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

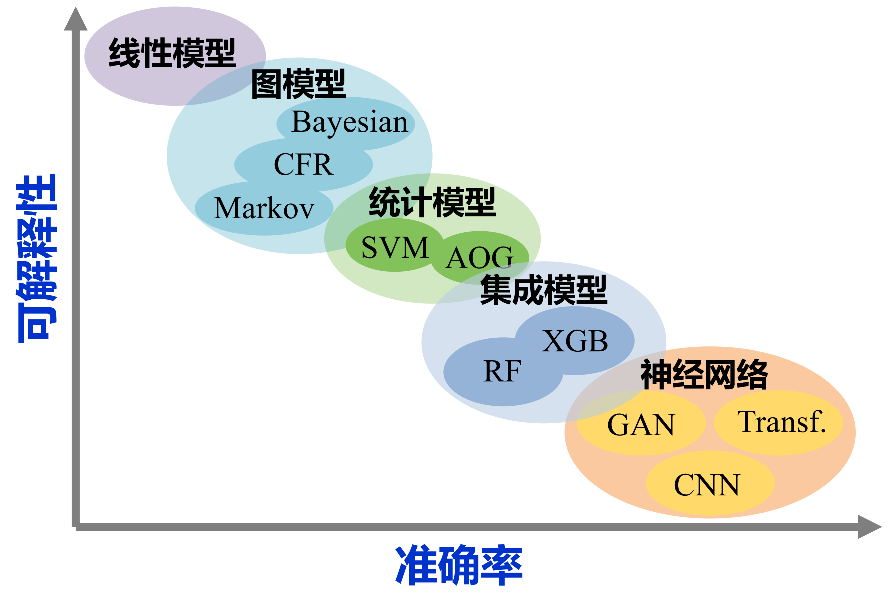
■ 基础模型的特点及其新挑战

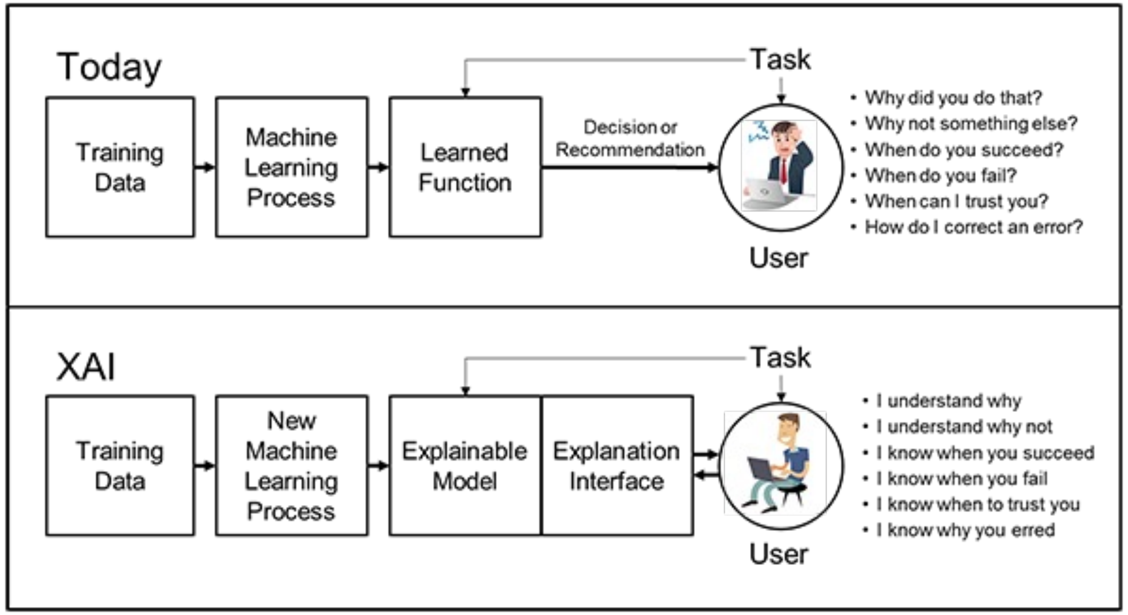
一、传统可解释AI与基础模型

- 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。

DARPA

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)

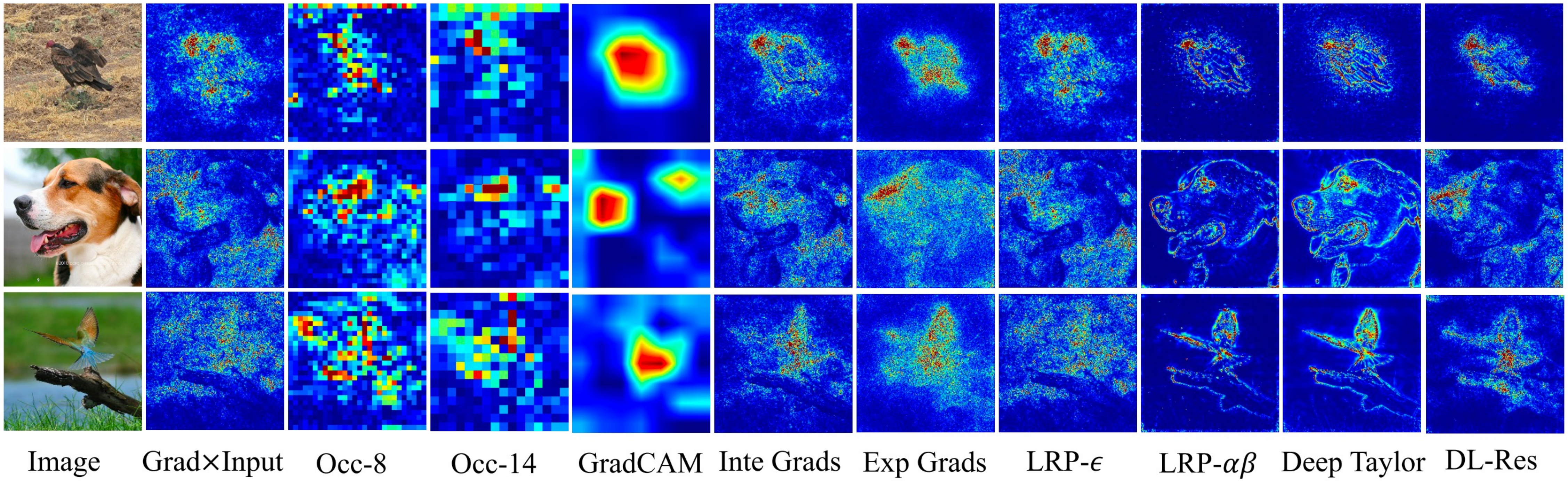
- 解释一个预训练模型或其决策的结果；
- 在模型做完决策后提供的解释。

传统可解释AI方法简介

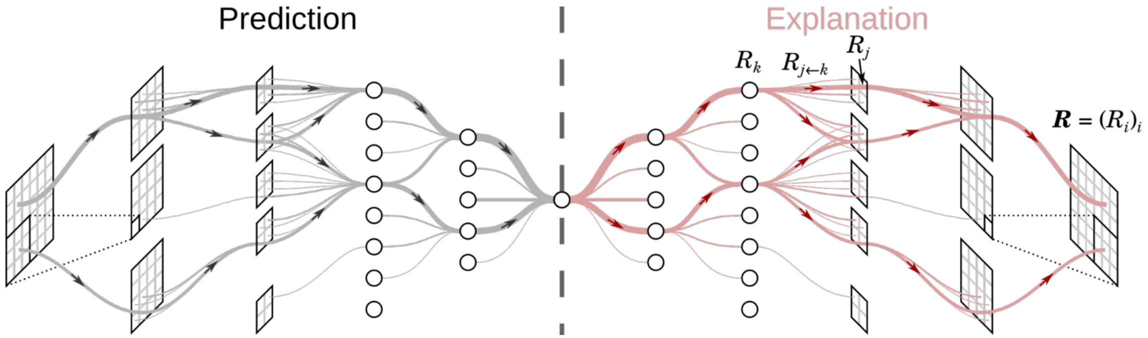
- 基于归因的方法
- 基于特征可视化的方法
- 基于语义概念的方法
- 基于关联模型的方法
- 基于其他模态模型的方法
- 基于设计的方法
- 基于因果的方法

传统可解释AI方法简介

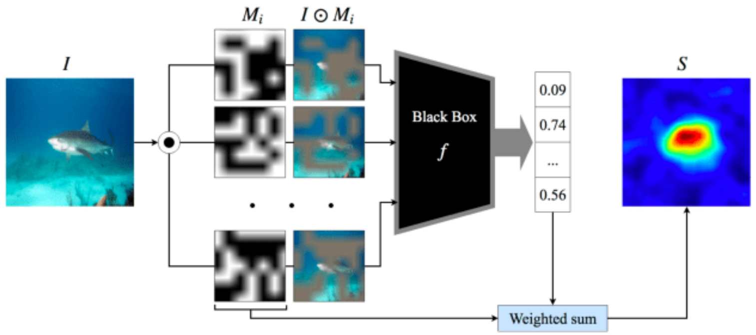
基于归因的方法



基于模型内部机理（白盒）

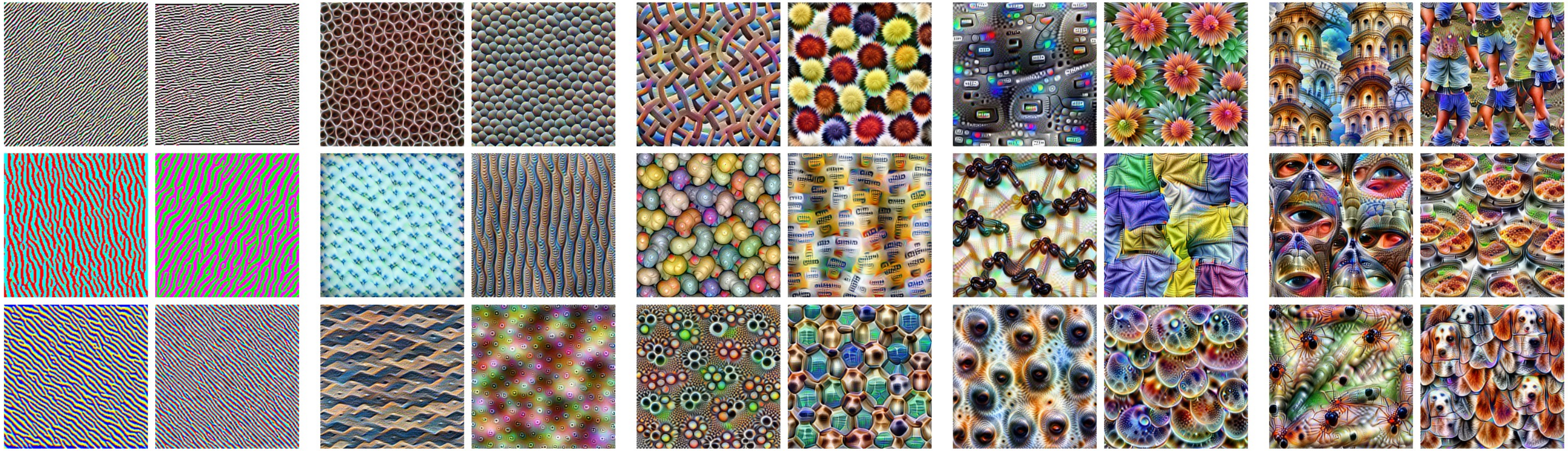


基于扰动（黑盒）

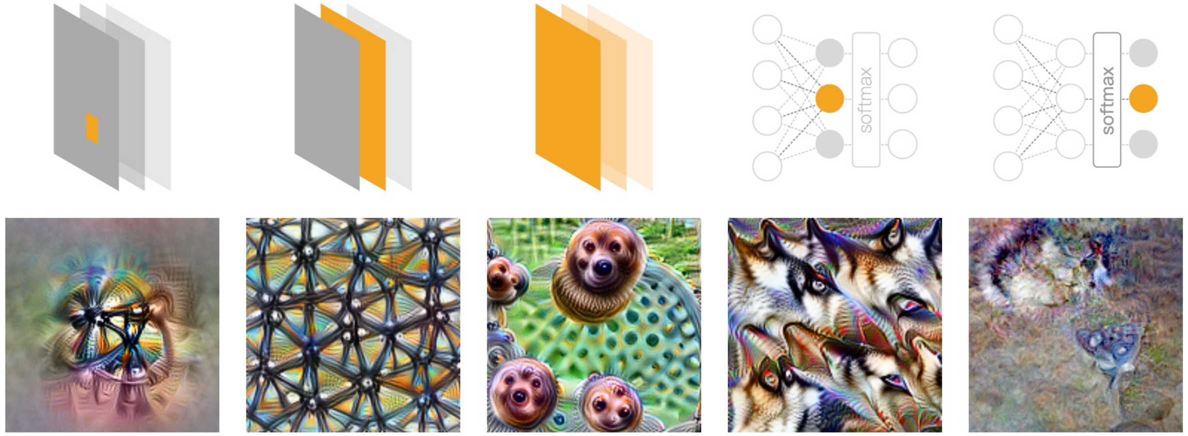


传统可解释AI方法简介

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)



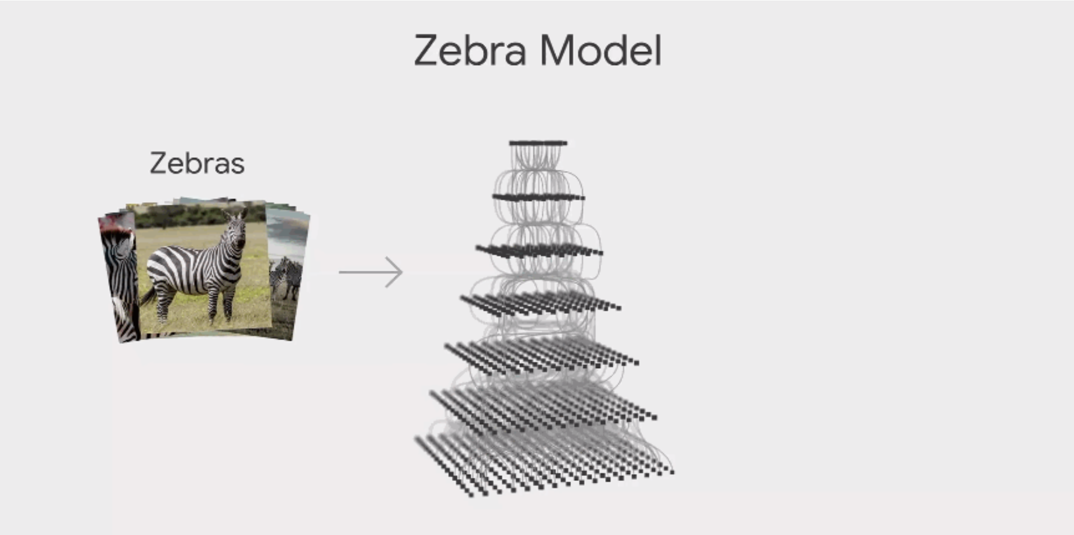
Neuron	Channel	Layer/DeepDream	Class Logits	Class Probability
$\text{layer}_n[x,y,z]$	$\text{layer}_n[:,:,z]$	$\text{layer}_n[:,:,:]^2$	pre_softmax[k]	softmax[k]

