Temporal Knowledge-Aware Image Captioning

Anonymous ACL submission

Abstract

Contextualized image captioning is a task that extends beyond generating a purely visual description of the image content and aims to produce a caption that is influenced by the context and informed by the real world knowledge. In this paper, we present an approach to knowledge-aware image captioning, with a specific focus on the temporal domain. We propose a way to identify relevant information in external data sources, such as geographic databases and common knowledge bases, and then encode it in a way that is most useful for the captioning network. We develop an end-to-end caption generation system that incorporates external knowledge into the captioning process at several stages. The system is trained and tested on our novel temporal knowledge-aware captioning dataset, achieving significant improvements over multiple baselines across standardly used metrics. We demonstrate that our approach is effective for generating highly contextualized captions with both relevant *and* accurate temporal facts.

1 Introduction

Image captioning is the task of automatically generating a natural language caption for a given image. A rapidly evolving modification of this task is contextualized image captioning (Lu et al., 2018; Biten et al., 2019; Zhao et al., 2019; Nikiforova et al., 2020; Tran et al., 2020; Bai et al., 2021) which aims to extend beyond a purely visual description and produce a caption that is influenced by the context and informed by the real world knowledge. Motivating this research is the stark contrast between the captions created by most automatic caption generators and the captions that humans produce naturally. Consider the image in Figure 1. The captions that were generated by two standard automatic captioning systems (Xu et al., 2015) and (Anderson et al., 2018) are almost identical and both accurately describe the visual content of the image. However, the human-generated caption is very different: it is much more contextualized (identifies this famous clock tower as Big Ben) and includes information that cannot be inferred from the image alone (the year of construction). In order to produce such captions, an

Human: Clock Tower, Palace of Westminster. Completed in 1859, the clock tower houses the bell known as Big Ben.

SAT (Xu et al., 2015): a very tall clock tower towering over a city

BUTD (Anderson et al., 2018): a tall clock tower towering over a city

Figure 1: An example image. https://www.geograph.org.uk/photo/2865824

automatic caption generator has to be able to access and utilize real world knowledge relevant to the image. This task presents a range of challenges, starting with identifying such knowledge in external data sources. Crucially, it needs to be done for every input image separately, since general pre-training would not cover the specific knowledge related to the objects in the individual images. Further, the extracted knowledge needs to be represented in a way that is useful for the image captioning network; distribution-based representation, which is standardly used for vocabulary words, is not particularly informative for the named entities and facts, as their semantics is conveyed poorly by their distribution patterns (e.g. the various contexts in which the token "1859" appears in a large scale corpus are too diverse for a good and precise representation of "1859" as the year when Big Ben was completed). The caption generation process needs to be adapted to produce image-specific facts along with the regular vocabulary words. Finally, the generated facts must be *accurate* in the context of the image and according to the external knowledge sources. This adds a new dimension to the evaluation of the generated captions: verifying their factual correctness, beyond what can be verified from the image itself. These challenges, although explored for general-purpose knowledge-aware language modeling (Logan et al., 2019; Liu et al., 2019; Hayashi et al., 2020), have not yet been tackled in the context of image captioning.

In this paper we present an approach to *knowledge-aware image captioning*, with relevant facts from an external knowledge base informing the caption generation process. We specifically concentrate on a subset of *temporal knowledge*, i.e. facts related to time indications, such as the "completed in 1859" fact in Figure 1. This restriction on the knowledge domain lets us limit the variability of data the captioning system is exposed to, ensuring a more focused and controllable study. Our proposed approach can be easily generalized to all types of facts. Our contributions are as follows:

(I) We present a novel method of identifying and retrieving relevant knowledge from multiple databases by using the *geographic* metadata of an image in order to construct an *image-specific* knowledge context.

(II) We develop a contextualized image captioning pipeline with extra knowledge incorporated at several stages. Specifically, the generation module is modified for working with geographic names and fact-related entities relevant for a given image, which appear in the captions alongside regular vocabulary words. To the best of our knowledge, this is the first time that the underlying language model in an image captioning system has been made knowledge-aware by integrating real world facts from an external knowledge base.

(III) We compile a new dataset of naturally created image captions, where each caption includes a contextually relevant temporal fact. We conduct extensive experiments on this dataset and show the effectiveness of our proposed framework based on multiple image captioning metrics and the *correctness* of the generated facts.

2 Related Work

In **image caption generation**, the seminal Show and Tell paper (Vinyals et al., 2015) introduced an end-to-end trainable neural caption generator structured as an encoder-decoder pipeline. It consists of two stages: in the first stage, a CNN encoder (usually pre-trained on an image classification task) provides a representation of the visual features of the image, and in the second stage, an RNN decoder is initialized with the encoder's output and generates a caption word by word. Further research presented many technical improvements to the standard architecture, such as the attention mechanism over the visual image features (Xu et al., 2015; You et al., 2016; Lu et al., 2017; Anderson et al., 2018), scene graph generation for the image representation (Wang et al., 2019; Li and Jiang, 2019; Lee et al., 2019), the Transformer network instead of a traditional RNN as the decoder (Zhu et al., 2018; Li et al., 2019; Yu et al., 2019). In this paper, we build on the previous advances in image captioning and use a de facto standard encoder-decoder system with a pre-trained CNN in the encoder and a Transformer network in the decoder.

