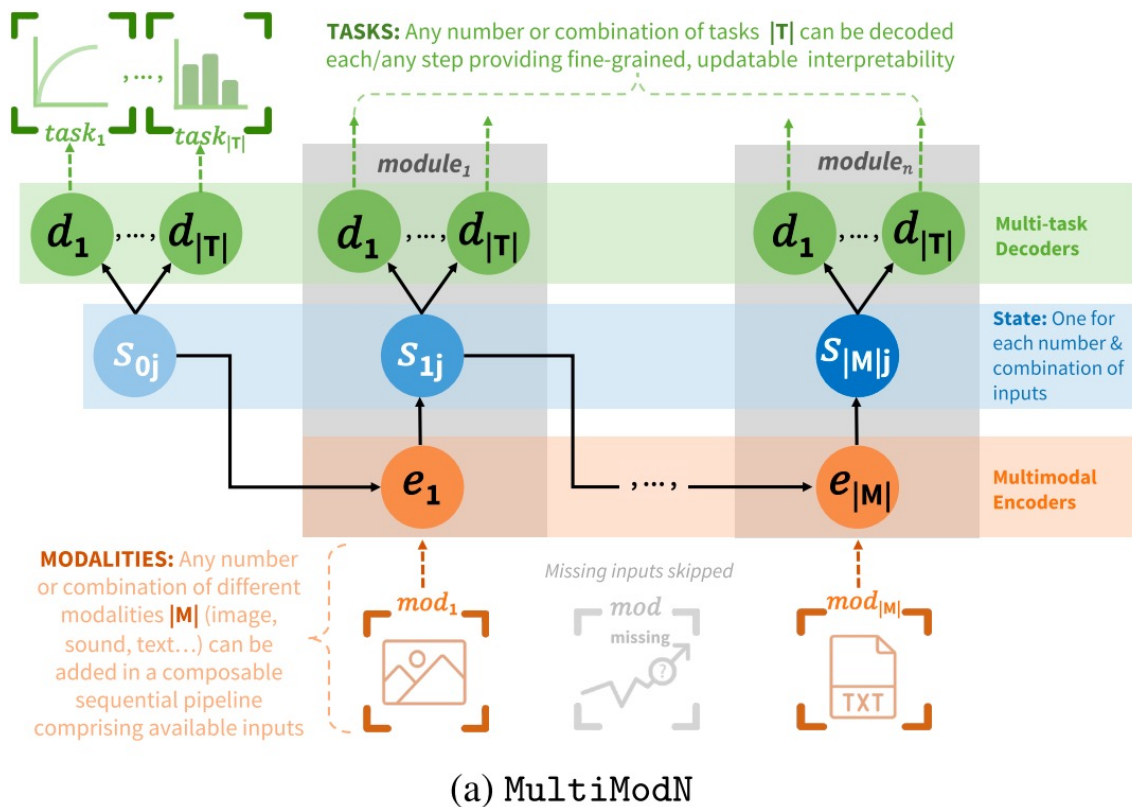
目标即为学习一个方程 $g: X \rightarrow M$ 。



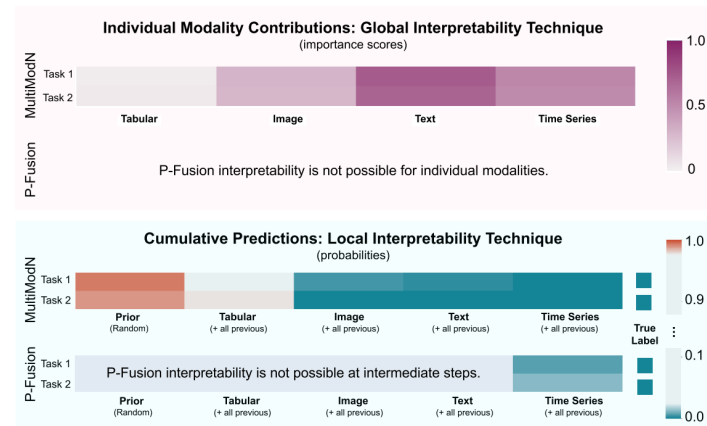
CIMI的框架。唯一可训练的组件是解码器 ϕ ，这是一个简单的神经网络，可以用相对较少的样本进行训练。

