"Caption" as a Coherence Relation: Evidence and Implications

Malihe Alikhani
Computer Science
Rutgers University
malihe.alikhani@rutgers.edu

Matthew Stone
Computer Science
Rutgers University
matthew.stone@rutgers.edu

Abstract

We study verbs in image–text corpora, contrasting *caption* corpora, where texts are explicitly written to characterize image content, with *depiction* corpora, where texts and images may stand in more general relations. Captions show a distinctively limited distribution of verbs, with strong preferences for specific tense, aspect, lexical aspect, and semantic field. These limitations, which appear in data elicited by a range of methods, restrict the utility of caption corpora to inform image retrieval, multimodal document generation, and perceptually-grounded semantic models. We suggest that these limitations reflect the discourse constraints in play when subjects write texts to accompany imagery, so we argue that future development of image–text corpora should work to increase the diversity of event descriptions, while looking explicitly at the different ways text and imagery can be coherently related.

1 Introduction

Researchers interested in modeling relations between language and the world are increasingly starting from multimodal corpora that combine text with visual information; see Bernardi et al. (2017) for review.

A key benchmark problem, which we explore here, is to learn to produce an appropriate text caption to accompany an image. This problem brings fundamental scientific and engineering challenges, and has immediate applications, particularly in making online content more accessible. At the same time, the problem lends itself to appealing high-level characterizations—learning to describe in words what's happening in an image—which suggests that the line of research affords sweeping insights into depiction, image retrieval, and real-world commonsense inference.

In this paper, we offer a theoretically-situated but empirically-motivated critique of this broader understanding of captioning. We argue that current image–caption corpora systematically suffer from key deficits in coverage, and therefore cannot underpin general models for linking images and text. Instead, we suggest that these deficits might be remedied through attention to different corpora and different image–text relationships.

Our starting point is the observation that images and text in multimodal documents are used coherently together: like all contributions to discourse, they stand in particular relations to one another, which guide readers toward the inferential connections intended by the author (Hobbs, 1990). **Captioning**, we argue, is such a relation. A text that is presented as the caption to an image presents restricted kinds of information about the image and adopts a distinctive perspective. In particular, we suggest, captions characteristically describe imagery as though what we see has been going on indefinitely in the past, is happening now, and will continue indefinitely into the future.

We justify this account of captioning with an empirical study of action descriptions in English image captioning corpora. Our central finding is that they are disproportionately **atelic**, meaning that they describe an ongoing process in a general way, without invoking its possible goal, endpoint or culmination; see Hamm and Bott (2018). This is the difference between *painting an advertisement* (telic) and *using oils* (atelic); *performing their hit song* (telic) and *performing on stage* (atelic); *running a 5K* (telic) and simply *running* (atelic). Of course, captions frequently feature **stative** descriptions, which evoke conditions rather than activities: *names are etched on a wall*, *the building towers over the skyline*.

Captioning is just one of many possible coherence relations connecting text and imagery: we

Proceedings of the Second Workshop on Shortcomings in Vision and Language, pages 58–67
Minneapolis, USA, June 6, 2019.

(a) People are standing outside next to a food truck.

(b) A man is sitting in front of a bunch of fruit.

(c) It was a beautiful day for him.

(d) Actor and guest arrive at the premiere.

(e) Score small X at base of each peach with paring knife.

(f) Lower peaches into boiling water and simmer until skins loosen, 30 to 60 seconds.

