Paper and codes:

PHyCLIP: ℓ_1-Product of Hyperbolic Factors Unifies Hierarchy and Compositionality in Vision-Language Representation Learning

Daiki Yoshikawa & Takashi Matsubara

HOKKAIDO UNIVERSITY

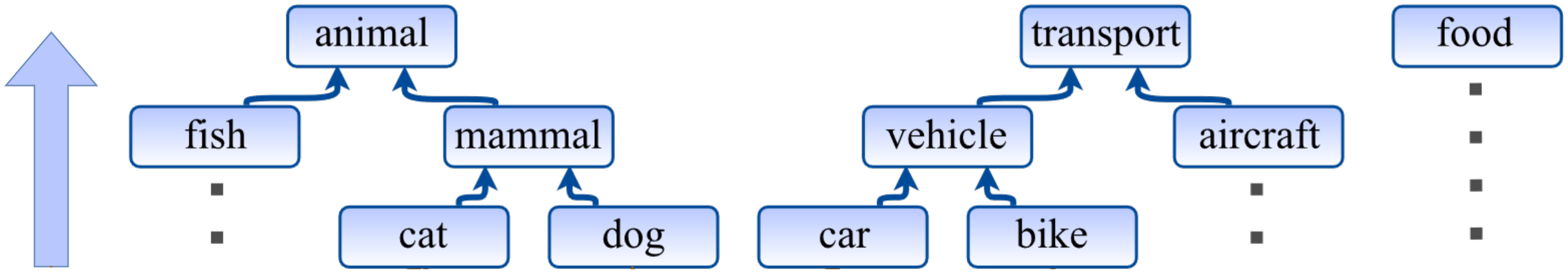
CyberAgent AI Lab

Background

Theorem 1 (Sarkar, 2011):

- *A metric tree is quasi-isometrically embedded into a 2D hyperbolic space.*
  - A hyperbolic space is effective to capture a taxonomic hierarchy of concepts. (Nickel & Kiela, 2017)
  - Non-hyperbolic spaces cannot capture hierarchies effectively with few dimensions.

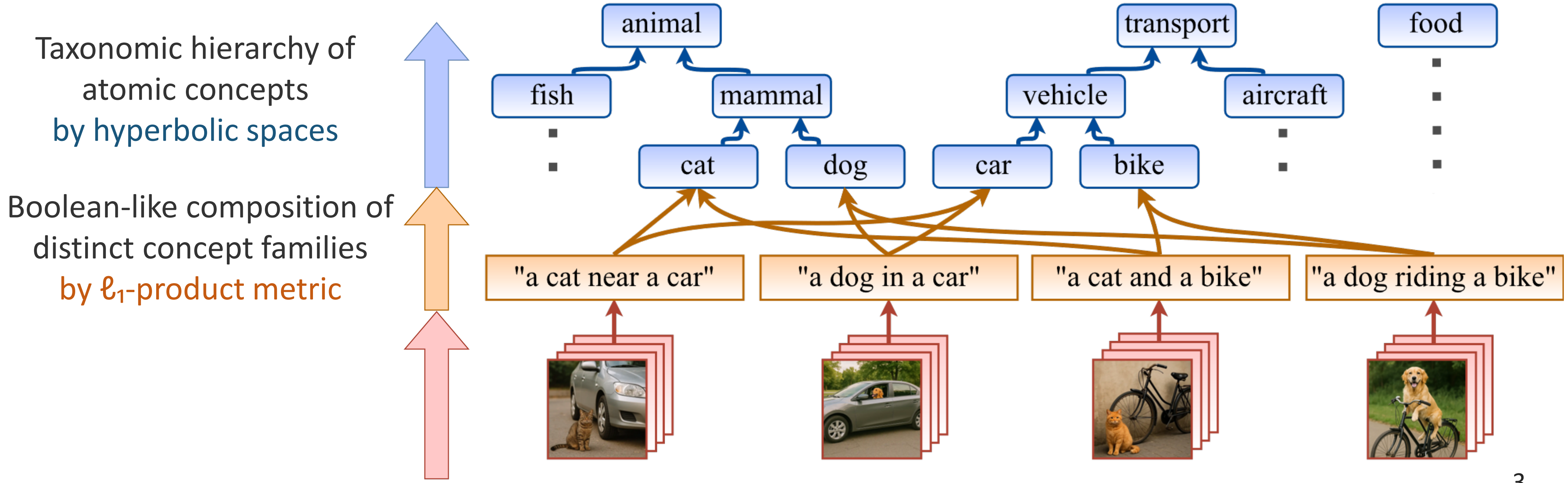
Taxonomic hierarchy of
atomic concepts
by hyperbolic spaces



