Toward Scalable Social Alt Text: Conversational Crowdsourcing as a Tool for Refining Vision-to-Language Technology for the Blind

Elliot Salisbury[*], **Ece Kamar**[+], **Meredith Ringel Morris**[+]
[*]University of Southampton, [+]Microsoft Research
[*]e.salisbury@ecs.soton.ac.uk, [+] {eckamar, merrie}@microsoft.com

Abstract

The access of visually impaired users to imagery in social media is constrained by the availability of suitable alt text. It is unknown how imperfections in emerging tools for automatic caption generation may help or hinder blind users' understanding of social media posts with embedded imagery. In this paper, we study how crowdsourcing can be used both for evaluating the value provided by existing automated approaches and for enabling workflows that provide scalable and useful alt text to blind users. Using real-time crowdsourcing, we designed experiences that varied the depth of interaction of the crowd in assisting visually impaired users at caption interpretation, and measured trade-offs in effectiveness, scalability, and reusability. We show that the shortcomings of existing AI image captioning systems frequently hinder a user's understanding of an image they cannot see to a degree that even clarifying conversations with sighted assistants cannot correct. Our detailed analysis of the set of clarifying conversations collected from our studies led to the design of experiences that can effectively assist users in a scalable way without the need for real-time interaction. They also provide lessons and guidelines that human captioners and the designers of future iterations of AI captioning systems can use to improve labeling of social media imagery for blind users.

Introduction

Social media is becoming pervasive in American culture; as of 2014, 74% of online adults in the U.S. use social networking sites (Duggan et al. 2015). The opportunity to engage with social media is an important part of social, professional, and political life, making it important that people who are blind or visually impaired (BVI) can access the entirety of content shared in social media. For example, Twitter has more than 313 million active users per month (Twitter 2016); Twitter is particularly popular among blind users, in part because it evolved from a very simple, text-based interface (Morris et al. 2016; Brady et al. 2013). However, embedded imagery is becoming more prevalent in social media; a study of Twitter found that more than 40% of popular (retweeted) posts contained embedded multimedia as of June 2015 (Morris et al. 2016), which constrains the accessibility of the content in Twitter by BVI users. As a response, Twitter recently began to offer limited capabilities to augment images with alternative text (a.k.a. alt text or captions) that can be read aloud by the screen reader technology (e.g., JAWS, VoiceOver, Narrator, etc.) that provides computer access to people who are BVI (Kloots 2016); however, while no official numbers on alt text compliance for Twitter are yet available, alt text compliance and quality on the web in general is low (Bigham et al. 2006; Goodwin et al. 2011; Shi 2006), and this trend is likely to be exacerbated by quickly-created, user-generated content such as tweets.

Recently, automated approaches that combine computer vision and natural language processing to describe image content have emerged as a potential solution for improving the accessibility of social media imagery for BVI users. Examples include the automatic alt text system deployed by Facebook (Wu, Pique, and Wieland 2016) and automated image captioning systems (Fang et al. 2015; Karpathy and Fei-Fei 2015). Although assisting blind users is a motivating application domain for these systems, the value these imperfect systems provide to BVI users is unclear. While existing systems are tested in the lab within constrained data sets, the performance of these systems in the context of social media (which incorporates a wide variety of professional and casual quality imagery and covers a range of subjects and styles) is not yet studied. The levels of detail, accuracy, or confidence expected from BVI users may not be attainable with current vision-to-language technologies. Unexpected imperfections in automated system output may degrade user trust, or may hurt users instead of helping them.

In this work, we explore ways for combining crowd input and existing automated approaches to assist BVI users in accessing social media with visual content. Our studies focus on the following research questions: (1) What value is provided by a state-of-the-art vision-to-language API in assisting BVI users, and what are the areas for improvement? (2) What are the trade-offs between alternative workflows for the crowd assisting BVI users? (3) Can human-in-the-loop workflows result in reusable content that can be shared with other BVI users?

To study these research questions, we designed and experimented with workflows that varied the level of human engagement and the involvement of an automated system to better understand the requirements for creating good-quality, scalable, automated or semi-automated alt text for BVI con-

sumers of social media. The results show that the negative impact of erroneous system output on user understanding is so significant that it cannot be completely erased even through free-form conversation with a sighted assistant. On the positive side, human input, either assisting users alone or correcting/complementing the automated system, is effective in increasing user satisfaction. Our structured Q&A workflow emerged as an effective workflow for enabling scalable, lower-cost assistance to BVI users. We complement the large-scale crowdsourcing study with a small-scale evaluation of TweetTalk with real BVI users. We conclude with a set of guidelines that human captioners or future AI captioning systems can use to improve labeling of social media imagery for blind users.

