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]Damo Academy, Alibaba Group
[3]Nanyang Technological University [4]Singapore Management University

wangt97@hotmail.com, jianqiang.jqh@gmail.com, hanwangzhang@ntu.edu.sg, qianrusun@smu.edu.sg

Outline

- What is Common Sense (very brief)
- Visual Common Sense in CV
- Previous Work
- Challenge —— Observational Bias
- A Toy Experiment for Causal Intervention
- Proposed VC R-CNN

Outline

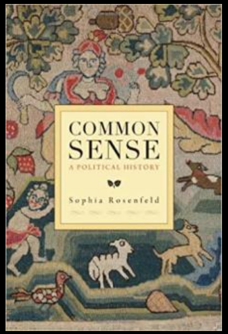
- What is Common Sense (very brief)
- Visual Common Sense in CV
- Previous Work
- Challenge —— Observational Bias
- A Toy Experiment for Causal Intervention
- Proposed VC R-CNN

What is Common Sense

- Many philosophers try to explain “Common Sense”

The ability with which animals (including humans) process **sense perceptions, memories and imagination** in order to reach many types of **basic judgments**.

——Aristotle, The first person to discuss “commonsense”



Those **plain, self-evident truths** that one needed **no proof to accept precisely** because they accorded so well with the basic intellectual capacities and experiences of the whole social body.

——Rosenfeld, Common Sense: A Political History

Outline

- What is Common Sense (very brief)
- Visual Common Sense in CV
- Previous Work
- Challenge —— Observational Bias
- A Toy Experiment for Causal Intervention
- Proposed VC R-CNN

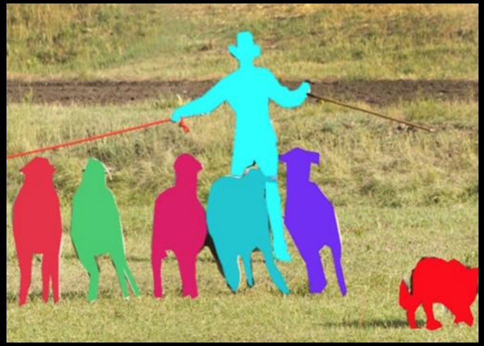
Visual Common Sense in CV

- Today's CV systems are very good at answering……

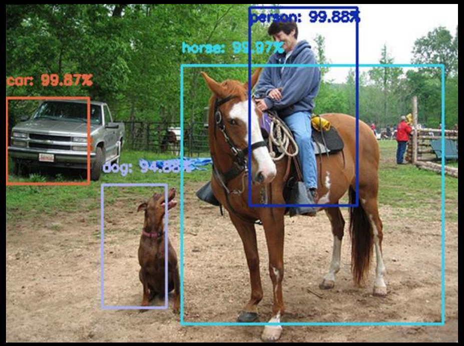
What ?

Where ?

Classification



Segmentation



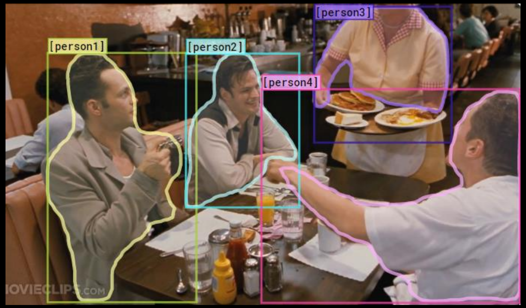
Detection

Tracking

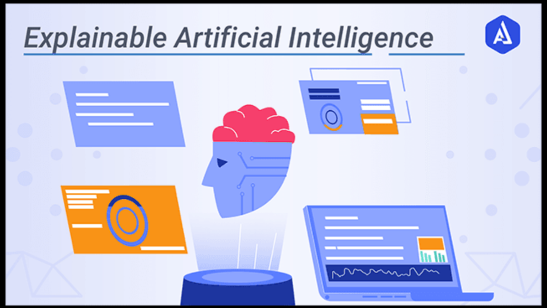
Visual Common Sense in CV

- However, not good at answering ……

Why ?



Visual Reasoning



Explainable AI

Cognition & Common Sense

Our Machine needs Visual Common Sense

Visual Common Sense in CV

- How to define Visual Common Sense in CV ?

Visual + Common + Sense-making

Visual Common Sense in CV

- How to define Visual Common Sense in CV ?

Visual + Common + Sense-making

Visual Common Sense in CV

- How to define Visual Common Sense in CV ?

Visual : Large Scale Visual Data

Unsupervised Fashion (No Common Sense Label)

