Visual Commonsense R-CNN

Tan Wang[1,3] Jianqiang Huang[2,3] Hanwang Zhang[3] Qianru Sun[4]

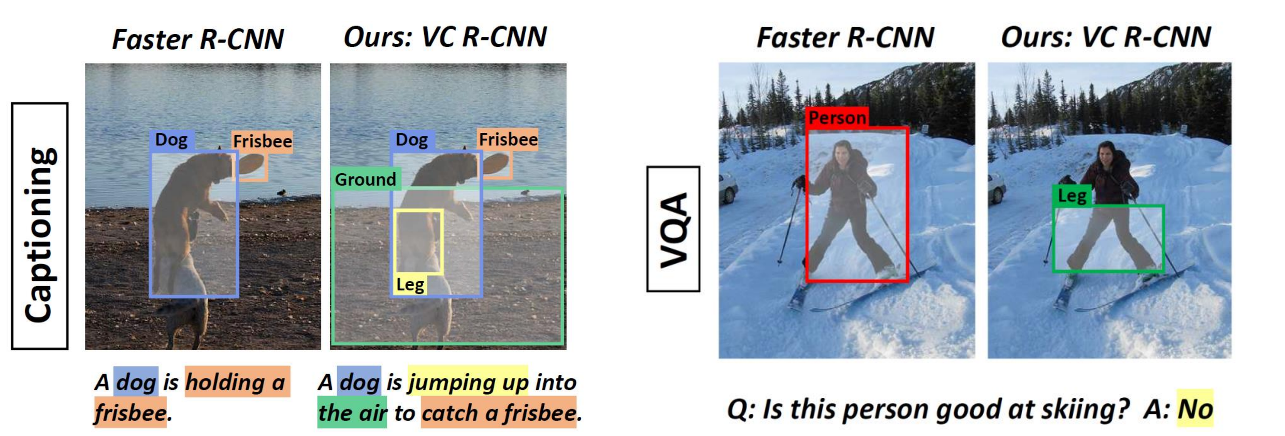
[1]University of Electronic Science and Technology of China, [2]Alibaba Damo Academy
[3]Nanyang Technological University, [4]Singapore Management University

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

Key Words: 1) Common Sense; 2) Causality; 3) Un-/Self-supervised Learning; 4) Representation Learning

