中国科学院大学
University of Chinese Academy of Sciences

Concept Bottleneck Models

2024 年 12 月 20 日

杨泽康

中国科学院大学
University of Chinese Academy of Sciences

目录

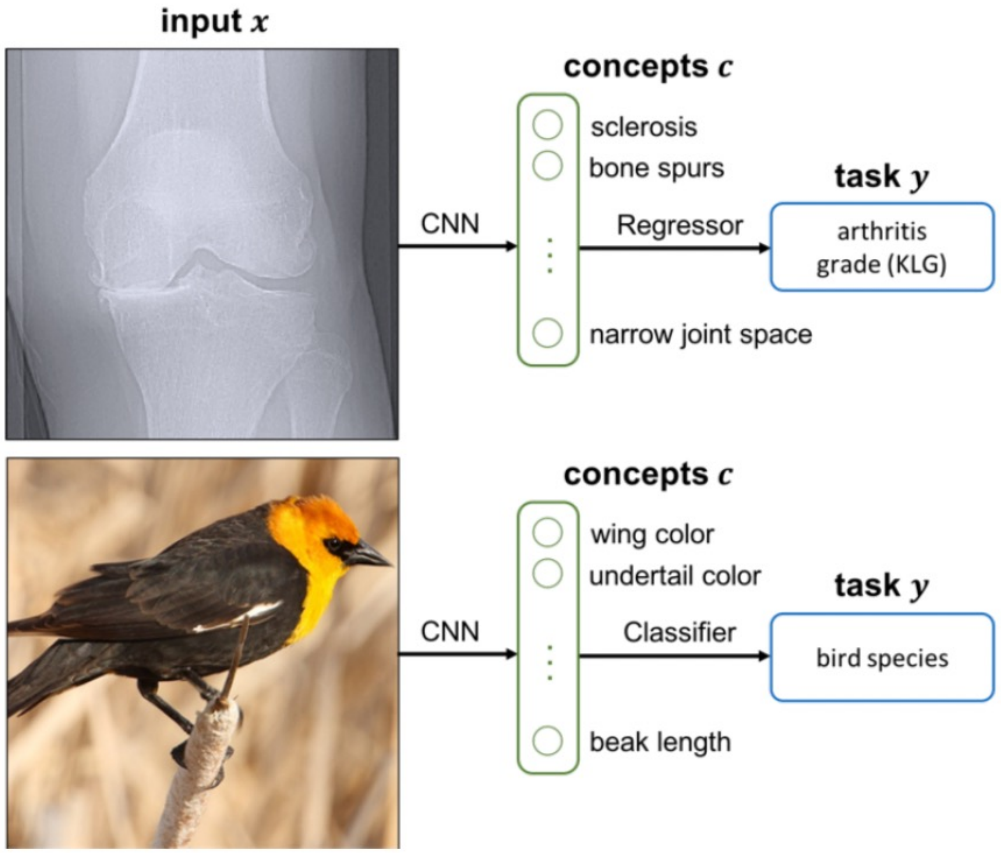
CONTENTS

Concept Bottleneck Models

PART ONE

Concept Bottle Models

CBMs：先由输入数据预测中间概念c=f(x)，再由中间概念预测下游任务y=G(c)。中间概念c为下游任务的预测提供了很好的**可解释性。**



1. 分阶段训练，先训练c->y, 再训练x->c

$$\hat{f} = \arg\min_f \sum_i L_Y(f(c^{(i)}), y^{(i)})$$

$$\hat{g} = \arg\min_g \sum_{i,j} L_C(g(x^{(i)}), c_j^{(i)})$$

2. 序列端到端训练

$$\hat{f}, \hat{g} = \arg\min_{f,g} \sum_i L_Y(f(g(x^{(i)})), y^{(i)})$$

3. 联合训练

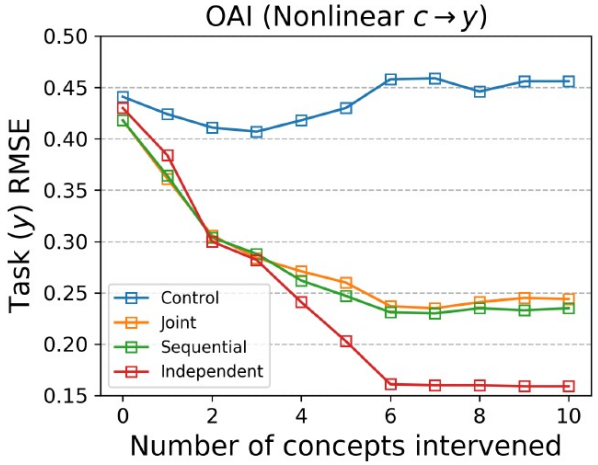
$$\hat{f}, \hat{g} = \arg\min_{f,g} \sum_i (L_Y(f(g(x^{(i)})), y^{(i)}) + \sum_j \lambda L_{C_j}(g(x^{(i)}), c_j))$$

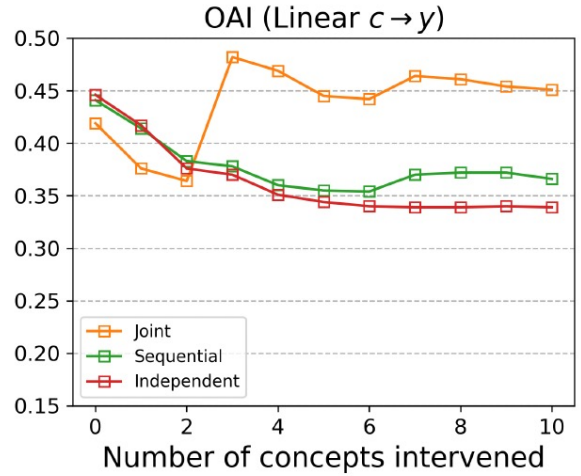
Koh, Pang Wei, et al. "Concept bottleneck models." *International conference on machine learning*. PMLR, 2020.