Figure 1: The difference in instruction results in different captions. People take a particular perspective when writing captions. (a) and (b) are examples from COCO. (c) shows one step of a story in VIST. (d) is an example from the Google caption dataset. (e) and (f) are examples of two steps of a multimodal recipe.
Photo credits: (a) by Gary Soup, (b) by Carol Mitchell, (c) by Jeff Kravitz/FilmMagic/GettyImages, (e) and (f) by Kate Kelly/AmericasTestKitchen.

can find diverse relations considering a broader range of corpus data. Figure 1 illustrates these possibilities. Figure 1(a) and (b), from MSCOCO (Lin et al., 2014), are typical descriptive examples from caption data sets, describing imagery in terms of open-ended activities. Figure 1(c), from (Huang et al., 2016), and (d), from (Sharma et al., 2018), exhibit another possibility: these images are accompanied by **play-by-play text**, written in the narrative present (Pullum et al., 2002, 129), which suggests that the photo catches the moment that makes the captions true. Many other cases, we argue, are best analyzed in terms of an **illustration** relation connecting text to an accompanying image. As shown in Figure 1(e) and (f), from (Yagcioglu et al., 2018), illustration relations allow for diverse verbs—telic, atelic and stative alike—to be described in the text.

Thus, where vision–language applications involve this illustration relation, as is plausible in many cases of image retrieval, document synthesis, and grounded language use, caption corpora will systematically lack the full range of action descriptions that general solutions must handle. We conclude by arguing that future researchers should focus on naturally-occurring examples, where text and images connect in diverse ways, and should explicitly model the coherence relationships between text and images.

2 Related Work

Vision–language corpora have inspired a range of approaches for image retrieval and language generation, and increasing awareness of the biases of corpora and models is bringing increased attention to the linguistic characteristics of the corpora (Bernardi et al., 2017; Ferraro et al., 2015). For example, van Miltenburg et al. (2018a) present a tax-

K	COCO	Flickr	VIST	CC	Recipe	ANC
Top 10	0.599	0.594	0.538	0.390	0.392	0.443
Top 30	0.724	0.723	0.669	0.535	0.511	0.563
Top 100	0.864	0.840	0.822	0.834	0.715	0.709
Top 300	0.948	0.934	0.920	0.930	0.862	0.840

Table 1: Fraction of verbal part-of-speech tokens accounted for by top K verb lemmas, by corpus. Frequent verbs disproportionately dominate in captions.

onomy of the ways that subjects refer to people in the images, while van Miltenburg et al. (2018b) investigate the difference between spoken and written image descriptions. We continue this trend by offering a comparative study of verb use in multimodal corpora for the first time.

Authors intend contributions to play specific roles in multimodal discourse. Previous works characterized the inferences that guide interpretations between images in terms of coherence relations (McCloud, 1993; Cohn, 2013; Cumming et al., 2017). In this work, we explore relations between images and text, with a particular emphasis on the link between images and captions.

Gella et al. (2019) presented a model for disambiguating verb senses in images (e.g. playing guitar v.s. children playing) using a single verb and the related image as the inputs of the system. Our work is different because we are investigating how people write captions for images and not a single verb.

We investigate the relationship between tense, aspect and discourse structure in image–text corpora. This will naturally raise the question of whether we can distinguish between what information is in an image caption and how that relates to existing verb classes. We draw on existing verb classifications to capture lexical and grammatical aspects for our empirical study. (Vendler, 1957; Levin, 1993; Baker et al., 1998; Schuler, 2005; Dowty, 1986; Comrie, 1976; Krifka, 1998).

3 Method

We study five prominent image–text corpora that vary in how constrained the relationship is between image and text:

- Microsoft Common Objects in Context (COCO) (Lin et al., 2014);
- Flickr30K (Flickr) (Young et al., 2014);
- Visual Storytelling (VIST) (Huang et al., 2016);
- Google's Conceptual Captions (CC) (Sharma et al., 2018); and
- the Recipe dataset (Yagcioglu et al., 2018).

COCO, Flickr and VIST are crowdsourced corpora, while CC and the Recipe dataset collect user-generated text. These corpora are designed to focus on the captioning relations exhibited in Figure 1. VIST asks for descriptive texts to link five images into a short narrative; CC pairs web images with relevant text from associated ALT-TEXT HTML attributes. These corpora may exhibit a broader range of inferential connections between image in text, such as the cases of play-by-play narrative in Figure 1. Finally, the Recipe dataset collects naturally-occurring text and images developed in combination, and includes a wide range of illustration relations (and a range of other strategies for achieving coherence across modalities which offer possibilities for future research).