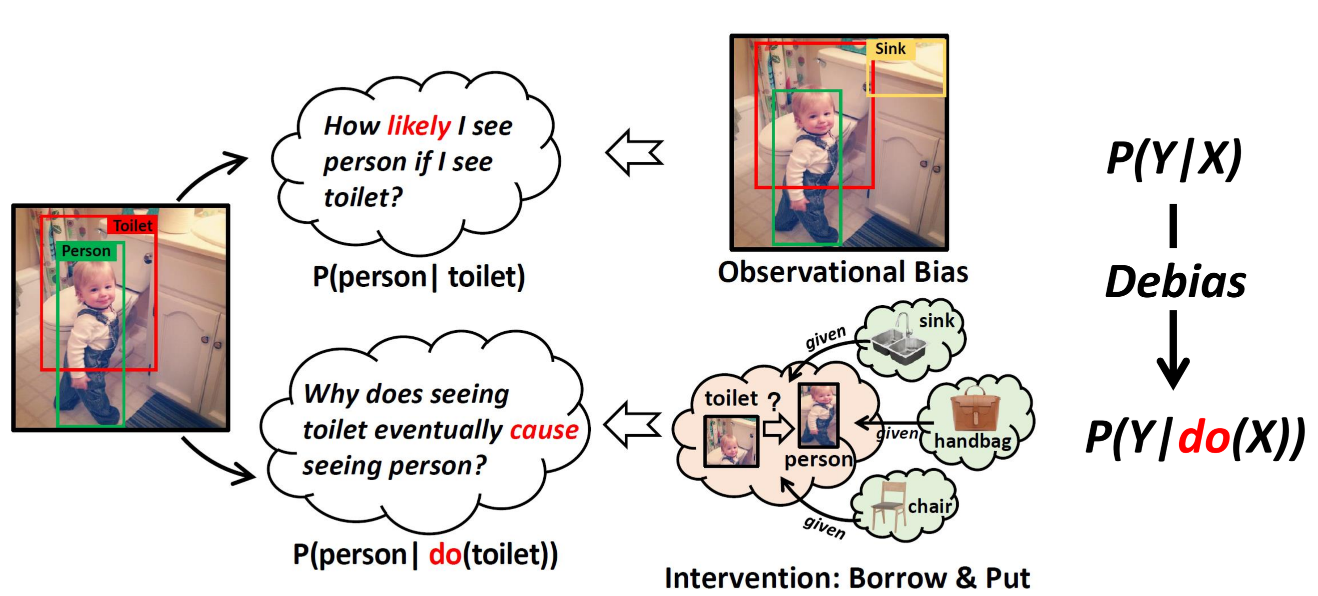
Motivation & Solution

- Today's CV systems are good at telling us "what" (e.g., classification) and "where" (e.g., detection), yet bad at knowing "why". --- ***Machine needs Common Sense!!!***
- ***Cognitive Errors*** due to the lack of common sense.



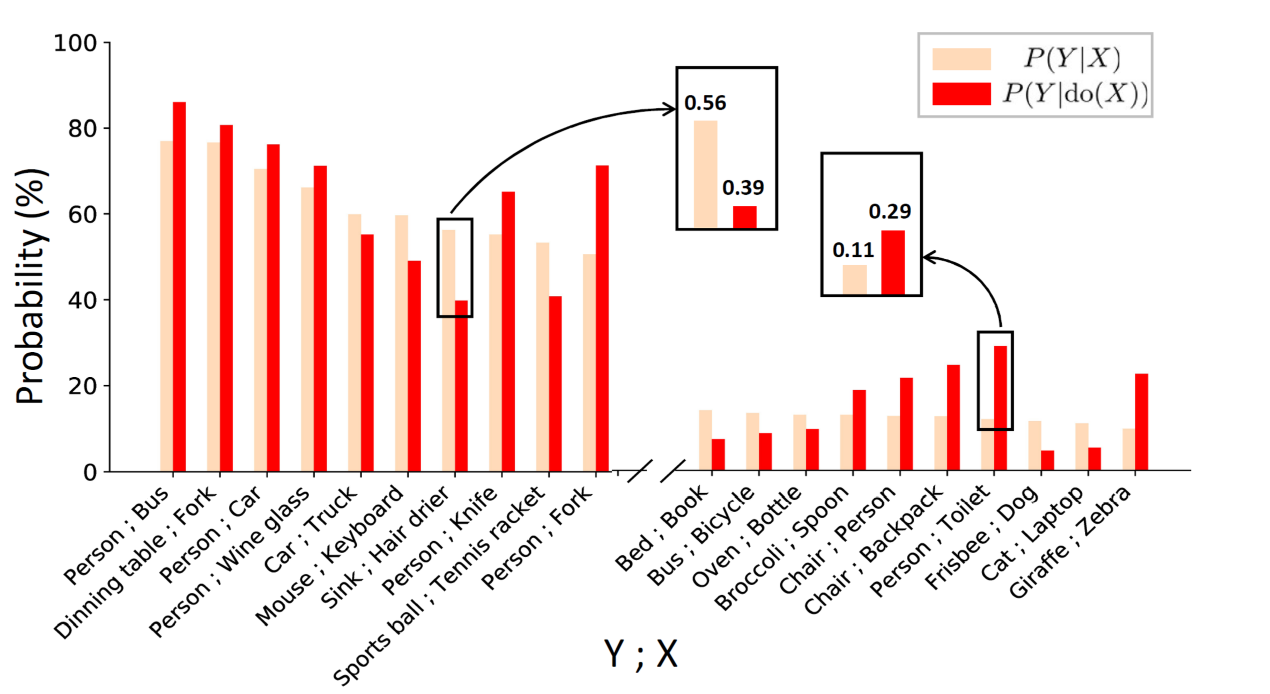
(a) Inexact Visual Relationships (b) Unreasonable Visual Attention