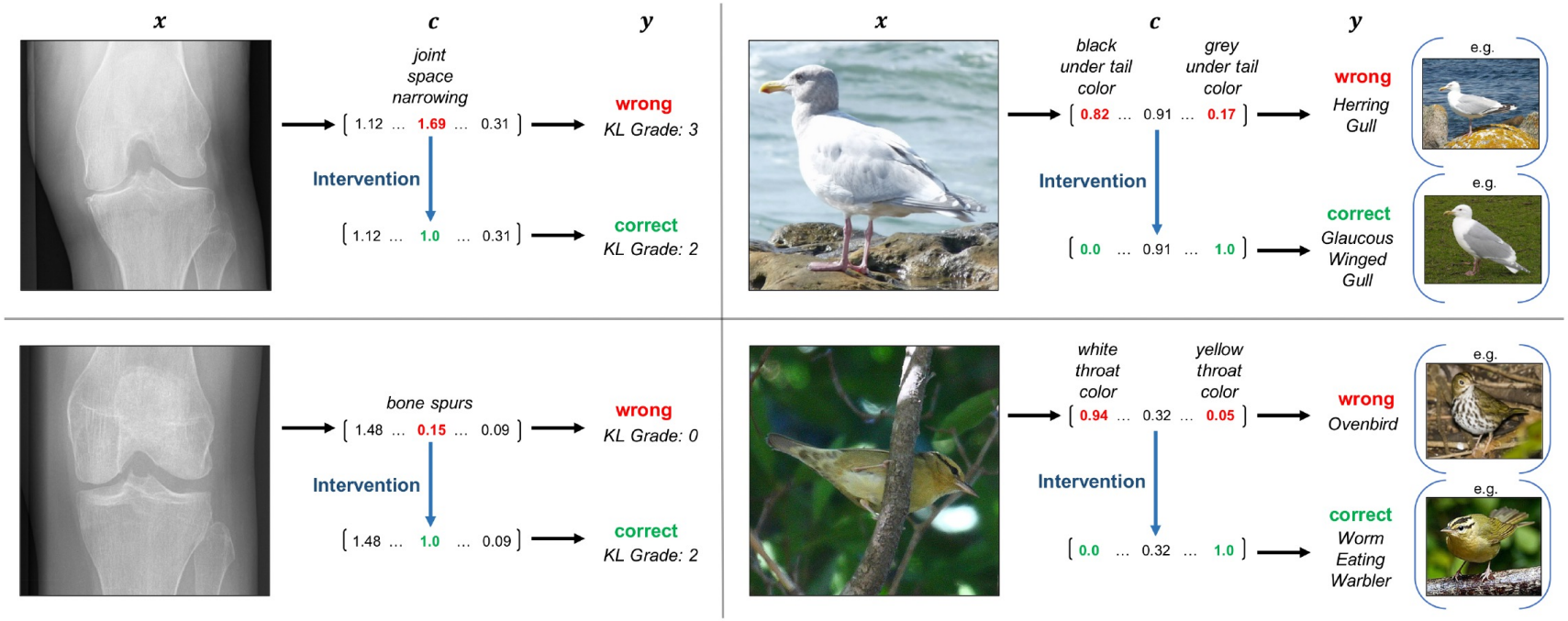
Concept Bottle Models

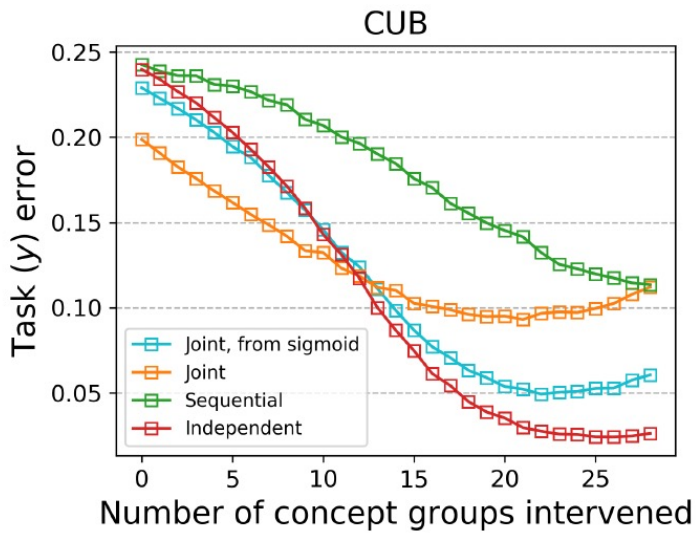
CBMs——Test-time Intervention

在推理时可以通过人工矫正预测错误的概念c来矫正错误预测的y。









Koh, Pang Wei, et al. "Concept bottleneck models." *International conference on machine learning*. PMLR, 2020.

