Object Detection Meets Knowledge Graphs

Yuan Fang, Kingsley Kuan, Jie Lin,

Cheston Tan and Vijay Chandrasekhar

*Agency for Science, Technology and Research (A*STAR), Singapore*

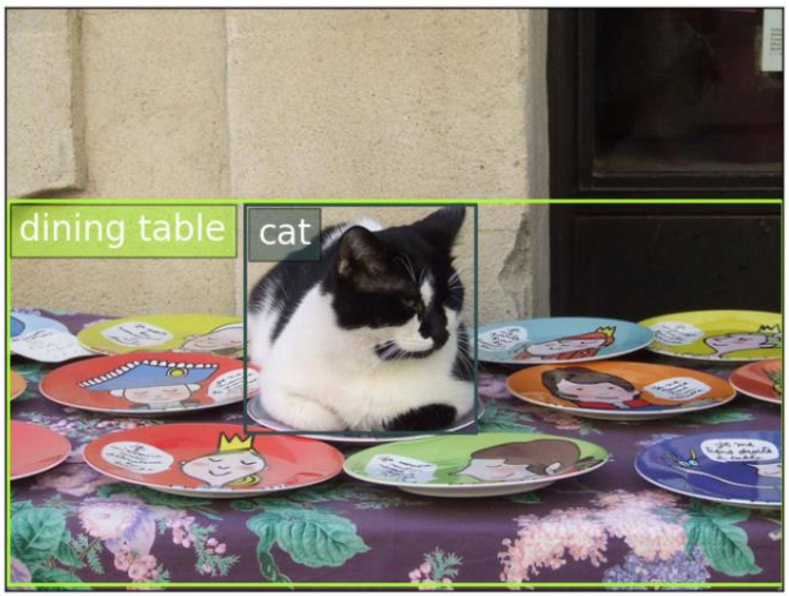
Institute for Infocomm Research

Outline

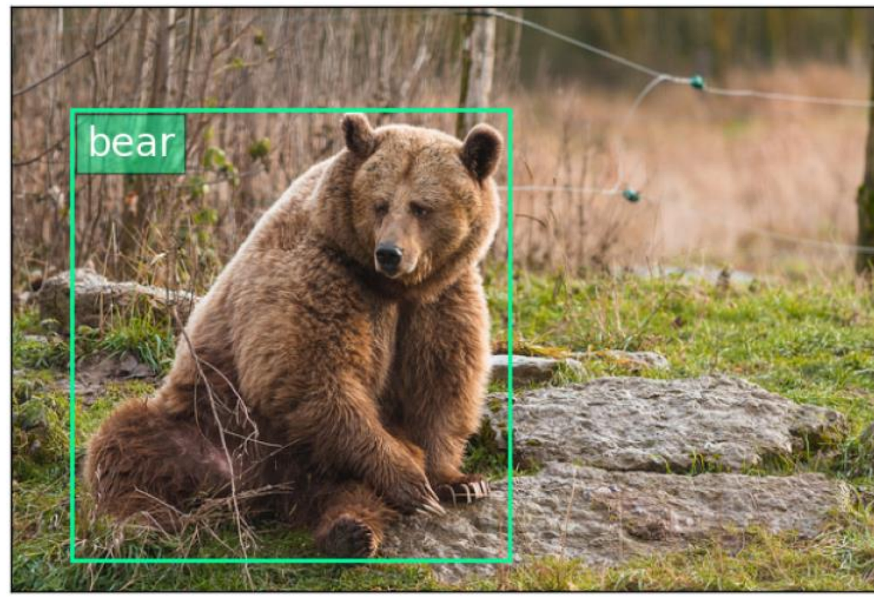
- **Problem & Motivation**
- Approach: Semantic consistency
- Approach: Re-optimization
- Results
- Case studies
- Conclusion

Problem

(a) Detecting `cat` and `table`



(b) Detecting `bear`



Motivation

- Most existing methods
  - only utilize image features
  - Ignoring external knowledge: common sense or domain specific expertise
- Example knowledge
  - Cat sits on table 
  - Bear sits on table 

Outline

- Problem & Motivation
- **Approach: Semantic consistency**
- Approach: Overall framework
- Results
- Case studies
- Conclusion