- Our Solution: ***Likelihood*** ⟶ ***Causal Intervention***



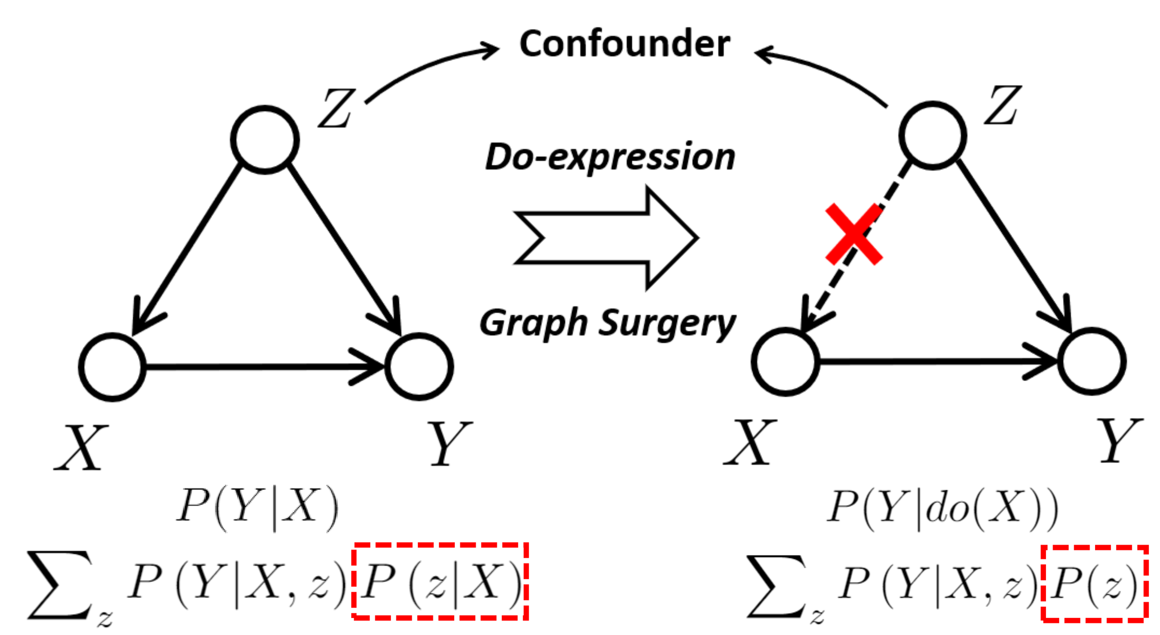
A Toy Experiment

- We performed a toy experiment on MS-COCO with GT labels to compare the ***difference*** between ***P(Y|X)*** and ***P(Y|do(X))***. --- ***Here comes the Confounder Bias!!!***