University of Chinese Academy of Sciences

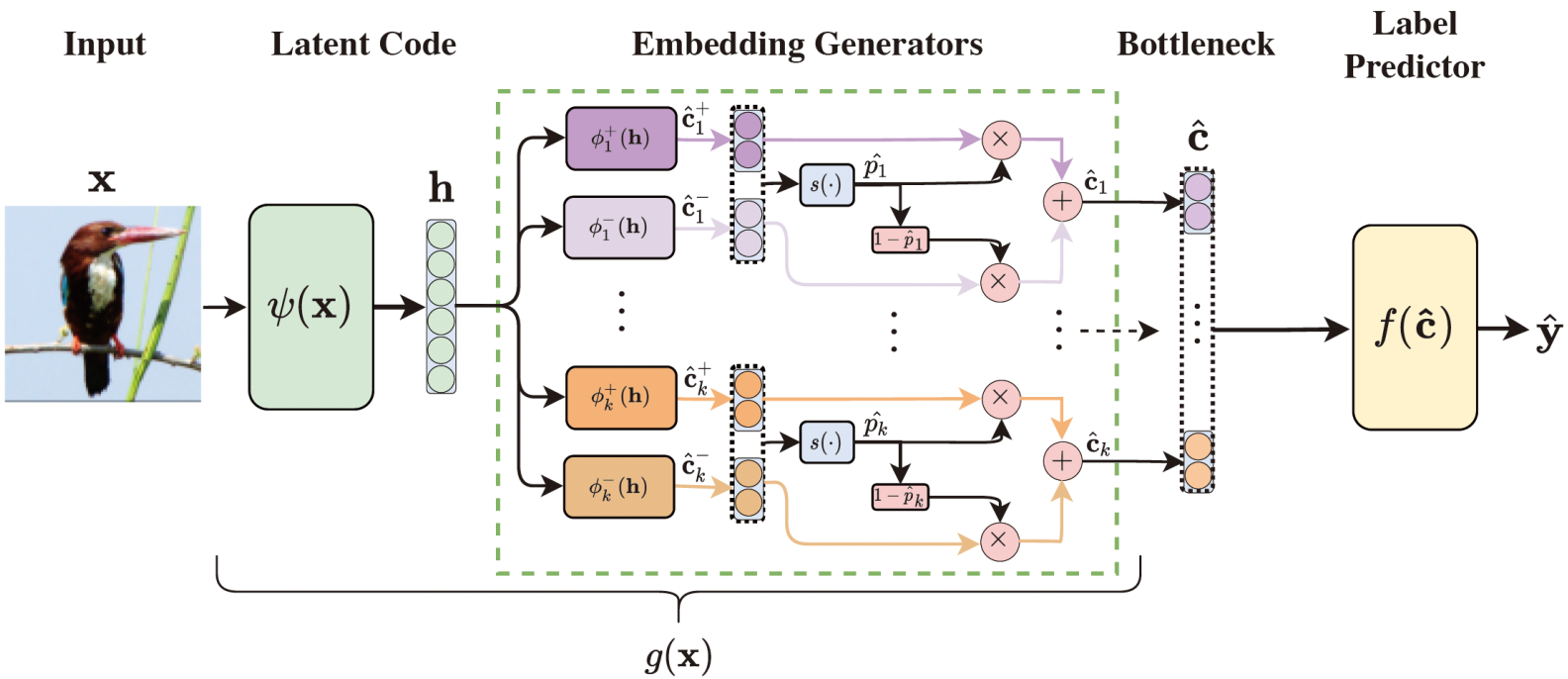
Performance vs Interpretability

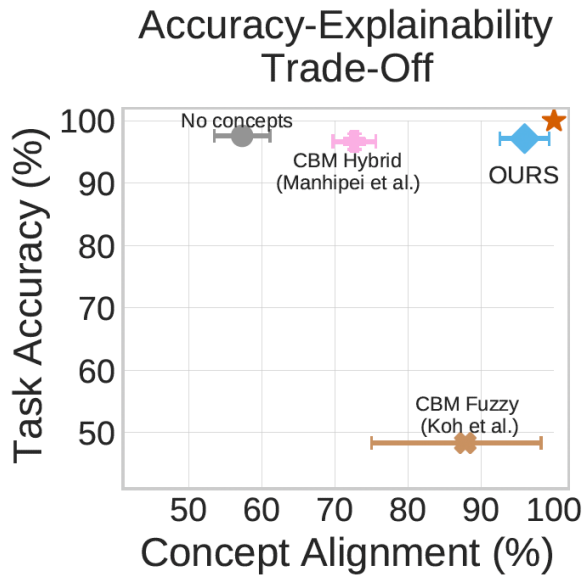
PART TWO

CBMs有很好的可解释性，但牺牲了一些模型表现。

1. 不再用一个单一的激活值表示是否存在某个concept，用两个embedding表示是否存在这个概念，并用一个MLP预测这两个embedding的相对重要性，对embedding加权后作为这个concept的特征进行下游任务。

