A Formative Study on Designing Accurate and Natural Figure Captioning Systems

Xin Qian
University of Maryland
College Park, MD, USA
xinq@umd.edu

Sungchul Kim
Adobe Research
San Jose, CA, USA
sukim@adobe.com

Eunyee Koh
Adobe Research
San Jose, CA, USA
eunyee@adobe.com

Joel Chan
University of Maryland
College Park, MD, USA
joelchan@umd.edu

Fan Du
Adobe Research
San Jose, CA, USA
fdu@adobe.com

CHI'20 Extended Abstracts, April 25–30, 2020, Honolulu, HI, USA
ACM ISBN 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.3382946

Abstract
Automatic figure captioning is widely useful for improving the readability and accessibility of figures. Despite recent advances in figure question answering and parsing figure elements that enable machines to accurately read information from figures, the machine learning community still lacks sufficient understanding of this problem, on what contents are important to include in a caption and how to make it sound natural. In this work, we crawled, annotated, and analyzed a corpus of real-world human-written figure captions. Our study results show that real-world captions usually consist of a finite set of caption units and that automatic figure captioning should be formulated as a multi-stage task. The first stage is to generate caption units with high accuracy and the second is to stitch together the units with diverse stitching patterns, to form a natural caption.

Author Keywords
Figure captioning; text generation; data visualization.

CCS Concepts
•Human-centered computing → Natural language interfaces;

Introduction
Figures are visual representations of data to communicate patterns such as trends, comparisons, and rates. Scien-

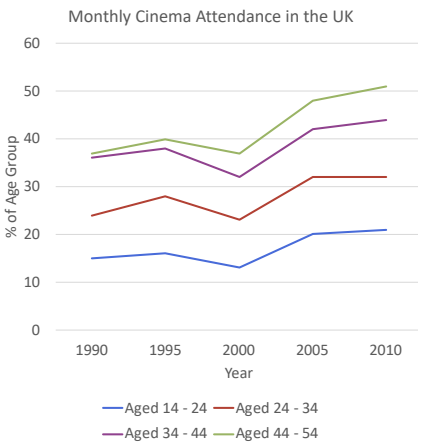

Figure 1: An example figure created from Microsoft Excel and to generate a caption with.

tific papers, business reports, websites, and other documents include them to supplement the textual contents. Common figure types include line charts, bar charts, pie charts, and scatter plots.[1] Figures are often displayed along with a caption summarizing their key information to improve their readability and accessibility. A descriptive caption can be used as the alternative text, enabling visually impaired users to read the figure [27, 28], improving search engine indexing, or allowing busy users to skim read a report. Despite the importance of figure captions, human-written captions can sometimes be trivial, non-informative, or absent altogether [2]. The arising call for web accessibility [3] inspired us to explore automatic figure captioning, which could be potentially useful to existing publishing tools such as WordPress, Google Web Designer, and Adobe Acrobat.

To automatically generate captions, machines need to parse figure elements, reason over their relationships, then describe in natural language. Recent progress in figure question answering [15, 16] and parsing figure elements [22, 24] demonstrated machines' capability in these processes, making figure captioning an emerging "last mile" problem. However, to formulate the problem and provide a viable solution, the machine learning community still lacks sufficient understanding of what contents are important to include in a caption that helps readers to know about a figure and how to make a caption sound natural.

We seek to fill this gap by providing design implications and guidelines toward designing accurate and natural figure captioning systems. We began by crawling a corpus of 95 real-world human-written figure captions that were written by expert tutors and considered as high-quality captions. Then, to understand which caption contents are important and investigate how different content categories are perceived by readers, we recruited nine annotators to serve the role of caption readers and annotate the corpus from a set of tags representing different types of information.

Our analysis found that real-world captions are composed of caption units from a finite set of types that are naturally stitched together. Correspondingly, system designers shall consider formulating the problem of figure captioning as a multi-stage task. The first stage focuses on generating caption units of specific types with high accuracy, and the second stage stitches together the caption units with diverse stitching patterns to form a natural caption. Our direct contributions are:

- A summary of a formative study that crawled, annotated, and analyzed a corpus[2] of real-world human-written figure captions.
- A discussion on design implications and guidelines toward designing accurate and natural figure captioning systems.