Visual Common Sense in CV

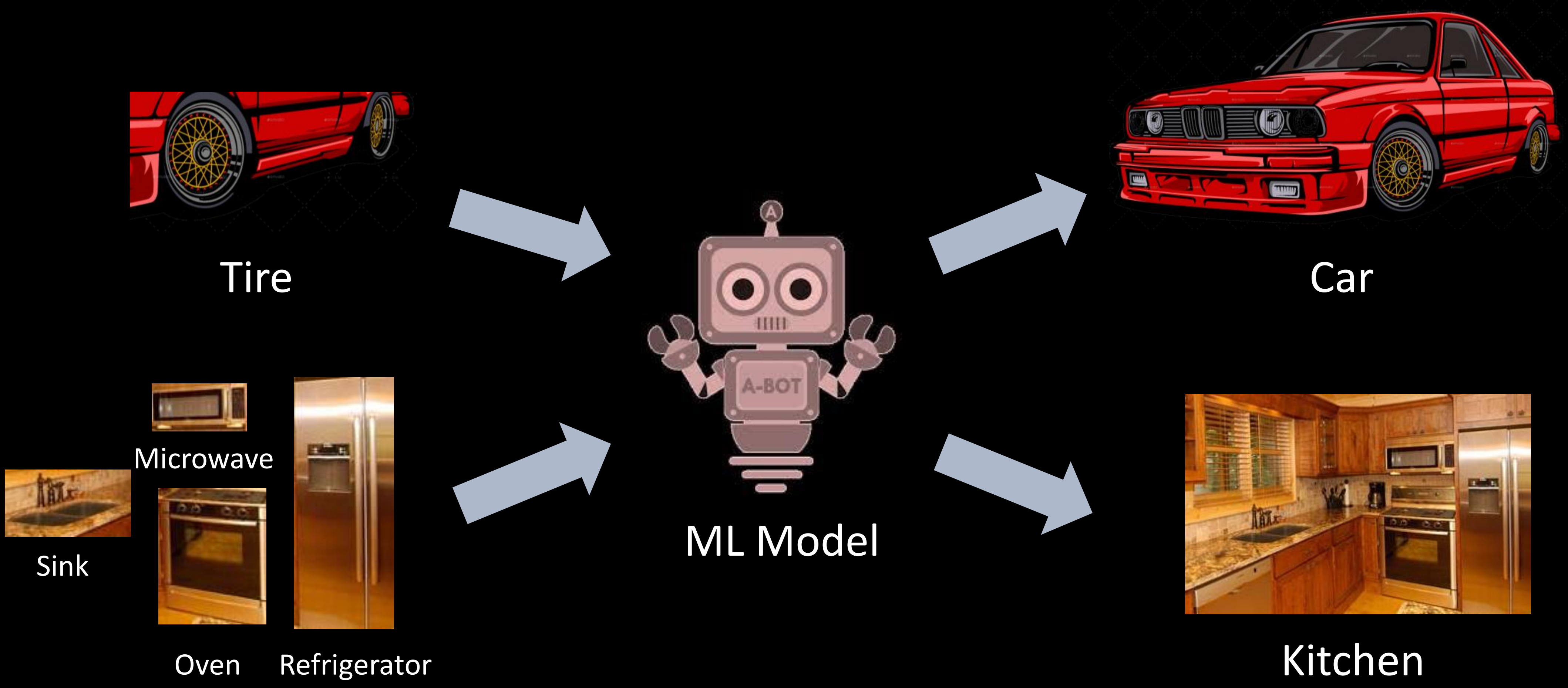
- How to define Visual Common Sense in CV ?

Visual + **Common** + Sense-making

Visual Common Sense in CV

- How to define Visual Common Sense in CV ?

  Common: Correlation; The Cornerstone of ML



Visual Common Sense in CV

- How to define Visual Common Sense in CV ?

Visual + Common + **Sense-making**

Visual Common Sense in CV

- How to define Visual Common Sense in CV ?

Sense-making: Cognitive Reasoning; Affordance

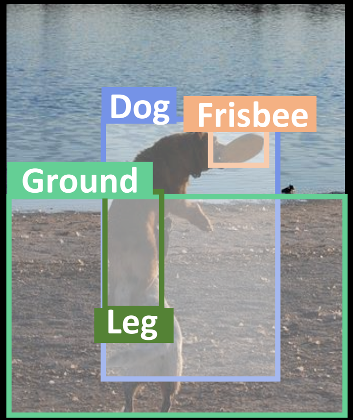
Non-VC vs. VC

- Cognition Errors in nowadays Vision & Language Tasks —— Image Captioning

Without VC

A dog is holding a frisbee.

With VC



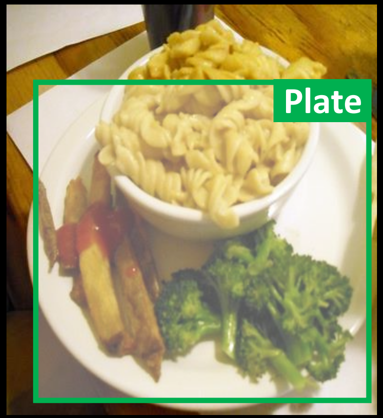
A dog is jumping up into the air to catch a frisbee.

Non-VC: Inexact Visual Relationships

Without VC

A plate of food on the table.

With VC



A plate of food with a bowl of pasta.

Non-VC: Non-reasonable Visual Attention

Non-VC vs. VC

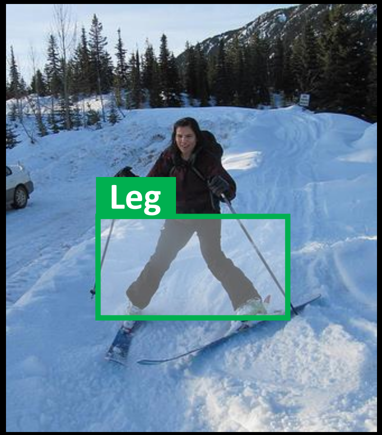
- Cognition Errors in nowadays Vision & Language Tasks —— VQA

Without VC ***With VC***

Q: Is the girl excited to have a hotdog?

A:Yes

Without VC ***With VC***



Q: Is this person good at skiing?

