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

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

[†]University of Toronto

[‡]Pontificia Universidad Católica de Chile

[§]Chilean Center for Semantic Web Research

IJCAI 2017

Summary

Motivation

Computer Vision

Simple $\rightarrow$ Complex

Computer Vision

Image classification: Kitchen

Computer Vision

Simple $\rightarrow$ Complex

Image classification: Kitchen
Image captioning: "a woman in a chef coat holding bread loaves"

Computer Vision

Simple $\rightarrow$ Complex

Image classification: Kitchen
Image captioning: "a woman in a chef coat holding bread loaves"
Image Q&A:
Q: What is the chef holding?
A: bread loaves

Computer Vision

Simple → Complex

Computer Vision

Image classification: Kitchen
Image captioning: "a woman in a chef coat holding bread loaves"
Image Q&A:
Q: What is the chef holding?
A: bread loaves

Computer Vision

- Simple → Complex
- Learning everything from examples

Computer Vision

Image classification: Kitchen
Image captioning: “a woman in a chef coat holding bread loaves”
Image Q&A:
Q: What is the chef holding?
A: bread loaves

Computer Vision

Simple → Complex

Learning everything from examples

It does not scale (96.4% classification → 32.2% captioning)

Image classification: Kitchen
Image captioning: "a woman in a chef coat holding bread loaves"
Image Q&A:
Q: What is the chef holding?
A: bread loaves

Computer Vision

Simple → Complex

Learning everything from examples

It does not scale (96.4% classification → 32.2% captioning)

Idea: Prior knowledge can fill the holes in our datasets.

Prior Knowledge in Computer Vision

Main research trends

- Small hand-crafted ontologies
- Free form text (e.g. Wikipedia)
- Lexical ontologies (e.g. WordNet)

Prior Knowledge in Computer Vision

Main research trends

- Small hand-crafted ontologies
- Free form text (e.g. Wikipedia)
- Lexical ontologies (e.g. WordNet)

What about commonsense ontologies, such as ConceptNet?

ConceptNet (CN)

CN is a commonsense ontology.

ConceptNet (CN)

CN is a commonsense ontology.

Format

$\text{Concept}_1 - \text{Relation type} \rightarrow \text{Concept}_2$

Relation types

AtLocation, HasProperty, IsA, SimilarSize, UsedFor, CapableOf, ...

ConceptNet

ConceptNet (CN)

CN is a commonsense ontology.

Format

Concept_1 – Relation type $\rightarrow$ Concept_2

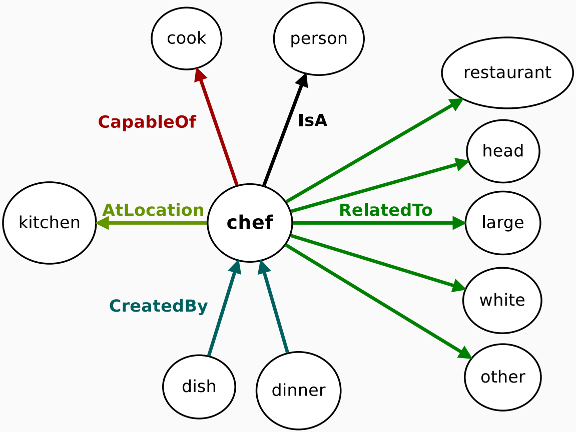
Relation types

AtLocation, HasProperty, IsA, SimilarSize, UsedFor, CapableOf, ...

Examples

desk – RelatedTo $\rightarrow$ office
computer – AtLocation $\rightarrow$ office
office – UsedFor $\rightarrow$ work
... and 8 million more

ConceptNet



cook
person
restaurant
CapableOf
IsA
head
kitchen
AtLocation
chef
RelatedTo
large
white
CreatedBy
dish
dinner
other

ConceptNet

CN is a great source of prior knowledge for Computer Vision.

ConceptNet

CN is a great source of prior knowledge for Computer Vision.

- CN has millions of assertions (vs hand-crafted ontologies)

ConceptNet

CN is a great source of prior knowledge for Computer Vision.

- CN has millions of assertions (vs hand-crafted ontologies)
- CN provides key knowledge to computers (vs Wikipedia)

ConceptNet

CN is a great source of prior knowledge for Computer Vision.