In the subtask of **contextualized image captioning**, external knowledge is incorporated into the caption generation system, providing image-specific information relevant for generating the captions. The sources of external knowledge can include related textual data (e.g. when captions are generated for the news article images), a common knowledge base such as ConceptNet (Speer et al., 2017) or DBpedia (Auer et al., 2007), or a specialized database, for example, a geographic one. In this work, we utilize the OpenStreetMap[1] database for geographic knowledge and DBpedia for general facts.

Existing datasets for contextualized image captioning usually include either relevant contextual information directly (Biten et al., 2019; Whitehead et al., 2018; Tran et al., 2020) or the metadata needed for extracting it from external sources (Lu et al., 2018; Nikiforova et al., 2020). The recently released Wikipedia-based Image Text (WIT) Dataset (Srinivasan et al., 2021) contains images from Wikipedia articles accompanied by metadata and related texts. The included metadata is mostly low-level (mime type, height, width) and does not cover, for example, related geographic information even when it is available. In a running example from Srinivasan et al. (2021), which is a photograph of Half Dome in Yosemite Valley, the location where the photograph was taken is available on the Wikimedia page of the image but it is

[1] https://www.openstreetmap.org/

not included in the metadata for this image in the WIT dataset. The geographic metadata (latitude and longitude coordinates of the image location) is likely to be easily available for many real-life photographs due to the built-in GPS in modern cameras and phones, and it can be extremely helpful in identifying information relevant for the contextualized image description in external data sources, which is why we include it in our dataset for temporal knowledge-aware captioning.

Some contextualized image captioning systems use a template approach: a caption is generated with placeholder token slots that are later filled with the most fitting named entities extracted from an external knowledge source (Biten et al., 2019; Jing et al., 2020; Hu et al., 2020; Bai et al., 2021). The template approach is reported to be effective for producing more informative image descriptions; however, it can be problematic if no relevant entities of the required type are present in the available external data. A more flexible approach involves encoding external knowledge and using it as a context that informs the caption generation process (Mogadala et al., 2018; Zhou et al., 2019; Huang et al., 2020; Tran et al., 2020) and, in some works, as an additional vocabulary for the decoder (Whitehead et al., 2018; Chen and Zhuge, 2020; Nikiforova et al., 2020). In this paper, we use the relevant knowledge in two ways: first, as the additional context alongside the visual representation of the image and, second, for building the image-specific vocabularies of real world entities and facts that can be generated in the caption.

Our approach to generating facts in the captions draws from **knowledge-aware language modeling** (Logan et al., 2019; Liu et al., 2019; Hayashi et al., 2020). Multiple LMs have been developed that make use of external knowledge bases, such as Freebase (Bollacker et al., 2008) and Wikidata (Vrandečić and Krötzsch, 2014). These models are able to choose between generating a knowledge base entity or a regular vocabulary word based on the preceding context. We propose a novel application of this idea to the image captioning task, providing our underlying LM with three types of tokens to choose from (names of the geographic entities around the image location, temporal facts about these entities and regular vocabulary words), with our newly developed separate ways of encoding and generating the different token types to best address their specific properties.

3 The Temporal Knowledge-Aware Dataset

For our task of temporal knowledge-aware image captioning, we develop a novel dataset[2] with naturally created image captions that include facts from the temporal domain. We collected the images and the related data from the website of the Geograph project[3], which aims to photograph and document every square kilometer of Great Britain. An advantage of this data source is the rich metadata that is provided for the photographs, including the latitude and longitude coordinates of the location where each photograph was taken.

Our dataset consists of 6788 Geograph images with the captions and the location metadata. Each caption in the dataset contains a reference to a temporal fact (a fact related to a date or a year) about a topical geographic entity (e.g. a building, a bridge, a park, etc.), for example, "Theatre Royal Haymarket. Dating back to 1720". Each image is paired with the latitude and longitude coordinates of its location, which makes it possible to identify information relevant to the image in various external knowledge resources. For example, in our knowledge-aware captioning system we utilize the coordinates to retrieve a list of objects around the image location from a geographic database and then extract facts about these objects from a general knowledge base (see Section 4). The details regarding the dataset split into train, validation and test sets are given in Appendix A.

4 Modeling Context

We introduce two types of context into the image captioning system: the geographic context and the (temporal) knowledge context. The geographic context of a given image is approximated as a set of relevant geographic entities around the image location, which may or may not be depicted in the image itself. We use the geographic context to build the knowledge context — a collection of facts about the relevant geographic entities. Both contexts, along with the visual features of the image, inform the caption generation process. The contexts also act as image-specific vocabularies of geographic names and facts that can appear in the caption.