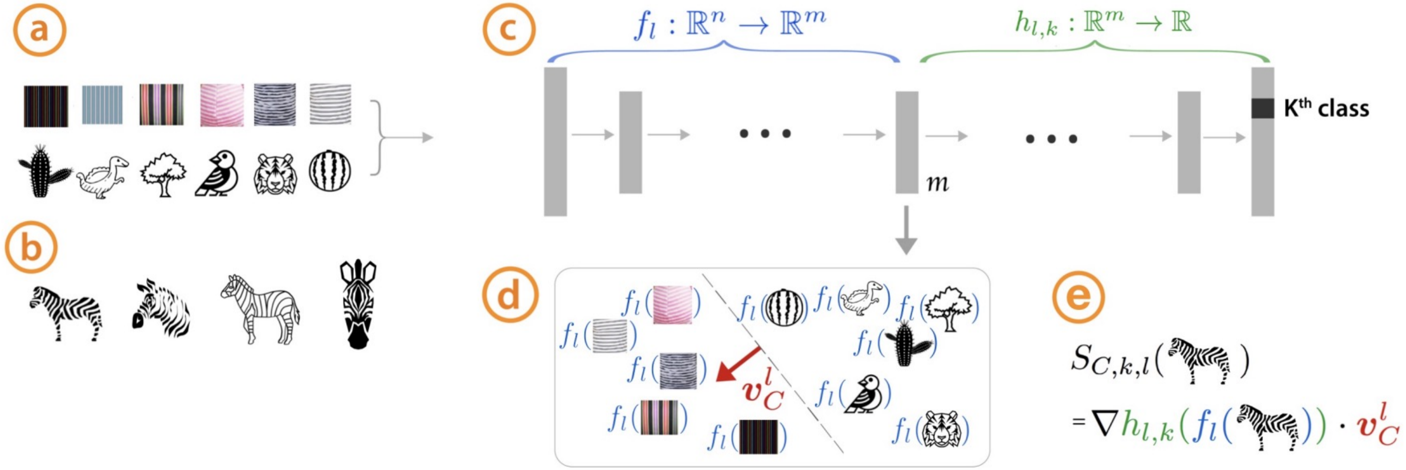
Feature Visulization[2]:
指定中间单元，优化输入，使目标单元有最大激活响应，观察优化的输入图像。

[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：

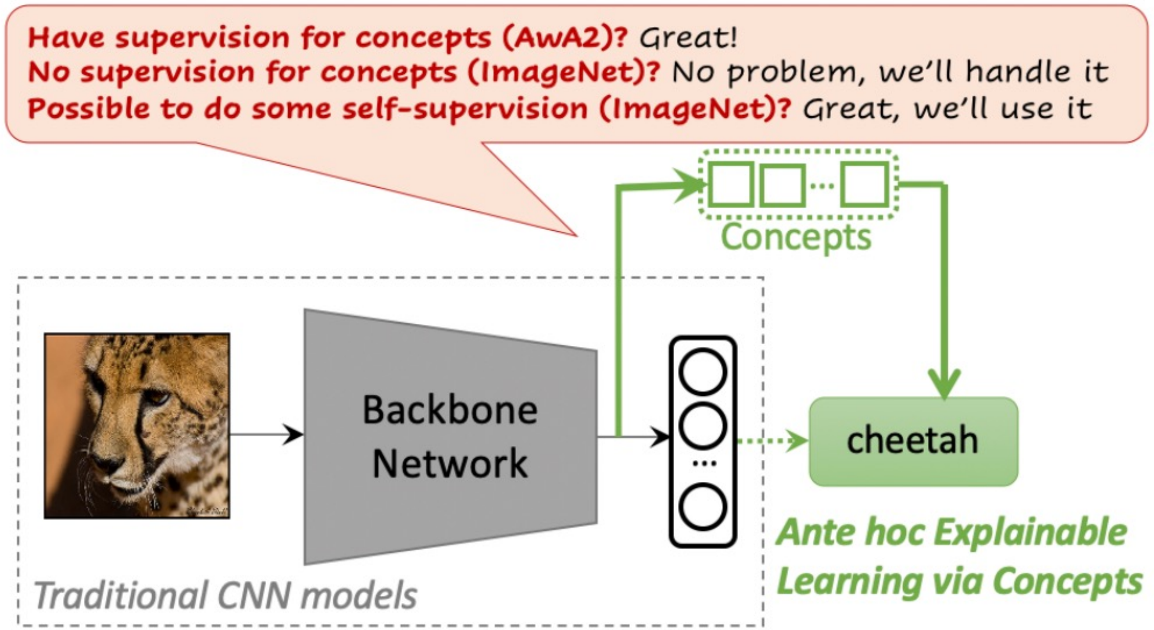
$$TCAV_c = \frac{|x \in \chi^c : S_c(x) > 0|}{|\chi^c|}.$$

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



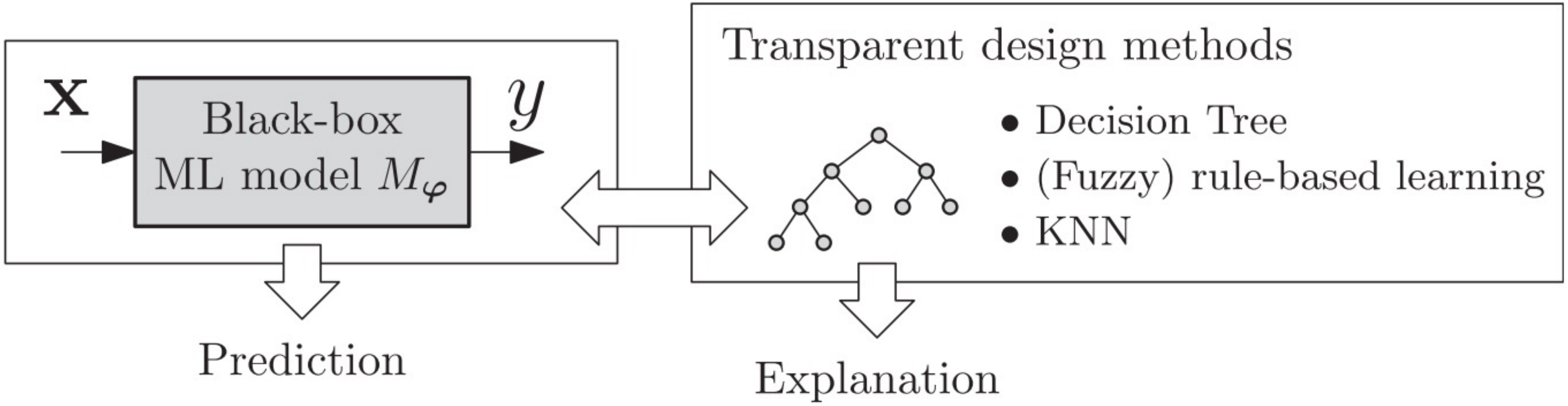
Self-Explaining Neural Networks

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。

[5] Sarkar, Anirban, et al. "A framework for learning ante-hoc explainable models via concepts." *CVPR*. 2022.

传统可解释AI方法简介

基于关联模型的方法

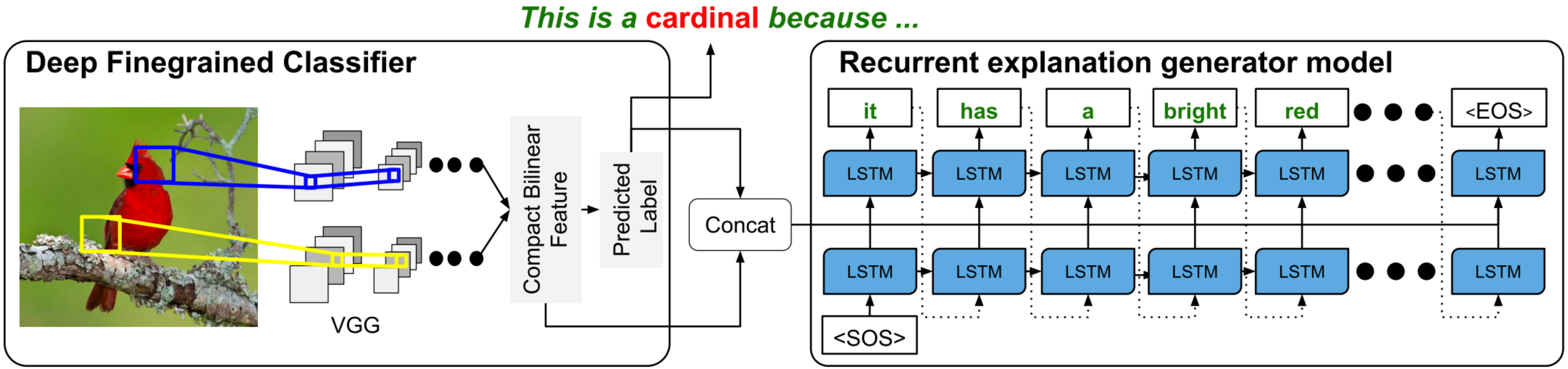


映射一个不可解释的黑盒系统变成了白盒孪生体，更易于解释。但通常会影响最终模型的性能[6]。

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

传统可解释AI方法简介

基于其他模态模型的方法

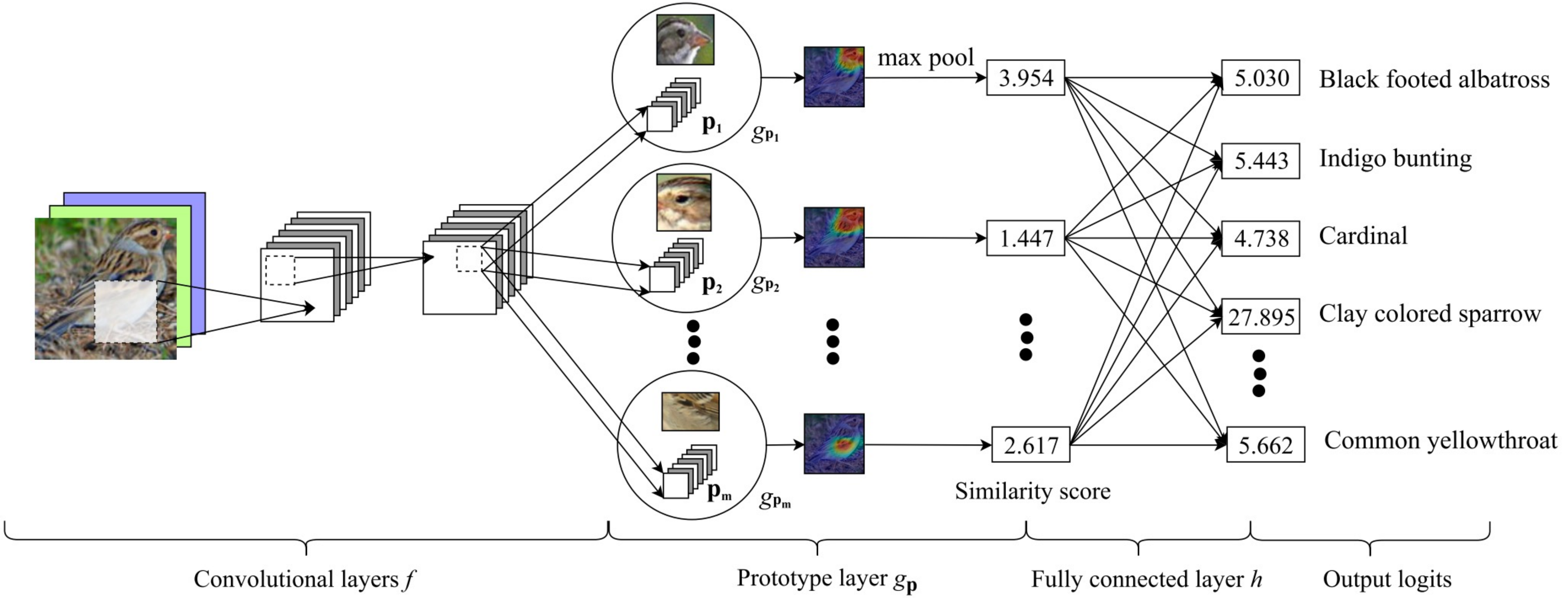


用一个不可解释的模型来解释一个黑盒模型是令人担忧的。

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

传统可解释AI方法简介

基于设计的方法

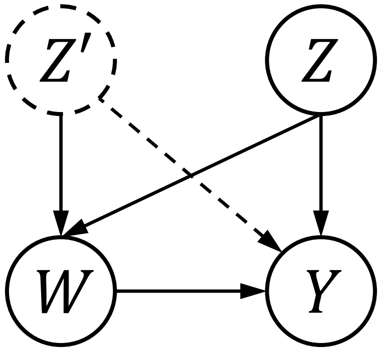


需要给定指定的概念原型特征，通用性、可扩展性差。

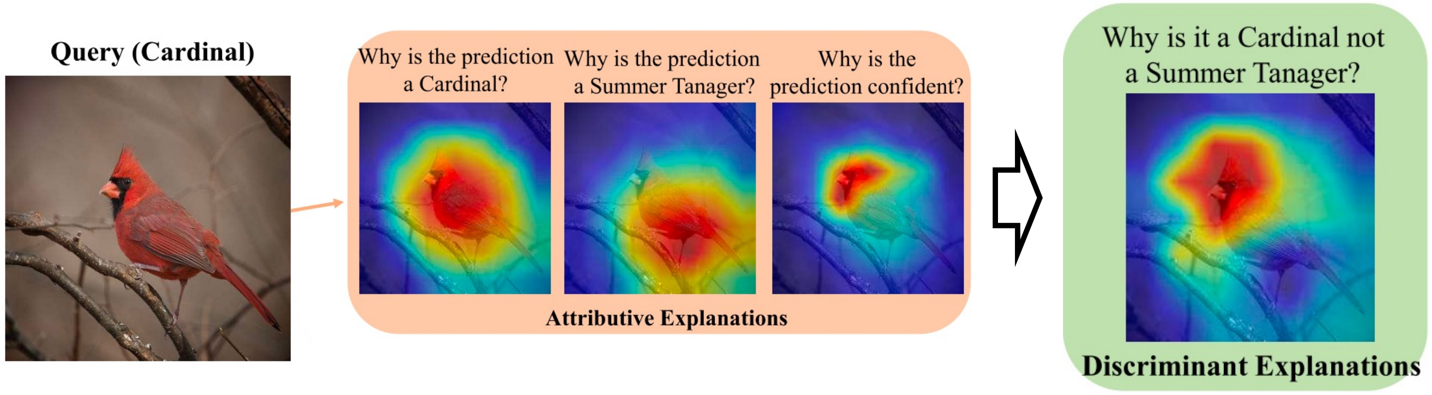
[8] Chen, Chaofan, et al. "This looks like that: deep learning for interpretable image recognition." *NeurIPS* 32 (2019).

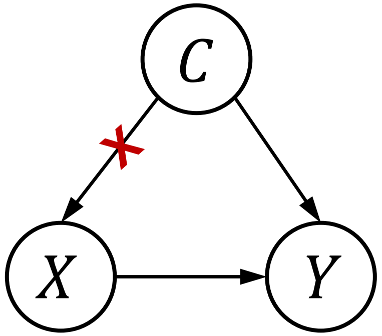
传统可解释AI方法简介

基于因果的方法



反事实推理





因果干预

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

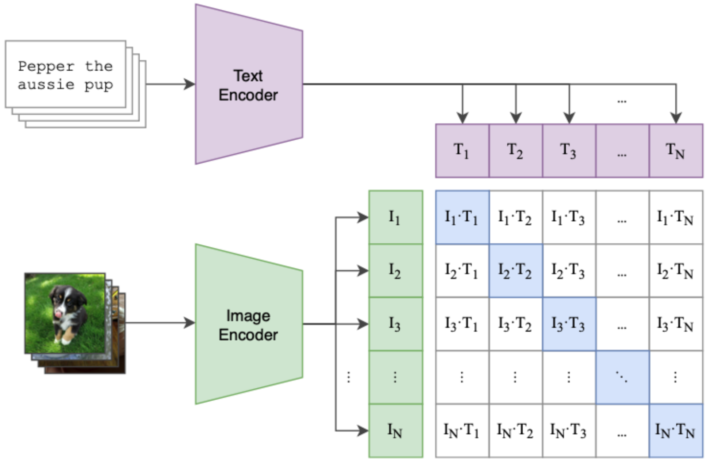
先验的可能导致模型产生偏见的因素

[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.

