Aligning Human Knowledge with Visual Concepts Towards Explainable Medical Image Classification

(Share in the Group of Paper Lecture)

Haifei Zhang

LABORATOIRE HUBERT CURIEN
UMR • CNRS • 5516 • SAINT-ETIENNE

18 February, 2025

Outline

1. Background
2. Explicd method: Explainable language-informed criteria-based diagnosis

Outline

XAI: Explainable Artificial Intelligence

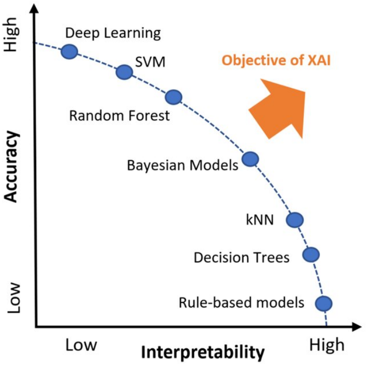


Figure: Trade off between explainability and performance of different AI models.[1]

XAI methods

1. Intrinsic interpretable models
2. Post-hoc explanations

[1] González-Alday, R., García-Cuesta, E., Kulikowski, C. A., & Maojo, V. (2023). A Scoping Review on the Progress, Applicability, and Future of Explainable Artificial Intelligence in Medicine. Applied Sciences, 13(19), 10778.

XAI: Explainable Artificial Intelligence

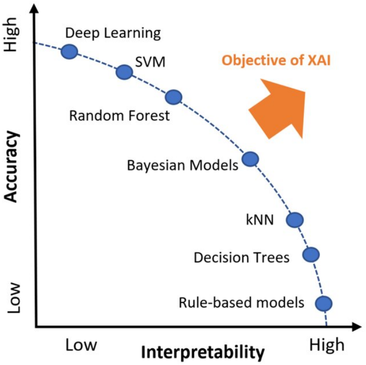


Figure: Trade off between explainability and performance of different AI models.[1]

XAI methods

1. Intrinsic interpretable models
2. Post-hoc explanations
3. **Combination of the two above?**

[1] González-Alday, R., García-Cuesta, E., Kulikowski, C. A., & Maojo, V. (2023). A Scoping Review on the Progress, Applicability, and Future of Explainable Artificial Intelligence in Medicine. Applied Sciences, 13(19), 10778.

CBMs: Concept Bottleneck Models

Hybrid strategy of black-box models and self-interpretable models

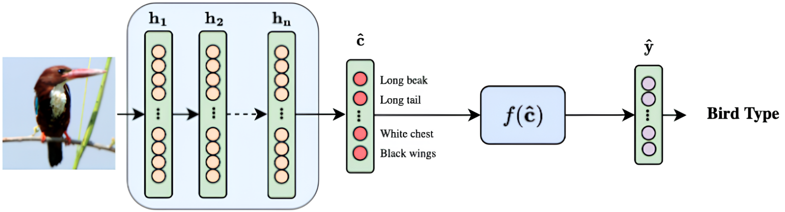


Figure: Concept Bottleneck Models learn predictions as a function of concepts.[2]

- **Training data:** $\{x_i,\ c_i,\ y_i\}_{i=1}^{N}$, where $c_i \in \mathbb{R}^K$
- **To learn** concept predictor $\hat{c}_i = \hat{g}(x_i)$ and label predictor $\hat{y}_i = \hat{f}(\hat{c}_i)$

[2] Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020, November). Concept bottleneck models. In International conference on machine learning (pp. 5338-5348). PMLR.

VLMs: Visual-Language Models

- A Vision Encoder – A deep learning model (e.g., CNNs, ViTs) that processes images and extracts meaningful features.
- A Language Encoder – A transformer-based model (e.g., BERT, GPT) that processes text and understands context.
- A Fusion block and decoder – A method (e.g., cross-attention) that aligns visual features with text, allowing the model to relate image content to language.

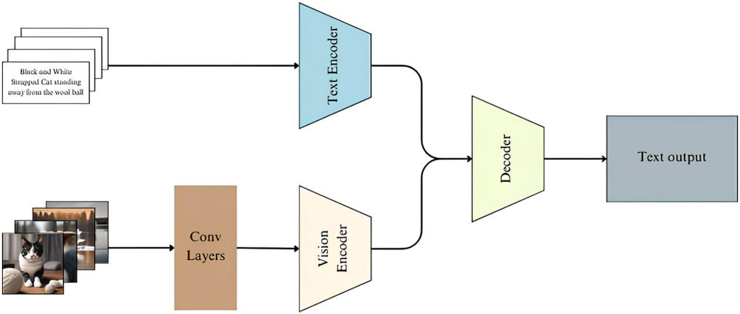
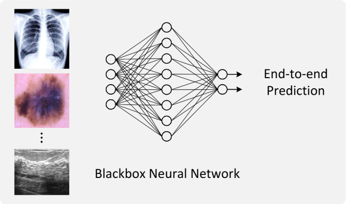
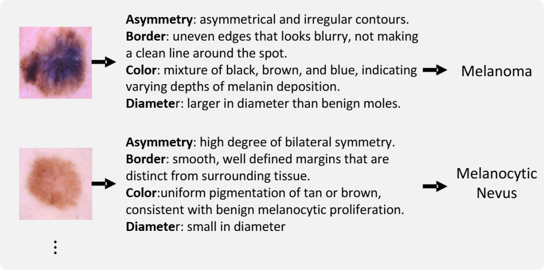


Figure: Typical structure of visual-language models.

