A Survey of Current Datasets for Vision and Language Research

Francis Ferraro[1*], **Nasrin Mostafazadeh**[2*], **Ting-Hao (Kenneth) Huang**[3], **Lucy Vanderwende**[4], **Jacob Devlin**[4], **Michel Galley**[4], **Margaret Mitchell**[4]

Microsoft Research

1 Johns Hopkins University, 2 University of Rochester, 3 Carnegie Mellon University,
4 Corresponding authors: {lucyv,jdevlin,mgalley,memitc}@microsoft.com

Abstract

Integrating vision and language has long been a dream in work on artificial intelligence (AI). In the past two years, we have witnessed an explosion of work that brings together vision and language from images to videos and beyond. The available corpora have played a crucial role in advancing this area of research. In this paper, we propose a set of quality metrics for evaluating and analyzing the vision & language datasets and categorize them accordingly. Our analyses show that the most recent datasets have been using more complex language and more abstract concepts, however, there are different strengths and weaknesses in each.

1 Introduction

Bringing together language and vision in one intelligent system has long been an ambition in AI research, beginning with SHRDLU as one of the first vision-language integration systems (Winograd, 1972) and continuing with more recent attempts on conversational robots grounded in the visual world (Kollar et al., 2013; Cantrell et al., 2010; Matuszek et al., 2012; Kruijff et al., 2007; Roy et al., 2003). In the past few years, an influx of new, large vision & language corpora, alongside dramatic advances in vision research, has sparked renewed interest in connecting vision and language. Vision & language corpora now provide alignments between visual content that can be recognized with Computer Vision (CV) algorithms and language that can be understood and generated using Natural Language Processing techniques.

Fueled in part by the newly emerging data, research that blends techniques in vision and in language has increased at an incredible rate. In just the past year, recent work has proposed methods for image and video captioning (Fang et al., 2014; Donahue et al., 2014; Venugopalan et al., 2015), summarization (Kim et al., 2015), reference (Kazemzadeh et al., 2014), and question answering (Antol et al., 2015; Gao et al., 2015), to name just a few. The newly crafted large-scale vision & language datasets have played a crucial role in defining this research, serving as a foundation for training/testing and helping to set benchmarks for measuring system performance.

Crowdsourcing and large image collections such as those provided by Flickr[1] have made it possible for researchers to propose methods for vision and language tasks alongside an accompanying dataset. However, as more and more datasets have emerged in this space, it has become unclear how different methods generalize beyond the datasets they are evaluated on, and what data may be useful for moving the field beyond a single task, towards solving larger AI problems.

In this paper, we take a step back to document this moment in time, making a record of the major available corpora that are driving the field. We provide a quantitative analysis of each of these corpora in order to understand the characteristics of each, and how they compare to one another. The quality of a dataset must be measured and compared to related datasets, as low quality data may distort an entire subfield. We propose a set of criteria for analyzing, evaluating and comparing the quality of vision & language datasets against each other. Knowing the details of a dataset compared to similar datasets allows researchers to define more precisely what task(s) they are trying to solve, and select the dataset(s) best suited to their goals, while being aware of the implications and biases the datasets could impose on a task.

We categorize the available datasets into three major classes and evaluate them against these cri-

[*] F.F. and N.M. contributed equally to this work.

[1] http://www.flickr.com

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 207–213,
Lisbon, Portugal, 17-21 September 2015. ©2015 Association for Computational Linguistics.

teria. The datasets we present here were chosen because they are all available to the community and cover the data that has been created to support the recent focus on image captioning work. More importantly, we provide an evolving website[2] containing pointers and references to many more vision-to-language datasets, which we believe will be valuable in unifying the quickly expanding research tasks in language and vision.

2 Quality Criteria for Language & Vision Datasets

The quality of a dataset is highly dependent on the sampling and scraping techniques used early in the data collection process. However, the content of datasets can play a major role in narrowing the focus of the field. Datasets are affected by both *reporting bias* (Gordon and Durme, 2013), where the frequency with which people write about actions, events, or states does not directly reflect real-world frequencies of those phenomena; they are also affected by *photographer's bias* (Torralba and Efros, 2011), where photographs are somewhat predictable within a given domain. This suggests that new datasets may be useful towards the larger AI goal if provided alongside a set of quantitative metrics that show how they compare against similar corpora, as well as more general "background" corpora. Such metrics can be used as indicators of dataset bias and language richness. At a higher level, we argue that clearly defined metrics are necessary to provide quantitative measurements of how a new dataset compares to previous work. This helps clarify and benchmark how research is progressing towards a broader AI goal as more and more data comes into play.

