Improved Explainability and Robustness of Foundation Models: Principled Methodology to Incorporate Domain Knowledge

Dimitris Metaxas

Board of Governors Professor

CBIM Center

Computer Science

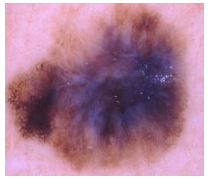
dnm@cs.rutgers.edu

Diagnosing Like a Human Expert: Data and Knowledge

RUTGERS–NEW BRUNSWICK
Artificial Intelligence
and Data Science Collaboratory

- **Use Domain Knowledge**
  - Specific data-based criteria (images, blood test, genomic data)
  - Other type of knowledge (types of disease, severity, outcomes, correlations between data and disease)
- **Grounding:** Explainable solution (concepts) based on data and domain knowledge
- **Alignment**: Decision-making (Human like given data and knowledge)

Human Knowledge



Eg. Melanoma

Color: mixture of pigmentation, including but not limited to black, brown, red, white, and blue, indicative of varying depths of melanin deposition and potential regression areas.

Shape: lack of uniformity, often presenting with a non-symmetrical outline indicative of its aggressive growth nature.

Border: uneven edge and might look a bit fuzzy, making it hard to tell where the spot ends and normal skin begins.

Pattern: heterogeneity in pigmentation with multiple colors, alongside irregular streaks, dots, and globules. An atypical pigment network and the presence of a blue-white veil are indicative of malignancy.

Texture: heterogeneous surfaces, often presenting as a combination of smooth, scaly, and irregularly eroded areas, indicative of rapid and atypical cellular proliferation with potential areas of regression.

Our Method

Explicd: Explainable language-informed criteria-based diagnosis

Domain Knowledge Query & Diagnostic Criteria Formulation

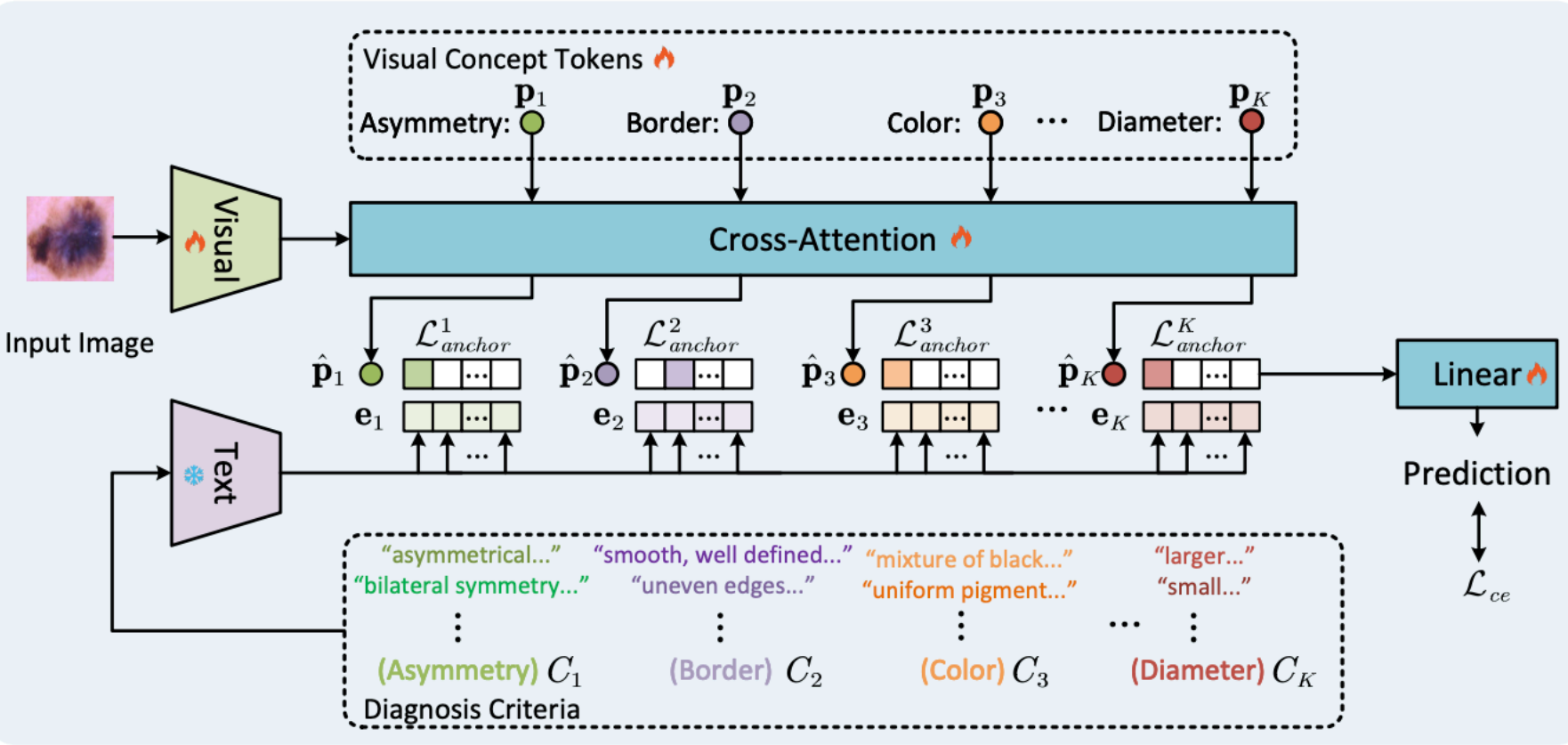
Prompt: describe the clinical criteria to diagnose skin lesion from dermoscopic images

LLM/Human Experts: the criteria, encapsulated within the ABCDE rule, help to identify skin lesion types, including asymmetry, border, color, diameter, evolving.