Background

Proposition 1:

- Additive operations are effective to capture the composition of concepts. (e.g., Boolean algebra, bag-of-words, and vector addition)
  - A hyperbolic space struggles to capture the composition, as it lacks such operation.
  - A Boolean algebra with the Hamming distance is isometrically embedded into an ℓ_1-product metric space, but *NOT* into a hyperbolic space.



Background

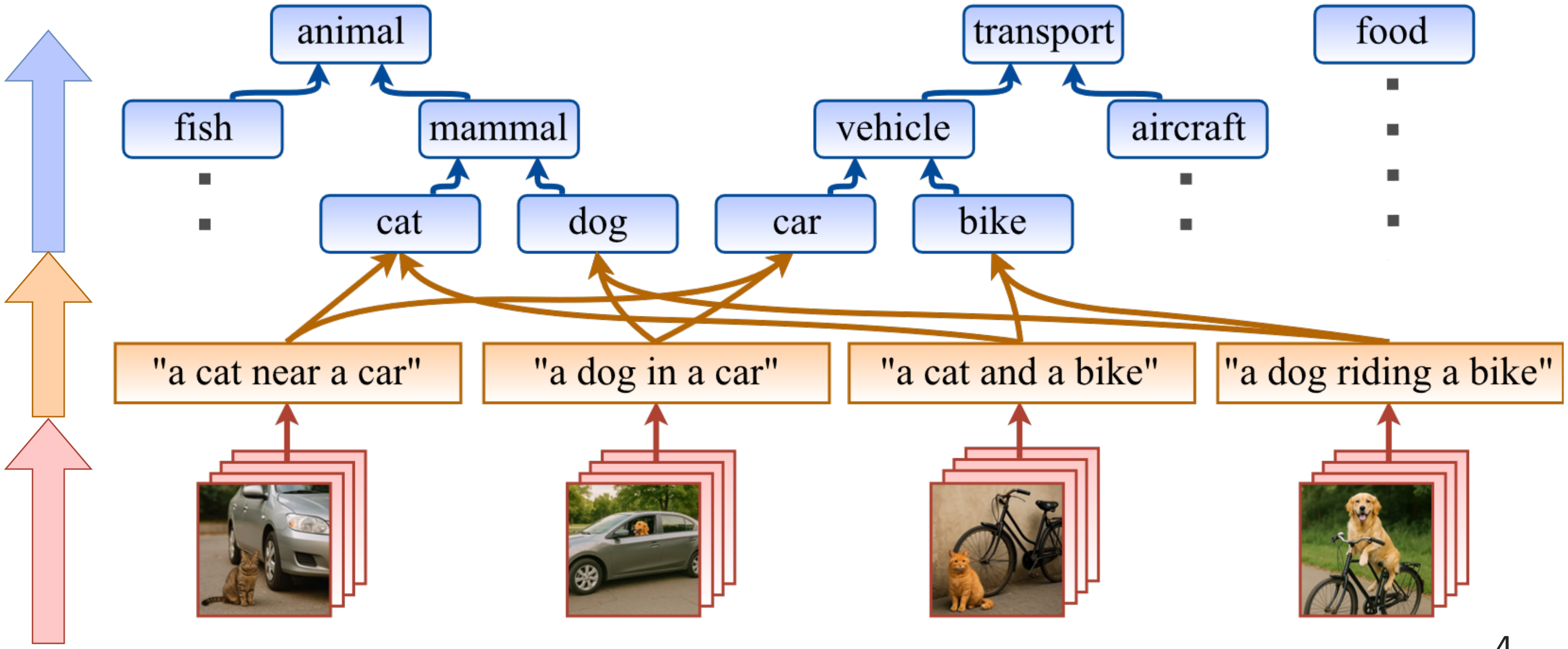
Images and texts have two aspects:

- Tree-like taxonomic hierarchy, embedded into a hyperbolic space.
- Boolean-like compositionality, captured by an ℓ_1-product metric.