多模态编码式基础模型的可解释性

设计新式模型，但大参数模型上可能难训练



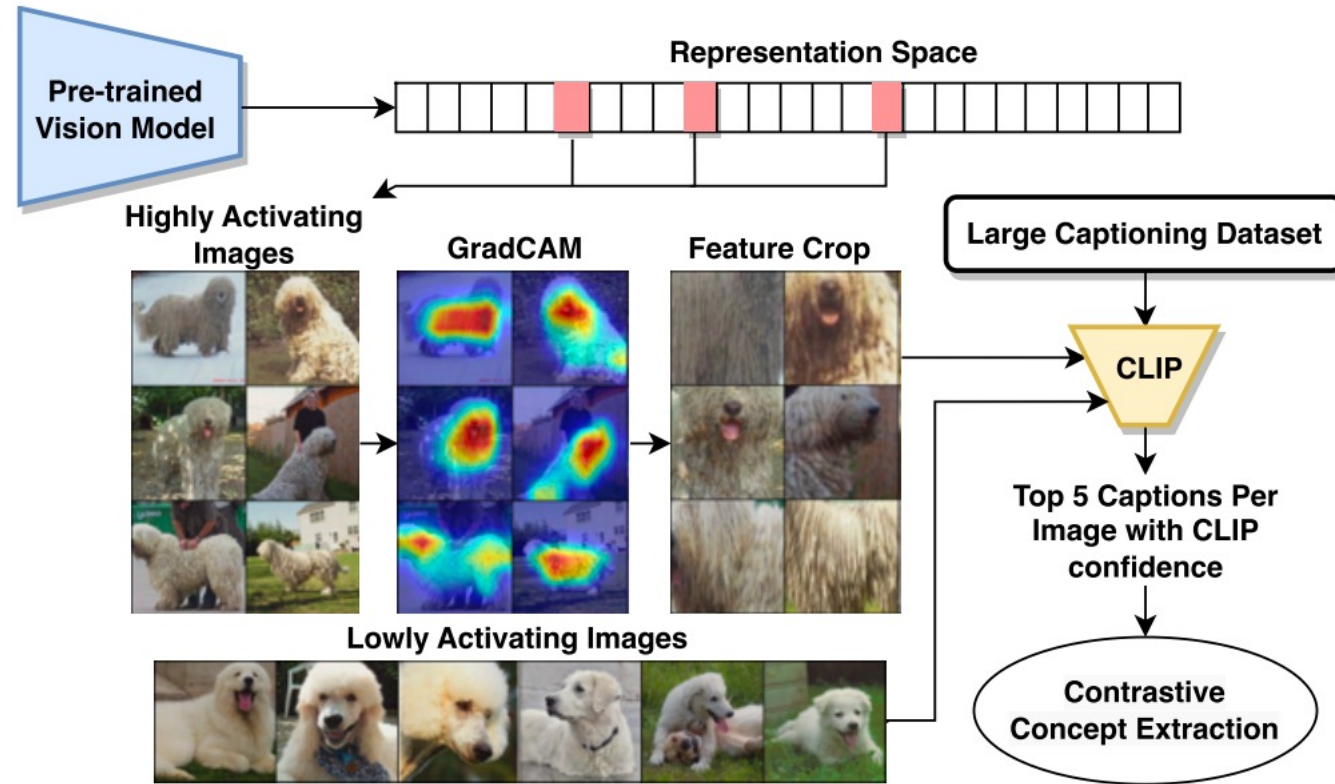
当前的多模态模型提出了模态的并行集成，其中表示被同时融合和处理。并行聚变(以下简称p-Fusion)产生了限制。