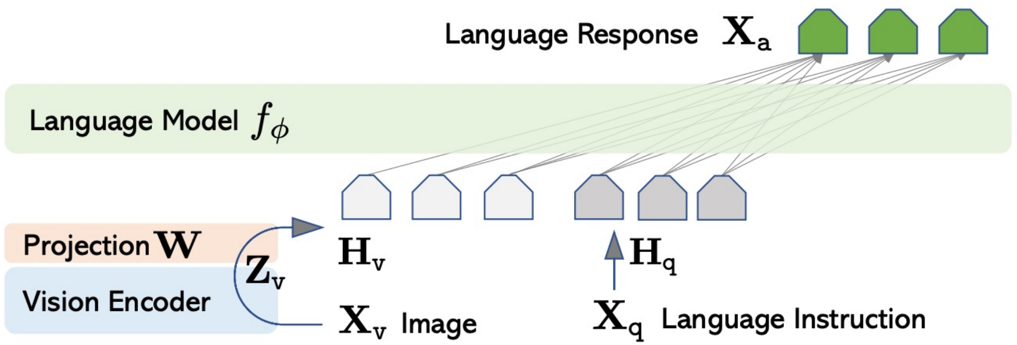
一、传统可解释AI与基础模型

- 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

多模态编码式基础模型：
- ✓ 双流结构
- ✓ 编码式模型
- ✓ Zero-Shot能力

多模态问答式基础模型：
- ✓ 参数量较大
- ✓ 编码+生成式模型
- ✓ 双模态输入
- ✓ 提示学习能力

大语言模型：
- ✓ 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

多模态编码式基础模型：
- □ 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！

多模态问答式基础模型：
- □ 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？

大语言模型：
- □ 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。

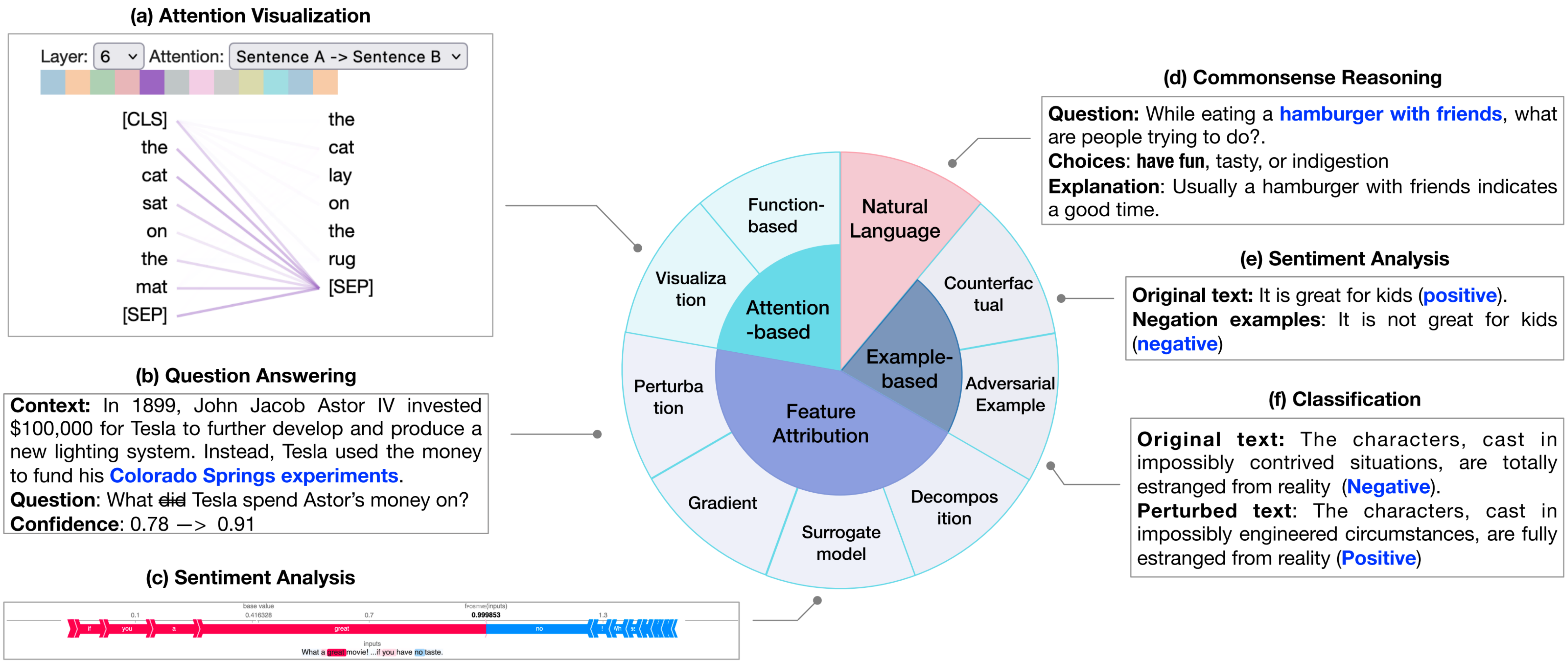
二、基础模型的可解释性研究现状

- 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

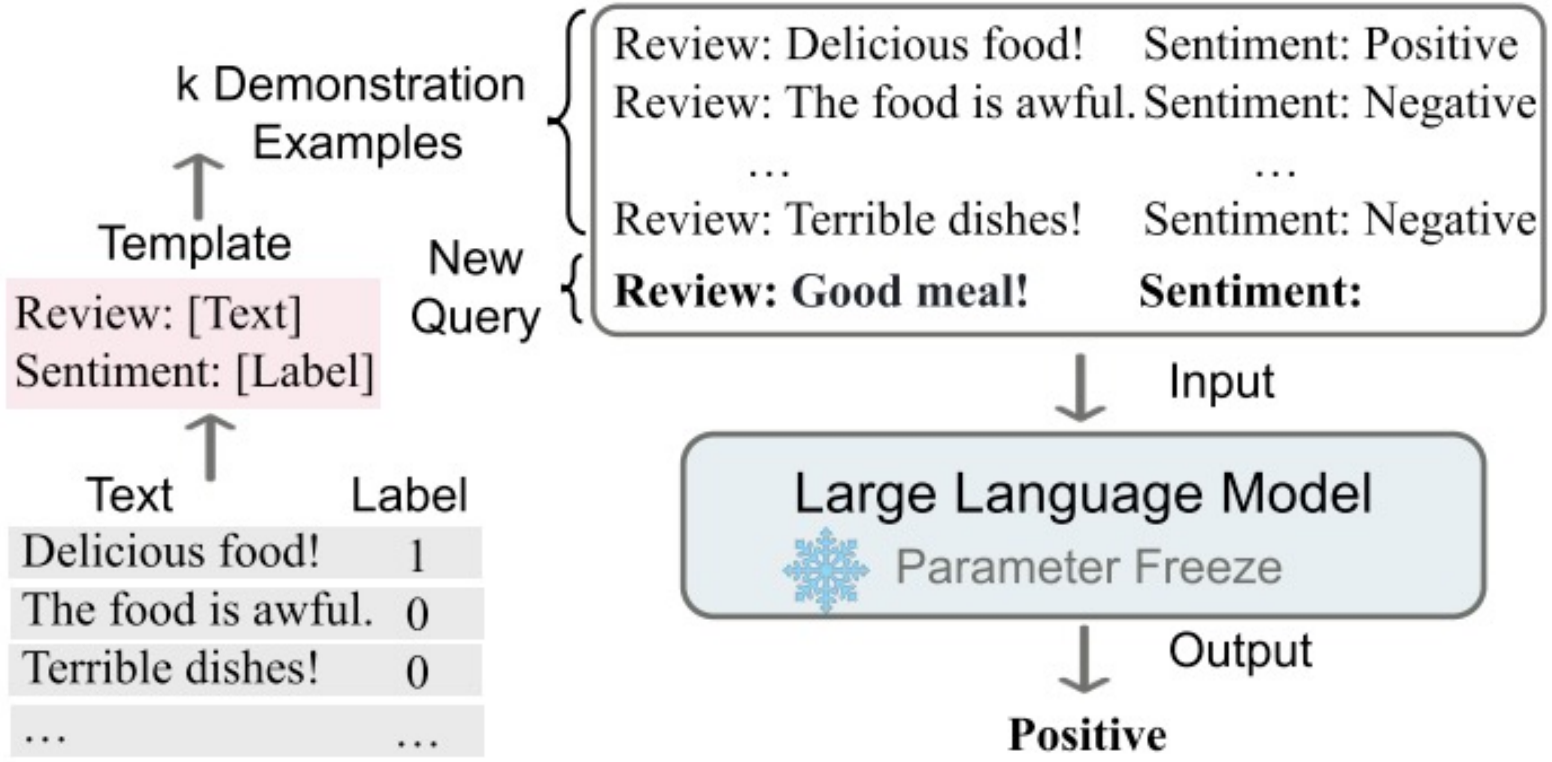
大语言模型的可解释性



传统语言模型的局部可解释方法汇总

[11] Zhao, Haiyan, et al. "Explainability for Large Language Models: A Survey." *arXiv preprint arXiv:2309.01029* (2023).

大语言模型的可解释性

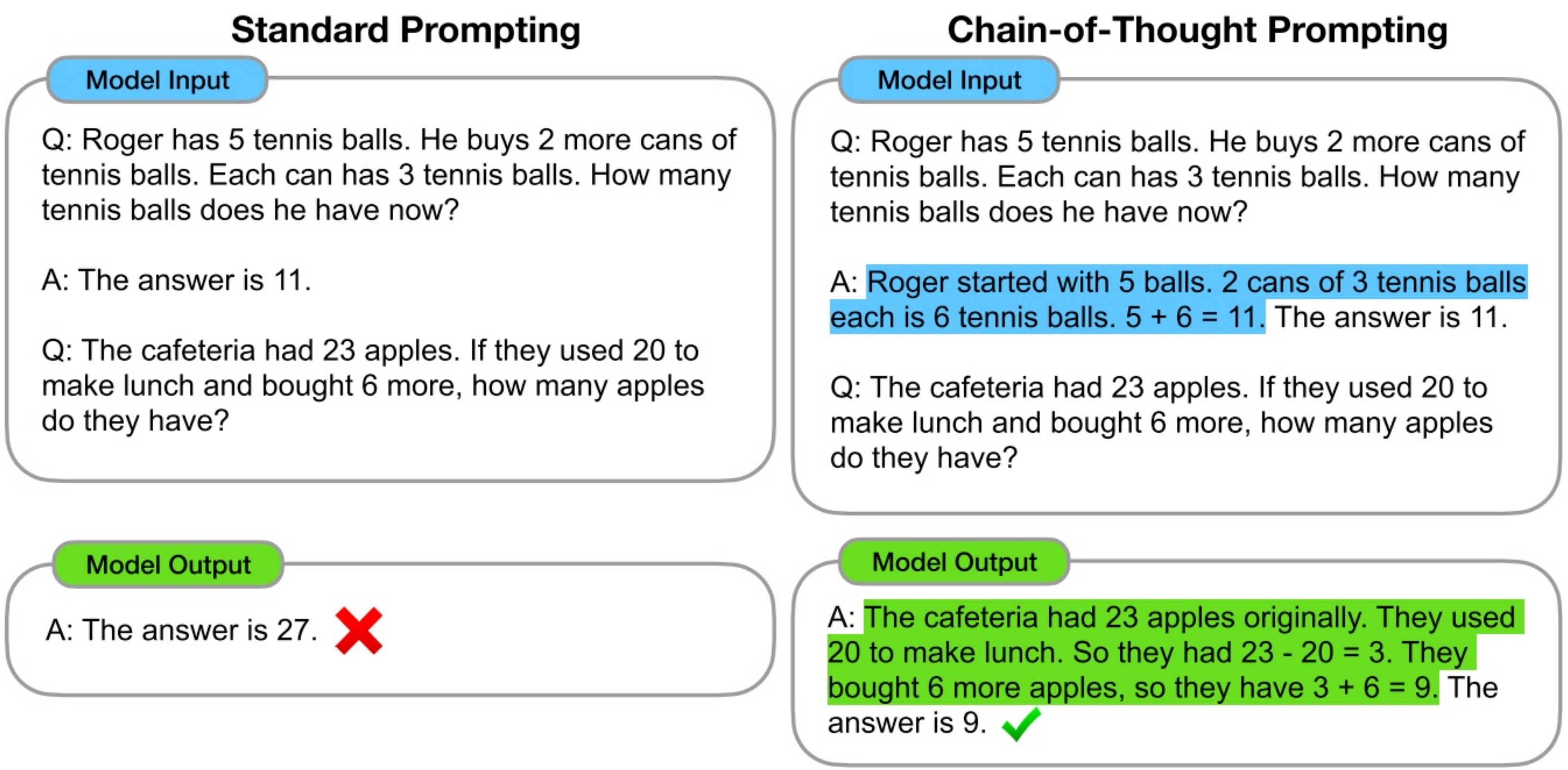


上下文学习的释例。

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

大语言模型的可解释性

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：

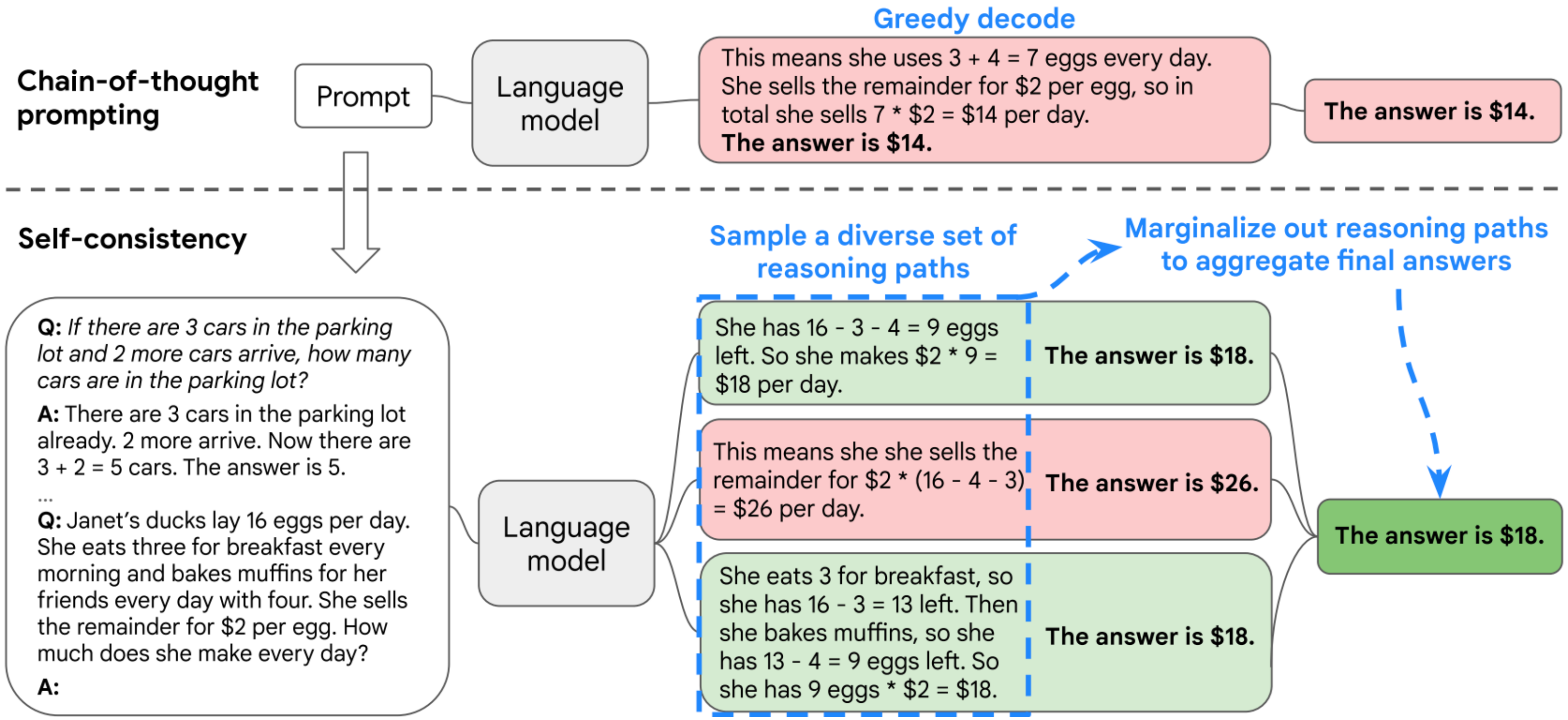
1. 思维链原则上允许模型将多步骤问题分解为中间步骤；
2. 为模型的行为提供了一个**可解释的窗口**，表明如何得出特定的答案，并提供调试推理路径出错的地方的机会；
3. 可能适用于（至少原则上）人类可通过语言解决的任何任务。
4. 在**足够大**的语言模型中，只要将思维链序列的示例包含到少数提示的示例中，就可以很容易地推导出思维链推理。

