

# Geo-Aware Image Caption Generation

Sofia Nikiforova, Tejaswini Deoskar, Denis Paperno, Yoad Winter

Utrecht University, Utrecht, the Netherlands

{s.nikiforova, t.deoskar, d.paperno, y.winter}@uu.nl

## Abstract

Standard image caption generation systems produce generic descriptions of images and do not utilize any contextual information or world knowledge. In particular, they are unable to generate captions that contain references to the *geographic context* of an image, for example, the location where a photograph is taken or relevant geographic objects around an image location. In this paper, we develop a geo-aware image caption generation system, which incorporates geographic contextual information into a standard image captioning pipeline. We propose a way to build an image-specific representation of the geographic context and adapt the caption generation network to produce appropriate geographic names in the image descriptions. We evaluate our system on a novel captioning dataset that contains contextualized captions and geographic metadata and achieve substantial improvements in BLEU, ROUGE, METEOR and CIDEr scores. We also introduce a new metric to assess generated geographic references directly and empirically demonstrate our system’s ability to produce captions with relevant and factually accurate geographic referencing.

## 1 Introduction

Image caption generation is a popular task that aims at producing a natural language description of a given image. A standard neural image captioning system consists of two stages: an “encoder”, a Convolutional Neural Network that encodes the visual features of an image as a vector, and a “decoder”, a language model that is initialized with this vector and generates a caption word by word. Recent research (Lu et al., 2018; Whitehead et al., 2018; Biten et al., 2019) has drawn attention to the fact that the standard approach is insufficient to imitate captions naturally produced by humans. People tend to describe images interpreting them based on context factors and world knowledge, while standard encoder-decoder captioning systems do not take any contextual or world knowledge into account.

One of the aspects that are missing from standard caption generation systems is the ability to produce image descriptions influenced by the geographic context, i.e. geographic objects surrounding the image location. For example, consider the photograph in Figure 1:



Figure 1: An example image.

**Ground Truth:** A path through Pitshanger Park, near Ealing in the west London suburbs

**Automatically generated:** a park bench sitting in the middle of a park

The person who took this photograph<sup>1</sup> captions it referring to the location of the photograph (“Pitshanger Park”) and the most relevant parts of the surroundings (“Ealing”, “west London suburbs”). However, a neural caption generation system (Xu et al., 2015) merely describes the objects in the image (“park bench”, “park”).

In this paper we present geo-aware image captioning, where geographic contextual information is incorporated into the generated captions. The contributions of this paper are as follows:

- We compile a new captioning dataset, the GeoRic dataset<sup>2</sup> (Section 2). Unlike standard captioning datasets, it contains naturally produced contextualized captions, accompanied by geographic metadata, which makes it suitable for training and testing a geo-aware captioning system.
- We propose a way to construct an image-specific geographic context using a geographic database (Section 3). The geographic context contains information about relevant objects around the image location, including their names, which can potentially appear in the caption. Along with the visual features of the image, this geographic context informs the text generation network.
- We propose modifications to the text generation part of the standard captioning pipeline to distinguish between the generation of the regular vocabulary words and the names of geographic entities (Section 4). The geographic names, being very rare in the corpora, have to be represented in a way that does not rely on their distributional properties but rather characterizes them as entities with specific features (e.g. size, distance to the image location).
- We develop a novel metric for measuring the correctness of the generated geographic references, i.e. spatial expressions with geographic names (Section 5). This metric does not rely on comparing the generated references with the ones that appeared in the ground truth captions, which is too restrictive in practice, but instead aims to assess to what extent the system has learned the semantic requirements of the spatial expressions.

Experiments on the GeoRic dataset demonstrate the ability of our geo-aware caption generation system to produce image descriptions that include meaningful and contextually relevant geographic information and show considerable improvements in several image captioning metrics (BLEU, ROUGE, METEOR, CIDEr).

## 2 The GeoRic Dataset

A dataset for geo-aware image captioning has to contain not only images with captions but also the geographic information related to the image locations. Since the generally used image captioning datasets (MSCOCO (Lin et al., 2014), Flickr8k (Hodosh et al., 2013), Flickr30k (Young et al., 2014)) do not include any geographic metadata, we compiled our own dataset – the GeoRic dataset (Geo-aware Rocky Image Captioning).

We gathered data from Geograph, an on-going project that aims to collect photographs of every square kilometer in Great Britain and Ireland. The project’s website<sup>3</sup> currently stores more than 6 million images with naturally produced captions. The website also provides rich metadata for every image including the latitude and longitude coordinates of the image location.

Our GeoRic dataset consists of 29,038 images from the Geograph project website, with captions and location coordinates. We selected captions that are exactly one sentence long (multi-sentence caption generation, although a promising research direction (Mao et al., 2018; Wu et al., 2019), is not addressed in this work) and include at least one spatial expression, such as “near”, “north of”, “across”, etc. (in order to ensure that the captions contain enough geographic referencing). An example entry in the dataset is shown in Table 1.