Related Work

Previous research on rule-based figure captioning and visual question answering for figures inspired our work.

Parsing Figures and Rule-based Figure Captioning

Previous work studied parsing and extracting elements from figures [6, 22, 24]. This work did not transform extracted elements into natural language descriptions. A few corpus studies examined the communicative goals behind single-sentence figure captions [10, 11, 19]. Rule-based figure captioning systems such as `PostGraphe` [12], `SAGE` [20], and `SelTex` [8] applied text planning on figure attributes to

[1] In this paper, we focus our discussion on the four common types.

[2] Available at bit.ly/figcap-corpus-analysis.

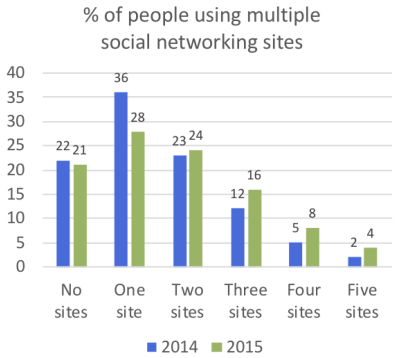

Figure 3: A figure being annotated.

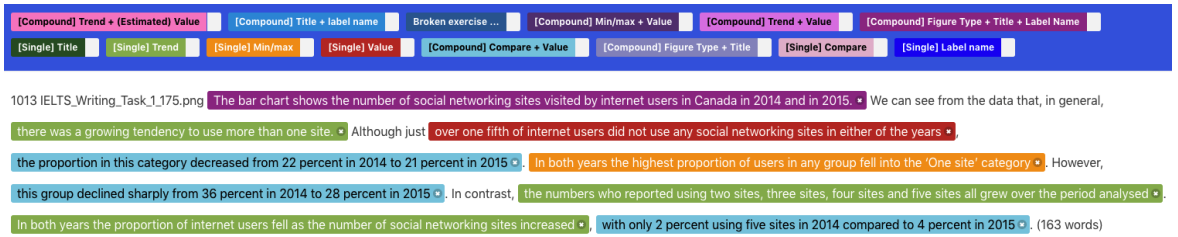

Figure 2: The annotation interface. This model answer corresponds to Figure 3.

Tag	Definition
Title	Describing or paraphrasing the title
Label name	Mention of text labels, (axis, legend, etc.)
Value	Specifying the value of an element in the figure
Min/max	Specifying the maximum or the minimum element
Compare	Comparing the value of two elements
Trend	Describing a high-level trend in the figure

Table 1: Single tag definitions.

create captions of pre-defined categories. Our analysis focuses on real-world captions that are more comprehensive and have multiple sentences and communicative goals. We aim to generate natural captions beyond pre-defined rules.

Figure Question Answering

Figure question answering (FQA) [4, 15, 16] is the visual question answering [1, 25] (VQA) task on figures. FQA has unique challenges, including handling figure-specific vocabulary and visual-semantic alignment [15]. Figure captioning can benefit from FQA challenges and solutions to accurately read information from figures. Existing figure captioning work [5] proposed a model to generate caption paragraphs, trained from aggregated figure question-answer pairs [16]. In comparison, we take one step back to understand what characterizes as useful captions that are also feasible for machine learning tasks, whether FQA datasets could help, and how to evaluate.

Corpus Analysis

To understand what contents are important in a figure caption, and how to make it natural, we crawled and annotated a corpus of real-world, human-written figure captions.

Corpus Details

The corpus consists of 95 model answers to an academic writing task in the IELTS Test.[3] The task requires ESL students to summarize the most important and relevant information in a given figure. The corpus is directly available, high-quality descriptions of figures, which are ideal outputs for our system. All model answers have at least 150 words, written by anonymous tutors from a public test preparation website.[4] They cover four common figure types (line, bar, pie charts, and scatter plots) and discuss topics in economics, environment, and education.

Participants and Procedure

To categorize caption contents and to investigate how different caption categories are perceived by caption readers, we recruited nine annotators to serve the role of caption readers and annotate the model answers. The annotators (P1-9) include five males and four females, aged 20-35 (M = 25.6, SD = 2.95). Six are graduate students in university research labs, and three are working professionals in the industry. All are based in the US, having graduate degrees and professional working proficiency in English.