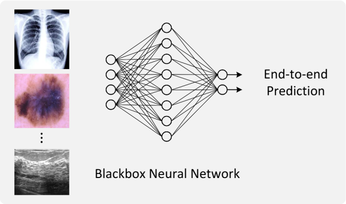
Outline

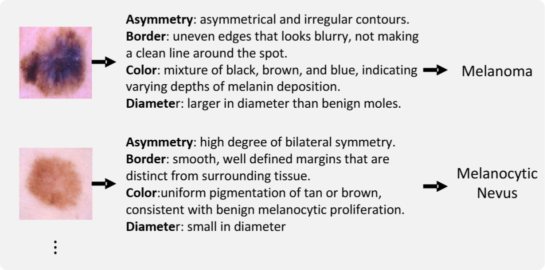
What is the problem?





What is the problem?





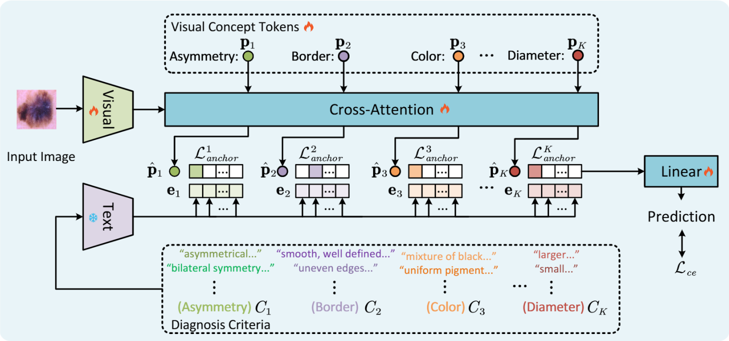
How to mimic the diagnostic process of human experts?

How to align human knowledge with visual concepts?[3]

[3] Gao, Y., Gu, D., Zhou, M., & Metaxas, D. (2024, October). Aligning human knowledge with visual concepts towards explainable medical image classification. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 46-56). Cham: Springer Nature Switzerland.

Main idea

1. Querying domain knowledge (diagnostic criteria) from LLMs or human experts.
2. Encoding these criteria as knowledge anchors using a pre-trained VLM's text encoder.
3. Learning visual concepts associated with these criteria via contrastive loss.
4. Using an interpretable model based on learned concepts to make predictions.



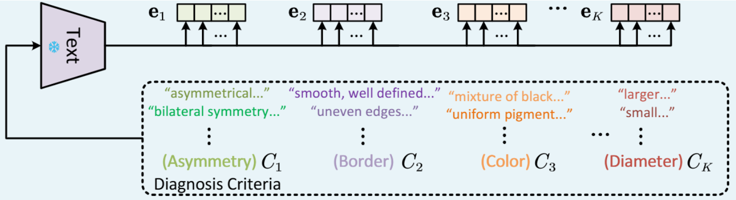
Domain knowledge and its embedding

Dataset: $\{(x_i, y_i)\}_{i=1}^N$, where x_i is an image and $y_i \in \mathcal{Y}$ is its label.

Domain knowledge (from LLMs or human experts):

- Problem-specific diagnosis criteria axes: $\{C_j\}_{j=1}^K$, e.g., for skin lesions, the criteria axes include asymmetry, border, color, diameter, texture, pattern, etc.
- Possible options within a particular criterion axis: $C_j = \{c_j^1, \dots, c_j^{k_j}\}$.
- For each image, according to its class, the ground truth value for each diagnostic criterion is recorded.
- Criteria anchor embeddings: $\{\boldsymbol{e}_j = \mathcal{T}(C_j)\}_{j=1}^K$, where $\mathcal{T}$ is a text encoder and $\boldsymbol{e}_j \in \mathbb{R}^{k_j \times d}$.

Note that, there is nothing need to learn in this step.