---

<sup>1</sup><https://www.geograph.org.uk/photo/5802332>

<sup>2</sup>The dataset is publicly available online at <https://rocky.sites.uu.nl/datasets#georic-dataset>

<sup>3</sup><http://www.geograph.org.uk/>


Image	URL	Caption	Latitude	Longitude
	<a href="https://www.geograph.org.uk/photo/3079623">https://www.geograph.org.uk/photo/3079623</a>	Farmland to the west of Burnham Market.	52.93659	0.70376

Table 1: An entry in the GeoRic dataset.

Table 2 contains quantitative statistics of the GeoRic dataset, including overall numbers as well as the numbers for the train, validation and test sets separately.

	In Total	Train	Validation	Test
Number of captions	29,038	21,778	3,630	3,630
Number of tokens	289,028	215,883	36,409	36,736
Average caption length (in tokens)	9.95	9.91	10.03	10.12
Number of geographic named entities per caption	2.05	2.04	2.05	2.06

Table 2: Quantitative statistics of the GeoRic dataset.

Table 3 shows the distribution of the eight most common spatial expressions in the dataset.

	In Total	Train	Validation	Test
Near	9,602	7,122	1,253	1,227
In	7,090	5,292	894	904
Along	2,359	1,814	263	282
Across	1,776	1,343	212	221
North of	2,233	1,715	253	265
South of	2,058	1,583	254	221
East of	1,267	947	149	171
West of	1,374	1,009	180	185

Table 3: Number of captions per spatial expression.

### 3 Geographic Context

We take a geographic context of an image to be a set of relevant geographic objects around the image location. This geographic context is used for compiling an *image-specific vocabulary* of geographic names as well as for complementing the image representation used by the text generation network (for which it needs to be additionally compressed into a single vector). In this section, we describe the process of constructing, encoding and compressing the geographic context.

#### 3.1 Constructing

In order to obtain a reasonable approximation of the geographic context of a given image (i.e. a set of relevant geographic objects in the area), we use OpenStreetMap<sup>4</sup>, a freely available and highly detailed geographic database. This database stores information about billions of objects: their locations, names, types and various specific attributes (e.g. speed limit for an object of the type “highway”). Using OpenStreetMap, we first identify all geographic objects within the radius of 10 kilometers from the image location (anything further is likely to be less relevant to the image description). An example entry of an object from OpenStreetMap is shown in Table 4.

We develop a ranking algorithm and use it to further restrict the list of objects to the ones that are especially relevant to a particular image and can therefore potentially appear in its caption. The objects

<sup>4</sup><https://www.openstreetmap.org/>

Name	Type	Size, km <sup>2</sup>	Latitude	Longitude
Cambourne	town	7.264	52.219984	-0.070078

Table 4: An OpenStreetMap geographic object.

are ranked according to the estimated probability of them being explicitly mentioned in the caption, which serves as a measure of their relevance. The probability is estimated by a logistic regression model, which takes into account the object features extracted from OpenStreetMap (type, size, distance from the object to the image location). The geographic context  $G$  is then formally defined as the top  $n$  objects ( $o_1 \dots o_n$ ) of this ranked list, with  $n$  as a hyperparameter of the system, which we set at 300.

### 3.2 Encoding

The objects in the geographic context  $G$  are represented in a vector form, which allows them to be accessed by a caption generation network. These vector representations are mapped to the names of the objects and used for selecting the geographic name to generate in the caption. Each object  $o_i$  in  $G$  is represented by encoding the following geographic features:

- the distance  $d_i$  between the object and the image location;
- the azimuth  $a_i$  of the direction from the object to the image location;
- the object’s size (area)  $s_i$ ;
- the object’s type  $t_i$ .

These features are intended to ensure a valid usage of spatial expressions, e.g. the distance value is needed for the correct usage of the preposition “near”, the azimuth value – for “north of”, “south of”, etc. The features are combined to create “geographic embeddings”, or GeoEmb, for every object  $o_i$  in ( $o_1 \dots o_n$ ).

$$\text{GEOEMB}(o_i) = d_i \vec{w}_d + a_i \vec{w}_a + s_i \vec{w}_s + E_t(t_i) \quad (1)$$

where  $\vec{w}_d$ ,  $\vec{w}_a$ ,  $\vec{w}_s$  are trainable linear transformation vectors and  $E_t$  is a separate embedding<sup>5</sup> for object types.

### 3.3 Compressing

Besides serving as a vocabulary of geographic names, the geographic context  $G$  is also used as an additional input to the text generation network, along with the vector representation of the image’s visual features. For that, it needs to be compressed into a single vector that encodes the most important aspects of the objects in  $G$ , proportional to their individual contribution to the geographic context. This vector  $e_g$  is built using a heuristic rule motivated by two assumptions.

The first assumption is that the types of the objects around the image location are helpful in characterizing the image location itself. This assumption is based on the fact that objects of similar types tend to be grouped together in space (e.g. shops in shopping streets, residential buildings in residential areas) and there are objects of different types that are commonly placed near each other (e.g. a church and a graveyard, a theater and a restaurant). Thus, we utilize the type embeddings  $E_t$  of the objects ( $o_1 \dots o_n$ ) in  $G$  to compute the vector  $e_g$ .