An example annotation is shown in Figure 2. We deployed the interface as a Heroku app based on the doccano text annotation tool [21]. Each annotator followed the guideline to annotate 30 model answers on their PC remotely. It took, on average, 2.5 hours to complete. Each annotator received a $40 Amazon gift card at a rate of $16/hr.

[3] The task is officially called as the Academic Writing Task 1, as in www.ielts.org/en-us/about-the-test/test-format.

[4] www.ielts-exam.net/academic_writing_samples_task_1/

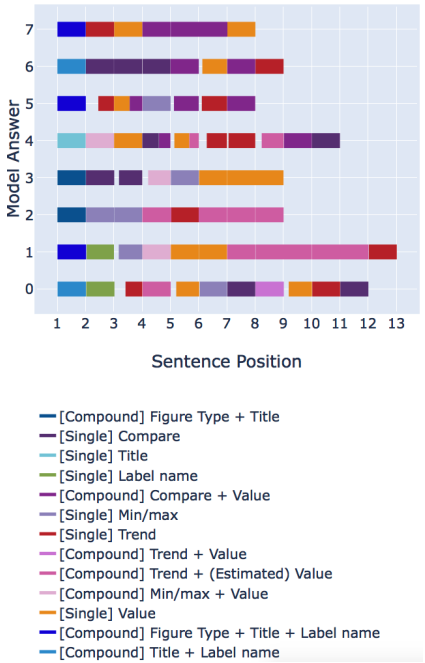

Figure 4: Tagging sequences of eight random model answers.

Group	Simple	Complex
Fleiss' κ	0.623	0.505

Table 2: We divide commonly annotated sentences into two groups, simple and complex, based on the simple heuristic of whether the sentence contains a comma. Simple sentences have higher Fleiss' kappa value.

At a high-level, the annotation guideline asks annotators to select spans of words from the model answer, and choose corresponding tags for the spans based on the types of information being covered. If no existing tag fits, they are strongly encouraged to add new tags.

The annotators and the authors jointly determined the guidelines in more detail. First, the authors drew inspirations from tagging tasks such as named entity recognition [18, Chapter 18] and dialog act recognition [26], where one single label tags each text instance. The authors piloted ten exercises and, until saturation, generated a set of six single tags in Table 1. Each tag represents one type of information in captions. Annotators gave email consent to this tag set. One annotator further proposed the minimum length for a span as no shorter than a clause.[5] This requirement allows spans to be semantically meaningful and independent units to users and are amenable to modeling. Accordingly, we limit the maximum length for a span as not exceeding the sentence boundary. Then, the tag set expanded to include compound tags. They are combinations of multiple single tags, to adapt to some clauses that cover multiple types of information. The multiple types in those clauses are intertwined, which cannot be separately tagged. For example, the clause *the bar chart shows the number of social media* covers both *figure type* and *title*. The initial tag set became six single tags and seven compound tags.

Results

Results suggest that captions can be tagged into spans of a finite set of types, which we define as "caption units." A caption unit is a span of words that cover one specific type of information about a figure. Caption units are not isolated from each other in captions. Instead, they are naturally stitched together in diverse patterns ("stitching patterns").

[5]A clause contains a subject and a predicate, and can stand by itself.

Figure 4 shows the tagging sequences for eight random model answers by one random annotator (P1).

Inter-annotator agreement: We calculate it on a common set of 20 model answers, at the sentence level, based on the tag sequence. Each sentence has a tag sequence from all spans that are being tagged in the sentence. We treat two tag sequences by different annotators as the same, if one tags "[Compound] Trend + Value" as a whole while the other tags "[Single] Trend" then "[Single] Value" for each clause. The Fleiss' kappa value [13] is 0.551 (moderate).[6]

The current level of agreement is promising despite two challenges in the annotation. First, annotators reported that *trend* and *compare* tags are sometimes ambiguous. Both are reasonable for sentences such as *the number of books read by men increased steadily between 2011 and 2012*, and *air pollution was a bigger problem in the early 20th century than it is now*. Merging the two tags improves the Fleiss' kappa value to 0.651 (substantial). This ambiguity also inspires us to create distinctive training data for the machine learning tasks on *trend* and *compare*.

