A causal view of compositional zero-shot recognition

Yuval Atzmon[1] **Felix Kreuk** [1,2] **Uri Shalit** [3] **Gal Chechik**[1,2]
[1]NVIDIA Research, Tel Aviv, Israel
[2]Bar-Ilan University, Ramat Gan, Israel
[3]Technion - Israel Institute of Technology
yatzmon@nvidia.com, gchechik@nvidia.com,

NIPS 2020

Task and Problems

University of Chinese Academy of Science

- Train: Seen attribute-object pairs
- Test: Seen(Familiar compositions) + Unseen(New compositions)

- Distribution-shift： train: red tomato; test: tomato → red
- Entanglement： white-cauliflower: which attribute is white and which is cauliflower

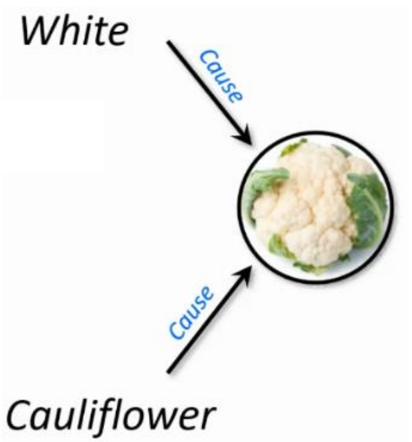
Solution

Most standard models:

$$p(Attr = a, Obj = o \mid Image = x)$$

The author model the causal direction: Physical entities "cause" image features

$$p(Image = x \mid Attr = a, Obj = o)$$

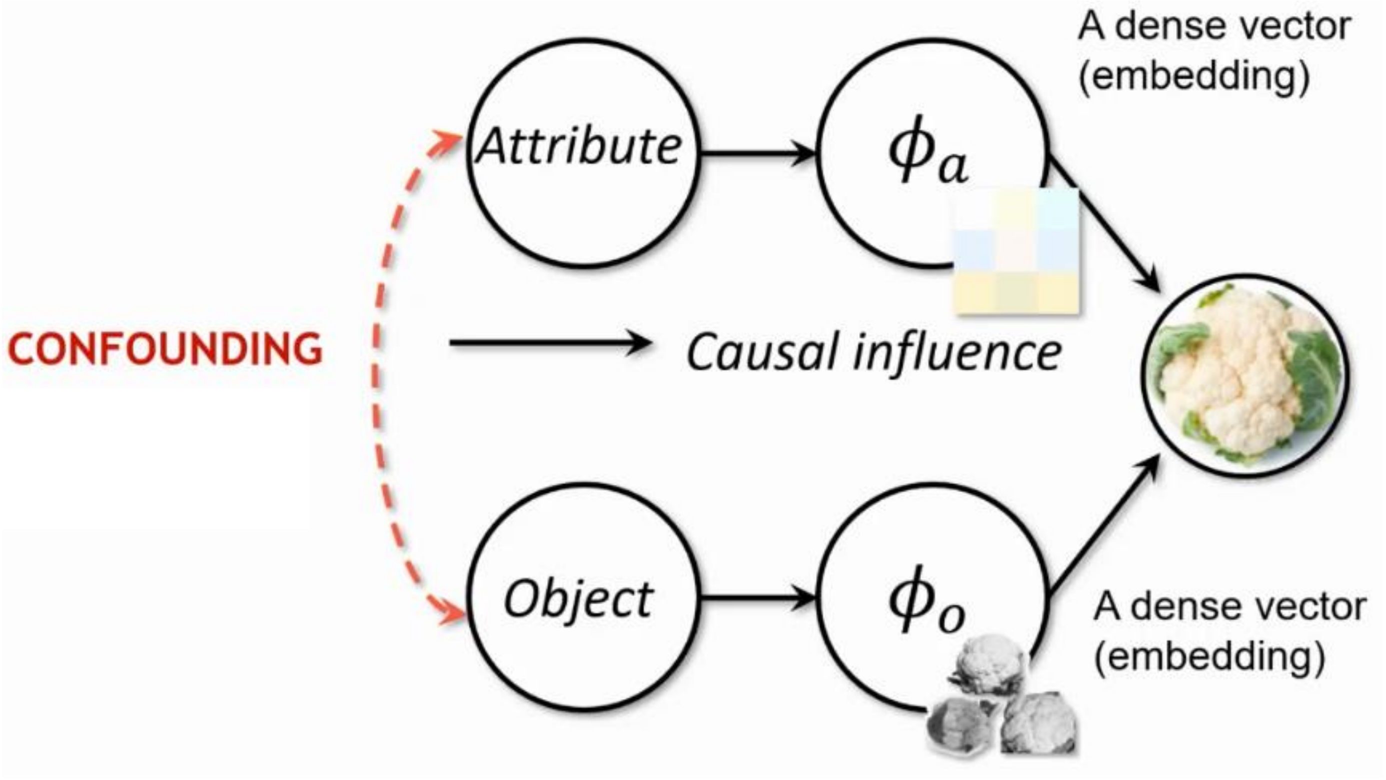


Confounding

University of Chinese Academy of Science

The spurious correlation of attr-obj varies between different domains.