Related Work

Image understanding and automated image captioning have emerged as challenging problems for Artificial Intelligence researchers in recent years (Lin et al. 2014). Example systems include Microsoft's CaptionBot (Tran et al. 2016), Google's Show and Tell (Vinyals et al. 2015), and many more (Wu, Pique, and Wieland 2016; Fang et al. 2015; Karpathy and Fei-Fei 2015). These AI image captioning systems attempt to describe the contents of an image in natural language and have been motivated by various applications, such as semantic image search, bringing visual intelligence to chatbots, and assisting BVI users in understanding the visual world around them. However, the adequacy of these systems for these tasks are not yet understood. Given limitations of existing AI approaches and the limited data sets they are trained with, these systems are prone to errors when they are used in the open world (e.g., captioning the wide variety of images posted to social media). Recent experiments have shown that the error rate of existing systems doubles when the systems are evaluated over images sampled from Instagram rather than being evaluated on existing constrained data sets (Tran et al. 2016). For semantic image search or chatbots, inaccurate or wrong results are easily detectable by sighted users and the consequences may not be significant. Whereas, in the case of helping the visually impaired, there is little understanding of how imperfections in the AI's description would affect BVI users' understanding of an image and their consequent actions. Moreover, the level of detail provided by the caption may not be appropriate for understanding images posted on social media, a context that may require nuanced explanations of abstract concepts (e.g., aesthetics, humor) that are an important currency within social platforms (Morris et al. 2016). In this work, we combine human intervention (via workers on Amazon's Mechanical Turk) with automated captions (via Microsoft's Cognitive Services API (Microsoft 2016)), to better understand the requirements for creating good-quality, scalable, automated or semi-automated alt text for BVI consumers of social media.

Other approaches for enhancing the accessibility of social media imagery for BVI users are powered by human computation. Crowdsourced conversational interfaces have been developed in previous work for assisting BVI users with their daily tasks. VizWiz allows blind users to take pictures with their phone and ask a question about the image, crowdsourcing the answer (Bigham et al. 2010). Chorus expands on this idea, allowing for longer conversations, where multiple crowd members are present, and can interact with the user for more reliable interactions (Lasecki et al. 2013). Social Microvolunteering (Brady, Morris, and Bigham 2015) uses third-party friendsourcing to achieve low-cost, high-quality answers to visual questions from people who are BVI, but it is unclear that the technique is scalable to the level needed to provide alt text to large sets of online images. In this work, we apply the ideas from crowdsourced conversational interfaces to the generation of scalable alt-text for social media. Different from previous work, we investigate trade-offs between various workflows that utilize varying combinations of automation and human input, and evaluate reusability as well as value to the immediate user.

Large-scale investigation of conversational interfaces for alt-text generation through the approach of VizWiz and Chorus would require many blind users to interact with our apps daily, and deploying large-scale systems such as this comes with its own challenges (Huang et al. 2016). Instead, we follow an approach that simulates all parties of the interaction with crowd workers using a real-time interactive crowdsourcing toolkit (Mao et al. 2012). With this approach, we study different workflows that enable conversations between two crowd workers with different interfaces, allowing for more data to be collected, quicker than previously possible.

Researchers have investigated workflows that combine crowd input with automated systems to overcome the shortcomings of automation (Yan, Kumar, and Ganesan 2010; Kamar, Hacker, and Horvitz 2012). However, existing work is suitable for well-defined tasks such as image labeling with objective ground truth answers, and it is unclear how human input can be combined with automated image captioning for alt-text generation for social media posts. We investigate this question through a conversational interface between BVI (or simulated-BVI) users and sighted assistants to discover what additional information BVI users would wish to know about an image. Researchers have investigated how sighted users ask questions when attempting to search for an image (Collins 1998); instead we are investigating the questions users ask when they are presented with visual content they cannot observe. Recent work has focused on developing AI chatbots, which can carry on free-from conversations within the shared context of images (Das et al. 2017), whereas our work focuses specifically on the questions towards assisting BVI users in the context of social media posts.

A recent study by (MacLeod et al. 2017) investigated how BVI users perceive captions generated by automated approaches for a curated set of image tweets. MacLeod, et al. showed that BVI users trust auto-generated captions even when they are inaccurate, and studied how to convey skepticism to prevent over-trusting. In this paper, we focus on human-in-the-loop workflows to improve the value BVI users get from alt text, focusing on what types of detail this audience values in captions.

Workflows for Alt Text Generation

We designed and studied four workflows for providing an understanding of images accompanying tweets, with BVI

users as the target audience. The inputs to each workflow are a single tweet, containing the tweet's text and the accompanying image. Then each workflow attempts to explain the tweet's image to BVI users within the context of the tweet.

The first two workflows provide a baseline state of the art approach to captioning images. The first workflow, Vision-to-Language, uses captions generated by the CaptionBot system for the tweet's image (Fang et al. 2015). CaptionBot is based off the technology that won the 2015 CVPR captioning challenge, and uses Microsoft Cognitive Services (Microsoft 2016), a set of APIs used for understanding imagery and text. Current Vision-to-Language systems cannot yet use the additional context of the tweet text and purely caption the image instead. The second workflow, Human-Corrected Captions, provides crowd workers with the original tweet text, accompanying image, and the Vision-to-Language-generated alt text. Workers were paid $0.05 to improve the automated caption to explain the image to a blind user, given the context of the tweet, ensuring the tweeter's intention was clear from the caption. While human corrections may fix factual errors in the automatically generated captions, the value of human-corrected captions to BVI users may be limited since workers may not foresee the type of information or the level of detail required for high-quality alt text desired by BVI users.

We developed two subsequent experiences, the TweetTalk conversational assistant workflow and the Structured Q&A workflow, that build upon and enhance the baseline captions. These four workflows allow us to investigate what key information end users desire in a caption for a social media image, how effective deeper human assistance is, and whether the information desired by a single consumer of the alt text will satisfy a larger audience of end users.

