Geometry of Categorical & Hierarchical Concepts in LLMs

Kiho Park, Yo Joong Choe, Yibo Jiang, Victor Veitch

University of Chicago

Problems

How are categorical concepts represented?

How are hierarchical relations between concepts represented?

Goal: extend the linear representation hypothesis to answer these problems

Concepts in LLMs

LLMs generate the next word using the softmax distribution.
We use the unified space induced by the causal inner product.

Embedding	Softmax	Unembedding
$l(x) \in \mathbb{R}^d$	$\mathbb{P}(y \mid x) \propto \exp(l(x)^\top g(y))$	$g(y) \in \mathbb{R}^d$

Context: "He is the" → X → W, Z (Concepts) → Y → Next word: "king", "man", "PhD", ⋮

Categorical Concepts {Mammal, Bird, Fish, Reptile}

Binary Concepts

- **Binary Feature**: {not_animal, is_animal}
- **Binary Contrast**: Plant ⇒ Animal

Hierarchical Structure: z is subordinate to w (denoted by $z \prec w$) if $\mathcal{Y}(z) \subseteq \mathcal{Y}(w)$. We say that Z is subordinate to W if there exists a value w_Z of W such that each value z_i of Z is subordinate to w_Z.

Linear Representation of Binary Concept

A vector $\bar{l}_W$ is a linear representation of a binary concept W if

$$\mathbb{P}(W = 1 \mid l + \alpha \bar{l}_W) > \mathbb{P}(W = 1 \mid l)$$

$$\mathbb{P}(Z \mid l + \alpha \bar{l}_W) = \mathbb{P}(Z \mid l)$$

for all contexts $l, \alpha > 0, Z$ subordinate to or causally separable with W.

Vector Representation of Binary Feature (Theorem 4)

If there exists a linear representation of a binary feature W for a value w, then, with a canonical origin, there exists a constant $b_w > 0$ such that

$$\bar{l}_W^\top g(y) = \begin{cases} b_w & \text{if } y \in \mathcal{Y}(w) \\ 0 & \text{if } y \notin \mathcal{Y}(w) \end{cases}$$

Now the linear representation has a magnitude, and we define a *vector* representation $\bar{l}_w$ of a binary feature for a value w where $\|\bar{l}_w\|_2 = b_w$.

Hierarchical Orthogonality (Theorem 6)

(a) $\bar{l}_{w_1} - \bar{l}_{w_0}$ is a linear representation $\bar{l}_{w_0 \Rightarrow w_1}$

(b) $\bar{l}_w \perp \bar{l}_z - \bar{l}_w$ for $z \prec w$

(c) $\bar{l}_w \perp \bar{l}_{z_1} - \bar{l}_{z_0}$ for $Z \in_R \{z_0, z_1\} \prec W \in_R \{\text{not\_w}, \text{is\_w}\}$

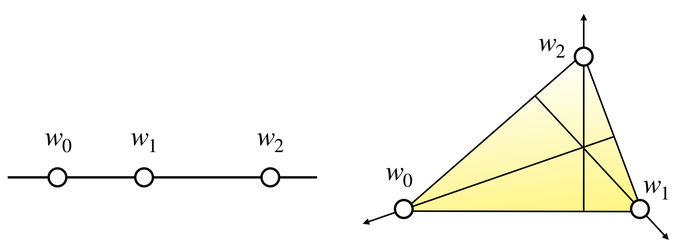
(d) $\bar{l}_{w_1} - \bar{l}_{w_0} \perp \bar{l}_{z_1} - \bar{l}_{z_0}$ for $Z \in_R \{z_0, z_1\} \prec W \in_R \{w_0, w_1\}$

(e) $\bar{l}_{w_1} - \bar{l}_{w_0} \perp \bar{l}_{w_2} - \bar{l}_{w_1}$ for $w_2 \prec w_1 \prec w_0$