[2]The dataset will be publicly available online at ANONYMIZED

[3]http://www.geograph.org.uk/

4.1 Geographic Context

In constructing the geographic context, we modify an approach proposed in Nikiforova et al. (2020) to adapt it to knowledge-aware captioning. The geographic context G of a given image is a set of n geographic entities $(e_1 \ldots e_n)$ located within a radius r from the image location. We set n at 300 and r at 1 kilometer as the hyperparameters of our system.

Each geographic entity e_i is associated with its name and a set of geographic features proposed in Nikiforova et al. (2020): distance d_i and azimuth a_i between the entity and the image location, the entity's size s_i and type t_i (as provided in the OpenStreetMap database). In addition to that, we introduce two new features, intended to reflect the salience of the entity through the amount of information available about it in a knowledge base: a binary indicator $\exists f_i$ that shows whether or not the entity corresponds to any facts in the knowledge context, and the number of facts $\#f_i$ that correspond to the entity in the knowledge context. A sample fragment of a geographic context, with the entities mapped to their names and features, is shown in Figure 2.

e_1 – Theatre Royal – (d_1=0.02km, a_1= -132°, s_1=0.001km², t_1=theatre, $\exists f_1$=1, $\#f_1$=2)
e_2 – Orange Street – (d_2=0.04km, a_2= -170°, s_2=0.0009km², t_2=tertiary, $\exists f_2$=0, $\#f_2$=0)
...
e_n – Charing Cross – (d_n=0.37km, a_n= -81°, s_n=0.0km², t_n=station, $\exists f_n$=1, $\#f_n$=1)

Figure 2: A fragment of a geographic context.

The features are combined in vector representations for the entities, which we call "geographic embeddings". For an entity e_i a geographic embedding is computed as follows:

$$\begin{aligned}\textsc{GeoEmb}(e_i) = \ & \text{Concat}[d_i,\, norm(a_i),\, s_i, \\ & \exists f_i,\, \#f_i, Emb_t(t_i)]\end{aligned} \quad (1)$$

where $norm$ is an azimuth normalization function, Emb_t is an embedding function for the entities' types, with the embeddings initialized randomly and optimized during training.

4.2 Knowledge Context

The knowledge context K is defined for a given image with the geographic context G as a set of m facts $(f_1 \ldots f_m)$ about the entities $(e_1 \ldots e_n) \in G$.

We obtain the facts from the DBpedia knowledge base, where they are stored as triples of the form <subject, predicate, object>. First, we select all the facts, in which the subject is one of the geographic entities from G, e.g. <Theatre Royal, built_in, 1720>, <Theatre Royal, architect, John Nash>, <Theatre Royal, rebuilt, 1879>, etc. We further restrict the list of facts to the ones where the object is a a date or a year, thus removing, for example, the architect fact above. We train a logistic regression model to rank the remaining facts based on how likely they are to be mentioned in the caption. The model takes into account the fact's predicate, the ranking of the fact's subject in the geographic context and its geographic features. The top m facts of the ranked list constitute the knowledge context of the image, with m as another hyperparameter of the system, which we set at 50.

Figure 3 shows a fragment of the knowledge context corresponding to the geographic context in Figure 2. We consider the year tokens, which were originally the objects in the fact triples, to be the "labels", by which the facts are realized in the captions. Each fact is therefore mapped to a year token, which can appear in a caption, and to a pair <subject, predicate> where the subject is a geographic entity from G.

f_1 – 1720 – <Theatre Royal, built_in>
f_2 – 1879 – <Theatre Royal, rebuilt>
...
f_m – 1906 – <Charing Cross, opened_in>

Figure 3: A fragment of a knowledge context.

Similarly to the geographic context, each fact in the knowledge context is represented in a vector form. A "fact embedding" for a fact f_i is calculated as follows:

$$\textsc{FactEmb}(f_i) = \textsc{GeoEmb}(e_i) + Emb_p(p_i) \quad (2)$$

where e_i is the subject of the fact f_i (an entity from the geographic context), p_i is its predicate and Emb_p is an embedding function for the predicates, with the embeddings initialized randomly and optimized during training.

This approach to encoding facts in the knowledge context provides the captioning system with the information it needs to select an appropriate fact based on the previously generated tokens. For any given fact, the system can take into account whether or not its subject is already