The workflows were tested on Amazon's Mechanical Turk (AMT), with recruitment restricted to U.S. workers only, due to the collection of tweets being mainly U.S.-centric and the description and conversations these workers take part in requiring sufficient understanding of the English language. Because current crowdsourcing platforms are largely inaccessible to users with disabilities and therefore lack a sufficiently large pool of BVI workers (Zyskowski et al. 2015), when testing our workflows, we simulated the experience of being BVI by employing (presumably) sighted turkers (whom we will refer to as the simulated-BVI workers) and making the images unavailable to them. While necessitated by practical constraints of testing these workflows at scale, we recognize that simulated-BVI workers may have different captioning preferences than people who are BVI; hence, we conducted additional testing with seven people who are blind or visually impaired to validate the generalizability of our findings.

We experiment with a data set of 85 tweets that was curated by previous work (MacLeod et al. 2017). Each tweet contains embedded image attachments from a set of popular accounts (e.g., @HillaryClinton, @nytimes, @TaylorSwift) and/or trending hashtags (e.g., #tbt [throw-back Thursday]). Tweets were selected to cover a broad range of topics (e.g., humor, news, celebrities, memes, etc.), representing the varied interests reported by blind users of Twitter (Morris et al. 2016). As described in (MacLeod et al. 2017), the tweets vary in terms of confidence of the automated system in generating an auto-caption.

Conversational Assistant Workflow

Evaluating the satisfaction of BVI users with assistive technologies for visual content is challenging; what makes a caption valuable for a BVI user is unknown. To understand these concepts and assess the value that crowd workers can generate for BVI users, we developed a Conversational Assistant workflow. This workflow uses TweetTalk, a scalable conversational platform between BVI (or simulated-BVI) users and human assistants. TweetTalk allows BVI users to have free-form conversations with sighted workers to find out about visual content. Analyses of conversations collected from TweetTalk show us what kind of information BVI users are interested in, extract key classes of information that can help enhance captions, and measure the value gained from unconstrained human assistance.

Our conversational assistant platform, TweetTalk, was built on top of the architecture described in (Mao et al. 2012), and enables us to investigate conversations between sighted and simulated-BVI crowd workers about a given tweet. This workflow connects two workers, but provides each worker with a different interface. One worker, whom we will refer to as the sighted assistant, can see the imagery associated with the tweet, while the other (the simulated-BVI worker) cannot. The simulated-BVI worker must then have a conversation with the sighted assistant in order to understand the image accompanying the tweet and write a description of it.

After accepting a TweetTalk HIT, the crowd workers are asked to read the instructions, and click the ready button; once ready they are put in a waiting room where they remain until another worker connects. Two workers are then paired up, randomly assigned a tweet from our data set and a role (sighted or simulated-BVI), notified and forwarded to the conversational interface (Figure 1).

The Conversational Assistant workflow follows the following steps, each designed to gain insights about the value automated systems and human assistance can provide:

1. **Read the tweet:** Both workers are shown the tweet's text and a Baseline Image Caption, that could either be empty, generated from Vision-to-Language, or a Human-Corrected caption. This baseline caption seeds the simulated-BVI worker's understanding of the image. Only the sighted assistant is shown the image associated with the tweet.
2. **Rate the caption:** We ask only the simulated-BVI worker to rate the utility of the baseline caption (if there is one), as they have not yet seen the image, so we can assess the initial trust the BVI user has for the baseline caption, and how this assessment later changes as a result of gaining more information about the image through the following conversation.
3. **Ask/Answer questions:** Both workers have access to a chat box; the simulated-BVI worker is asked to initiate the conversation by asking one or more questions about

