7 Appendix

This appendix primarily provides extra details on the model and data collection process. This is included to enusre our results are easily reproducable and to clarify exactly how the data was collected.

We first provide additional details on the LSTM units used by our approach in Section 7.1. Section 7.2 discusses the differences between 1^{st}, 2^{nd}, and 3^{rd} person sentiment. See Section 7.3 for a discussion of how the ANPs with sentiment where chosen. For details on rewriting sentences to incorporate ANPs see Section 7.4. Details on validating the rewritten sentences are in Section 7.5. The crowd sourced evaluation of generated sentences is described in Section 7.6.

7.1 The LSTM unit

The LSTM units we have used are functionally the same as the units used by Vinyals et al. (2015). This differs from the LSTM unit used by Xu et al. (2015a) because we do not concatenate contextual information to the units input. A graphical representation of our LSTM units is shown in Figure 5; for a more complete definition see Equation 2 in the companion paper. In Figure 5, note that only the LSTM unit is shown, without the fully connected output layers or word embedding layers.

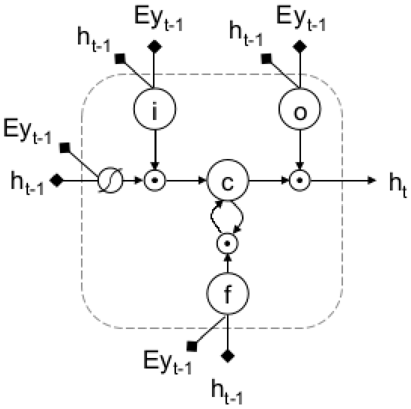
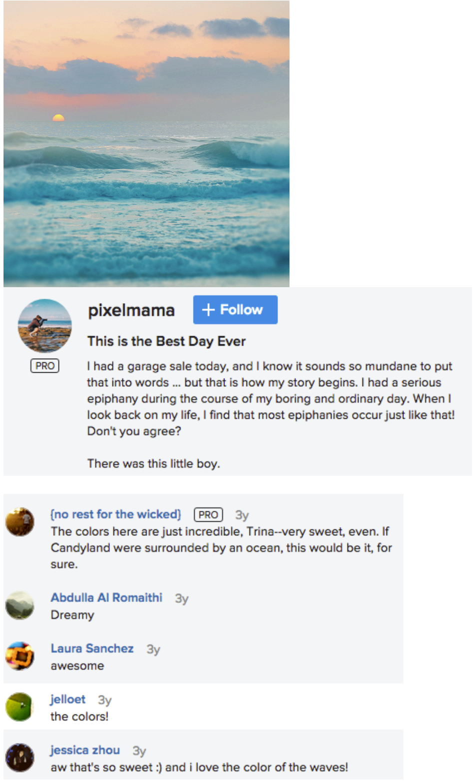


Figure 5: LSTM unit used in our paper, as in Equation 2. The filled diamond and square blocks on input nodes represent learn-able weights; in this case parts of the $\mathbf{T}^k$ matrix. Note that the weights on these inputs are not the same, they are learned separately.

7.2 Sentimental descriptions in the first, second, and third person

There are many ways a photo could evoke emotions, they can be referred to as sentiments from the first, second, and third person.

A first person sentiment is for a photo to elicit the emotions of its owner / author / uploader, who then records such sentiment for personal organization or communication to others (Ames and Naaman 2007). Such as the Flickr photo titled "This is the best day ever"[1], see Figure 6. The title and the caption describes a story but not the contents of the photo.



pixelmama

This is the Best Day Ever

I had a garage sale today, and I know it sounds so mundane to put that into words ... but that is how my story begins. I had a serious epiphany during the course of my boring and ordinary day. When I look back on my life, I find that most epiphanies occur just like that! Don't you agree?

There was this little boy.

{no rest for the wicked} PRO 3y
The colors here are just incredible, Trina--very sweet, even. If Candyland were surrounded by an ocean, this would be it, for sure.