Knowledge incorporation through semantic consistency

- Semantic consistency matrix *S*
  - $S_{l,l'}$: how related concepts l, l' are
  - $S_{\text{cat,table}} \gg S_{\text{bear,table}}$

- Object detection probability

Semantic consistency	**Probability in the same image**	**Example** (*b, b'* are bounding boxes in the same image)
Large	Comparable	$\lvert p(\text{cat}\mid b) - p(\text{table}\mid b') \rvert \approx 0$
Small	Different	$\lvert p(\text{bear}\mid b) - p(\text{table}\mid b') \rvert \gg 0$

$P_{b,l} \equiv p(l|b)$: probability of concept l given bounding box b

Constructing semantic consistency: Frequency-based

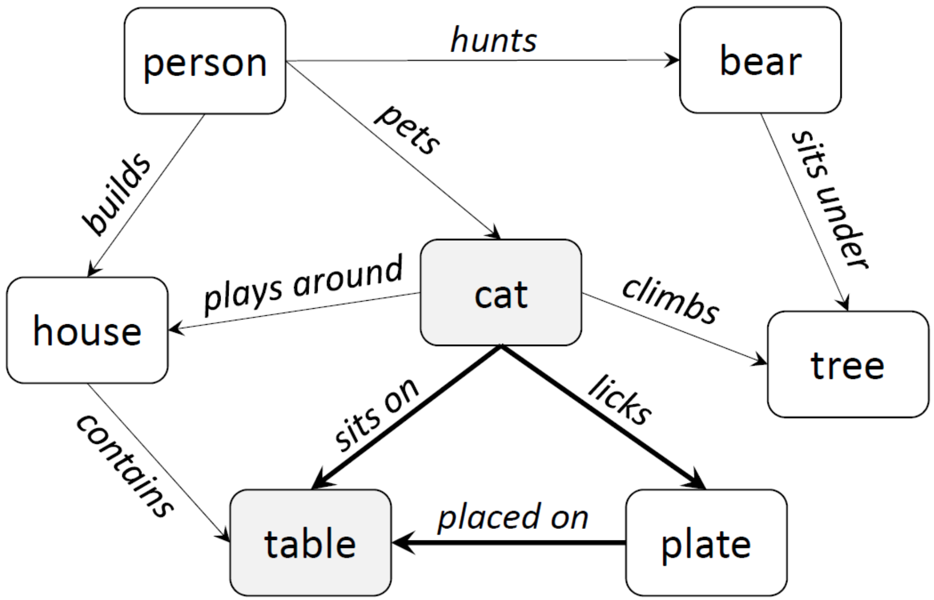
- Co-occurrence frequency based on training set

$$S_{\ell,\ell'} = \max\left(\boxed{\log \frac{n(\ell,\ell')N}{n(\ell)n(\ell')}}, 0\right)$$

Pointwise mutual information

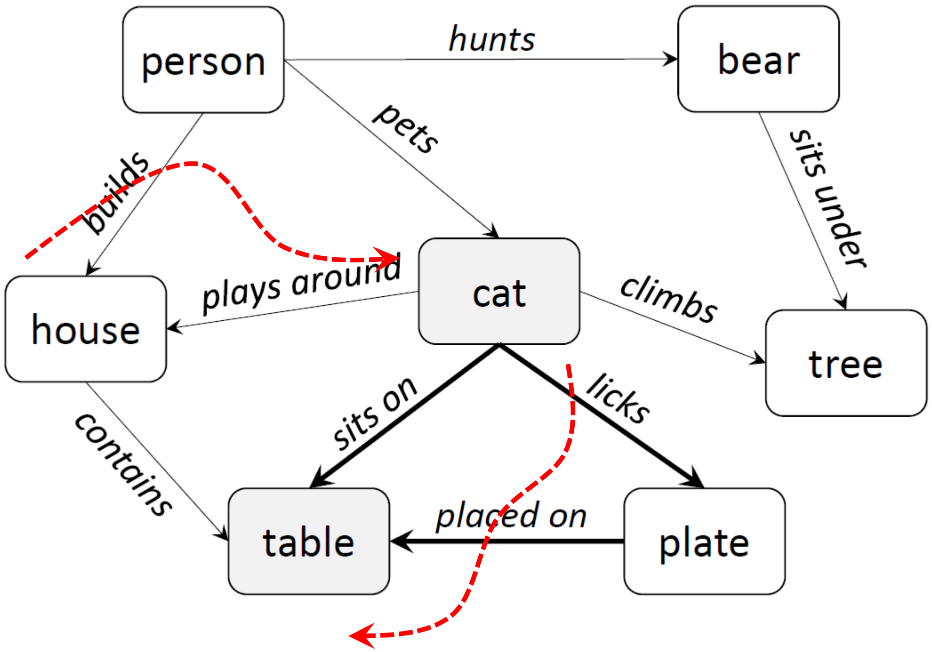
- Weakness:
  - Cannot generalize to new co-occurrences
  - The need of a training set