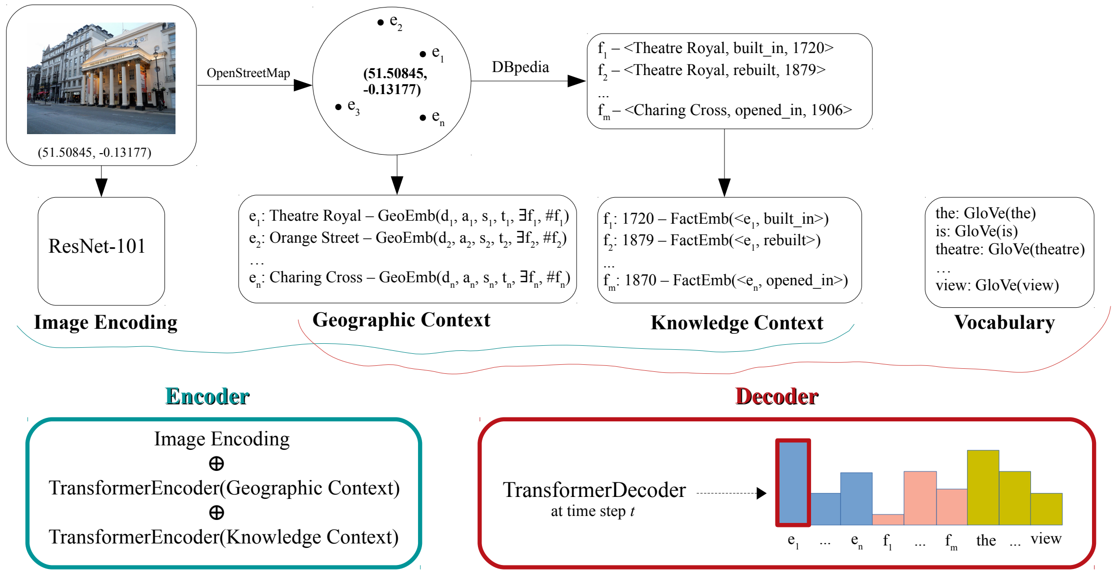

Figure 4: An overview of the knowledge-aware system architecture (best viewed in color).

present in the caption and estimate whether the previous caption tokens are consistent with the fact's predicate. The approach is not specific to the temporal domain: a fact of any type can be represented as a combination of its subject and predicate, e.g. <Theatre Royal, owner, Access Industries> $\rightarrow$ FACTEMB(Access Industries) = GEOEMB(Theatre Royal) + Emb_p(owner).

5 Knowledge-Aware Captioning Model

Our knowledge-aware caption generation system is an end-to-end trainable neural network with an encoder-decoder architecture. The overview of the system's architecture is shown in Figure 4. As seen in the figure, both encoder and decoder in the system make use of the geographic and knowledge contexts to produce knowledge-rich captions.

5.1 Encoder

The encoder's function is to convert input data into an informative representation that is subsequently used by the decoder to generate a caption. In a standard image captioning pipeline, the input data consists only of the image itself, and its encoding is a dense representation of its visual features. In our system, we also use geographic and knowledge contexts as the additional sources of input data.

For the image encoding E_{image}, we use a deep convolutional neural network (CNN), pre-trained on an image classification task, which is standard in image captioning applications. The particular CNN that we use is ResNet-101 (He et al., 2016), trained on the images from the ImageNet database (Russakovsky et al., 2015).

In addition, we encode information contained in the geographic and knowledge contexts. First, each of their elements is embedded with the embedding functions introduced in Section 4:

$$\begin{aligned} EmbG &= (\text{GEOEMB}(e_1)\dots \\ &\quad \dots\text{GEOEMB}(e_n)), e_i \in G \\ EmbK &= (\text{FACTEMB}(f_1)\dots \\ &\quad \dots\text{FACTEMB}(f_m)), f_i \in K \end{aligned} \tag{3}$$

They are subsequently encoded with two separate Transformer encoders (TrEnc), with a standard structure proposed in Vaswani et al. (2017).

$$\begin{aligned} E_{geo} &= \text{TRENC}(EmbG) \\ E_{fact} &= \text{TRENC}(EmbK) \end{aligned} \tag{4}$$

Finally, we concatenate the encodings of the image, the geographic context and the knowledge context:

$$E_{context} = \text{Concat}[E_{image},\ E_{geo},\ E_{fact}] \tag{5}$$

The result of the concatenation is the combined representation of the visual features of the image and the relevant information from the geographic and knowledge contexts.

5.2 Decoder

The decoder accepts the combined context representation from the encoder and generates an output sequence — the caption. The goal of the decoder is to produce a caption that would be fitting to the image and include accurate references to the geographic and knowledge contexts.

The decoder generates a caption token by token, at each step t taking into account the previously generated tokens $w_1 \dots w_{t-1}$ and the context representation $E_{context}$. In the process, each input token is represented by a sum of its vector embedding and the encoding of its position in the sequence.

$$PosEmb(w_i) = Emb(w_i) + Pos(w_i) \quad (6)$$

We use pre-trained GloVe word embeddings (Pennington et al., 2014) for the regular vocabulary tokens. However, the geographic entity names and fact-related tokens require a different kind of representation. It is important that the decoder can utilize information about their most meaningful characteristics: physical properties of the geographic entities and the facts' subjects and predicates. For this reason, we use the GEOEMB and FACTEMB embedding functions to represent them.

$$Emb(w_i) = \begin{cases} \text{GEOEMB}(w_i), \text{if } w_i \in G \\ \text{FACTEMB}(w_i), \text{if } w_i \in K \\ \text{GLOVE}(w_i), \text{otherwise} \end{cases} \quad (7)$$

We employ a Transformer decoder (TrDec) with a standard structure. It attends to the positional embeddings of the previously generated tokens and to the encoder's output, the combined representation of the image contexts.