Abdulla Al Romaithi 3y
Dreamy

Laura Sanchez 3y
awesome

jelloet 3y
the colors!

jessica zhou 3y
aw that's so sweet :) and i love the color of the waves!

Figure 6: The Flickr photo as discussed in Section 7.2. The title and caption are an example of first person sentiment, where a story is told rather than describing the contents of the photo. The comments are second-person sentiments.

A second person sentiment is expressed by someone whom the photo is *communicated to*,such as the comments "awesome" and "so sweet" for the photo above.

The third person sentiment is one expressed by an objective viewer, who has information about its visual content but does not know the backstory, such as describing the photo above as "Dreamy sunset by the sea".

It will be difficult to learn the correct sentiments for the first or second person, since the computer lacks knowledge of the personal and communication context – to the extent that a change in context and assumptions could completely flip the polarity of the sentiment (See Figure 3). In this work, we focus on learning possible sentiments from the third person. We collect descriptions with sentiment by people who are asked to describe them – this setting is close to that of recent collections of subjectively descriptive image captions (Chen et al. 2015; Hodosh, Young, and Hockenmaier 2013).

7.3 Customizing Visual Sentibank for captions

Visual SentiBank (Borth et al. 2013) is a database of Adjective-Noun Pairs (ANP) that are frequently used to describe online images. We adopt its methodology to build the

[1] https://www.flickr.com/photos/pixelmama/7612700314/

sentiment vocabulary. We take the title and the first sentence of the description from the YFCC100M dataset (Thomee et al. 2015), keep entries that are in English, tokenize, and obtain all ANPs that appear in at least 100 images. We score these ANPs using the average of SentiWordNet (Esuli and Sebastiani 2006) and SentiStrength (Thelwall et al. 2010), with the former being able to recognize common lexical variations and the latter designed to score short informal text. We keep ANPs that contain clear positive or negative sentiment, i.e., having an absolute score of 0.1 and above. We then take a union with the Visual SentiBank ANPs. This gives us 1,027 ANPs with a positive emotion, 436 with negative emotions. A full set of these ANPs are released online, along with sentences containing these ANPs written by AMT workers.

7.4 AMT interface for collecting image captions with sentiment

We went through three design iterations for collecting relevant and succinct captions with the intended sentiment.

Our first attempt was to invite workers from Amazon Mechanical Turk (AMT) to compose captions with either a positive or negative sentiment for an image – which resulted in overly long, imaginative captions. A typical example is: "*A crappy picture embodies the total cliche of the photographer 'catching himself in the mirror,' while it also includes a too-bright bathroom, with blazing white walls, dark, unattractive, wood cabinets, lurking beneath a boring sink, holding an amber-colored bowl, that seems completely pointless, below the mirror, with its awkward teenage-composition of a door, showing inside a framed mirror (cheesy, forced perspective,) and a goofy-looking man with a camera.*"

We then asked turkers to place ANPs into an existing caption, which resulted in rigid or linguistically awkward captions. Typical examples include: "a bear that is inside of the great water" and "a bear inside the beautiful water".

These prompts us to design the following re-writing task: we take the available MSCOCO captions, perform tokenization and part-of-speech tagging, and identify nouns and their corresponding candidate ANPs. We provide ten candidate ANPs with the same sentiment polarity and asked AMT worker to rewrite *any one of the original captions* about the picture using at least one of the ANPs. The form that the AMT workers are shown is presented in Figure 7. We obtained three positive and three negative descriptions for each image, authored by different Turkers. As anecdotal evidence, several turkers emailed to say that this task is *very interesting*.

The instructions given to workers are shown in Figure 7. We based these instructions on those used by Chen et al. (2015) to construct the MSCOCO dataset. They were modified for brevity and to provide instruction on generating a sentence using the provided ANPs. We found that these instructions were clear to the majority of workers.