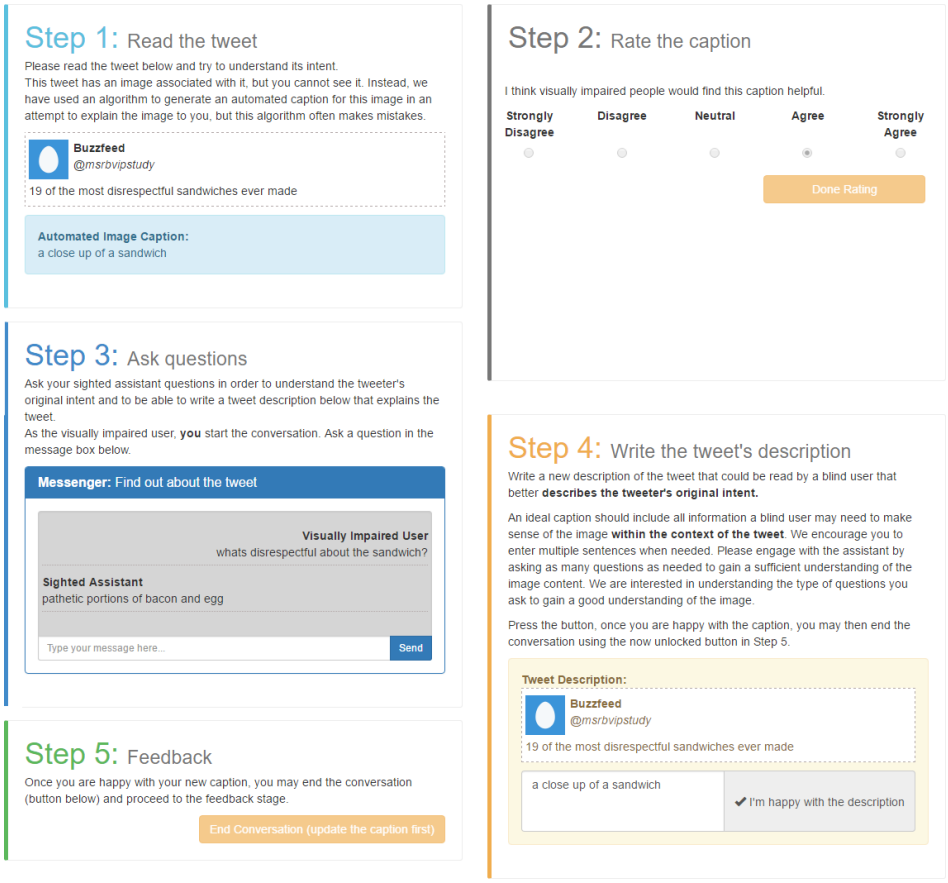

Figure 1: An example of the TweetTalk interface shown to workers in the simulated-BVI role.

the image. The sighted assistant is asked to reply sensibly to these questions, but without writing their own complete description of the image, because we are interested in capturing the simulated-BVI worker's questions. This step has two purposes; it informs us about the information users would like, and allows us to quantify the effectiveness of free-form human assistance.

4. **Write a Description:** After they feel they have had sufficient opportunity to ask questions of the sighted assistant, the simulated-BVI worker is then asked to write a new description of the tweet's image, so that we can evaluate their understanding gained through conversation.
5. **Feedback:** In the final step (Not shown in Figure 1), we show the simulated-BVI worker the image for the first time, and ask them to rerate the baseline caption and the new description generated in step 4. As such, we can gain insight into their assessment of the effectiveness of the conversation and how well they believe the new description describes the image.

Execution Details We tested the system on the 85 tweets from (MacLeod et al. 2017), recruiting 235 unique workers for the conversational assistant. Due to the time it takes for two workers to complete this task (ranging between 2 to 20 minutes, and on average taking 8 minutes), we paid a base rate of $0.75, with the option of earning a bonus payment up to $0.20, depending on the quality of the conversation and resulting description, and an additional $0.10 bonus was paid unconditionally to the simulated-BVI worker, as they had more workload and typically took longer on the task.

To rate the baseline image captions and the simulated-BVI worker's generated descriptions, workers are asked the same Likert-type question (i.e., "I think visually impaired people would find this caption helpful.") used in (MacLeod et al. 2017), using a five-point scale (Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree). During the conversational workflow, the simulated-BVI workers taking part in the conversations are asked to rate the baseline caption twice, once before the conversation having not seen the image (First-Party Before for Baseline Captions), and again after the conversation once the image is revealed to them (First-Party After for Baseline Captions). By asking the same question before and after seeing the image, we can then measure any change in ratings after seeing the image, which we refer to as the satisfaction factor. This satisfaction factor captures how misleading a caption may be to a BVI user. For example, a given caption may be wrong, but without seeing the image, may make sense to the reader and give them a false impression about the tweet. By then showing the image to the rater, we can capture not only the accuracy of the caption, but how

Code	Description	%
Describe	Give me more detail about this.	16%
Location	Where is this picture taken? (websites can be locations)	15%
Specific	A question very specific to the tweet, hard to generalize.	9%
People	Who is present in the image? (including animals and characters)	9%
Action	What action is happening in the image.	8%
Confirming	Checking the original caption is correct.	7%
Pose	What pose are the people/things in?	7%
Wearing	What are the people in the image wearing or holding?	7%
Emotion	What emotion is evoked by the scene, or by the people in it?	5%
Color	Asking about the color of something.	4%
Text	Read the text in the image.	3%
Background	What is present in the image that isn't part of the notable subject?	3%
Famous	Is this a famous photo or person?	2%
Count	How many things are there?	1%
Context	Asking for information about the context of the tweet.	1%
Notable	Asking for anything else they should be aware of in the image.	1%

Table 1: Coded conversations

their understanding of the tweet changed.

Once the simulated-BVI worker finishes conversing with the sighted assistant, the simulated-BVI worker is asked to write a description of the image to the best of their understanding, gained through conversing with the sighted assistant without seeing the image. This allows us to extract how well they have understood the imagery. Once the image is revealed to the simulated-BVI user, they then must rate their description (First-Party After for Descriptions Generated Through TweetTalk Conversations).

Structured Questions Workflow

After analyzing the conversations between workers using the Conversational Assistant workflow, we were able to develop a streamlined workflow using the most common question types present in TweetTalk interactions. This eliminates the need for a conversational pairing, thus reducing the temporal and monetary costs, and removing the common conversational problem of user dropouts.

Conversational Analysis Over 2700 TweetTalk messages were sent by our users; we filter these messages to just those sent by simulated-BVI workers (1359 messages), and for those conversations that were seeded by the Human Corrected Captions (429), as these questions were most likely to be informative given the simulated-BVI workers' greater initial understanding. We analysed these 429 questions using an iterative, open coding approach in which we identified and refined thematic categories in the questions. The categories were not exhaustive (we allowed a "none of the above" category for long-tail or difficult-to-classify questions), but were mutually exclusive. Two of the authors then coded the questions using this coding scheme, achieving an IRR of .67 (Cohen's Kappa), a reasonable level of inter-rater reliability given the high number of categories (17). Table 1 describes these question categories, and the percentage of TweetTalk conversations in which they appeared.

