How a General-Purpose Commonsense Ontology can Improve Performance of Learning-Based Image Retrieval

Rodrigo Toro Icarte[†] Jorge A. Baier[‡§] Cristian Ruz[‡] Alvaro Soto[‡]

[†]University of Toronto [‡]Pontificia Universidad Católica de Chile

[§]Chilean Center for Semantic Web Research

[†]rntoro@cs.toronto.edu, [‡]{jabaier, cruz, asoto}@ing.puc.cl

Motivation

Computer Vision (CV)

Simple → Complex

- Image classification, captioning, Q&A, ...

  Learning everything from examples **does not scale**
- 96.4% classification → 32.2% captioning

Prior knowledge can fill the holes in our datasets

- Small hand-crafted ontologies, free form text (e.g. Wikipedia), and lexical ontologies (e.g. WordNet)

What about commonsense ontologies, such as ConceptNet?

ConceptNet (CN)

Format: Concept$_1$ — Relation type → Concept$_2$
Relation types: AtLocation, HasProperty, IsA, UsedFor, ...
Examples:

- desk — RelatedTo → office
- computer — AtLocation → office
- office — UsedFor → work
- ... and 8 million more

Great source of prior knowledge for CV

- Millions of assertions
- Key knowledge for computers
- Rich source of commonsense knowledge
- Simple to use

... we do not know how to exploit it

Task	w/o CN	w/ CN	CN gain
Image Tagging [8]	7.3%	7.6%	**0.3%**
Video Retrieval [2]	3.9%	3.1%	**-0.8%**
Image Riddles [1]	68.0%	68.7%	**0.7%**

Sentence Based Image Retrieval

Task: Rank n images according to their *relevance* with respect to a text query

Example

"a woman in a chef coat holding bread loaves"

A Baseline for Image Retrieval

Given a text query t and an image I, we define:

$$\mathrm{MIL}(t,I) = \prod_{w \in V \cap S_t} P(w|I)$$

where:

- S_t is the set of words in t
- V is a set of detectable words
- $P(w|I)$ is the score of detector w over I

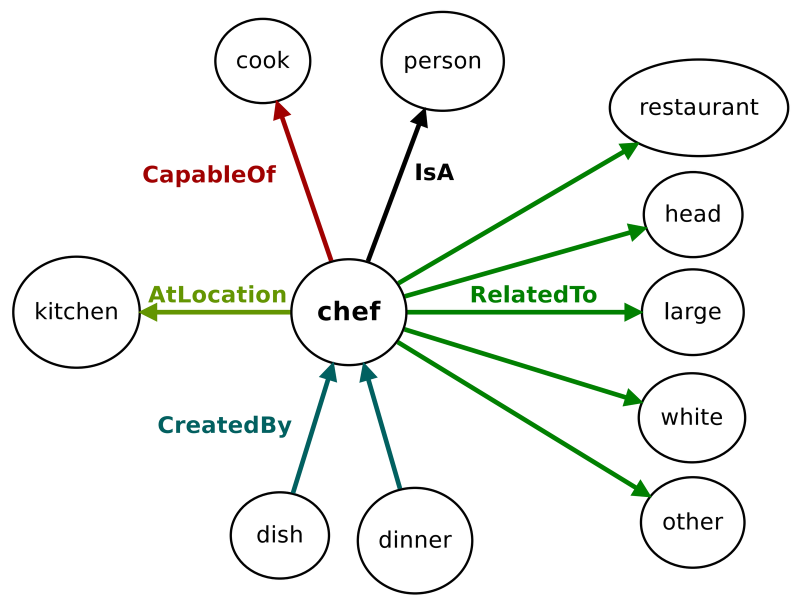
We used the 1000 Concept detectors trained by Fang et al. [3]

Example

t = "a woman in a chef coat holding bread loaves"

1 2 3 4 ... 512

CN-based Detector Enhancement



Main idea:

Augment the set of detectors using CN

CN score

For each word $w \notin V$, we define:

$$\mathrm{CN}_{\mathrm{Agg}}(w) = \underset{w_i \in \{V \cap \mathrm{cn}(w)\}}{\mathrm{Agg}} P(w_i|I)$$

where:

- $\mathrm{Agg} \in \{\min, \mathrm{avg}, \max\}$
- $\mathrm{cn}(w)$ is the set of neighbors of w in CN

Example

w_i	$P(w_i\|I)$	w_i	$P(w_i\|I)$
kitchen	0.996	dish	0.126
cook	0.796	white	0.091
restaurant	0.374	other	0.043
person	0.340	dinner	0.023
large	0.152	head	0.003

$\mathrm{CN}_{\min}(\mathrm{chef}) = 0.003$
$\mathrm{CN}_{\mathrm{avg}}(\mathrm{chef}) = 0.294$
$\mathrm{CN}_{\max}(\mathrm{chef}) = 0.996$

Results

Database ⊂ COCO 5K	r@1	r@5	r@10	median rank	mean rank
Baseline					
MIL	**13.2**	33.4	45.2	**13**	82.2
CN					
$\mathrm{CN}_{\min}$	12.2	31.4	43.4	15	77.0
$\mathrm{CN}_{\mathrm{avg}}$	**13.2**	**33.7**	**46.0**	**13**	**66.3**
$\mathrm{CN}_{\max}$	12.2	32.1	44.1	14	73.0
CN gain	**0.0%**	**0.3%**	**0.8%**	**0**	**15.9**

CN + ESPGAME score

For each word $w \notin V$, we define:

$$\mathrm{CNE}_{\mathrm{Agg}}(w) = \underset{w_i \in \{V \cap \mathrm{cn}(w)\}}{\mathrm{Agg}} P(w|w_i) \cdot P(w_i|I) + P(w|\neg w_i) \cdot P(\neg w_i|I)$$