In this section, we propose a set of such metrics that characterize vision & language datasets. We focus on methods to measure *language quality* that can be used across several corpora. We also briefly examine metrics for *vision quality*. We evaluate several recent datasets based on all proposed metrics in Section 4, with results reported in Tables 1, 2, and Figure 1.

2.1 Language Quality

We define the following criteria for evaluating the captions or instructions of the datasets:

- **Vocabulary Size** (*#vocab*), the number of unique vocabulary words.
- **Syntactic Complexity** (*Frazier*, *Yngve*) measures the amount of embedding/branching in a sentence's syntax. We report mean Yngve (Yngve, 1960) and Frazier measurements (Frazier, 1985); each provides a different counting on the number of nodes in the phrase markers of syntactic trees.
- **Part of Speech Distribution** measures the distribution of nouns, verbs, adjectives, and other parts of speech.
- **Abstract:Concrete Ratio** (*#Conc*, *#Abs*, *%Abs*) indicates the range of visual and non-visual concepts the dataset covers. Abstract terms are ideas or concepts, such as 'love' or 'think' and concrete terms are all the objects or events that are mainly available to the senses. For this purpose, we use a list of most common abstract terms in English (Vanderwende et al., 2015), and define concrete terms as all other words except for a small set of function words.
- **Average Sentence Length** (*Sent Len.*) shows how rich and descriptive the sentences are.
- **Perplexity** provides a measure of data skew by measuring how expected sentences are from one corpus according to a model trained on another corpus. We analyze perplexity (*Ppl*) for each dataset against a 5-gram language model learned on a generic 30B words English dataset. We further analyze pair-wise perplexity of datasets against each other in Section 4.

2.2 Vision Quality

Our focus in this survey is mainly on language, however, the characteristics of images or videos and their corresponding annotations is as important in vision & language research. The quality of vision in a dataset can be characterized in part by the variety of visual subjects and scenes provided, as well as the richness of the annotations (e.g., segmentation using bounding boxes (*BB*) or visual dependencies between boxes). Moreover, a vision corpus can use abstract or real images (*Abs/Real*).

3 The Available Datasets

We group a representative set of available datasets based on their content. For a complete list of datasets and their descriptions, please refer to the supplementary website.[2]

3.1 Captioned Images

Several recent vision & language datasets provide one or multiple captions per image. The captions

[2] http://visionandlanguage.net

of these datasets are either the original photo title and descriptions provided by online users (Ordonez et al., 2011; Thomee et al., 2015), or the captions generated by crowd workers for existing images. The former datasets tend to be larger in size and contain more contextual descriptions.

3.1.1 User-generated Captions

• **SBU Captioned Photo Dataset** (Ordonez et al., 2011) contains 1 million images with original user generated captions, collected in the wild by systematic querying of Flickr. This dataset is collected by querying Flickr for specific terms such as objects and actions and then filtered images with descriptions longer than certain mean length.
• **Déjà Images Dataset** (Chen et al., 2015) consists of 180K unique user-generated captions associated with 4M Flickr images, where one caption is aligned with multiple images. This dataset was collected by querying Flickr for 693 high frequency nouns, then further filtered to have at least one verb and be judged as "good" captions by workers on Amazon's Mechanical Turk (Turkers).

3.1.2 Crowd-sourced Captions

• **UIUC Pascal Dataset** (Farhadi et al., 2010) is probably one of the first datasets aligning images with captions. Pascal dataset contains 1,000 images with 5 sentences per image.
• **Flickr 30K Images** (Young et al., 2014) extends previous Flickr datasets (Rashtchian et al., 2010), and includes 158,915 crowd-sourced captions that describe 31,783 images of people involved in everyday activities and events.
• **Microsoft COCO Dataset (MS COCO)** (Lin et al., 2014) includes complex everyday scenes with common objects in naturally occurring contexts. Objects in the scene are labeled using per-instance segmentations. In total, this dataset contains photos of 91 basic object types with 2.5 million labeled instances in 328k images, each paired with 5 captions. This dataset gave rise to the CVPR 2015 image captioning challenge and is continuing to be a benchmark for comparing various aspects of vision and language research.
• **Abstract Scenes Dataset (Clipart)** (Zitnick et al., 2013) was created with the goal of representing real-world scenes with clipart to study scene semantics isolated from object recognition and segmentation issues in image processing. This removes the burden of low-level vision tasks. This dataset contains 10,020 images of children playing outdoors associated with total 60,396 descriptions.