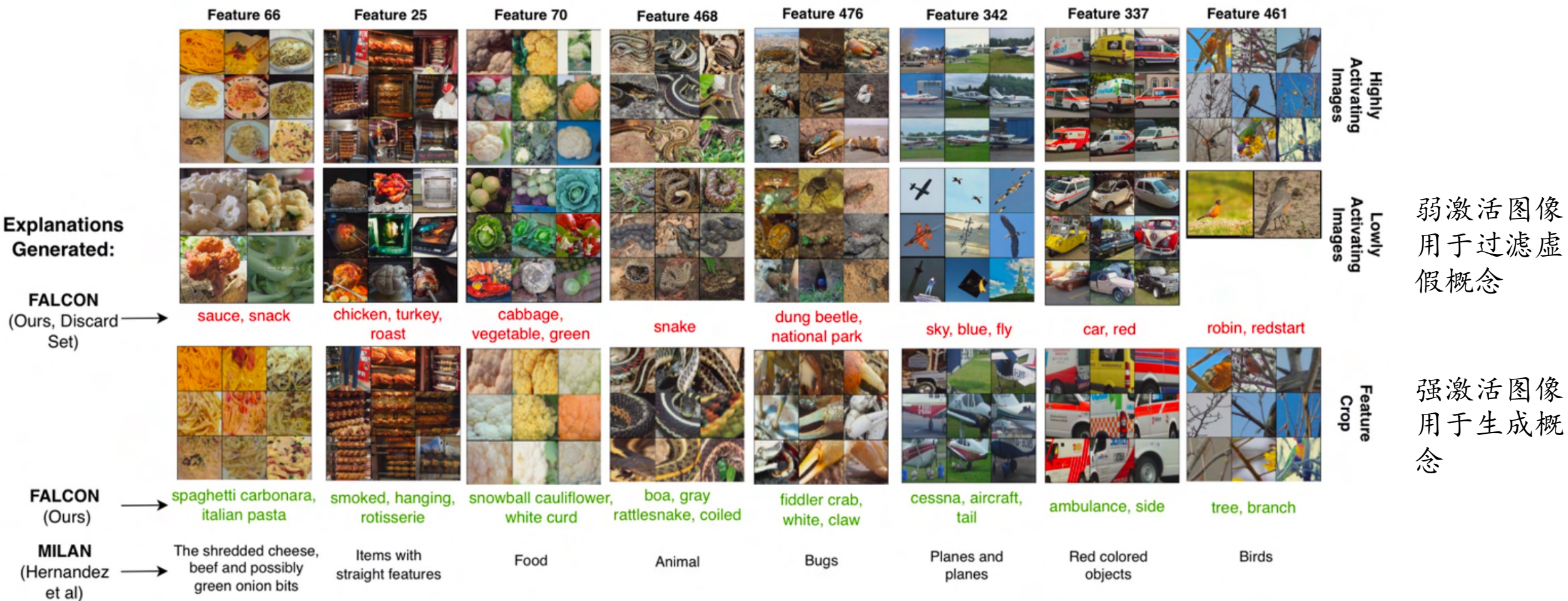
MultiModN中固有的特定于模式的模型可解释性。

多模态编码式基础模型的可解释性

解释多模态特征的表征

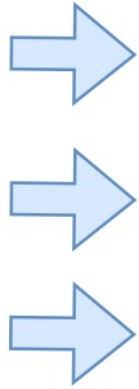
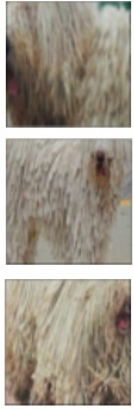


多模态编码式基础模型的可解释性



多模态编码式基础模型的可解释性

Highly Activating Images

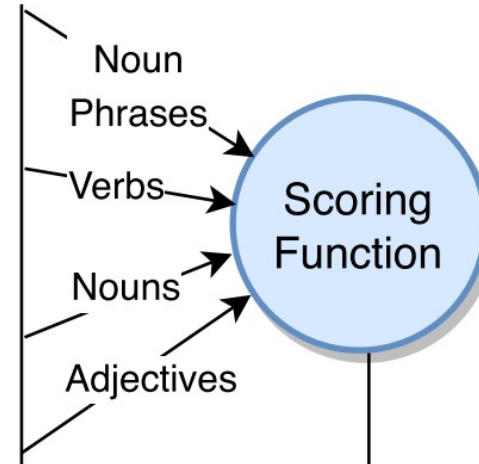


Top 5 Captions with CLIP Confidence

a closeup picture of a Bearded Collie panting - SCORE: 0.723
Closeup Of A Bearded Collie... - SCORE: 0.708
A close up of a Bearded Collie's healthy, long coat - SCORE: 0.677
image of a shaggy dog's profile with patio furniture blurred out in the background. - SCORE: 0.664
Bearded Collie close up - SCORE: 0.652

White Shaggy Dog Mascot Costume - SCORE: 0.888
A close up of a Komondor with a white corded coat - SCORE: 0.886
File:Komondor.png - SCORE: 0.837
Komondor head Corded White Long Showing - SCORE: 0.811
A komondor participates in the annual event on February 11. - SCORE: 0.809

Komondor pup in grass - close up Stock Footage - SCORE: 0.562
The Puli: A Dreadlocked Dog from the Carpathians - SCORE: 0.53
image of a shaggy dog's profile with patio furniture blurred out in the background. - SCORE: 0.516
The Puli, Hungary's Indispensable Herd Dog - SCORE: 0.485
A close up of a Komondor with a white corded coat - SCORE: 0.437