To assess what's distinctive about these corpora, we compare them to two points of reference: the American National Corpus (ANC) which is a balanced corpus of spoken and written English (Leech et al., 2014) and Facebook's children's stories (FS) (Hill et al., 2015), a corpus of written narrative.

To measure different verb forms, we used part-of-speech tags, parses, and dependency labels, computed using the SpaCy natural language processing toolkit (Honnibal and Johnson, 2015), to find verbs and their associated auxiliaries. We then applied rules to classify the verb groups into past or non-past forms (including present, modal, and non-finite forms), and separately into simple (e.g., *ran*), progressive (e.g., *was running*) or perfect aspect (e.g., *has run*). Perfect progressive forms (*has been running*) are classed with perfect, since they share the focus on a result state not an ongoing activity. We keep a separate count for **copular** (copula) forms of the verb *be*—those that relate a subject to a predicate expressed as a noun phrase,

adjective phrase or prepositional phrase.

4 The Simplicity of Caption Corpora

We begin with the overall finding that motivates our research: Verb use in image–caption corpora is markedly rarer and less diverse than in ANC.

Verbs are less frequent overall in image–caption corpora. In ANC, 0.184 of the tokens have verb POS tags; that drops to 0.065 in CC, 0.026 in COCO, 0.017 in VIST and 0.012 in Flickr. (The difference seems wild, but remember captions won't have helper verbs for modals, passive, and negation, and may be bare noun phrases.) But the frequency of verbs also drops off faster in image–caption corpora, particularly across the most frequent 100 verbs. Table 1 shows how strongly the top 10 and top 30 lemmas dominate in image–caption corpora. By comparison, image–text data sets that allow for more varied links between images and text, particularly the Recipe dataset, show more diverse verb usage. This suggests that it's not just the connection between text and image that limits verb use, but the particular constraints of caption content.

Looking at the frequent verbs from Flickr and COCO gives a sense of the uniformity of captions. The 17 Frequent Caption Verbs listed in Table 2

is/are	wearing	sitting	standing
has/have	walking	holding	looking
playing	jumping	watching	smiling
talking	doing	eating	carrying
running	driving	laying	

Table 2: Verbs occurring at least 100 times per million words in COCO (Lin et al., 2014) or Flickr (Young et al., 2014), shown in their most frequent forms: *be* and *have* (simple present), plus 17 verbs we call the Frequent Caption Verbs (FCVs) (present participle).

make up 40.4% of verbs in COCO but only 6.30% of verbs in AN (not counting *be*, 23.3% of ANC and 23.0% of COCO; or *have*, 6.5% of ANC and 2.8% of COCO). Note how almost all the FCVs involve sustained activities associated with distinctive poses.

Not surprisingly, similar vocabulary is found in image captioning systems trained on these data sets. Table 3 tabulates the kinds of verbs produced across the COCO development set by eight successful image captioning models (Dai et al., 2017; Tavakoli et al., 2017; Liu et al., 2017; Mun et al., 2017). We can see that the outputs of these models also exhibit a preponderance of descriptions with FCVs and *be/have*.

models	FCVs	be/have	other
Dai et al., 2017	0.572	0.231	0.197
Liu et al., 2017	0.571	0.271	0.158
Mun et al., 2017	0.638	0.266	0.095
Tavakoli et al., 2017	0.609	0.231	0.160
Shetty et al., 2016	0.535	0.282	0.183
Shetty et al., 2017	0.609	0.231	0.160
Zhou et al., 2017	0.609	0.256	0.135
Wu et al., 2017	0.561	0.257	0.181

Table 3: Relative frequency of different kinds of verbs produced by eight captioning models trained on COCO.