The second assumption is that the relative importance of a given object depends on its size and the distance to the image location. The bigger and the closer the object is, the more influence it should have on  $e_g$ . Therefore, we weigh the objects’ type embeddings  $E_t(t_i)$  by the objects’ sizes  $s_i$  and distances to the image location  $d_i$ . The weighted embeddings are then averaged<sup>6</sup>.

<sup>5</sup>A mapping from type indices to vectors of real numbers, initialized randomly and optimized during training.

<sup>6</sup>Averaging was the most effective compressing strategy in our experiments.

The compressed geocontext  $e_g$  is calculated as follows:

$$e_g = \frac{1}{n} \left( \sum_{i=1}^n (s_i + \varepsilon) \frac{1}{d_i + \varepsilon} E_t(t_i) \right) \quad (2)$$

where  $s_i$  and  $t_i$  are the size and the type of the object  $o_i$  respectively,  $d_i$  is the distance between  $o_i$  and the image location,  $\varepsilon$  is a smoothing term, and the mean is taken over all the objects ( $o_1 \dots o_n$ ) in  $G$ .

## 4 Geo-Aware Caption Generation

Our geo-aware caption generation system is based on the Show, Attend and Tell caption generator (Xu et al., 2015), which is structured as an encoder-decoder pipeline with an added visual attention component. We introduce the geographic contextual information at both the encoder and the decoder stage. Figure 2 shows the overall architecture of our system. Concrete implementation details (vector sizes, learning rates, etc.) are provided in Appendix A.

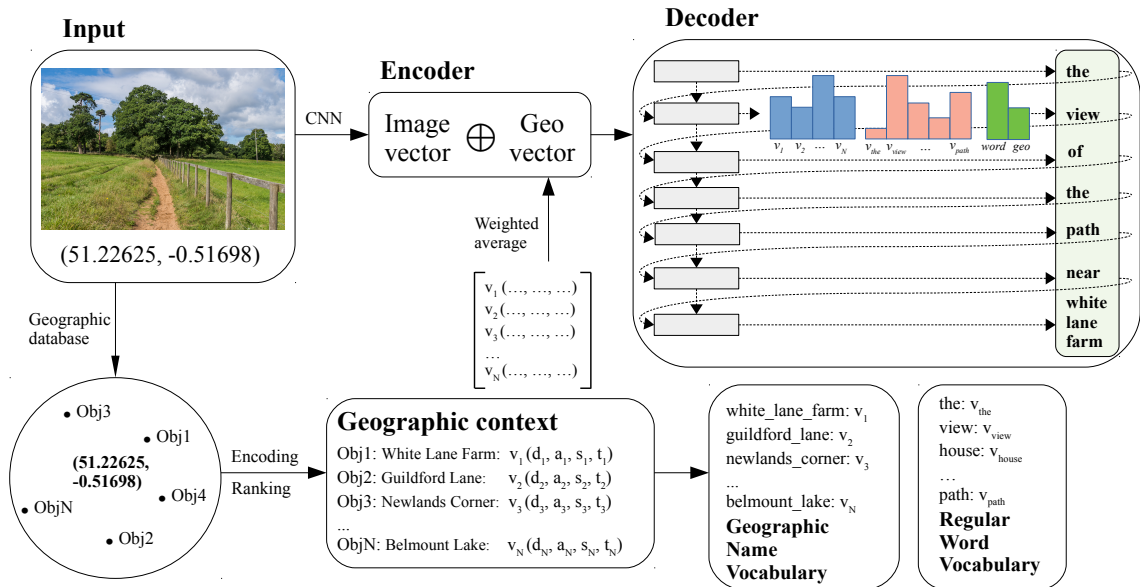


Figure 2: An overview of the geo-aware system architecture.

### 4.1 Encoder

The purpose of this stage is to encode the context that is needed for subsequent caption generation. In our system, the visual context for the caption generation is retrieved in a standard way, using a pre-trained Convolutional Neural Network. The image goes through a 101-layered Residual Network (He et al., 2016), pre-trained on ImageNet (Russakovsky et al., 2015). The output is a vector  $e_v$  that represents the visual content of the image.

In addition to that, the geographic context  $G$  is constructed based on the coordinates of the image location, as described in section 3.1. The full geographic context is then compressed into a single vector  $e_g$ , as described in section 3.3.

Finally, the two vectors  $e_v$  and  $e_g$  are concatenated and the result,  $e_v \oplus e_g$ , which represents both visual and geographic features of the input, is passed to the decoder, where it is used to initialize the text generation network.

## 4.2 Decoder

The purpose of the decoder is to generate a caption based on the output of the encoder. In a geo-aware caption generation system, the main challenge is to adapt the decoding mechanism to generate out-of-vocabulary geographic names as well as regular vocabulary words.