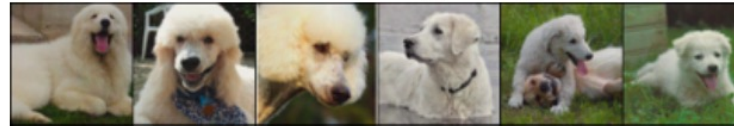
Discard Set

Words containing digits and special characters

Stopwords

Non-conceptual words

Lowly Activating Images



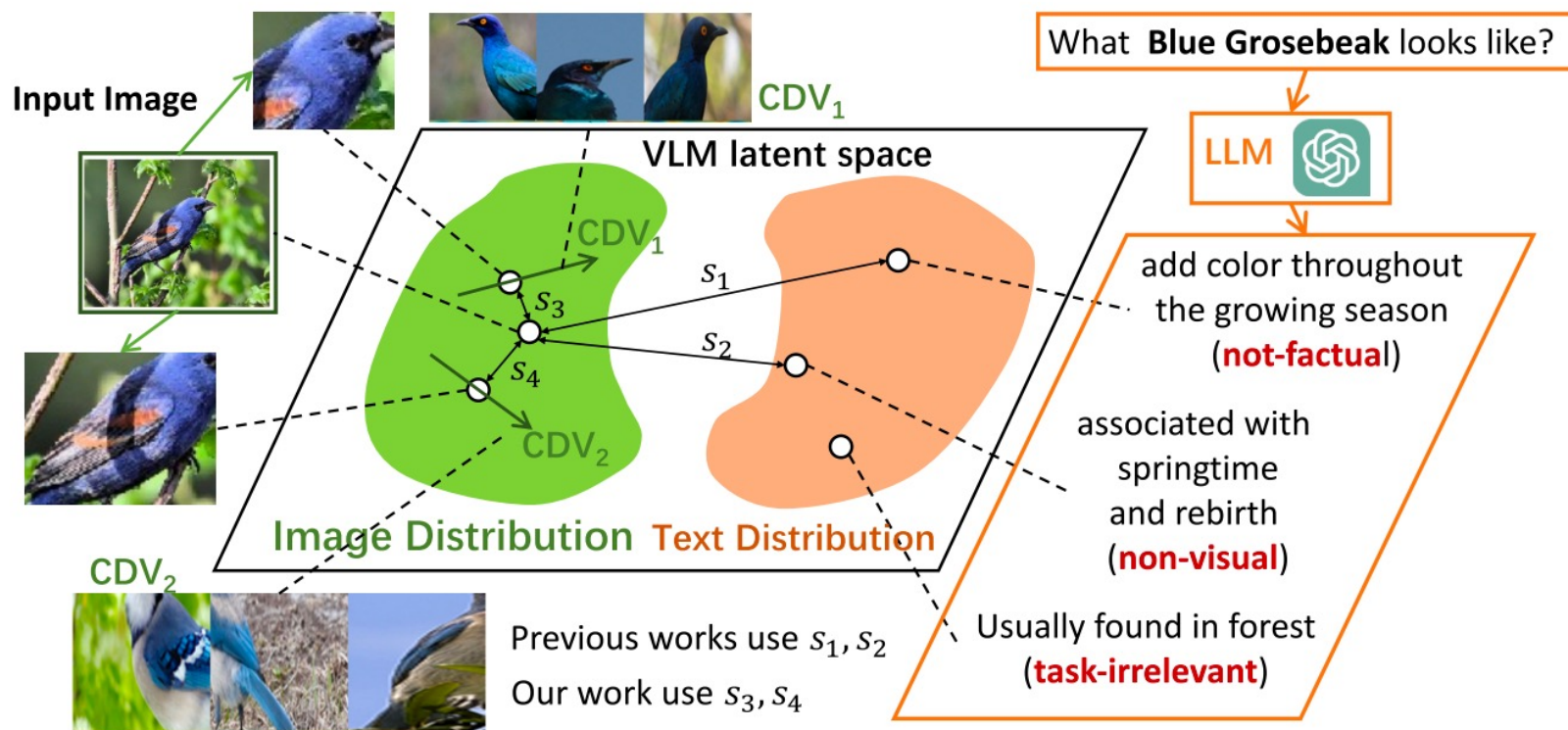
Contrasting (Spurious) concepts:
DOG, WHITE

Top scoring concepts:

WHITE (0.37)
DOG (0.30)
SHAGGY (0.29)
COAT (0.26)
KOMONDOR (0.20)
CORDED (0.16)

多模态编码式基础模型的可解释性

CLIP ViT-L-14 latent space illustration



VLM潜在空间中不同模式的概念示意图。在这项工作中，预计将使用视觉概念而不是LLM给出的文本概念进行解释。

多模态编码式基础模型的可解释性

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？一些假设是否正确？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果，同时适应在大模型训练数据量巨大的情况。

二、基础模型的可解释性研究现状

- 大语言模型的可解释性
- 多模态编码式基础模型的可解释性
- 多模态问答式基础模型的可解释性

多模态问答式基础模型的可解释性

VisProg CVPR 2023 Best Paper 美国Allen Institute for AI (西雅图)

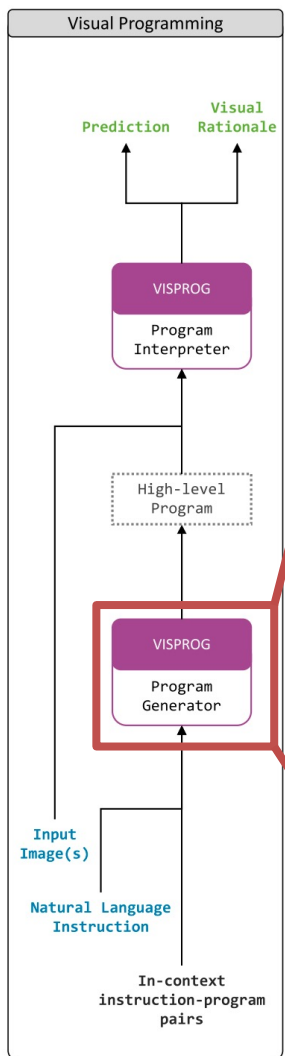


VisProg是一个模块化和可解释的神经符号系统, 用于组合视觉推理。

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

多模态问答式基础模型的可解释性

VisProg CVPR 2023 Best Paper 美国Allen Institute for AI (西雅图)



In-context Examples

```

Instruction: Hide the face of Nicole Kidman with :p
Program:
OBJ0=Facedet(image=IMAGE)
OBJ1=Select(image=IMAGE, object=OBJ0, query='Nicole Kidman')
IMAGE0=Emoji(image=IMAGE, object=OBJ1, emoji='face_with_tongue')
RESULT=IMAGE0

Instruction: Create a color pop of the white Audi
Program:
OBJ0=Seg(image=IMAGE)
OBJ1=Select(image=IMAGE, object=OBJ0, query='white Audi')
IMAGE0=ColorPop(image=IMAGE, object=OBJ1)
RESULT=IMAGE0

Instruction: Replace the red car with a blue car
Program:
OBJ0=Seg(image=IMAGE)
OBJ1=Select(image=IMAGE, object=OBJ0, query='red car')
IMAGE0=Replace(image=IMAGE, object=OBJ1, prompt='blue car')
RESULT=IMAGE0

Instruction: Replace the BMW with an Audi and cloudy sky with clear sky
Program:

```



```

OBJ0=Seg(image=IMAGE)
OBJ1=Select(image=IMAGE, object=OBJ0, query='BMW')
IMAGE0=Replace(image=IMAGE, object=OBJ1, prompt='Audi')
OBJ1=Seg(image=IMAGE0)
OBJ2=Select(image=IMAGE0, object=OBJ1, query='cloudy sky')
IMAGE1=Replace(image=IMAGE0, object=OBJ2, prompt='clear sky')
RESULT=IMAGE1

```

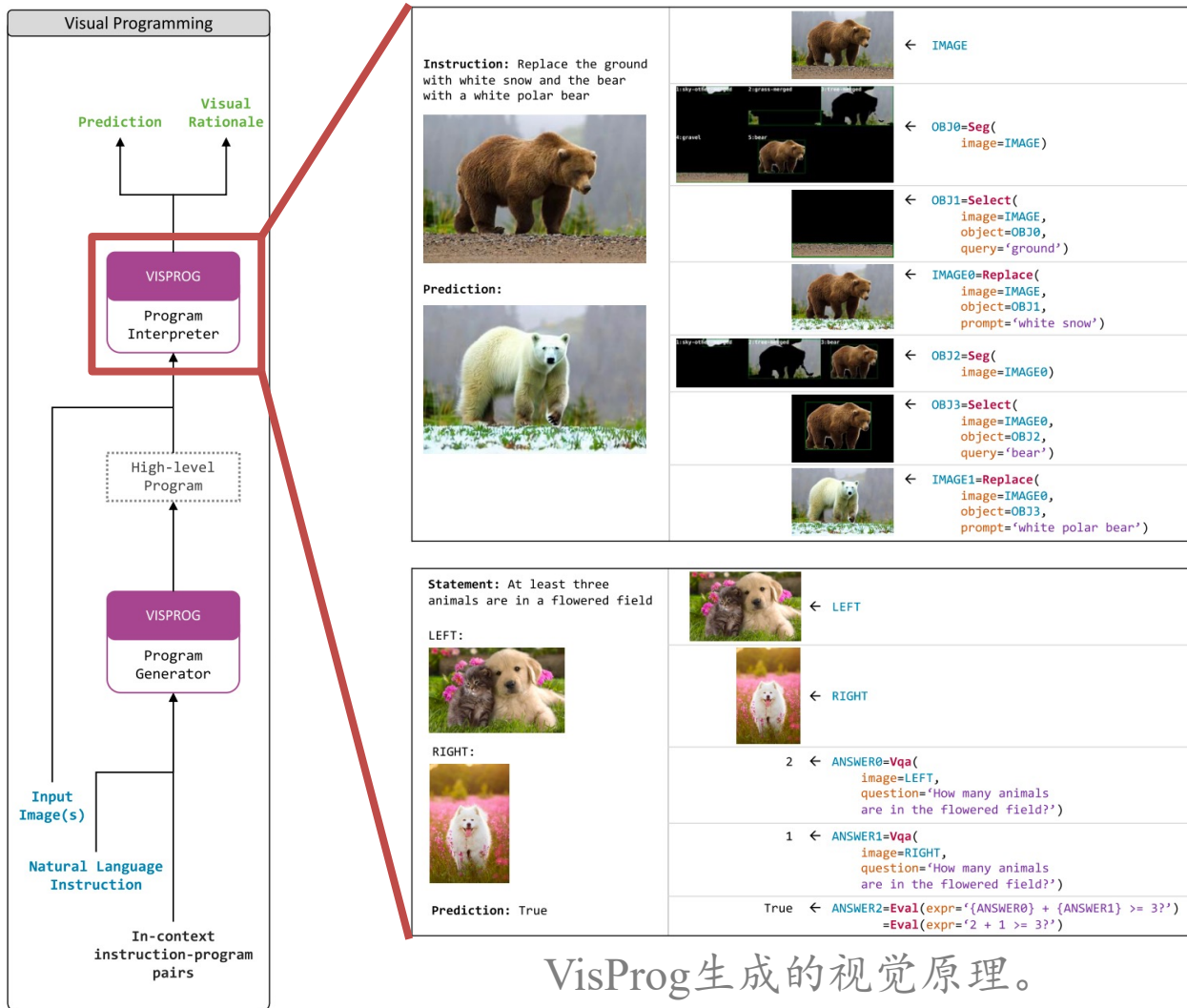
VisProg的程序生成过程。

Image Understanding	Loc OWL-ViT	FaceDet DSFD (pypi)	Seg MaskFormer	Select CLIP-ViT	Classify CLIP-ViT	Vqa ViLT
Image Manipulation	Replace Stable Diffusion	ColorPop PIL.convert() cv2.grabCut()	BgBlur PIL.GaussianBlur() cv2.grabCut()	Tag PIL.rectangle() PIL.text()	Emoji AugLy (pypi)	
	Crop PIL.crop()	CropLeft PIL.crop()	CropRight PIL.crop()	CropAbove PIL.crop()	CropBelow PIL.crop()	
Knowledge Retrieval	List GPT3	Arithmetic & Logical	Eval eval()	Count len()	Result dict()	

