# SentenceRacer: A Game with a Purpose for Image Sentence Annotation

**Kenji Hata*, Sherman Leung*, Ranjay Krishna, Michael S. Bernstein, Li Fei-Fei**
Stanford University
{kenjihata, sherman, rak248, msb, feifeili}@cs.stanford.edu

## ABSTRACT
Recently datasets that contain sentence descriptions of images have enabled models that can automatically generate image captions. However, collecting these datasets are still very expensive. Here, we present SentenceRacer (SR), an online game that gathers and verifies descriptions of images at no cost. Similar to the game hangman, players compete to uncover words in a sentence that ultimately describes an image. SR both generates and verifies that the sentences are accurate descriptions. We show that SR generates annotations of higher quality than those generated on Amazon Mechanical Turk (AMT).

## ACM Classification Keywords
H.5.m. Design, Human Factors

## Author Keywords
Crowdsourcing; Games; Annotation; Tagging; Images

## INTRODUCTION
The ability of describing images with sentences has numerous applications like helping the visually impaired independently browse the Internet. Recently, with competitions like Microsoft COCO's Image Captioning [2], there has been an increased interest in the task of automatic image description generation [4]. With this interest, there is a dire need for large scale datasets that can be used for training these sentence generation models. Datasets like COCO [2] and Flickr30M [6] have been collected by crowdsourcing the description task to human workers on Amazon Mechanical Turk. Once the sentences are generated by one crowd worker, both [2, 6] send their sentences to additional crowd workers to verify the accuracy of the sentences. The biggest bottleneck in growing these datasets to a much larger scale has been the cost of generating these sentences and verifying their accuracy.

Humans are strikingly proficient at "filling in the blanks" — whether it be crosswords, hangman, or Wheel of Fortune. We enjoy partially filled puzzles because of the feeling of simultaneously knowing and not knowing the full answer [3]. Previous research by von Ahn and Dabbish show gamification to be an effective vehicle for labeling images in datasets for free [5]. However, such games were only limited to single word annotations [5]. This paper explores another game mechanism in order to generate more complex, full-sentence annotations. We propose a gamification method to crowdsource sentence annotations of images by having players write sentences for other players to guess. Additionally, we show that there exists a direct correlation between the accuracy of the sentence description and a player's ability to guess the sentence.

Motivated to reduce the cost of collecting a large image captioning dataset, we present a game that achieves the following: (1) Generates sentence descriptions of images (2) Verifies that these sentences are accurate (3) Captures sentences of better quality than those collected by Amazon Mechanical Turk (AMT).

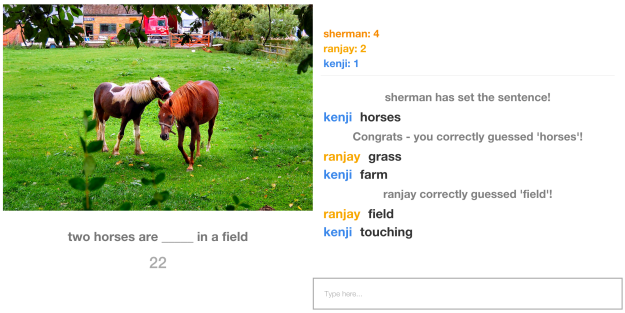


**Figure 1. A screenshot of SR's interface. The left side displays the image and the state of the verified words so far. The right side displays the chat interface and the scoreboard for guessing.**

## SYSTEM DESCRIPTION
SR is played with a minimum of three players. Each round of the game rotates a leader position. The leader sets a sentence describing the image for other players to guess. After eliminating stop words from the sentence, we allow all players to see guesses made by other players while blocking the leader's communication with the guessers. Players only have limited time to guess the words set by the leader. A correct guess rewards both the guesser and the leader with points and reveals the guess' position in the sentence. This reward system implicitly motivates the leader to write descriptive sentences about the image, as they will be easily guessed by the other players.

*These authors contributed equally to the publication.



UIST13 Adjunct, October 811, 2013, St. Andrews, United Kingdom.
ACM 978-1-4503-2406-9/13/10.
http://dx.doi.org/10.1145/2508468.2514726

## DATA AND ANALYSIS

To gather the data, we took ten groups of four volunteers and ran each group on the same ten randomly sampled images from Microsoft's COCO dataset [2].