Who are the main subjects of the image (people, animals, notable objects, etc.)? Describe their physical characteristics (notable features, clothes, poses, relative positions, etc.)
Where is this set? Describe the location and the prominent features of the background.
What are the subjects of the image doing? Describe their actions, and their intent.
What emotion does this image evoke? Or what are the emotions of those present in the image?
Describe any noteworthy aspects of the image's visual style.
Is this tweet intended to be humorous? Explain how.
Is this a famous or well-known image?
Does this tweet contain a meme (meme images, #hashtags, etc.)? If so, describe what the meme is about.

Table 2: Structured Questions

From these coded questions, we identified the core concepts that our users were interested in. We used these concepts to create a set of questions to extract desired details about social media images. We expressed common questions that had yes/no answers (e.g., Is she wearing a dress?) more generally to improve the value from an answer (e.g., Describe what they are wearing). The resulting question list for the Structured Questions Workflow is given in Table 2.

Execution Details We created a HIT for collecting answers for each question and for each tweet in our collection. We repeated the answer collection three times with unique workers (2040 assignments at $0.05 each). The time taken to answer the questions ranged between 3 seconds to 14 minutes, and on average took 1 minute. We ask workers to decide if this question was useful to ask given the tweet, and to write an answer for the question. For each tweet, if most workers identify a particular question as useful, we take the longest answer (other mechanisms for choosing the highest quality answer could be substituted at this stage if desired).

To evaluate the effectiveness of these answers to understanding the image, we adopt a similar interface as that shown to the simulated-BVI worker in TweetTalk, replacing the chat box with the list of answered questions and, as before, the worker is asked to write a description of the tweet. Furthermore, without the need for pairing with a conversational partner and waiting for their replies; the time taken on

the task was far less, so each HIT cost $0.15.

Experiments

We evaluated the Conversational Assistant workflow with 235 unique crowd workers. We held one conversation per tweet per initial seeding Baseline Image Caption, and with no seeding captions at all (i.e., 3 treatments), leading to a total of 255 conversations. Evaluating the Structured Questions workflow does not require worker pairing and thus was faster and easier to recruit workers; we performed three repeats per tweet, for a total of 255 runs.

The baseline image captions and the descriptions generated through our workflows were evaluated by the first-party workers taking part in the conversational workflows, as previously explained. However, as these ratings may be subjective, we also rate all descriptions using crowd workers who did not participate in the conversation (Third-party Evaluation). Third-party evaluations were elicited in a HIT, paying $0.03, in which we first showed only the tweet and the caption or description to be rated, without revealing the source image. Once the Likert question is answered, we reveal the image to the worker and ask them to rate it again. The Third-party Evaluation is completed for all baseline captions as well as the descriptions created by the simulated BVI users participating in the Conversational Assistance and Structured Q&A workflows.

Results

Table 3 presents the first-party and third-party ratings of both the simple baseline image captions, the descriptions generated through the Conversational Assistant workflow (for each treatment seeding the conversation with a different caption), and the Structured Questions workflow. All results stated as significant have been found as such using Friedman's test with a follow-up pairwise comparison using Wilcoxon's test with Bonferroni correction.

We found no significant difference across conditions for the first-party ratings (i.e., the rating they give their own description after the task). We also observed that the first-party ratings were higher than those given by third-parties uninvolved in their creation. To further investigate this disparity, we designed a quick follow up study in which we showed both the description and the conversation to third-party raters and checked if this disparity was due to intrinsic valued gained through conversation that wasn't relayed in their description. The results suggested that showing the conversation did not provide additional value and the disparity results from workers rating their own work higher.

Next, we evaluate the captions and descriptions generated by various workflows based on their third-party evaluations (i.e., collected after seeing the imagery). The results show that current Vision-to-Language systems have significantly worse accuracy when compared to even a simple human-in-the-loop approach (Human-Corrected Captions, z=7.45, $p < .001$) and to our caption improvement workflows (Conversational Assistant and Structured Questions, z=2.20, $p < .001$ and z=3.19, $p < .001$). This suggests that automatic image captioning systems require more work before they are ready for use by social networking platforms.

We observe no significant difference in the accuracy between the Human-Corrected Captions and the description generated after using TweetTalk, on the treatments seeded with either no caption or the Human-Corrected captions (those seeded with Vision-to-Language captions are discussed below). However, the Structured Questions approach significantly ($0.52 <= z <= 0.99$, $p <= .03$) improves understanding against all approaches.

Additionally, we observed that seeding the conversation with Vision-to-Language creates significantly less satisfaction, (i.e., the captions are believable, but turn out to be inaccurate) than simply providing a Human-Corrected Caption ($z = -0.78$, $p < .001$), or conversations seeded with Human-Corrected Captions ($z = -0.63$, $p = .003$).