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。

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

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。

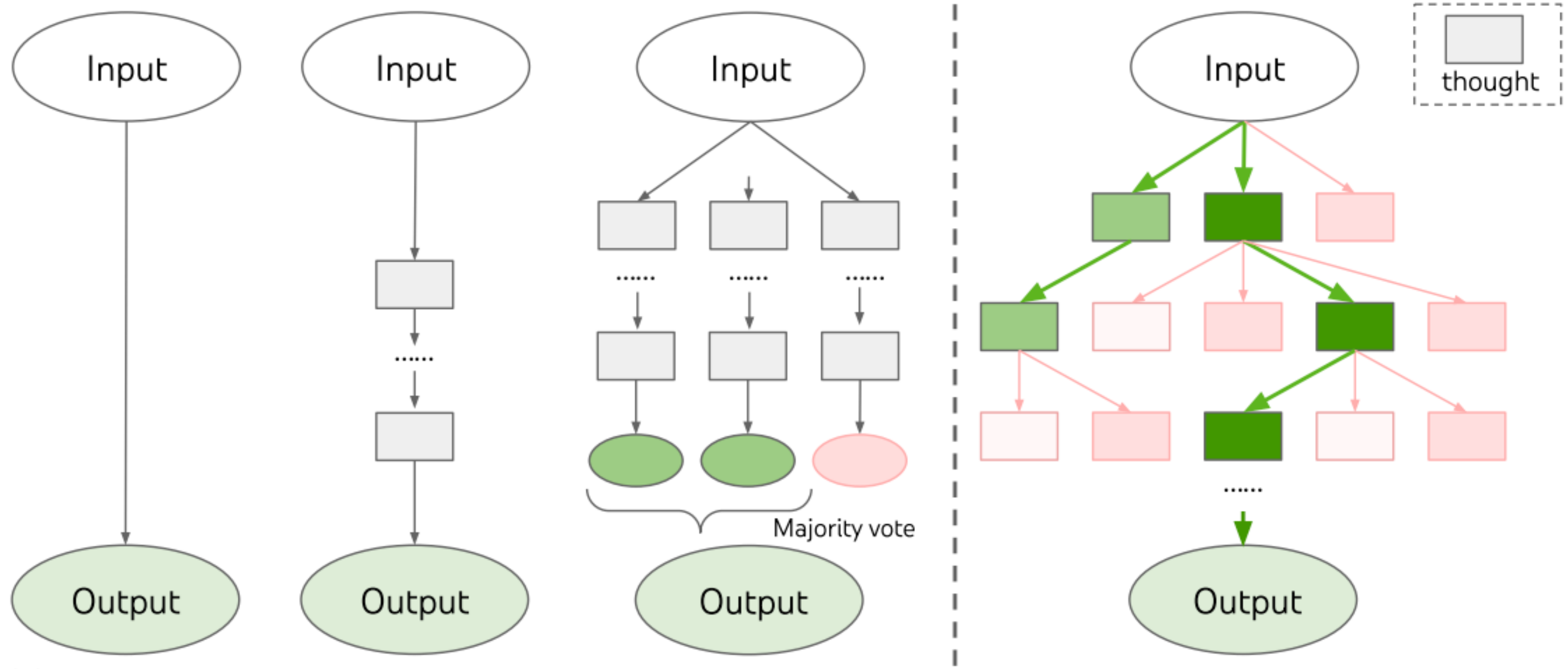
[14] Wang, Xuezhi, et al. "Self-Consistency Improves Chain of Thought Reasoning in Language Models." *ICLR*. 2023.

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.*



(a) Input-Output Prompting (IO) (c) Chain of Thought Prompting (CoT) (c) Self Consistency with CoT (CoT-SC) (d) Tree of Thoughts (ToT)

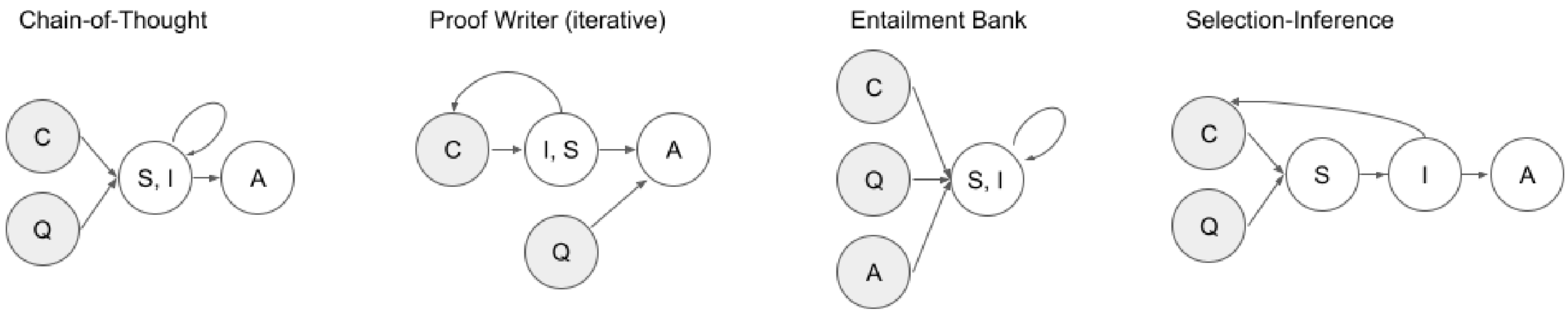
ToT步骤：

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。

[16] Creswell, Antonia, Murray Shanahan, and Irina Higgins. "Selection-Inference: Exploiting Large Language Models for Interpretable Logical Reasoning." *ICLR*. 2022.

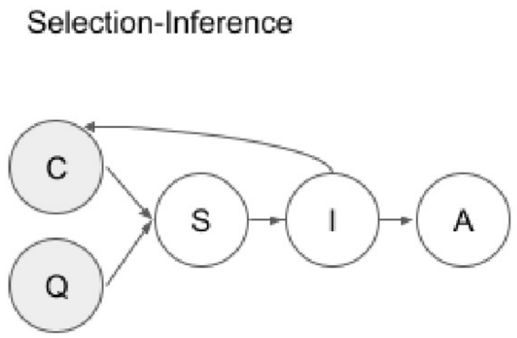
大语言模型的可解释性

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, $\mathcal{C}^0$,made up of statements e.g. facts and rules.
Require: The question, q.
Require: Language model, LLM.
Require: The number of reasoning steps, H.

$t=0$ ▷ Start at step 0.
while $t<H$ **do**
  $s^t \leftarrow$ Selection_Module(p_{select},C^t,q,LLM) ▷ Do selection.
  $i^t \leftarrow$ Inference_Module(p_{infer},s^t) ▷ Do inference.
  $C^{t+1} \leftarrow C^t \cup i^t$ ▷ Add the newly inferred fact to the context.
  $t \leftarrow t+1$ ▷ Move onto the next step of reasoning
end while
return s^t



