Geographic Context for Image Captioning

Sofia Nikiforova — s.nikiforova@uu.nl — Utrecht University

Introduction

Problem: Existing caption generation systems cannot produce **contextualized** captions.

Show, Attend and Tell system [1]:
”a park bench sitting in the middle of a park”

Human-generated:
”A path through Pitshanger Park, near Ealing in the west London suburbs.”

Solution: **image-specific geographic context** added to the standard captioning architecture.

Challenges

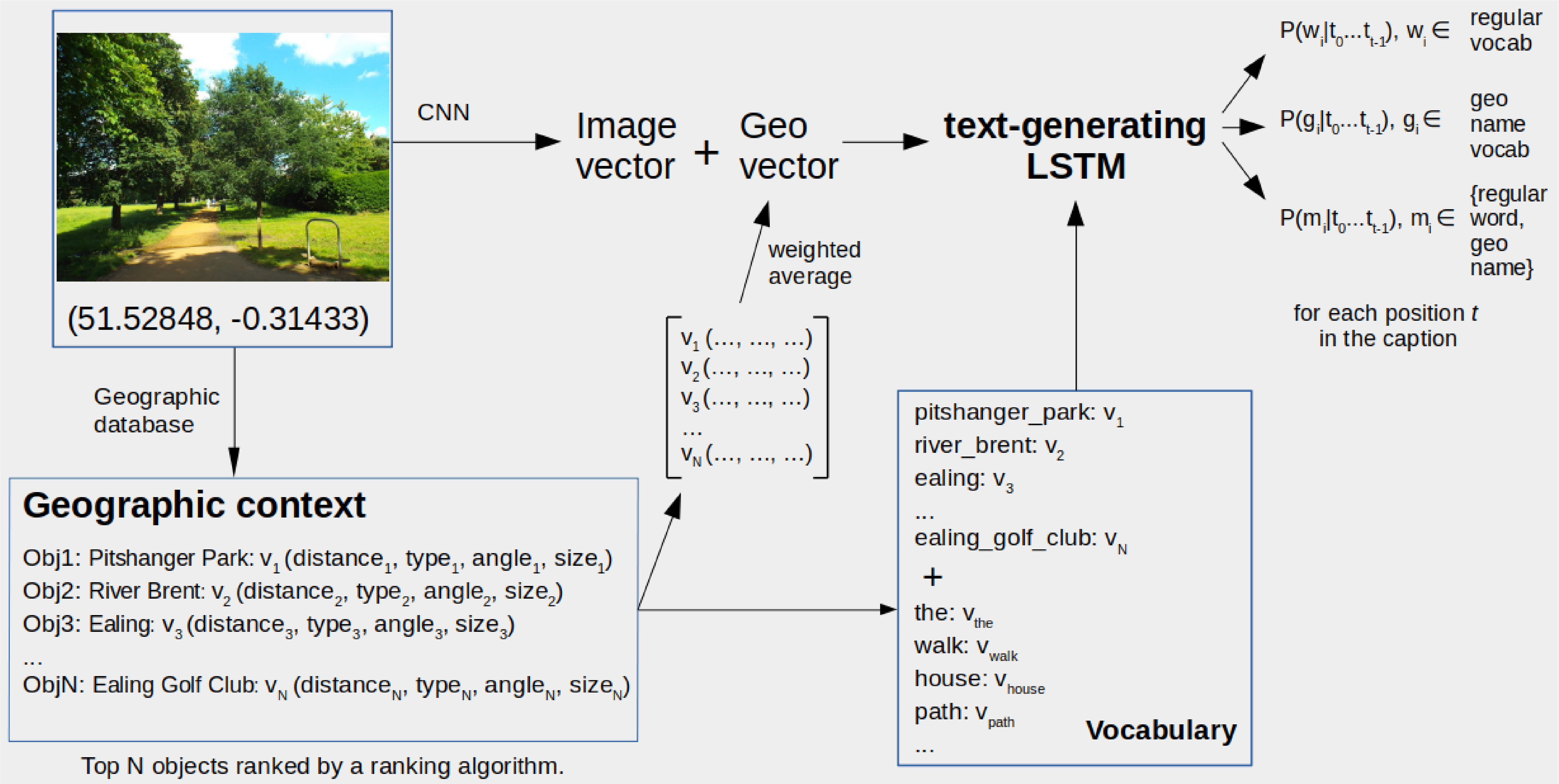
- Which objects from a geographic database should be included in the geographic context?
- What information related to geographic objects should be considered?
- How to make a text generation network generate appropriate geographic names?

