
# 模式识别
# Pattern Recognition

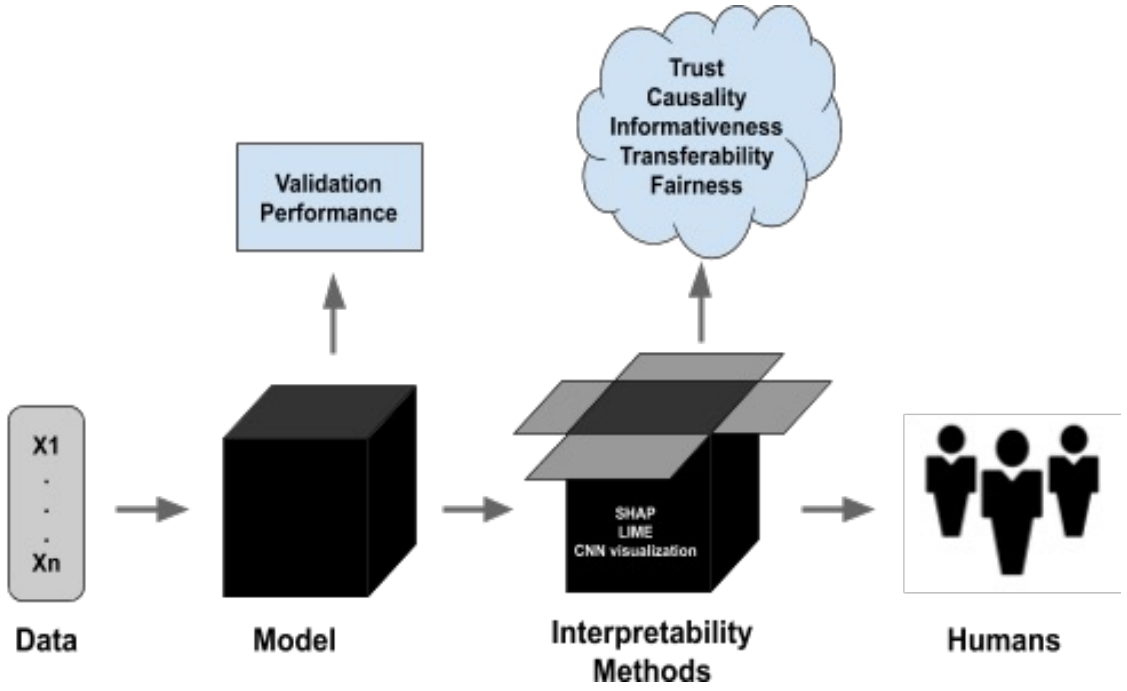
李泽榉，复旦大学 生物医学工程与技术创新学院

# 目录

# 模型可解释性

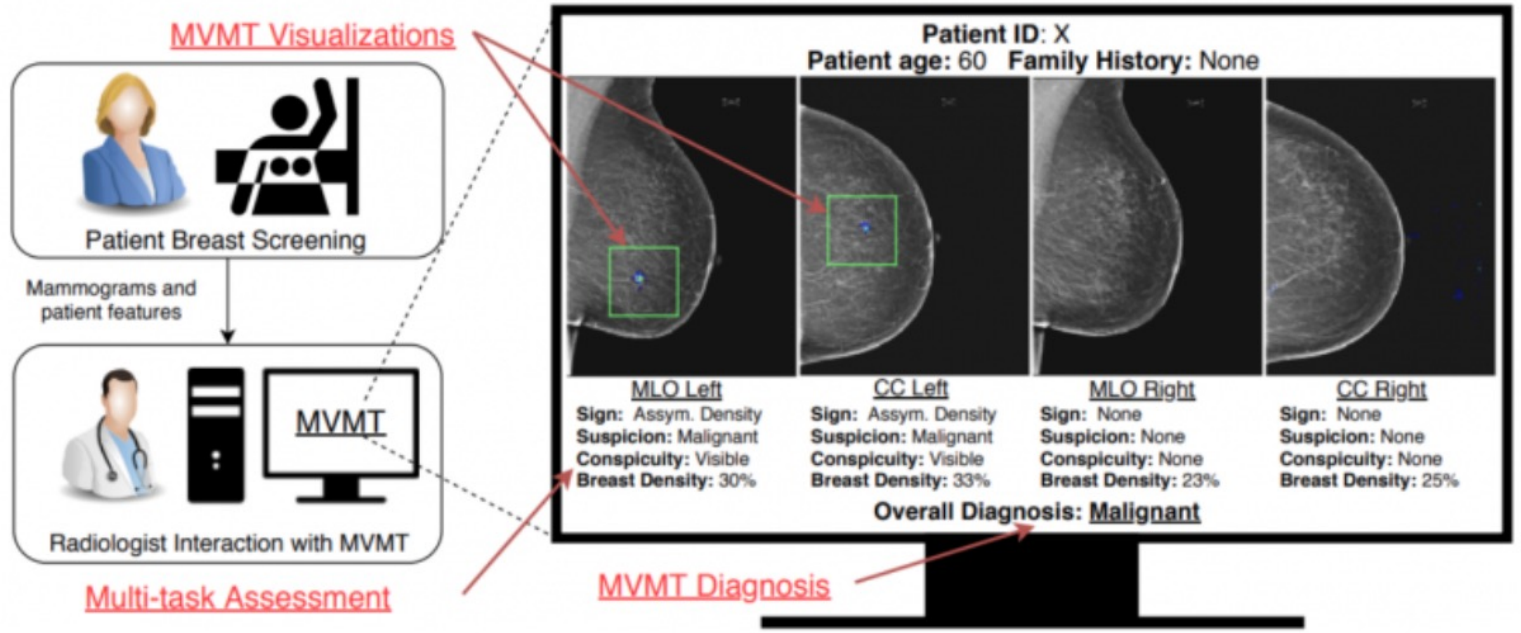
- 增强模型决策的透明度与可解释性，使其**与人类认知和价值观保持一致**

- 可解释性使我们能够理解模型究竟在学习什么、模型还能提供哪些额外信息，以及其决策背后的依据。



# 模型可解释性

- 举例：机器学习系统输出额外**图像描述特征**，供放射科医生结合最终诊断结果进行核查。

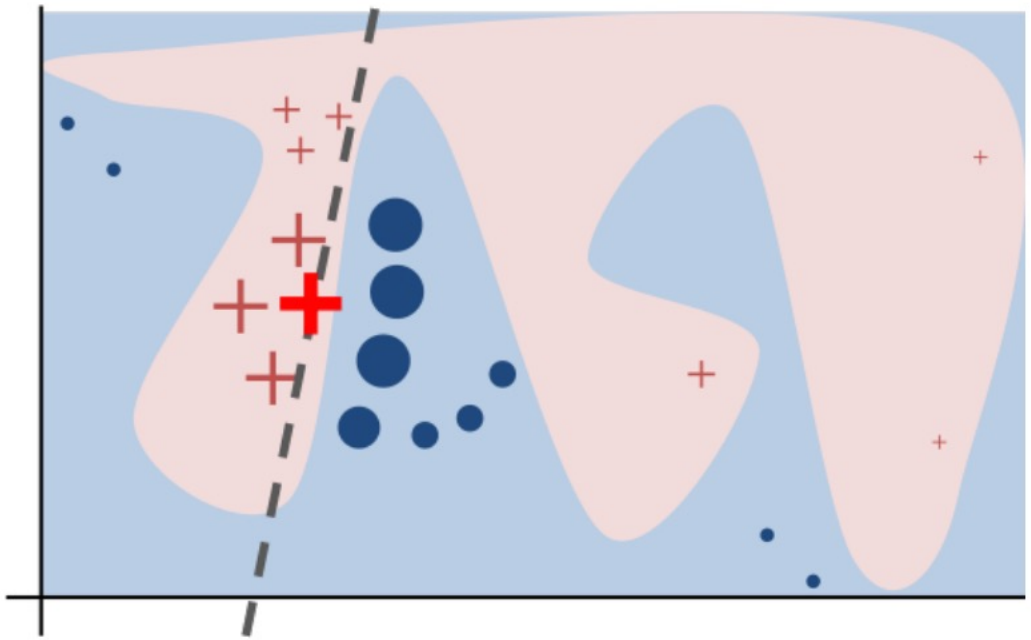



Multi-view Multi-task Learning for Improving Autonomous Mammogram Diagnosis

# 模型可解释性

- **LIME** (Local Interpretable Model-Agnostic Explanations)

- 在预测的局部范围内学习一个可解释的模型



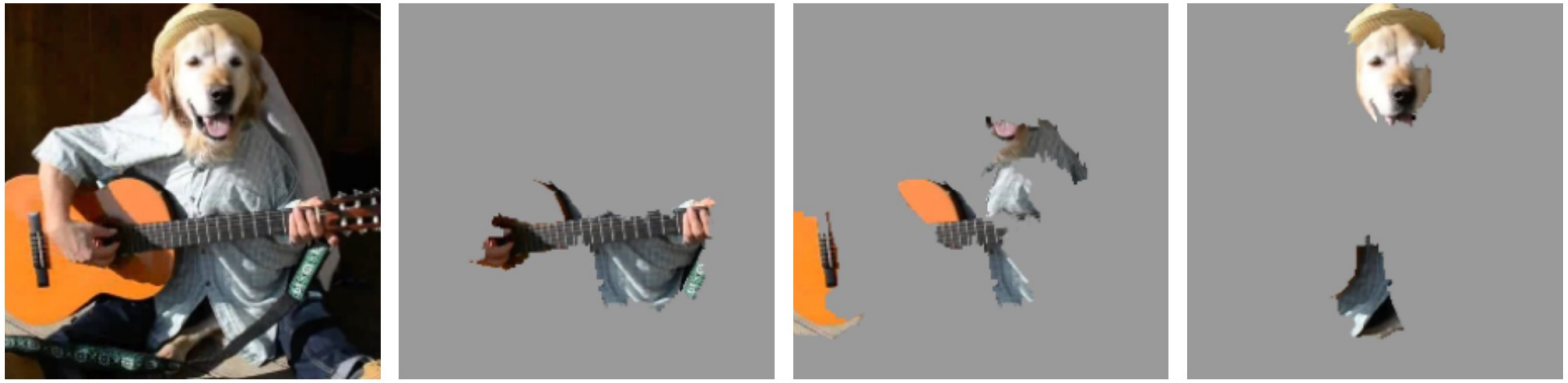
"Why Should I Trust You?" Explaining the Predictions of Any Classifier