3.1.3 Captions of Densely Labeled Images

Existing caption datasets provide images paired with captions, but such brief image descriptions capture only a subset of the content in each image. Measuring the magnitude of the reporting bias inherent in such descriptions helps us to understand the discrepancy between what we can learn for the specific task of image captioning versus what we can learn more generally from the photographs people take. One dataset useful to this end provides image annotation for content selection:
• **Microsoft Research Dense Visual Annotation Corpus** (Yatskar et al., 2014) provides a set of 500 images from the Flickr 8K dataset (Rashtchian et al., 2010) that are densely labeled with 100,000 textual labels, with bounding boxes and facets annotated for each object. This approximates "gold standard" visual recognition.

To get a rough estimate of the reporting bias in image captioning, we determined the percentage of top-level objects[3] that are mentioned in the captions for this dataset out of all the objects that are annotated. Of the average 8.04 available top-level objects in the image, each of the captions only reports an average of 2.7 of these objects.[4] A more detailed analysis of reporting bias is beyond the scope of this paper, but we found that many of the biases (e.g., people selection) found with abstract scenes (Zitnick et al., 2013) are also present with photos.

3.2 Video Description and Instruction

Video datasets aligned with descriptions (Chen et al., 2010; Rohrbach et al., 2012; Regneri et al., 2013; Naim et al., 2015; Malmaud et al., 2015) generally represent limited domains and small lexicons, which is due to the fact that video processing and understanding is a very compute-intensive task. Available datasets include:
• **Short Videos Described with Sentences** (Yu and Siskind, 2013) includes 61 video clips (each 35 seconds in length, filmed in three different

[3] This visual annotation consists of a two-level hierarchy, where multiple Turkers enumerated and located objects and stuff in each image, and these objects were then further labeled with finer-grained object information (*Has* attributes).

[4] We did not use an external synonym or paraphrasing resource to perform the matching between labels and captions, as the dataset itself provides paraphrases for each object: each object is labeled by multiple Turkers, who labeled *Isa* relations (e.g., "eagle" is a "bird").

		Size(k)		Language								Vision	
	Dataset	Img	Txt	Frazier	Yngve	Vocab Size (k)	Sent Len.	#Conc	#Abs	%Abs	Ppl	(A)bs/ (R)eal	BB
Balanced	Brown	-	52	18.5	77.21	47.7	20.82	40411	7264	15.24%	194	-	-
User-Gen	SBU	1000	1000	9.70	26.03	254.6	13.29	243940	9495	3.74%	346	R	-
	Deja	4000	180	4.13	4.71	38.3	4.10	34581	3714	9.70%	184	R	-
Crowd-sourced	Pascal	1	5	8.03	25.78	3.4	10.78	2741	591	17.74%	123	R	-
	Flickr30K	32	159	9.50	27.00	20.3	12.98	17214	3033	14.98%	118	R	-
	COCO	328	2500	9.11	24.92	24.9	11.30	21607	3218	12.96%	121	R	Y
	Clipart	10	60	6.50	12.24	2.7	7.18	2202	482	17.96%	126	A	Y
Video	VDC	2	85	6.71	15.18	13.6	7.97	11795	1741	12.86%	148	R	-
Beyond	VQA	10	330	6.50	14.00	6.2	7.58	5019	1194	19.22%	113	A/R	-
	CQA	123	118	9.69	11.18	10.2	8.65	8501	1636	16.14%	199	R	Y
	VML	11	360	6.83	12.72	11.2	7.56	9220	1914	17.19%	110	R	Y

Table 1: Summary of statistics and quality metrics of a sample set of major datasets. For Brown, we report Frazier and Yngve scores on automatically acquired parses, but we also compute them for the 24K sentences with gold parses: in this setting, the mean Frazier score is 15.26 while the mean Yngve score is 58.48.

outdoor environments), showing multiple simultaneous events between a subset of four objects: a person, a backpack, a chair, and a trash-can. Each video was manually annotated (with very restricted grammar and lexicon) with several sentences describing what occurs in the video.