The decoder generates one word at a time, at each iteration using the image representation  $e_v$  produced by the encoder and the sequence of previous words. For generating vocabulary words, we use the same approach as in a standard captioning system, where the generated word  $w_t$  is the word with the highest probability of being next in the sequence, as in Equation 3. The probability distribution is estimated over all the words in the vocabulary  $V$ , as in Equation 4. Equation 5 shows the computation of the hidden state  $h_t$  of the Long Short-Term Memory network (LSTM) at time  $t$  (initialized by a concatenation of the visual vector  $e_v$  and the geographic vector  $e_g$  at  $t = 0$ ).

$$w_t = \arg \max_{w_i} P(w_i | w_0 \dots w_{t-1}) \quad (3)$$

$$P(w_i | w_0 \dots w_{t-1}) = \text{softmax}_i(h_t W_w), w_i \in V \quad (4)$$

$$h_t = \begin{cases} \text{LSTM}(e_v, \text{EMB}(w_0), \dots, \text{EMB}(w_{t-1}), h_{t-1}), & \text{if } t > 0 \\ e_v \oplus e_g, & \text{otherwise} \end{cases} \quad (5)$$

where  $W_w$  is a trainable linear transformation matrix,  $w_0, \dots, w_{t-1}$  are previous caption words represented through vector embeddings EMB.

To adapt this workflow to generate geographic names, three issues have to be addressed.

(1) **Geographic name embedding.** Caption words  $w_0, \dots, w_{t-1}$  can include geographic names as well as regular vocabulary words. Geographic names are most likely absent from the available collections of pre-trained word vector embeddings; moreover, the co-occurrence based approach to word embedding, standard for the vocabulary words, is not ideal for geographic names, which are better characterized through their real-world parameters (type, size, etc.). So, we use pre-trained GloVe embeddings (Pennington et al., 2014) for vocabulary words, whereas geographic names are encoded with a special geographic embedding GeoEmb (see section 3.2) that represents their geographic features.

$$\text{EMB}(w_i) = \begin{cases} \text{GLOVE}(w_i), & \text{if } w_i \in V \\ \text{GEOEMB}(w_i), & \text{if } w_i \in G \end{cases} \quad (6)$$

(2) **Geographic name selection.** The vocabulary  $V$  is compiled from the words observed in the training captions, which means that the geographic names that did not appear in the training data will not be included in  $V$  at all. However, like all named entities, geographic names are quite rare in the corpora and a lot of the names that could be relevant to the captions in the test sample will never be encountered during training. Therefore, the most probable geographic name should be selected from the names of the objects in the geographic context  $G$ , which is constructed for each image specifically, rather than the vocabulary  $V$ . We compute the probability distribution over the objects in  $G$  in the same way as we do for the vocabulary words:

$$P(o_i | w_0 \dots w_{t-1}) = \text{softmax}_i(h_t W_o), o_i \in G \quad (7)$$

where  $W_o$  is a trainable linear transformation matrix,  $h_t$  is calculated as in Equation 5.

(3) **Generation choice.** The adapted decoder needs to have a way of choosing between generating a vocabulary word and a geographic name at time  $t$ . In our system it is done through another estimation of a probability distribution. The probability of a binary “mask” ( $\{0$  for a vocabulary word,  $1$  for a geographic name $\}$ ) is computed as shown in Equation 8.

$$P(\text{mask}_i | w_0 \dots w_{t-1}) = \text{softmax}_i(h_t W_{\text{mask}}), \text{mask}_i \in \{\text{vocab}, \text{geo}\} \quad (8)$$

where  $W_{mask}$  is a trainable linear transformation matrix and  $h_t$  is calculated as in Equation 5. If the probability  $P(\text{vocab})$  is higher, then the most probable vocabulary word is generated, otherwise the output is the most probable geographic name.

## 5 Results and Discussion

In this section we first compare the performance of our geo-aware image captioning system with that of a non-geographic one, the Show, Attend and Tell system, trained on the same GeoRiC dataset (which we will refer to as the “standard” system). Second, we analyze the quality of the generated geographic references with a newly developed metric, which aims to directly assess the correctness of the spatial expressions usage instead of comparing the generated geographic references with the ground truth ones.

### 5.1 System Evaluation: Comparing Geo-aware with Standard

The non-geographic baseline we are comparing the geo-aware system with is the same one we used as a base for creating our system; thus, any difference in the performance is assumed to be a consequence of the added geographic component. Both the geo-aware and the standard system were trained and tested on the GeoRiC dataset (described in section 2). Out of 29,038 captions in the GeoRiC dataset, we used 21,778 randomly selected captions for training, 3,630 for validation and 3,630 for testing.