5 Properties of Captions

Why are the verbs of captions so impoverished? The commonalities of the verbs in Table 2 suggests that it's because captions present specific kinds of information, in characteristic ways. We hypothesize that these constraints are associated with a **Caption** coherence relation that authors can use to link image and text into a coherent whole. In this section, we identify key semantic and pragmatic properties of this Caption relation.

Caption verbs show morphological commonalities: *ing*-forms predominate, which suggests that caption writers prefer progressive aspect, describing events as ongoing throughout some topic time—here, presumably, the moment of the photo. The progressive form combines with the auxiliary *be*: the predominance of *is* and *are* over *was* and *were* indicates that caption writers prefer present tense descriptions, construing the moment of the photo as "now" that anchors the speaker's perspective. Section 5.1 confirms that these are distinctive and characteristic features specifically cued by captioning tasks.

Caption verbs also show semantic commonalities. Not surprisingly, all involve visible events; Section 5.2 quantifies this preference. In addition, the verbs generally either are stative or describe unbounded activities without an inherent culmination or end-point; this is known in linguistics as atelic aktionsart (Vendler, 1957; Verkuyl, 2005). Section 5.3 reports an analysis confirming that captions prefer atelic descriptions over telic ones.

	progressive	perfect	simple	copula	past	non-past
COCO	0.493	0.121	0.193	0.187	0.140	0.850
Flickr	0.481	0.065	0.208	0.339	0.120	0.879
VIST	0.112	0.081	0.702	0.104	0.517	0.482
CC	0.207	0.161	0.528	0.103	0.139	0.860
Recipe	0.121	0.109	0.667	0.103	0.219	0.781
ANC	0.075	0.188	0.621	0.109	0.403	0.592
FS	0.076	0.126	0.647	0.137	0.606	0.382

Table 4: Grammatical tense and aspect across corpora. Progressive and non-past dominate in Flickr and COCO whereas the simple form dominates in Recipe, ANC and FS. The dataset from the image–text corpora that is the closest to ANC with respect to aspect is the Recipe dataset.

Overall then, we conclude that Caption texts offer present-tense descriptions anchored to the moment depicted in the related image and appeal to temporally unbounded eventualities to summarize the information explicitly visible in that image.

5.1 Captions prefer present progressive

We report the percentages of realization of tense and aspect on verbs that project full sentences across corpora in Table 4. Progressive verbs make 49% and 48% of COCO and Flickr respectively. The linguistic expressions in these captions mainly include reference to here and now, describing the situation in a progressive form. ANC on the other hand, includes only around 8% progressive verbs. For all the pairs, the distributions of tense and aspect are reliably different ($\chi^2 > 39.03, \quad p < 0.04$).

COCO and Flickr show a preponderance of progressive and non-past forms. The effect is even larger in the results of the models that are trained on COCO. As we can see in Table 5 progressive form makes up to 74% of the output of the models. Note that we know from Table 3 that these models have between 23% to 28% *be* and *have*.

models	non-past	progressive
Dai et al.,2017	0.994	0.550
Liu et al.,2017	0.995	0.709
Mun et al.,2017	0.998	0.691
Tavakoli et al.,2017	0.999	0.731
Shetty et al.,2016	0.998	0.728
Shetty et al.,2017	0.992	0.519
Zhou et al., 2017	0.998	0.739
Wu et al., 2017	0.998	0.678

Table 5: Relative frequency of non-past and progressive in verbs produced by eight captioning models trained on COCO.

CC shows a greatly increased use of simple forms in the present, while VIST shows simple forms in a mix of present and past. The instructions in VIST to tell a story, and the genre conventions of ALT-TEXT, lead to play-by-play descriptions in the narrative present (or sometimes for VIST, past) rather than the progressive descriptions provided by crowd-workers who just describe what they see.

Table 4 shows that VIST has a different distribution of tense and aspect in comparison to FS. Overall, FS includes 10% more past verbs. This involves more past perfect and simple past verbs where VIST includes more present progressive and simple present.

