Intuitive Image Descriptions are Context-Sensitive

Shayan Hooshmand
Department of Computer Science
Columbia University
shayan.hooshmand@columbia.edu

Elisa Kreiss
Department of Linguistics
Stanford University
ekreiss@stanford.edu

Christopher Potts
Department of Linguistics
Stanford University
cgpotts@stanford.edu

Introduction Images are pervasive on the Web (Bigham et al., 2006; Guinness et al., 2018) but their presence poses a serious accessibility challenge. When a user is not able to see an image, a *description* makes it accessible (Morris et al., 2016; Gurari et al., 2020). According to users, a useful description is context-sensitive, e.g., the same image appearing on a shopping website vs. a news website is expected to be described differently (Stangl et al., 2021). Previous crowdsourced efforts to construct datasets with text from images have only presented images in isolation (Lin et al., 2014; Dognin et al., 2020). However, to create NLP models that can generate useful descriptions, we need to efficiently create large description datasets that are context-sensitive. In this work, we investigate whether untrained crowdworkers intuitively generate descriptions that are context-sensitive when images are presented as embedded within an article.

Methods Participants were shown an image displayed within the first paragraph of a Wikipedia article and asked to write image descriptions that would make them non-visually accessible. To manipulate an image's context, every image was randomly chosen to appear with one of three articles that it could plausibly appear in; likewise, every article appeared with three images (see Appendix A). Overall, there were 54 unique image–article pairs. Each participant saw 6 of them, randomly sampled ensuring that image topics were unrelated to each other. We hypothesized that the descriptions provided by untrained crowdworkers would be sensitive to the article the images were presented in.

Results We recruited 74 participants on Amazon's Mechanical Turk and excluded 7 participants on two bases: 1) indicating confusion about the task and 2) providing nonsensical or unrelated descriptions. Overall, we collected 273 descriptions with on average 5 descriptions per image–article pair (see Appendix A). To investigate whether descriptions written for the same image and article are more similar to each other than descriptions written for the same image but different articles, we conducted similarity analyses based on vector representations and keyword matching. Across methods, we find that descriptions are significantly more likely to reflect contents of the article they were presented with than mismatched articles. Appendix B provides full details on the similarity-based methods and associated statistical tests.

Conclusion Our results provide initial evidence that untrained description writers intuitively take context into account when providing descriptions. Further work will determine whether the context-sensitivity that human description writers tend toward is the same that users find useful. These insights can help us understand the extent and usefulness of training crowdworkers when developing large context-sensitive description corpora. Moreover, our findings have implications for the development of deep learning models for automatic image description generation: in particular, how best can these models accept contextual input, and will the effect of context in neural models be comparable to that in humans? Some encouraging initial results suggest that context helps the model of Kreiss et al. (2021) produce relevant information and influences which image features are mentioned first. The way context informs the generated descriptions, however, isn't always intuitive, since we also find that it can constrain the syntactic structure instead of content (see Appendix C). These initial findings reinforce the conclusion that image description generation for accessibility is a deep and challenging problem requiring innovations in both modeling and dataset creation.

Meaning in Context: Pragmatic Communication in Humans and Machines Workshop at 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

References

Jeffrey P. Bigham, Ryan S. Kaminsky, Richard E. Ladner, Oscar M. Danielsson, and Gordon L. Hempton. 2006. WebInSight:: Making web images accessible. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility - Assets '06*, page 181, Portland, Oregon, USA. ACM Press.

Pierre Dognin, Igor Melnyk, Youssef Mroueh, Inkit Padhi, Mattia Rigotti, Jarret Ross, Yair Schiff, Richard A. Young, and Brian Belgodere. 2020. Image Captioning as an Assistive Technology: Lessons Learned from VizWiz 2020 Challenge. *arXiv:2012.11696 [cs].*

Darren Guinness, Edward Cutrell, and Meredith Ringel Morris. 2018. Caption Crawler: Enabling Reusable Alternative Text Descriptions using Reverse Image Search. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, pages 1–11, Montreal QC, Canada. ACM Press.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning images taken by people who are blind. In *European Conference on Computer Vision*, pages 417–434. Springer.

Elisa Kreiss, Noah D. Goodman, and Christopher Potts. 2021. Concadia: Tackling image accessibility with context. *arXiv:2104.08376 [cs].*

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*, volume 8693, pages 740–755, Cham. Springer International Publishing.

Meredith Ringel Morris, Annuska Zolyomi, Catherine Yao, Sina Bahram, Jeffrey P. Bigham, and Shaun K. Kane. 2016. " With most of it being pictures now, I rarely use it" Understanding Twitter's Evolving Accessibility to Blind Users. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5506–5516.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Abigale Stangl, Nitin Verma, Kenneth R Fleischmann, Meredith Ringel Morris, and Danna Gurari. 2021. Going Beyond One-Size-Fits-All Image Descriptions to Satisfy the Information Wants of People Who are Blind or Have Low Vision. page 23.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2016. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *arXiv:1502.03044 [Cs].*

A Additional Dataset Information

A.1 Trial Data

Images were organized under one of six topics: Mountains, Soccer, Church, Laid Table, Living Room, Meeting. Each topic had three associated articles that the images could be paired with. The articles associated with each topic are found in Table 1, and an example article is provided in Table 2.