Acknowledgements

Work on this paper was funded by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 742204).

Geographic context

Geographic context of an image — a set of relevant objects around the image location.



(51.52848, -0.31433)

CNN

Image vector + Geo vector

text-generating LSTM

$P(w_i|t_0...t_{t-1})$, $w_i \in$ regular vocab

$P(g_i|t_0...t_{t-1})$, $g_i \in$ geo name vocab

$P(m_i|t_0...t_{t-1})$, $m_i \in$ {regular word, geo name}

for each position *t* in the caption

weighted average

v_1 (..., ..., ...)
v_2 (..., ..., ...)
v_3 (..., ..., ...)
...
v_N (..., ..., ...)

Geographic database

Geographic context

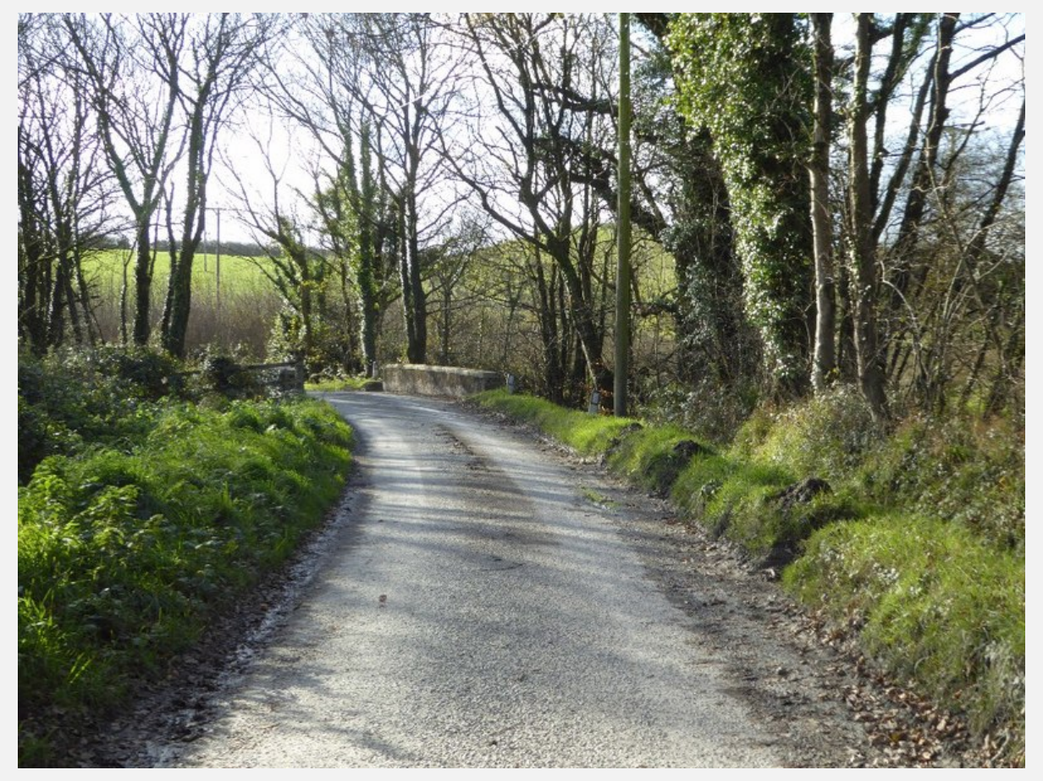
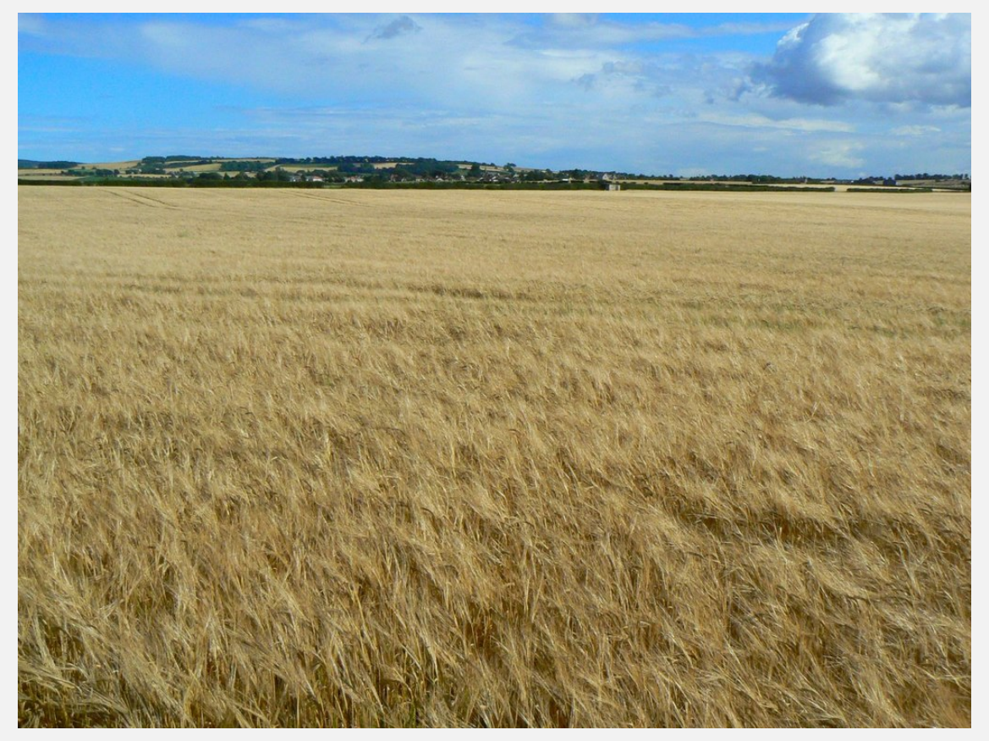
Obj1: Pitshanger Park: v_1 (distance$_1$, type$_1$, angle$_1$, size$_1$)
Obj2: River Brent: v_2 (distance$_2$, type$_2$, angle$_2$, size$_2$)
Obj3: Ealing: v_3 (distance$_3$, type$_3$, angle$_3$, size$_3$)
...
ObjN: Ealing Golf Club: v_N (distance$_N$, type$_N$, angle$_N$, size$_N$)