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,
```

[16] Creswell, Antonia, Murray Shanahan, and Irina Higgins. "Selection-Inference: Exploiting Large Language Models for Interpretable Logical Reasoning." *ICLR*. 2022.

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。

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

大语言模型的可解释性

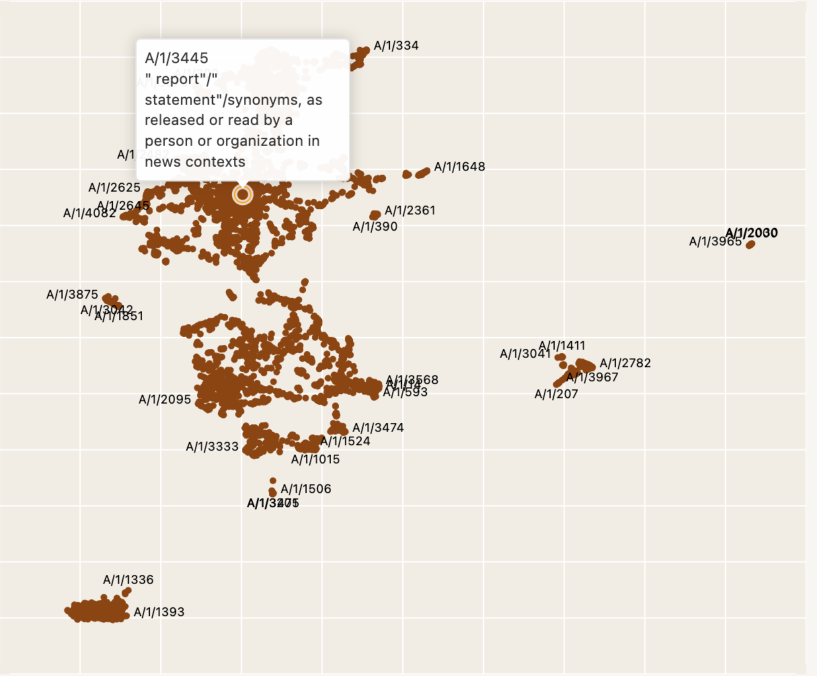
2023.10.4 Claude背后公司Anthropic发布Poster：

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	Feature	Description
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"



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

大语言模型的可解释性

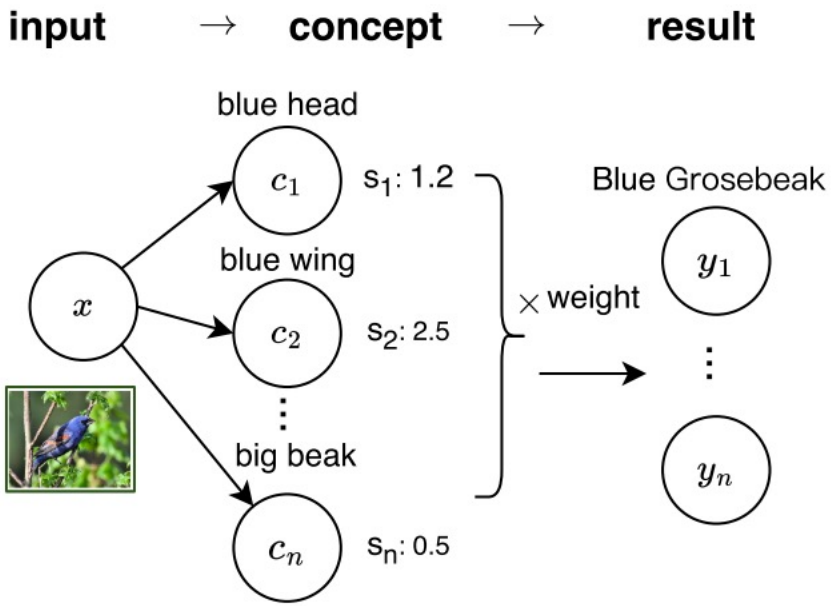
总结

- 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

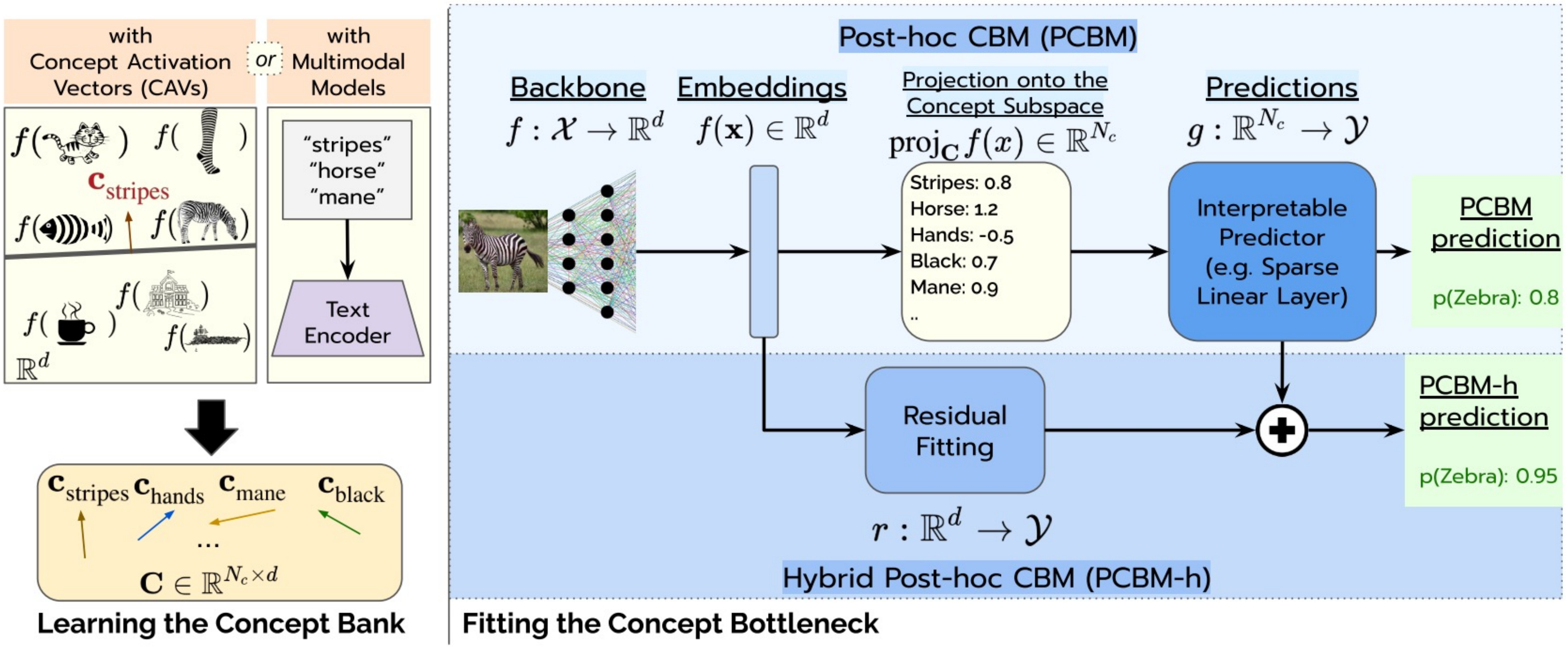


传统的概念瓶颈模型

缺点：

- 如何确定语义概念集？
- 需要人工密集的语义标注。

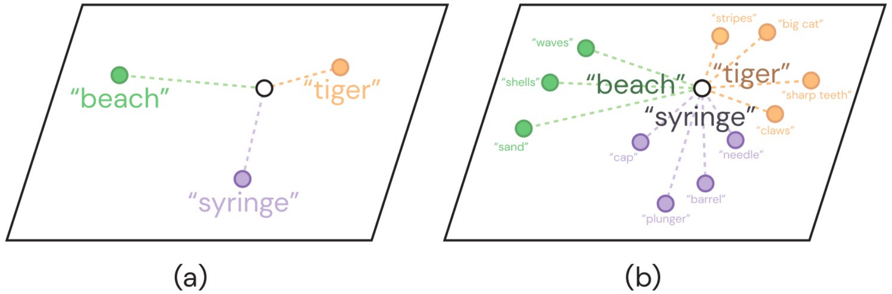
多模态编码式基础模型的可解释性



Post-hoc Concept Bottleneck Models.

[18] Yuksekgonul, Mert, Maggie Wang, and James Zou. "Post-hoc Concept Bottleneck Models." *ICLR*. 2023.

多模态编码式基础模型的可解释性



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

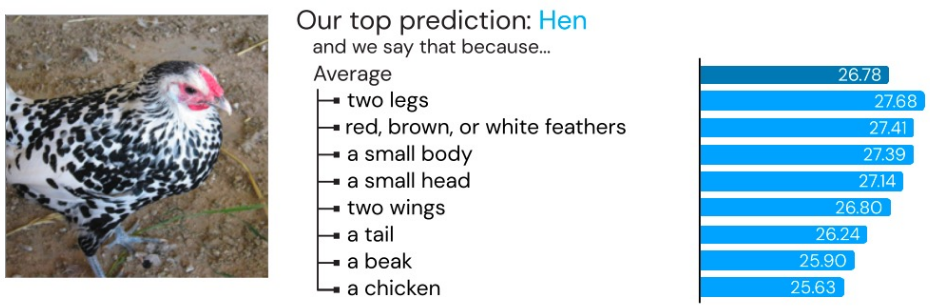
Pirate ship
- a large, sailing vessel
- a flag with a skull and crossbones
- cannons on the deck
- a wooden hull
- portholes
- rigging
- a crow's nest

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 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

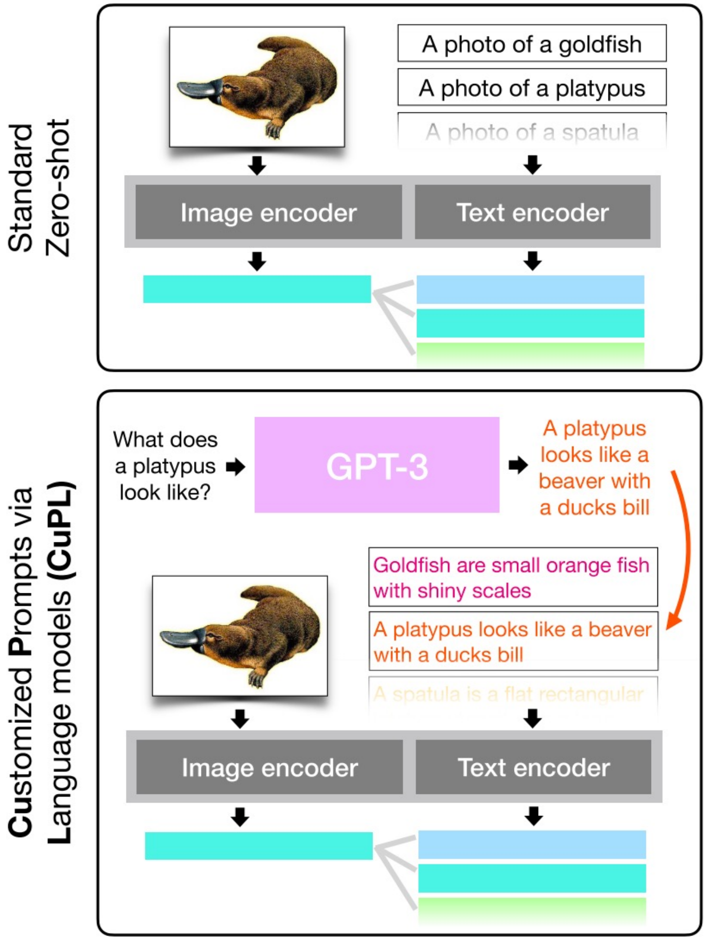
ImageNet和ImageNetV2的模型有一致的~ 3-5%的改进，CUB有~ 1%的改进。

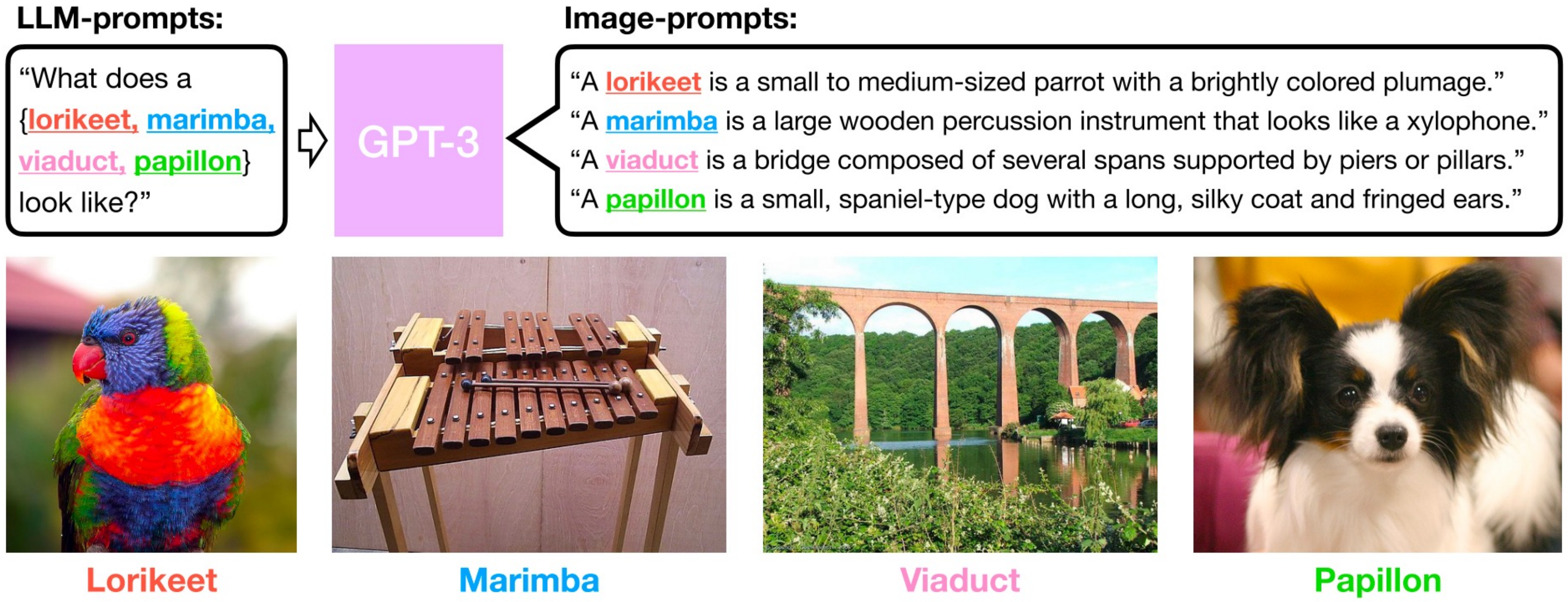


CLIP通过描述符进行决策。

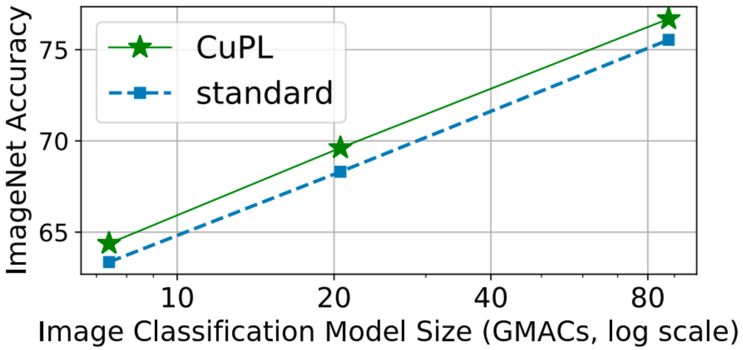
[19] Menon, Sachit, and Carl Vondrick. "Visual Classification via Description from Large Language Models." *ICLR*. 2022.

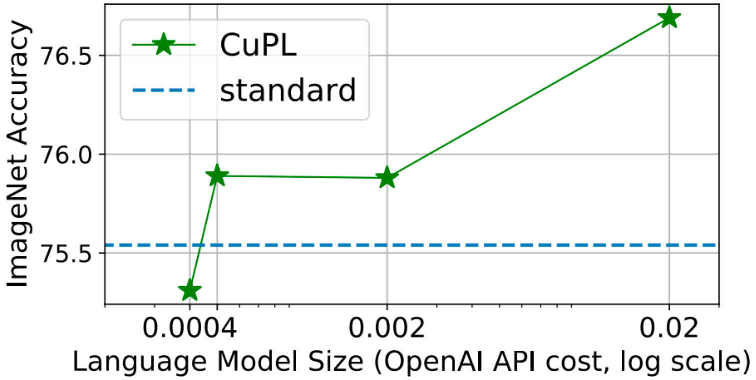
多模态编码式基础模型的可解释性





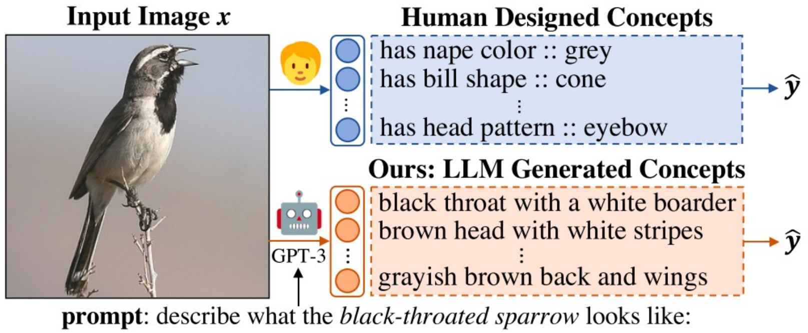
显示了 LLM 生成的图像提示和来自 ImageNet 的相关图像的示例。 仅图像提示用于下游图像分类。



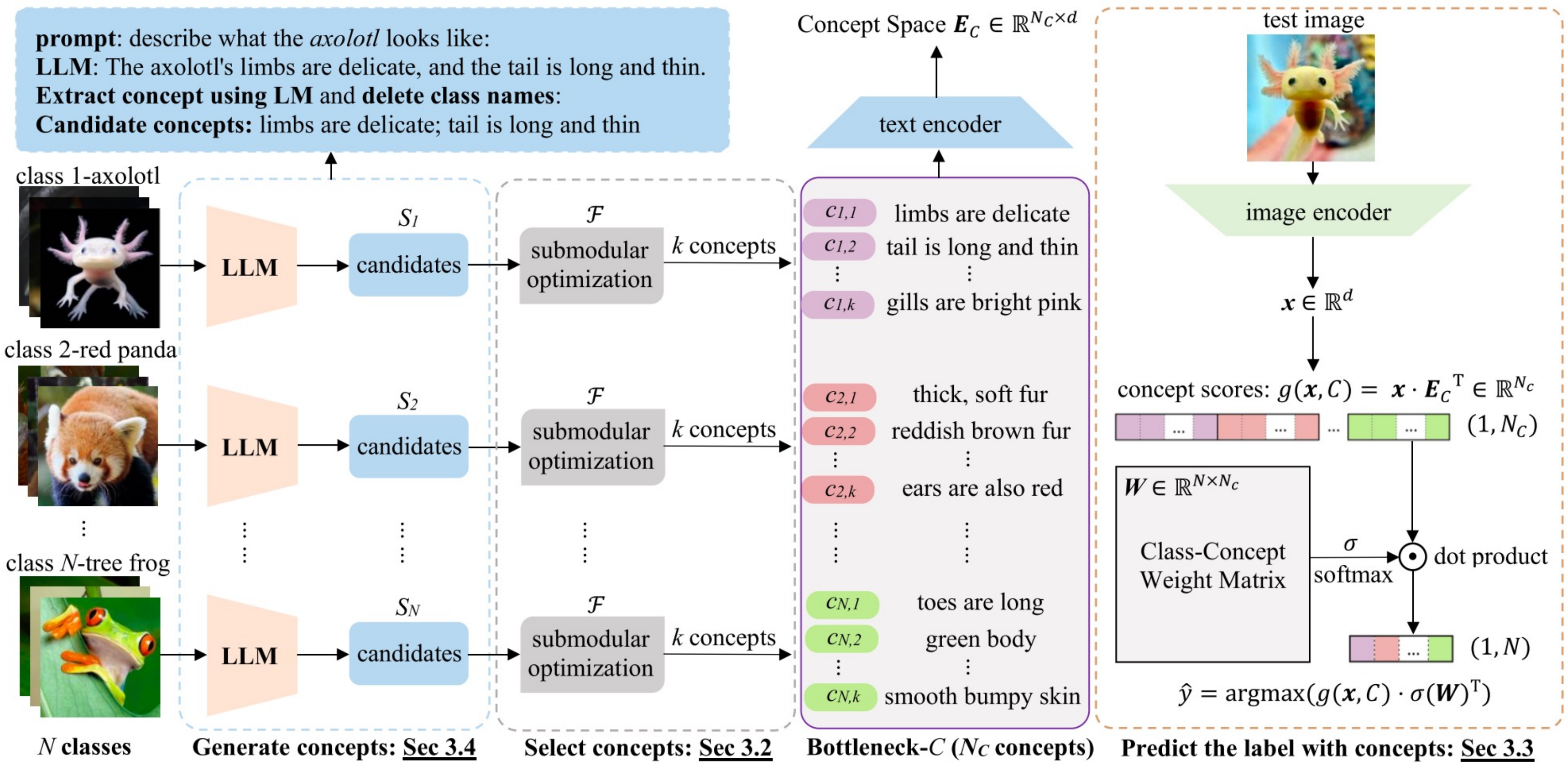


