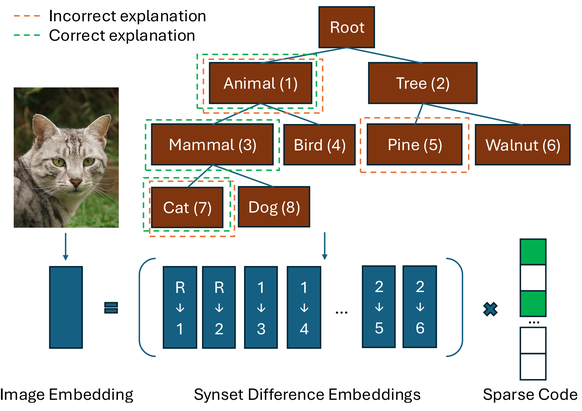
Hierarchical Concept Embedding & Pursuit for Interpretable Image Classification

Nghia Nguyen Tianjiao Ding René Vidal

University of Pennsylvania

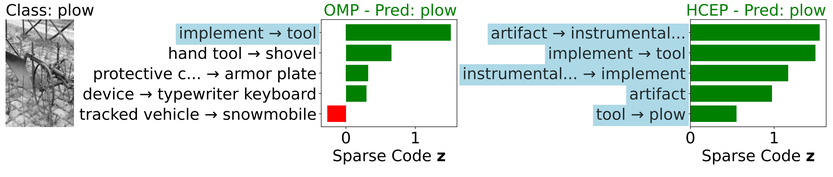
Motivation: Interpretable Image Classification

- Interpretable-by-design models extract **concepts** from images and classify the image based on these concepts.
- The distinguishing features of an image form a small set of concepts, so **sparse coding** is a natural tool for concept extraction.
- Sparse coding methods represent an image embedding as a sparse combination of concept embeddings.
- Hierarchy provides a natural structure to organize semantic concepts for classification.
- **Problem:** Existing methods ignore the **hierarchical structure** of semantic concepts, leading to **hierarchy-violating** explanations.



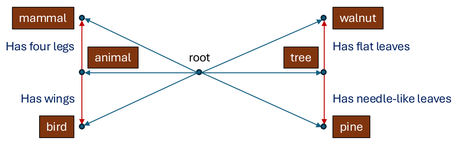
Contributions

- **Hierarchical Concept Embedding:** Geometric framework with identifiability guarantees for hierarchical concept representations.
- **Hierarchical Concept Pursuit:** Hierarchical dictionary construction & HB-OMP algorithm that recovers rooted paths in the hierarchy.
- **Experiments:** Best trade-off b/w classification accuracy and concept support precision for interpretable image classification.



Hierarchical Concept Embedding

Two **ideal geometric properties** for hierarchical embeddings:



- **Well-clustered synsets:** Descendants near parents; sibling subtrees are disjoint.
- **Hierarchical orthogonality:** $(\boldsymbol{a}^{(j)} - \boldsymbol{a}^{(i)})^\top \boldsymbol{a}^{(i)} = 0$ for child j of i.

Hierarchical Concept Pursuit

Synset Embedding Construction:

- **Leaf nodes** (e.g., `cat`, `dog`): $\boldsymbol{a}^{(i)}$ = avg CLIP image embeddings in class i.
- **Internal nodes** (e.g., `mammal`): $\boldsymbol{a}^{(i)}$ = avg of children embeddings.

Hierarchical Dictionary:

- $\boldsymbol{D} = \left[\, \boldsymbol{a}^{(j)} - \boldsymbol{a}^{(\mathsf{par}(j))} \,\right]_{j \in \mathcal{A}}$, where $\mathcal{A}$ is the set of synsets.
- Each atom captures the **difference between a synset and its parent**. E.g., the atom for `cat` captures the difference from `animal`.
- A synset embedding decomposes as a sum along a rooted path:

$$\boldsymbol{a}^{(i)} = \sum_{j \in \mathsf{anc}(i) \cup \{i\}} \left(\boldsymbol{a}^{(j)} - \boldsymbol{a}^{(\mathsf{par}(j))} \right)$$

Hierarchical Beam OMP (HB-OMP):

- Given image embedding $\boldsymbol{x}$ and concept dictionary $\boldsymbol{D}$, find sparse $\boldsymbol{z}$:

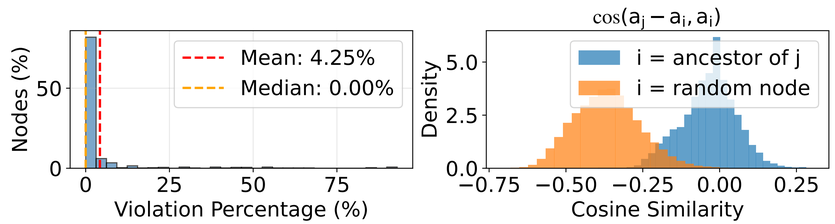
$$\min_{\boldsymbol{z}} \| \underbrace{\boldsymbol{x} - \boldsymbol{D}\boldsymbol{z}}_{\text{residual}} \|_2^2 + \lambda \|\boldsymbol{z}\|_0$$

- Iteratively, Orthogonal Matching Pursuit (OMP) selects the atom **most correlated with the current residual** *and* updates the residual by projecting out the selected atoms.
- We modify OMP so that the next atom is restricted to children of the last selected node.
- We also maintain B hypotheses with the smallest residual norm.

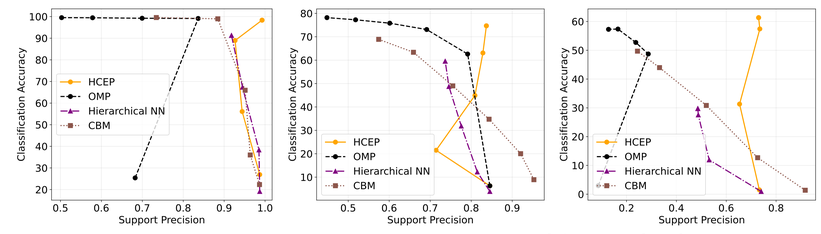
Theoretical Justification

Proposition: If HB-OMP extends a hypothesis that is a prefix of a true path, it is **less likely to introduce an incorrect atom** than OMP would.

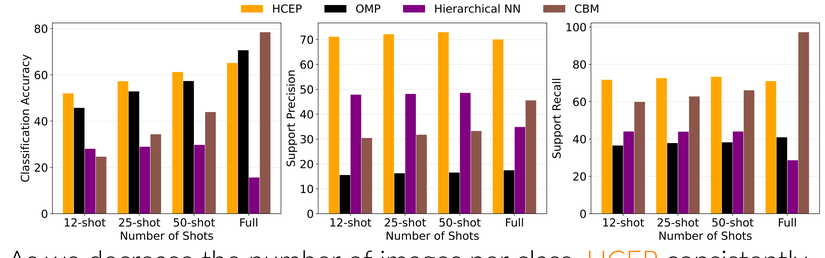
Experiments



CLIP embeddings empirically satisfy both geometric conditions: **well-clustered synsets** (left) and **hierarchical orthogonality** (right).



ImageNette, CIFAR-100, ImageNet: HCEP **achieves the best trade-off between classification accuracy and support precision** over interpretable baselines across sparsity levels.



As we decrease the number of images per class, HCEP consistently **outperforms in classification accuracy, support precision, and support recall**.