5.2 Captions prefer visible event verbs

Caption verbs also show semantic commonalities. Not surprisingly, they tend to involve visible events; that rules out a rich array of verbs that generally occur frequently.

To quantify this, we counted the occurrences of verbs in five Levin classes (Levin, 1993): desire verbs (e.g. need, want), verbs of psychological states (e.g. cheer, worry), declare verbs (e.g. believe, suppose), learn verbs (e.g. learn, memorize) and conceal verbs (e.g. screen, hide). The complete list can be found in the appendix. These verbs occur with a frequency of more than 20 per thousand words in ANC. They occur just 10.2, 15.7 and 16.6 times per million words in COCO, Flickr and VIST respectively. The differences are stark: even in telling a story, crowd workers confine themselves to the imagery, and stick to the visible facts. Other genres are less constrained; we find these verbs in CC and Recipe at a rate of 1080 and 1087 per million. Anecdotally, this reflects the additional relations that can link images and text

	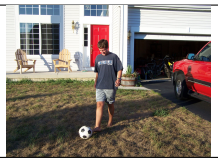			
A black frisbee is sitting on top of a roof.	A man playing soccer outside of a white house with a red door.	The boy is throwing a soccer ball by the red door.	A soccer ball is over a roof by a frisbee in a rain gutter.	Two balls and a frisbee are on top of a roof.
A discus got stuck up on the roof.	Why not try getting it down with a soccer ball?	Up the soccer ball goes.	It didn't work so we tried a volley ball.	Now the discus, soccer ball, and volleyball are all stuck on the roof.

Table 6: An example from VIST dataset that illustrates the difference between descriptive captions (middle row) and narrative (bottom row) and different uses of verbal tense and aspect in multimodal corpora. Photo credit: Ron Bieber

in these data sets. For example, ALT-TEXT fields often report first-person evaluations commenting on the imagery—prototypically, *I love it [what's shown]*, or *I want it [what's shown]*.

Do all visible verbs occur equally in image–text corpora? Of course not. Verbs differ in many different ways, most notably in their "image prior", how likely they are to happen during photo opportunities or to be featured and mentioned when images are published online. However, if someone says an event is common and interesting to watch and describe, but also says that it's rare to photograph it, you should be skeptical.

With that in mind, consider the verbs in Table 7. Truly invisible verbs, like *worry* and *wonder*, are not only missing from Flickr, COCO and VIST, but yield almost no hits on the web in the pattern *saw them V*. We also find frequent FCVs, like *walk* and *sit*, that occur widely across genres. The challenge are cases like *build* and *draw*. Google Ngram counts for *saw them build* and *saw them draw* confirm that they describe visible events with high frequency across text corpora, but these verbs are nevertheless rare in image–caption corpora. Maybe there's more to say here.

5.3 Captions prefer atelic descriptions

Our hypothesis is that the **lexical aspect** of verbs (Hamm and Bott, 2018) plays an important role in image captions. Lexical aspect describes the temporal structure of described eventualities. There are three main cases. **Stative** descriptions characterize ongoing conditions that do not involve dynamic activity, like *being* or *having*. **Atelic** ones characterize processes that can continue indefinitely, like *waiting* or *standing*. **Telic** ones characterize events that reach a definite endpoint and stop, like *arriving* or *winning*. What's relevant here is that a moment in time suffices to see that stative and atelic eventualities are under way. Telic descriptions can be established only by seeing the endpoint being realized, perhaps after an appropriate preparatory process.

Lexical aspect is partly due to the lexical meaning of the verb, but it also depends on whether relevant arguments are described in a delimited way or not—which gives rise to the linguistic problem of aspectual composition (Verkuyl, 2005). *Running* is an unbounded, atelic process. But *running the race* is a telic description: it ends when the race is run. And *running races* is again atelic: you can keep running new races indefinitely. The difference between telic and atelic descriptions thus has to be labeled by human annotators, based on the verb and its arguments.