How can we enjoy the best of both worlds?

Taxonomic hierarchy of atomic concepts by hyperbolic spaces

Boolean-like composition of distinct concept families by ℓ_1-product metric

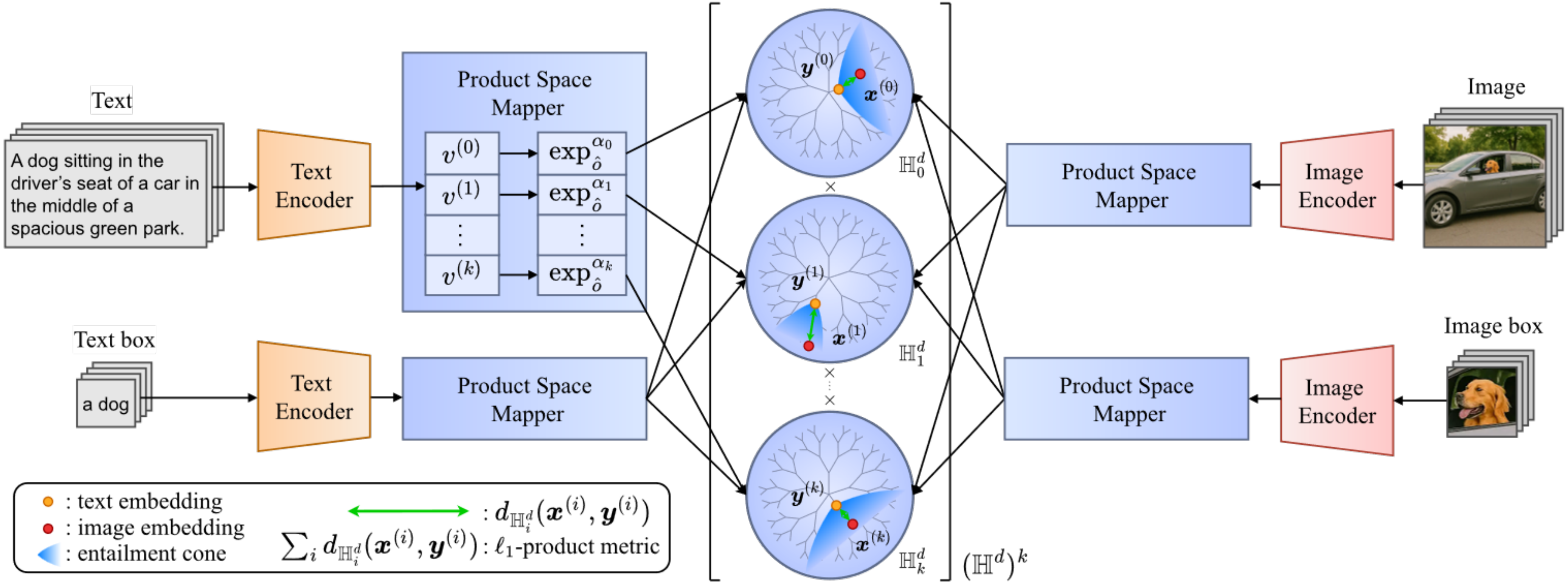


PHyCLIP: ℓ_1-Product of Hyperbolic Factors

Product-of-Hyperbolic (PHy) embedding

- Representation learning that embeds instances into $((\mathbb{H}^d)^k, d_1)$
  - $(\mathbb{H}^d)^k$: a Cartesian product of k hyperbolic factors $\mathbb{H}^d$
  - d_1: an ℓ_1-product metric $d_1(\boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{k} d_{\mathbb{H}_i^d}(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})$

    for embeddings $\boldsymbol{X} = (\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k)})$ and $\boldsymbol{Y} = (\boldsymbol{y}^{(1)}, \dots, \boldsymbol{y}^{(k)})$