- CN has millions of assertions (vs hand-crafted ontologies)
- CN provides key knowledge to computers (vs Wikipedia)
- CN is a rich source of commonsense knowledge (vs WordNet)

ConceptNet

CN is a great source of prior knowledge for Computer Vision.

- CN has millions of assertions (vs hand-crafted ontologies)
- CN provides key knowledge to computers (vs Wikipedia)
- CN is a rich source of commonsense knowledge (vs WordNet)
- CN is simple to use (vs CYC)

Summary

Motivation

Summary

Motivation

Prior knowledge has a key role in Computer Vision.

Summary

Motivation

Prior knowledge has a key role in Computer Vision.

ConceptNet (CN) is a rich source of prior knowledge.

Summary

Motivation

Prior knowledge has a key role in Computer Vision.

ConceptNet (CN) is a rich source of prior knowledge.

Previous works

ConceptNet in Computer Vision

Task	w/o CN	w/ CN	CN gain
Image Tagging Xie and He (2013)	7.3%	7.6%	**0.3%**
Video Retrieval de Boer et al. (2016)	3.9%	3.1%	**-0.8%**
Image Riddles Aditya et al. (2016)	68.0%	68.7%	**0.7%**

More examples: Bicocchi et al. (2012), Le et al. (2013), others

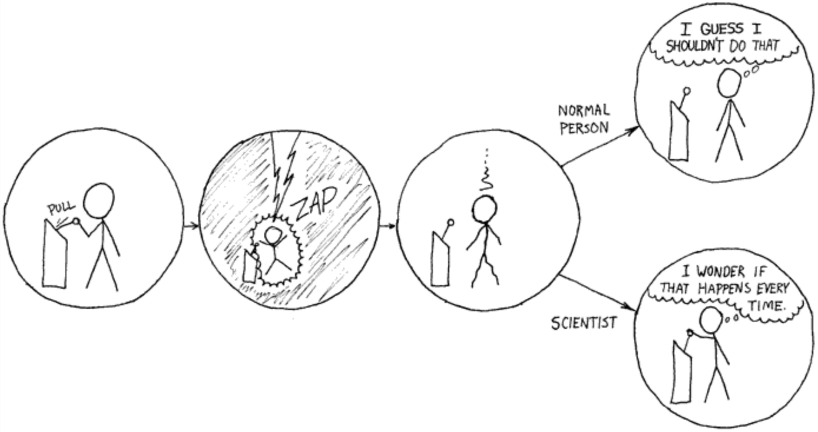
ConceptNet in Computer Vision

Task	w/o CN	w/ CN	CN gain
Image Tagging Xie and He (2013)	7.3%	7.6%	**0.3%**
Video Retrieval de Boer et al. (2016)	3.9%	3.1%	**-0.8%**
Image Riddles Aditya et al. (2016)	68.0%	68.7%	**0.7%**

More examples: Bicocchi et al. (2012), Le et al. (2013), others

... but we wanted to give CN another try.

… because we are scientists



Source: https://xkcd.com/242/

Summary

Motivation

Prior knowledge has a key role in Computer Vision.

ConceptNet (CN) is a rich source of prior knowledge.

Previous works

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.

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 🤓

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

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...

Sentence Based Image Retrieval

"a woman in a chef coat holding bread loaves"






Sentence Based Image Retrieval

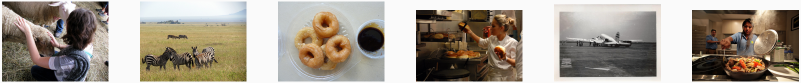
"a woman in a chef coat holding bread loaves"

Rank *n* images according to their *relevance*
with respect to a natural language query.

Sentence Based Image Retrieval

“a woman in a chef coat holding bread loaves”



Rank *n* images according to their *relevance* with respect to a natural language query.

