Salient Image Caption Generation with Initial Sentences of Localized Narratives

Pat Healy
PhD Student in Information Science
University of Pittsburgh School of Computing and Information
pat.healy@pitt.edu

1 Localized Narratives

The *Localized Narratives* multimodal image annotation protocol and dataset (Pont-Tuset et. al., 2020) offers a clear opportunity for work in salient image caption generation. The dataset pairs image descriptions with sets of coordinates that represent the locations in the image that are being described by a given sub-string of the description. This project focuses on describing the dataset itself, implementing a captioning model to incorporate it for the task of salient image description generation, and assessing the performance of that model.

1.1 Localized Narratives Annotations

Annotations in the *Localized Narratives* dataset come in seven parts: dataset id, image id, annotator id, caption, timed caption, traces, and voice recording. The dataset and image id come to define the image this annotation is describing, indicating the external dataset the image is retrieved from and the specific image within that dataset. The annotator id identifies the worker who created this annotation. The caption is the simple full text of the image description created by the worker. The timed caption breaks the full text of the caption into a list of tuples representing utterances, which includes the text of that utterance and its start and end time. The traces are a list of tuples indicating image coordinates of the cursor and a timestamp of that placement. Voice recording indicates the url path to a recording of the annotator speaking the caption.

1.2 Analysis of Localized Narratives

In the work that introduced this dataset, there was significant analysis to understand the quality of the data itself (Pont-Tuset et. al., 2020). This was primarily semantic and localization accuracy, which primarily concerned itself with whether the nouns and verbs annotators used to describe the images were actually present in those images and whether the location of their cursor matched the objects the text of their utterances were describing. Those researchers found the dataset to score particularly high in both of these measures, with 98.0% semantic accuracy and only a very small portion of mouse trace segments fell outside bounding boxes.

Additionally, the researchers analyzed richness and diversity of their dataset. Richness analysis focused on mean length of descriptions and mean number of nouns, pronouns, adjectives, verbs, and adpositions; this dataset included higher mean length and mean occurrence of those categories, which the researchers choose to interpret as a richer use of natural language in connection to the images. Diversity analysis focused on simply comparing the plotted number of nouns per caption to similar distributions from other datasets; this dataset was found to be noticeably more diverse than most other datasets surveyed, but comparable to the Stanford Visual Paragraphs.

It seems to me that these measures of quality have a very clear blindspot, which may or may not actually have any meaningful impact on the performance of any applications that use it. These measures have no ability to capture whether the given captions describe *important* information. Take, for instance, the included example from the dataset, named 'Route 9 Dinosaur' in its original dataset (Figure 1).

This image's caption was "In this image I can see the ground, few buckets, few poles, the orange colored net, few persons standing, few boards and a building. I can see few lights to the building and a board. I can see the sky to the right top of the image." This description would score relatively high in the measures the researchers used to assess their dataset, seeing as it explains, in many words, a series of elements that do indeed appear in the image, and the trace provided is similarly high quality, mapping directly to the elements the annotator

Figure 1: An example image from the Localized Narratives dataset. Image source: Route 9 Dinosaur. Author: LancerE. Dataset: Open Images. ID: 7ca4fbc972ee3e4a.

has listed. However, a layperson would likely see this image and note that it is an image of a blue dinosaur, an obvious detail completely omitted from the annotator's caption.

Though this is necessarily only anecdotal evidence from my own brief exploration of the dataset, it seems this is an issue the occurs frequently. Though we would hope the annotator would begin the annotation process by describing salient imagery, this clearly does not seem to be the case in this dataset.

If this were a typical captioning dataset that gave only the simple text caption for a given image, this may be a problem, since we may typically care to have data that indicates primarily the most important objects in an image. However, in this particular dataset it may not actually be a disqualifying fault. Clearly, the dataset provides adequate information for generating descriptions given an image and cursor path through the image. This means it could be used for accurate salient image captioning, but only if one was able to identify a salient path separately; this dataset likely does not provide adequate data to identify salient objects themselves.

2 Captioning

2.1 Theoretical Justification

Given the issue described in the previous section, namely that *Localized Narratives* does not necessarily include descriptions of salient objects, a best-case captioning model would require some other mechanism, unrelated to *Localized Narratives* to identify these salient objects. I imagine a method that uses some common method of salient object detection to determine a trace segment to give to Pont-Tuset et. al.'s controlled captioning model. According to an issue reply on the *Localized Narratives* GitHub repository (https://github.com/google/localized-narratives/issues/4), they should be releasing this controlled captioning model code in the near future, which would make this particular method trivial to implement. Lacking this code as a starting point, I instead chose to implement a much less intelligent model, if only as an experiment to test a hypothesis my own analysis has already doubted.

To make another wildly anecdotal observation, among the *Localized Narratives* captions that did successfully mention salient objects it seems they tended to describe the most salient objects in their first sentence. From this, it seems reasonable that a captioning model trained only on the first sentences of the *Localized Narratives* captions may give a rough estimate of salient captioning. Thus, this is my chosen method for this project. I chose to not include the traces themselves in my model. Though they're obviously critical to the *Localized Narratives* dataset itself and what I previously described as my optimal solution, I see no obvious way to integrate them into this particular model, which we hope to use to generate captions for images which we have no trace[1].

2.2 The Dataset

The model was trained and tested on the Flickr30K training and testing sets provided by *Localized Narratives*. Unfortunately, due to time and hardware constraints, I had to greatly limit the number of items to be used in training. Rather than include the full almost thirty thousand training annotations and one thousand testing annotations, it was only feasible to use 100 of each set. Of course, one with greater time and/or hardware resources could easily train with more data [2].