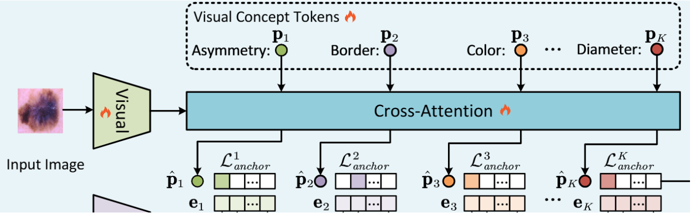
Visual concept learning

- Visual concept tokens: $\boldsymbol{p} \in \mathbb{R}^{K\times d}$, with each token designated to represent one of the criteria axes.
- For image x, its feature map given by a visual encoder $\mathcal{V}$: $\mathcal{V}(x)$.
- Visual concept encoding: $\hat{\boldsymbol{p}}(x) = \text{cross-attention}(\boldsymbol{p}, \mathcal{V}(x), \mathcal{V}(x))$.
- Criteria anchor contrastive loss (note $\hat{\boldsymbol{p}} = \hat{\boldsymbol{p}}(x)$ for simplification):

$$\mathcal{L}_{anchor}(\hat{\boldsymbol{p}}, \boldsymbol{e}_1, \ldots, \boldsymbol{e}_K) = -\frac{1}{K}\sum_{j=1}^{K} \log \frac{\exp(sim(\hat{\boldsymbol{p}}_j, \boldsymbol{e}_j^{\text{positive}})/\tau)}{\sum_{l=1}^{k_j} \exp(sim(\hat{\boldsymbol{p}}_j, \boldsymbol{e}_j^{l})/\tau)},$$

where τ controls the softness of the softmax function and the dot product calculates the similarity.

Note that, we need to learn $\boldsymbol{p}$ and $\mathcal{V}$.



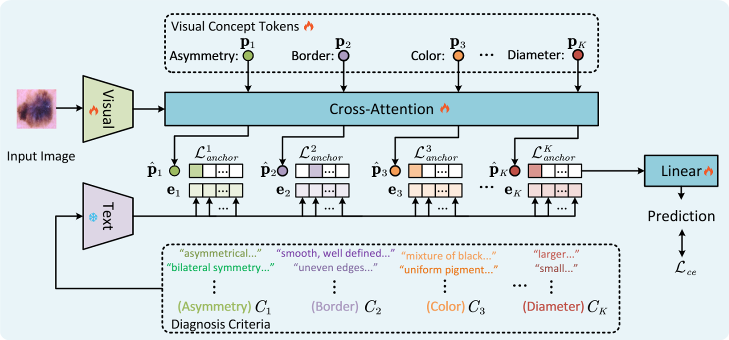
Interpretable classification

Linear layer to make prediction

$$\hat{y} = W(sim(\hat{\boldsymbol{p}}_j, \boldsymbol{e}_j^l),\ l = 1, \ldots, k_j;\ j = 1, \ldots, K)^\intercal$$

Overall loss function

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \{\mathcal{L}_{ce}(\hat{y}_i, y_i) + \mathcal{L}_{anchor}(\hat{\boldsymbol{p}}(x_i), \boldsymbol{e}_1, \ldots, \boldsymbol{e}_K)\}$$



Comparison with other models

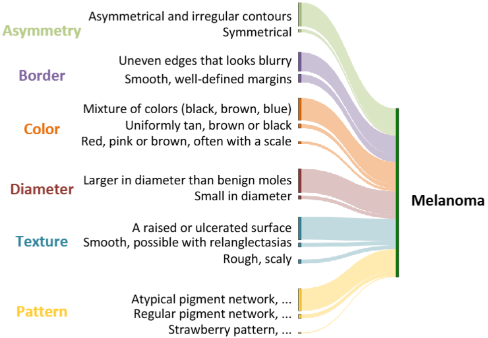
Setting	Model	ISIC2018	NCT	IDRiD	BUSI	CM	Edema
Zero-shot	CLIP	11.6	9.9	31.1	30.8	49.5	51.4
	BioViL	8.5	7.7	26.2	30.8	70.8	76.9
	BiomedCLIP	21.2	35.3	37.9	37.2	69.3	77.1
Black-box	ResNet50	82.6	93.4	53.4	84.6	79.7	77.4
	ViT-Base	89.0	94.4	57.3	88.5	79.2	80.9
Explainable	LaBo	80.9	90.2	48.4	75.8	73.5	74.2
	Explicd (ours)	**90.0**	**95.1**	**58.5**	**89.7**	**81.8**	**85.7**

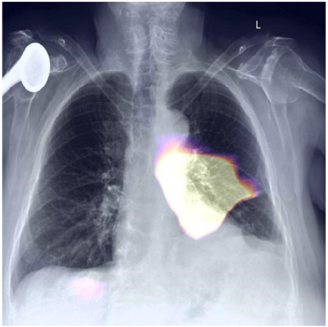
LaBo:[4]

1. Querying concepts for each class from LLMs.
2. Embedding each concept into $\mathbb{R}^d$.
3. Embedding the whole input image into $\mathbb{R}^d$.
4. In the embedding space, calculating the similarity of the input image to each concept.
5. Using a linear layer to make predictions based on these similarities.

[4] Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., & Yatskar, M. (2023). Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 19187-19197).

Interpretability





Discussions