• **Microsoft Research Video Description Corpus (MS VDC)** (Chen and Dolan, 2011) contains parallel descriptions (85,550 English ones) of 2,089 short video snippets (10-25 seconds long). The descriptions are one sentence summaries about the actions or events in the video as described by Amazon Turkers. In this dataset, both paraphrase and bilingual alternatives are captured, hence, the dataset can be useful translation, paraphrasing, and video description purposes.

3.3 Beyond Visual Description

Recent work has demonstrated that n-gram language modeling paired with scene-level understanding of an image trained on large enough datasets can result in reasonable automatically generated captions (Fang et al., 2014; Donahue et al., 2014). Some works have proposed to step beyond description generation, towards deeper AI tasks such as question answering (Ren et al., 2015; Malinowski and Fritz, 2014). We present two of these attempts below:

• **Visual Madlibs Dataset (VML)** (Yu et al., 2015) is a subset of 10,783 images from the MS COCO dataset which aims to go beyond describing which objects are in the image. For a given image, three Amazon Turkers were prompted to complete one of 12 fill-in-the-blank template questions, such as 'when I look at this picture, I feel –', selected automatically based on the image content. This dataset contains a total of 360,001 MadLib question and answers.

• **Visual Question Answering (VQA) Dataset** (Antol et al., 2015) is created for the task of open-ended VQA, where a system can be presented with an image and a free-form natural-language question (e.g., 'how many people are in the photo?'), and should be able to answer the question. This dataset contains both real images and abstract scenes, paired with questions and answers. Real images include 123,285 images from MS COCO dataset, and 10,000 clip-art abstract scenes, made up from 20 'paperdoll' human models with adjustable limbs and over 100 objects and 31 animals. Amazon Turkers were prompted to create 'interesting' questions, resulting in 215,150 questions and 430,920 answers.

• **Toronto COCO-QA Dataset (CQA)** (Ren et al., 2015) is also a visual question answering dataset, where the questions are automatically generated from image captions of MS COCO dataset. This dataset has a total of 123,287 images with 117,684 questions with one-word answer about objects, numbers, colors, or locations.

4 Analysis

We analyze the datasets introduced in Section 3 according to the metrics defined in Section 2, using the Stanford CoreNLP suite to acquire parses and part-of-speech tags (Manning et al., 2014). We also include the Brown corpus (Francis and Kucera, 1979; Marcus et al., 1999) as a reference point. We find evidence that the VQA dataset captures more abstract concepts than other datasets, with almost 20% of the words found in our abstract concept resource. The Deja corpus has the least number of abstract concepts, followed by COCO and VDC. This reflects differences in col-

	Brown	Clipart	Coco	Flickr30K	CQA	VDC	VQA	Pascal	SBU
Brown	237.1	99.6	560.8	405.0	354.039	187.3	126.5	47.8	621.5
Clipart	233.6	11.2	117.4	109.4	210.8	82.5	114.7	28.7	130.6
Coco	274.6	59.2	36.2	75.3	137.0	87.1	236.9	39.3	111.0
Flickr30K	247.8	78.5	54.3	37.8	181.5	72.1	192.2	39.9	125.0
CQA	489.4	186.1	137.0	244.5	33.8	259.0	72.1	74.9	200.1
VDC	200.5	52.4	61.5	51.1	289.9	30.0	180.1	28.7	154.5
VQA	425.9	368.8	366.8	665.8	317.7	455.0	19.6	119.3	281.0
Pascal	265.2	64.5	43.2	63.4	174.2	83.0	228.2	36.0	105.3
SBU	473.9	107.1	346.4	344.0	328.5	230.7	194.3	78.2	119.8
#vocab	*14.0k*	*1.1k*	*13k*	*9.4k*	*5.3k*	*4.9k*	*1.4k*	*1.0k*	*65.1k*

Table 2: Perplexities across corpora, where rows represent test sets (20k sentences) and columns training sets (remaining sentences). To make perplexities comparable, we used the same vocabulary frequency cutoff of 3. All models are 5-grams.