As is common in the image captioning task, we use the BLEU metric (Papineni et al., 2002) to compare the systems. BLEU counts the number of matching n-grams between the generated caption and the ground truth caption and computes the precision score, i.e. estimates how much of the generated caption is found in the ground truth caption. In addition, we report scores in ROUGE (Lin, 2004), METEOR (Denkowski and Lavie, 2014) and CIDEr (Vedantam et al., 2015), which were suggested as improvements upon BLEU. ROUGE computes recall rather than precision, estimating how much of the ground truth caption is captured in the generated caption. METEOR uses stemming to preprocess words and adds the ability to match synonyms. CIDEr gives a higher weight to the words that are more informative according to the TF-IDF score. The results of the comparison between the systems are shown in Table 5.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr
Standard	13.38	2.82	0.64	0.33	15.79	5.55	7.38
Geo-aware	<b>18.12</b>	<b>8.42</b>	<b>3.42</b>	<b>1.46</b>	<b>22.61</b>	<b>10.35</b>	<b>70.53</b>

Table 5: Metric scores of the standard and the geo-aware systems, measured on the test set.

The geo-aware system shows statistically significant improvements in all the reported metrics (two-sample t-test,  $p < 0.0001$ ). The most substantial improvement is observed in CIDEr, which increased almost tenfold compared to the standard system’s score. This surge could be explained by the fact that the geo-aware system is able to generate correct geographic names, which are highly informative and have a high TF-IDF and therefore contribute a lot to the CIDEr score. This claim is supported by a manual study of 100 caption pairs with the biggest difference in CIDEr. All the geo-aware captions in this sample contained geographic names that were also found in the corresponding ground truth captions.

The scores in Table 5 are considerably lower than those of the state-of-the-art image captioning systems applied to the standard datasets. For example, the original Show, Attend and Tell paper reports a BLEU-4 score of 25.0 and a METEOR score of 23.9 on the MSCOCO dataset. However, the highly contextualized captions in the GeoRiC dataset are much harder for the caption generation system to imitate. Moreover, the standard datasets provide several ground truth captions for each image and a match with any of them counts towards the system’s scores, whereas the GeoRiC dataset only has one caption per image. The few captioning systems developed on the datasets of contextualized captions and with a single caption per image (Lu et al., 2018; Whitehead et al., 2018; Biten et al., 2019) demonstrate results that are comparable to ours, with BLEU-4 scores ranging from 0.83 to 4.7, ROUGE – from 12.11 to 21.1, METEOR – from 4.34 to 11.0, CIDEr – from 12.79 to 29.9. A more detailed account of these systems is given in Section 6.

Table 6 shows some examples of the captions generated by both systems for the images in the test set, along with the ground truth captions for these images. In these examples, both the standard and geo-aware systems produce captions that accurately or almost accurately describe the visual content of the images. However, only the geo-aware system is able to generate correct geographic references; all the geographic references generated by the standard system are incorrect. Images for the examples in Table 6 and more examples of the generated captions are given in Appendix B.

<b>Ground truth</b>	Beach near Ethel Point at Bembridge
<b>Standard</b>	the beach <i>at roker near sunderland</i> on the horizon
<b>Geo-aware</b>	the view of the beach <b>near ethel point</b>
<b>Ground truth</b>	Grand Union Canal locks near Hatton Country World taken on a wet day
<b>Standard</b>	the bridge carries the over the canal <i>just west of horton village</i>
<b>Geo-aware</b>	the view of the lock <b>on the grand union canal near hatton</b>
<b>Ground truth</b>	The memorial is situated in Buccleuch Park
<b>Standard</b>	the war memorial <i>at the junction with &lt;unk&gt; road in kinver</i>
<b>Geo-aware</b>	a war memorial <b>in buccleuch park</b>

Table 6: Examples of the generated captions. Correct geographic references are given in **bold**; incorrect ones are given in *italics*.

## 5.2 Geographic References Analysis

A common way of assessing the quality of the generated caption is comparing it to the ground truth caption. But this method is too restrictive for establishing whether the generated geographic references are factually accurate: there are numerous ways to describe a single image location. For example, the ground truth caption may describe a house in the image as located “near Victoria Street” while the generated caption describes it as located “in Kent”, where both descriptions are correct.

Instead of comparing the generated geographic references to the ones that were used in the ground truth captions, we directly assess the accuracy of the generated references by verifying that the spatial expressions are combined with geographic names that satisfy the expression’s semantic requirements. We specifically target the expressions that occur most frequently in our dataset: “near”, “in”, “along”, “across”, “north of”, “south of”, “east of”, “west of”.

Based on the theoretical research on spatial prepositions (Gahegan, 1995; Garrod et al., 1999; Take-mura et al., 2005; Zwarts, 2017), we use certain parameters of a geographic object  $X$  to estimate whether it is possible to use  $X$  in connection with a given spatial expression. The parameters for each of the selected spatial expressions are listed below:

- **near, in:** distance (*between the image location and  $X$* )
- **along, across:** distance (*between the image location and  $X$* ) and type (*of  $X$* )
- **north (south, east, west) of:** azimuth (*of an angle between the image location and  $X$* )

For each spatial expression, we obtain the distributions of these parameter values from training data. These distributions approximate the conditions in which the expressions have been naturally produced by humans. Then, in a similar fashion, we obtain the distributions of the parameter values from the captions generated by the system for the test images and their geographic contexts. Using the Wasserstein metric (Wasserstein, 1969), we determine how close these distributions are to each other. The Wasserstein metric is equal to an area between the two empirical cumulative distribution functions (ECDFs) and is calculated as follows:

$$\text{WM}(p_{\text{train}}, p_{\text{gen}}) = \int_{-\infty}^{+\infty} |\text{ECDF}_{\text{train}} - \text{ECDF}_{\text{gen}}| \quad (9)$$



where  $p_{train}$  is the distribution of the parameter values in the training data,  $p_{gen}$  is the distribution of the parameter values in the generated captions, and  $ECDF_{train}$  and  $ECDF_{gen}$  are the respective ECDFs of these distributions.