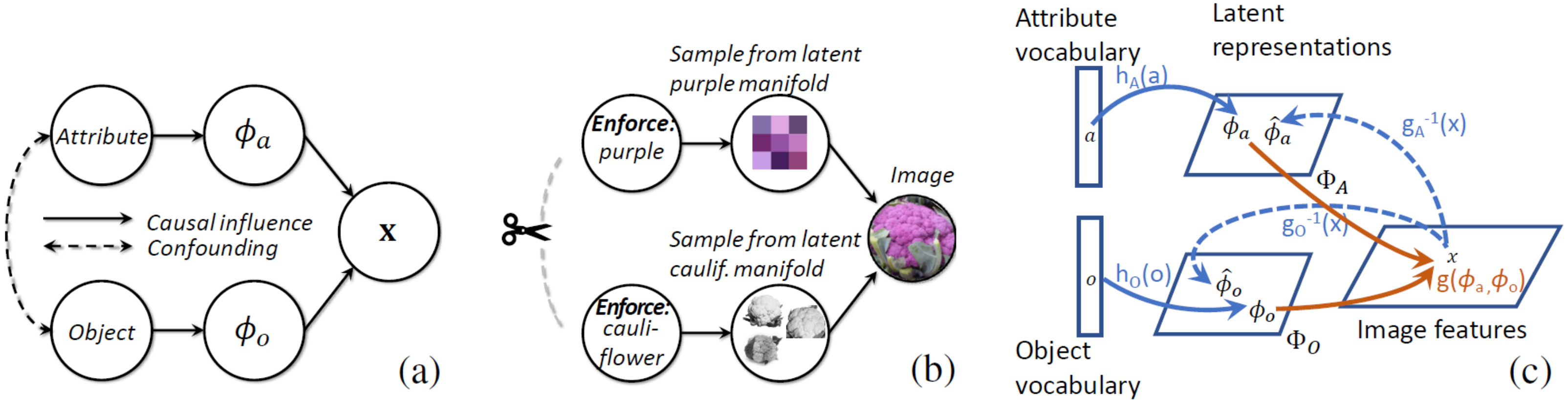
Intervention: $p^{do(A=a,O=o)}(x)$



Generation

Given interdependent attribute and object: $a \in \mathcal{A}, o \in \mathcal{O}$

$p(\phi_a|a)$, $p(\phi_o|o)$: Gaussian distribution, $\phi_a \sim \mathcal{N}(h_a, \sigma_a^2 I)$



$$x \sim \mathcal{N}(g(\phi_a, \phi_o), \sigma_x^2 I)$$

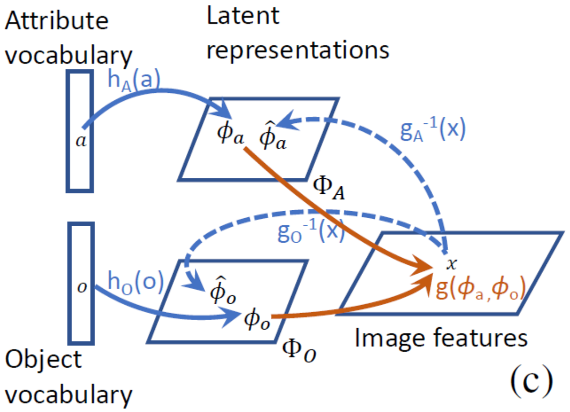
Inference

$$(\hat{a}, \hat{o}) = \underset{a,o \in \mathcal{A}\times\mathcal{O}}{\operatorname{argmax}} \; p^{do(A=a,O=o)}(\mathbf{x})$$

How to maximize the log likelihood of the conditional distribution?

$$p(\mathbf{x}|a,o) = \iint_{\phi_a,\phi_o} p(\mathbf{x},\phi_a,\phi_o|a,o) = \iint_{\phi_a,\phi_o} p(\mathbf{x}|\phi_a,\phi_o)p(\phi_a|a)p(\phi_o|o)d\phi_o d\phi_a$$

Approximate by image: $\hat{\phi}_a = g_A^{-1}(x)$



(c)

$$\hat{L}(a,o) = \frac{1}{\sigma_a^2}||\hat{\phi}_a - h_a||^2 + \frac{1}{\sigma_o^2}||\hat{\phi}_o - h_o||^2 + \frac{1}{\sigma_x^2}||\mathbf{x} - g(h_a,h_o)||^2$$

Train

$$\mathcal{L} = \mathcal{L}_{data} + \lambda_{indep}\mathcal{L}_{indep} + \lambda_{invert}\mathcal{L}_{invert}$$