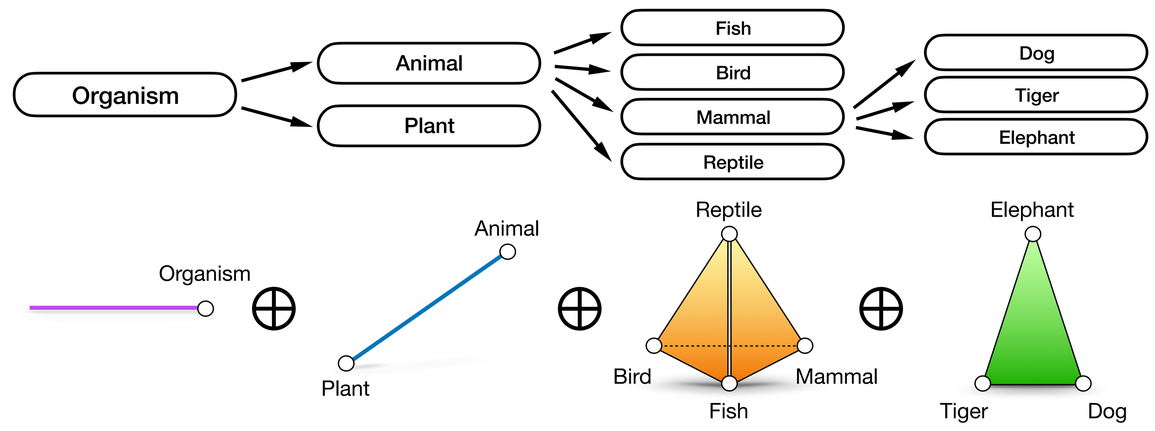
Categorical Concepts as Simplices (Theorem 8)

The *polytope* representation of a categorical concept Z is the convex hull of the vector representations for the elements of the concept.

For every joint distribution $Q(w_0, \ldots, w_{k-1})$, if there exists some l_i such that $\mathbb{P}(W = w_i \mid l_i) = Q(W = w_i)$ for every i, the vector representations $\bar{l}_{w_0}, \ldots, \bar{l}_{w_{k-1}}$ form a $(k-1)$-simplex in the representation space. In this case, we take the *simplex* to be the representation of the categorical concept $W = \{w_0, \ldots, w_{k-1}\}$.

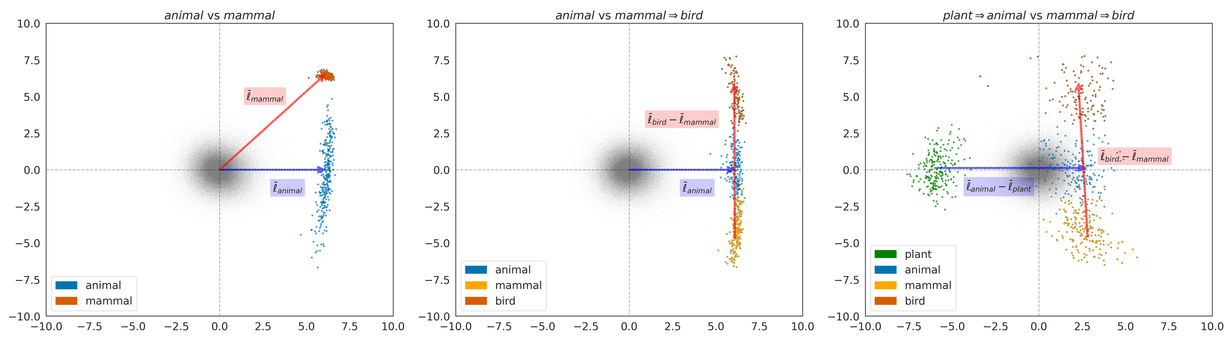


Results



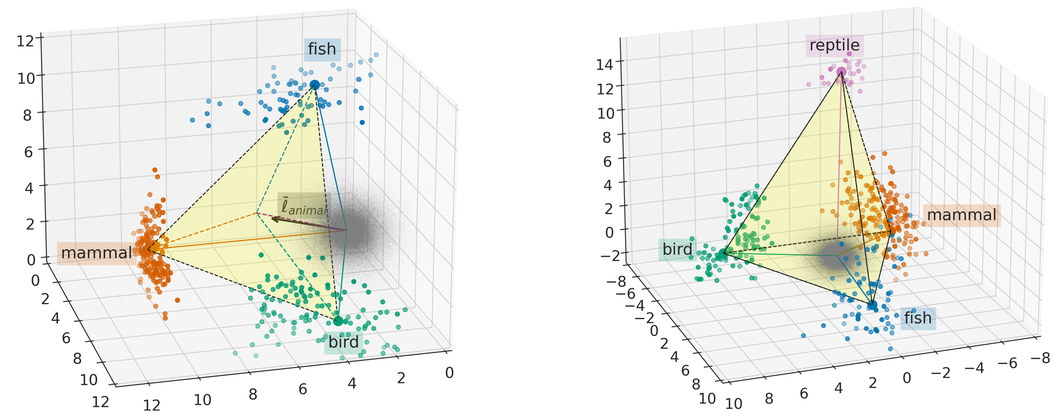
Pictorial depiction of the representation of hierarchically related concepts

