A Concept-Based Explainability Framework for Large Multimodal Models

Jayneel Parekh

ISIR, Sorbonne Université

Collaborators: Pegah Khayatan, Mustafa Shukor, Alasdair Newson, Matthieu Cord

NeurIPS 2024

Multimodal learning tasks and models

- Default human perception is multimodal
- Often aligned or correlated modalities $\Longrightarrow$ Multimodal learning tasks
- Focus on Large Multimodal Models (LMMs) processing visual data (eg. Flamingo, LLaVA)
- Popular for solving captioning, question-answering, reasoning tasks

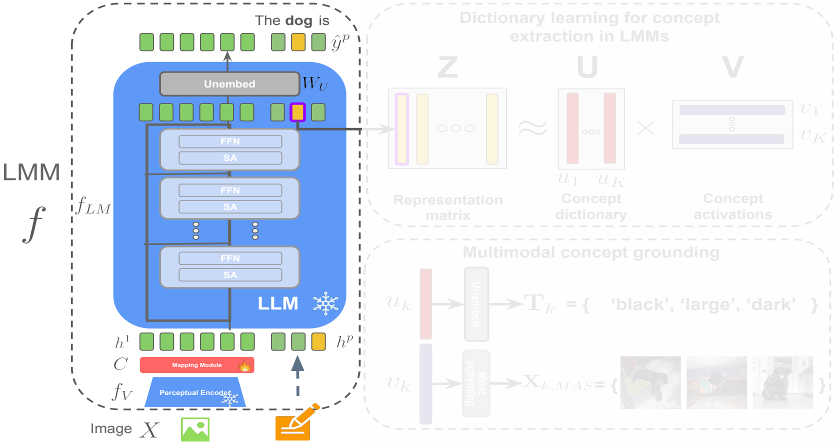
What is our aim?

- Pretrained LMM f = Visual encoder (f_V) + Connector (C) + Language model (f_{LM})
- Captioning dataset $\mathcal{S} = \{(X_i, y_i)\}_{i=1}^N$. Images $X_i \in \mathcal{X}$ and captions $y_i \subset \mathcal{Y}$
- A token of interest $t \in \mathcal{Y}$ (Eg. 'Dog', 'Cat' etc.)

Aim: Understand internal representations of f about t in terms of high-level concepts

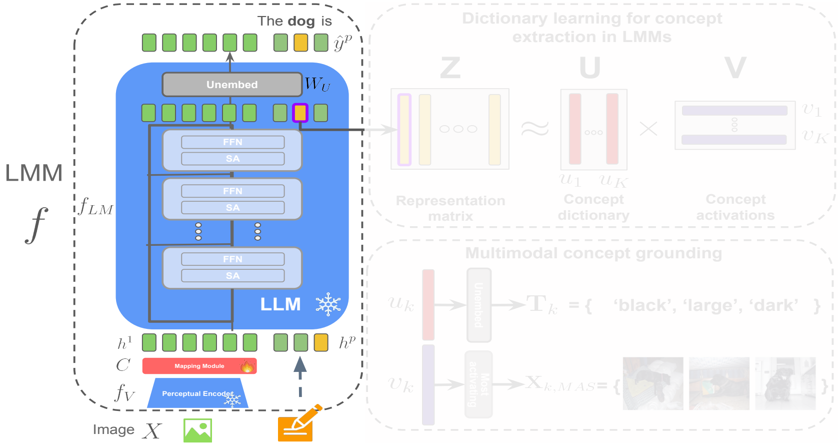
Contribution: A Concept based eXplainability framework for LMMs (CoX-LMM)

Design of CoX-LMM



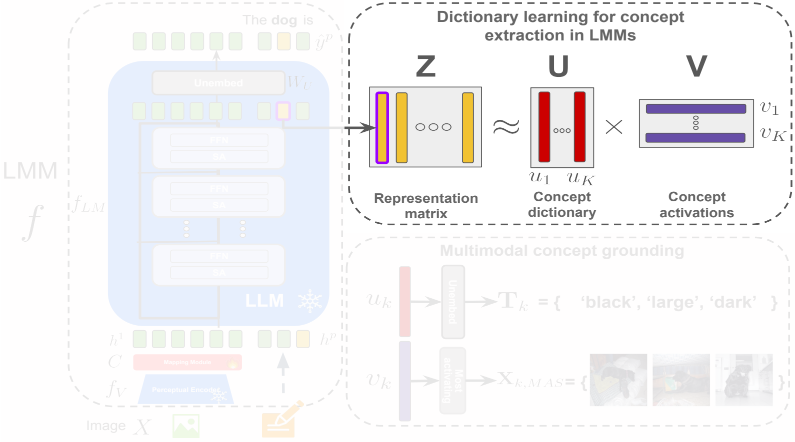
- Input to f_{LM} - Concatenated sequence of tokens: (1) Visual tokens $C(f_V(X))$, (2) textual tokens previously predicted by f_{LM}
- Caption predicted by f_{LM} trained for next-token prediction task