VisProg已有的所支持的功能模块。

多模态问答式基础模型的可解释性

VisProg CVPR 2023 Best Paper 美国Allen Institute for AI (西雅图)



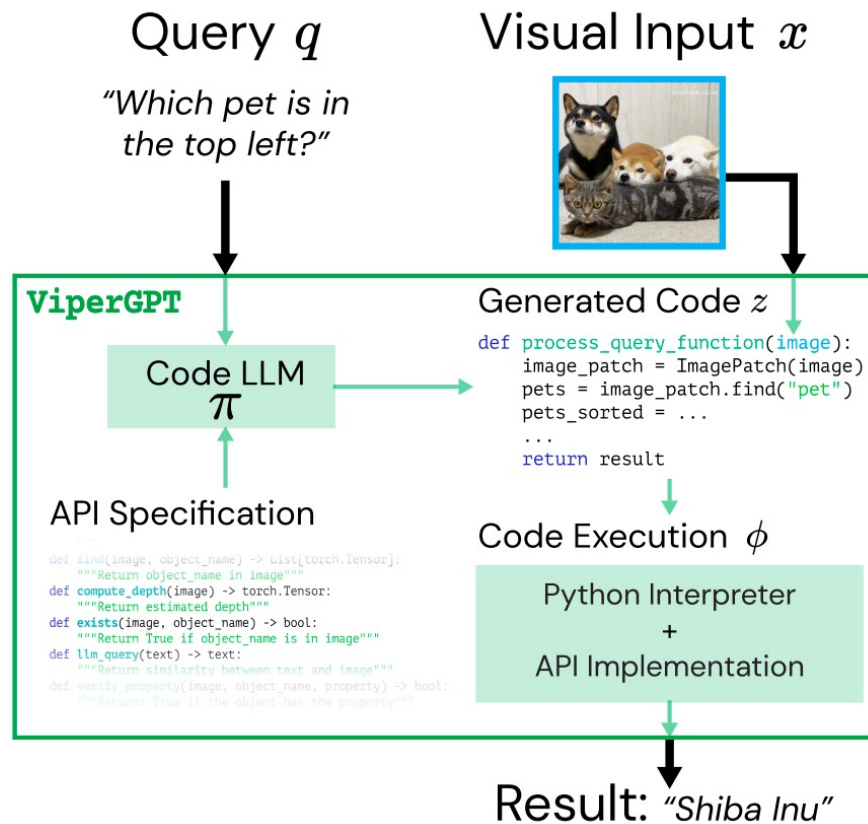
Task	Input	Output	Modules
Compositional Visual QA (GQA)	Image + Question	Text	Loc, Vqa, Eval, Count, Crop, CropLeft, CropRight, CropAbove, CropBelow
Reasoning on Image Pairs (NLVR)	Image Pair + Statement	True/False	Vqa, Eval
Factual Knowledge Object Tagging	Image + Instruction	Image	FaceDet, List, Classify, 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。

VisProg生成的视觉原理。

多模态问答式基础模型的可解释性

ViperGPT 美国 Columbia大学



ViperGPT是一个以编程方式解决复杂可视化查询的框架。

多模态问答式基础模型的可解释性

ViperGPT 美国 Columbia大学

Query: How many muffins can each kid have for it to be fair?



Generated Code

```
def execute_command(image):  
    image_patch = ImagePatch(image)  
    muffin_patches = image_patch.find("muffin")  
    kid_patches = image_patch.find("kid")  
    return str(len(muffin_patches) // len(kid_patches))
```

Execution



```
kid_patches =  
image_patch.find("kid")  
▶ len(muffin_patches)=8  
▶ len(kid_patches)=2  
▶ 8//2 = 4  
Result: 4
```

Query: Drink with zero alcohol



```
def execute_command(image):  
    image_patch = ImagePatch(image)  
    drink_patches = image_patch.find("drink")  
    for drink_patch in drink_patches:  
        drink_name = drink_patch.simple_query("What is this?")  
        alcoholic = llm_query(f"Does the {drink_name} have alcohol?")  
        if alcoholic == "no":  
            return drink_patch  
    return None
```



```
▶ drink_name = 'tullamore dew'  
▶ alcoholic = 'yes'  
▶ drink_name = 'bacardi'  
▶ alcoholic = 'yes'  
▶ drink_name = 'gin'  
▶ alcoholic = 'yes'  
▶ drink_name = 'dr pepper'  
▶ alcoholic = 'no'
```

Result: 