[20] Pratt, Sarah, et al. "What does a platypus look like? generating customized prompts for zero-shot image classification." CVPR. 2023.

多模态编码式基础模型的可解释性

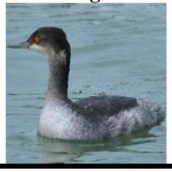
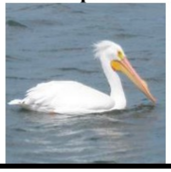
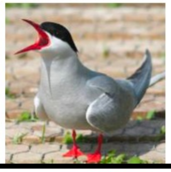
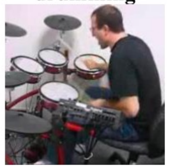
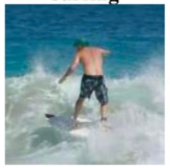
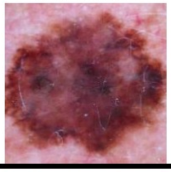


通过提示大型语言模型 (LLM) (如GPT-3)减轻了对人工设计概念的需求



[21] 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

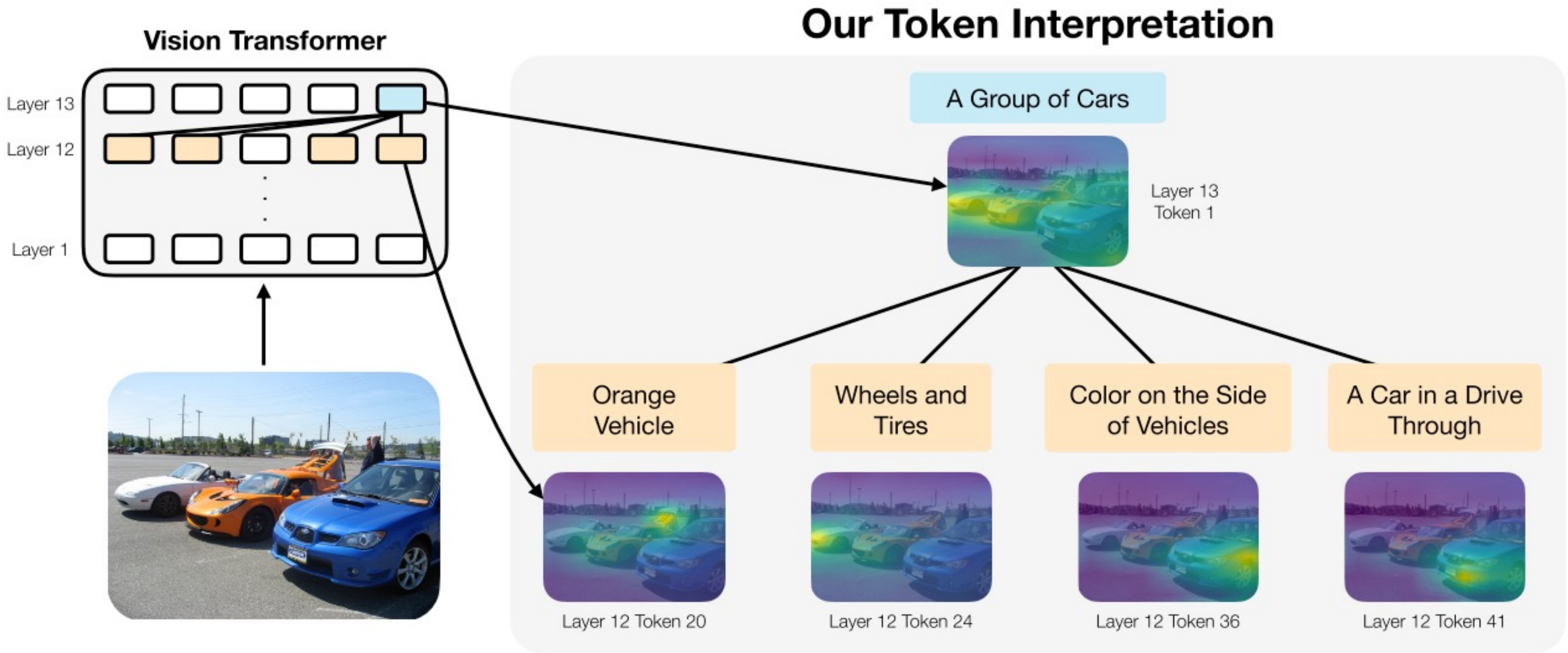
	Class Name	Top-3 Concepts	Class Name	Top-3 Concepts	Class Name	Top-3 Concepts	Class Name	Top-3 Concepts
ImageNet	**badger** 	1.short legs and long body make it an excellent digger 2. black-and-white striped fur 3. coat is very shaggy	**ant** 	1. black and red stinger 2. small, black insect with six legs 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
HAM100000	**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

[21] 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

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。

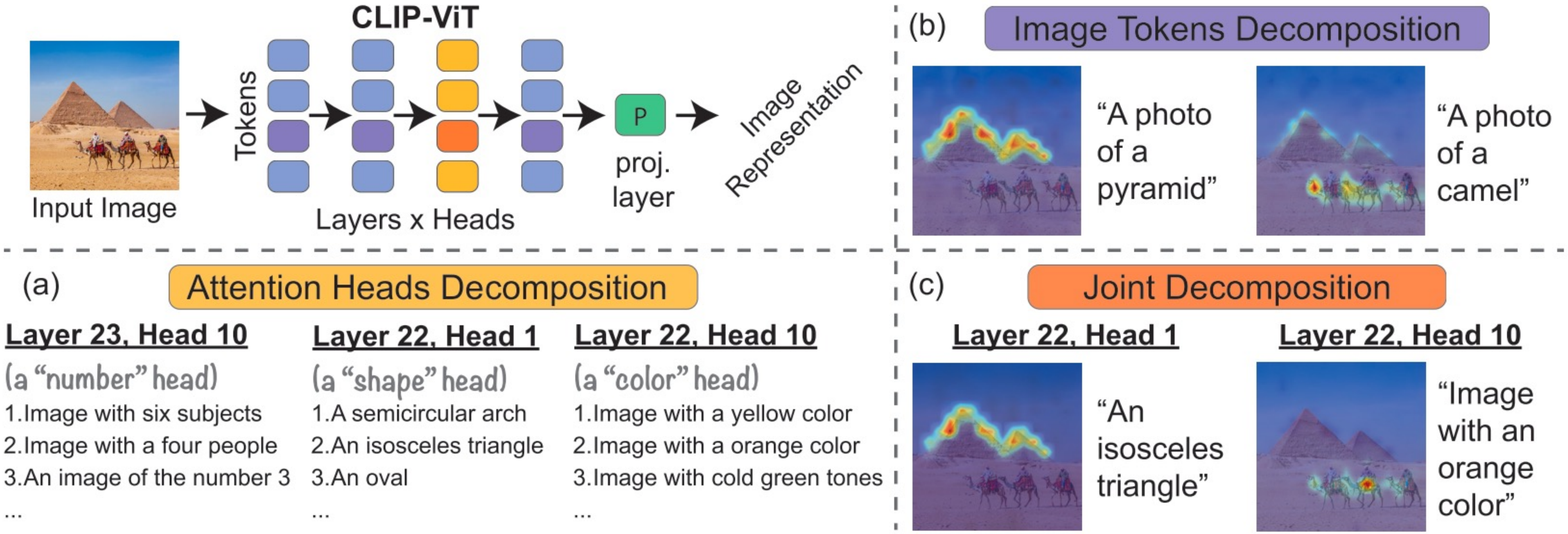
关键假设是，当Transformer中的潜在token不关注其他token时，它们在后续层中保持相同的语义信息。所提出的方法利用这个属性来解释潜在token，通过前向传播将潜在嵌入映射到最后一层，而无需自注意力操作。

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！

[22] Chen, Haozhe, et al. "Interpreting and Controlling Vision Foundation Models via Text Explanations." *arXiv preprint arXiv:2310.10591* (2023).

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。

[23] Gandelsman, Yossi, Alexei A. Efros, and Jacob Steinhardt. "Interpreting CLIP's Image Representation via Text-Based Decomposition." *arXiv preprint arXiv:2310.05916* (2023).

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。解释大模型带来了很多独特挑战：（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？

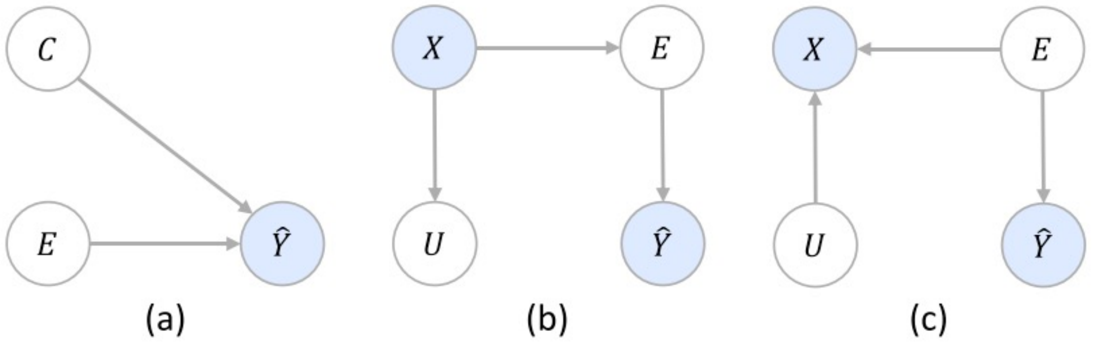
[24] Wu, Chenwang, et al. "A Causality Inspired Framework for Model Interpretation." *KDD*. 2023.