2. Test-time干预时，不再加权，直接用相应的embedding替换相应concept的特征。





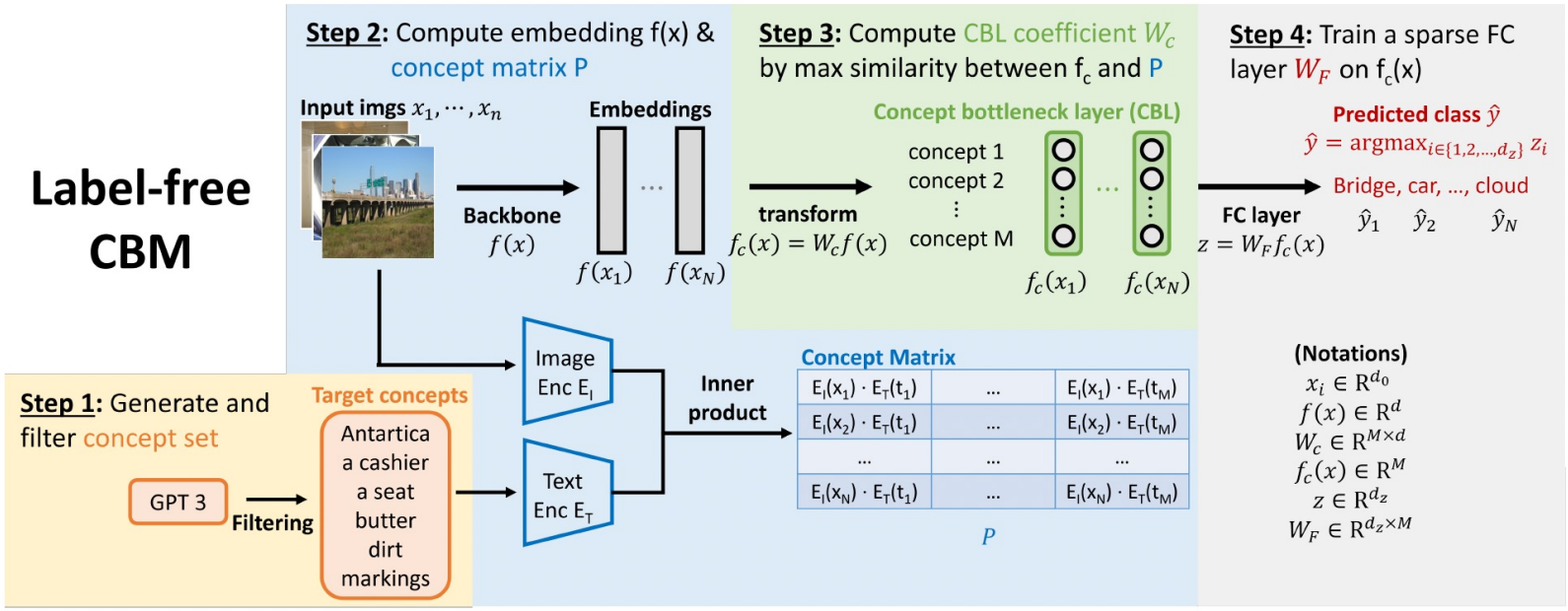
Espinosa Zarlenga, Mateo, et al. "Concept embedding models: Beyond the accuracy-explainability trade-off." *Advances in Neural Information Processing Systems* 35 (2022): 21400-21413.