Another consideration for the comparison of workflows is the time and monetary costs of assisting BVI users. Real-time crowdsourcing for free-form conversation is time consuming and expensive, on average taking 8 minutes per tweet, with no significant difference in duration between the seeding captions, and costing up to $0.95 for compensating the sighted assistant. Whereas, the structured questions do not suffer from the challenges of real-time crowdsourced conversation; the time taken to answer a question on average takes 1 minute, and although we need multiple workers to answer the same question, these can be performed simultaneously. The total cost of these HITs, to get 3 answers to the 8 questions, was $1.20. Although more expensive than the human-corrected captions and the conversational assistant, the structured Q&A workflow is more general purpose, results in a much greater satisfaction, and the cost can be amortized across multiple BVI users, while the conversation is an individual experience. In future versions of the of the structured Q&A workflow, different strategies such as answering multiple questions per HIT, or predicting relevant questions to ask per tweet, can be taken to reduce its cost.

Validation with BVI Users

To validate our approach, and that our observations from simulated-BVI users generalize to real BVI users, we ran a follow-up study with seven blind and visually impaired adults. Participants were recruited from an email list of BVI members in our organization and on Twitter. Participants ranged in age from 21-33 years old, six were male, one female. All had significant visual impairments (requiring the use of screen readers or screen magnifiers), and most used social media multiple times a day.

Given the limited size of our subject pool, we preferred to use TweetTalk over Structured Q&A in experimenting with real BVI users so that we could collect more detailed information than just assessments, including what BVI users ask about and their preferences about interactive crowd experiences. Since the common questions identified through running TweetTalk form the basis of the Structured Q&A workflow, observing what BVI users ask about and and the overlap with the set of questions in Table 2 inform us the potential effectiveness of helping BVI users with the Structured Q&A workflow.

Each BVI participant was presented with a random tweet from our collection, and the uncorrected Vision-to-

	Baseline Image Captions		Descriptions Generated Through TweetTalk Conversations			
	Vision-to-Language	Human-Corrected Captions	No Caption	Vision-to-Language	Human-Corrected Captions	Structured Questions
First-Party Before	2.56	3.36				
First-Party After	1.92	3.48	4.11	3.97	4.22	4.11
First-Party Satisfaction	-0.64	0.12				
Third-Party Before	2.91	3.63	3.70	3.92	3.81	3.42
Third-Party After	1.85	3.74	3.65	3.64	3.83	4.10
Third-Party Satisfaction	-1.06	0.11	-0.05	-0.28	0.02	0.68

Table 3: Average Likert Ratings: Before/After ratings are on a 1-5 scale; Satisfaction ratings are on a -4 to 4 scale (i.e., how much an individual changes their rating after seeing the image); higher ratings are better.

Language caption, and paired with a sighted assistant in order to answer questions users may have about the imagery. This is similar to the typical TweetTalk workflow, except that the role of the sighted assistant was fulfilled by a member of our research team (the experiment could not be conducted through AMT due to screen reader accessibility issues). The researcher followed the same instructions given to the sighted assistant workers on Amazon Mechanical Turk, only answering the BVI users' questions directly and not providing any additional information other than that requested by the question. While using members of the research team in the sighted assistant role is not ideal, it enabled us to discover the types of questions asked by these BVI users. This process was repeated three times per participant on different tweets.

The description generated by the BVI users at the end of TweetTalk was evaluated by third-parties, and was not found to be significantly different in quality than those generated by our simulated-BVI workers for the same Vision-to-Language seed (before: 3.83, after: 3.80, satisfaction: -0.03). The questions the BVI users asked were coded using the same scheme as before; no new types of questions were asked, and there was no significant difference in the frequency of these question types, indicating that the Structured Q&A workflow would be informative for real BVI users.

All seven participants stated they enjoyed conversing with the sighted assistant; however, one BVI participant suggested that perhaps some imagery is harder to describe than others: *Some [sighted assistants] were better at giving information than others, or perhaps some images are easier to explain. I felt like it was hard to get the answers I needed on the first tweet.*

Another user was shown a tweet about a sunset, and suggested that there's nothing more about that image he needed to know, describing its beauty and colors was not needed, in his opinion: *I ask people these kind of questions all the time, but it's hard to think of questions about the sunset image.*

The level of description end users desire may vary based on factors such as their level of interest in a tweet, time pressure (are they browsing tweets quickly or in a leisurely fashion), and/or whether they have been blind since birth or lost their sight later in life (in which case they may have different knowledge of and interest in certain types of detail, such as color). We asked the BVI participants if they would use a conversational interface like TweetTalk to get more information about images they encounter in social media, such as Twitter or Facebook; all seven said yes.

Facebook's automatic descriptions are sometimes helpful, but it would be good to supplement them at times.

Yes, I usually ask original tweeters what they'd shown but would use [TweetTalk] if it were there.

While our BVI participants stated that they would indeed use conversational assistants to understand imagery, we believe these approaches to be too time consuming and expensive to be a practical and scalable solution. Our goal is to generate an alt text such that a conversational assistant need only be used in rare instances complementing more scalable workflows such as the Structured Q&A.