Second, when sentences get more complicated, it becomes more challenging for annotators to make full sense of writer intent. Table 2 shows that annotators have higher agreement on simple sentences than complex ones. They sometimes failed to notice the major or other minor types of information or multiple clauses that should be separately tagged. One sentence *in contrast, the price of fresh fruits and vegetables rose significantly throughout this period* connects to previous sentences using words like "in contrast". Annotators chose the tag *compare* based on this connection that is more secondary and high-level than the

[6]We also calculate Fleiss' kappa at the word level as 0.516 (moderate). However, word-level calculation is strict. For the above example, all words are considered as having different tags.

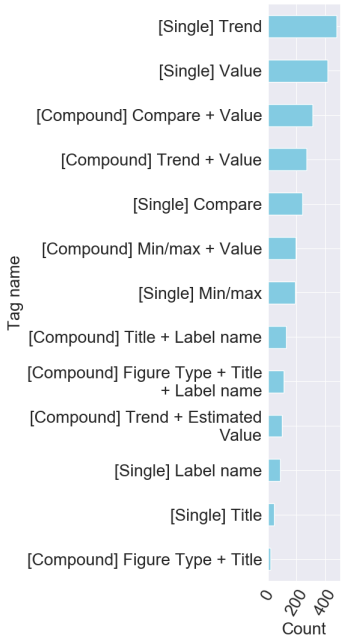

Figure 5: Counts for each tag (excluding new tags from P5).

major type *trend* conveyed in the rest of the sentence. We attribute high-level connections as "stitching patterns" later.

Tag statistics and "blank spans": Tagged spans are more and longer than "blank spans." For each model answer, there are 9.89 tagged spans with 17.53 words on average. For any consecutive sequence of words not being tagged, we treat them as "blank spans." Each model answer has an average of 5.92 blank spans, where each span has 2.73 words. All except one (P5) annotators solely used the initial tag set. P5 added four new compound tags that combine among *compare*, *min/max*, *trend*, and *value*.

Analyzing tag frequency shows us important ones to model as machine learning tasks. Figure 5 shows the counts for each tag. Single tags ranked by frequency are *trend*, *value*, *compare*, *min/max*, *label name*, and *title*. For compound tags, analyzing the combinations of different types shows us the most meaningful combinations to model as stitching patterns (e.g., *figure type* does not stand alone as a single tag but always associates with other single tags like *title*). For "blank spans," Figure 6 lists the top 10 unigrams, bigrams, and trigrams among all "blank spans." These "blank spans" fall into two major categories: **determiners** (e.g., "the," "this") and **conjunctions** (e.g., "and," "but").

Stitching patterns: Analyzing "blank spans" and their surrounding tagged spans (which we refer to as "caption units") shows how humans stitch together different caption units to make captions natural ("stitching patterns"). There are three major ones. The first is **co-reference**. When one caption unit follows another, humans replace duplicated subjects or objects with co-reference. In the sentence *this decreased by 0.3% in 2014*, "this" refers to *percentage of unemployed women* mentioned in the previous sentence.

The second is **subordination**, where humans strengthen a subjective statement (e.g., *compare*) with a subordinate clause on supporting evidence (e.g., *value*). Relative clauses are one type of subordination, where determiners (e.g. "which," "that") often lead those clauses. An example is *the only subject where boys' results were better than girls was Geography, where they achieved a pass rate of 30.4%.*

The third is **conjunction**. Humans add proper conjunction to connect values in two caption units, such as "however," "but," and "on the other hand." Depending on different relationships, the conjunction words can be "while" in *food came in 2nd place in Japan while in Malaysia the proportion was the third*, or "contrast" in *in stark contrast to this, for 2015, being underweight was only a problem among 20–29-year olds.* The three patterns can happen in various scopes: caption units are stitched together into sentences; sentences are stitched together into captions.

Implications and Guidelines

We summarize our findings as three implications below.

Implication 1: Accurate Units Make Accurate Captions
Captions are composed of caption units. The first stage is to accurately generate caption units. One basic modeling formulation is controllable image captioning [9]. Given a figure and a caption type, the model outputs a sequence of words as a caption unit for that type. Our study identified two additional formulation variations.

First, metadata information is a useful model input. Metadata has been shown effective in visual reasoning tasks for regular images [29]. For figures, metadata information includes the bounding boxes and values of figure elements and text labels. They could often be accurately extracted [14] or obtained (e.g., from Vega-lite [23] or Microsoft Excel), which could offload model burden without it doing feature extraction but only focusing reasoning.