Constructing semantic consistency: knowledge graph (KG) based



- Generalization: indirect relationships (person-plate)
- Robustness: multiple relationships (cat-table, cat-plate-table)

Constructing semantic consistency: knowledge graph (KG) based



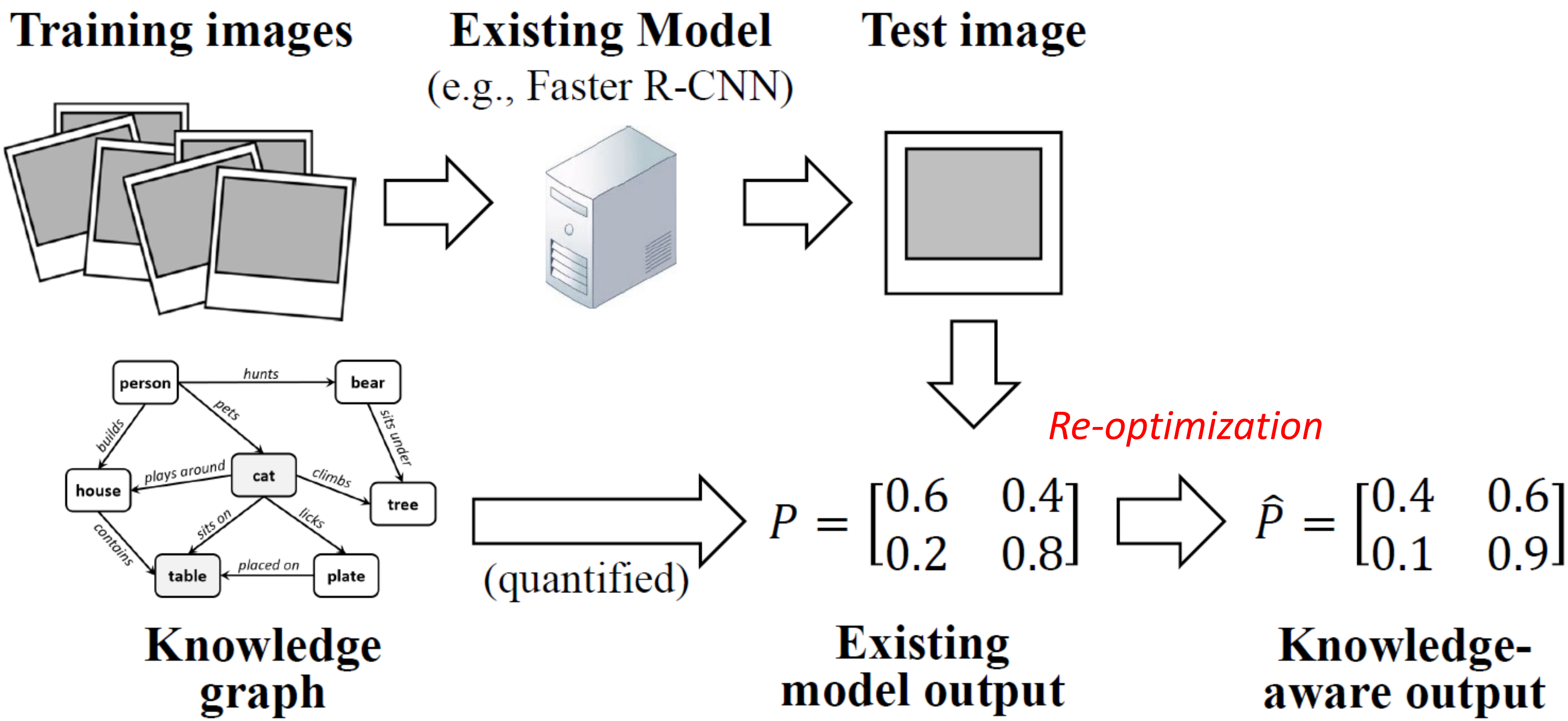
A random walk $v_0, v_1, \ldots, v_t$ with restart