# 模型可解释性

- 在x附近采样**随机扰动**
- 图片分类中用**超像素**来采样扰动

**Algorithm 1** Sparse Linear Explanations using LIME

**Require:** Classifier $f$, Number of samples $N$
**Require:** Instance $x$, and its interpretable version $x'$
**Require:** Similarity kernel $\pi_x$, Length of explanation $K$
  $\mathcal{Z} \leftarrow \{\}$
  **for** $i \in \{1, 2, 3, ..., N\}$ **do**
    $z'_i \leftarrow sample\_around(x')$
    $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
  **end for**
  $w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$ ▷ with $z'_i$ as features, $f(z)$ as target
  **return** $w$



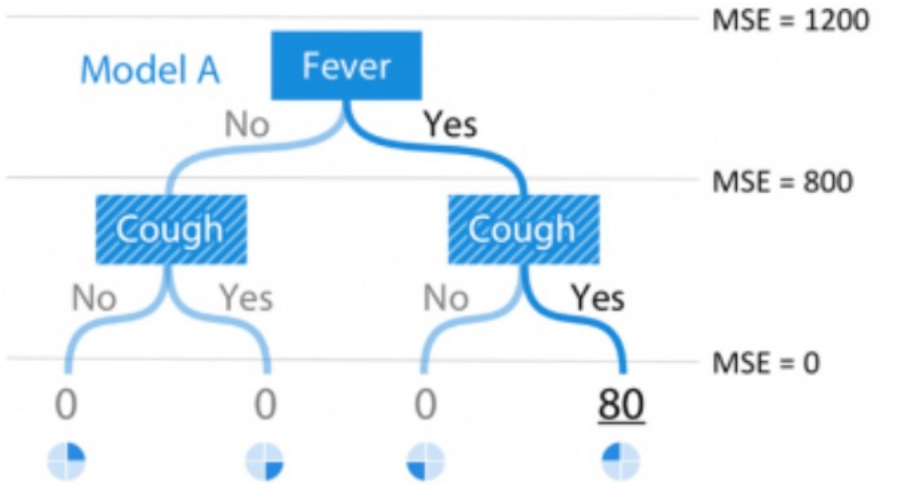
(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

**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" ($p = 0.21$)**

# 模型可解释性

- **LIME的问题**：**不稳定**，多次运行LIME，可能会生成差异显著的解释

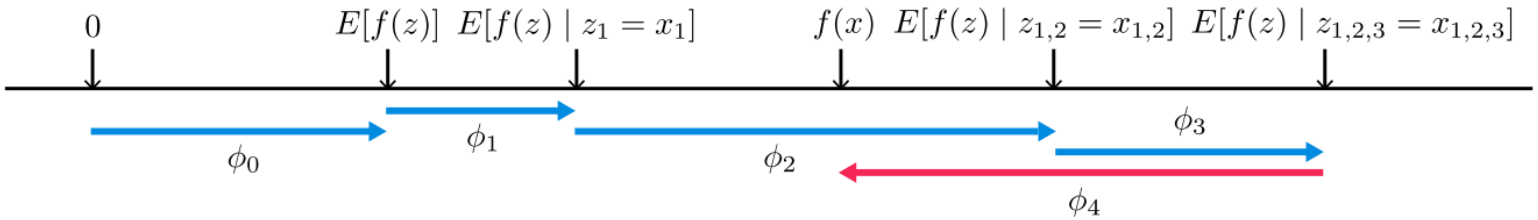
- 这是因为LIME在生成局部扰动样本时存在随机性（如样本采样、特征扰动），难以解释模型决策**依赖特征**间的复杂组合




- 能知道Fever和Cough对risk判断重要
- 但不能分析出Fever和Cough的同时出现最重要

# 模型可解释性

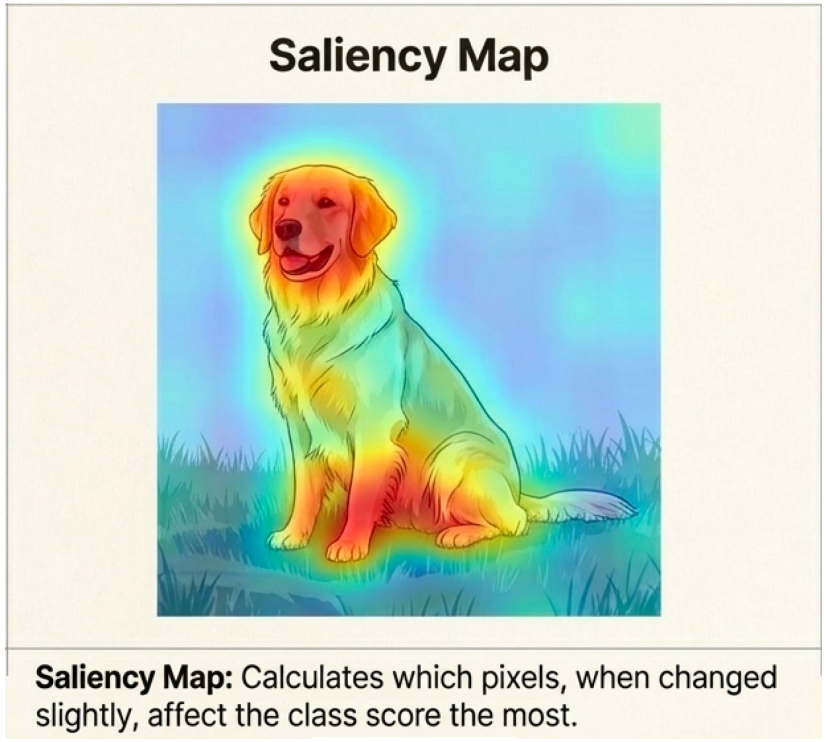
- **SHAP** (SHapley Additive exPlanations)
- 基于博弈论通过计算每个特征对模型预测的贡献值，可以证明他是**唯一满足**Local accuracy、Missingness和Consistency性质的（LIME不行）
- 本质想法是通过**遍历**所有可能的特征组合，精确计算每个特征在所有组合中的平均边际贡献