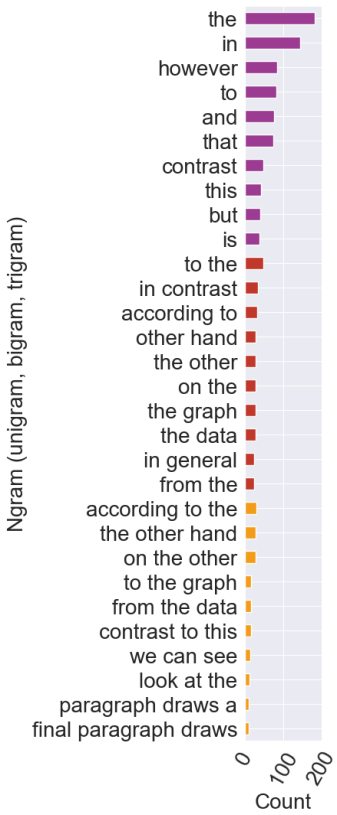

Figure 6: Top 10 unigrams (purple), bigrams (red), and trigrams (orange) among all "blank spans".

Second, we can define slots and values [18, Chapter 24] within ground-truth caption units and add them as additional outputs to the model, which can help us measure and improve the accuracy in system-generated caption units. In a system output *60 is the accuracy of the SVM model*, "60" is a wrong value for the slot "SVM". Correspondingly, models could have additional objective functions that optimize the accuracy in predicting slots and values.

Implication 2: Pattern Stitching Makes a Caption Natural
Once the system generates a set of accurate caption units, the next stage is to stitch them together to make a natural caption. System designers shall consider simulating the three stitching patterns from our corpus analysis results: co-reference, subordination, and conjunction. Besides, it needs to revisit caption units generated by the system and correct unnatural cases. We exemplify two unnatural cases.

First, figure captions for accessibility shall avoid abbreviations and be as explicit as possible. A screen reader might speak an abbreviated caption word (e.g., "ETA", estimated time of arrival) like a regular English word ("eta," the Greek letter), which is confusing. Punctuation is another unnatural case to be corrected. For example, the title of Figure 3 is *% of people using multiple social networking sites*. In Figure 2, it was corrected during stitching as *the bar chart shows the number of social networking sites visited by...*, where "the number of" replaces "%". Doing these corrections prevents unnatural units from cascading into later stage.

Implication 3: Keep Human in the Loop
We identified two avenues where humans could help improve the system. First, system developers may consider involving crowd-workers in the stitching task. Although our corpus covers figures of varying complexity, the three stitching patterns identified in the study may not cover all corner patterns on how humans write a natural caption. Gathering extra data from crowd-workers may improve the results. Second, developers may consider deploying a premature system to overcome the cold-start problem. The system could generate a caption and ask the figure author to accept, modify, or reject as feedback. With sufficient volume, the feedback can serve as additional training data. The system could also choose which figure caption to query users in an active learning manner.

Conclusion and Future Work

Automatic figure captioning is promising for improving the readability and accessibility of figures. In this paper, we collected and annotated a corpus of real-world figure captions written by expert tutors. Through an analysis of the corpus, we identified three design implications and guidelines for designing accurate and natural figure captioning systems.

In future work, we plan to release a figure captioning benchmark dataset for the research community. The dataset will include pairs of figures and caption units. One option to create it is by converting three existing figure question answering (FQA) datasets [4, 15, 16]. We will integrate figure metadata, convert question-answer pairs into caption units, and then categorize them into different types. A pilot study showed substantial overlaps between the types of question-answer pairs and our caption unit types in Table 1. Another option is to crawl figure-caption pairs from real scientific papers using PDFFigures 2.0 [7], categorize the captions, and parse figure metadata using FigureSeer [24]. We will develop a system from the dataset to demonstrate our design guidelines. We propose to integrate figure metadata information as additional system inputs. Sequence-to-sequence models on table-to-text applications [17] could be one inspiration, to encode the structured metadata information and decode a natural language caption unit.

REFERENCES

[1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2016. VQA: Visual question answering. *International Journal of Computer Vision* 1, 123 (2016), 4–31.