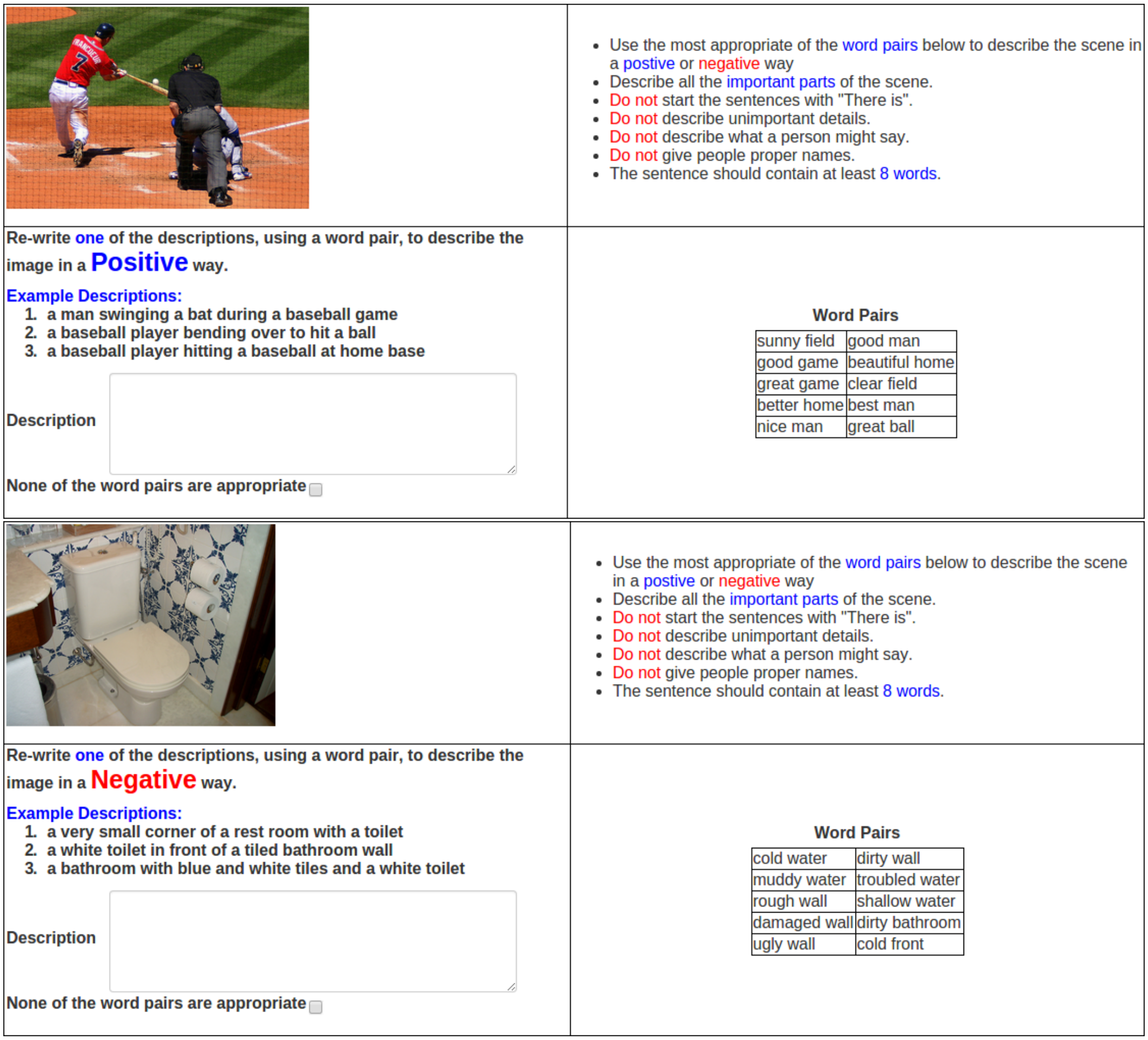
7.5 AMT interface validating image captions with sentiment

The AMT validation interface, in Figure 8 was designed to determine what effect adding sentiment into the ground truth captions effects their descriptiveness. Additionally we wanted to understand the fraction of images that could reasonably be described using either positive or negative sentiment. Each task presents the user with three MSCOCO captions and three positive or negative sentences, and asks users to rate them. Our four point descriptiveness scale is based on schemes used by other authors (Hodosh, Young, and Hockenmaier 2013; Vinyals et al. 2015).