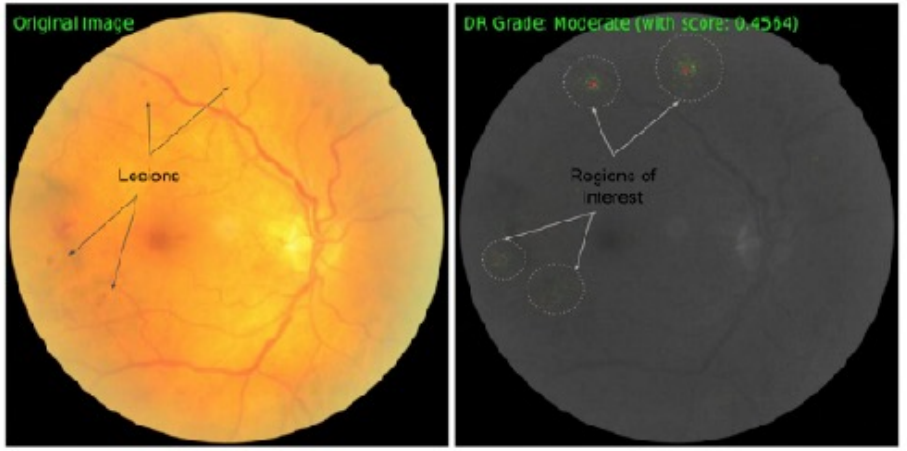

A Unified Approach to Interpreting Model Predictions

# 模型可解释性

- **显著性图**：直接对输入图片求导



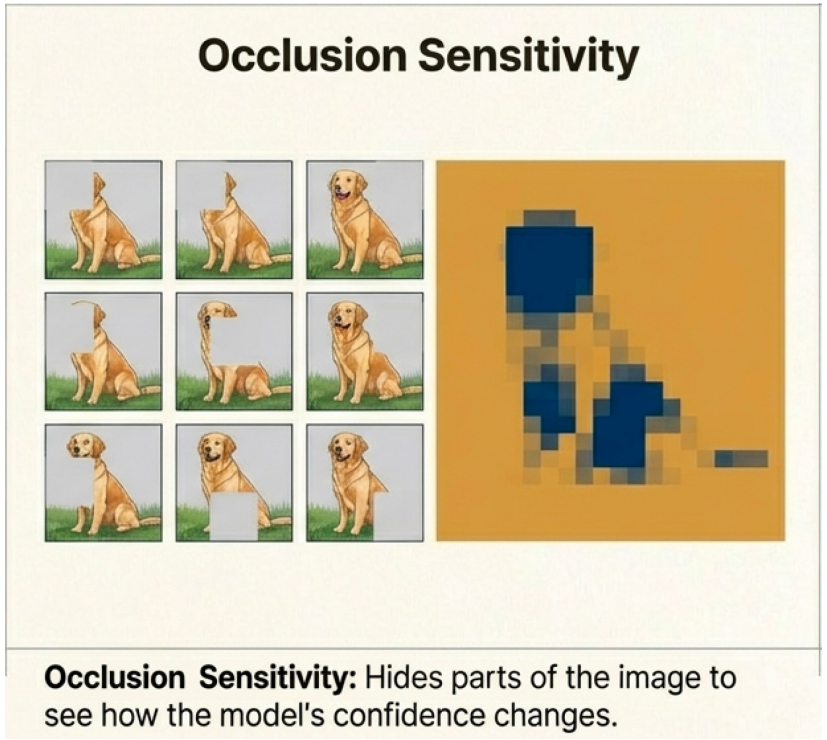

$$\frac{\partial s_{dog}(x)}{\partial x}$$

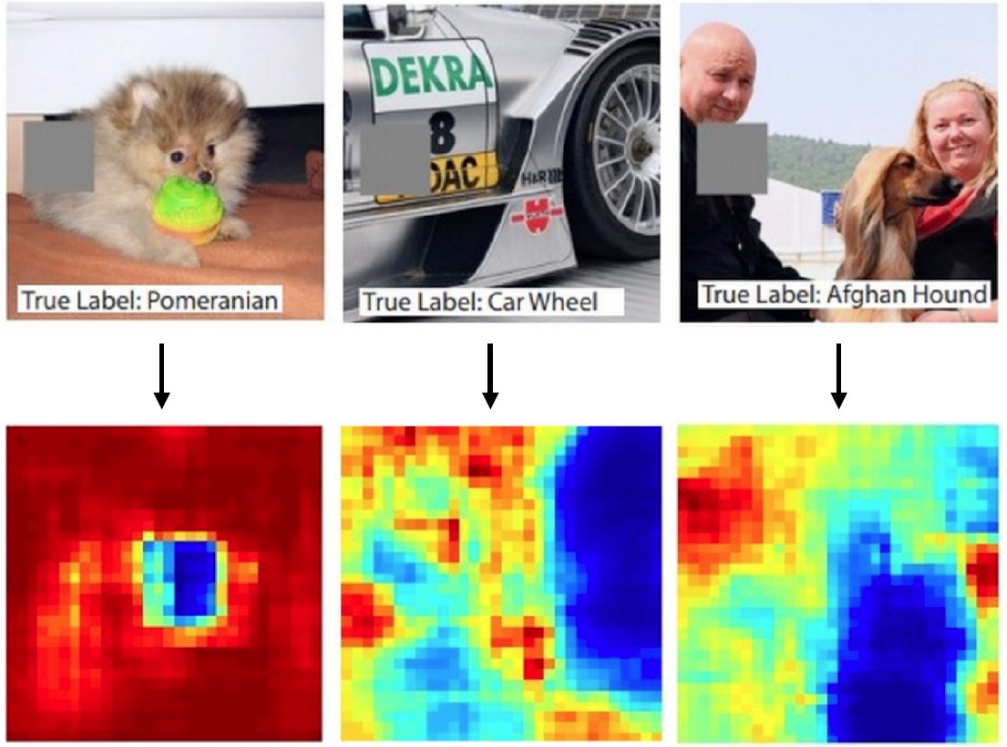



Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps

# 模型可解释性

- **遮盖敏感度**：每次遮盖一点

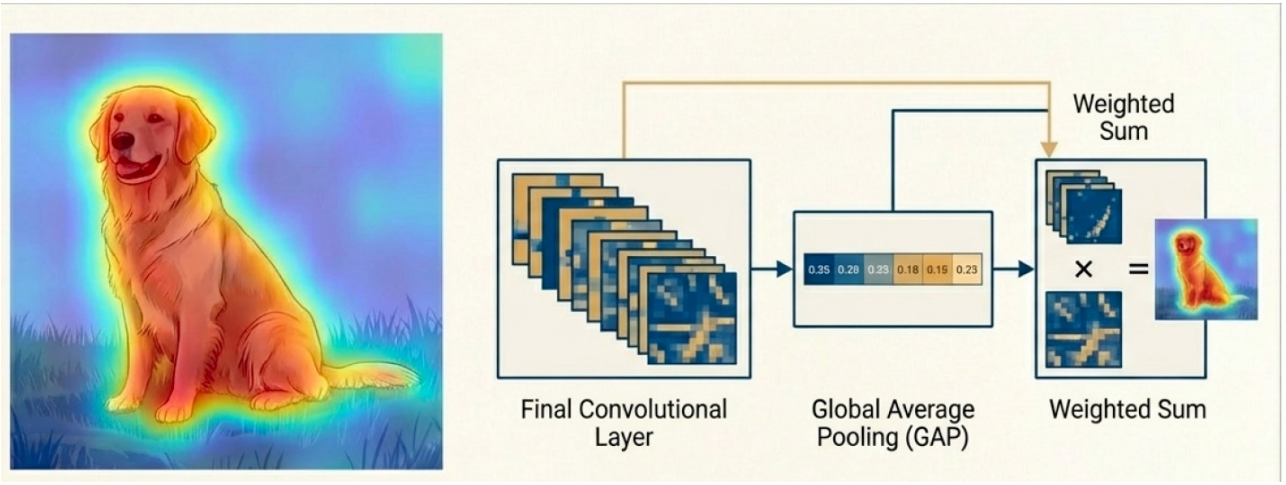



Visualizing and Understanding Convolutional Networks

# 模型可解释性

- **CAM** (Class Activation Mapping)
- 利用卷积特征图的加权组合来定位重要区域
- 更高效、更连续、更常用

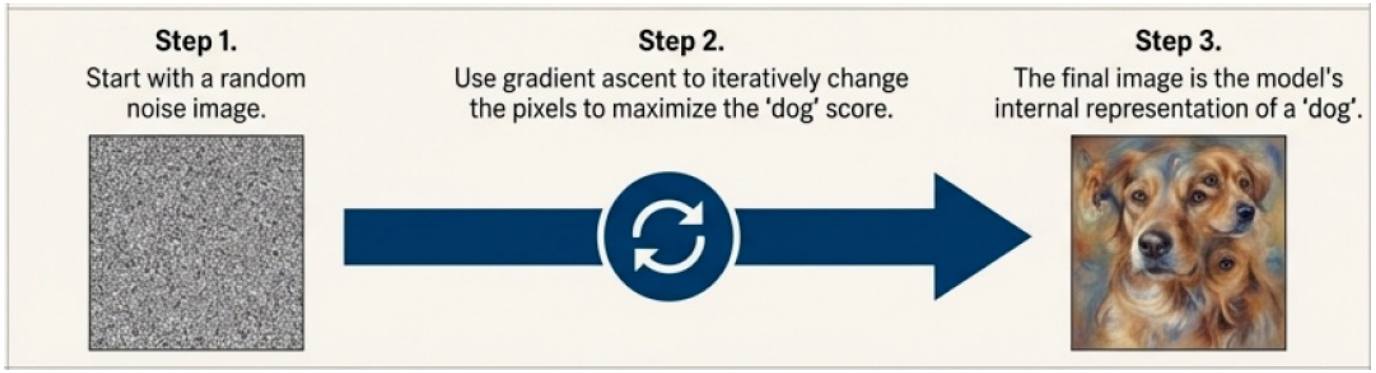


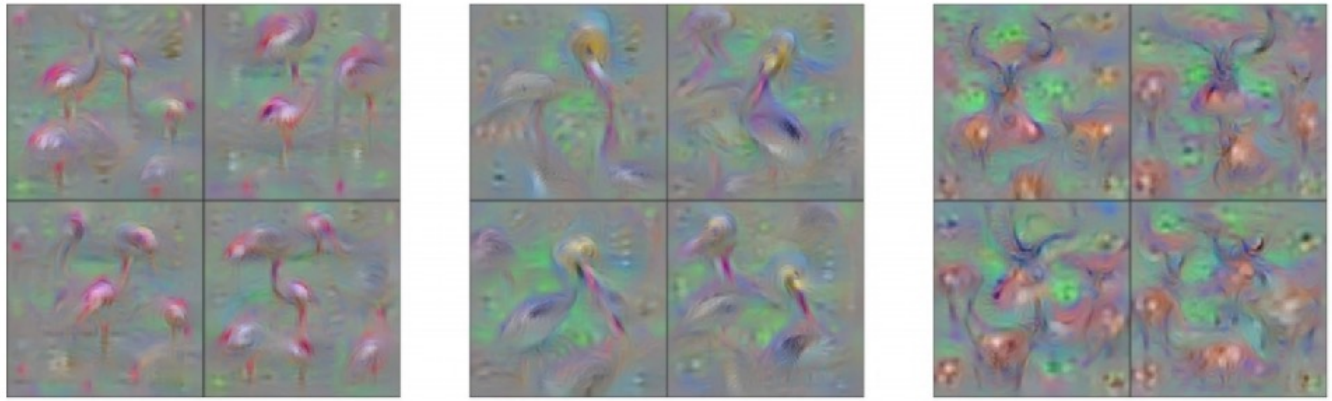


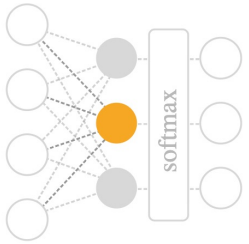
Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps

# 模型可解释性

- **概念可视化**：对于**每一种类别**求导（注意是softmax之前），优化得到似然度最大的输入图片




Class Logits

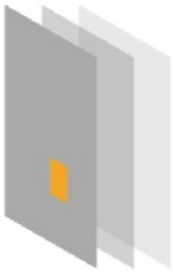
`pre_softmax[k]`

Understanding Neural Networks Through Deep Visualization

# 模型可解释性

- **概念可视化**：可以对某**一个神经元**求导

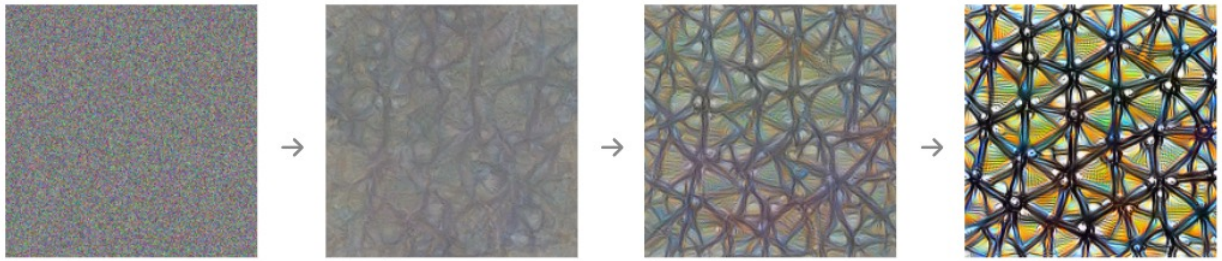
- 对于神经网络的每一个神经元（第n层、第z个通道、x，y位置），可以通过优化来找到激活模式



Neuron

$layer_n[x,y,z]$

Starting from random noise, we optimize an image to activate a particular neuron (layer mixed4a, unit 11).

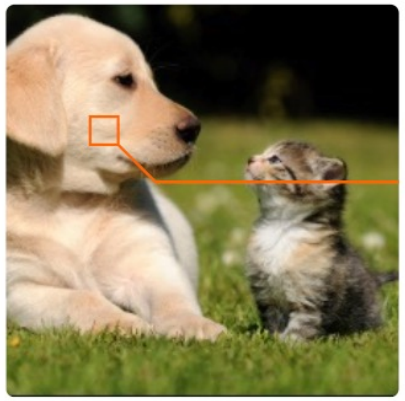


Step 0 → Step 4 → Step 48 → Step 2048

The Building Blocks of Interpretability

# 模型可解释性

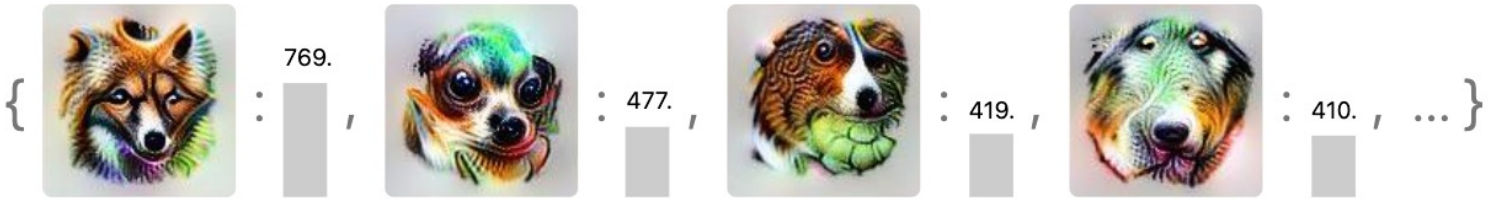
- **概念可视化**：第n层、第z个通道、x，y位置的解释



Making sense of these activations is hard because we usually work with them as abstract vectors:

$a_{4,3}$ = [0, 0, 0, 0, 89.2, 0, 0, 0, 17.7, 0, 0, 0, ...]