Baseline

Baseline

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

Baseline

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

Prob	Concept
0.996	kitchen
0.920	preparing
0.800	food
0.796	cooking
0.590	making
...	...
0.236	woman

Baseline

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

Prob	Concept
0.996	kitchen
0.920	preparing
0.800	food
0.796	cooking
0.590	making
...	...
0.236	woman

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

Baseline

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

Prob	Concept
0.996	kitchen
0.920	preparing
0.800	food
0.796	cooking
0.590	making
...	...
0.236	woman

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

$\mathbf{MIL}(t, I) = \mathrm{P}(\text{woman}|I) \cdot \mathrm{P}(\text{coat}|I) \cdot \mathrm{P}(\text{holding}|I) \cdot \mathrm{P}(\text{bread}|I)$

Baseline

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


1


2


3


4

...


512

$$\textbf{MIL}(t, I) = \mathrm{P}(\text{woman}|I) \cdot \mathrm{P}(\text{coat}|I) \cdot \mathrm{P}(\text{holding}|I) \cdot \mathrm{P}(\text{bread}|I)$$

Baseline

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

1

2

3

4

...

512

$$\textbf{MIL}(t, I) = \mathrm{P}(\text{woman}|I) \cdot \mathrm{P}(\text{coat}|I) \cdot \mathrm{P}(\text{holding}|I) \cdot \mathrm{P}(\text{bread}|I)$$

What are the limitations of this approach?

Baseline

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


1


2


3


4

...

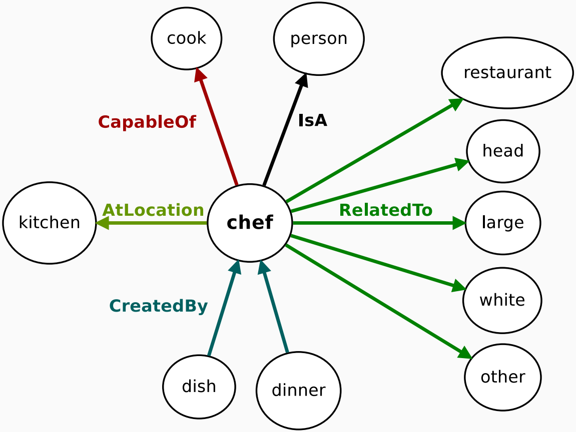

512

$$\textbf{MIL}(t, I) = \mathrm{P}(\text{woman}|I) \cdot \mathrm{P}(\text{coat}|I) \cdot \mathrm{P}(\text{holding}|I) \cdot \mathrm{P}(\text{bread}|I)$$

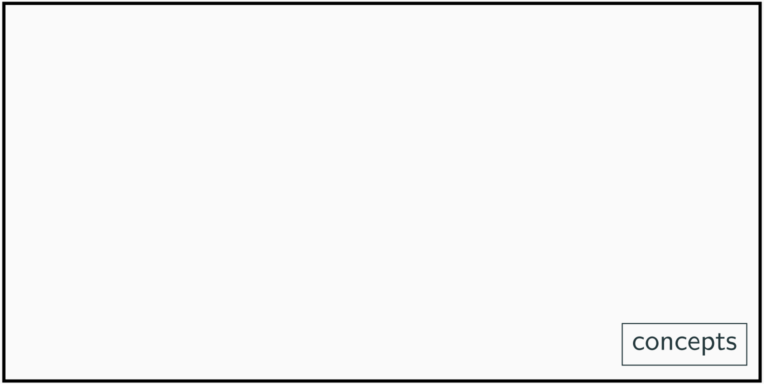
What are the limitations of this approach?

How can we detect a *chef* without a *chef* detector?