7.6 AMT interface for rating captions with a sentiment

The AMT rating interface shown in Figure 9 was used to evaluate the performance of the four different methods. Each task consists of three different types of rating: most positive, most interesting and descriptiveness. The most positive and most interesting ratings are done pair-wise, comparing a sentence generated from one of the four methods to a sentence generated by *CNN+RNN*. The descriptiveness rating uses the same four point scale as the validation interface from Section 7.5. There are 5 images to rate per task; this is essential because of the way AMT calculates prices.

We found that asking Turkers to rate sentences using this method initially produced very poor results, with many Turkers selecting random options without reading the sentences. We suspect that in a number of cases bots were used to complete the tasks. Our first solution was to use more skilled Turkers, called masters workers, although this lead to cleaner results the smaller number of workers meant that a large batch of tasks took far too long to complete. Instead we used workers with a 95% or greater approval rating. To combat the quality issues we randomly interspersed the manual sentiment captions from our dataset, and then rejected all tasks from worker who failed to achieve 60% accuracy for the most positive rating. This was found to be an effective way of filtering out the results. We note that there were very few cases where workers were close to the 60% accuracy cut-off, they were typically much higher or much lower than the threshold, this validates the idea that some workers were not completing the task correctly.



- Use the most appropriate of the word pairs below to describe the scene in a postive or negative way
- Describe all the important parts of the scene.
- Do not start the sentences with "There is".
- Do not describe unimportant details.
- Do not describe what a person might say.
- Do not give people proper names.
- The sentence should contain at least 8 words.

Re-write one of the descriptions, using a word pair, to describe the image in a Positive way.

Example Descriptions:

1. **a man swinging a bat during a baseball game**
2. **a baseball player bending over to hit a ball**
3. **a baseball player hitting a baseball at home base**

Description

None of the word pairs are appropriate □

Word Pairs

sunny field	good man
good game	beautiful home
great game	clear field
better home	best man
nice man	great ball

- Use the most appropriate of the word pairs below to describe the scene in a postive or negative way
- Describe all the important parts of the scene.
- Do not start the sentences with "There is".
- Do not describe unimportant details.
- Do not describe what a person might say.
- Do not give people proper names.
- The sentence should contain at least 8 words.

Re-write one of the descriptions, using a word pair, to describe the image in a Negative way.

Example Descriptions:

1. **a very small corner of a rest room with a toilet**
2. **a white toilet in front of a tiled bathroom wall**
3. **a bathroom with blue and white tiles and a white toilet**