The idea is that if the system has learned the semantic requirements of a given spatial expression, the distribution of its parameter values in the generated captions will be close to the distribution observed in the training data, and therefore, its Wasserstein metric score will be low.

Table 7 shows how often each of the spatial expressions was generated and the Wasserstein metric scores for their parameters. For most expressions, these scores are substantially lower than the Wasserstein metric scores for a random baseline, in which every generated geographic name was replaced with a random one. The system however struggles with the less frequent azimuth-related expressions “south of”, “east of”, “west of”.

Note that the random baseline is still a strong baseline, since we selected the random names from the geographic contexts, which means that these names are quite likely to appear in the captions.

Spatial expression	Number of occurrences	Wasserstein metric	
		Our system	Random baseline
Near	2233	<b>4.14</b>	20.54
In	533	<b>16.64</b>	32.19
Along	307	<b>32.72</b> (distance)	56.24 (distance)
		<b>9.92</b> (type)	14.8 (type)
Across	54	<b>18.41</b> (distance)	38.24 (distance)
		<b>11.87</b> (type)	13.98 (type)
North of	241	<b>37.59</b>	46.97
South of	42	51.69	<b>41.4</b>
East of	181	67.78	<b>56.99</b>
West of	35	58.45	<b>56.39</b>

Table 7: The number of generated spatial expressions and the Wasserstein metric scores.

The lower Wasserstein metric scores achieved by our system indicate that it has to some extent learned the semantic requirements of the spatial expressions, e.g. the fact that “near” should be combined with the objects that are close to the image location, or that “along” requires the geographic objects of certain types, such as roads and rivers but not buildings or towns.

## 6 Related Work

Most modern image captioning systems use the standard encoder-decoder approach with a CNN encoding the visual features of the image and an RNN, usually an LSTM, decoding the image features into a textual description (Vinyals et al., 2015; Xu et al., 2015; You et al., 2016). Much effort is directed at improving the components of the standard caption generation pipeline (Lu et al., 2017; Wang and Chan, 2018; Zhu et al., 2018; Hossain et al., 2019; Tan et al., 2019); however, the standard approach itself has several shortcomings, such as producing captions that are too generic (Tran et al., 2016), inability to generate words that are not present in the training data (Hendricks et al., 2016) and disregarding pragmatic factors and world knowledge (van Miltenburg, 2019).

There have been a few attempts to create a context-aware, less generic caption generation system. Lu et al. (2018) and Biten et al. (2019) proposed generating image descriptions as templates with blank slots and filling them afterwards with named entities, including locations and dates. This template-based method might be problematic in case some of the generated blank slots cannot be filled based on the available data.

Closer to our work, Whitehead et al. (2018) used a more flexible approach<sup>7</sup> to enrich video descriptions

<sup>7</sup>A “pointer-generator” technique that allows the system to select between generating a vocabulary word and a named entity, adapted from See et al. (2017) where it was used for text summarization.

with event and entity names that are extracted from related news articles. The authors also introduced a “knowledge gate vector”, the purpose of which is to provide the description generation network with information about the types of events and entities mentioned in the article. Similarly to Whitehead et al. (2018), we use an external source (a geographic database) to create a representation of the context and adapt the text generator to produce out-of-vocabulary context-related words. Our geographic context representation not only registers the presence of certain entity types but also incorporates various features of the relevant entities.

To the best of our knowledge, none of the context-aware caption generation systems have had a specific focus on the geographic context of the image location. On the other hand, geographic information in particular has been used as an additional input to the systems developed for other tasks, such as image classification (Tang et al., 2015; Chu et al., 2019) and creating distributional word embeddings (Cocos and Callison-Burch, 2017).

Both Tang et al. (2015) and Chu et al. (2019) reported increased performance in the image classification task after providing their systems with geographic knowledge. However, the content of the geographic knowledge provided to their systems was very different. Chu et al. (2019) used only the normalized latitude and longitude coordinates of the image location. Opposite to that, Tang et al. (2015) supplied to their system all the geological, ecological, sociological, demographic and other types of data that could be obtained from the external sources. Our approach to constructing the geographic context strikes a balance between the two: we include explicit information about the most relevant geographic objects around the image location and ensure that this information is only what is most useful for the captioning process.

Cocos and Callison-Burch (2017) used geographic information for creating contextualized word embeddings, specifically information about the types of the objects around the word usage location. The authors assigned the same weight to the objects within the same radius from the location, however, arguably, such objects can still be more or less significant depending on their size and relative distance to the location. In our system, we use these characteristics of the objects to construct the maximally relevant geographic context.