Query: What would the founder of the brand of the car on the left say to the founder of the brand of the car on the right?



```
def execute_command(image):  
    image_patch = ImagePatch(image)  
    car_patches = image_patch.find("car")  
    car_patches.sort(key=lambda car: car.horizontal_center)  
    left_car = car_patches[0]  
    right_car = car_patches[-1]  
    left_car_brand = left_car.simple_query("What is the brand of this car?")  
    right_car_brand = right_car.simple_query("What is the brand of this car?")  
    left_car_founder = llm_query(f"Who is the founder of {left_car_brand}?")  
    right_car_founder = llm_query(f"Who is the founder of {right_car_brand}?")  
    return llm_query(f"What would {left_car_founder} say to {right_car_founder}?")
```

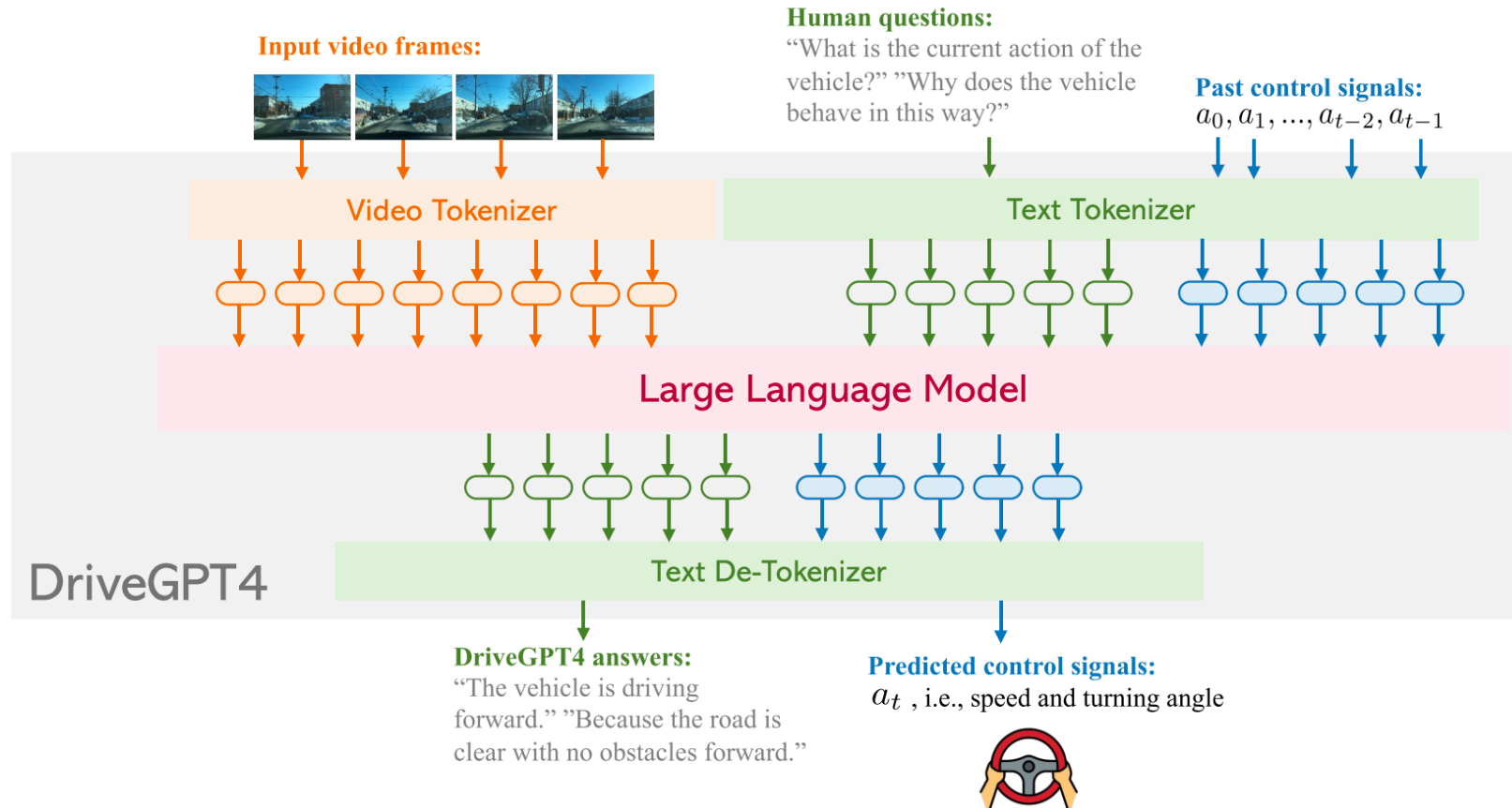


```
▶ left_car_brand='Lamborghini'  
▶ right_car_brand='Ferrari'  
▶ left_car_founder='Ferruccio Lamborghini'  
▶ right_car_founder='Enzo Ferrari'
```

Result: "Ferruccio Lamborghini might say, 'It's been an honor to be a rival of yours for so many years, Enzo. May our cars continue to push each other to be better and faster!'"