Baseline + ConceptNet

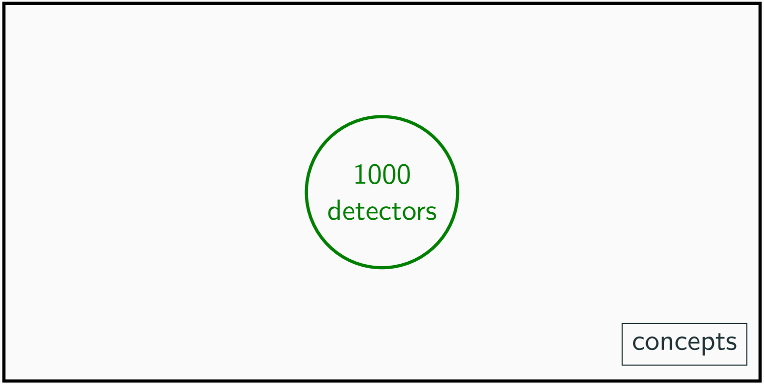


Baseline + ConceptNet



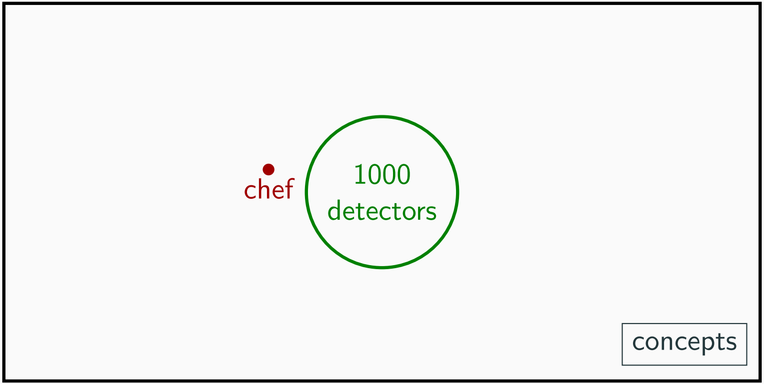
Idea: Augment the set of detectors using CN

Baseline + ConceptNet



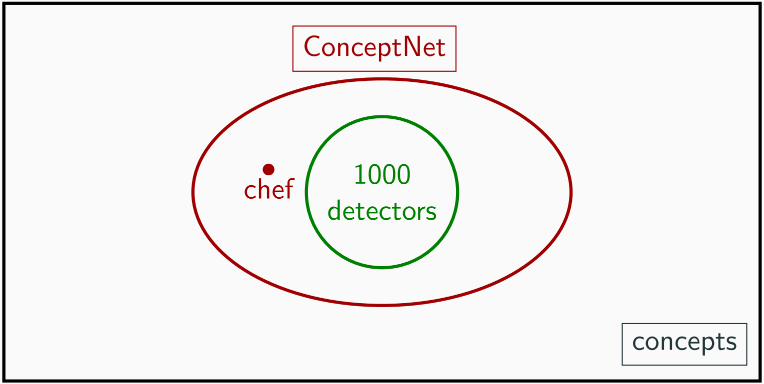
Idea: Augment the set of detectors using CN

Baseline + ConceptNet



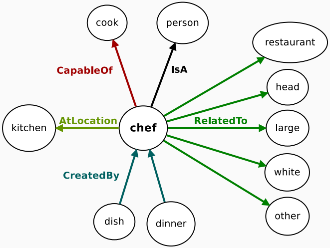
Idea: Augment the set of detectors using CN

Baseline + ConceptNet

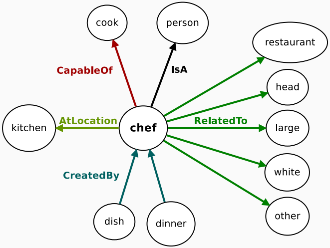


Idea: Augment the set of detectors using CN

CN Score

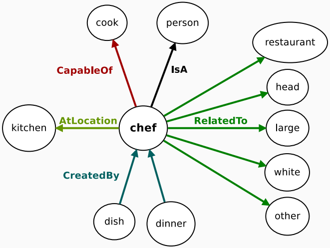

cook
person
restaurant
CapableOf
IsA
head
kitchen
AtLocation
chef
RelatedTo
large
white
CreatedBy
dish
dinner
other

CN Score



Word	Prob	Word	Prob
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

CN Score



Word	Prob	Word	Prob
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

$$\mathrm{CN}_{\mathrm{MIN}}(\mathrm{chef}) = 0.003$$
$$\mathrm{CN}_{\mathrm{AVG}}(\mathrm{chef}) = 0.294$$
$$\mathrm{CN}_{\mathrm{MAX}}(\mathrm{chef}) = 0.996$$

CN Score

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					
CN_{MIN}	12.2	31.4	43.4	15	77.0
CN_{AVG}	**13.2**	**33.7**	**46.0**	**13**	**66.3**
CN_{MAX}	12.2	32.1	44.1	14	73.0
CN Gain	**0.0%**	**0.3%**	**0.8%**	**0**	**15.9**

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...

Summary

Motivation

Prior knowledge has a key role in Computer Vision.