A: No

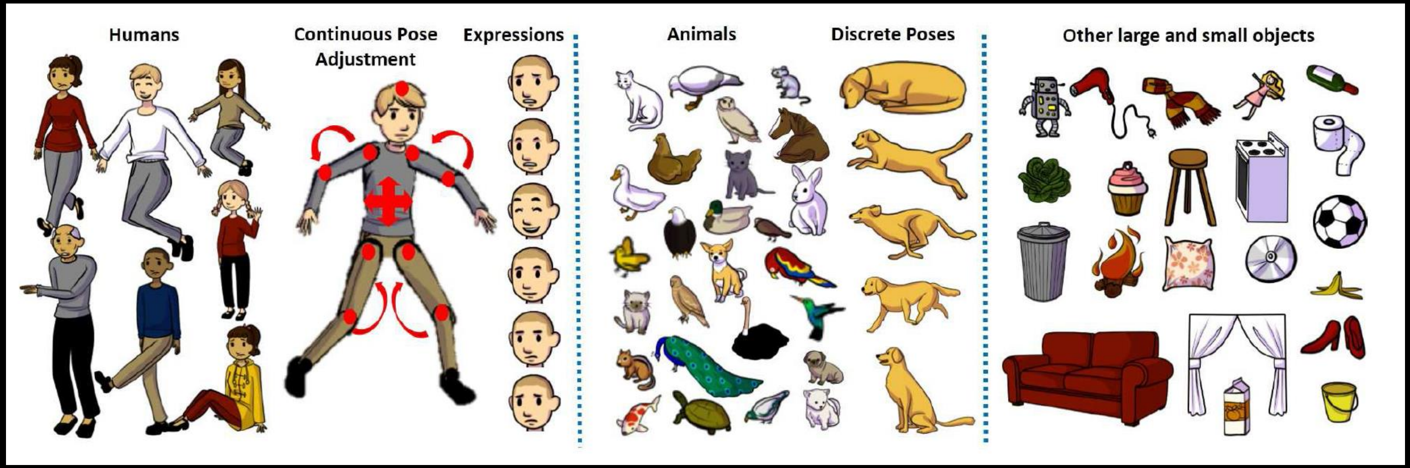
Non-VC: Non-reasonable Visual Attention

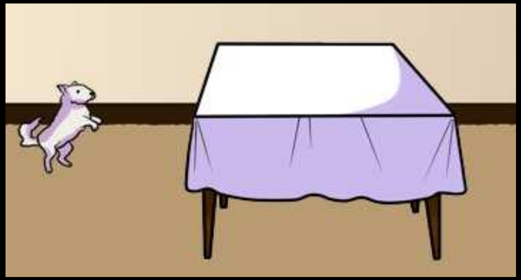
Outline

- What is Common Sense (very brief)
- Visual Common Sense in CV
- Previous Work
- Challenge —— Observational Bias
- A Toy Experiment for Causal Intervention
- Proposed VC R-CNN

Supervised Visual Commonsense Learning

- Pre-defined VC Knowledge Base





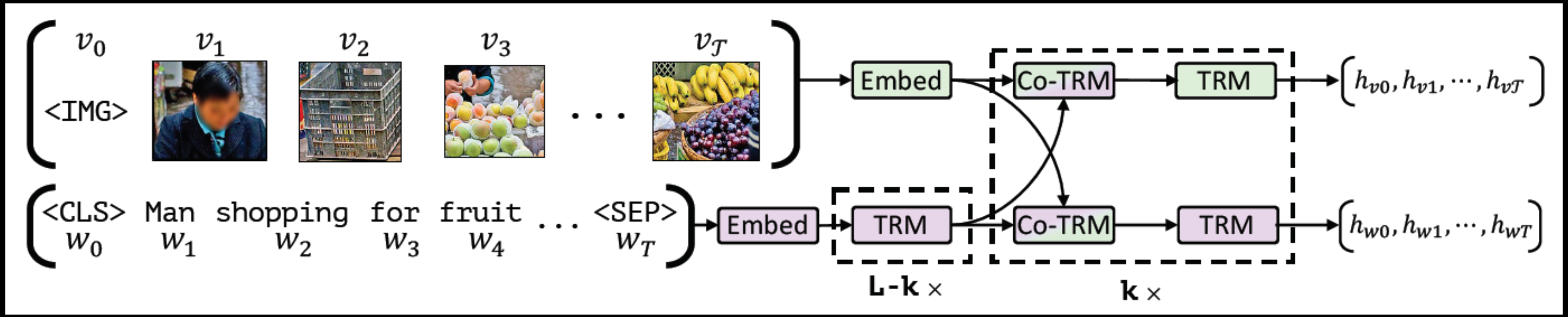
dogs gather around table

- Abstract scene dataset containing 213 relations and 2466 nouns.
- The commonsense assertation needs manual annotation ⟶ **Weakly supervised Learning**

Vedantam, Ramakrishna, et al. Learning Common Sense Through Visual Abstraction, *ICCV 2015*

Weekly Supervised Visual Commonsense Learning

- Vision-and-language Corpus



- Reporting Bias

Most : *Many people are walking on the street.*

Little : *Many people are walking with legs.*

→ **Unsupervised Learning**

Lu, Jiasen, et al. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS 2019*
Su, Weijie, et al. Vl-bert: Pre-training of generic visual-linguistic representations. *ICLR 2020*

Unsupervised Word Vector Learning in NLP

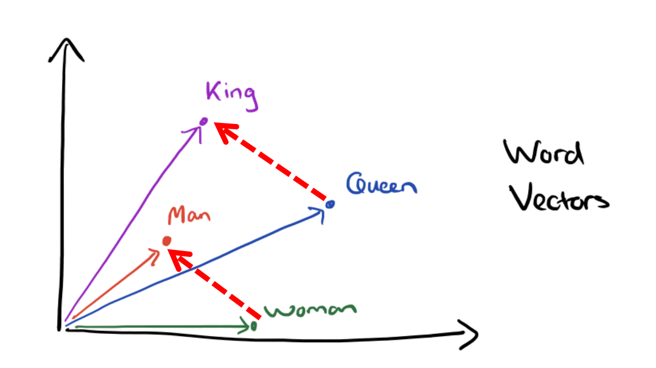
Pineapples are yellow and spikey.

Skip-gram Model

Pineapples are yellow and spikey.