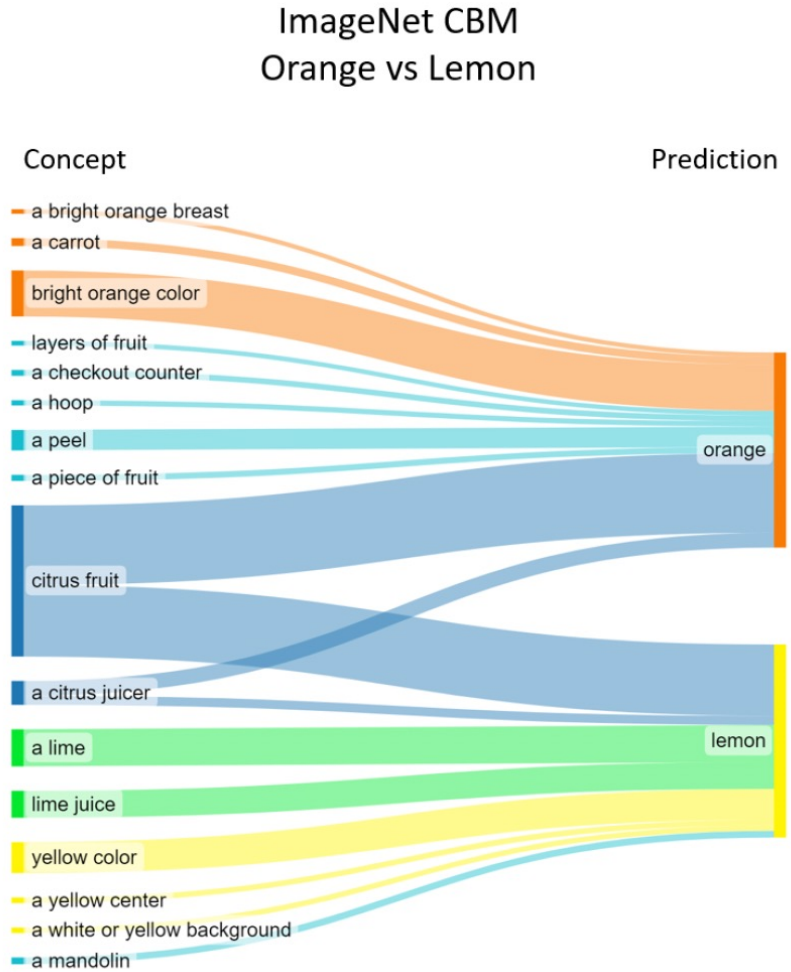
Label-free CBMs

PART THREE

Challenge: 训练CBM模型需要concept标注。

1. 用GPT 3针对图像的label自动生成一些concept。
2. 用预训练的CLIP模型为每张图片提供concept标注。