Estimate $P(w|w_i)$ and $P(w|\neg w_i)$ from **ESPGAME**

Example

$P(\mathrm{chef}|I) = P(\mathrm{chef}|\mathrm{cook}, I) P(\mathrm{cook}|I) +$
$\qquad P(\mathrm{chef}|\neg\mathrm{cook}, I) P(\neg\mathrm{cook}|I)$
$P(\mathrm{chef}|I) = P(\mathrm{chef}|\mathrm{cook}, I) \cdot 0.8 + P(\mathrm{chef}|\neg\mathrm{cook}, I) \cdot 0.2$
$P(\mathrm{chef}|I) \approx P(\mathrm{chef}|\mathrm{cook}) \cdot 0.8 + P(\mathrm{chef}|\neg\mathrm{cook}) \cdot 0.2$
$P(\mathrm{chef}|I) \approx 0.1413 \cdot 0.8 + 0.0003 \cdot 0.2$
$P(\mathrm{chef}|I) \approx 0.112$

Results

Database ⊂ COCO 5K	r@1	r@5	r@10	median rank	mean rank
Baseline					
MIL	13.2	33.4	45.2	13	82.2
CN + ESPGAME					
$\mathrm{CNE}_{\min}$	14.3	34.6	46.6	**12**	68.3
$\mathrm{CNE}_{\mathrm{avg}}$	**14.6**	35.6	48.0	**12**	61.2
$\mathrm{CNE}_{\max}$	14.3	**35.9**	**48.2**	**12**	**60.6**
CN gain	**1.4%**	**2.5%**	**3.0%**	**1**	**21.6**

Example

t = "a woman in a chef coat holding bread loaves"

1 2 3 4 ... 35

More Experiments

ESPGAME alone

Database ⊂ COCO 5K	r@1	r@5	r@10	median rank	mean rank
Baseline					
MIL	13.2	33.4	45.2	**13**	82.2
ESPGAME					
$\mathrm{ESP}_{\min}$	12.6	30.7	41.1	17	122.4
$\mathrm{ESP}_{\mathrm{avg}}$	**13.6**	**34.2**	**46.2**	**13**	69.0
$\mathrm{ESP}_{\max}$	13.5	33.7	45.7	**13**	**66.2**
ESP gain	**0.4%**	**0.8%**	**1.0%**	**0**	**16.0**

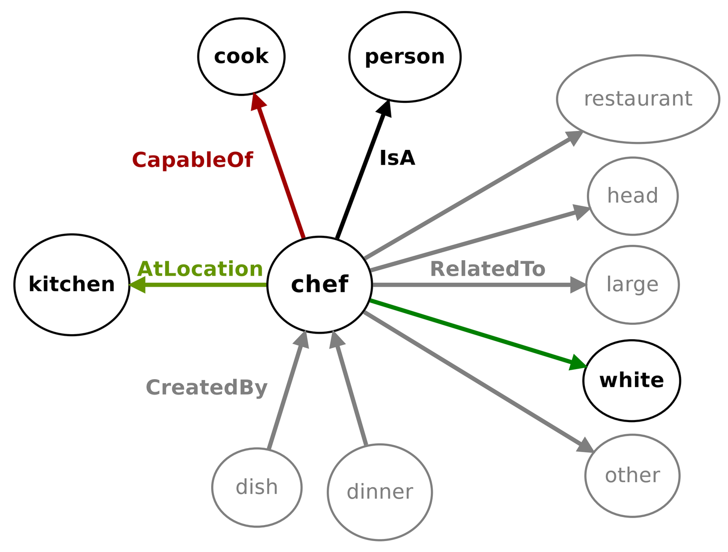
COCO 5K

Database COCO 5K	r@1	r@5	r@10	median rank	mean rank
GMM+HGLMM [5]	10.8	28.3	40.1	17	49.3
BRNN [4]	10.7	29.6	42.2	14	–
MIL (our baseline)	15.7	37.8	50.5	10	53.6
$\mathbf{CNE}_{\mathbf{MAX}}$ (our method)	16.2	39.1	51.9	10	44.4
LVQ [6]	16.7	40.5	53.8	–	–
OE [7]	18.0	–	57.6	7.0	35.9

In the paper

- Zero-shot learning
- COCO 22K

Why is CN helping this time?



Summary

Motivation

- Prior knowledge has a key role in Computer Vision
- ConceptNet (CN) is a rich source of prior knowledge

Previous works

- They suggest that CN sucks
- We don't care, we think CN is cool 😎

Method

- CN for image retrieval... sucks!
- CN + ESPGAME for image retrieval... works!

Contribution

- We can exploit commonsense ontologies in Computer Vision, but this knowledge must be filtered in a meaningful way (e.g. using ESPGAME)

References

[1] S. Aditya, Y. Yang, C. Baral, and Y. Aloimonos. Answering image riddles using vision and reasoning through probabilistic soft logic. *arXiv*, 2016.

[2] M. de Boer, K. Schutte, and W. Kraaij. Knowledge based query expansion in complex multimedia event detection. *Multimedia Tools and Applications*, 75(15), 2016.

[3] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, et al. From captions to visual concepts and back. In *CVPR*, 2015.

[4] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[5] B. Klein, G. Lev, G. Sadeh, and L. Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *CVPR*, 2015.

[6] X. Lin and D. Parikh. Leveraging visual question answering for image-caption ranking. In *ECCV*, 2016.

[7] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun. Order-embeddings of images and language. In *ICLR*, 2016.

[8] L. Xie and X. He. Picture tags and world knowledge: learning tag relations from visual semantic sources. In *ACMMM*, 2013.