ConceptNet (CN) is a rich source of prior knowledge.

Previous works

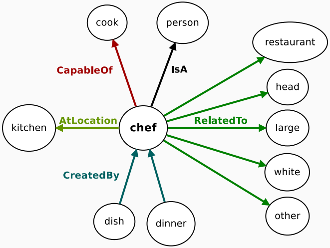
They suggest that CN sucks.

We don't care, we think CN is cool 🤓

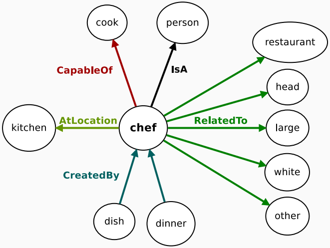
Method

CN for image retrieval... sucks!

CN + ESPGAME Score

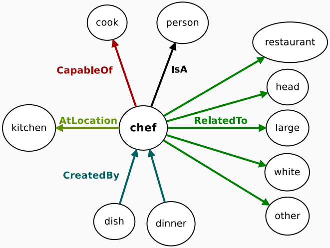


CN + ESPGAME Score



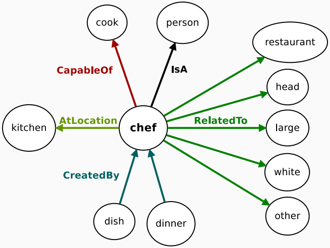
$P_c(\text{chef}|I)$

CN + ESPGAME Score



$$P_c(\text{chef}|I) = P(\text{chef}|\text{cook}, I)P(\text{cook}|I) + P(\text{chef}|\neg\text{cook}, I)P(\neg\text{cook}|I)$$

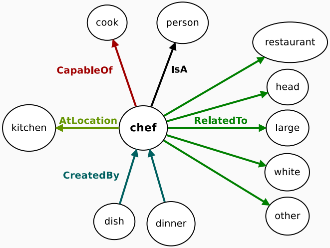
CN + ESPGAME Score



$$P_c(\text{chef}|I) = P(\text{chef}|\text{cook}, I)P(\text{cook}|I) + P(\text{chef}|\neg\text{cook}, I)P(\neg\text{cook}|I)$$

$$P_c(\text{chef}|I) = P(\text{chef}|\text{cook}, I) \cdot 0.796 + P(\text{chef}|\neg\text{cook}, I) \cdot 0.204$$

CN + ESPGAME Score

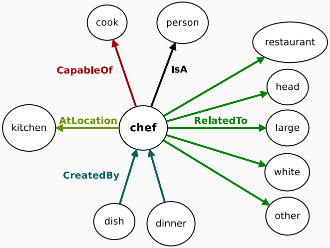


$$P_c(\text{chef}|I) = P(\text{chef}|\text{cook}, I)P(\text{cook}|I) + P(\text{chef}|\neg\text{cook}, I)P(\neg\text{cook}|I)$$

$$P_c(\text{chef}|I) = P(\text{chef}|\text{cook}, I) \cdot 0.796 + P(\text{chef}|\neg\text{cook}, I) \cdot 0.204$$

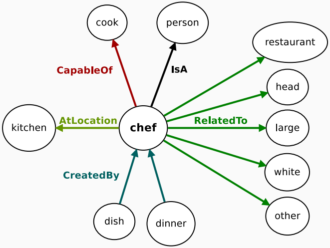
$$P_c(\text{chef}|I) \approx P(\text{chef}|\text{cook}) \cdot 0.796 + P(\text{chef}|\neg\text{cook}) \cdot 0.204$$

CN + ESPGAME Score



$$P_c(\text{chef}|I) = P(\text{chef}|\text{cook}, I)P(\text{cook}|I) + P(\text{chef}|\neg\text{cook}, I)P(\neg\text{cook}|I)$$
$$P_c(\text{chef}|I) = P(\text{chef}|\text{cook}, I) \cdot 0.796 + P(\text{chef}|\neg\text{cook}, I) \cdot 0.204$$
$$P_c(\text{chef}|I) \approx P(\text{chef}|\text{cook}) \cdot 0.796 + P(\text{chef}|\neg\text{cook}) \cdot 0.204$$
$$P_c(\text{chef}|I) \approx 0.1413 \cdot 0.796 + 0.0003 \cdot 0.204$$