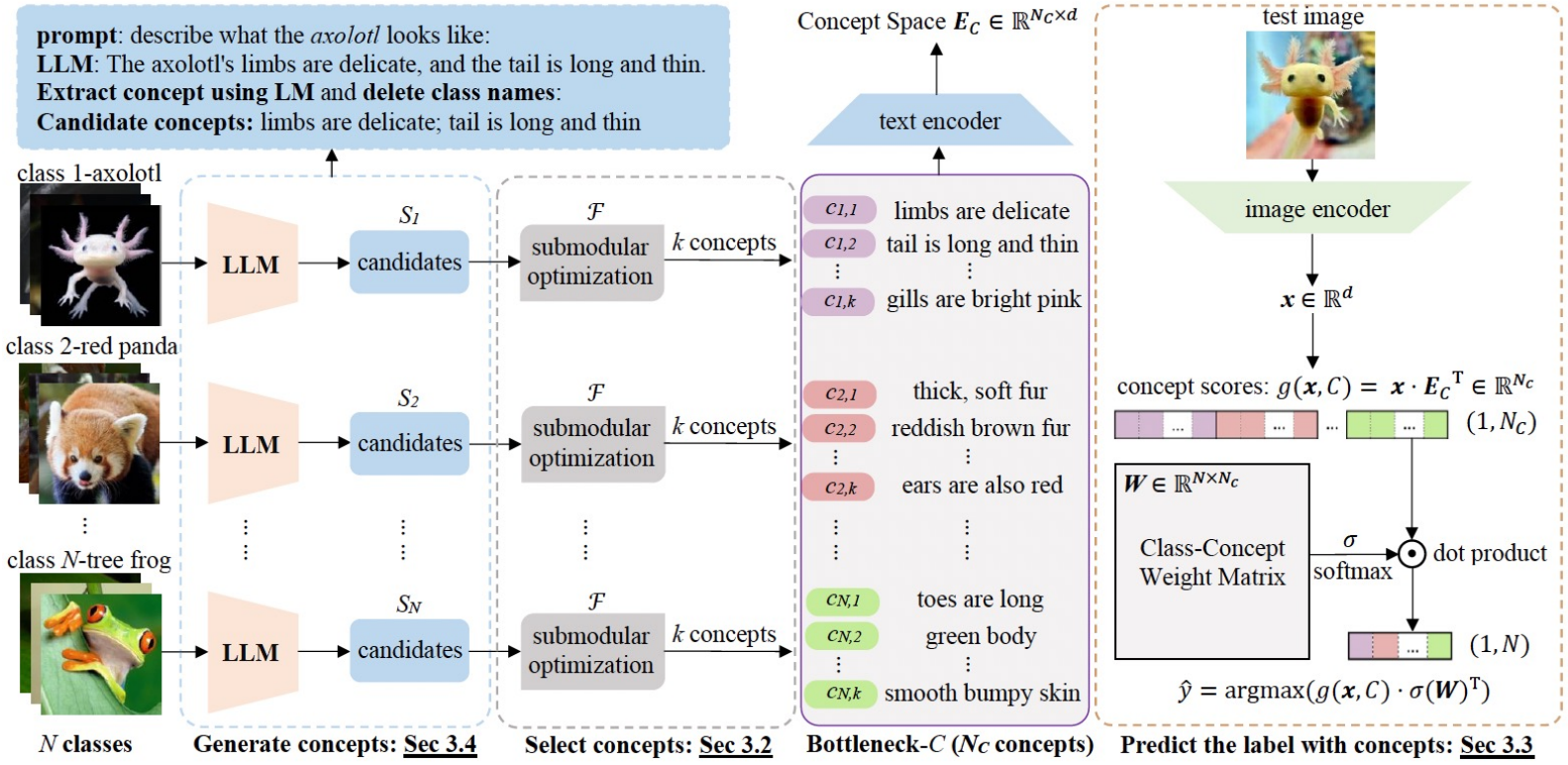


Oikarinen, Tuomas, et al. "Label-free Concept Bottleneck Models." *International Conference on Learning Representations*. 2023.