CBOW Model

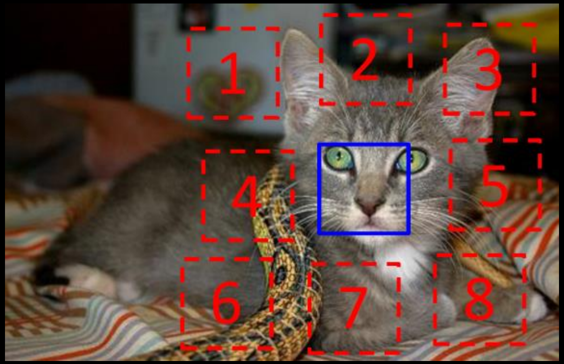
Unsupervised Contextual Prediction



King Man Woman Queen

Mikolov, Tomas, et al. Distributed Representations of Words and Phrases and their Compositionality, *NIPS 2013*
Devlin, Jacob, et al. Bert: Pre-training of deep bidirectional transformers for language understanding, *NAACL 2019*

Why not Vision ?



X = ( , ) —> Y = 3

Contextual Patch Prediction

Solving Jigsaw Puzzles

Context Encoder

NOT effective in downstream tasks

Just *the Correlation Prediction* —— Observational Bias

Doersch, Carl, et al. Unsupervised Visual Representation Learning by Context Prediction, *ICCV 2015*
Noroozi, Mehdi, and Paolo Favaro. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles, *ECCV 2016*
Pathak, Deepak, et al. Context encoders: Feature learning by inpainting, *CVPR 2016*

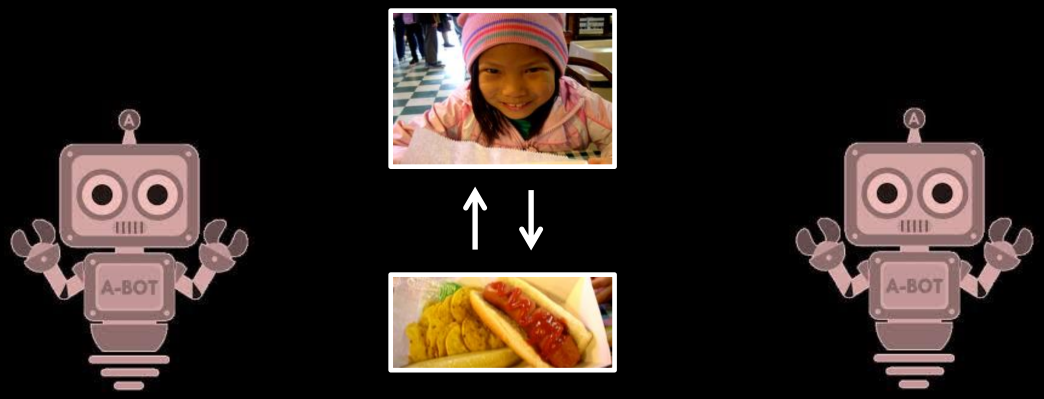
Outline

- What is Common Sense (very brief)
- Visual Common Sense in CV
- Previous Work
- Challenge —— Observational Bias
- A Toy Experiment for Causal Intervention
- Proposed VC R-CNN

What is the Observational Bias

- Correlation ≠ Sense-making

Observed Images



$$\frac{P(\text{Person}, \text{Hotdog})}{P(\text{Hotdog})} = \frac{45174}{82783} \approx 0.5457$$

Visual & Common

VQA Task

Q: Is the girl excited to have a hotdog?

A: Yes

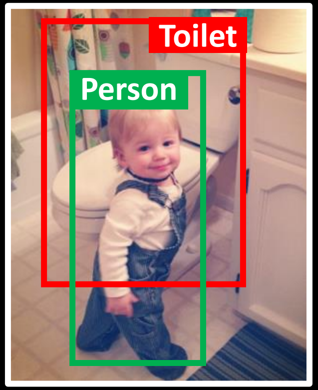
NOT Sense-making

How Can We Reach "Sense-making" ?

Predict

Object A → Object B

Learning the VC about A causes the existence of B



Predict

Person → Toilet

→ Learning the VC about Person causes the existence of Toilet → Person can use Toilet