$$h_t = \text{TRDEC}(PosEmb(w_{1 \dots t-1}); E_{context}) \quad (8)$$

In a standard captioning pipeline, the output of the decoder h_t is then passed to a final linear layer that acts as a classifier, estimating the probability distribution over all the tokens in the vocabulary V. The vocabulary is usually fixed and consists of the words from the training dataset. In our case, captions also include entity names and facts from the geographic and knowledge contexts, which are image-specific, and therefore, not all relevant entity names and fact-related tokens would necessarily be a part of V. So, we modify the last stage of the decoding process by computing three sets of scores: the scores for the vocabulary tokens from V, the scores for the geographic entity names from G and the scores for the facts from K.

$$\begin{aligned} y_{v_1} \dots y_{v_k} &= h_t\, W_{vocab},\ v_1 \dots v_k \in V \\ y_{e_1} \dots y_{e_n} &= (EmbG\ \text{DIAG}(h_t))\, \vec{w}_{geo},\ e_1 \dots e_n \in G \\ y_{f_1} \dots y_{f_m} &= (EmbK\ \text{DIAG}(h_t))\, \vec{w}_f,\ f_1 \dots f_m \in K \end{aligned} \quad (9)$$

where W_{vocab} is a trainable linear transformation matrix, $\vec{w}_{geo}$ and $\vec{w}_f$ are trainable linear transformation vectors, and $\text{DIAG}(h_t)$ denotes a diagonal matrix with the h_t vector in the main diagonal.

The three sets of scores are then concatenated and fed to a softmax layer, which produces an overall probability distribution over the tokens $(v_1 \dots v_k) \in V$, $(e_1 \dots e_n) \in G$ and $(f_1 \dots f_m) \in K$ (see the diagram in Figure 4). The token with the highest probability is generated at position t.

$$\begin{aligned} & w_t = \arg\max_{w_i} P(w_i), w_i \in V \cup G \cup K \\ & \text{where } P(w_i) = \\ & = \sigma_i(\text{Concat}[y_{v_1} \dots y_{v_k}, y_{e_1} \dots y_{e_n}, y_{f_1} \dots y_{f_m}]) \end{aligned} \quad (10)$$

6 Results and Discussion

We trained and tested our knowledge-aware captioning system on our dataset (described in Section 3). To evaluate its performance, we employ the standardly used metrics that compare the automatically generated captions to the human written captions for the same images: BLEU (Papineni et al., 2002) and its extensions, METEOR (Denkowski and Lavie, 2014), ROUGE (Lin, 2004) and CIDEr (Vedantam et al., 2015). In addition to that, we measure the correctness of the generated temporal facts by estimating their precision and recall.

For baselines, we train two other caption generation systems on the same dataset. Both baseline systems share the overall encoder-decoder Transformer-based architecture with our knowledge-aware system; however, we reduce the amount of context available to them. The first one, the "decontextualized" system, has both geographic and knowledge components removed, so, its performance represents the level that can be achieved by a standard image captioning pipeline with no additional contextualization. The second baseline, which we call "geo-aware", has access only to the geographic context: the output of the encoder is the concatenation of the image representation and the geographic context encoding, and during caption generation the decoder can only pick from the vocabulary tokens and the geographic entity names. The difference between the performance level of the geo-aware and the knowledge-aware systems will demonstrate the impact of the external knowledge component on the generated captions. We also run a standard

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr
Standard (Xu et al., 2015)	0.22	0.06	0.01	0.00	5.88	1.65	0.09
Decontextualized	20.64	10.42	5.41	2.81	23.19	7.77	2.12
Geo-aware	22.97	13.32	8.17	4.92	28.50	10.18	13.65
Knowledge-aware	**29.98**	**19.10**	**12.97**	**8.96**	**31.59**	**13.44**	**25.06**

Table 1: Metric scores of the different systems, measured on the test set.

Ground truth: bellgrove station , glasgow . opened in 1871 as a joint station to connect several of the main lines entering glasgow . view east towards airdrie .
Standard (Xu et al., 2015): a train traveling down train tracks next to a bridge
Decontextualized: *thatcham station* , *berkshire* . opened in *1847* by the great western railway on its line from reading to taunton . view south towards thatcham and taunton .
Geo-aware: **bellgrove station** , **glasgow** . opened in *1868* by the glasgow brighton & south coast railway on the line from glasgow bridge to croydon . view north towards glasgow .
Knowledge-aware: **bellgrove station** , **glasgow** . opened in **1871** by the glasgow & dorchester railway . view south east towards sway and dorchester .
Ground truth: hexham bridge . the current nine-arched stone bridge across the river tyne at hexham was built by william johnson and robert thompson , designed by engineer-architect robert mylne , following previous bridge designs by john smeaton but with piling and other measures to improve its stability . the bridge was completed between 1793 and 1795 . in 1967 the parapets were moved from their original position in order to widen the bridge to allow foot passengers . hexham bridge is a listed building grade ii* link link
Standard (Xu et al., 2015): a view of a bridge over a body of water
Decontextualized: *bewdley bridge* . *bewdley bridge* was designed by thomas telford and built in *1798*
Geo-aware: **hexham** . the bridge was designed by sir joseph bazalgette and opened in *1921* .
Knowledge-aware: **hexham bridge** , **hexham** . **the hexham bridge** was built in **1793** , the designs of architect reginald h . uren and cost of devonshire .