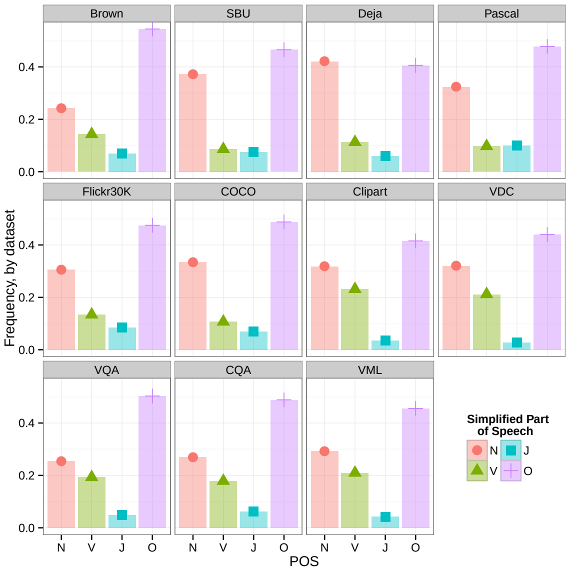

Figure 1: Simplified part-of-speech distributions for the eight datasets. We include the POS tags from the balanced Brown corpus (Marcus et al., 1999) to contextualize any very shallow syntactic biases. We mapped all nouns to "N," all verbs to "V," all adjectives to "J" and all other POS tags to "O."

lecting the various corpora: For example, the Deja corpus was collected to find *specifically* visual phrases that can be used to describe multiple images. This corpus also has the most syntactically simple phrases, as measured by both Frazier and Yngve; this is likely caused by the phrases needing to be general enough to capture multiple images.

The most syntactically complex sentences are found in the Flickr30K, COCO and CQA datasets. However, the CQA dataset suffers from a high perplexity against a background corpus relative to the other datasets, at odds with relatively short sentence lengths. This suggests that the automatic caption-to-question conversion may be creating unexpectedly complex sentences that are less reflective of general language usage. In contrast, the COCO and Flickr30K dataset's relatively high syntactic complexity is in line with their relatively high sentence length.

Table 2 illustrates further similarities between datasets, and a more fine-grained use of perplexity to measure the usefulness of a given training set for predicting words of a given test set. Some datasets such as COCO, Flickr30K, and Clipart are generally more useful as out-domain data compared to the QA datasets. Test sets for VQA and CQA are quite idiosyncratic and yield poor perplexity unless trained on in-domain data. As shown in Figure 1, the COCO dataset is balanced across POS tags most similarly to the balanced Brown corpus (Marcus et al., 1999). The Clipart dataset provides the highest proportion of verbs, which often correspond to actions/poses in vision research, while the Flickr30K corpus provides the most nouns, which often correspond to object/stuff categories in vision research.

We emphasize here that the distinction between a qualitatively good or bad dataset is task dependent. Therefore, all these metrics and the obtained results provide the researchers with an objective set of criteria so that they can make the decision whether a dataset is suitable to a particular task.

5 Conclusion

We detail the recent growth of vision & language corpora and compare and contrast several recently released large datasets. We argue that newly introduced corpora may measure how they compare to similar datasets by measuring *perplexity*, *syntactic complexity*, *abstract:concrete* word ratios, among other metrics. By leveraging such metrics and comparing across corpora, research can be sensitive to how datasets are biased in different directions, and define new corpora accordingly.

Acknowledgements We thank the three anonymous reviewers for their feedback, and Benjamin Van Durme for discussions on reporting bias.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. *arXiv preprint arXiv:1505.00468*.

Rehj Cantrell, Matthias Scheutz, Paul W. Schermerhorn, and Xuan Wu. 2010. Robust spoken instruction understanding for hri. In Pamela J. Hinds, Hiroshi Ishiguro, Takayuki Kanda, and Peter H. Kahn Jr., editors, *HRI*, pages 275–282. ACM.

David L. Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 190–200, Stroudsburg, PA, USA. Association for Computational Linguistics.

David L. Chen, Joohyun Kim, and Raymond J. Mooney. 2010. Training a multilingual sportscaster: Using perceptual context to learn language. *J. Artif. Int. Res.*, 37(1):397–436, January.

Jianfu Chen, Polina Kuznetsova, David Warren, and Yejin Choi. 2015. Déjà image-captions: A corpus of expressive descriptions in repetition. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 504–514, Denver, Colorado, May–June. Association for Computational Linguistics.

Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2014. Long-term recurrent convolutional networks for visual recognition and description. *CoRR*, abs/1411.4389.

Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2014. From captions to visual concepts and back. *CoRR*, abs/1411.4952.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, pages 15–29, Berlin, Heidelberg. Springer-Verlag.