Experiments and Results

PHyCLIP

- Trained on GRIT (Peng et al., 2023) with the contrastive & entailment losses.
- Better at hierarchical classifications and object compositions, but worse at object relations.
  - because of its Boolean-like behavior

	w/ boxes	General datasets: ImageNet	CIFAR-10	CIFAR-100	SUN397	Caltech-101	STL-10	Fine-grained datasets: Food-101	CUB	Cars	Aircraft	Pets	Flowers	Specialized datasets: DTD	EuroSAT	RESISC45	Country211
CLIP		38.87	76.26	48.19	50.70	73.62	93.03	51.19	12.90	7.82	3.01	45.89	21.16	22.02	35.73	42.03	5.13
CLIP	✓	38.81	76.53	48.59	50.80	74.29	93.34	51.05	12.70	8.40	2.89	46.19	**21.32**	21.74	37.49	41.78	5.10
MERU		37.96	77.63	46.37	49.39	72.10	93.14	51.67	11.09	7.80	3.53	43.36	19.98	22.18	**38.81**	41.77	4.86
MERU	✓	38.08	78.14	46.80	49.59	72.69	93.28	51.92	10.70	7.77	3.53	43.22	18.31	22.07	37.31	41.73	5.01
HyCoCLIP	✓	43.80	89.00	58.59	54.49	76.14	**94.96**	52.64	14.90	10.24	**3.57**	53.33	19.41	**25.90**	36.36	46.97	**5.64**
PHyCLIP	✓	**44.31**	**89.33**	**59.05**	**55.32**	**76.35**	94.84	**57.26**	**15.90**	**10.89**	3.24	**54.18**	19.98	25.50	36.29	**48.22**	5.56

	w/ boxes	VL-CheckList–Object: Location: Center	Location: Mid	Location: Margin	Size: Large	Size: Medium	Size: Small	SugarCrepe: Replace: Obj	Replace: Att	Replace: Rel	Swap: Obj	Swap: Att	Add: Obj	Add: Att	Overall
CLIP		61.9	60.3	60.4	63.9	60.2	58.2	89.37	79.95	69.54	60.54	66.02	80.39	73.36	77.72
CLIP	✓	61.9	59.3	60.8	63.7	60.8	58.1	89.69	80.33	69.49	**61.63**	**66.47**	80.62	73.55	77.97
MERU		61.3	59.0	59.0	64.0	57.7	56.1	89.10	80.50	69.44	60.82	65.32	80.47	74.90	77.81
MERU	✓	61.0	58.5	58.7	62.6	58.7	56.5	89.39	79.95	**69.65**	60.41	66.07	80.41	**75.34**	77.93
HyCoCLIP	✓	70.4	69.5	67.8	72.6	66.1	67.2	**91.38**	79.74	67.24	54.69	63.66	82.57	74.23	77.99
PHyCLIP	✓	**71.2**	**70.3**	**70.4**	**73.7**	**68.1**	**67.8**	91.06	**81.05**	66.36	57.41	65.87	**83.24**	73.80	**78.32**