Discussion

We have shown that, often, providing state-of-the-art (as of late 2016) AI-generated image captions to simulated-BVI users hinders their ability to accurately understand imagery posted on social media, despite the extra context given in the text of the post. Additionally, given access to a conversational assistant with which to understand the imagery, it is often better to provide no caption at all than to seed the understanding with an incorrect AI caption, an observation also noted by (Jeong et al. 2013), that precision is very important for social systems. These findings are in line with recent findings that BVI users place too much trust in AI captions (MacLeod et al. 2017), and our results suggest that this trust in the initial caption may mislead people to the extent that they are unable to properly frame questions to improve their understanding. Our findings underscore the importance of future research by those working on AI-based captioning systems toward the topic of creating and conveying user-understandable quality and confidence metrics that can accompany automated captions, to potentially mitigate the confusion we saw arising from inaccurate caption seeds.

Observing a large set of conversations about unseen images via the TweetTalk tool allowed us to construct a set

of eight canonical questions (given in Table 2) relevant to understanding social media imagery that human captioners or future AI captioning systems can use to improve labeling of social media imagery for BVI users, such that they can describe the concepts of the image most relevant for understanding in this scenario. Many of the questions considered important by our audience cover subjective issues (e.g., context, and emotion), and understanding these concepts is currently an unsolved problem in AI. Human-in-the-loop solutions such as ours can serve an important role in both the gathering of training data, and in fulfilling this end-user need. Future AI captioning systems may need to use this training data and additional metadata associated with the social media post (e.g., the geolocation, the timing, the hashtags, and the tweet text), to more accurately caption the image.

In addition to the set of eight questions distilled from TweetTalk, we found that 9% of questions asked were highly specific to the tweet, and could not be generalized into a reusable question. For example, on a tweet revealing the tour dates for a musical performance, one worker asked "Can I only buy tickets with an American Express card?". This suggests that despite creating a set of generalized questions suitable for most situations, conversational assistants for answering specific questions about social media posts may still be useful to a BVI end user.

While we use CaptionBot in this work as an example of a system whose output can be improved via our human-in-the-loop method, this method would be generally applicable to any Vision-to-Language tool, all of which have room for improvement (Wu et al. 2017; Nieva 2015). Even as AI systems evolve, it is unlikely (or extremely distant) that they would have human-level accuracy, particularly on nuanced aspects of captions that may be subjective (e.g., emotional valence, aesthetics), so corrective techniques will remain important for improving and training future systems. This is particularly true for our scenario of image captioning for people who are blind, which requires a much higher threshold of accuracy than for sighted users (MacLeod et al. 2017; Wu et al. 2017).

Tradeoffs

Evaluations of workflows with varying complexity reveal trade-offs with respect to their implementation for assisting BVI users. Conversational assistants are able to provide a more specific and tailored answer to the end-user, but require managing real-time crowds, a challenging problem that LegionTools, (Gordon, Bigham, and Lasecki 2015) or TurkServer (Mao et al. 2012) attempt to address; however, these tools still require significant manual oversight, and are not as of yet suitable for large scale autonomous systems. Furthermore, conversational assistants must deal with workers dropping out mid-conversation, and ensuring a high-quality response, these are challenges addressed by Chorus (Lasecki et al. 2013), but increase the costs of running such a system. In contrast, adding a human corrective step to automated AI generated captions is cheaper and easier to scale. Furthermore, with a set of guiding questions, workers can produce detailed and meaningful captions in a fraction of the time of having a real-time conversation, and the resulting captions can be shared amongst many end users.

Limitations

These experiments were run with workers on the crowdsourcing website Amazon Mechanical Turk. The vast majority of these workers are unlikely to have any visual impairment (Zyskowski et al. 2015). The types of questions asked about an image may be different depending on whether a user is BVI or whether they're sighted. We did not observe a noticeable difference in the type of questions asked by crowd workers and the seven BVI users in our validation study, but it is possible that a larger sample of BVI users might reveal important differences that did not come to light in our study.

Furthermore, the tweets we used in this experiment were selected to be representative of a range of topics and image types typically seen in a standard Twitter feed (MacLeod et al. 2017; Morris et al. 2016). However, to an individual user, BVI or crowd worker, these tweets may have been uninteresting, and thus the worker may not be motivated to understand the tweet as they would be had the post been from a friend or someone they cared about.

Some social media platforms, such as Twitter, likely require less personalization than others (e.g., Facebook), as users tend to follow public figures whose content likely contains less personalized content than that of friends. Human-in-the-loop workflows powered by the crowd may raise privacy concerns for platforms in which content is privately shared to friend groups. However, our workflows for alt text generation would generalize to friendsourcing (Brady, Morris, and Bigham 2015) where the worker pool is sourced from the social network more broadly. Analyzing what percentage of images require personalized captions and developing tools or incentives to increase author alt-text generation are avenues for future work.

Conclusions and Future Work

We have shown how current AI captioning systems may hinder, rather than help, BVI users' understanding of social media posts. We developed workflows that incorporate different levels of automation and human involvement to improve this understanding, and to analyze the information that BVI users wish to know about. We identified guideline questions that answer the main interests BVI users had, and show that redistributing these answers can improve user understanding of social media posts containing imagery.

Given that alt-text in a social media context should be brief to facilitate fast browsing, we foresee there may be value in exploring alternative alt text formats, such as interactive formats in which users can query additional information about the image should they wish, perhaps using our set of structured questions. For popular imagery and posts, the guideline questions could be pre-asked, anticipating the details users would want. For those questions not yet asked before, real-time crowdsourcing could be used to respond quickly, and any future similar question can return the same answer, reducing the workload and distributing the cost.