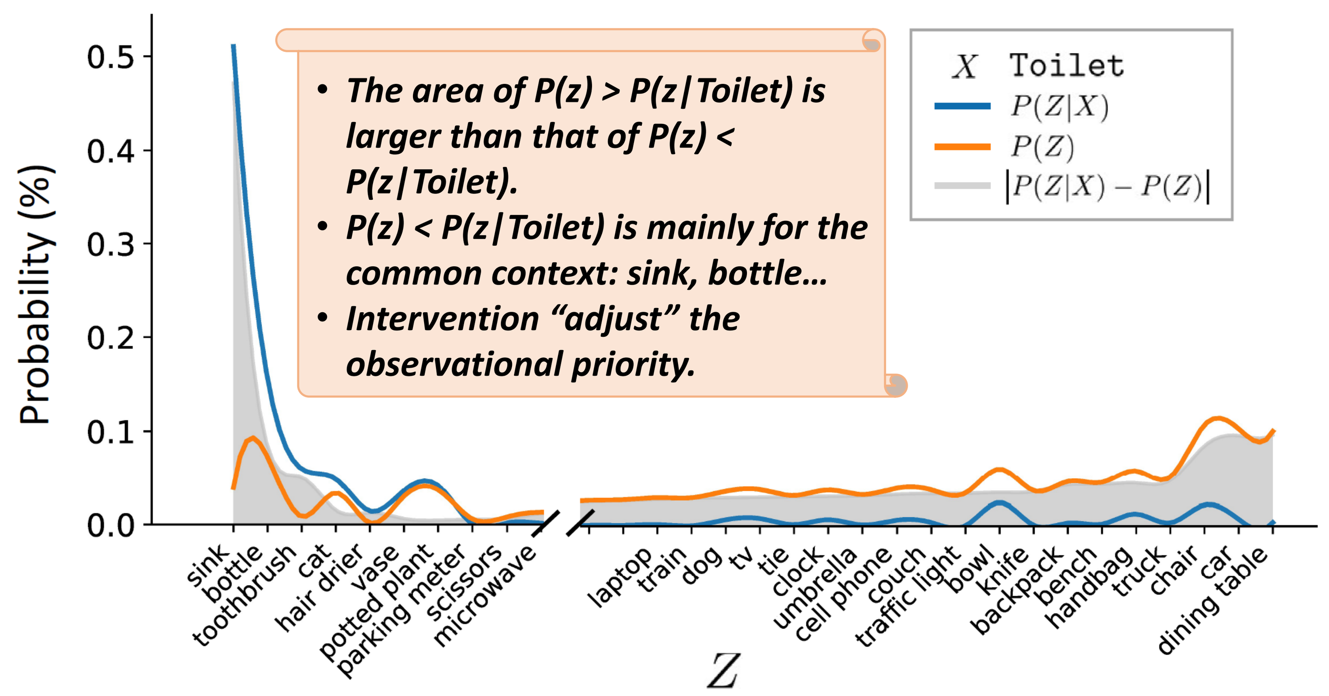
Approach & Framework

- ***Theory: Causal Intervention***

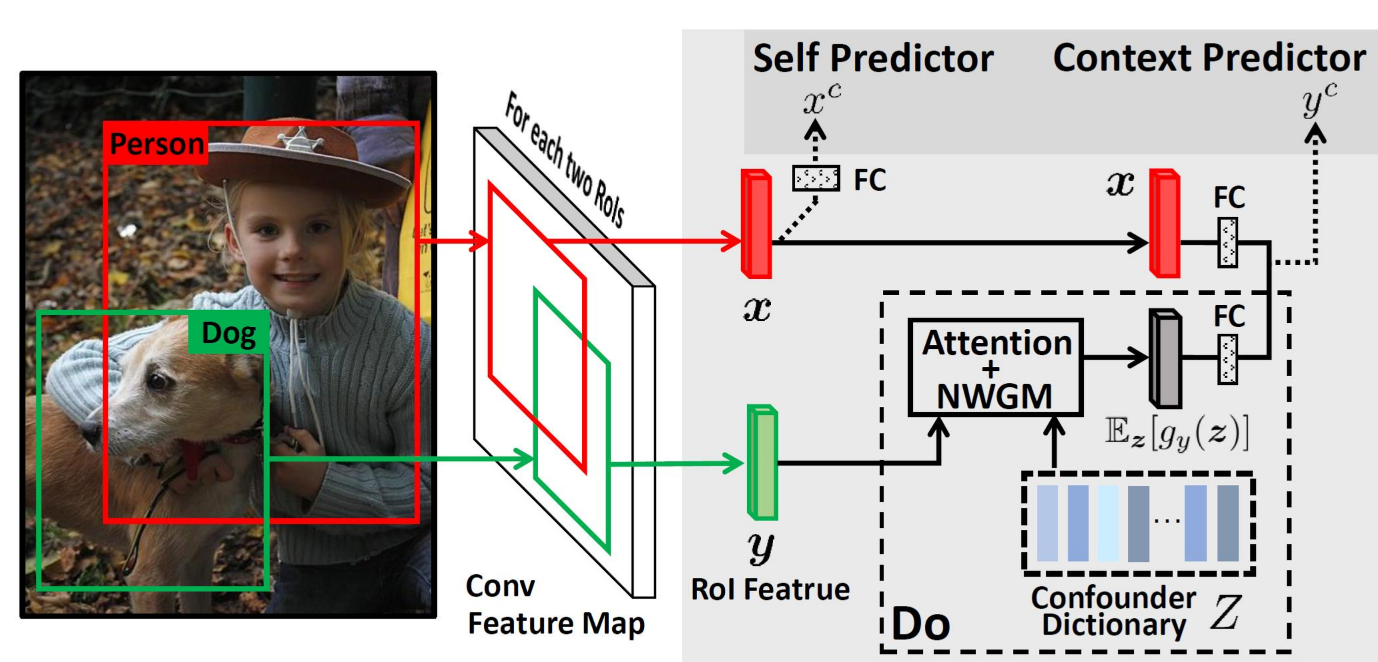


- ***Where is the Confounder Bias from?***

Here we take X = "Toilet" as the example



- ***Implementation: VC R-CNN Framework***



- **Visual Backbone**: R-CNN; **Training Objective**: Causal Intervention
- The **Delivery** of VC R-CNN is a **region feature extractor** for any region proposal.
- **Light**: fast & memory-efficient; **Non-intrusive**: just concatenate