Table 2: Examples of the generated captions. Correct geographic references and temporal facts are given in **bold**; incorrect ones are given in *italics*.

pre-trained caption generation system (Xu et al., 2015) on our test set. The standard system has no contextual component and was trained on the out-of-domain images from the MSCOCO dataset (Lin et al., 2014). Table 1 shows the comparison between the metric scores of the knowledge-aware system and the three baselines (decontextualized, geo-aware and standard)[4].

The standard system trained on the out-of-domain images produces captions that are very different from the ground truth ones, which is reflected in the particularly low metric scores. This is expected, since the dataset it was trained on, as well as the architecture of the system itself, did not account for the context of the images and instead focused on their visual descriptions only. Overall, the metrics indicate that the more context is available to the system, the better it can reproduce the ground truth captions. The geo-aware system improves upon the decontextualized baseline, and the knowledge-aware system outperforms all the baselines across all metrics[5]. All the improvements are statistically significant (two-sample t-test, p <0.001). The most radical improvements are in the CIDEr metric, which gives a higher weight to the words that are more informative according to the TF-IDF score; geographic names and fact-related tokens are usually rare in the corpus and highly informative, so they contribute a lot to this metric.

Table 2 shows examples[6] of the captions generated by the knowledge-aware system and the baselines, as well as the original human written captions for the same images. Here, the standard system from Xu et al. (2015) successfully produces

[4]It would also be informative to compare our system to those that use alternative ways to encode and produce knowledge base entities during caption generation, for example, the ways proposed in general-purpose knowledge-aware language modeling; we leave the development of such systems and further comparison to future research.

[5]We also note that our system's metric scores are on par with those achieved on average by the other contextualized image captioning systems (Biten et al., 2019; Nikiforova et al., 2020; Tran et al., 2020; Bai et al., 2021), although a direct comparison is not possible due to the differences in the datasets and the task specifics.

[6]The images for these examples and additional examples from the test set are given in Appendix C.

accurate descriptions of what can be seen in the images but includes no references to their context. The decontextualized baseline system fails to generate correct geographic entity names and facts, as it has no access to the context of the images and simply draws all the caption tokens out of the general vocabulary. The geo-aware system can utilize the context available to it to produce accurate geographic references but, not being able to use the knowledge context, does not produce correct facts. The captions generated by the knowledge-aware system demonstrate that it is able to successfully use both geographic and knowledge contexts and produce relevant references to geographic entities and accurate temporal facts about them. Although it does produce incorrect facts from outside of the temporal domain (e.g. "designs of architect Reginald H. Uren" for the Hexham Bridge, which was actually designed by Robert Mylne), it is expected since the knowledge context only includes facts related to dates and years in the scope of this paper.

6.1 Generated Facts Accuracy

In our evaluation of the system, we specifically focus on the temporal facts in the generated captions. We test the correctness of the facts against the DBpedia knowledge base and measure precision and recall to quantify it. We take precision to be the number of times a correct temporal fact was generated, divided by the overall number of times any temporal fact was generated.

$$Precision = \frac{\textit{\# correct facts}}{\textit{\# all facts}} \tag{11}$$

We take recall to be the number of times a correct temporal fact was generated, divided by the number of times that the system generated a geographic entity that had a temporal fact in the knowledge context. A low value of recall would mean that a system does not generate temporal facts when they are available. This is not necessarily a fault in general; however, in this paper, the goal is to create a system with a high tendency to generate accurate temporal facts, which should correspond to a high level of recall.

$$Recall = \frac{\textit{\# correct facts}}{\textit{\# all geo entities with facts}} \tag{12}$$

In addition to the decontextualized and geo-aware baselines introduced earlier, we also create a "random fact" baseline. It takes the captions generated by the knowledge-aware system and replaces the fact token (the year) in each caption with a year randomly picked from the knowledge context. This creates quite a strong baseline because the probability of any year from the knowledge context to be relevant to the image and to appear in the caption is high by design. Table 3 shows the precision and recall scores of the three baselines and our knowledge-aware system.

	Precision	Recall
Decontextualized	0.0	0.0
Geo-aware	0.7	1.07
Random fact	48.75	46.96
Knowledge-aware	**84.40**	**81.31**

Table 3: Precision and recall scores.

Unsurprisingly, the geo-aware and the decontextualized baselines produce near to no accurate temporal facts, resulting in extremely low scores (the geo-aware system had a few coincidental correct guesses). The strong random fact baseline's scores are much higher, but are still greatly outperformed by the knowledge-aware system.