Independence loss:

$$\begin{aligned}
\mathcal{L}_{indep} &= \mathcal{L}_{oh} + \lambda_{rep}\mathcal{L}_{rep} \\
\mathcal{L}_{oh} &= \mathcal{I}\left(\hat{\phi}_a, O|A\right) + \mathcal{I}\left(\hat{\phi}_o, A|O\right) \\
\mathcal{L}_{rep} &= \mathcal{I}\left(\hat{\phi}_a, \hat{\phi}_o|A\right) + \mathcal{I}\left(\hat{\phi}_a, \hat{\phi}_o|O\right)
\end{aligned}$$

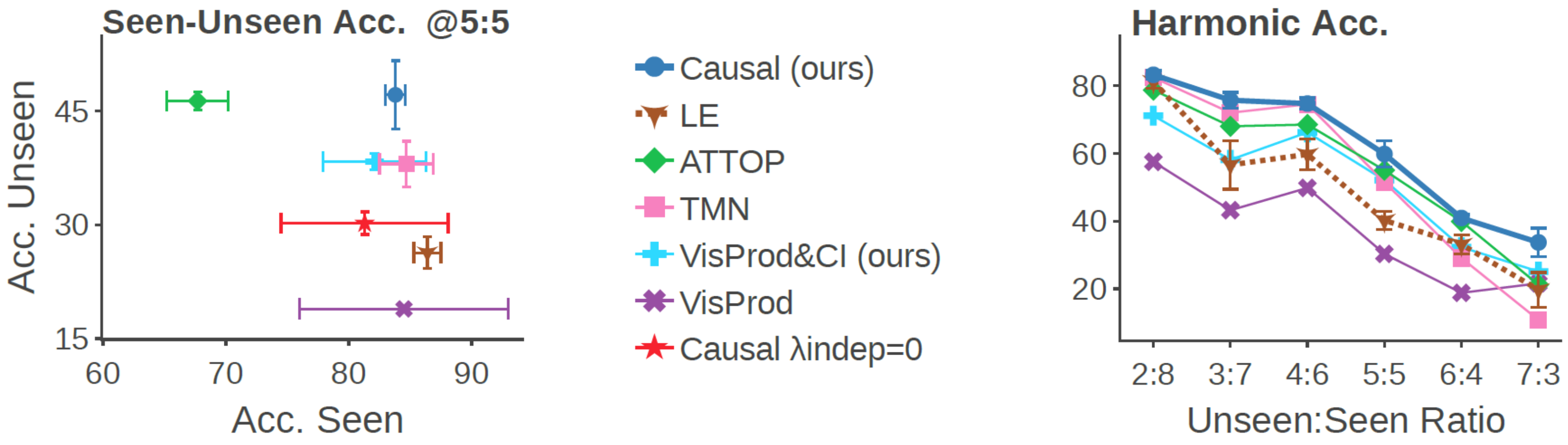
Hilbert-Schmidt Information Criterion

Invertible embedding loss:

$$\mathcal{L}_{invert} = CE(a, f_a(h_a)) + CE(o, f_o(h_o)) + \lambda_g\left[CE(a, f_{ga}(g(h_a, h_o))) + CE(a, f_{go}(g(h_a, h_o)))\right]$$

Experiments

Data bases: Zappos (real), AO-CLEVr (synthetic)



	UNSEEN	SEEN	HARMONIC	CLOSED	AUSUC
WITH PRIOR EMBEDDINGS					
LE	10.7 ± 0.8	52.9 ± 1.3	17.8 ± 1.1	55.1 ± 2.3	19.4 ± 0.3
ATTOP	22.6 ± 2.9	35.2 ± 2.7	26.5 ± 1.4	52.2 ± 1.8	20.3 ± 1.8
TMN	9.7 ± 0.6	51.9 ± 2.4	16.4 ± 1.0	$\mathbf{60.9 \pm 1.1}$	$\mathbf{24.6 \pm 0.8}$
NO PRIOR EMBEDDINGS					
LE*	15.6 ± 0.6	52.0 ± 1.0	24.0 ± 0.7	58.1 ± 1.2	22.0 ± 0.9
ATTOP*	16.5 ± 1.5	15.8 ± 1.9	15.8 ± 1.4	42.3 ± 1.5	16.7 ± 1.1
TMN*	6.3 ± 1.4	$\mathbf{55.3 \pm 1.6}$	11.1 ± 2.3	58.4 ± 1.5	24.5 ± 0.8
CAUSAL λ_{indep}=0	22.5 ± 2.0	45.5 ± 3.7	29.4 ± 1.5	55.3 ± 1.1	22.2 ± 0.9
CAUSAL	$\mathbf{26.6 \pm 1.6}$	39.7 ± 2.2	$\mathbf{31.8 \pm 1.7}$	55.4 ± 0.8	23.3 ± 0.3

Summary

University of Chinese Academy of Science

- A new causal perspective: "which intervention on attribute and object caused the image".
- Disentangled representations of attributes and objects.
- The attributes and objects have distinct and stable generation processes.
- Attributes and objects are fully disentangled.