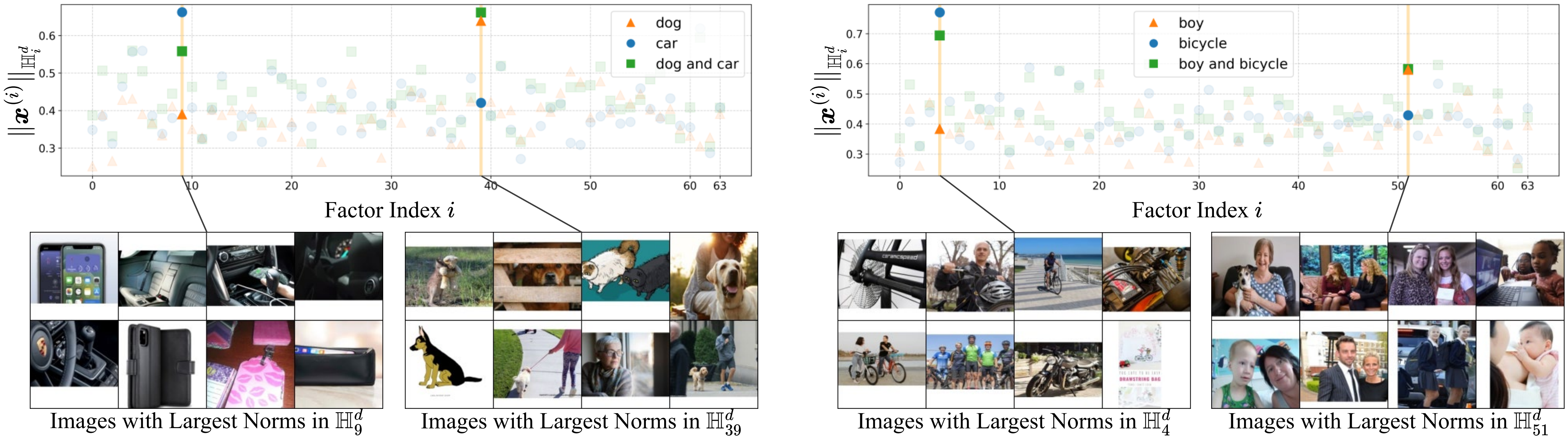
	w/ boxes	Text → Image: COCO R@5	COCO R@10	Flickr R@5	Flickr R@10	Image → Text: COCO R@5	COCO R@10	Flickr R@5	Flickr R@10	Hierarchical Classification (WordNet): TIE(↓)	LCA(↓)	J(↑)	P_H(↑)	R_H(↑)
CLIP		56.29	67.53	83.15	89.58	70.32	80.09	**91.60**	95.60	3.750	2.276	0.7774	0.8471	0.8483
CLIP	✓	56.20	67.50	82.75	89.42	70.35	80.19	91.10	95.63	3.736	2.279	0.7784	0.8473	0.8501
MERU		55.73	67.02	82.15	89.05	69.57	79.33	90.77	**95.83**	3.815	2.294	0.7733	0.8454	0.8450
MERU	✓	55.87	67.21	81.96	88.89	69.70	79.69	91.20	**95.83**	3.802	2.289	0.7740	0.8457	0.8455
HyCoCLIP	✓	57.11	68.32	83.06	89.63	69.51	79.73	91.47	95.63	3.319	2.092	0.8043	0.8676	0.8661
PHyCLIP	✓	**58.03**	**69.05**	**83.39**	**89.93**	**70.94**	**80.86**	91.20	95.53	**3.294**	**2.083**	**0.8059**	**0.8684**	**0.8672**

# of factors, k	# of dims., d	product metric	curvature	classification: ImageNet	classification: Food-101	retrieval COCO, R@5: Image	retrieval COCO, R@5: Text	hierarchical: TIE	hierarchical: J
1	512	–	hyp.	43.80	52.64	57.11	69.51	3.319	0.8043
8	64	ℓ_1	hyp.	44.38	54.61	57.80	70.80	3.273	0.8072
16	32	ℓ_1	hyp.	44.09	55.29	57.26	69.22	3.287	**0.8066**
32	16	ℓ_1	hyp.	43.90	54.48	56.70	66.92	3.324	0.8035
64	8	ℓ_1	hyp.	**44.31**	**57.26**	**58.03**	70.94	3.294	0.8059
128	4	ℓ_1	hyp.	44.16	53.96	57.79	**71.18**	**3.284**	0.8064
64	8	ℓ_2	hyp.	43.32	53.39	57.09	70.53	3.367	0.8011
64	8	ℓ_∞	hyp.	6.55	10.33	8.77	14.51	9.697	0.4247
-	-	ℓ_2	mixed	39.34	49.05	56.72	70.81	3.712	0.7797