Top N objects ranked by a ranking algorithm.

pitshanger_park: v_1
river_brent: v_2
ealing: v_3
...
ealing_golf_club: v_N
+
the: v_{the}
walk: v_{walk}
house: v_{house}
path: v_{path}
...

Vocabulary

Examples of generated captions

Image	Human-generated	System	Automatically generated
	Country road crossing a bridge over a tributary of the River Tamar near Moreton Pound.	MSCOCO-trained, no Geo	a long road that has some trees on it
		Geograph-trained, no Geo	the bridge carries a road over a small stream near <unk> farm
		Geograph-trained, plus Geo	the view of the road near river tamar
	A ripening field of barley near Newton of Lathrisk.	MSCOCO-trained, no Geo	a large field with a field in the background
		Geograph-trained, no Geo	a field of wheat to the north of <unk>
		Geograph-trained, plus Geo	a crop of barley to the north of freuchie

Data sources

Source	Source type	Data
Geograph [2]	image hosting website	images with captions and coordinates
OpenStreetMap [3]	geographic database	information about the objects in the geo context

Results

System	Trained on	CIDEr score
Show, Attend and Tell	MSCOCO	6.77
Show, Attend and Tell	Geograph	8.88
Show, Attend and Tell + Geo	Geograph	**18.18**

Conclusion

A caption generation system with an added geographic component produces contextualized captions that are more informative and relevant to the images without compromising the quality of the image description.

References

[1] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[2] Geograph® Britain and Ireland. *http://www.geograph.org.uk/*.

[3] OpenStreetMap. *https://www.openstreetmap.org/*.