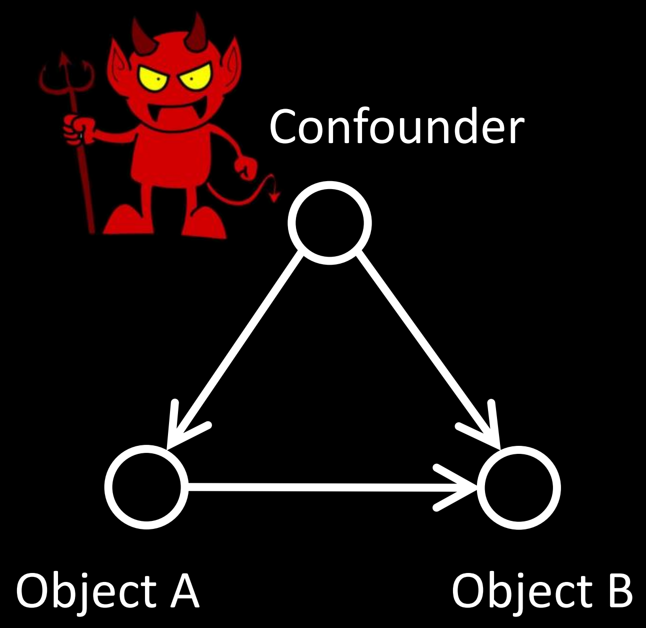
Predict

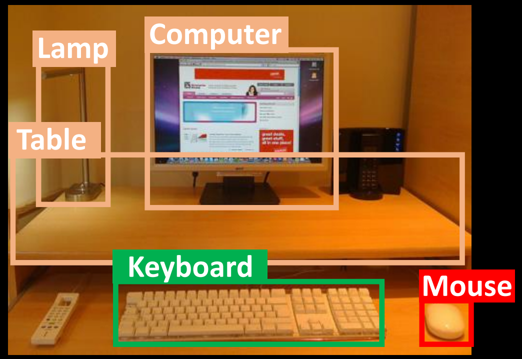
Toilet → Person

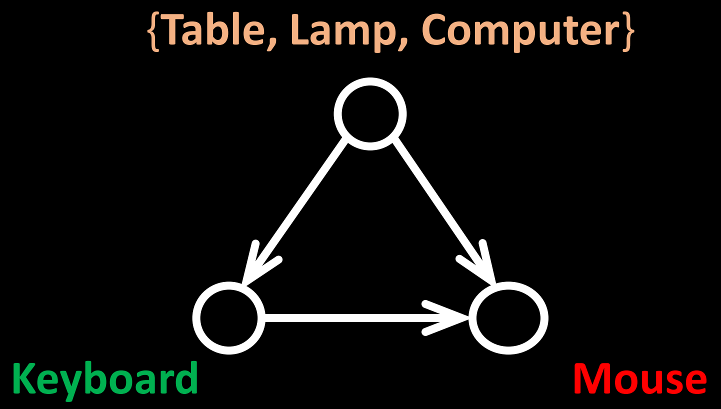
→ Learning the VC about Toilet causes the existence of Person → Toilet can be used by Person