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



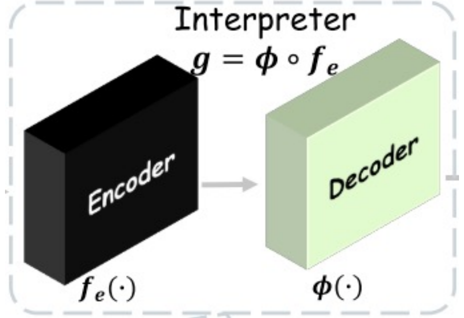
$E \perp U \mid X$　　$E \perp U$

(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。

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

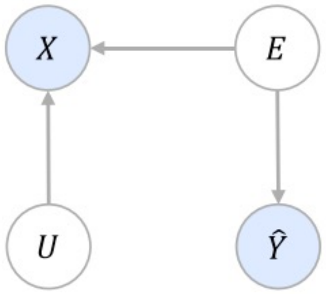
目标即为学习一个方程 $g: X \rightarrow M$。



[24] Wu, Chenwang, et al. "A Causality Inspired Framework for Model Interpretation." *KDD*. 2023.

多模态编码式基础模型的可解释性

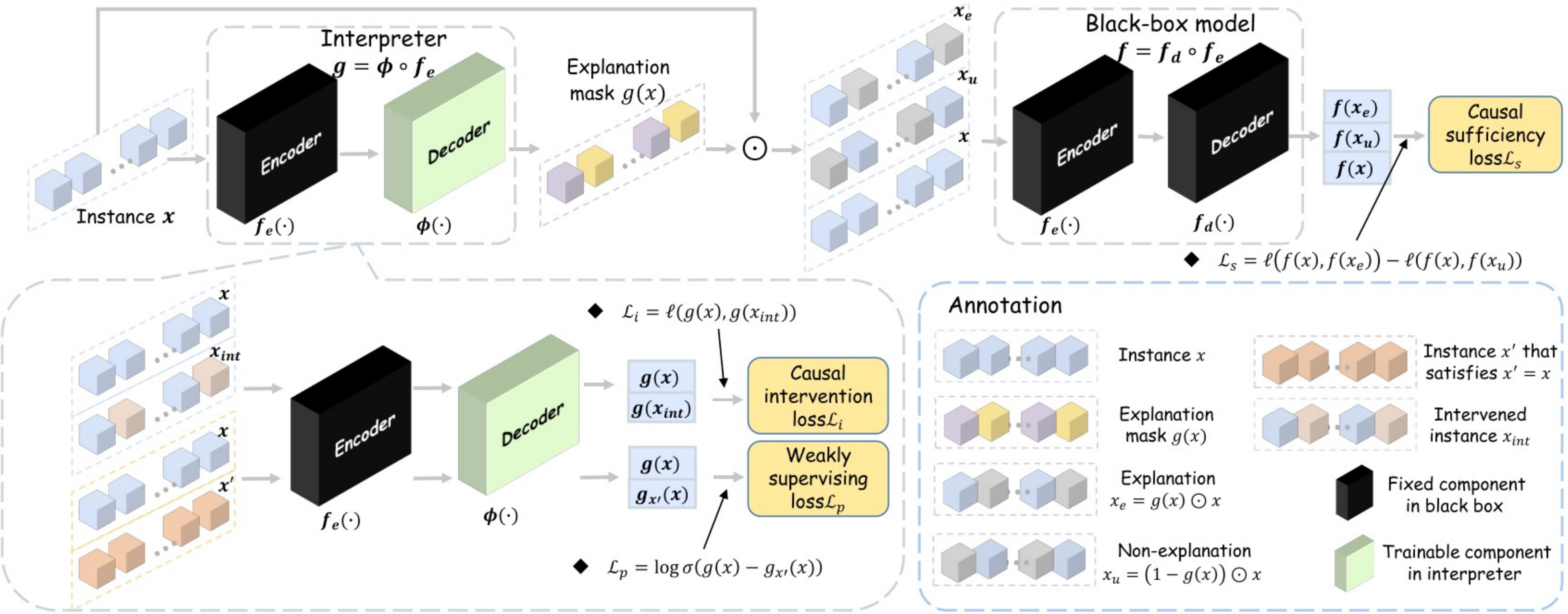
Causality Inspired Model Interpreter (CIMI)　中国科学技术大学、微软亚洲研究院



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$g: X \rightarrow M$。

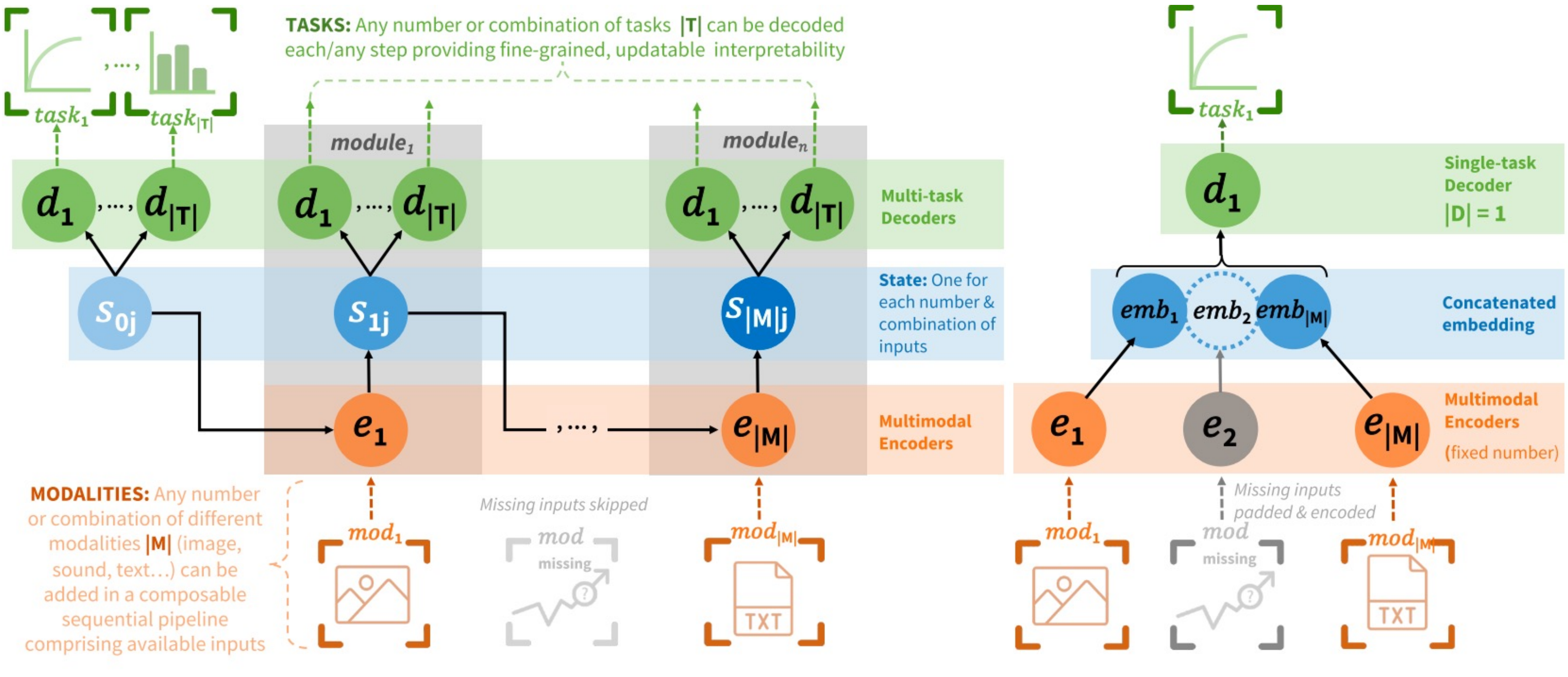


CIMI的框架。唯一可训练的组件是解码器ϕ，这是一个简单的神经网络，可以用相对较少的样本进行训练。

[24] Wu, Chenwang, et al. "A Causality Inspired Framework for Model Interpretation." *KDD*. 2023.

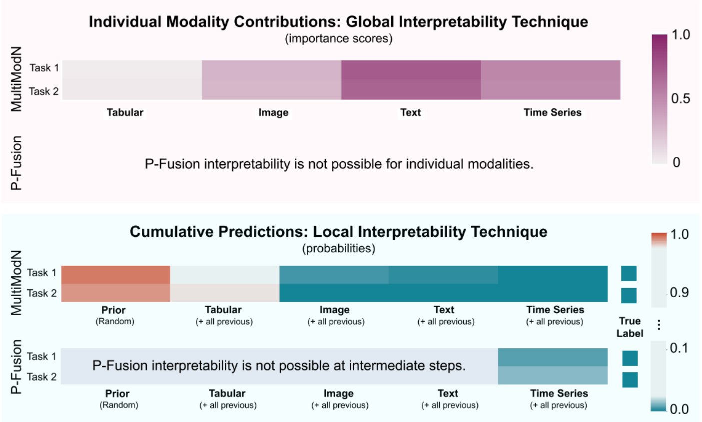
多模态编码式基础模型的可解释性

设计新式模型，但大参数模型上可能难训练



(a) MultiModN (b) P-Fusion

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。

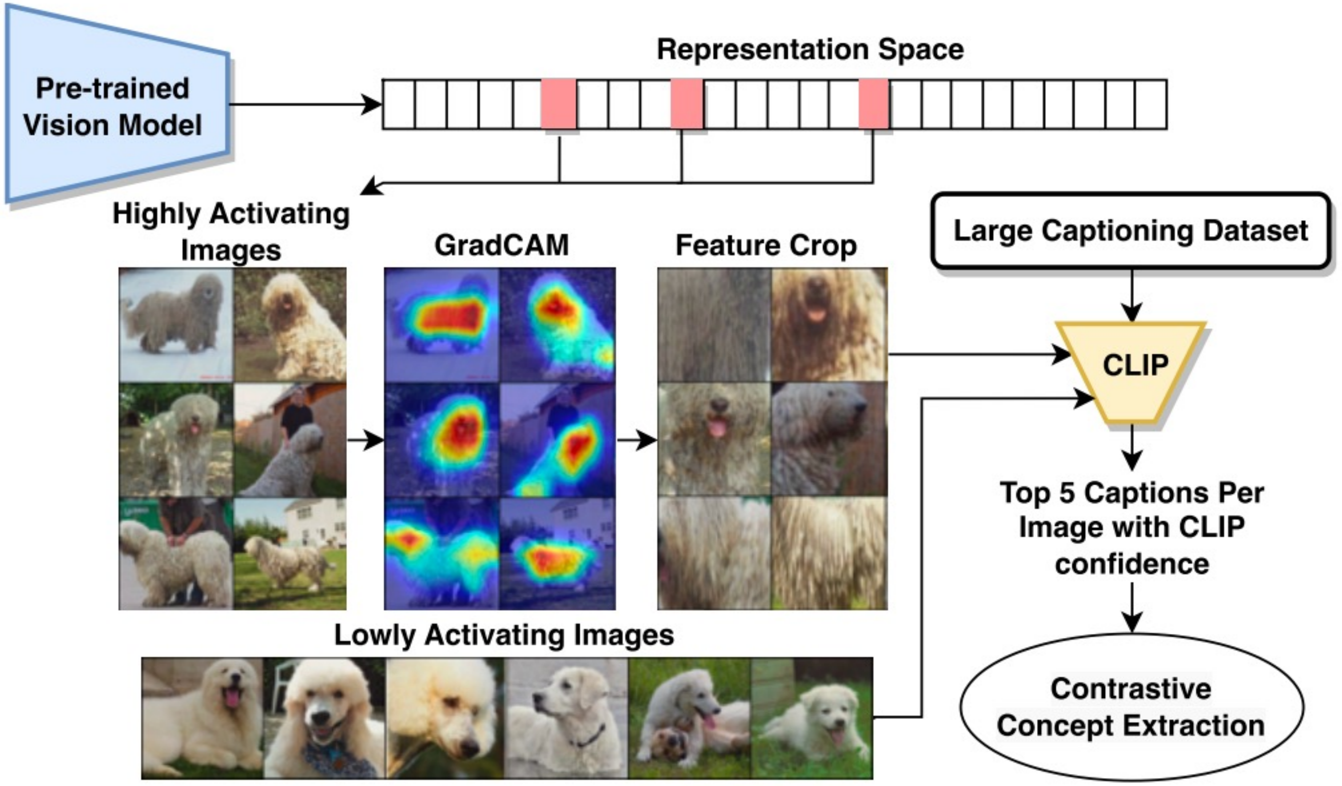


MultiModN中固有的特定于模式的模型可解释性。

[25] Swamy, Vinitra, et al. "MultiModN-Multimodal, Multi-Task, Interpretable Modular Networks." *arXiv preprint arXiv:2309.14118* (2023).

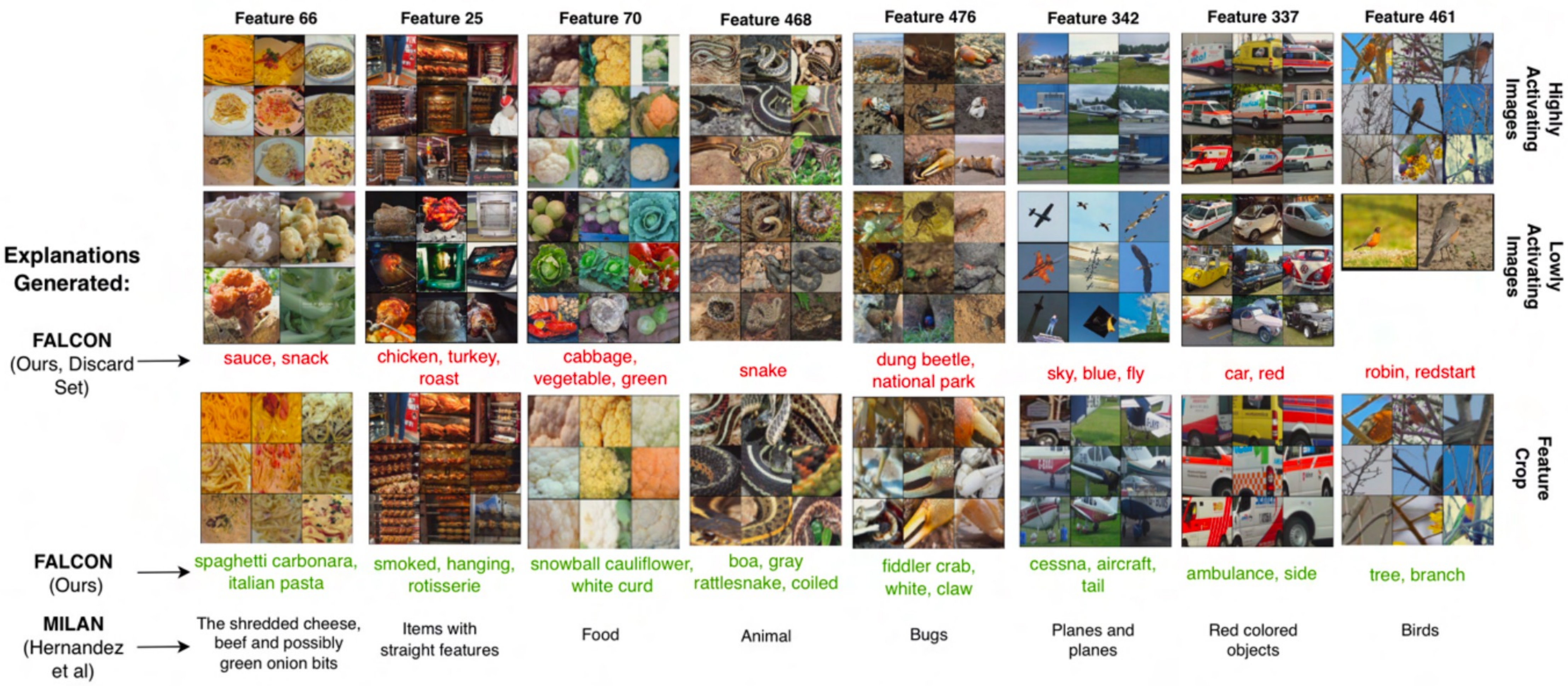
多模态编码式基础模型的可解释性

解释多模态特征的表征



[26] Kalibhat, Neha, et al. "Identifying Interpretable Subspaces in Image Representations." (2023).

多模态编码式基础模型的可解释性

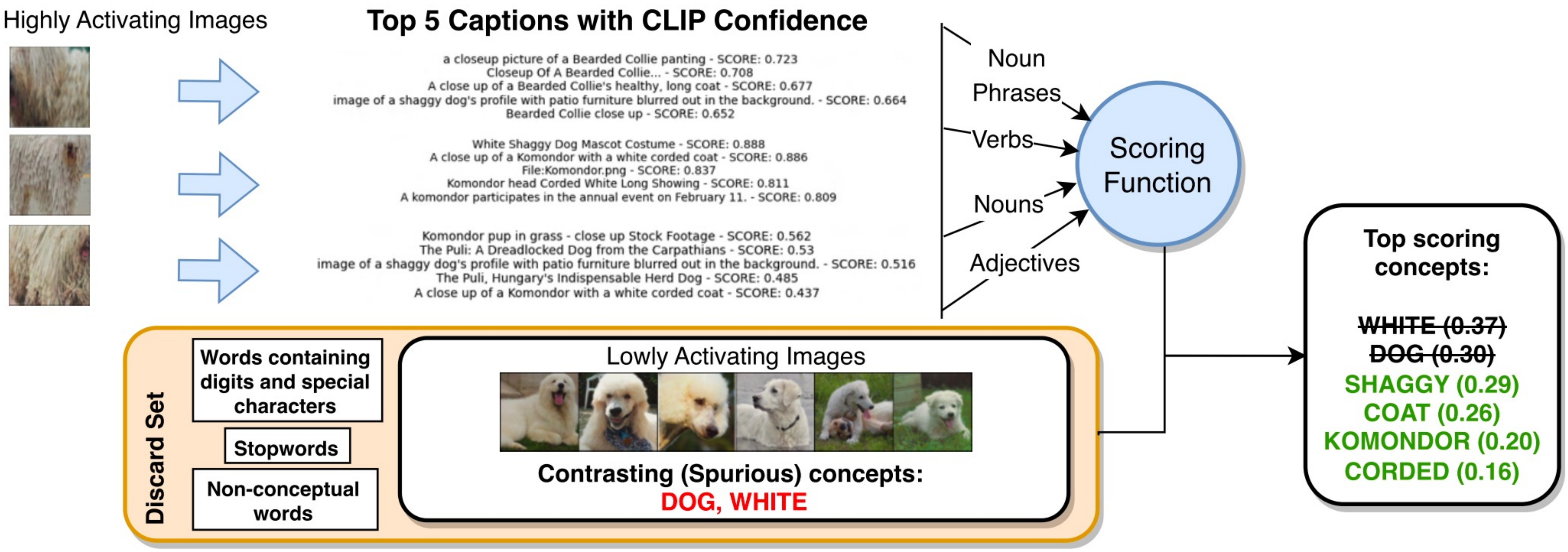


弱激活图像用于过滤虚假概念

强激活图像用于生成概念

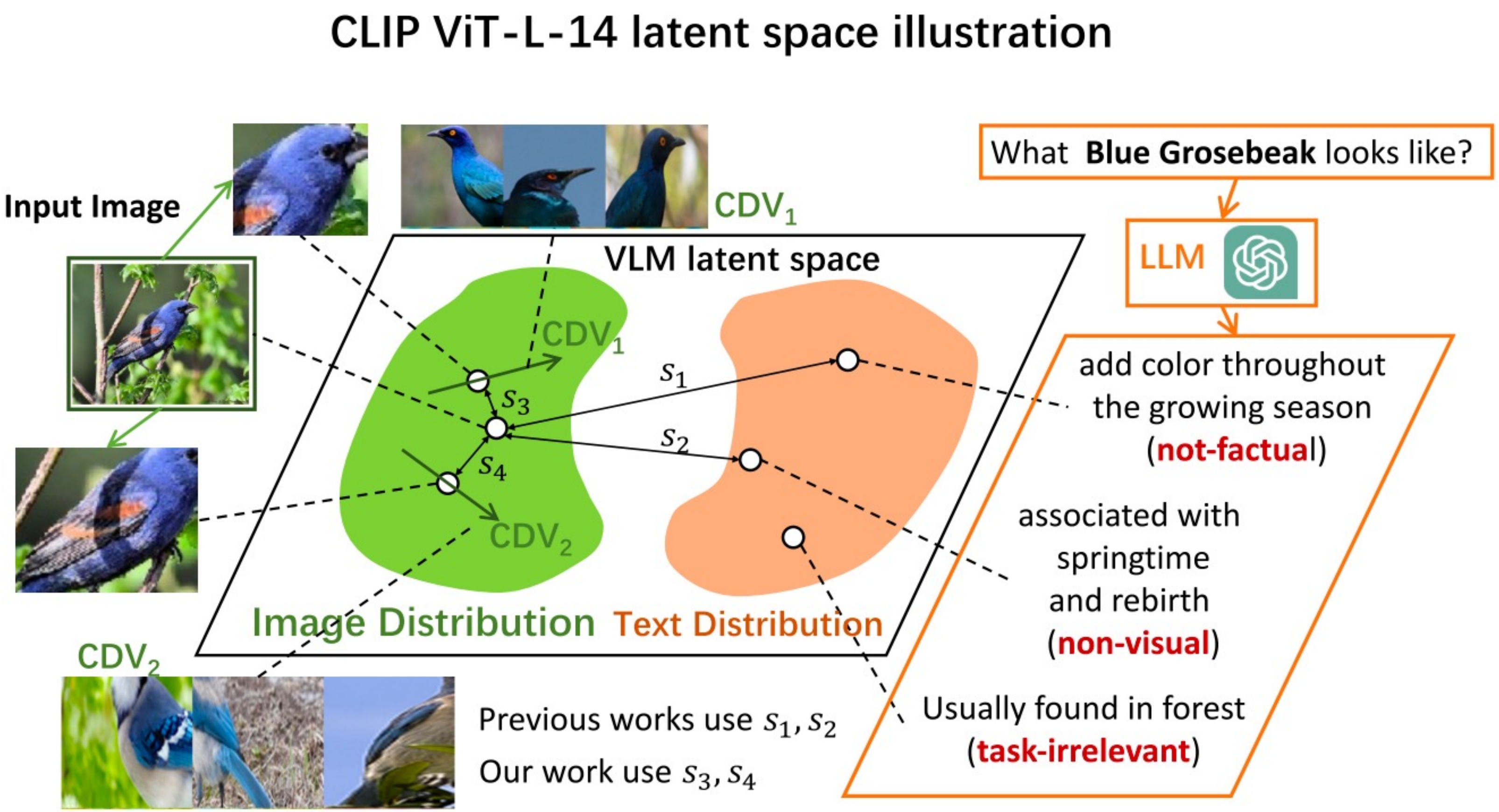
[26] Kalibhat, Neha, et al. "Identifying Interpretable Subspaces in Image Representations." (2023).

多模态编码式基础模型的可解释性



[26] Kalibhat, Neha, et al. "Identifying Interpretable Subspaces in Image Representations." (2023).

多模态编码式基础模型的可解释性



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。

[27] Pratt, Sarah, et al. "Cross-modality Interpretable image classification via Concept Decomposition Vector of Visual Language Models." (2023).