### People Find SR to be more fun

Qualitative results suggest that SR is more fun in comparison to the task of image captioning. Surveys comparing SR and a standard AMT image captioning task show that players find SR more fun and engaging particularly because of the social and fast-paced aspects of the game.

### SR's Sentences are Confirmed by AMT

Sentences collected by SR were sent to AMT for verification by three crowd workers. A sentence is verified by AMT if at least two out of the three workers agree that the sentence accurately describes the image. A sentence is considered verified by SR if all the words in the sentence can be guessed by the players. We found that 87.8% of sentences verified by SR were also verified by AMT, while only 54.9% of the sentences not verified by SR were verified by AMT. We also found that the sentences collected from SR have a higher percentage of verified sentences (87.8%) than those collected from AMT workers (85.5%) on the same images.

We investigated the relation of the percentage of sentences verified with the number of remaining blanks left in the game. Table 2 shows that the percentage verified increases as the number of blanks decreases. The tail end of the blanks is sparse, causing the data to have high variance. However, we believe that this trend still shows that SR's verification process adequately determines whether a sentence accurately describes an image. The number of blank spaces are directly correlated with how likely a sentence will be verified.

| Source | Total Blanks | # Sentences | Verified (%) |
|---|---|---|---|
| SR | 4 | 7 | 42.80 |
| | 3 | 12 | 50.00 |
| | 2 | 9 | 33.30 |
| | 1 | 12 | 75.00 |
| | 0 | 49 | **87.80** |
| AMT | - | 200 | 85.50 |

**Table 1. Correlation between number of blanks and percentage of verified sentences. As the number of blanks decreases, the percentage of sentences verified by AMT increases. Also in comparison, sentences collected from AMT have a lower verification percentage than the sentences collected by SR.**

### SR has Higher Sentence Quality than AMT

Figure 2 shows the quality of some sentences we received from both tasks on AMT and from playing SR. We measure sentence quality by the amount of information we can extract from the sentence describing an image. The average number of objects, object attributes, and pairwise-object relationships per sentence is a basic indicator of sentence quality [1]. Table 2 shows that SR has statistically significant more objects and relationships and may suggest that SR provides more attributes as well. We believe SR's sentence quality stems from the rule that correct guesses reward both the guesser and the leader. Players are incentivized to write longer sentences, leading to higher averages of objects, relationships, and attributes, than AMT tasks, where this incentive is absent.


**Figure 2. Comparing verified sentences from AMT and SR .**

| | Objects | Relationships | Attributes |
|---|---|---|---|
| **AMT** (n=200) | 2.30 | 1.02 | 1.17 |
| **SR** (n=49) | 2.98 | 1.88 | 1.45 |
| P-values | $< 0.01$ | $< 0.001$ | 0.1 |

**Table 2. T-test showing that increased number of objects, relationships, and (potentially) attributes.**

## CONCLUSION

In this paper, we demonstrate how SR is able to collect sentences describing images, an expensive task for Computer Vision research. This system introduces the idea of collecting and verifying sentences through a game that uses contextual cues as a means of entertainment and verification. Our evaluations suggest that this game is more enjoyable than standard methods of collecting sentences. SR can also simultaneously perform the collection and verification of sentences. Finally, we show that the sentences collected by SR are of higher quality than those collected by AMT.

## REFERENCES

1. Johnson, J., Krishna, R., Stark, M., Li, L.-J., Shamma, D. A., Bernstein, M., and Fei-Fei, L. Image retrieval using scene graphs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015).
2. Lin, T.-Y., and et al. Microsoft COCO: Common objects in context. *Computer VisionECCV* (2014), 740–755.
3. Nickerson, R. S. Five down, absquatulated : Crossword puzzle clues to how the mind works. *Psychonomic Bulletin & Review* (2011), 217–241.
4. Vinyals, O., Toshev, A., Bengio, S., Erhan, D., and et al. Show and Tell: A Neural Image Caption Generator. *CoRR abs/1411.4555* (2014).
5. von Ahn, L., and Dabbish, L. Labeling images with a computer game. In *Proceedings of SIGCHI*, ACM Press (2001), 9–18.
6. Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*.