Visualizations

Single-concept prompts activate certain hyperbolic factors

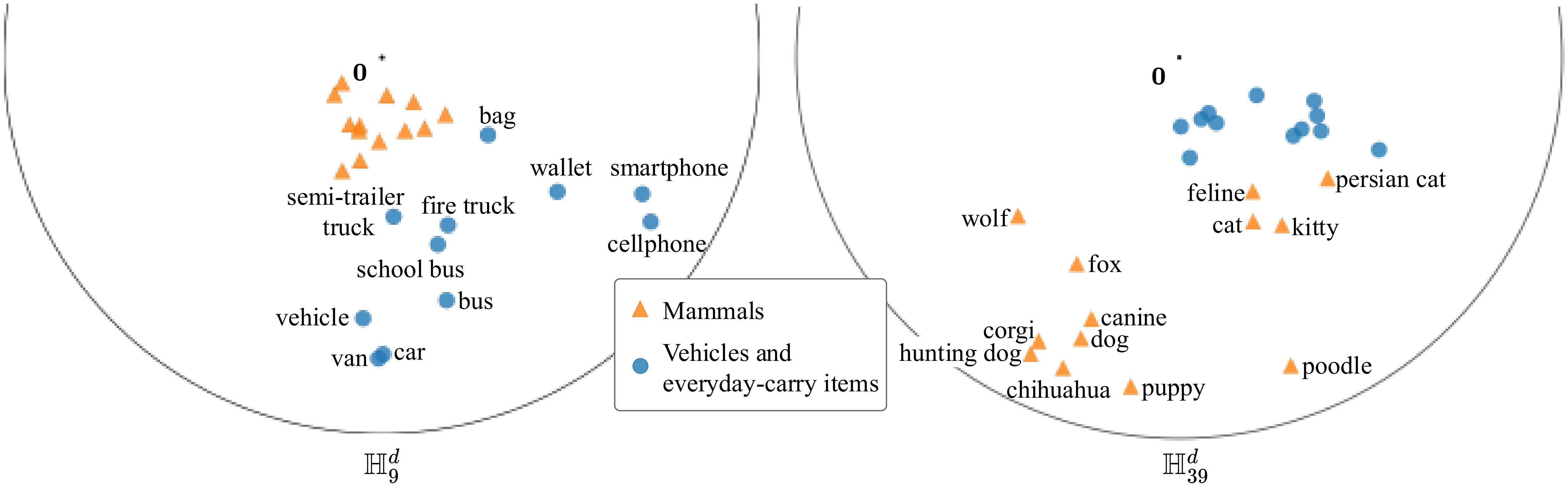
- We took factor-wise embeddings $\boldsymbol{x}^{(i)}$.
  - Prompt “a dog” leads to a large embedding $\boldsymbol{x}^{(39)}$ in $\mathbb{H}^d_{39}$, in which animal images also have large norms.
  - Prompt “a car” and images of vehicles and everyday-carry items do the same in $\mathbb{H}^d_9$.
  - Bikes and wheels in $\mathbb{H}^d_4$, humans in $\mathbb{H}^d_{51}$, etc.