$$\lim_{t \to \infty} P(v_t = l' | v_0 = l)$$

Outline

- Problem & Motivation
- Approach: Semantic consistency
- **Approach: Re-optimization**
- Results
- Case studies
- Conclusion

Overall Framework



Approach: Re-optimization

$$E(\widehat{P}) = (1-\epsilon)\sum_{b=1}^{B}\sum_{\substack{b'=1\\ b'\neq b}}^{B}\sum_{\ell=1}^{L}\sum_{\ell'=1}^{L}\boxed{S_{\ell,\ell'}\left(\widehat{P}_{b,\ell}-\widehat{P}_{b',\ell'}\right)^2}$$

$$+\,\epsilon\sum_{b=1}^{B}\sum_{\ell=1}^{L} B\|S_{\ell,*}\|_1\boxed{\left(\widehat{P}_{b,\ell}-P_{b,\ell}\right)^2}$$

Bounding box $b \in \{1,2,\dots,B\}$
Object labels $l \in \{1,2,\dots,L\}$
$S_{l,l'}$: semantic consistency between l, l'
$P_{b,l}$: original probability of label l given bounding box b
$\widehat{P}_{b,l}$: re-optimized probability of label l given bounding box b

Weakness of the proposed approach

- The re-optimization step based on knowledge is a ***post processing*** step
- Independent of the object detection model
- Cannot feedback into the detection model (eg. through backpropagation)
- Thesis of this paper: only intends to ***demonstrate the benefits of utilizing knowledge*** in deep learning models

Outline

- Problem & Motivation
- Approach: Semantic consistency
- Approach: Re-optimization
- **Results**
- Case studies
- Conclusion

Results – MSCOCO dataset

	mAP	**Recall**		**Recall**@100 by area		
	@100	@100	@10	small	medium	large
minival-4k						
FRCNN	24.5	35.9	35.2	14.2	41.5	55.6
KF-500	24.4	37.1	35.6	14.3	42.8	57.3
KF-All	24.5	37.9	36.2	**14.6**	43.9	58.6
KG-CNet	24.4	**38.9**	**36.6**	14.4	**45.2**	**60.0**
test-dev						
FRCNN	24.2	34.6	34.0	12.0	38.5	54.4
KF-500	24.3	37.4	35.9	13.7	42.1	58.0
KF-All	24.3	38.2	36.4	14.2	43.0	59.2
KG-CNet	24.2	**39.2**	**36.9**	**14.5**	**44.0**	**60.7**
test-std						
FRCNN	24.2	34.7	34.1	11.5	38.9	54.4
KG-CNet	24.1	**39.2**	**37.0**	**14.2**	**44.4**	**60.5**

Up to 4.6% in recall

FRCNN: Faster RCNN (knowledge-free)
KF-500: Frequency based knowledge (500 images)
KF-All: Frequency based knowledge (all)
KG-CNet: knowledge graph based on ConceptNet

Results – PASCAL VOC dataset

	mAP @100	**Recall**@100 by concepts																				
		all	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
FRCNN	66.5	81.9	76.1	89.0	74.3	73.4	64.6	89.7	85.8	90.5	69.0	88.9	85.4	91.6	92.0	85.2	82.4	60.8	83.1	89.1	84.4	82.1
KF-500	66.6	83.8	80.0	91.7	**79.1**	**76.0**	67.0	89.7	88.8	92.5	69.7	92.6	85.9	90.8	**94.0**	86.8	82.0	59.6	87.2	90.0	89.7	82.8
KF-All	66.5	84.6	**80.7**	**93.5**	**79.1**	**76.0**	**67.6**	**90.1**	88.8	**93.6**	68.1	**93.0**	**86.9**	**94.1**	93.1	**89.5**	83.1	65.4	**88.0**	89.1	**90.1**	81.8
KG-CNet	66.6	**85.0**	80.4	92.3	78.6	**76.0**	**67.6**	**90.1**	**89.1**	92.2	**74.2**	**93.0**	86.4	93.0	92.2	88.6	**87.7**	**66.9**	87.6	**90.4**	89.7	**83.4**