With feature visualization, however, we can transform this abstract vector into a more meaningful "semantic dictionary".




There seem to be detectors for floppy ears, dog snouts, cat heads, furry legs, and grass.

The Building Blocks of Interpretability

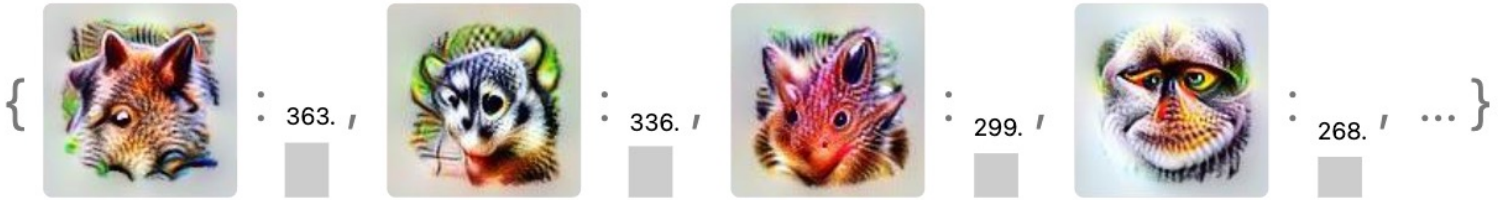
# 模型可解释性

- **概念可视化**：第n层、第z个通道、x′ ，y′ 位置的解释

Making sense of these activations is hard because we usually work with them as abstract vectors:

$a_{6,10}$ = [0, 0, 8.12, 0, 82.0, 0, 0, 20.4, 0, 0, 0, 0, ...]

With feature visualization, however, we can transform this abstract vector into a more meaningful "semantic dictionary".




There seem to be detectors for floppy ears, dog snouts, **cat heads**, furry legs, and grass.