多模态问答式基础模型的可解释性

DriveGPT4 香港大学、浙江大学、华为Noah's Ark Lab、悉尼大学



多模态问答式基础模型的可解释性

DriveGPT4 香港大学、浙江大学、华为Noah's Ark Lab、悉尼大学



Human: This is a 8-frame video. In this video, you are sitting in a vehicle on the road. The vehicle speed (m/s) of each frame is 9.86 9.1 8.18 7.24 6.18 5.21 4.22 3.11. The vehicle driving direction (degree) of each frame is 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0. what is the current action of this vehicle?

Assistant: The car slows down to a stop.

Human: What's driving the vehicle to behave in this way?

Assistant: since the light ahead became red.

Human: Forecast the speed and turning angle of the vehicle in the ensuing frame.

Assistant: Speed: 2.09; Turning angle: 0.0

多模态问答式基础模型的可解释性

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？

三、我们思考的一些方法

参考文献

- [1] Explainable Artificial Intelligence, <https://www.darpa.mil/program/explainable-artificial-intelligence>
- [2] Feature Visualization, <https://distill.pub/2017/feature-visualization/>
- [3] Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)." *ICML*, 2018.
- [4] Ramaswamy, Vikram V., et al. "Overlooked Factors in Concept-Based Explanations: Dataset Choice, Concept Learnability, and Human Capability." *CVPR*. 2023.
- [5] Sarkar, Anirban, et al. "A framework for learning ante-hoc explainable models via concepts." *CVPR*. 2022.
- [6] Arrieta, Alejandro Barredo, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." *Information Fusion* 58 (2020): 82-115.
- [7] Hendricks, Lisa Anne, et al. "Generating visual explanations." *ECCV*, 2016.
- [8] Chen, Chaofan, et al. "This looks like that: deep learning for interpretable image recognition." *NeurIPS* 32 (2019).
- [9] Wang, Pei, and Nuno Vasconcelos. "Scout: Self-aware discriminant counterfactual explanations." *CVPR*. 2020.
- [10] Wang, Tan, et al. "Causal attention for unbiased visual recognition." *ICCV*. 2021.
- [11] Zhao, Haiyan, et al. "Explainability for Large Language Models: A Survey." *arXiv preprint arXiv:2309.01029* (2023).
- [12] Dong, Qingxiu, et al. "A survey for in-context learning." *arXiv preprint arXiv:2301.00234* (2022).
- [13] Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." *NeurIPS* 35 (2022): 24824-24837.
- [14] Wang, Xuezhi, et al. "Self-Consistency Improves Chain of Thought Reasoning in Language Models." *ICLR*. 2023.
- [15] Yao, Shunyu, et al. "Tree of thoughts: Deliberate problem solving with large language models." *arXiv preprint arXiv:2305.10601* (2023).
- [16] Creswell, Antonia, Murray Shanahan, and Irina Higgins. "Selection-Inference: Exploiting Large Language Models for Interpretable Logical Reasoning." *ICLR*. 2022.
- [17] Trenton Bricken, *et al.*, "Towards Monosemanticity: Decomposing Language Models With Dictionary Learning." <https://transformer-circuits.pub/2023/monosemantic-features>. 2023.
- [18] Yuksekgonul, Mert, Maggie Wang, and James Zou. "Post-hoc Concept Bottleneck Models." *ICLR*. 2023.
- [19] Menon, Sachit, and Carl Vondrick. "Visual Classification via Description from Large Language Models." *ICLR*. 2022.

参考文献

- [20] Pratt, Sarah, et al. "What does a platypus look like? generating customized prompts for zero-shot image classification." *CVPR*. 2023.
- [21] Yang, Yue, et al. "Language in a bottle: Language model guided concept bottlenecks for interpretable image classification." *CVPR*. 2023.
- [22] Chen, Haozhe, et al. "Interpreting and Controlling Vision Foundation Models via Text Explanations." *arXiv preprint arXiv:2310.10591* (2023).
- [23] Gandelsman, Yossi, Alexei A. Efros, and Jacob Steinhardt. "Interpreting CLIP's Image Representation via Text-Based Decomposition." *arXiv preprint arXiv:2310.05916* (2023).
- [24] Wu, Chenwang, et al. "A Causality Inspired Framework for Model Interpretation." *KDD*. 2023.
- [25] Swamy, Vinitra, et al. "MultiModN-Multimodal, Multi-Task, Interpretable Modular Networks." *arXiv preprint arXiv:2309.14118* (2023).
- [26] Kalibhat, Neha, et al. "Identifying Interpretable Subspaces in Image Representations." (2023).
- [27] Pratt, Sarah, et al. "Cross-modality Interpretable image classification via Concept Decomposition Vector of Visual Language Models." (2023).
- [28] Gupta, Tanmay, and Aniruddha Kembhavi. "Visual programming: Compositional visual reasoning without training." *CVPR*. 2023.
- [29] Surís, Dídac, Sachit Menon, and Carl Vondrick. "ViperGPT: Visual inference via python execution for reasoning." *ICCV* (2023).
- [30] Xu, Zhenhua, et al. "DriveGPT4: Interpretable End-to-end Autonomous Driving via Large Language Model." *arXiv preprint arXiv:2310.01412* (2023).

谢谢各位聆听，
敬请批评指正！

汇报人：陈若愚

日期：2023.10.20