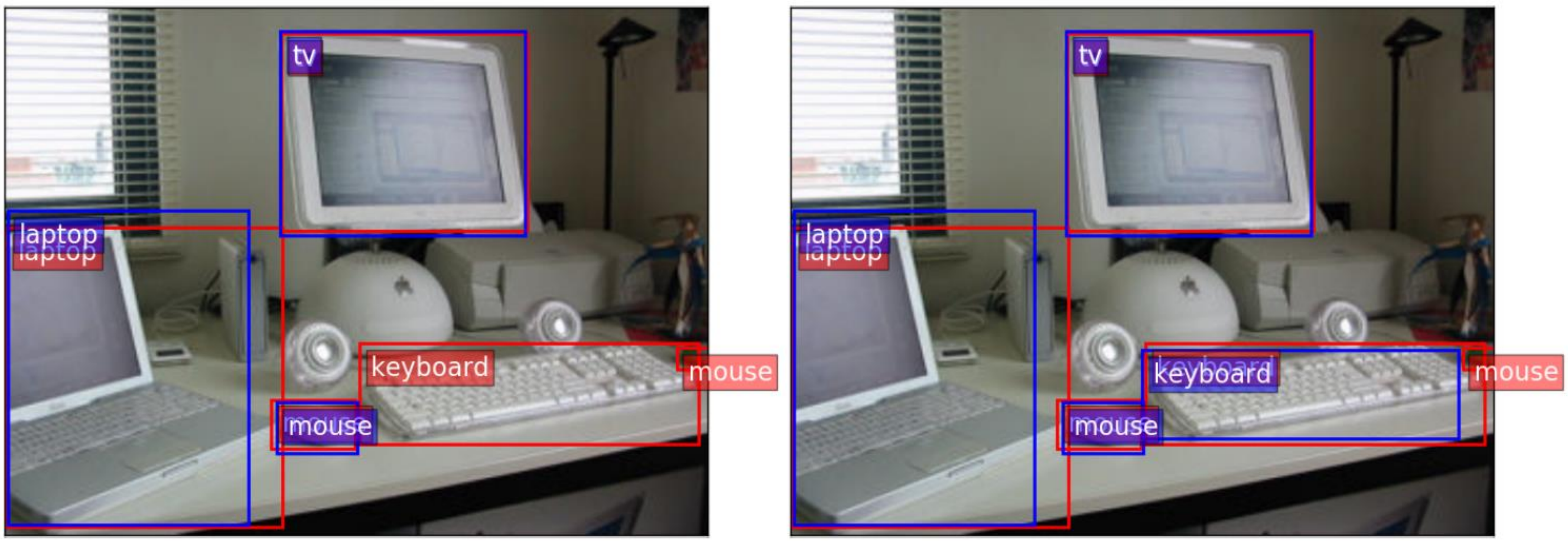
Up to 3.1% in recall

Outline

- Problem & Motivation
- Approach: Semantic consistency
- Approach: Re-optimization
- Results
- **Case studies**
- Conclusion

Case study – office scene

(a) Office scene: FRCNN (left) fails to detect `keyboard`, but KG-CNet (right) does due to the presence of `laptop`.



groundtruth

detected

$S(\text{keyboard}, \text{laptop}) \approx$ 135x median value

Case study – outdoor scene

(b) Outdoor scene: FRCNN (left) fails to detect `surfboard`, but KG-CNet (right) does due to the presence of `person`.

groundtruth

detected

$S(\text{surfboard}, \text{person}) \approx$ 5x median value

Conclusion & future work

- External knowledge is helpful
- Complement existing methods to achieve better prediction results
- Next step: end-to-end learning with knowledge