W Nelson Francis and Henry Kucera. 1979. Brown Corpus manual: Manual of information to accompany a standard corpus of present-day edited American English for use with digital computers. *Brown University, Providence, Rhode Island, USA*.

L. Frazier. 1985. Syntactic complexity. In D. R. Dowty, L. Karttunen, and A. M. Zwicky, editors, *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*, pages 129–189. Cambridge University Press, Cambridge.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question answering. *CoRR*, abs/1505.05612.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge extraction. In *Automated Knowledge Base Construction (AKBC) 2013: The 3rd Workshop on Knowledge Extraction, at CIKM 2013*, AKBC'13.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar, October. Association for Computational Linguistics.

Gunhee Kim, Seungwhan Moon, and Leonid Sigal. 2015. Joint Photo Stream and Blog Post Summarization and Exploration. In *28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*.

Thomas Kollar, Jayant Krishnamurthy, and Grant Strimel. 2013. Toward interactive grounded language acquisition. In *Robotics: Science and Systems*.

Geert-Jan M. Kruijff, Hendrik Zender, Patric Jensfelt, and Henrik I. Christensen. 2007. Situated dialogue and spatial organization: What, where...and why? *International Journal of Advanced Robotic Systems, Special Issue on Human and Robot Interactive Communication*, 4(2), March.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems 27*, pages 1682–1690.

Jonathan Malmaud, Jonathan Huang, Vivek Rathod, Nicholas Johnston, Andrew Rabinovich, and Kevin Murphy. 2015. Whats cookin? interpreting cooking videos using text, speech and vision. In *North American Chapter of the Association for Computational Linguistics Human Language Technologies (NAACL HLT 2015), May 31 - June 5, 2015, Denver, Colorado USA*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Mitchell Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. Brown corpus, treebank-3.

Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. 2012. A Joint Model of Language and Perception for Grounded Attribute Learning. In *Proc. of the 2012 International Conference on Machine Learning*, Edinburgh, Scotland, June.

Iftekhar Naim, Young C. Song, Qiguang Liu, Liang Huang, Henry Kautz, Jiebo Luo, and Daniel Gildea. 2015. Discriminative unsupervised alignment of natural language instructions with corresponding video segments. In *North American Chapter of the Association for Computational Linguistics Human Language Technologies (NAACL HLT 2015), May 31 - June 5, 2015, Denver, Colorado USA*.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 139–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics (TACL)*, 1:25–36.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Question answering about images using visual semantic embeddings. In *Deep Learning Workshop, ICML 2015*.

Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. 2012. A database for fine grained activity detection of cooking activities. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, IEEE, June.

Deb Roy, Kai-Yuh Hsiao, and Nikolaos Mavridis. 2003. Conversational robots: Building blocks for grounding word meaning. In *Proceedings of the HLT-NAACL 2003 Workshop on Learning Word Meaning from Non-linguistic Data - Volume 6*, HLT-NAACL-LWM '04, pages 70–77, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2015. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*.

A. Torralba and A. A. Efros. 2011. Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 1521–1528, Washington, DC, USA. IEEE Computer Society.

Lucy Vanderwende, Arul Menezes, and Chris Quirk. 2015. An amr parser for english, french, german, spanish and japanese and a new amr-annotated corpus. Proceedings of NAACL 2015, June.

Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. 2015. Translating videos to natural language using deep recurrent neural networks. In *Proceedings the 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT 2015)*, pages 1494–1504, Denver, Colorado, June.

Terry Winograd. 1972. *Understanding Natural Language*. Academic Press, New York.

Mark Yatskar, Michel Galley, Lucy Vanderwende, and Luke Zettlemoyer. 2014. See no evil, say no evil: Description generation from densely labeled images. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 110–120, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Victor H. Yngve. 1960. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104:444–466.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Haonan Yu and Jeffrey Mark Siskind. 2013. Grounded language learning from video described with sentences. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 53–63, Sofia, Bulgaria. Association for Computational Linguistics. Best Paper Award.

Licheng Yu, Eunbyung Park, Alexander C. Berg, and Tamara L. Berg. 2015. Visual Madlibs: Fill in the blank Image Generation and Question Answering. *arXiv preprint arXiv:1506.00278*.

C. Lawrence Zitnick, Devi Parikh, and Lucy Vanderwende. 2013. Learning the visual interpretation of sentences. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 1681–1688.