[2] Jeffrey P Bigham, Erin L Brady, Cole Gleason, Anhong Guo, and David A Shamma. 2016. An uninteresting tour through why our research papers aren't accessible. In *CHI-EA '16*.

[3] Erin Brady, Yu Zhong, and Jeffrey P. Bigham. 2015. Creating accessible PDFs for conference proceedings. In *W4A '15*.

[4] Ritwick Chaudhry, Sumit Shekhar, Utkarsh Gupta, Pranav Maneriker, Prann Bansal, and Ajay Joshi. 2019. LEAF-QA: Locate, encode & attend for figure question answering. *arXiv preprint arXiv:1907.12861* (2019).

[5] Charles Chen, Ruiyi Zhang, Eunyee Koh, Sungchul Kim, Scott Cohen, Tong Yu, Ryan A. Rossi, and Razvan C. Bunescu. 2019. Figure captioning with reasoning and sequence-level training. *arXiv preprint arXiv:1906.02850* (2019).

[6] Daniel Chester and Stephanie Elzer. 2005. Getting computers to see information graphics so users do not have to. In *International Symposium on Methodologies for Intelligent Systems*. Springer, 660–668.

[7] Christopher Clark and Santosh Divvala. 2016. PDFFigures 2.0: Mining figures from research papers. (2016).

[8] Marc Corio and Guy Lapalme. 1999. Generation of texts for information graphics. In *EWNLG '99*.

[9] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2019. Show, Control and Tell: A framework for generating controllable and grounded captions. In *CVPR '19*.

[10] Stephanie Elzer, Sandra Carberry, Daniel Chester, Seniz Demir, Nancy Green, Ingrid Zukerman, and Keith Trnka. 2005. Exploring and exploiting the limited utility of captions in recognizing intention in information graphics. In *ACL '05*.

[11] Stephanie Elzer, Sandra Carberry, and Ingrid Zukerman. 2011. The automated understanding of simple bar charts. *Artif. Intell.* 175, 2 (Feb. 2011), 526–555.

[12] Massimo Fasciano and Guy Lapalme. 1996. Postgraphe: a system for the generation of statistical graphics and text. In *INLG '96*.

[13] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[14] Morten Jessen, Falk Böschen, and Ansgar Scherp. 2019. Text localization in scientific figures using fully convolutional neural networks on limited training data. In *DocEng '19*.

[15] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. DVQA: Understanding data visualizations via question answering. In *CVPR '18*.

[16] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300* (2017).

[17] Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. 2018. Table-to-text generation by structure-aware seq2seq learning. In *AAAI '18*.

[18] James H Martin and Daniel Jurafsky. 2009. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson/Prentice Hall Upper Saddle River.

[19] Kathleen F McCoy, Sandra Carberry, Tom Roper, and Nancy Green. 2001. Towards generating textual summaries of graphs.

[20] Vibhu O. Mittal, Johanna D. Moore, Giuseppe Carenini, and Steven Roth. 1998. Describing Complex Charts in Natural Language: A caption generation system. *Computational Linguistics* 24, 3 (1998), 431–467.

[21] Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. (2018). `https://github.com/doccano/doccano`

[22] Jorge Poco and Jeffrey Heer. 2017. Reverse-Engineering Visualizations: Recovering visual encodings from chart images. *Computer Graphics Forum (Proc. EuroVis)* 36, 3 (2017), 353–363.

[23] Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. 2016. Vega-lite: A grammar of interactive graphics. *IEEE transactions on visualization and computer graphics* 23, 1 (2016), 341–350.

[24] Noah Siegel, Zachary Horvitz, Roie Levin, Santosh Divvala, and Ali Farhadi. 2016. FigureSeer: Parsing result-figures in research papers. In *ECCV '16*.

[25] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *CVPR '19*.

[26] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics* 26, 3 (2000), 339–374.

[27] W3C 2016. Applying text alternatives to images with the Alt entry in PDF documents. (2016). `https://www.w3.org/TR/WCAG20-TECHS/pdf#PDF1`

[28] W3C 2019. A first review of web accessibility. (2019). `https://www.w3.org/WAI/test-evaluate/preliminary/#images`

[29] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. 2018. Neural-symbolic VQA: Disentangling reasoning from vision and language understanding. In *NeurIPS '18*.