Prompt: please describe the typical features for the ABCDE rule of each class

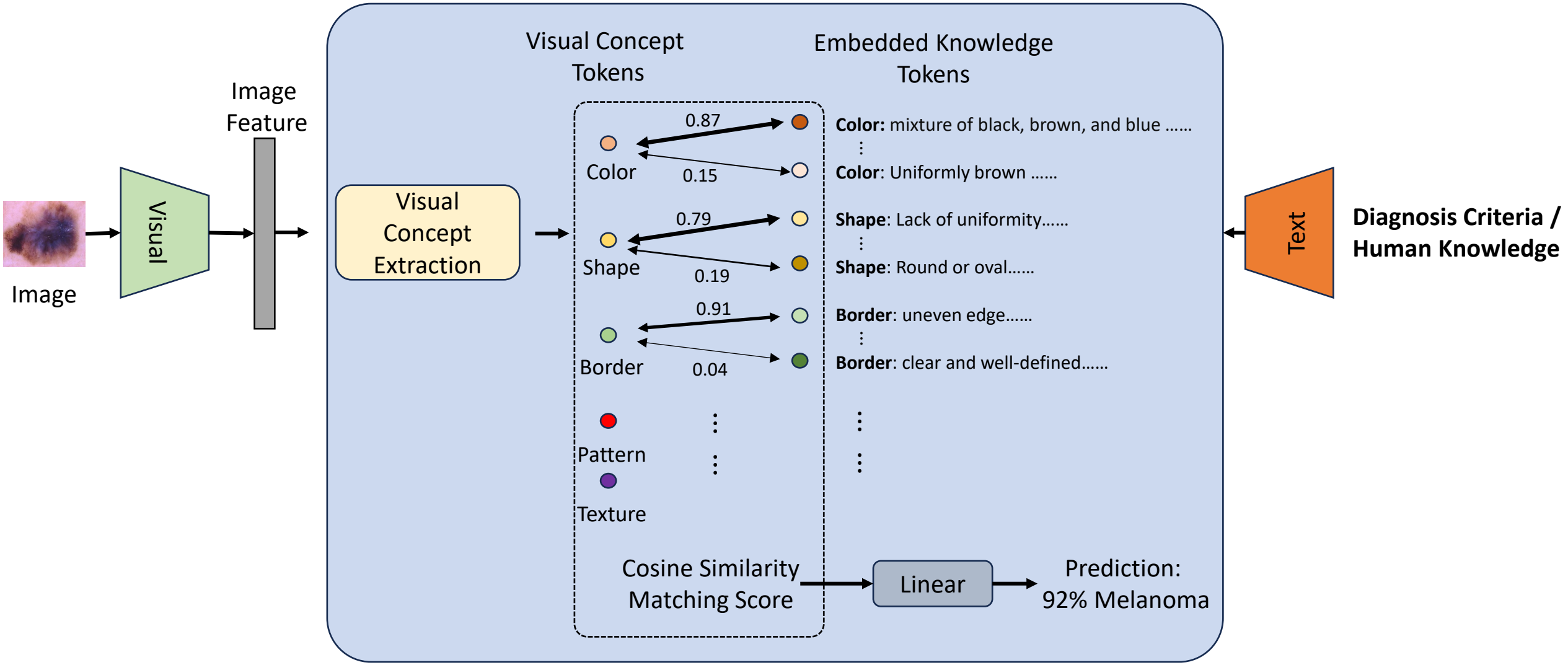
LLM/Human Experts: melanoma exhibits `asymmetry`: ...; `border`: ...; `color`: ...; `diameter`: ...; melanocytic nevus shows: ...; basal cell carcinoma ……

Gather and catalog these criteria of each class as knowledge anchors



Example of Interpretable Diagnosis (Melanoma): Fine-grained Aligned Vision-Language Model

RUTGERS–NEW BRUNSWICK
Artificial Intelligence
and Data Science Collaboratory



Results

- Dataset
  - **ISIC2018**: **Dermoscopic** skin lesion images, 7 disease states classification
  - **NCT**: **Histological images**. 9 tissue classes classification
  - **IDRiD**: **Diabetic retinopathy** (DR) images, 5 DR severity level classification
  - **BUSI**: **Breast Ultrasound images**, 3 breast cancer stage classification: normal/benign/malignant
  - **MIMIC-CXR**: **Chest X-Ray images**, cardiomegaly (CM) and Edema classification

- Results
  - Our method significantly outperforms general/specific VLM
  - Our method even outperforms supervised learning black-box models
  - Our method is interpretable

Table 1: Performance comparison across five benchmarks. Balanced accuracy is reported for CM and edema in MIMIC-CXR due to class imbalance; accuracy is reported for the other datasets.

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**

Future AI

- AI++: Go beyond Data
- AI for Science is evolving where domain knowledge needs to be used in decision making
- AI meets Humans to go beyond human capabilities
- Can we discover new knowledge?