Overall, our system is a novel combination of the two trends that have emerged in the recent research: the context-aware image caption generation and the usage of geographic information to enhance the performance of various NLP and vision and language models.

## 7 Conclusions

In this paper, we have presented an approach to incorporate geographic contextual information into the caption generation pipeline. To the best of our knowledge, this is the first approach that focuses on generating geographically grounded image descriptions. In addition, we have introduced the GeoRic dataset for image captioning, which includes contextualized captions and geographic metadata. Our experiments on the GeoRic dataset showed the effectiveness of our geo-aware captioning system and its advantage over a standard captioning network, with large improvements in several standardly used evaluation metrics, as well as a good performance on a metric designed specifically for geographic expressions. Our system is thus able to produce contextualized captions that include correct geographic referencing without compromising the overall quality of the image description. In future work, we plan to refine the process of selecting the most appropriate geographic object in a given context and to experiment with including other types of relevant contextual information, as well as with other vision and language systems (such as visual question answering) that could benefit from geographic grounding.

## Acknowledgements

This work was funded by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 742204). We thank the anonymous reviewers for their helpful comments. We would also like to thank colleagues in the ROCKY project and all members of the NLP Reading Group at UiL OTS, Utrecht University, for many fruitful discussions and valuable suggestions.

## References

- Ali Furkan Biten, Lluís Gomez, Marçal Rusiñol, and Dimosthenis Karatzas. 2019. Good news, everyone! Context driven entity-aware captioning for news images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12466–12475.
- Grace Chu, Brian Potetz, Weijun Wang, Andrew Howard, Yang Song, Fernando Brucher, Thomas Leung, and Hartwig Adam. 2019. Geo-aware networks for fine-grained recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*.
- Anne Cocos and Chris Callison-Burch. 2017. The language of place: Semantic value from geospatial context. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 99–104.
- Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380.
- Mark Gahegan. 1995. Proximity operators for qualitative spatial reasoning. In *International Conference on Spatial Information Theory*, pages 31–44. Springer.
- Simon Garrod, Gillian Ferrier, and Siobhan Campbell. 1999. In and on: investigating the functional geometry of spatial prepositions. *Cognition*, 72(2):167–189.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.
- Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. 2016. Deep compositional captioning: Describing novel object categories without paired training data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–10.
- Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.
- Md Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, Hamid Laga, and Mohammed Bannamoun. 2019. Bi-SAN-CAP: Bi-directional self-attention for image captioning. In *2019 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–7. IEEE.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 375–383.
- Di Lu, Spencer Whitehead, Lifu Huang, Heng Ji, and Shih-Fu Chang. 2018. Entity-aware image caption generation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4013–4023.
- Yuzhao Mao, Chang Zhou, Xiaojie Wang, and Ruifan Li. 2018. Show and tell more: Topic-oriented multi-sentence image captioning. In *IJCAI*, pages 4258–4264.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

- Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, pages 1073–1083.
- Celina Maki Takemura, Roberto Cesar, and Isabelle Bloch. 2005. Fuzzy modeling and evaluation of the spatial relation “along”. In *Iberoamerican Congress on Pattern Recognition*, pages 837–848. Springer.
- Jia Huei Tan, Chee Seng Chan, and Joon Huang Chuah. 2019. Comic: Toward a compact image captioning model with attention. *IEEE Transactions on Multimedia*, 21(10):2686–2696.
- Kevin Tang, Manohar Paluri, Li Fei-Fei, Rob Fergus, and Lubomir Bourdev. 2015. Improving image classification with location context. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1008–1016.
- Kenneth Tran, Xiaodong He, Lei Zhang, Jian Sun, Cornelia Carapcea, Chris Thrasher, Chris Buehler, and Chris Sienkiewicz. 2016. Rich image captioning in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*, pages 49–56.
- Emiel van Miltenburg. 2019. *Pragmatic factors in [automatic] image description*. Ph.D. thesis, Vrije Universiteit Amsterdam.
- Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164.
- Qingzhong Wang and Antoni B. Chan. 2018. CNN+CNN: Convolutional decoders for image captioning. *Computing Research Repository*, arXiv:1805.09019.
- Leonid N Wasserstein. 1969. Markov processes over denumerable products of spaces describing large systems of automata. *Problems of Information Transmission*, 5(3):47–52.
- Spencer Whitehead, Heng Ji, Mohit Bansal, Shih-Fu Chang, and Clare Voss. 2018. Incorporating background knowledge into video description generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3992–4001.
- Siyang Wu, Zheng-Jun Zha, Zilei Wang, Houqiang Li, and Feng Wu. 2019. Densely supervised hierarchical policy-value network for image paragraph generation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 975–981. AAAI Press.
- Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *32nd International Conference on Machine Learning, ICML 2015*, pages 2048–2057.
- Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4651–4659.
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association of Computational Linguistics*, 2(1):67–78.
- Xinxin Zhu, Lixiang Li, Jing Liu, Haipeng Peng, and Xinxin Niu. 2018. Captioning transformer with stacked attention modules. *Applied Sciences*, 8(5):739.
- Joost Zwarts. 2017. Spatial semantics: Modeling the meaning of prepositions. *Language and linguistics compass*, 11(5).