A.2 Collected Descriptions

Figure 1 presents the distribution of the number of descriptions per image-article pair (range: [2, 12], mean: 5.34, median: 5, mode: 4).

Topic	Articles
Mountains	Body of Water, Orogeny, Montane Ecosystems
Soccer	Competition, Hairstyle, Advertising
Church	Roof, Christian Cross, Building Material
Laid Table	Tableware, Dinner, American Cuisine
Living Room	Interior Design, Window, Furniture
Meeting	Consumer Electronics, Furniture, Cooperation

Table 1: The articles associated with each topic.

Hairstyle
A hairstyle, hairdo, or haircut refers to the styling of hair, usually on the human scalp. Sometimes, this could also mean an editing of facial or body hair. The fashioning of hair can be considered an aspect of personal grooming, fashion, and cosmetics, although practical, cultural, and popular considerations also influence some hairstyles.

Table 2: An example article from Wikipedia that occurred along the images. Participants saw a similar-length excerpt for all articles.

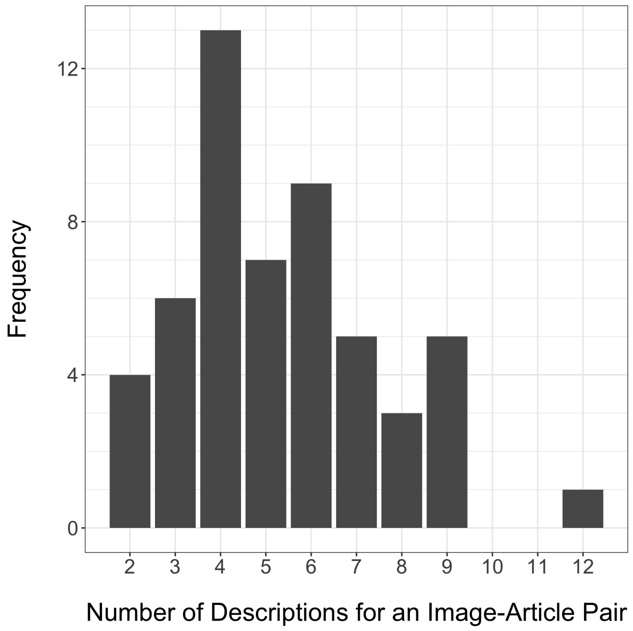

Figure 1: The distribution of the number of descriptions per image-article pair.

B Similarity-based Analyses

B.1 Vector Similarity

We hypothesized that descriptions written for the same image and article should be more similar to one another than those written for the same image presented in different articles. To test this hypothesis, we created TF-IDF vector representations for each description over three documents; each document was the concatenation of one of the articles the image could have appeared with and all the descriptions written for that image when paired with that article.

On average, the cosine similarities of descriptions written for the same image and article were higher, affirming our hypothesis (same article: 0.179, 95% CI: [0.171, 0.187]; different article: 0.165, 95% CI: [0.160, 0.170]; t-test: $p < 0.01$).

The magnitude of the similarity, however, is still quite low, which is likely because TF-IDF measures similarity based on exact token matches. For example, TF-IDF assigns the words “advertisements” and “ads” a similarity of 0. In order to capture graded similarities between words, we performed the same analysis using Sentence-BERT embeddings (Reimers and Gurevych, 2019), and saw an increase in overall similarities (same article: 0.674, 95% CI: [0.666, 0.682]; different article: 0.663, 95% CI: [0.657, 0.669]; t-test: $p < 0.02$).

While again statistically significant, the numeric difference in similarities for different- and same-article descriptions remains small. A better metric would merge the word-level relevance of TF-IDF with the gradable token similarities of word embeddings. We introduce a new metric, then, that makes use of both GloVe word embeddings and IDF scores (Pennington et al., 2014). As seen in equation 1, the vector representation **v** for a description **d** becomes an IDF-weighted average of GloVe word embeddings. This novel metric captures both graded similarity between words and the relative importance of contextually-sensitive words like “advertisement” over context-independent words like “soccer.”