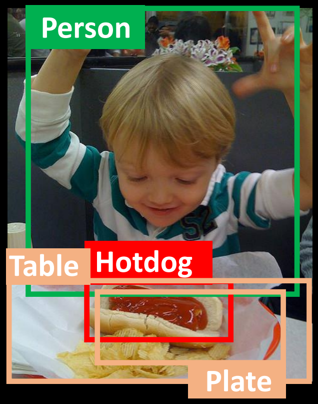
But in the Real World …

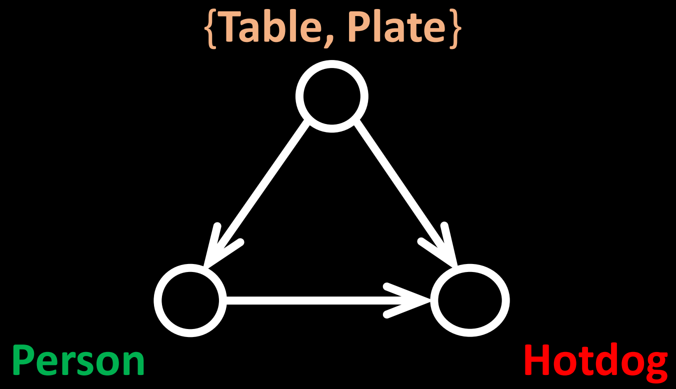
- The Hidden Evil —— Confounder





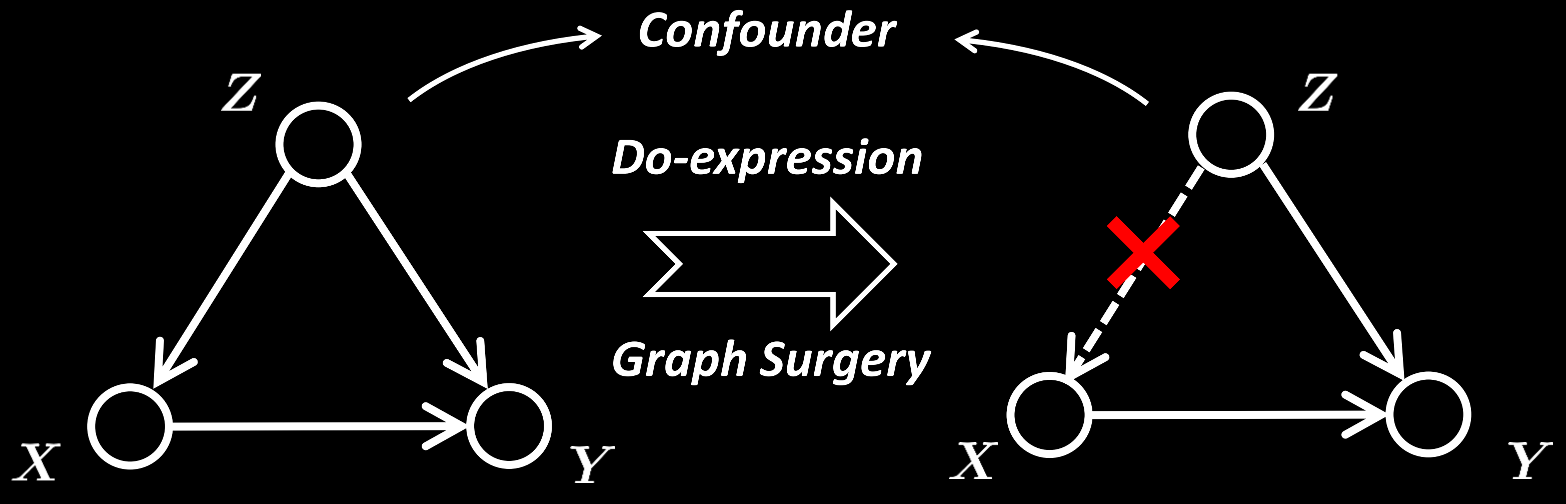






How to Eliminate the Observational Bias

- Causal Intervention —— Backdoor Adjustment



$P(Y|X)$ $\qquad\qquad$ $P(Y|do(X))$

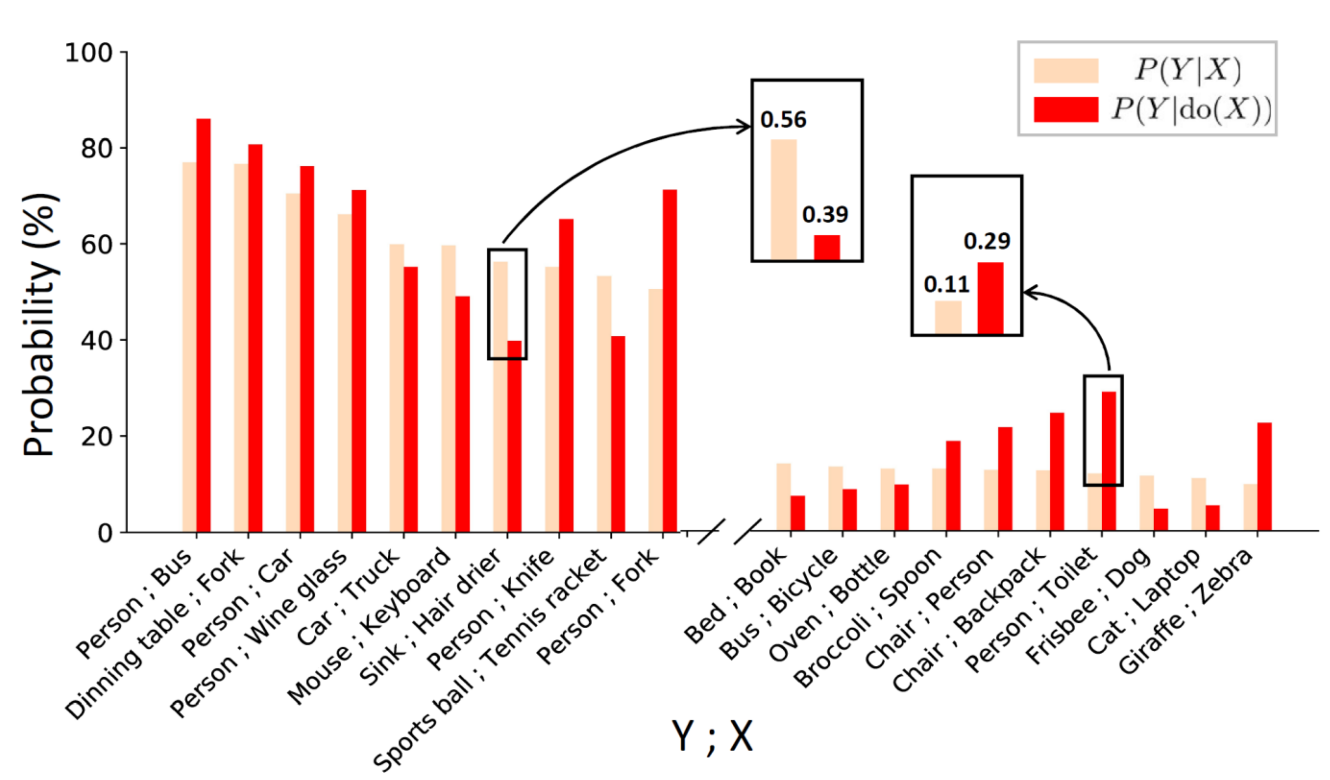
$$P(Y|X) = \sum_z P(Y|X,z)P(z|X) = \frac{P(Y,X)}{P(X)}$$

$$P(Y|\mathrm{do}(X)) = \sum_z P(Y|X,z)P(z) = \sum_z \frac{P(Y,X,z)P(z)}{P(X,z)}$$

Outline

- What is Common Sense (very brief)
- Visual Common Sense in CV
- Previous Work
- Challenge —— Observational Bias
- A Toy Experiment for Causal Intervention
- Proposed VC R-CNN

A Toy Experiment on MS-COCO



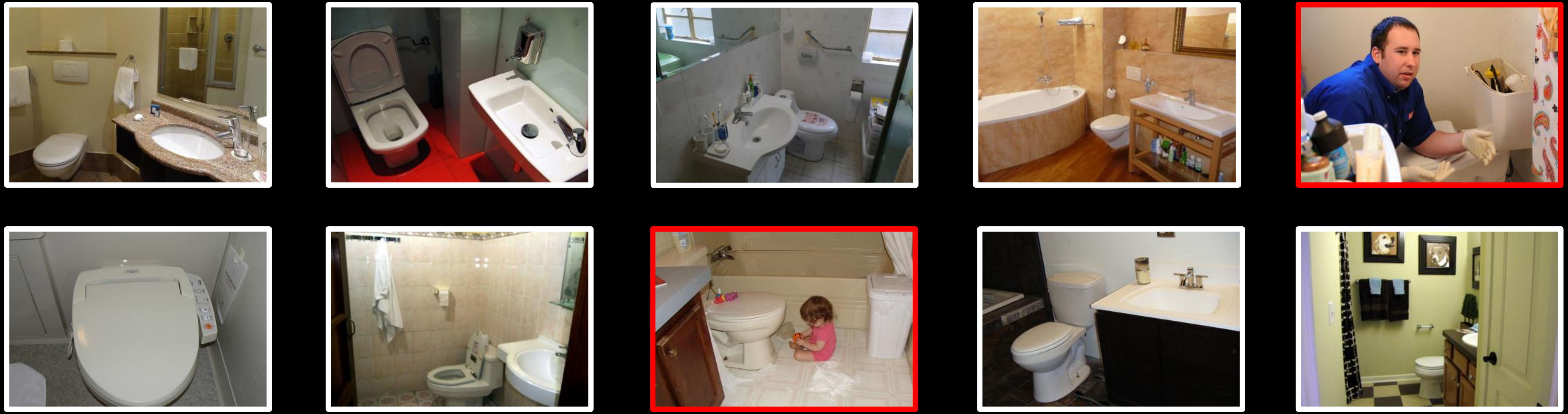
$$P(Y|X) = \sum_z P(Y|X,z)P(z|X) = \frac{P(Y,X)}{P(X)}$$

$$P(Y|\mathrm{do}(X)) = \sum_z P(Y|X,z)P(z) = \sum_z \frac{P(Y,X,z)P(z)}{P(X,z)}$$