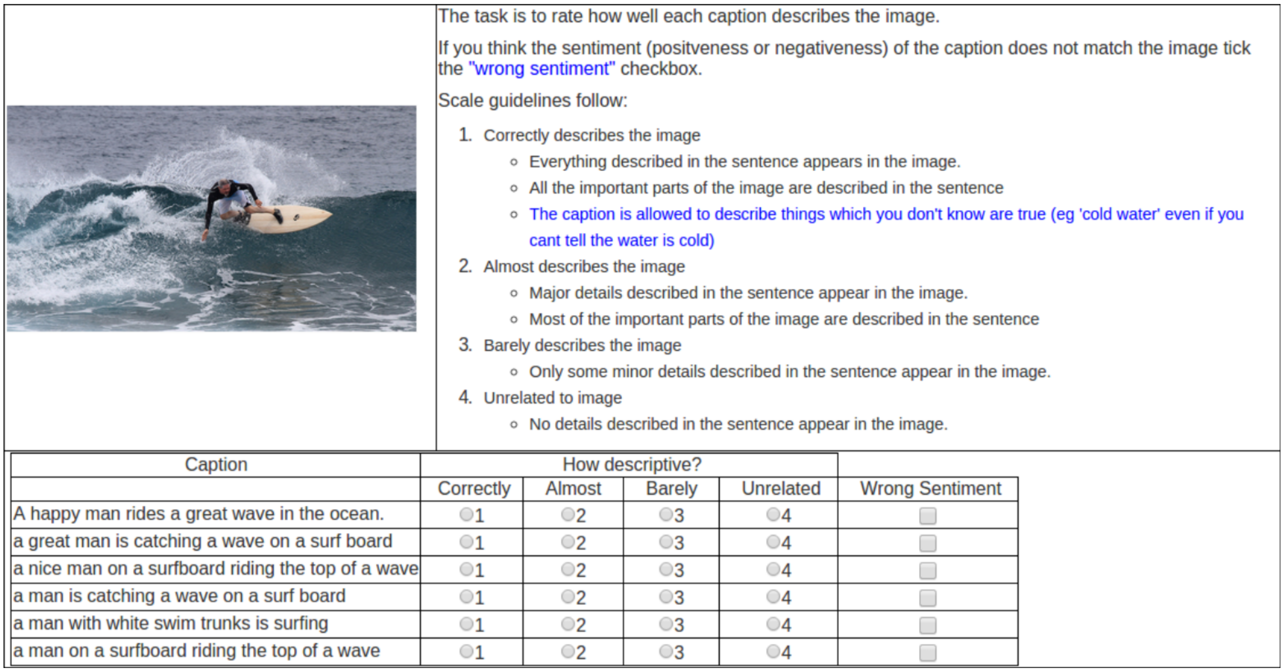
Description

None of the word pairs are appropriate □

Word Pairs

cold water	dirty wall
muddy water	troubled water
rough wall	shallow water
damaged wall	dirty bathroom
ugly wall	cold front

Figure 7: Mturk interfaces and instructions for *Collecting* sentences with a positive (top) and negative (bottom) sentiment.



The task is to rate how well each caption describes the image.

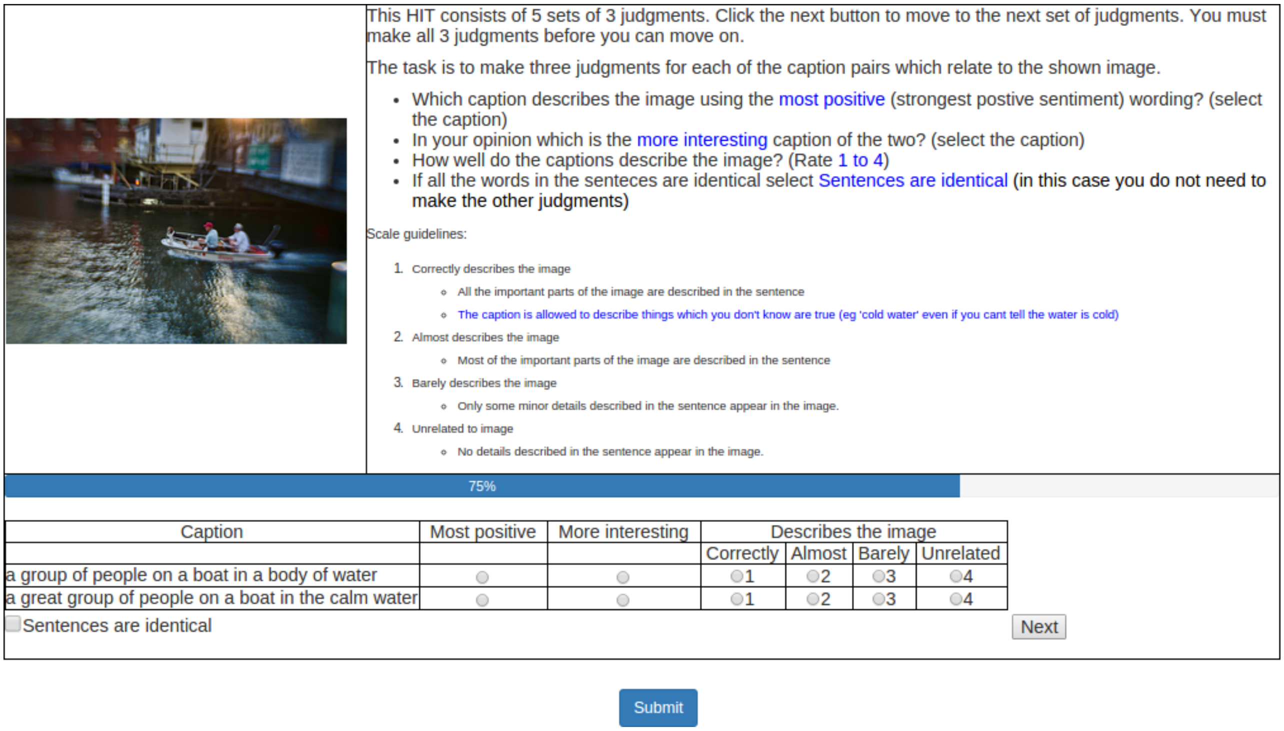
If you think the sentiment (positveness or negativeness) of the caption does not match the image tick the "wrong sentiment" checkbox.