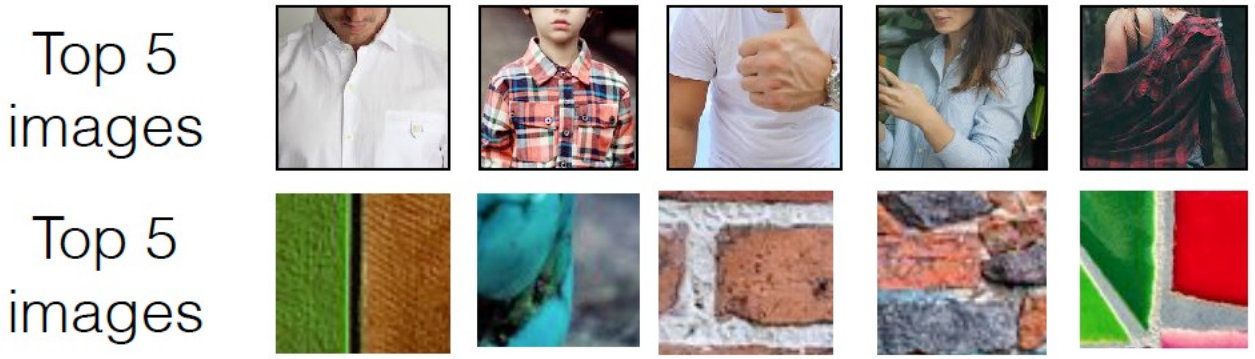
The Building Blocks of Interpretability

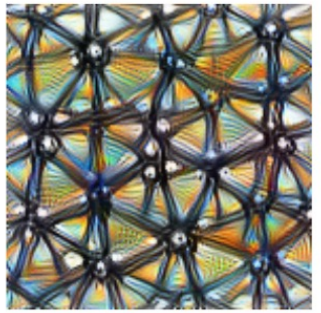
# 模型可解释性

- **概念可视化**：可以对某**一个通道**求导

- 对于神经网络的每一个通道（第n层第z个通道），可以通过优化来找到激活模式

- 匹配训练数据集中的最大激活的图片



Channel

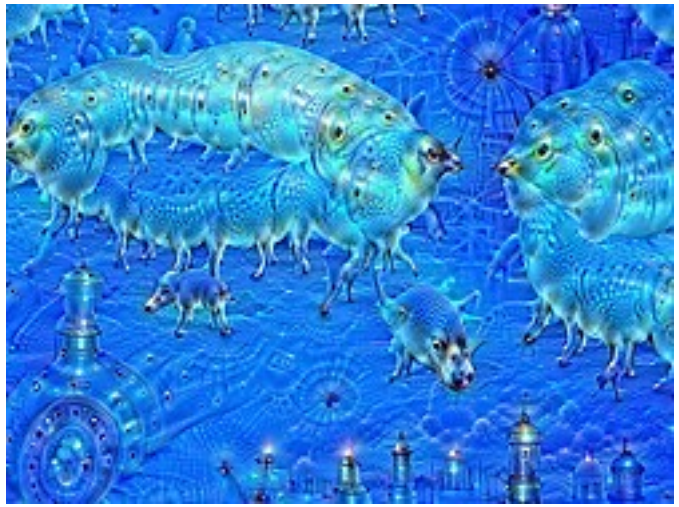
$layer_n[:,:,z]$

# 模型可解释性

- **概念可视化**：可以对某**一层神经网络**求导

- 对于神经网络的每一层（第n层）所有神经元，可以通过优化来找到激活模式

- 艺术成分比较高，找到网络最感兴趣的图片



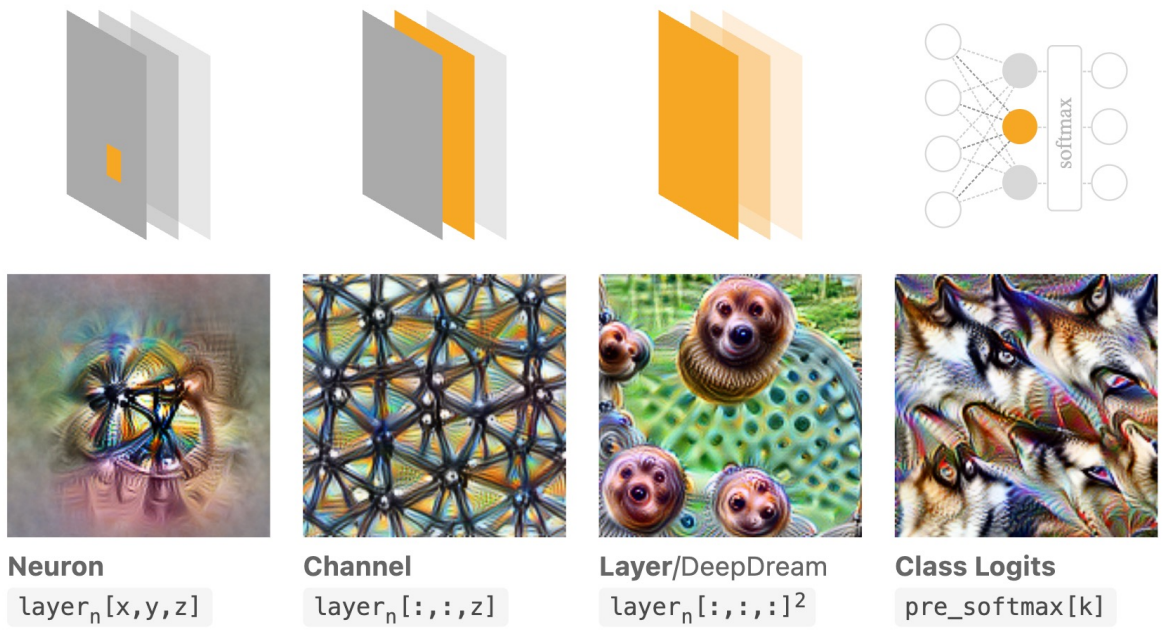
**Layer**/DeepDream

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

# 模型可解释性

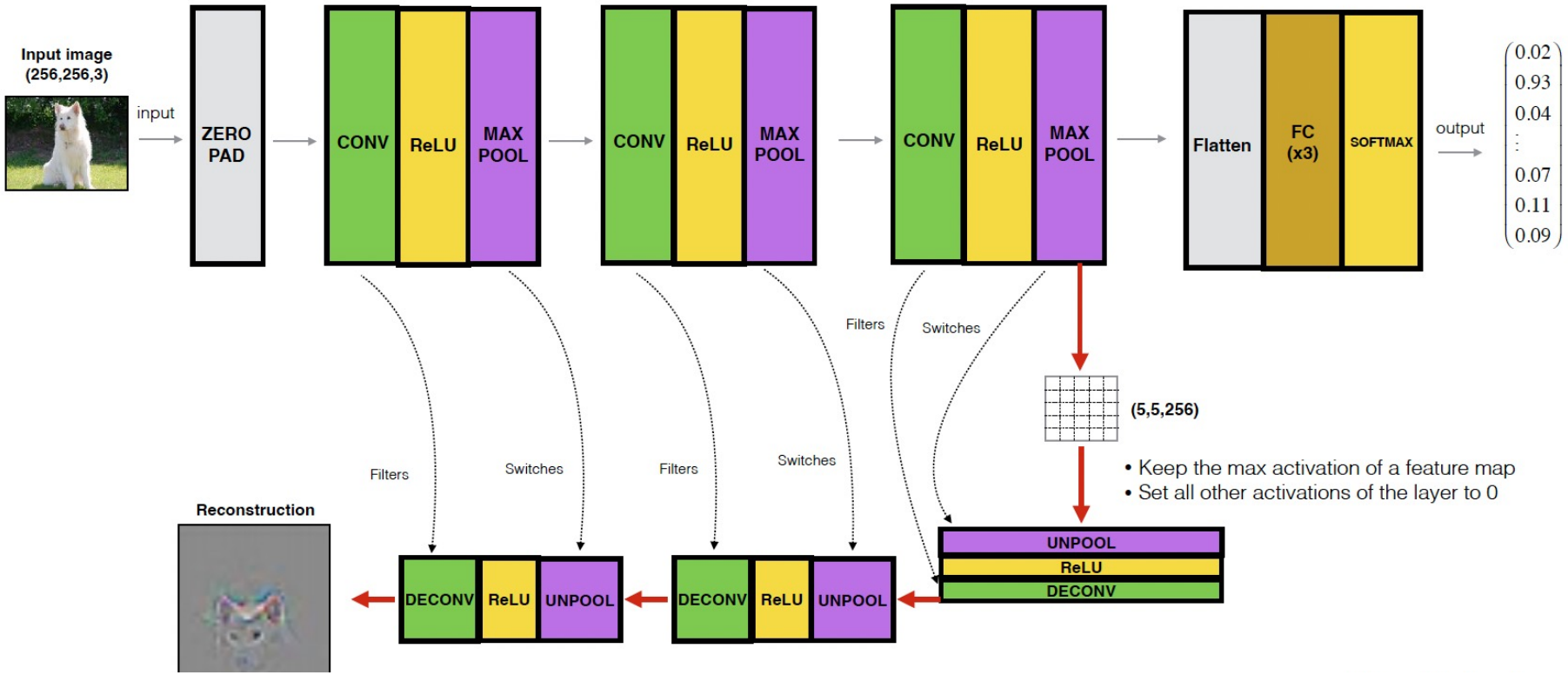
- **概念可视化**：可视化神经元、通道、深度和网络输出各个层面的概念

- 也可以进一步组合解释



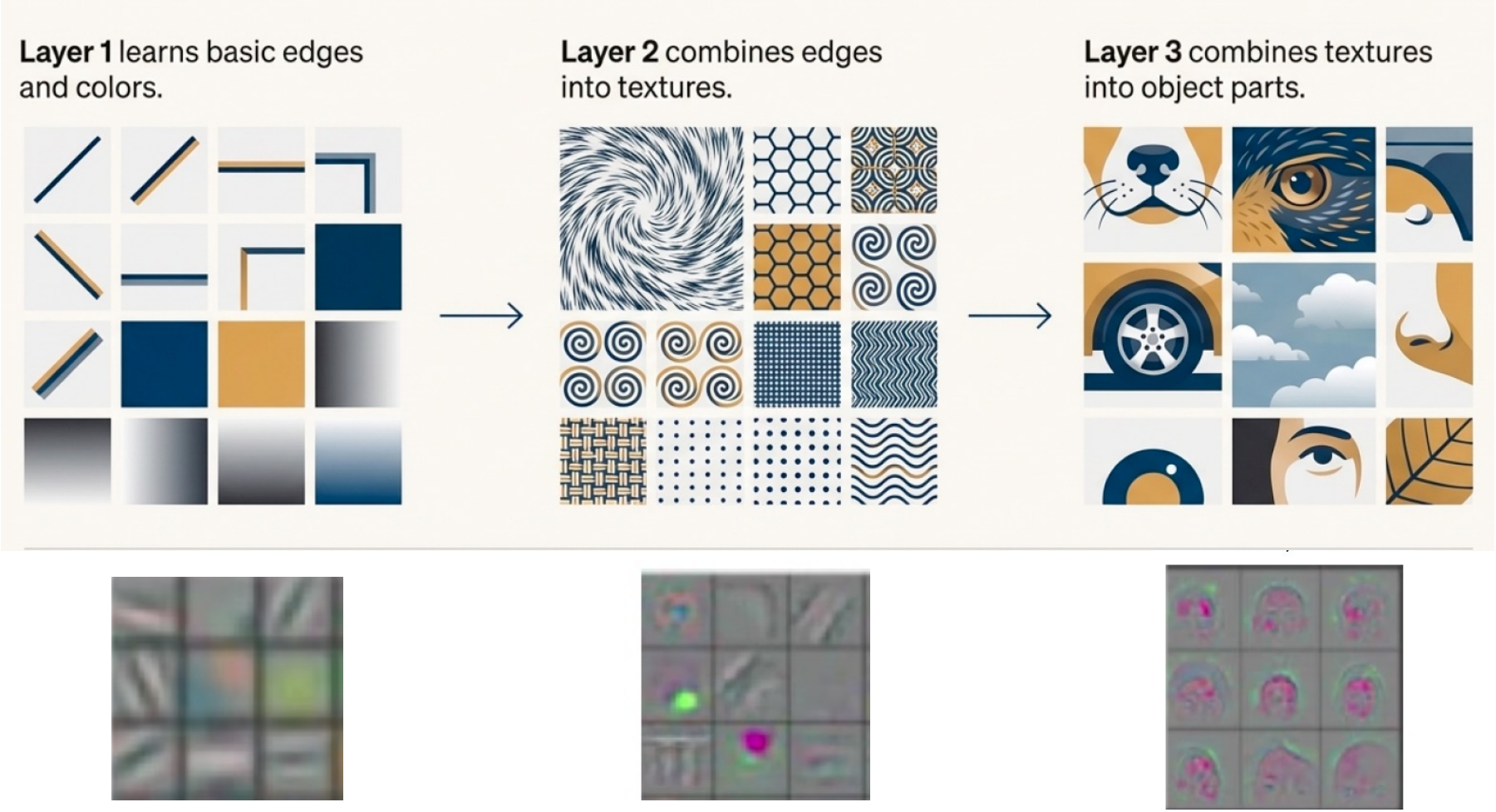

# 模型可解释性

- **反卷积（卷积转置）**：把卷积层y=Wx直接改写为x=$W^T$y
- 可以找到某个神经元对应的激活图



# 模型可解释性

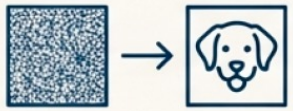
- 基于反卷积的**逐层分析**，随着网络加深，神经元感受野加大




The Building Blocks of Interpretability

# 小结

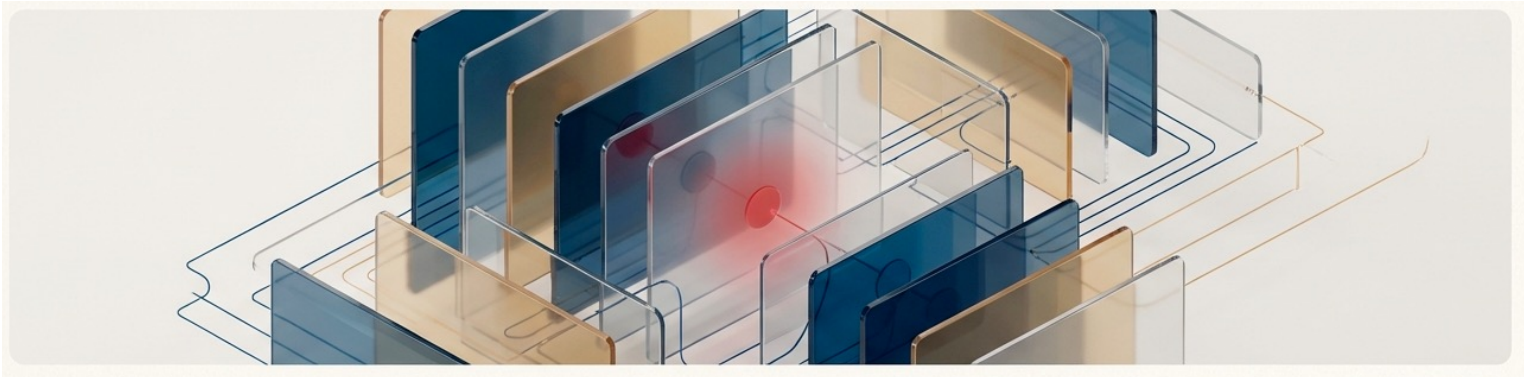
- 神经网络可解释性工具总结

| | | |
|---|---|---|
| Explain any classifier |  LIME |  SHAP |
| What part of the input is responsible for the output? |  Saliency Maps |  Occlusion Sensitivity |  Class Activation Maps (CAMs) |
| What is the model's internal "idea" of a concept? |  Gradient Ascent (Class Model Visualization) | |
| What is the role of a given neuron, filter, or layer? |  Deconvolution |  Dataset Search (finding images that maximize activation) |

# 目录

# 情景假设

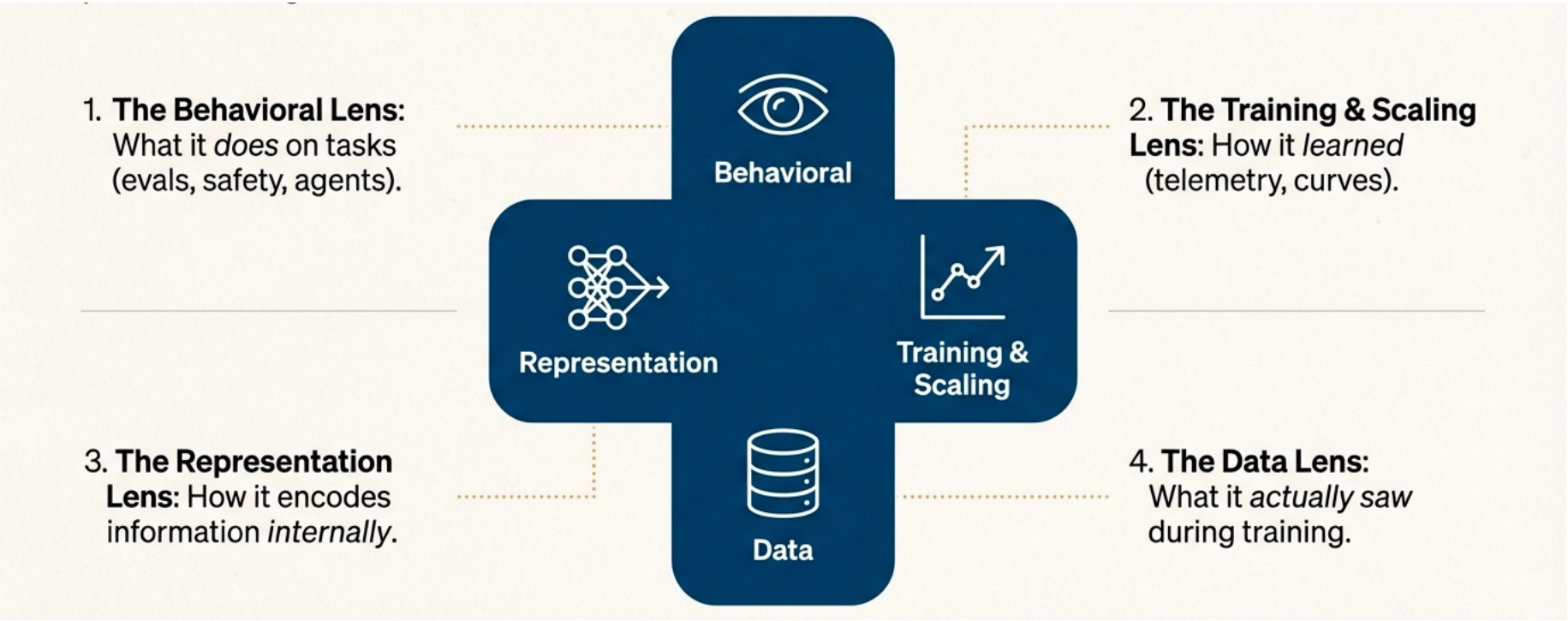
- 你是一位创业者，正在训练一个拥有200B参数的最新大语言模型（LLM），但发现其实际**表现未达预期**，主要问题包括：
  - 推理能力不佳；
  - 安全性弱，容易被越狱；
  - 在部分智能体（Agent）场景中延迟过高。