A Toy Experiment on MS-COCO

- The detailed analysis —— Person; Toilet

Passive Observation



Due to the privacy, the image with both Person and Toilet can be very little.

$$P(\text{Person}|\text{Toilet}) = \frac{P(\text{Person}, \text{Toilet})}{P(\text{Toilet})} = \frac{277}{2317} \approx 0.11$$

A Toy Experiment on MS-COCO

- The detailed analysis —— Person; Toilet

Causal Intervention

z = Sink

…

$$\frac{P(\text{Person}, \text{Toilet}, \text{Sink})P(\text{Sink})}{P(\text{Toilet}, \text{Sink})} = \frac{119 \times 0.0397}{1183} \approx 0.0039$$

z = Cup

…

$$\frac{P(\text{Person}, \text{Toilet}, \text{Cup})P(\text{Cup})}{P(\text{Toilet}, \text{Cup})} = \frac{13 \times 0.0787}{144} \approx 0.0071$$

⋮ ⋮ ⋮

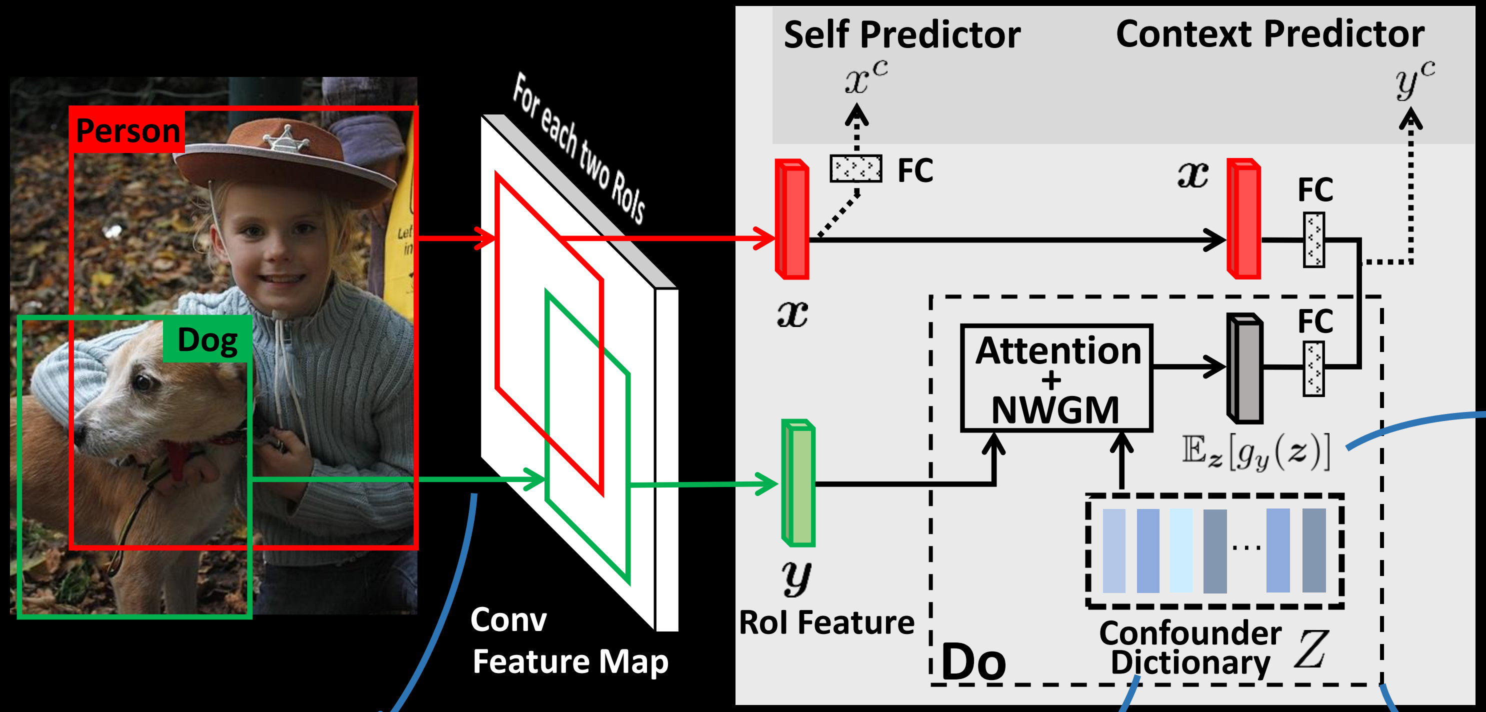
$$P(\text{Person}|do(\text{Toilet})) = \sum_z P(\text{Person}|\text{Toilet}, z)P(z) = \sum_z \frac{P(\text{Person}, \text{Toilet}, z)P(z)}{P(\text{Toilet}, z)} \approx 0.29$$

Outline

- What is Common Sense (very brief)
- Visual Common Sense in CV
- Previous Work
- Challenge —— Observational Bias
- A Toy Experiment for Causal Intervention
- Proposed VC R-CNN

Our Framework

$$L_{self}(p, x^c) = -log(p[x^c]) \qquad L_{cxt}(p_i, y_i^c) = -log(p_i[y_i^c])$$



$$Softmax(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{W}_2 \cdot \mathbb{E}_{\boldsymbol{z}}[g_y(\boldsymbol{z})])$$

$$\boldsymbol{q} = \boldsymbol{W}_3\boldsymbol{y},\ \boldsymbol{K} = \boldsymbol{W}_4\boldsymbol{Z}^T$$
$$\boldsymbol{a} = Softmax(\boldsymbol{q}^T\boldsymbol{K}/\sqrt{\sigma})$$
$$\boldsymbol{A} = [\boldsymbol{a}; \ldots; \boldsymbol{a}]$$
$$\mathbb{E}_{\boldsymbol{z}}[g_y(\boldsymbol{z})] = \sum_z [\boldsymbol{A} \odot \boldsymbol{Z}] P(\boldsymbol{z})$$