Experiment Results

- ***Image Captioning***

Performance on Karpathy Test Split

Model	Feature	MS-COCO				Open Images			
		B4	M	R	C	B4	M	R	C
Up-Down	Obj	36.7	27.8	57.5	122.3	36.7	27.8	57.5	122.3
	+Cor	38.1	28.3	58.5	127.5	38.3	28.4	58.8	127.4
	+VC	39.5	29.0	59.0	130.5	39.1	28.8	59.0	130.0
AoANet†	Obj	38.1	28.4	58.2	126.0	38.1	28.4	58.2	125.9
	+Cor	38.8	28.9	58.7	128.6	38.9	28.8	58.7	128.2
	+VC	**39.5**	**29.3**	**59.3**	**131.6**	39.3	29.1	59.0	131.5
SOTA	AoANet	38.9	29.2	58.2	129.8	38.9	29.2	58.2	129.8

※ AoANet† indicates the AoANet [1] without the refine encoder. The grey row highlight the results of our VC feature in each model.

Performance on MSCOCO Test Server

Model	BLEU-4		METEOR		ROUGE-L		CIDEr-D	
Metric	c5	c40	c5	c40	c5	c40	c5	c40
Up-Down	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
SGAE	37.8	68.7	28.1	37	58.2	73.1	122.7	125.5
CNM	37.9	68.4	28.1	36.9	58.3	72.9	123.0	125.3
AoANet	37.3	68.1	28.3	37.2	57.9	72.8	124.0	126.2
Up-Down+VC	37.8	69.1	28.5	37.6	58.2	73.3	124.1	126.2
AoANet†+VC	**38.4**	**69.9**	**28.8**	**38.0**	**58.6**	**73.8**	**125.5**	**128.1**

※ Up-Down+VC and AoANet†+VC are the short for concatenated on [4] in Up-Down and AoANet†.

- ***VQA***

Performance on VQA2.0 Dataset

Model	Feature	MS-COCO				Open Images			
		Y/N	Num	Other	All	Y/N	Num	Other	All
Up-Down	Obj	80.3	42.8	55.8	63.2	80.3	42.8	55.8	63.2
	+Cor	81.5	44.6	57.1	64.7	81.3	44.7	57.0	64.6
	+VC	82.5	46.0	57.6	65.4	82.8	45.7	57.4	65.4
MCAN	Obj	84.8	49.4	58.4	67.1	84.8	**49.4**	58.4	67.1
	+Cor	85.0	49.2	58.9	67.4	85.1	49.1	58.6	67.3
	+VC	**85.2**	**49.4**	**59.1**	**67.7**	85.1	49.1	58.9	67.5
SOTA	MCAN	84.8	**49.4**	58.4	67.1	84.8	**49.4**	58.4	67.1

※ The results on VQA2.0 Test Server can be found in our paper.