Hierarchical Semantics are Encoded as Orthogonality



The projection of the unembedding vectors on the 2D subspaces: $\text{span}\{\bar{l}_{animal}, \bar{l}_{mammal}\}$ (left, Thm 6 (b)), $\text{span}\{\bar{l}_{animal}, \bar{l}_{bird} - \bar{l}_{mammal}\}$ (middle, Thm 6 (c)), $\text{span}\{\bar{l}_{animal} - \bar{l}_{plant}, \bar{l}_{bird} - \bar{l}_{mammal}\}$ (right, Thm 6 (d))

Categorical Concepts are Represented as Simplices

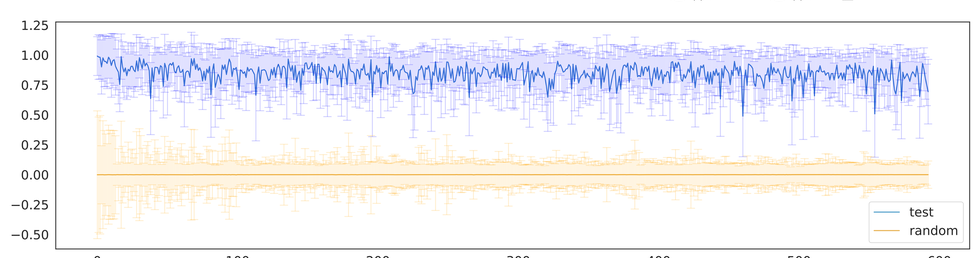


The projection of the unembedding vectors on the 3D subspaces: $\text{span}\{\bar{l}_{mammal}, \bar{l}_{bird}, \bar{l}_{fish}\}$ (left, $\bar{l}_{animal}$ is orthogonal to the simplex!), $\text{span}\{\bar{l}_{bird} - \bar{l}_{mammal}, \bar{l}_{fish} - \bar{l}_{mammal}, \bar{l}_{reptile} - \bar{l}_{mammal}\}$ (right)

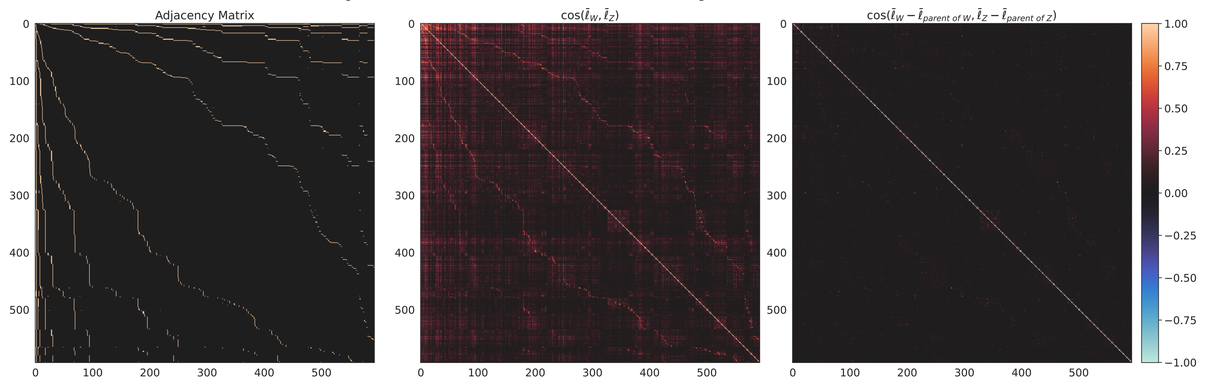
Further Experiments and Details

Using Gemma-2B model, we estimate the vector representation $\bar{l}_w$ of a binary feature for a value w using the collection of the tokens that have w:

$$\bar{l}_w = (\tilde{g}_w^\top \mathbb{E}(g_w))\tilde{g}_w, \text{ with } \tilde{g}_w = \frac{\text{Cov}(g_w)^\dagger \mathbb{E}(g_w)}{\|\text{Cov}(g_w)^\dagger \mathbb{E}(g_w)\|_2}$$



The projection of test and random words onto the estimated $\bar{l}_w$ shows that the vector representations for binary features in WordNet exist.



The adjacency matrix of WordNet hierarchy (left) is clearly visible in the middle heatmap. By contrast, consistent with Thm 6 (e), the child-parent and parent-grandparent vectors are orthogonal in the right heatmap.

Find us online! Email: parkkiho@uchicago.edu


Kiho Park


Yo Joong Choe


Yibo Jiang


Victor Veitch

arXiv:2406.01506