CNN Backbone
Ground-truth Bbox

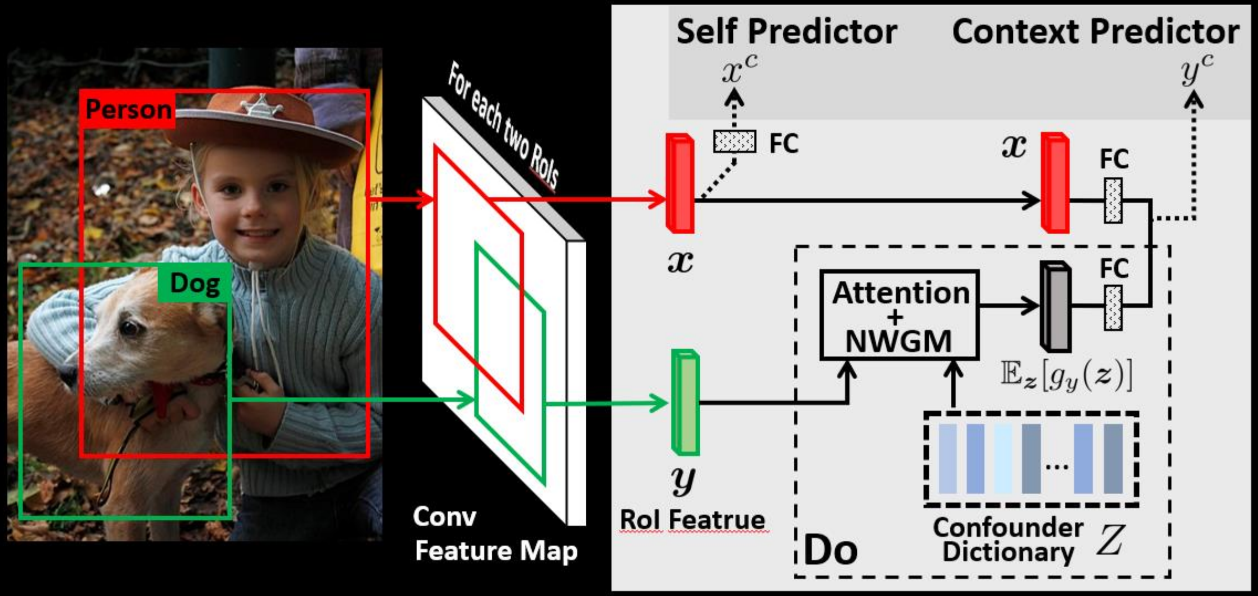
Precomputed
RoI Feature

$$P(\text{Dog}|\text{do}(\text{Person})) = \sum_{z \in Z} P(\text{Dog}|\text{Person}, z)P(z)$$

e.g., $\boldsymbol{z}$ = Car, Frisbee, Bench … **(Non-local Objects)**

* Details can be found in our paper “Visual Commonsense R-CNN”, CVPR 2020

Our Framework



Highlights

- Unsupervised representation learning via causal inference
- Fast ; Light ; Non-intrusive (Easy to Use)
- SOTA performance on 3 downstream tasks

Experimental Results

Image Captioning

Model	Feature	MS-COCO B4	MS-COCO M	MS-COCO R	MS-COCO C	Open Images B4	Open Images M	Open Images R	Open Images 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

Performance on Karpathy Test Split

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**

Performance on MSCOCO Test Server

VQA

Model	Feature	MS-COCO Y/N	MS-COCO Num	MS-COCO Other	MS-COCO All	Open Images Y/N	Open Images Num	Open Images Other	Open Images 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

Performance on VQA2.0 Dataset

VCR

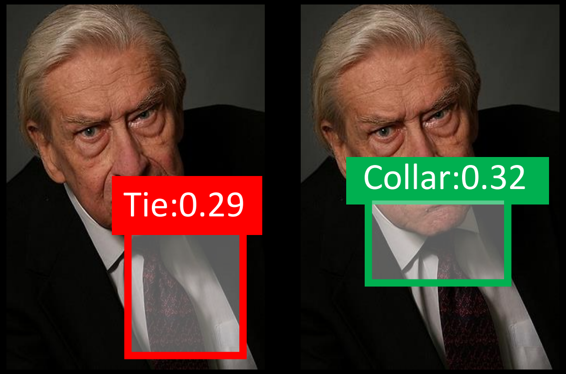
Model	Feature	MS-COCO Q→A	MS-COCO QA→R	Open Images Q→A	Open Images QA→R
R2C	Obj	65.9	68.2	65.9	68.2
	+Cor	66.5	68.9	66.6	69.1
	+VC	67.4	69.5	67.2	69.9
ViLBERT†	Obj	69.1	69.6	69.1	69.6
	+Cor	69.3	69.9	69.2	70.0
	+VC	**69.5**	70.2	**69.5**	70.3
SOTA	ViLBERT†	69.3	**71.0**	69.3	**71.0**

Performance on VCR Dataset

Qualitative Results

- Our Learned Visual Common Sense

Q: Is his collar buttoned?



A:Yes *A:Yes*

VQA

Q: Where are [person8] and [person2] ? A: They are at wedding.

R: They are surrounded by tables and wedding guests.

VCR

Reasonable Attention Weight in downstream Tasks

Take-home Message

More advanced VC:
Counterfactual → Humor, Empathy?

The only way to Artificial General Intelligence (AGI) ?

Correlation ≠ Causality

“BERT + X” is a wrong hype ?

One Small Step for Human
One Giant Leap for AI ?