Visualizations

A taxonomic tree of a concept family

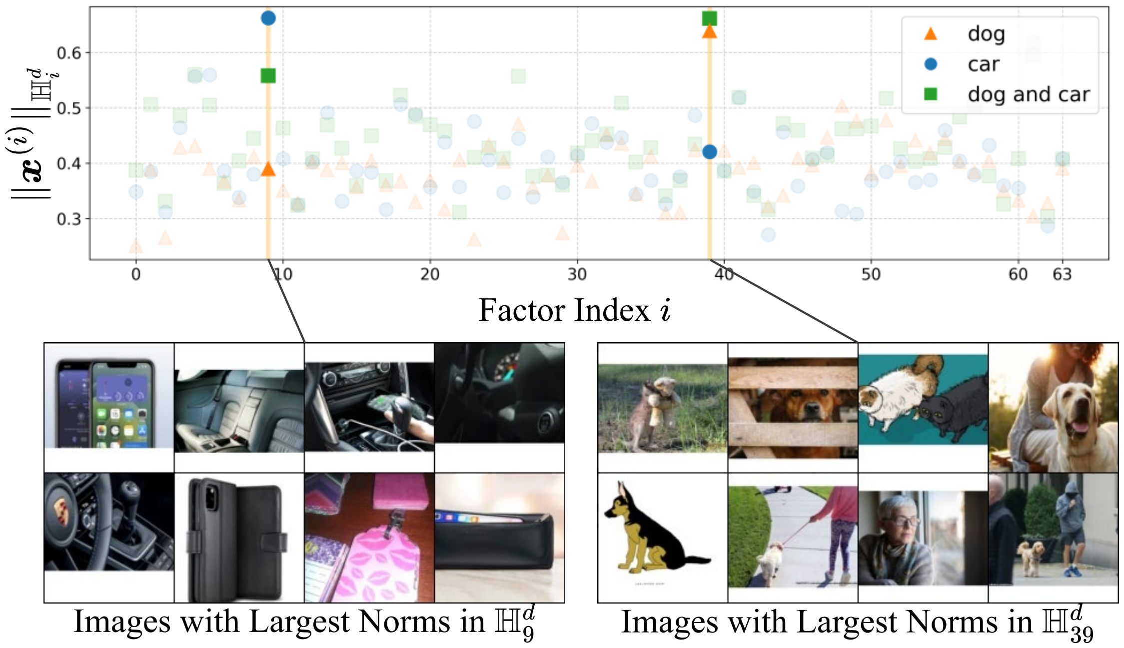
- emerges in each hyperbolic factor.
  - $\mathbb{H}_9^d$ is devoted for a taxonomy of vehicles and everyday-carry items
  - $\mathbb{H}_{39}^d$ is devoted for a taxonomy of mammals (especially, Carnivora)

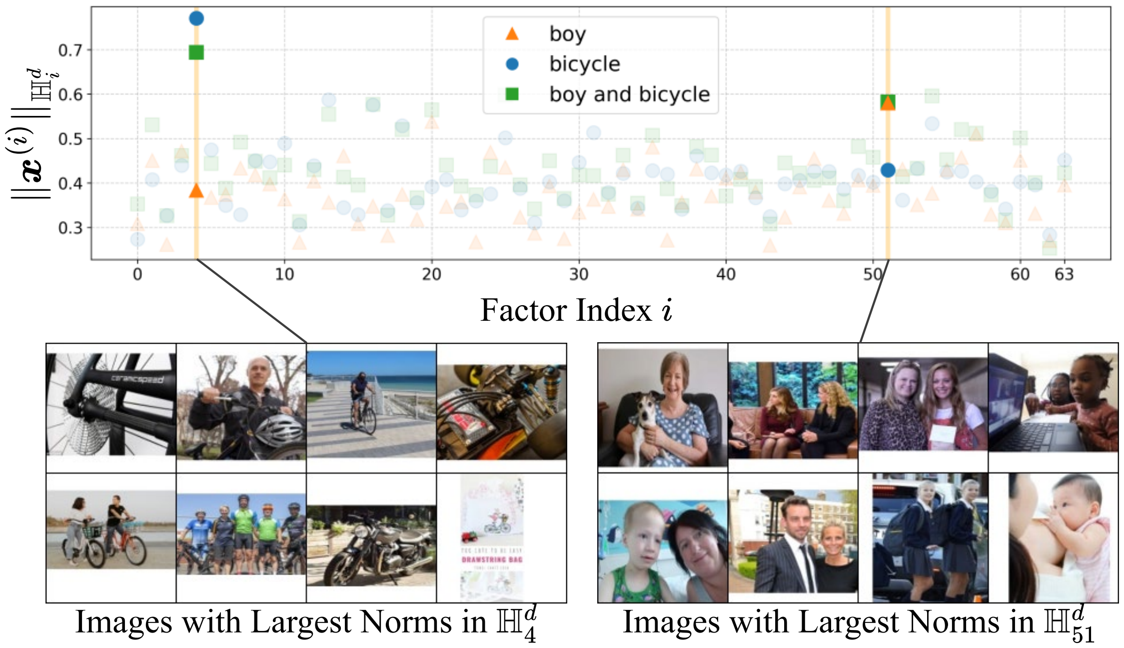


Visualizations

The ℓ_1-product metric captures the composition.