Design of CoX-LMM



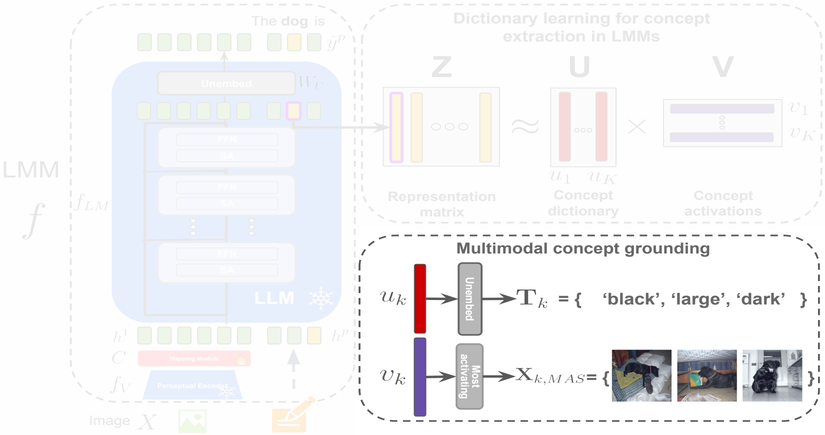
- Extract residual stream representations of t from f for a relevant set of images $\mathbf{X}$
- Collect all such B-dimensional representations as columns of matrix $\mathbf{Z} \in \mathbb{R}^{B \times M}$

Design of CoX-LMM



- Dictionary learning for concept extraction. Semi-NMF optimization: $\mathbf{U}^*, \mathbf{V}^* = \arg\min_{\mathbf{U},\mathbf{V}} \; ||\mathbf{Z} - \mathbf{U}\mathbf{V}||_F^2 + \lambda||\mathbf{V}||_1 \quad s.t.\ \mathbf{V} \geq 0,$ and $||u_k||_2 \leq 1 \; \forall k \in \{1, ..., K\}$
- Columns of $\mathbf{U}^* \in \mathbb{R}^{B \times K}$ – concept vectors. Rows of $\mathbf{V}^* \in \mathbb{R}^{K \times M}$ – concept activations

Design of CoX-LMM: Multimodal concept grounding!



- **Text grounding**: Decode concept vector u_k with f_{LM} head and extract top tokens
- **Visual grounding**: Extract most activating samples for u_k (via activations v_k)

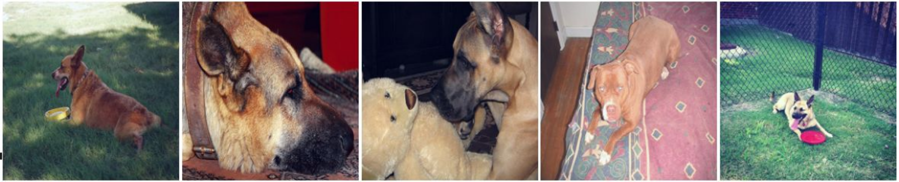
Example multimodal concepts

- Different types of ‘dogs’

Example multimodal concepts

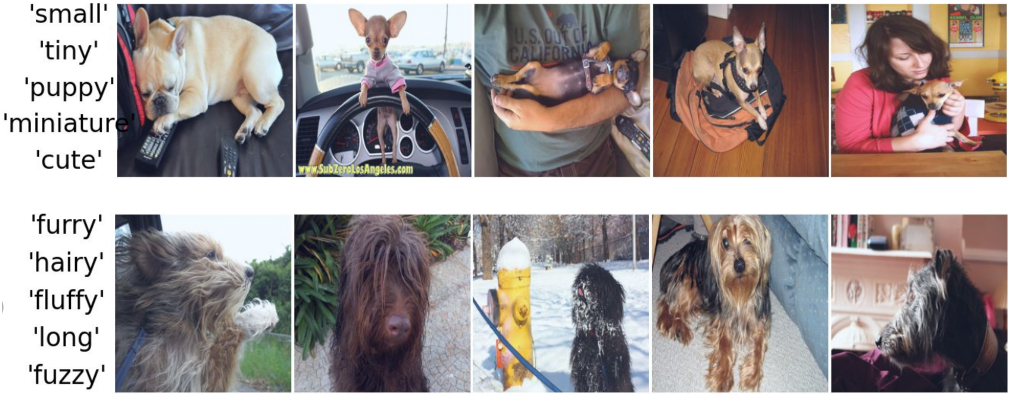
'black'
'large'
'dark'
'big'
'close'

'brown'
'large'
'dog'
'tan'
'golden'



- Different types of ‘dogs’
- Concepts about color

Example multimodal concepts



- Different types of ‘dogs’
- Concepts about color, characteristics

Example multimodal concepts

- Different types of ‘dogs’
- Concepts about color, characteristics, state

Example multimodal concepts

- Different types of ‘dogs’
- Concepts about color, characteristics, state, or related objects

The End

Thanks for the attention!

Project webpage: https://jayneelparekh.github.io/LMM_Concept_Explainability/

Code: https://github.com/mshukor/xl-vlms