Since I was interested in understanding whether looking only at the first sentences of the training set captions would improve the model's salience,

[1]Full disclosure: it is very likely there is an obvious way of integrating the trace data into the model that I am not seeing, either by my lack of expertise or creativity. I am, unfortunately, very new to the world of machine learning.

[2]To do this, one only needs to change lines 62 and 70 of prepare_text.py

I ended up training two models: one with the full captions from the *Localized Narratives* Flickr30k training set and another with only those captions' first sentence, found by performing substring operations from the start of the caption to the first instance of a period[3].

2.3 The Model

As a starting point, I used a captioning model developed for an online tutorial in Keras by Jason Brownlee (Brownlee, 2017). The structure of this model is shown in Figure 2.

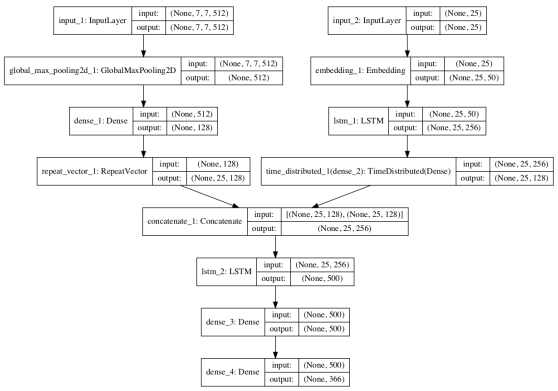

Figure 2: Plot of the captioning model sourced from Brownlee (Brownlee, 2017).

Due to hardware and time constraints, the models were both unfortunately trained much less than is ideal. One model, which I'm calling "First Sentences" was trained with only the first sentence of each caption in the training set. The second model, which I'm calling "Full Text" was trained with the full text of the caption. Both models were trained for 50 epochs, with 3 images per update, and 33 batches per epoch.

3 Evaluation

The two models were each trained and evaluated for three trials. They were primarily evaluated based on calculated Bilingual Evaluation Understudy (BLEU) score for each trial against both the training and test set. The table in Figure 3 shows the results.

Given that the first sentences model performed with slightly higher average BLEU Scores evaluated on the test set and much higher average BLEU scores evaluated on the training set, we may be led to support my initial hypothesis that training only on first sentences will lead to improved salient captions. Of course, there are several problems with this conclusion. For one, this was a very small number of trials with a particularly small dataset, clearly not up to the standard of rigor required to make strong comparisons. More importantly, though, we can't really understand this measure to be capture the thing we most clearly care about: salience.

[3]This was done by running the included scripts twice, modifying only the boolean variable on line 8 of prepare_text.py

	First Sentences		Full Text	
Trial	Train	Test	Train	Test
1	0.046	0.019	0.027	0.019
2	0.068	0.033	0.021	0.014
3	0.051	0.011	0.030	0.019
AVG	0.055	0.021	0.026	0.017

Figure 3: Table summarizing BLEU Score values for First Sentences and Full Text Models across three trials.

To my knowledge, there is no clear, quantifiable standard measure of salience. As in the Localized Narratives dataset, we may be able to say whether or not the information in our captions accurately describes our images, but we don't really understand if the objects the text is describing represents the most important objects in the image. To give some idea, we can at least qualitatively observe examples and reach anecdotal conclusions.

Figure 4: Image 2973269132 from the Flickr30k set. In the Localized Narratives test set, its caption is "in this picture can see lion running and trying to catch an animal on the right side can see the grass on the ground and the blurry background".

Take, as an example, the image in Figure 4.

The first sentences model generated the sentence "in this picture see dog like dog". Clearly this wouldn't be mistaken for an accurate description of the image, but it's still somewhat clear where the "dog" came from. As contrast, the following was a generated caption by the full text model: "in this picture can see dog wearing the and in the and and on the and and sky on the and and sky on the and and sky on the and and sky on the and and sky on the and and sky on the and and sky on the and and sky on the and and sky on the and and sky on the and and sky".

Clearly there are significant issues present in the full text model that are likely only the result of under-training the model, born out of hardware and time limitations. But still, in the contrast between these two captions, we can clearly see that by simply generating a shorter caption based on initial information, the model seems to be more likely to produce salient information and, more precisely, not provide extraneous information. Of course, this work does not have the rigor to make particularly strong conclusions, but as an exploratory work, it shows some promise.

4 Personal Takeaways

This was an incredibly difficult project for me, as someone with relatively little experience with Machine Learning in Python. Previously, I had only ever implemented very basic classification models in Python and taken a particularly theory-focused Machine Learning course with implementation only in R[4]. This project definitely pushed me outside of my comfort zone in a way that is probably essential, given my field[5].

We are clearly beyond me applying this work specifically to my main group project for the course, given my unfortunate delay in reaching this work, but I'd say I'm definitely going to apply what I learned in this project in other work. This was essentially my introduction to Keras and TensorFlow, and that seems to be a particularly useful technology.

[4]R proved both completely impractical and mostly painful to work in. I am very thankful to have had this project to push me into Python ML.

[5]My research interests often feel the least technical of my department.

Acknowledgments

I'd like to thank Malihe Alikhani for her thoughtful support throughout the course this work was completed for, as well as leniency in the face of NDS-related delays. I'd also like to thank Tristan Maidment and Anthony Sicilia for offering guidance on machine learning methods.

References

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. *Connecting Vision and Language with Localized Narratives*. ECCV (Spotlight).

Jason Brownlee 2017. *How to Use Small Experiments to Develop a Caption Generation Model in Keras*. MachineLearningMastery.com. Retrieved from https://machinelearningmastery.com/develop-a-caption-generation-model-in-keras/