多模态编码式基础模型的可解释性

总结

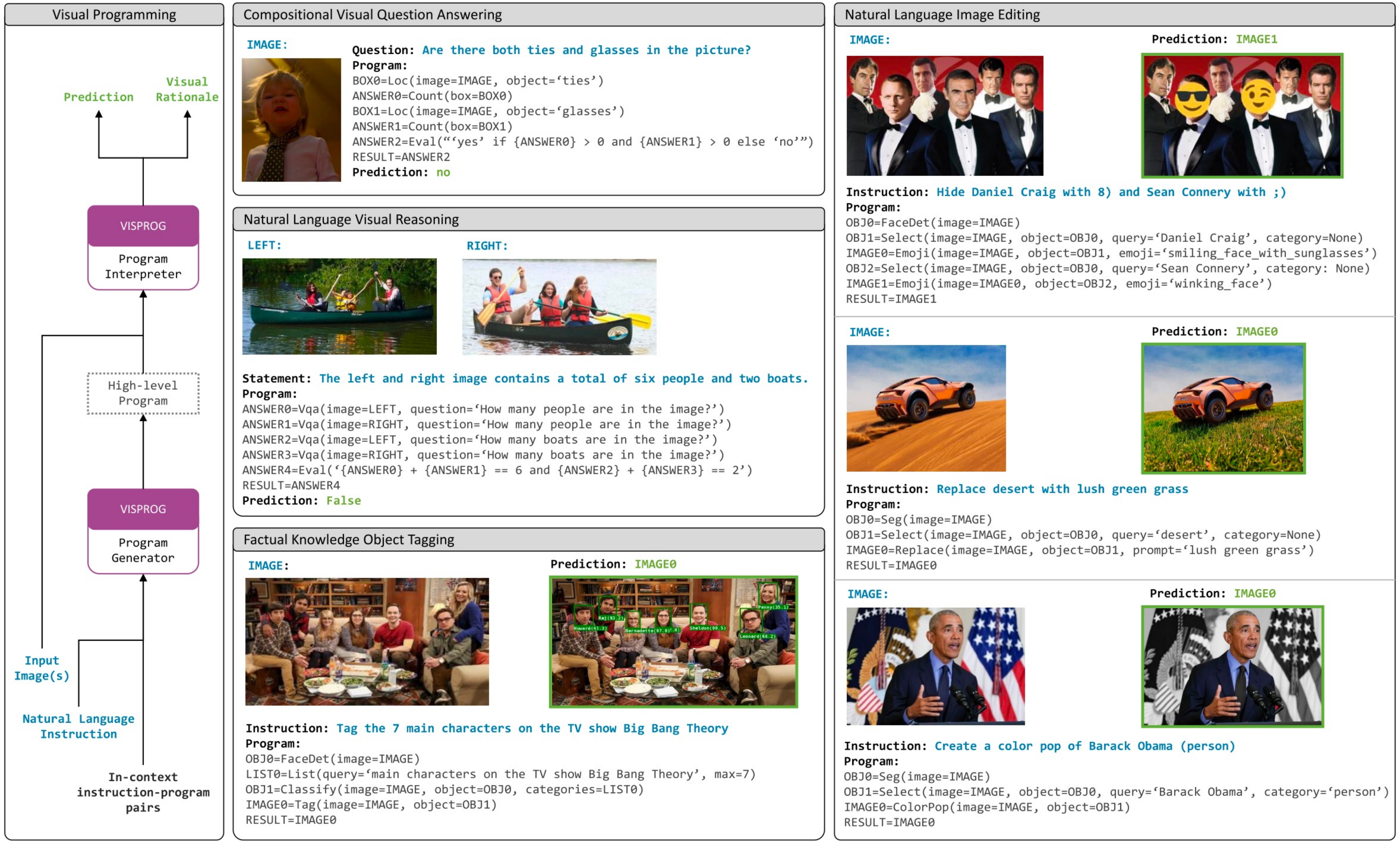
- 如何利用多模态编码基础模型的特性，利用人类容易理解的文本描述来辅助解释？
- 如何理解多模态基础模型内部的运行机理？一些假设是否是正确的？或者是不是应该这样被理解？
- 如何构建统一的因果图模型，以应对大模型中巨大的参数量与消耗的参数推理的挑战？
- 如何分解特征，以帮助人类的理解？
- 如何设计一个更方便可解释的模型结果，同时适应在大模型训练数据量巨大的情况。

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)

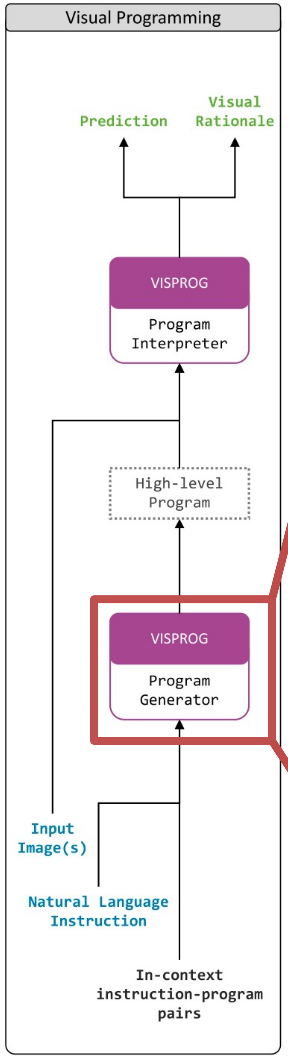


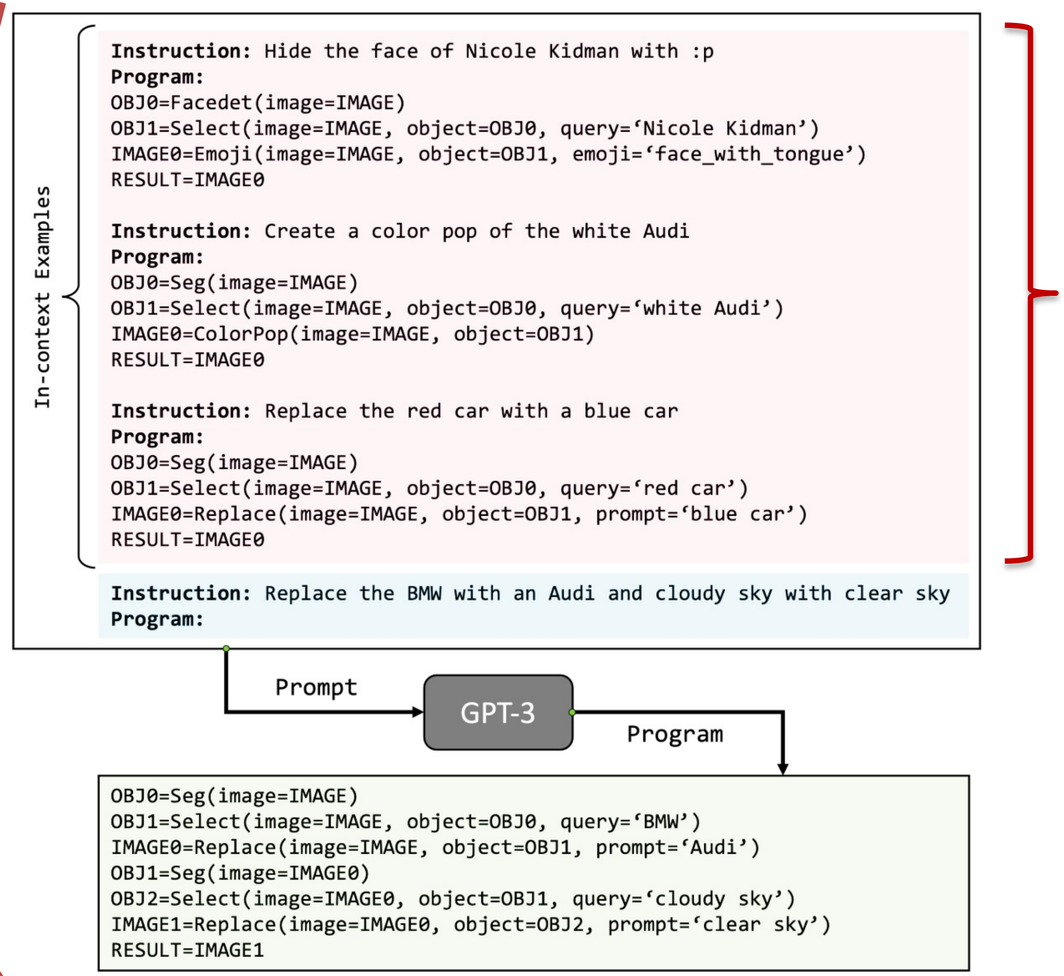
VisProg是一个模块化和可解释的神经符号系统，用于组合视觉推理。

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

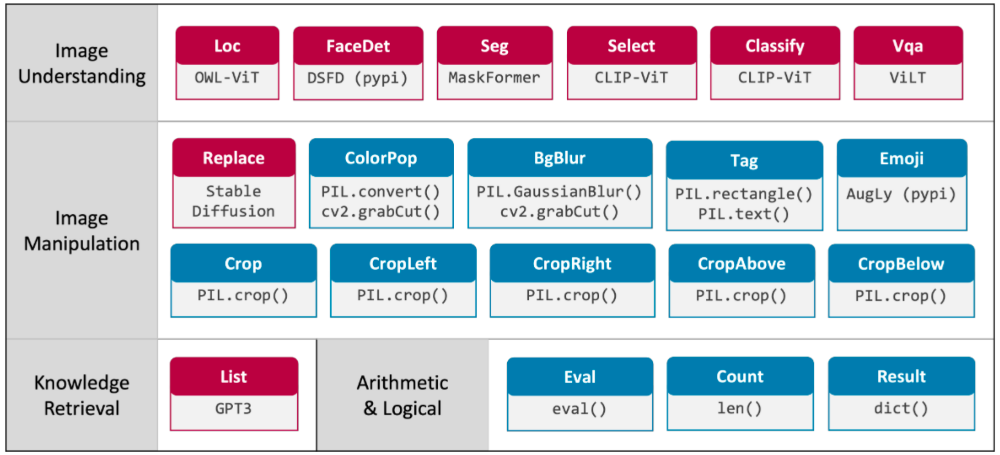
多模态问答式基础模型的可解释性

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





VisProg的程序生成过程。

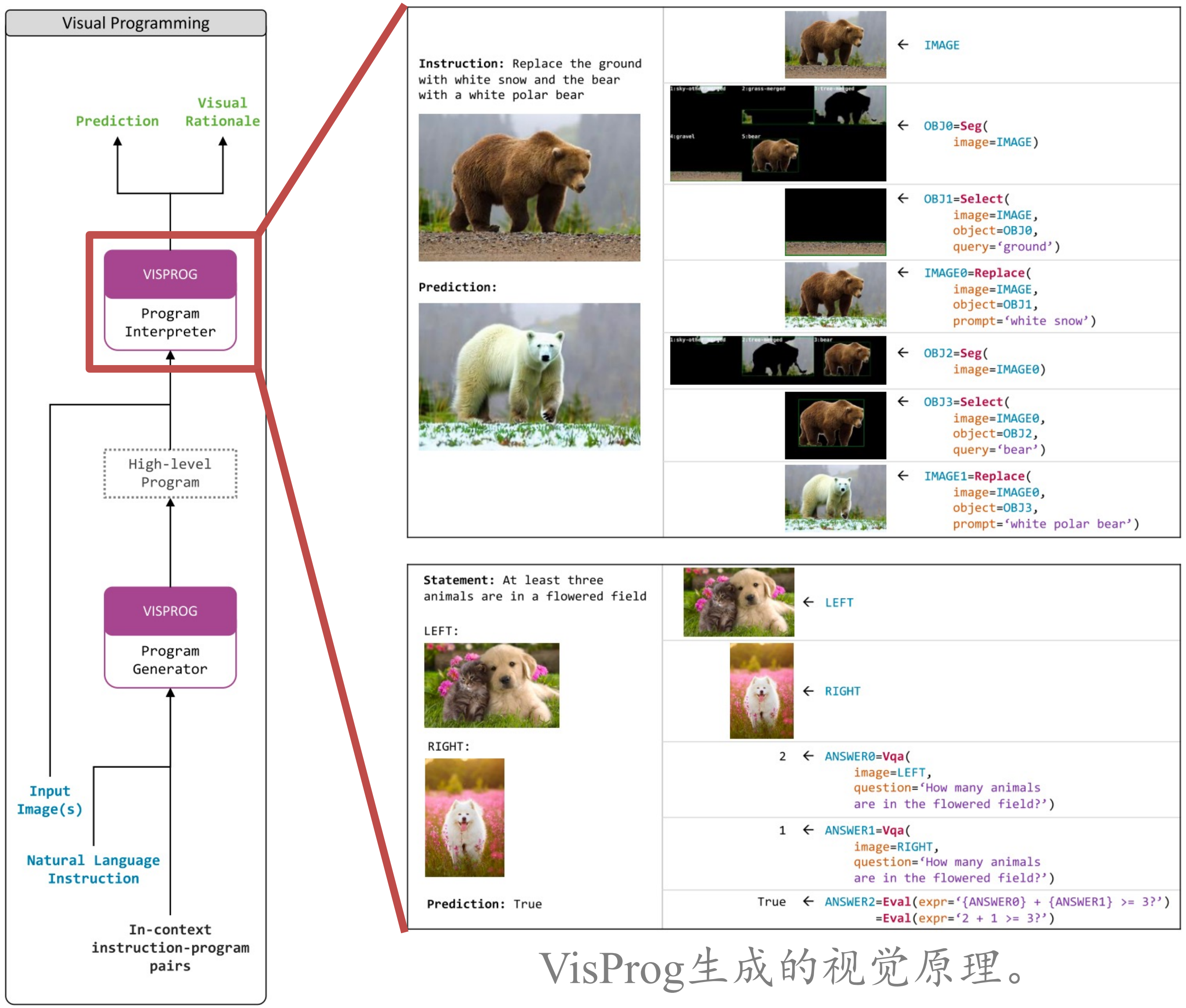


VisProg已有的所支持的功能模块。

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

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。

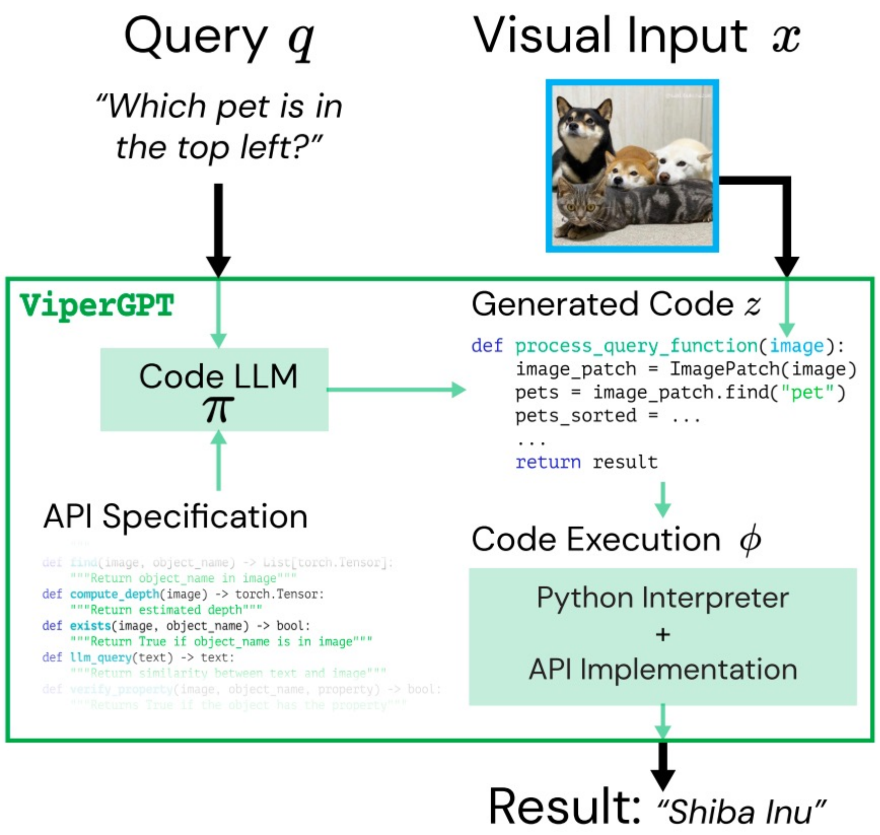
Task	Input	Output	Modules
Compositional Visual QA (GQA)	Image + Question	Text	Loc, Vqa, Eval, Count, Crop, CropLeft, CropRight, CropAbove, CropBelow
Reasoning on Image Pairs (NLVR)	Image Pair + Statement	True/False	Vqa, Eval
Factual Knowledge Object Tagging	Image + Instruction	Image	FaceDet, List, Classify,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

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。

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

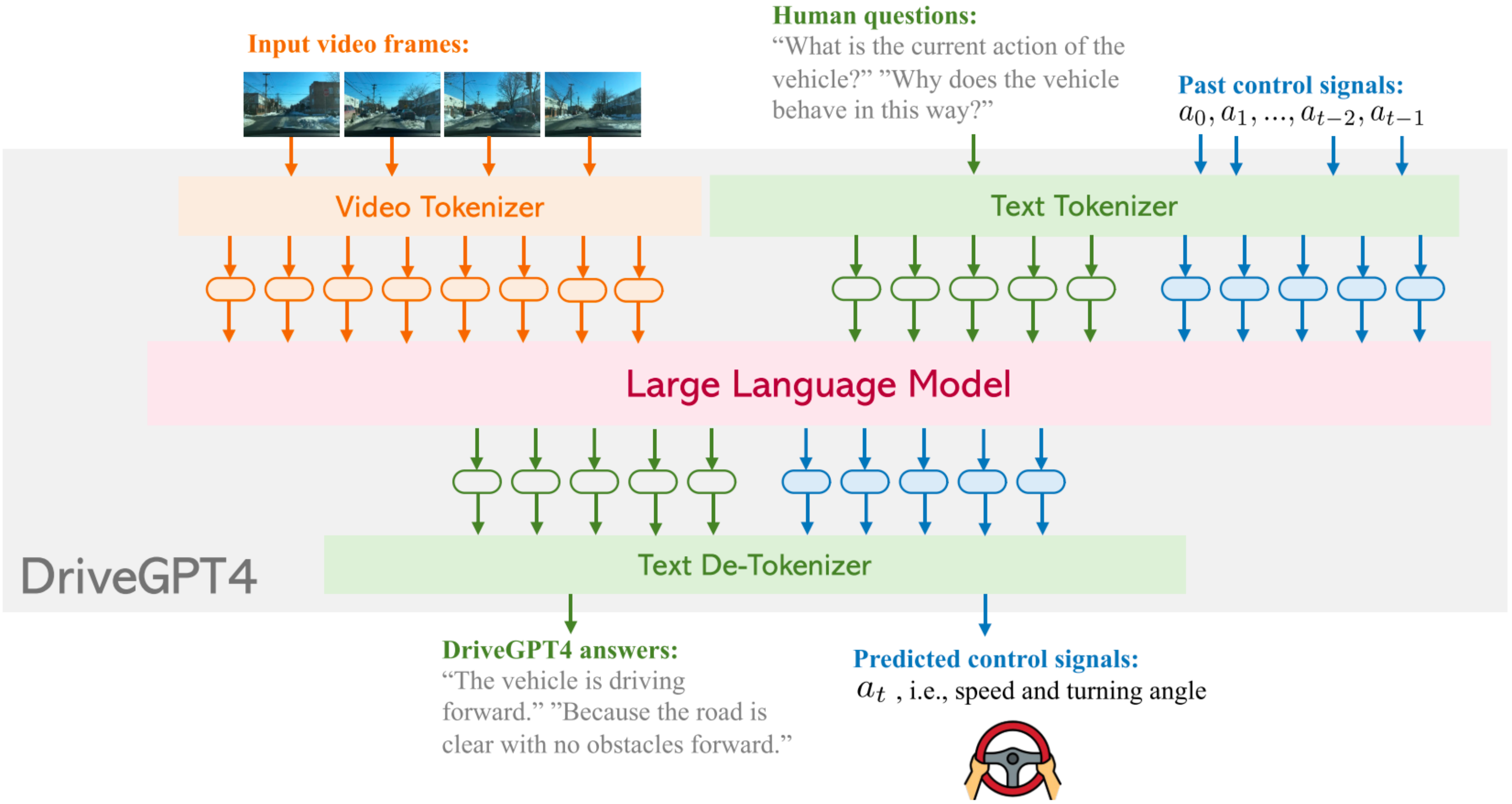
多模态问答式基础模型的可解释性

ViperGPT　　美国 Columbia大学

[29] 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

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



[30] Xu, Zhenhua, et al. "DriveGPT4: Interpretable End-to-end Autonomous Driving via Large Language Model." *arXiv preprint arXiv:2310.01412* (2023).

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

[30] Xu, Zhenhua, et al. "DriveGPT4: Interpretable End-to-end Autonomous Driving via Large Language Model." *arXiv preprint arXiv:2310.01412* (2023).

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