Because of the small number of documents, however, the range in possible IDF values is limited, thereby overestimating the importance of frequently occurring words. To address this issue, IDF values are transformed to increase the relevance of more unique words. As seen in equation 3, words that appear in all documents are given a weight of 0, while the difference in weights for words that appear in a subset of documents is increased. Overall, this metric ensures that frequent and likely context-insensitive words play a lesser role in determining a description’s context sensitivity.

$$\mathbf{v}(\mathbf{d}) = n(\mathbf{d}) \cdot \sum\nolimits_{w \in tokens(\mathbf{d})} h(\text{IDF}(w)) \cdot \mathbf{GloVe}(w) \tag{1}$$

where

$$n(\mathbf{d}) = \frac{1}{\sum_{w \in tokens(\mathbf{d})} \text{IDF}(w)} \tag{2}$$

$$h(x) = \exp(x) - \exp(1) \tag{3}$$

With this updated vector representation, cosine similarities of descriptions written for the same image and article are far higher than those written for the same image and different articles (same article: 0.847, 95% CI: [0.841, 0.853]; different article: 0.545, 95% CI: [0.536, 0.555]; t-test: $p < 0.01$). The results of all of these methods taken together provide converging evidence in favor of the hypothesis that description writers intuitively integrate context in their descriptions. In future work, we plan to further validate our custom vector representation analysis with new descriptions.

B.2 Keyword Analysis

To supplement our vector representation-based analyses, we conducted an analysis on the collected descriptions using keywords related to each article that we defined after inspecting the data. The full set of keywords is found in Table 3.

We find that, while only 7.3% contain keywords from unrelated articles, 55% of descriptions contain their associated article’s keywords. This difference reaffirms our hypothesis that context affects the descriptions crowdworkers provide.

Future work could define these keywords in a more systematic way, through lexical databases like WordNet, and count occurrences on a larger set of descriptions. It's also worth noting that 55% is a lower-bound estimate for context-sensitive descriptions, since the keyword list is not exhaustive. Furthermore, these metrics don't take other context-sensitive strategies such as the order in which features are presented into account, which suggests that the rate of context-sensitivity is likely higher.

Article	Keywords
Advertising	advert-, logo, banner
American Cuisine	american
Body of Water	water, lake, river
Building Material	stone, brick, wood
Christian Cross	cross
Consumer Electronics	device, laptop, phone
Cooperation	cooperation, team, together
Competition	competition, team
Dinner	dinner
Furniture	furniture, couch, table, chair
Hairstyle	hair
Interior Design	layout, design, surround, center
Montane Ecosystem	climate, ecosystem, vegetation, tree
Orogeny	rock, form
Roof	roof
Tableware	tableware, plate, utensil, cup
Window	window

Table 3: The keywords associated with each article used in section B.2.

C Model Results

As a preliminary assessment of how neural models perform with contextual input, we used an LSTM-based description generation model which takes an image in concatenation with the context as input and uses separate soft attention mechanisms over both inputs during decoding (Kreiss et al., 2021; Xu et al., 2016).

We trained our model on the Concadia dataset, a corpus of images from Wikipedia with associated descriptions and the Wikipedia paragraph closest to the image (Kreiss et al., 2021). For our qualitative evaluation, we presented the model with the stimuli from the human subject experiment (see Methods paragraph). The generated descriptions indicate contextual sensitivity, both in intuitive and non-intuitive ways.

We present the case of intuitive effects of context here with landscape-related images and contexts. Notably, all images contained mountains, rivers/lakes and forest, and they were paired with the first paragraph from the Wikipedia articles *Montane Ecosystems*, *Body of Water* and *Orogeny* (see Table 4). Firstly, image features that were contextually relevant appeared earlier in the description. For instance, when the images were paired with the article *Body of Water*, references to lakes and rivers were the first mentioned features in the descriptions, and never when the same images were paired with other articles. Moreover, the forests in the images were described first when the associated article was *Montane Ecosystems*, and were otherwise only used as modifiers, as in *forested area*.

In most cases, however, the model was context sensitive in non-intuitive ways. For example, in our soccer-related images, the article *Hairstyle* induced the following specific syntax, no matter the image it was paired with: "A photograph of a man wearing a **ADJ NN** and **ADJ NNS**," where **ADJ** is an adjective, **NN** a singular noun, and **NNS** a plural noun. Descriptions for the same soccer images but paired with other articles did not follow this syntax. While this syntactic effect again reflects context-sensitivity, it is not intuitive or useful.

One hypothesis for why the model was most successful with the landscape images is that the elements in those images—mountains, trees, bodies of water, etc.—were likely mentioned in descriptions for similar images from the training data. Our other images, though, contain elements that are much less

	Image 1	**Image 2**	**Image 3**
Montane Ecosystems	The mountain range	A forest in a montainous landscape	A forest with a lake in the foreground
Body of Water	A view of a lake on a lake	An aerial view of a lake surrounded by mountains	A river flowing through a forested valley
Orogeny	A mountain rising above a lake	An aerial view of a large tropical cyclone	A broad mountain rising above a forested area

Table 4: Model-generated description on landscape images.

likely to be described unless a specific context induces it, such as the advertising in the background of our soccer images. This might lead to the model acquiring undesired artifacts.

While these first results are promising in suggesting that models can in principle incorporate context for description generation, future work has to identify methods that ensure useful effects of context on automatic description generation.