ZeroC: A Neuro-Symbolic Model for Zero-shot Concept Recognition and Acquisition at Inference Time

Tailin Wu¹, Megan Tjandrasuwita², Zhengxuan Wu¹, Xuelin Yang¹, Kevin Liu¹, Rok Sosič¹, Jure Leskovec¹

Motivation

Humans have the remarkable ability to **recognize** and **acquire** novel visual concepts in a zero-shot manner.



Furthermore, the inferred knowledge (symbolic structure of a new concept) can be **transferred** across domains, allowing models in the second domain to acquire a concept **without ever seeing an example**.

Significance and prior works

Significance:

Our goal is to endowing machine learning (ML) models such zero-shot capabilities, allowing them to:

- Tackle **more complex** recognition tasks at inference time, without further training on those specific tasks.
- Acquire knowledge in inference and represent them as **universal, symbolic representation**, which can be transferred across domains.

Why is it hard:

• ML models typically generalize to examples drawn from **same/similar distribution** as in training.

Prior works:

- Visual compositionality: address factors of variation (e.g. color, position, smiling) without hierarchical structures; or addresses composition of transformation.
- **Concept or relation learning:** do not generalize to hierarchical concepts.
- Zero-shot learning: only generalize to new combinations of features (constituent concepts) while neglecting relation structures.



ZeroC method

Architecture:

ZeroC models a concept/relation with an energy-based model (EBM) that maps tuples of inputs onto a scalar energy:

Concept: concept EBM $E_{X,M,C}(x,m,c)$, $X \in \mathbb{R}^d$: image, $M \in [0,1]^d$: mask, $c \in C$: label **Relation:** relation EBM $E_{X,M_1,M_2,R}(x,m_1,m_2,r)$, $X \in \mathbb{R}^d$: image, $M_1, M_2 \in [0,1]^d$: mask, $r \in \mathbb{R}$: label (string)

The above EBMs define a joint probability: $P_{X,M,C}(x,m,c) = \frac{1}{Z_C} \exp(-E_{X,M,C}(x,m,c))$ If the mask *m* is actually masking an object that belongs to concept *c*, $P_{X,M,C}(x,m,c)$ will be high and $E_{X,M,C}(x,m,c)$ will be low.

Detection at inference: Stochastic Gradient Langevin Dynamics (SGLD)

$$\tilde{m}^{k} = \tilde{m}^{k-1} - \frac{\lambda}{2} \nabla_{m} E_{X,M,C}(x, \tilde{m}^{k-1}, c) + \omega^{k}, \ \omega^{k} \sim \mathcal{N}(0, \lambda)$$
cation: $c^{*} = \operatorname{argmin} E(x, \tilde{m}^{K}, c)$

Classification: $c^* = \operatorname{argmin} E(x, \widetilde{m}^K, c)$

Method:

1. Recognizing novel hierarchical concepts at inference time:

(a) training of models (b) inference: given a hierarchical concept graph, derive its model from trained models



(a) Training elementary concepts and relations with our improved objective.
(b) Inference: summing EBMs according to the concept graph. (c) Detection via SGLD.

2. Acquiring and communicating novel hierarchical concepts:



At inference, $ZeroC_1$ distills 2D demonstration into concept graphs, transfer them to domain 2, and $ZeroC_2$ in domain2 can zero-shot recognize those concepts in 3D image.





¹Stanford University, ²MIT

Results

We design zero-shot recognition and acquisition tasks that are easy for humans, but extremely challenging for machines.

Zero-shot concept recognition (trained with simpler concepts and relations):



	Classification (acc.)		Detection (IoU)		
Model	HD-Letter	HD-Concept	HD-Letter+distractor	HD-Concept+distractor	
Statistics	46.5	53.5	5.69	12.6	
Heuristics	(-)	(-)	42.3	29.2	
CADA-VAE [8]	18.0	66.0	(-)	(-)	
ZeroC (ours)	84.5	70.5	72.5	84.7	
ZeroC composition without R-EBM	62.5	32.5	45.3	84.3	
ZeroC composition without HC-EBM	67.0	55.0	67.7	78.4	
ZeroC without $L^{(\text{pos-std})}$	43.6	65.5	76.1	81.5	
ZeroC without $L^{(neg)}$	64.5	59.0	60.0	84.2	
ZeroC without $L^{(em)}$	81.5	61.0	68.0	86.0	
ZeroC with only $L^{(Improved)}$	27.5	55.5	49.1	81.7	

Zero-shot concept acquisition across domains (see method 2)

	Domain 1 (2D ima	age) Parsing	Domain 2 (3D image)		
Model	Isomorphism (acc.) \uparrow	Edit distance \downarrow	Classification (acc.) \uparrow	Detection (IoU) \uparrow	
Statistics	2.33	3.14	33.3	2.53	
Mask R-CNN [13]+relation classification	35.5	1.01	(-)	(-)	
$\mathbf{ZeroC}_1 ightarrow \mathbf{ZeroC}_2$ (ours)	72.7	0.50	60.7	94.4	

Zero-shot concept recognition for CLEVR

Model	Classification acc (%)
Statistics	33.4
CADA-VAE	45.3
ZeroC (ours)	56.0

Trained with simpler concepts and relations, ZeroC can classify hierarchical concepts



Discussion

- Experiments show that trained with simpler concepts and relations, ZeroC can recognize more complex, hierarchical concepts, given a high-level, symbolic specification of their structures.
- Future work: (1) multiple hierarchies; (2) more realistic images;
 (3) Application in graphs (e.g. molecules), science.



Paper



Code