CN + ESPGAME Score



$$P_c(\text{chef}|I) = P(\text{chef}|\text{cook}, I)P(\text{cook}|I) + P(\text{chef}|\neg\text{cook}, I)P(\neg\text{cook}|I)$$

$$P_c(\text{chef}|I) = P(\text{chef}|\text{cook}, I) \cdot 0.796 + P(\text{chef}|\neg\text{cook}, I) \cdot 0.204$$

$$P_c(\text{chef}|I) \approx P(\text{chef}|\text{cook}) \cdot 0.796 + P(\text{chef}|\neg\text{cook}) \cdot 0.204$$

$$P_c(\text{chef}|I) \approx 0.1413 \cdot 0.796 + 0.0003 \cdot 0.204$$

$$P_c(\text{chef}|I) \approx 0.112$$

CN + ESPGAME Score

$$
\begin{aligned}
P_{\text{cook}}(\text{chef}|I) &\approx 0.1125 \\
P_{\text{kitchen}}(\text{chef}|I) &\approx 0.0549 \\
P_{\text{dish}}(\text{chef}|I) &\approx 0.0016 \\
P_{\text{person}}(\text{chef}|I) &\approx 0.0011 \\
P_{\text{dinner}}(\text{chef}|I) &\approx 0.0011 \\
P_{\text{head}}(\text{chef}|I) &\approx 0.0009 \\
P_{\text{other}}(\text{chef}|I) &\approx 0.0009 \\
P_{\text{white}}(\text{chef}|I) &\approx 0.0006
\end{aligned}
$$

CN + ESPGAME Score

$$
\begin{aligned}
P_{\text{cook}}(\text{chef}|I) &\approx 0.1125\\
P_{\text{kitchen}}(\text{chef}|I) &\approx 0.0549\\
P_{\text{dish}}(\text{chef}|I) &\approx 0.0016\\
P_{\text{person}}(\text{chef}|I) &\approx 0.0011\\
P_{\text{dinner}}(\text{chef}|I) &\approx 0.0011\\
P_{\text{head}}(\text{chef}|I) &\approx 0.0009\\
P_{\text{other}}(\text{chef}|I) &\approx 0.0009\\
P_{\text{white}}(\text{chef}|I) &\approx 0.0006
\end{aligned}
$$

$$
\begin{aligned}
\text{CNE}_{\text{MIN}}(\text{chef}) &= 0.0006\\
\text{CNE}_{\text{AVG}}(\text{chef}) &= 0.0217\\
\text{CNE}_{\text{MAX}}(\text{chef}) &= 0.1125
\end{aligned}
$$

CN + ESPGAME Score

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					
CNE_{MIN}	14.3	34.6	46.6	**12**	68.3
CNE_{AVG}	**14.6**	35.6	48.0	**12**	61.2
CNE_{MAX}	14.3	**35.9**	**48.2**	**12**	**60.6**
CN Gain	**1.4%**	**2.5%**	**3.0%**	**1**	**21.6**

CN + ESPGAME Score

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

1

2

3

4

...

512

CN + ESPGAME Score

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

    ... 

1 2 3 4 **512**

    ... 

1 2 3 4 **35**

CN + ESPGAME Score

$t =$ "those bagels are plain with nothing on them"


1


2


3


4

...


360

CN + ESPGAME Score

t = "those bagels are plain with nothing on them"


1

2

3

4
...

360


1

2

3

4
...

2

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!

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!

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

Results and Discussion

Database COCO 5K	r@1	r@5	r@10	median rank	mean rank
NeuralTalk (Vinyals et al., 2015)	6.9	22.1	33.6	22	72.2
GMM+HGLMM (Klein et al., 2015)	10.8	28.3	40.1	17	49.3
BRNN (Karpathy and Fei-Fei, 2015)	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 (Lin and Parikh, 2016)	16.7	40.5	53.8	–	–
OE (Vendrov et al., 2016)	18.0	–	57.6	7.0	35.9
	higher is better			lower is better	