- 面对投资人的质疑：“问题到底出在哪里？”
  在深入检查复杂的训练代码之前，你应该如何系统性地排查模型问题？

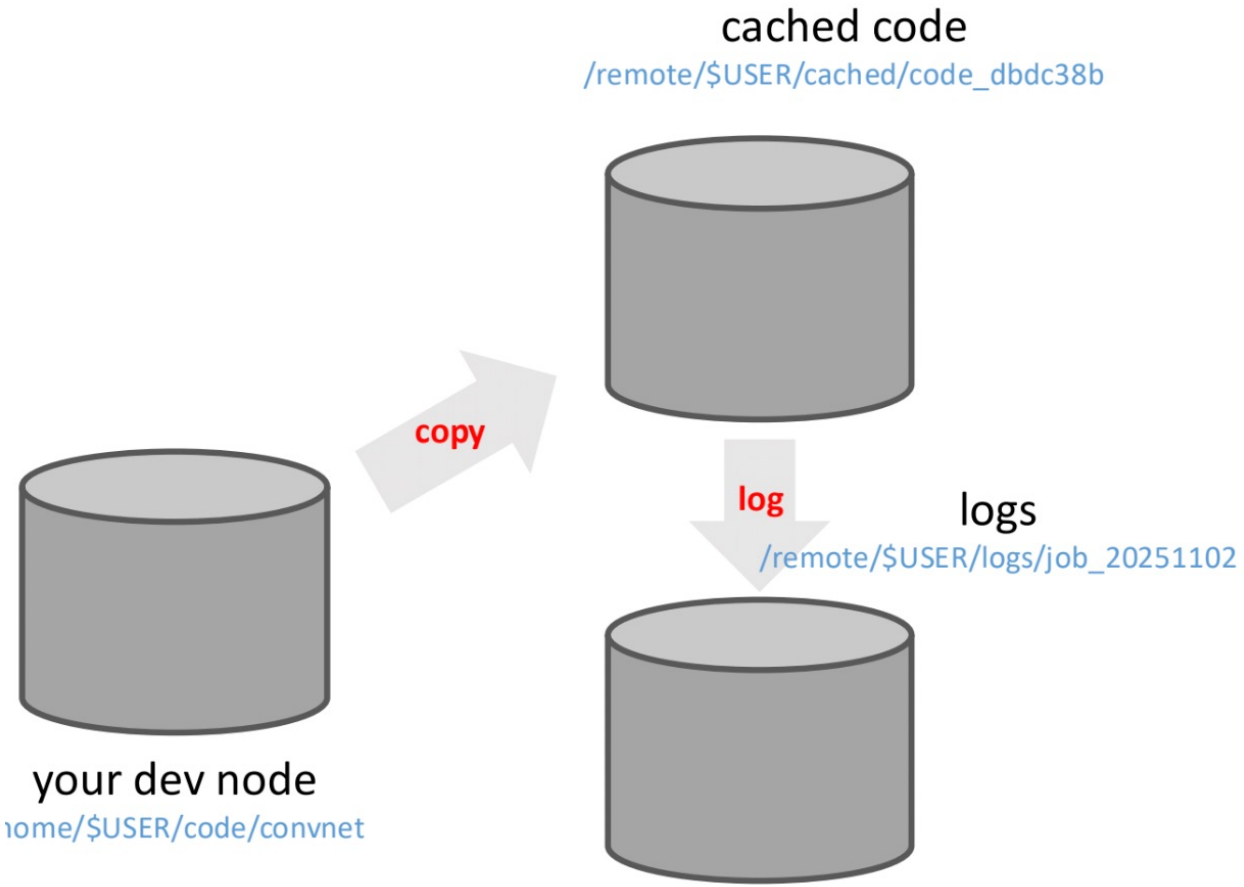
# 大语言模型中可解释性

- 检查LLM的四个维度：**表现**、**训练**、**表征**和**数据**



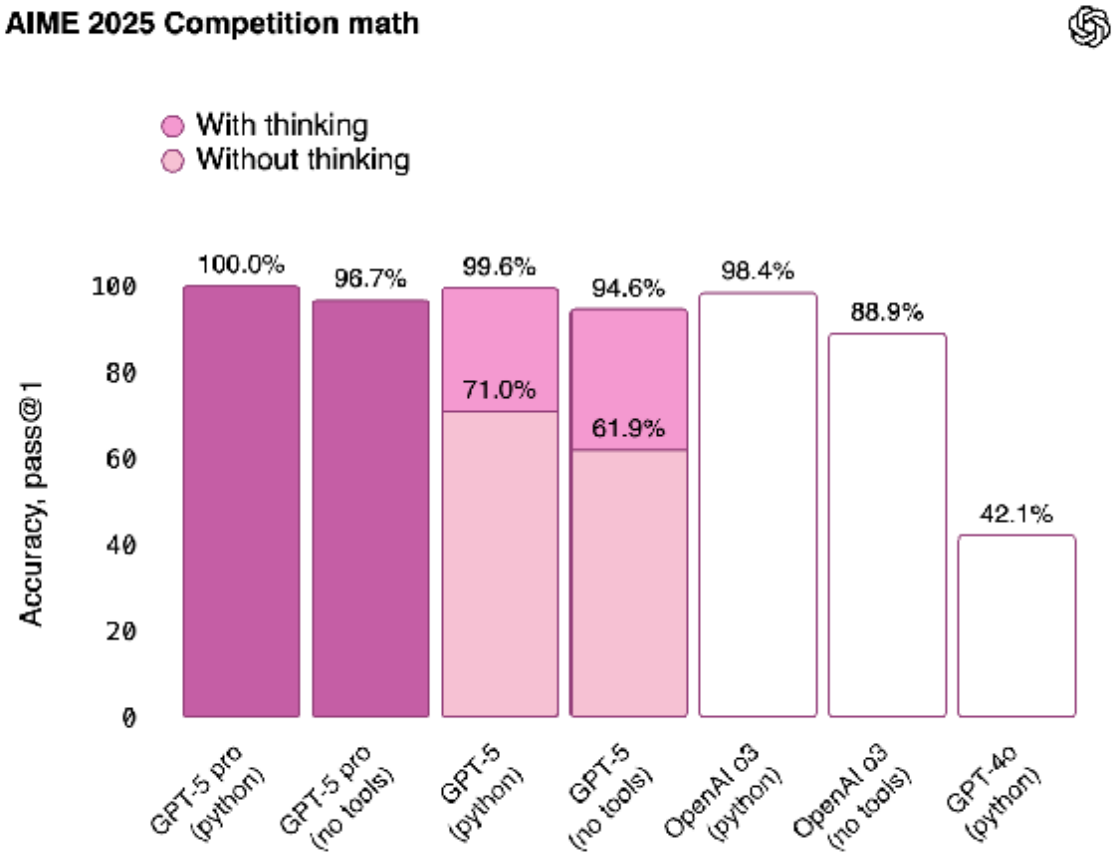
# 大语言模型中可解释性

- The Behavioral Lens
- 保存所有的代码修改过程（**Git**）
- 保证函数和commit的可读性
- 找到对应的过程代码，做好记录



# 大语言模型中可解释性

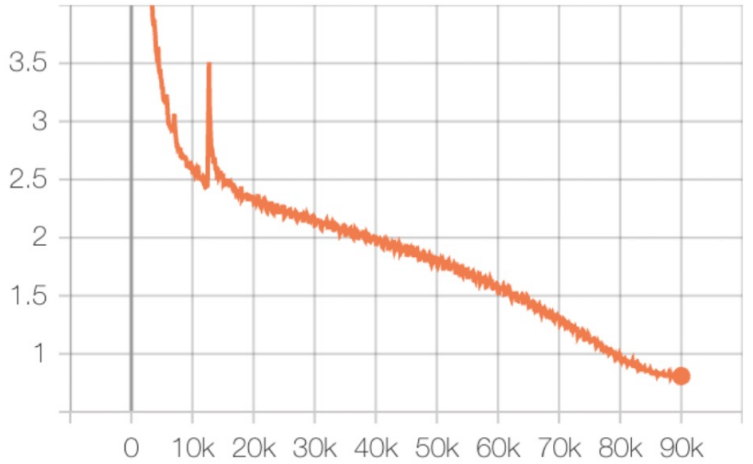
- The Behavioral Lens
- 比较**不同checkpoint的表现**
- 比较压力测试下的表现
- 比如jailbreak, misinformation, fairness, harmful content generation等




GPT-4 Technical Report

# 大语言模型中可解释性

- The Training and Scaling Lens
- 大部分时间不是在写代码，而是在检测**训练曲线**！
- **小心spike**，会极大影响模型效果




- 检查**学习率（往往最重要，在可承受范围内多尝试）**，warmup，优化器，batch size，初始化，计算精度
- 一次多跑几个控制变量的设定

# 大语言模型中可解释性

- The Training and Scaling Lens
- 是不是训练和验证曲线在下降
- 是不是**更多数据+计算->更好性能**
- 注意观察梯度范围
- 注意观察学习率、正则化改变的影响
- 更大的batch size，需同比增加学习率