7 Conclusions

In order to imitate natural human behavior in captioning an image, it is essential that automatic image captioning systems take into account the context of the image and related real world knowledge. In this paper, we have presented a novel way to contextualize a standard image captioning pipeline with real world data that is relevant to the image but is not directly inferable from it. We compiled a new image captioning dataset with naturally produced knowledge-rich captions and image metadata. Our experiments demonstrate the effectiveness of our approach: the trained knowledge-aware captioning system is able to generate captions with accurate references to relevant geographic entities and correct temporal facts about them. Compared to a range of baseline systems, it achieves substantial improvements in the standardly used metrics as well as in the precision and recall of the generated facts. The proposed approach is not specific to any particular domain and could be generalized to a wide range of fact types. In future work, we plan to extend the coverage of our contextualized image captioning system to other knowledge domains, taking it further in the direction of truly humanlike caption generation.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

S Auer, C Bizer, G Kobilarov, J Lehmann, R Cyganiak, and Z Ives. 2007. Dbpedia: A nucleus for a web of open data. in et al., a., ed.: The semantic web. In *6th International Semantic Web Conference (ISWC), 2nd Asian Semantic Web Conference (ASWC)*, pages 715–728.

Zechen Bai, Yuta Nakashima, and Noa Garcia. 2021. Explain me the painting: Multi-topic knowledgeable art description generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5422–5432.

Ali Furkan Biten, Lluis Gomez, Marçal Rusiñol, and Dimosthenis Karatzas. 2019. Good News, Everyone! Context driven entity-aware captioning for news images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12466–12475.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Jingqiang Chen and Hai Zhuge. 2020. A news image captioning approach based on multimodal pointer-generator network. *Concurrency and Computation: Practice and Experience*, page e5721.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Hiroaki Hayashi, Zecong Hu, Chenyan Xiong, and Graham Neubig. 2020. Latent relation language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7911–7918.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Anwen Hu, Shizhe Chen, and Qin Jin. 2020. Icecap: Information concentrated entity-aware image captioning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4217–4225.

Feicheng Huang, Zhixin Li, Haiyang Wei, Canlong Zhang, and Huifang Ma. 2020. Boost image captioning with knowledge reasoning. *Machine Learning*, 109(12):2313–2332.

Yun Jing, Xu Zhiwei, and Gao Guanglai. 2020. Context-driven image caption with global semantic relations of the named entities. *IEEE Access*, 8:143584–143594.

Kuang-Huei Lee, Hamid Palangi, Xi Chen, Houdong Hu, and Jianfeng Gao. 2019. Learning visual relation priors for image-text matching and image captioning with neural scene graph generators. *arXiv preprint arXiv:1909.09953*.

Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. 2019. Entangled transformer for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8928–8937.

Xiangyang Li and Shuqiang Jiang. 2019. Know more say less: Image captioning based on scene graphs. *IEEE Transactions on Multimedia*, 21(8):2117–2130.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 740–755.

Angli Liu, Jingfei Du, and Veselin Stoyanov. 2019. Knowledge-augmented language model and its application to unsupervised named-entity recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1142–1150.

Robert Logan, Nelson F Liu, Matthew E Peters, Matt Gardner, and Sameer Singh. 2019. Barack's wife hillary: Using knowledge graphs for fact-aware language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5962–5971.

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.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383.

Aditya Mogadala, Umanga Bista, Lexing Xie, and Achim Rettinger. 2018. Knowledge guided attention and inference for describing images containing unseen objects. *European Semantic Web Conference*, pages 415–429.

Sofia Nikiforova, Tejaswini Deoskar, Denis Paperno, and Yoad Winter. 2020. Geo-aware image caption generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3143–3156, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. *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.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. *Thirty-First AAAI Conference on Artificial Intelligence*.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2443–2449, New York, NY, USA. Association for Computing Machinery.

Alasdair Tran, Alexander Mathews, and Lexing Xie. 2020. Transform and tell: Entity-aware news image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13035–13045.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. *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. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3156–3164.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Dalin Wang, Daniel Beck, and Trevor Cohn. 2019. On the role of scene graphs in image captioning. In *Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pages 29–34.

Spencer Whitehead, Heng Ji, Mohit Bansal, Shih-Fu Chang, and Clare Voss. 2018. Incorporating Background Knowledge into Video Description Generation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3992–4001.

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. *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. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659.

Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. 2019. Multimodal transformer with multi-view visual representation for image captioning. *IEEE Transactions on Circuits and Systems for Video Technology*.

Sanqiang Zhao, Piyush Sharma, Tomer Levinboim, and Radu Soricut. 2019. Informative image captioning with external sources of information. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6485–6494.

Yimin Zhou, Yiwei Sun, and Vasant Honavar. 2019. Improving image captioning by leveraging knowledge graphs. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 283–293.

Xinxin Zhu, Lixiang Li, Jing Liu, Haipeng Peng, and Xinxin Niu. 2018. Captioning transformer with stacked attention modules. *Applied Sciences*, 8(5):739.