## Appendix A. Implementation

We use the PyTorch framework for the codebase of our system, specifically the PyTorch implementation of the Show, Attend and Tell system at <https://bit.ly/2Yv8gw3>, with some modifications to adapt it for our dataset.

The input to the visual encoder is an image that is resized to 256x256 pixels. The encoding has a size 14x14 with 2048 color channels, which makes the output of the encoder a (2048, 14, 14) size tensor. The pre-trained visual encoder is used without fine-tuning.

The decoder is based on a single layer LSTM with the output dimension size 512. For the word encoding, we use pre-trained GloVe embeddings of size 300, trained on the Common Crawl data. Words that are missing from the pre-trained vocabulary are initialized with random vectors. The word embeddings are then fine-tuned during training.




The geographic context that is used in the system consists of 300 objects, with an added extra <UNK\_ENT> “object” to handle unrecognized geographic names in the training data (the names that were not found in the geographic database). The original geographic features of the objects (distance, azimuth, size and type) are transformed into vectors of the same dimensionality as the vocabulary word embeddings.

For the fine-tuning both in the geographic context encoding and in the decoder, we use the Adam optimizer with the learning rate of  $4e^{-4}$ . During training, we compute the sum of the cross entropy losses calculated separately for words, geographic objects and mask generation. Training runs for 120 epochs with the batch size of 32 but if there is no improvement in loss for 20 consecutive epochs, early stopping is enabled. The best-performing system trained for 11 epochs (the early stopping was triggered after 31 epochs).

## Appendix B. Examples of the Generated Captions

Table 8 shows some examples of the captions generated by the geo-aware system and the standard system for the images in the test set. In (a)-(d), the correct geographic references generated by the geo-aware system include the same geographic names as the ground truth captions. In (e), the geographic reference generated by the geo-aware system does not match the ground truth one but is still accurate. In (f), the generated geographic name is the same as in the ground truth caption, but the overall reference is factually inaccurate (the image location is located to the west of Lumphanan, not to the east of it). In (g), both standard and geo-aware systems generated an incorrect geographic reference.

---

(a) <sup>8</sup>		<b>Ground truth:</b> Beach near Ethel Point at Bembridge <b>Standard:</b> the beach <i>at roker near sunderland</i> on the horizon <b>Geo-aware:</b> the view of the beach <b>near ethel point</b>
(b) <sup>9</sup>		<b>Ground truth:</b> Country road north of Sherfield on Loddon <b>Standard:</b> the view of the road junction <i>on the staffordshire and worcestershire canal near compton wolverhampton</i> <b>Geo-aware:</b> a minor road <b>to the north of sherfield</b>
(c) <sup>10</sup>		<b>Ground truth:</b> The memorial is situated in Buccleuch Park <b>Standard:</b> the war memorial <i>at the junction with &lt;unk&gt; road in kinver</i> <b>Geo-aware:</b> a war memorial <b>in buccleuch park</b>

---

<sup>8</sup><https://www.geograph.org.uk/photo/4037124>

<sup>9</sup><https://www.geograph.org.uk/photo/5421983>

<sup>10</sup><https://www.geograph.org.uk/photo/603837>





(d) <sup>11</sup>		<p><b>Ground truth:</b> Grand Union Canal locks near Hatton Country World taken on a wet day</p> <p><b>Standard:</b> the bridge carries the over the canal <i>just west of horton village</i></p> <p><b>Geo-aware:</b> the view of the lock <b>on the grand union canal near hatton</b></p>
(e) <sup>12</sup>		<p><b>Ground truth:</b> Grazing land near Evesbatch with the Malvern Hills on the horizon</p> <p><b>Standard:</b> a field of winter cereals <i>near &lt;unk&gt;</i></p> <p><b>Geo-aware:</b> a grassy field <b>near bishop's frome</b></p>
(f) <sup>13</sup>		<p><b>Ground truth:</b> Crossroad of minor roads, west of Lumphanan</p> <p><b>Standard:</b> the road <i>to the north of blackrod</i></p> <p><b>Geo-aware:</b> a small road <i>to the east of lumphanan</i></p>
(g) <sup>14</sup>		<p><b>Ground truth:</b> Fine buildings near Abbey Mill, Tewkesbury</p> <p><b>Standard:</b> the &lt;unk&gt; inn is situated <i>in station road</i></p> <p><b>Geo-aware:</b> a row of cottages on the coast <i>near bushley</i></p>

Table 8: Examples of the generated captions. Correct geographic references are given in **bold**; incorrect ones are given in *italics*.

<sup>11</sup><https://www.geograph.org.uk/photo/341036>

<sup>12</sup><https://www.geograph.org.uk/photo/3317792>

<sup>13</sup><https://www.geograph.org.uk/photo/5082917>

<sup>14</sup><https://www.geograph.org.uk/photo/5450719>