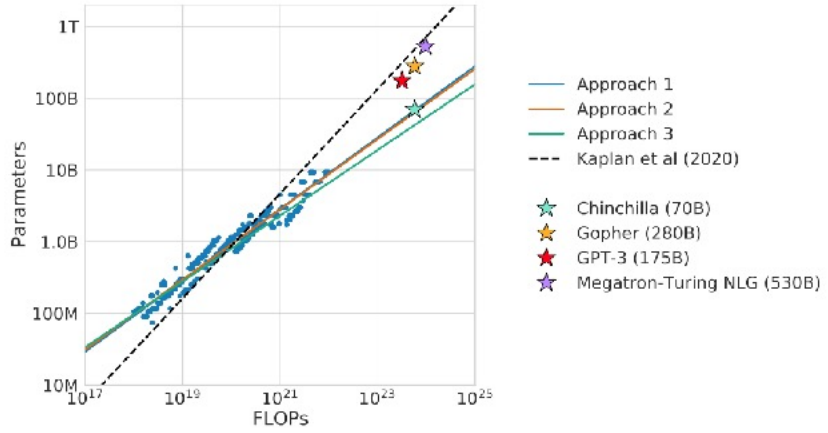


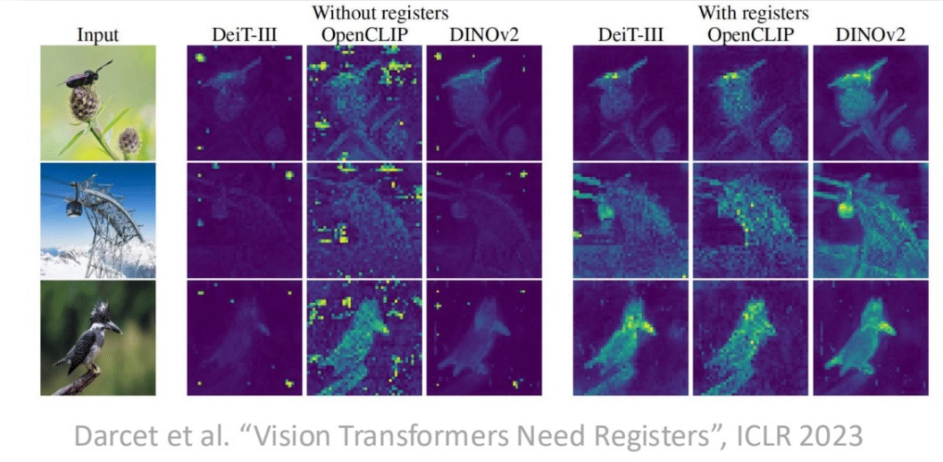

Figure 1 | **Overlaid predictions.** We overlay the predictions from our three different approaches, along with projections from Kaplan et al. (2020). We find that all three methods predict that current large models should be substantially smaller and therefore trained much longer than is currently done. In Figure A3, we show the results with the predicted optimal tokens plotted against the optimal number of parameters for fixed FLOP budgets. ***Chinchilla* outperforms *Gopher* and the other large models (see Section 4.2).**

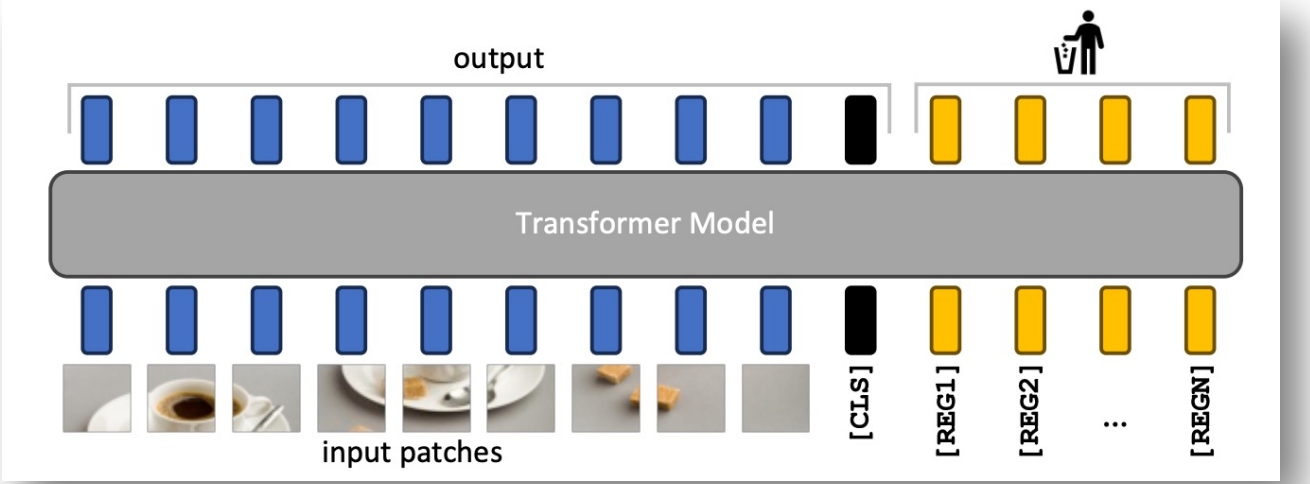
Training Compute-Optimal Large Language Models

# 大语言模型中可解释性

- The Representation Lens

- 通过**Attention map**来可视化（相关越高代表影响越大）




- 检查类似概念网络**embedding**是否也类似

# 大语言模型中可解释性

- The Representation Lens

- 针对LLMs，Anthropic提出思维环路的分析方法，用**稀疏特征替代神经元**

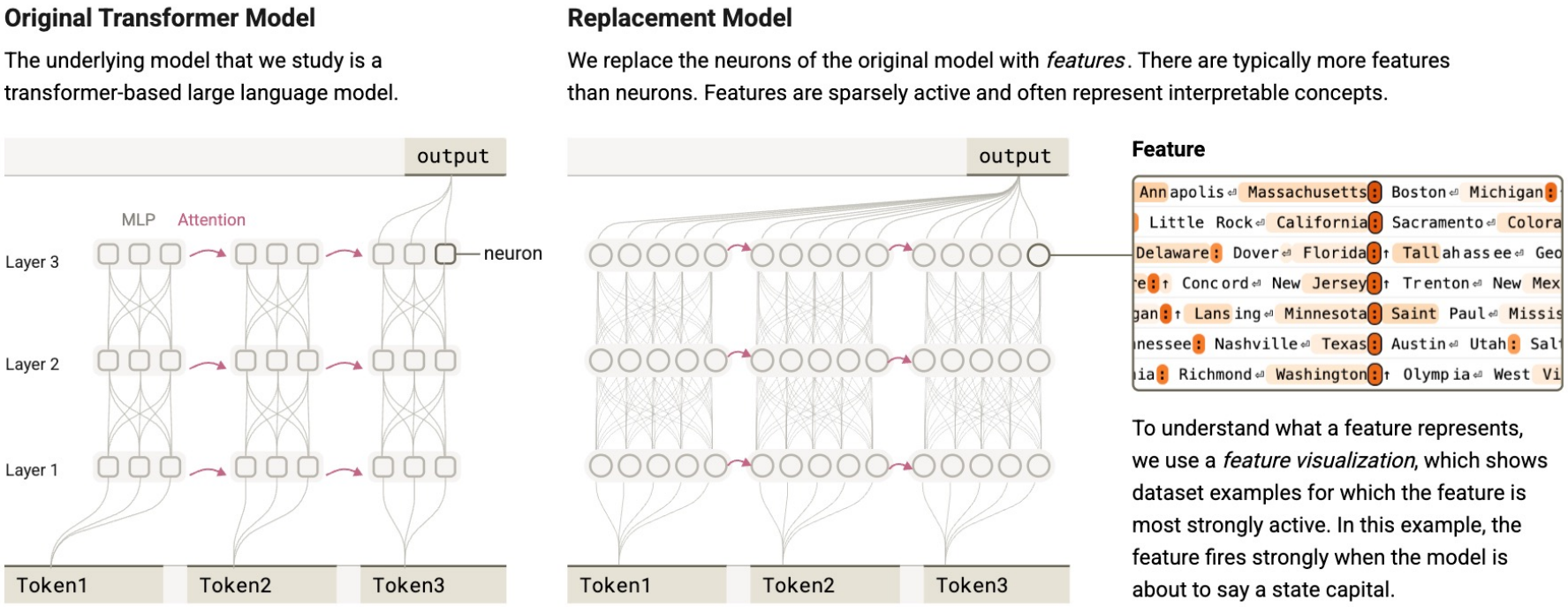
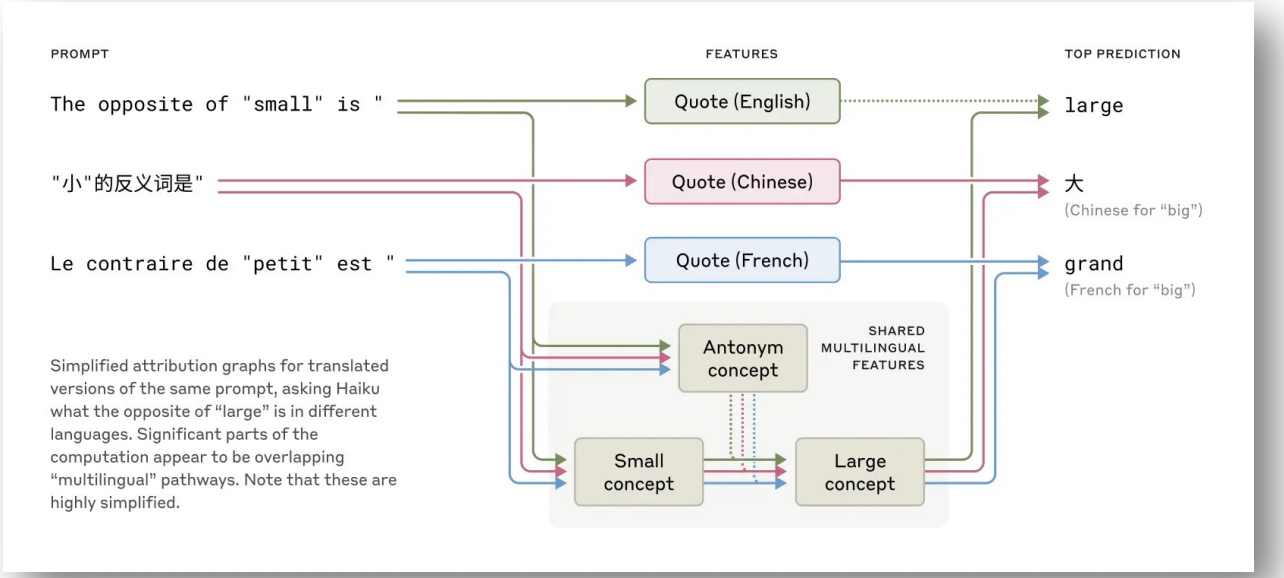


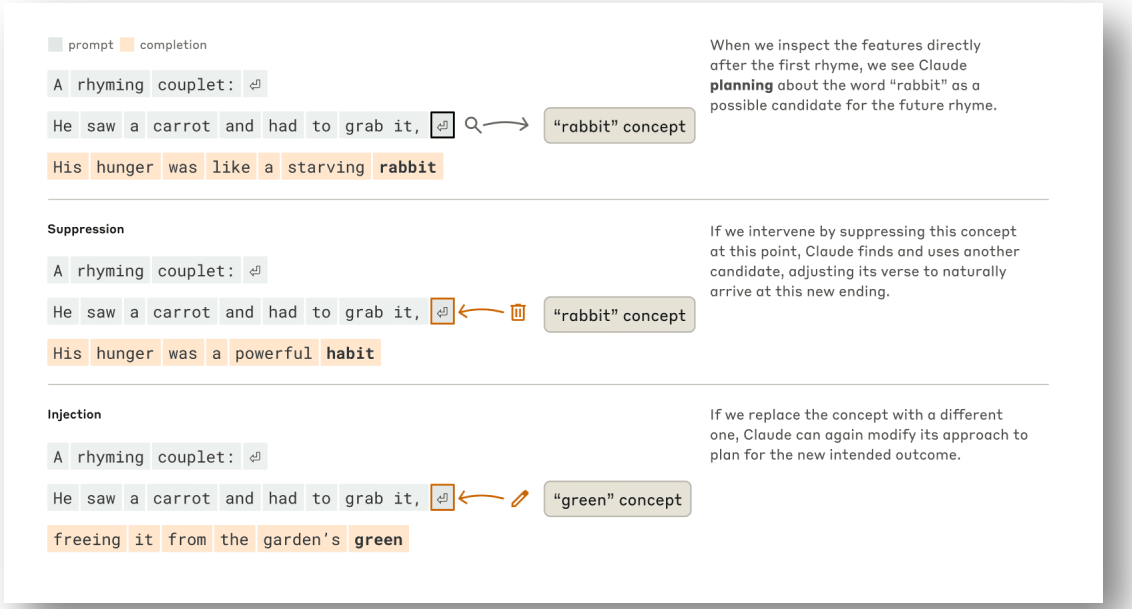

**Figure 2**: The replacement model is obtained by replacing the original model's neurons with the cross-layer transcoder's sparsely-active features.