Results and Discussion

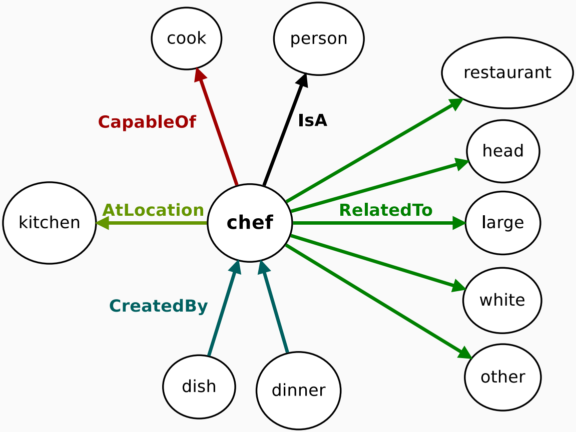
Database $\subset$ COCO 5K	r@1	r@5	r@10	median rank	mean rank
MIL (our baseline)	13.2	33.4	45.2	13	82.2
CNE_{MAX} (our method)	14.3	35.9	48.2	12	60.6
CN Gain	**1.1%**	**2.5%**	**3.0%**	**1**	**21.6**

Results and Discussion

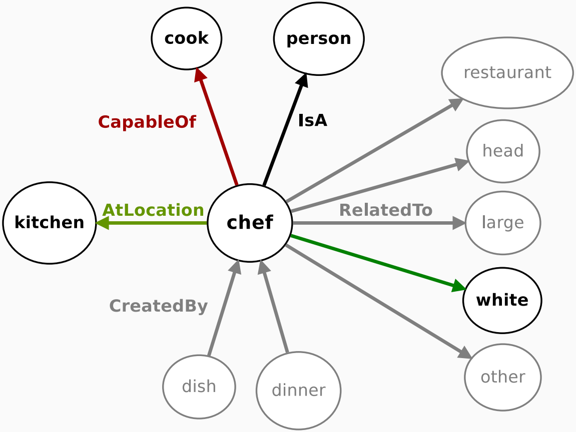
Database $\subset$ COCO 5K	r@1	r@5	r@10	median rank	mean rank
MIL (our baseline)	13.2	33.4	45.2	13	82.2
CNE_{MAX} (our method)	14.3	35.9	48.2	12	60.6
CN Gain	**1.1%**	**2.5%**	**3.0%**	**1**	**21.6**

Why is CN helping this time?

Why is CN helping this time?



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

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

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).

Acknowledgements

Becas Chile — Magister en el Extranjero

FONDECYT 1151018 and 1150328

UofT School of Graduate Studies

Canadian Artificial Intelligence Association

IJCAI travel grant

Conclusion

Our code: https://bitbucket.org/RToroIcarte/cn-detectors

Conclusion

Our code: https://bitbucket.org/RToroIcarte/cn-detectors

If you want to share ideas about commonsense knowledge in Computer Vision, please come to check my poster :)

Conclusion

Our code: https://bitbucket.org/RToroIcarte/cn-detectors

If you want to share ideas about commonsense knowledge in Computer Vision, please come to check my poster :)

Thank you!

References I

Somak Aditya, Yezhou Yang, Chitta Baral, and Yiannis Aloimonos. Answering image riddles using vision and reasoning through probabilistic soft logic. *arXiv preprint arXiv:1611.05896*, 2016.

Nicola Bicocchi, Matteo Lasagni, and Franco Zambonelli. Bridging vision and commonsense for multimodal situation recognition in pervasive systems. In *Pervasive Computing and Communications (PerCom), 2012 IEEE International Conference on*, pages 48–56. IEEE, 2012.

Maaike de Boer, Klamer Schutte, and Wessel Kraaij. Knowledge based query expansion in complex multimedia event detection. *Multimedia Tools and Applications*, 75(15):9025–9043, 2016.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4437–4446, 2015.

Dieu-Thu Le, Jasper RR Uijlings, and Raffaella Bernardi. Exploiting language models for visual recognition. In *EMNLP*, pages 769–779, 2013.

Xiao Lin and Devi Parikh. Leveraging visual question answering for image-caption ranking. In *European Conference on Computer Vision*, pages 261–277. Springer, 2016.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. In *ICLR*, 2016.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

Lexing Xie and Xuming He. Picture tags and world knowledge: learning tag relations from visual semantic sources. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 967–976. ACM, 2013.