Scale guidelines follow:

1. Correctly describes the image
   - Everything described in the sentence appears in the image.
   - All the important parts of the image are described in the sentence
   - The caption is allowed to describe things which you don't know are true (eg 'cold water' even if you cant tell the water is cold)
2. Almost describes the image
   - Major details described in the sentence appear in the image.
   - Most of the important parts of the image are described in the sentence
3. Barely describes the image
   - Only some minor details described in the sentence appear in the image.
4. Unrelated to image
   - No details described in the sentence appear in the image.

Caption	How descriptive?				
	Correctly	Almost	Barely	Unrelated	Wrong Sentiment
A happy man rides a great wave in the ocean.	○1	○2	○3	○4	□
a great man is catching a wave on a surf board	○1	○2	○3	○4	□
a nice man on a surfboard riding the top of a wave	○1	○2	○3	○4	□
a man is catching a wave on a surf board	○1	○2	○3	○4	□
a man with white swim trunks is surfing	○1	○2	○3	○4	□
a man on a surfboard riding the top of a wave	○1	○2	○3	○4	□

Figure 8: AMT interface and instructions for *Rating Groudtruth* sentences



This HIT consists of 5 sets of 3 judgments. Click the next button to move to the next set of judgments. You must make all 3 judgments before you can move on.

The task is to make three judgments for each of the caption pairs which relate to the shown image.

- Which caption describes the image using the most positive (strongest postive sentiment) wording? (select the caption)
- In your opinion which is the more interesting caption of the two? (select the caption)
- How well do the captions describe the image? (Rate 1 to 4)
- If all the words in the senteces are identical select Sentences are identical (in this case you do not need to make the other judgments)

Scale guidelines:

1. Correctly describes the image
   - All the important parts of the image are described in the sentence
   - The caption is allowed to describe things which you don't know are true (eg 'cold water' even if you cant tell the water is cold)
2. Almost describes the image
   - Most of the important parts of the image are described in the sentence
3. Barely describes the image
   - Only some minor details described in the sentence appear in the image.
4. Unrelated to image
   - No details described in the sentence appear in the image.

75%

Caption	Most positive	More interesting	Describes the image			
			Correctly	Almost	Barely	Unrelated
a group of people on a boat in a body of water	○	○	○1	○2	○3	○4
a great group of people on a boat in the calm water	○	○	○1	○2	○3	○4

□Sentences are identical

Figure 9: AMT interface and instructions for *comparative rating* of generated sentiment sentences