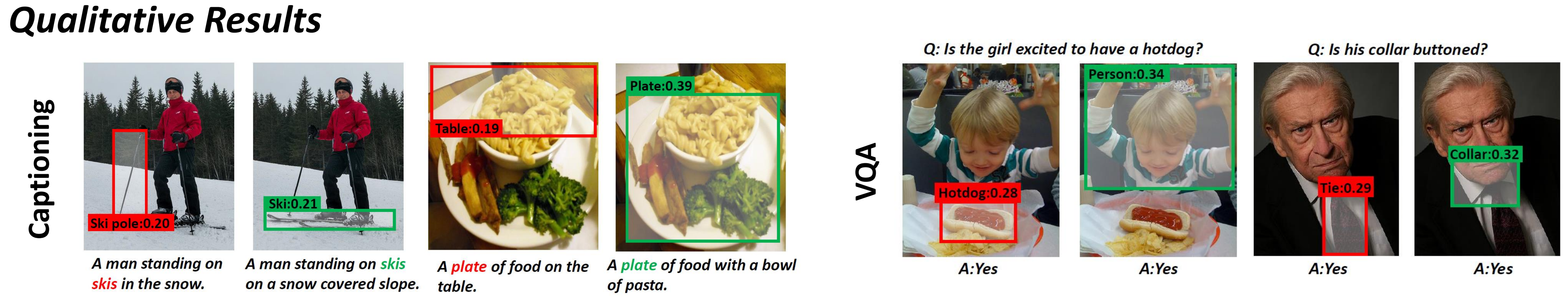
- ***VCR***

Performance on VCR Dataset

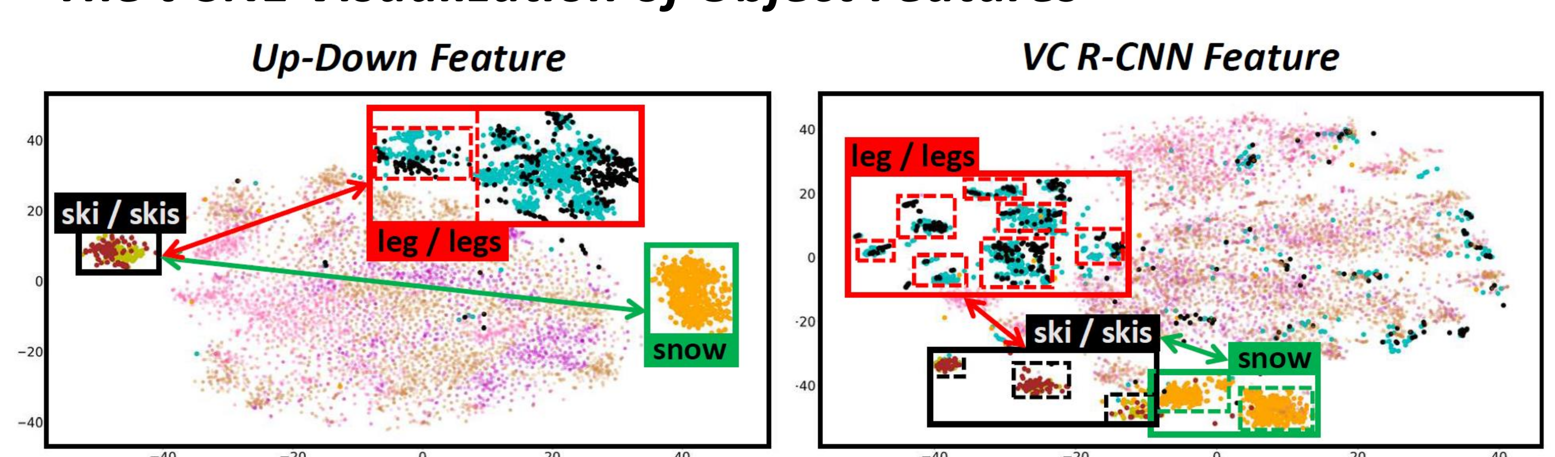
Model	Feature	MS-COCO		Open Images	
		Q→ A	QA→ R	Q→ A	QA→ R
R2C	Obj	65.9	68.2	65.9	68.2
	+Cor	66.5	68.9	66.6	69.1
	+VC	67.4	69.5	67.2	69.9
ViLBERT†	Obj	69.1	69.6	69.1	69.6
	+Cor	69.3	69.9	69.2	70.0
	+VC	**69.5**	70.2	**69.5**	70.3
SOTA	ViLBERT†	69.3	**71.0**	69.3	**71.0**

※ We utilized the ViLBERT† (the full ViLBERT [2] without the pretraining process) for fair comparison.

- ***Qualitative Results***



- ***The t-SNE Visualization of Object Features***



- The "ski" feature of our VC R-CNN is reasonably closer to "leg" and "snow" than the Up-Down feature [3].
- VC R-CNN feature merges into sub-clusters (dashed boxes), implying that the common sense is actually **multi-facet** and varies from context to context.

- ***Reference***

[1] L. Huang *et al.* Attention on attention for image captioning. In ICCV, 2019.

[2] J. Lu *et al.* Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In NeurIPS, 2019.

[3] P. Anderson *et al.* Bottom-up and top-down attention for image captioning and visual question answering. In CVPR, 2018.

- ***Links***

ArXiv: arxiv.org/abs/2002.12204

Code: github.com/Wangt-CN/VC-R-CNN

Zhihu Article: zhuanlan.zhihu.com/p/111306353

QR Code