References

Bigham, J. P.; Kaminsky, R. S.; Ladner, R. E.; Danielsson, O. M.; and Hempton, G. L. 2006. Webinsight:: making web images accessible. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, 181–188. ACM.

Bigham, J. P.; Jayant, C.; Ji, H.; Little, G.; Miller, A.; Miller, R. C.; Miller, R.; Tatarowicz, A.; White, B.; White, S.; et al. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, 333–342. ACM.

Brady, E. L.; Zhong, Y.; Morris, M. R.; and Bigham, J. P. 2013. Investigating the appropriateness of social network question asking as a resource for blind users. In *Proceedings of the 2013 conference on Computer supported cooperative work*, 1225–1236. ACM.

Brady, E.; Morris, M. R.; and Bigham, J. P. 2015. Gauging receptiveness to social microvolunteering. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1055–1064. ACM.

Collins, K. 1998. Providing subject access to images: a study of user queries. *The American Archivist* 61(1):36–55.

Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M.; Parikh, D.; and Batra, D. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Duggan, M.; Ellison, N. B.; Lampe, C.; Lenhart, A.; and Madden, M. 2015. Social media update 2014. pew research center.

Fang, H.; Gupta, S.; Iandola, F.; Srivastava, R. K.; Deng, L.; Dollár, P.; Gao, J.; He, X.; Mitchell, M.; Platt, J. C.; et al. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1473–1482.

Goodwin, M.; Susar, D.; Nietzio, A.; Snaprud, M.; and Jensen, C. S. 2011. Global web accessibility analysis of national government portals and ministry web sites. *Journal of Information Technology & Politics* 8(1):41–67.

Gordon, M.; Bigham, J. P.; and Lasecki, W. S. 2015. Legiontools: a toolkit+ ui for recruiting and routing crowds to synchronous real-time tasks. In *Adjunct Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, 81–82. ACM.

Huang, T.-H. K.; Lasecki, W. S.; Azaria, A.; and Bigham, J. P. 2016. is there anything else i can help you with?: Challenges in deploying an on-demand crowd-powered conversational agent.

Jeong, J.-W.; Morris, M. R.; Teevan, J.; and Liebling, D. J. 2013. A crowd-powered socially embedded search engine. In *ICWSM*.

Kamar, E.; Hacker, S.; and Horvitz, E. 2012. Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, 467–474. International Foundation for Autonomous Agents and Multiagent Systems.

Karpathy, A., and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3128–3137.

Kloots, T. 2016. Accessible images for everyone. `https://blog.twitter.com/2016/accessible-images-for-everyone`.

Lasecki, W. S.; Wesley, R.; Nichols, J.; Kulkarni, A.; Allen, J. F.; and Bigham, J. P. 2013. Chorus: a crowd-powered conversational assistant. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, 151–162. ACM.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 740–755. Springer.

MacLeod, H.; Bennett, C. L.; Ringel Morris, M.; and Cutrell, E. 2017. Understanding blind peoples experiences with computer-generated captions of social media images. In *Proceedings of the 2017 CHI Conference*.

Mao, A.; Chen, Y.; Gajos, K. Z.; Parkes, D.; Procaccia, A.; and Zhang, H. 2012. Turkserver: Enabling synchronous and longitudinal online experiments. In *Proceedings of the Fourth Workshop on Human Computation (HCOMP'12)*. AAAI Press.

Microsoft. 2016. Microsoft cognitive services api. `https://www.microsoft.com/cognitive-services/en-us/computer-vision-api`.

Morris, M. R.; Zolyomi, A.; Yao, C.; Bahram, S.; Bigham, J. P.; and Kane, S. K. 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*, 5506–5516. ACM.

Nieva, R. 2015. Google apologizes for algorithm kenly calling black people 'gorillas'. `https://www.cnet.com/news/google-apologizes-for-algorithm-mistakenly-calling-black-people-gorillas/`.

Shi, Y. 2006. E-government web site accessibility in australia and china: A longitudinal study. *Social Science Computer Review* 24(3):378–385.

Tran, K.; He, X.; Zhang, L.; Sun, J.; Carapcea, C.; Thrasher, C.; Buehler, C.; and Sienkiewicz, C. 2016. Rich image captioning in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 49–56.

Twitter. 2016. Twitter usage/company facts. `https://about.twitter.com/company`. Accessed: 2016-08-11.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3156–3164.

Wu, S.; Wieland, J.; Farivar, O.; and Schiller, J. 2017. Automatic alt-text: Computer-generated image descriptions for

blind users on a social network service. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1180–1192. ACM.

Wu, S.; Pique, H.; and Wieland, J. 2016. Using artificial intelligence to help blind people see facebook. http://newsroom.fb.com/news/2016/04/using-artificial-intelligence-to-help-blind-people-see-facebook/.

Yan, T.; Kumar, V.; and Ganesan, D. 2010. Crowdsearch: exploiting crowds for accurate real-time image search on mobile phones. In *Proceedings of the 8th international conference on Mobile systems, applications, and services*, 77–90. ACM.

Zyskowski, K.; Morris, M. R.; Bigham, J. P.; Gray, M. L.; and Kane, S. 2015. Accessible crowdwork? understanding the value in and challenge of microtask employment for people with disabilities. ACM Association for Computing Machinery.