A Dataset Split

We split out knowledge-aware captioning dataset into train, validation and test sets that constitute, respectively, 75%, 12.5% and 12.5% of the whole dataset. In order to avoid assigning different photographs of the same geographic entities to both train and validation/test sets, we base the split on the latitude of the image location instead of splitting the dataset randomly. The photographs that were taken to the north of the 54.8287° latitude are assigned to the test set, between the 53.534° and the 54.8287° latitude to the validation set, and the rest to the train set. With the latitude-based split, we ensure testing on the previously unseen data, which helps to detect possible overfitting.

B Mistake Analysis

The most typical mistake that the knowledge-aware system makes in temporal fact generation is producing a year that refers to a different event than what is stated in the caption, e.g. generating "st magnus cathedral [...] built in 1137" when in fact St. Magnus Cathedral was *founded* in 1137 and the year of construction is not specified in the knowledge context, see Table 4.

https://www.geograph.org.uk/photo/5987415

Ground truth: st magnus cathedral . established in 1137 , the cathedral is constructed of orkney red and yellow sandstone .

Knowledge-aware: st magnus cathedral , kirkwall . st magnus cathedral , built in 1137 , the village of kirkwall .

Knowledge context fragment

…
f_i – 1137 –
< St Magnus Cathedral, founded_in >
…

Table 4: An example of the caption generated with an incorrect fact.

This type of mistake occurs when the fact's predicate does not fit the previously generated tokens, e.g. the fact's predicate "founded_in" does not fit the previously generated expression "built in". Since there is no fact that refers to the cathedral's year of construction in the knowledge context at all, generating "built in" in the caption leads to a particularly high chance of producing an incorrect fact. Thus, this type of mistake highlights the importance of taking the knowledge context into account not only when the fact is being selected but also while generating other types of tokens, such as regular vocabulary words.

C Examples of the Generated Captions

(a)		Ground truth: bellgrove station , glasgow . opened in 1871 as a joint station to connect several of the main lines entering glasgow . view east towards airdrie . Standard (Xu et al., 2015): a train traveling down train tracks next to a bridge Decontextualized: *thatcham station* , *berkshire* . opened in *1847* by the great western railway on its line from reading to taunton . view south towards thatcham and taunton . Geo-aware: **bellgrove station** , **glasgow** . opened in *1868* by the glasgow brighton & south coast railway on the line from glasgow bridge to croydon . view north towards glasgow . Knowledge-aware: **bellgrove station** , **glasgow** . opened in **1871** by the glasgow & dorchester railway . view south east towards sway and dorchester .
(b)		Ground truth: hexham bridge . the current nine-arched stone bridge across the river tyne at hexham was built by william johnson and robert thompson , designed by engineer-architect robert mylne , following previous bridge designs by john smeaton but with piling and other measures to improve its stability . the bridge was completed between 1793 and 1795 . in 1967 the parapets were moved from their original position in order to widen the bridge to allow foot passengers . hexham bridge is a listed building grade ii* link link Standard (Xu et al., 2015): a view of a bridge over a body of water Decontextualized: *bewdley bridge* . *bewdley bridge* was designed by thomas telford and built in *1798* Geo-aware: **hexham** . the bridge was designed by sir joseph bazalgette and opened in *1921* . Knowledge-aware: **hexham bridge** , **hexham** . **the hexham bridge** was built in **1793** , the designs of architect reginald h . uren and cost of devonshire .
(c)		Ground truth: alnwick playhouse . built in 1925 as a cinema , the playhouse now presents a balanced programme of music , drama , dance , film and exhibitions . see link (archive link) Standard (Xu et al., 2015): a red and white building on a street corner Decontextualized: *the queens house* , *london* . *the queens house* was built in *1616* and is a grade i listed building . link Geo-aware: **alnwick** : **alnwick playhouse** . **the alnwick playhouse** is a theatre , **alnwick** complex , built in *1907* and is grade ii listed . link Knowledge-aware: **alnwick playhouse** . **the alnwick playhouse** is a grade ii listed building , built in **1925** . link
(d)		Ground truth: floodlit paisley town hall . looking along gauze street towards paisley cross . the town hall was built in 1882 to a design by belfast architect w.h.lynn and was funded by the coats threads family . Standard (Xu et al., 2015): a city at night with a clock tower Decontextualized: *st michaels church* , *southampton* . the church was built in *1070* , and is the oldest building in *1070* , it is a grade i listed building . it is now part of the oldest building . it is now grade i listed building . Geo-aware: the church of *paisley abbey* . the church was consecrated in *1873* . Knowledge-aware: **the paisley town hall** , **paisley** . **the paisley town hall** is a grade ii listed building , completed in **1882** , with details at : link

Table 5: Examples of the generated captions. Correct geographic references and temporal facts are given in **bold**; incorrect ones are given in *italics*. Correctness of the non-temporal facts is not assessed.

Image references:

(a) https://www.geograph.org.uk/photo/3373434

(b) https://www.geograph.org.uk/photo/5715314

(c) https://www.geograph.org.uk/photo/2459525

(d) https://www.geograph.org.uk/photo/2793844