获取高质量的concept标注困难。

1. 借助GPT针对每个label生成一些对应的concept。
2. 每个label对应大量concept，通过可学习的topK方法和优化一个目标函数，找到合适的NxK个concept。
   - 判别性：和label embedding相似性的最大似然。
   - 覆盖性：最大化和所有concept的相似度。
3. 只训练C->Y模型。预训练的CLIP得到图像预测的C，预测的C->Y用few-shot方式微调。

$$\mathcal{F}(C_y) = \underbrace{\alpha \cdot \sum_{c \in C_y} D(c)}_{\text{discriminability}} + \underbrace{\beta \cdot \sum_{c_1 \in S_y} \max_{c_2 \in C_y} \phi(c_1, c_2)}_{\text{coverage}},$$



Yang, Yue, et al. "Language in a bottle: Language model guided concept bottlenecks for interpretable image classification." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

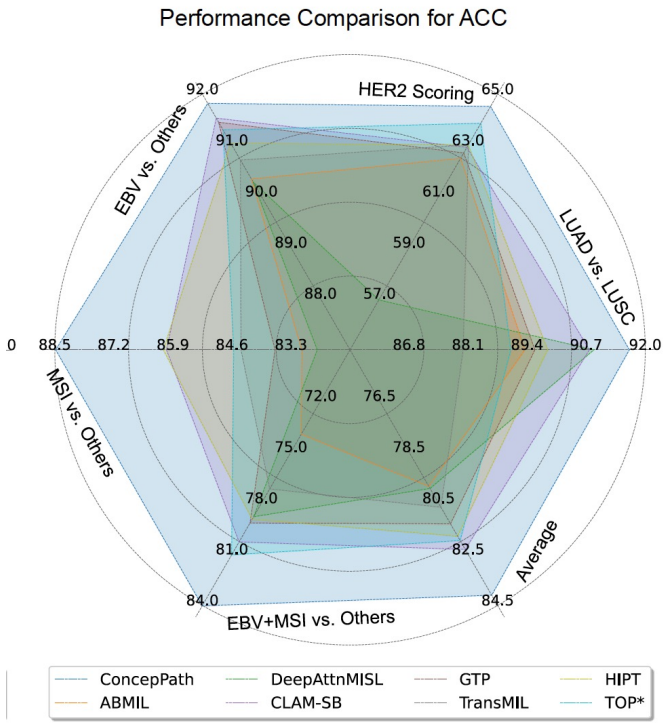
中国科学院大学
University of Chinese Academy of Sciences