- We made the textual composition like "a dog and a car."
  - This prompt activates the *both* factors devoted for "a dog" and "a car".
  - Its behavior is similar to the Boolean algebra,
    as the conjunction corresponds to the max operation for Boolean bits.





Visualizations

The ℓ_1-product metric captures the composition.

- The factor-wise "max" of two single-concept prompts retrieves images similar to the textual compositions.

"a dog and a car"

Factor-wise *max* of
"a dog" and "a car"

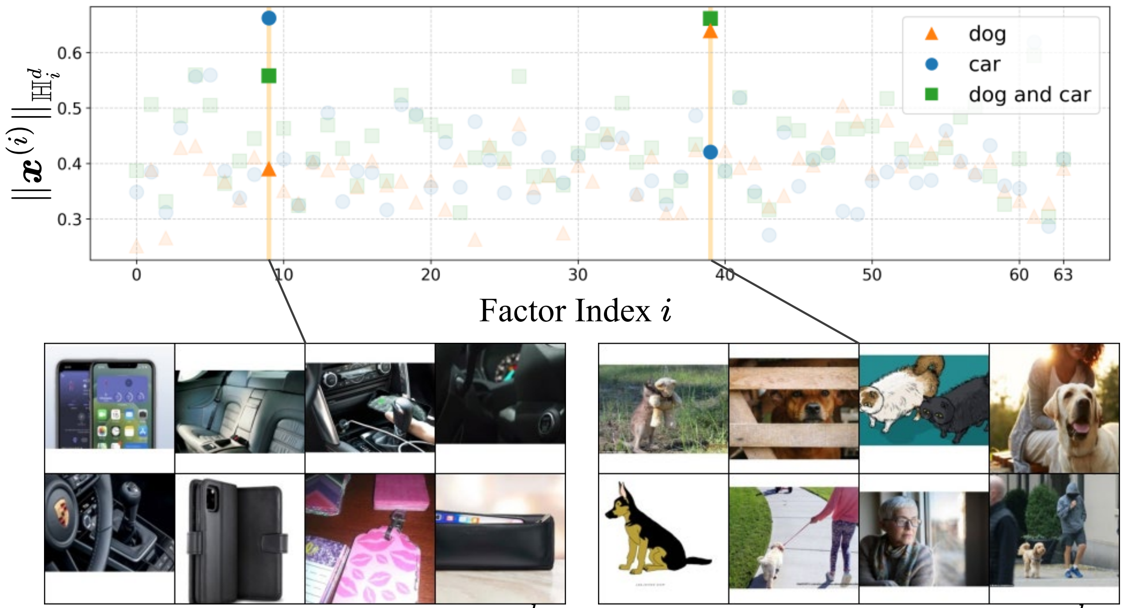
"a boy and a bicycle"

Factor-wise *max* of
"a boy" and "a bicycle"

Conclusion

PHyCLIP

- embeds instances into $((\mathbb{H}^d)^k, d_1)$
  - captures intra-family taxonomic hierarchies by hyperbolic factors $\mathbb{H}_i^d$
  - captures cross-family Boolean-like compositionality by an ℓ_1-product metric d_1.
- Even without explicit supervisions of hierarchies and compositions.



Images with Largest Norms in $\mathbb{H}_9^d$ Images with Largest Norms in $\mathbb{H}_{39}^d$

Images with Largest Norms in $\mathbb{H}_4^d$ Images with Largest Norms in $\mathbb{H}_{51}^d$