Tracing the thoughts of a large language model

# 大语言模型中可解释性

- The Representation Lens
- 可以**追踪**甚至**控制**LLM的想法




Tracing the thoughts of a large language model

# 大语言模型中可解释性

- The Data Len
- 检查数据分布：**数据混合**方法是否改变?
- 词元统计：一些词元是不是出现过于平凡?
- 污染统计：是不是产生测试数据泄漏?

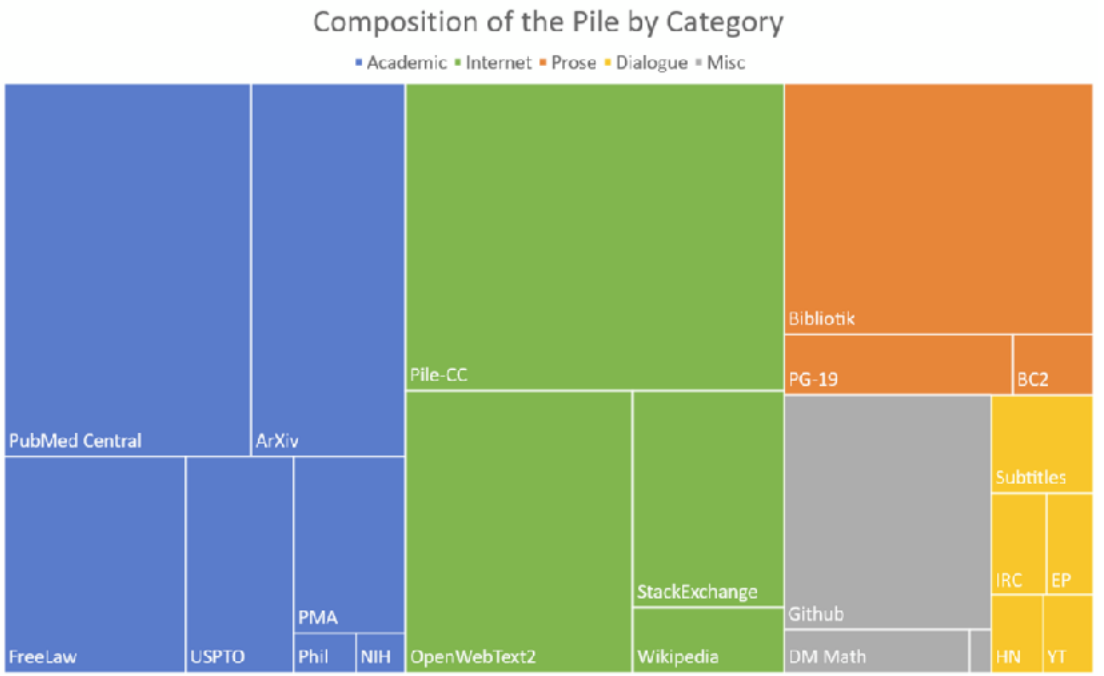
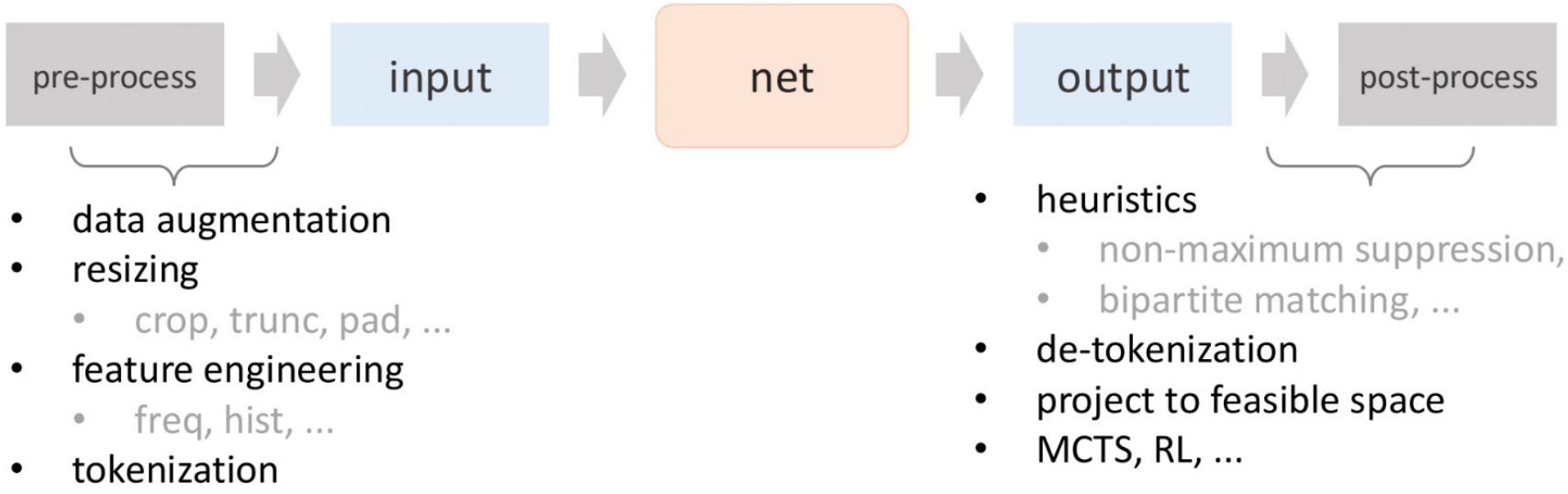



Figure 1: Treemap of Pile components by effective size.

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

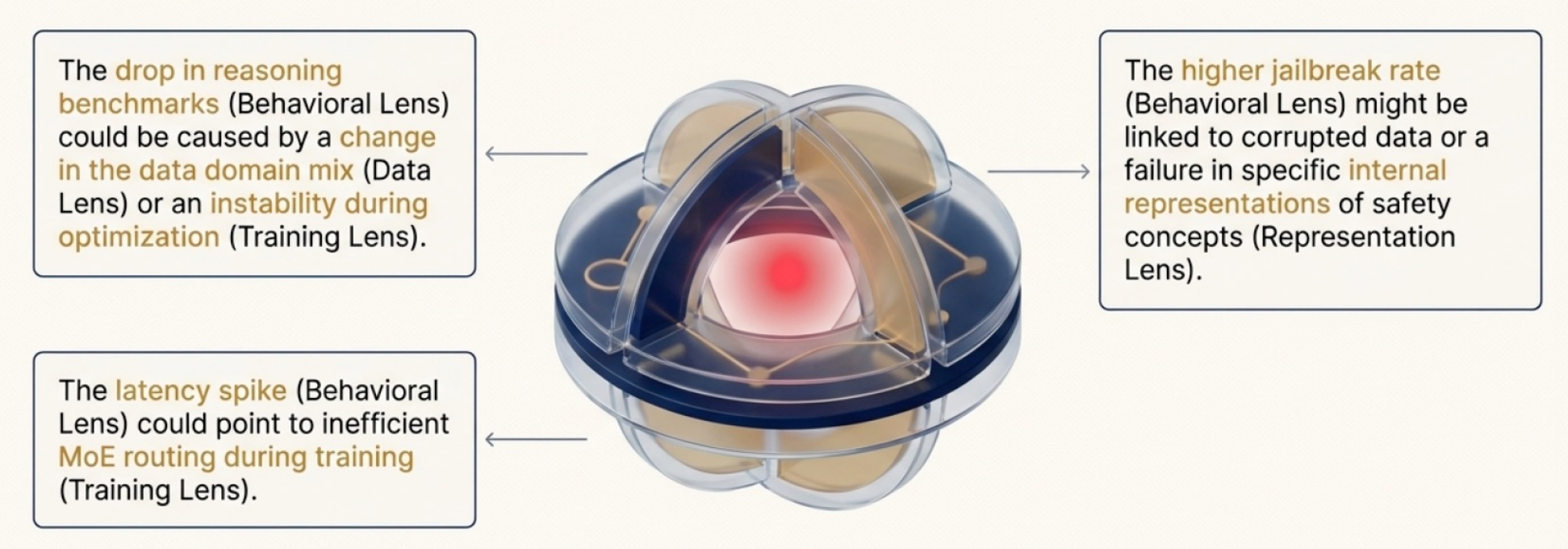
# 大语言模型中可解释性

- The Data Len
- 搞清楚输入输出是什么，知道你在做什么
- 搞清楚每一层的tensor形状，**手算网络参数和FLOPs**



# 小结

- 可解释分析不仅是一个学术练习，更是一种分析问题的方法论
- 多尝试，多体会，看看目标类似的开源项目



# 目录

# 期末项目

- **期末作业内容**：完成一篇课程论文。
- **提交内容一**：可以是科学博客文章或学术论文初稿，中英文皆可。
- **提交内容二**：至少**2篇**指定文献的阅读与注解文档。
- **截止日期**：北京时间1月1日 23:59。

# 评分政策

| 状态 | 提交时间 | 成绩说明 |
| --- | --- | --- |
| **按时提交** | 1月1日 23:59 之前 | **保障最低成绩为 B+** |
| **迟交提交** | 1月2日内 | 最高成绩不超过 B+ |
| | 1月3日内 | 最高成绩不超过 B |
| | 1月4日内 | 最高成绩不超过 B- |
| | 1月5日内 | 最高成绩不超过 C+ |
| | 1月6日内 | 最高成绩不超过 C |
| | 1月7日内 | 最高成绩不超过 C- |
| | 1月8日及之后 | D |

*注[1]：时间皆为北京时间，即CST
*注[2]：没有2篇注解文档成绩**降一档**
*注[3]：没参加课程汇报成绩**降两档**