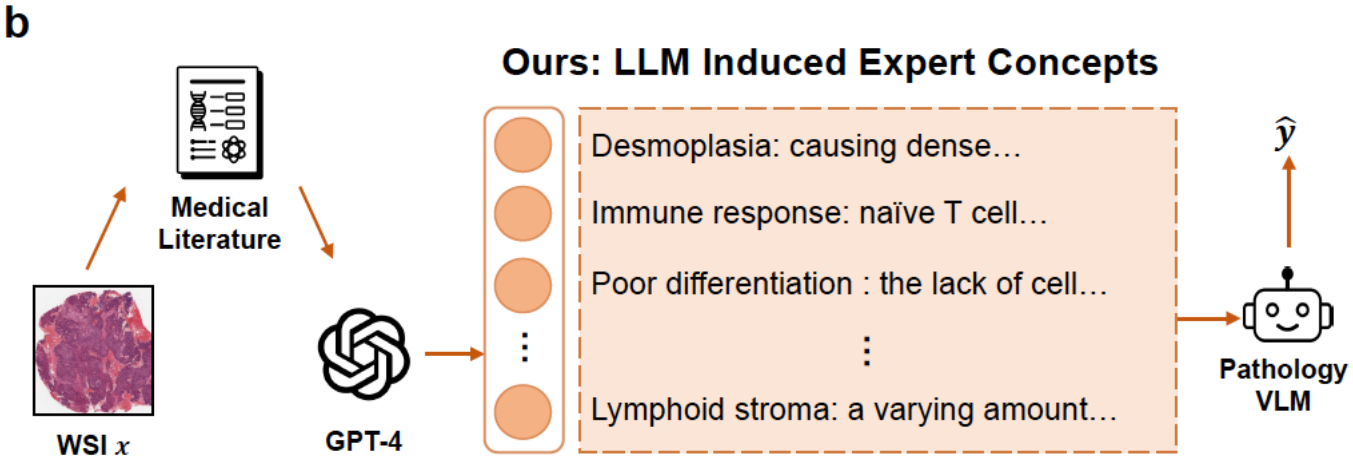
Application in Computational Pathology

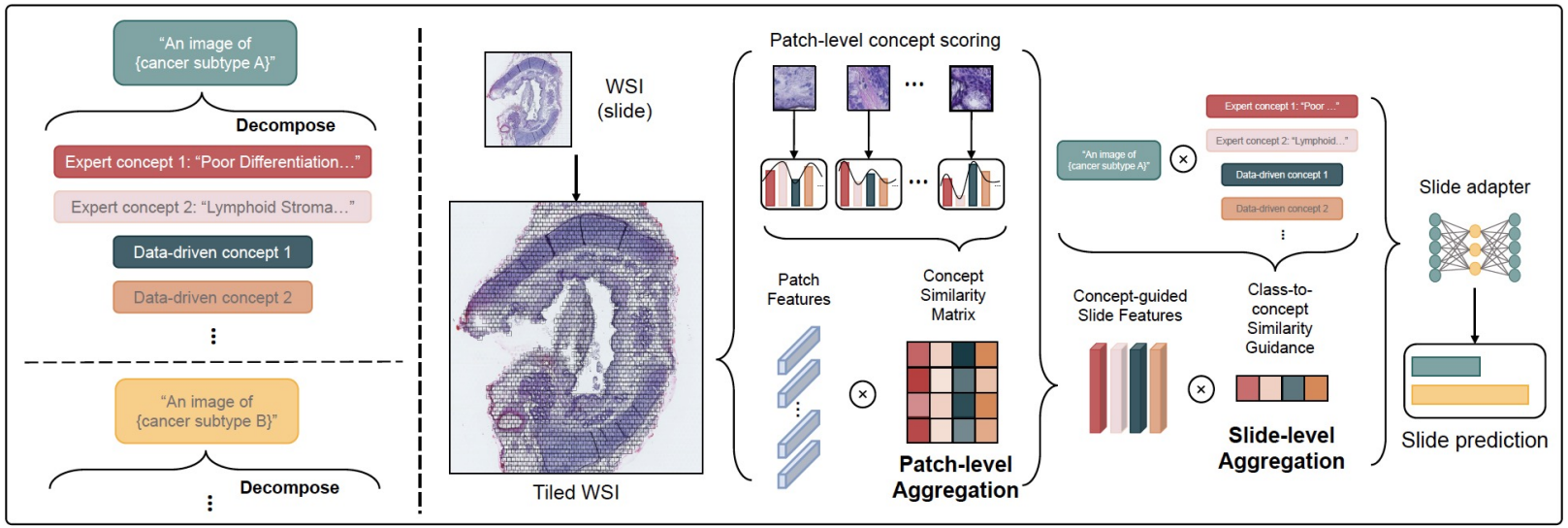
PART FOUR

ConcepPath

1. 借助医疗文献和GPT-4生成相关医学概念。
2. 预训练的CLIP得到每个patch各个概念的概率。
3. 与多示例学习结合进行下游任务预测。







Zhao, Weiqin, et al. "Aligning Knowledge Concepts to Whole Slide Images for Precise Histopathology Image Analysis." *arXiv preprint arXiv:2411.18101* (2024).

University of Chinese Academy of Sciences

End

What can we do ?

1. **Language in a Bottle方法存在的问题**
    1. 没有一个concept合并步骤，提取不同label之间共有的concept。
    2. Concept优化和模型优化是分开的步骤，可以考虑采用类似CoOp的方式合并两个步骤，做成端到端的有助于提升下游任务表现。
2. **医学中更关注算法的可解释性**
    1. 优化Test-time Intervention过程，用较少的干预带来较大的准确率提升，作为辅助诊断算法。

Thanks