If caption writers want to see the event they report, they should be reluctant to use telic descriptions. The image might not show the necessary culmination or the process leading up to it. However, this prediction depends on how speakers understand the progressive and narrative present forms. Semanticists often argue—on the basis of true examples like *In the '70s, Jodorowsky was making a film of "Dune" [but he never finished it]*—that a telic progressive description should be understood as a generic description of ongoing activities, **not** as a prediction of an eventual outcome. This is known as the imperfective paradox.

		worry	wonder	walk	sit	build	draw
corpus frequency	Flickr	0.1	0.4	524.6	675.0	10.4	10.5
	COCO	0.1	0.1	683.5	1991.5	3.2	2.1
	VIST	9.8	2.3	130.9	64.3	14.6	7.2
	CC	0	76.9	1745.6	1273.5	417.2	395.2
	ANC	143.6	196.1	264.4	269.1	323.6	167.5
Google Ngram	made them V	374	1975	2071	6121	919	1444
	saw them V	0	47	1586	412	193	713

Table 7: Corpus frequencies of select verbs (per million words) and counts from the Google Ngram dataset. The frequencies of *worry* and *wonder* are low in both image–text and the Google Ngram datasets. However, the frequencies of *build* and *draw*, while low in image–text corpora, are high in the Google Ngram dataset.

(Hamm and Bott, 2018). If this is captioners' understanding, they should use progressive telic descriptions freely, whenever they offer the best description of the activities visible in the image.

We (the authors) together with an undergraduate linguistics major at Rutgers drew 500 captions parsed as sentences from all of the datasets and derived a consensus annotation of whether those descriptions are stative, atelic, or telic. Verbs in telic and atelic classes are labeled as punctual or durative events (Moens, 1987; King, 1969). [1]

To calculate the effect size (a proxy for the difference of proportions of telic verbs across two data sets) that we are able to detect with 500 samples, we performed a sensitivity power analysis. The result of the analysis suggests that with a sample size of 500, we are able to detect effects sizes as small as 0.1650 with a power and significance level of 95% (Faul et al., 2014).

	durative	punctual
Flickr	22	7
COCO	23	5
VIST	79	33
CC	45	59
Recipe	189	110
ANC	197	97

Table 8: Counts of telic verbs out of 500 randomly selected sentences from each dataset. Pairwise comparisons of datasets suggest that every datasets is significantly different from others with the exception of two pairs; COCO and Flickr as well as Recipe and ANC. In general, the caption corpora contain fewer telic verbs in comparison to ANC and Recipe.

Table 8 presents the results of the annotation task. The results of t-test and f-test confirm that image–caption corpora emphasize atelic descriptions. For CC, noisy text meant our sample included only 412 relevant items, giving a telic rate of 0.252. In particular, an f-test shows that the distributions of telic verbs in these corpora are different ($f = 409.8$, $p = 1.1e - 644$). By t-test, Flickr is similar to COCO ($t = 0.12$, $p = 0.890$) and Recipe is similar to ANC ($t = -0.90$, $p = 0.366$), but all other datasets are two by two significantly different ($t>10$, $p < 0.0001$).

To calculate the inter-rater agreement, we determined Cohen's κ. We randomly selected 200 sentences from CC and assigned each to two annotators. The κ is 0.77, which indicates substantial agreement (Viera et al., 2005).

Our analysis depends on aspectual composition. In Flickr and COCO, FCVs contribute to atelic descriptions in 96% of occurrences whereas these verbs contribute to atelic descriptions only 39% of occurrences in ANC, because of different word senses and argument realizations. By contrast, verbs that contribute to telic descriptions in Flickr also contribute to telic descriptions in ANC in 98% of the cases. This underscores that the preference for atelic descriptions in image captions is a systematic phenomenon and not just an artifact of the small number of verbs found in the corpora.

6 Conclusions

By analyzing verb usage in image–caption corpora, we find that writers asked to caption an image take a particular perspective: they describe visible eventualities as present, continuing, and indefinite in temporal extent. These features help explain why verb use in captioning corpora is extremely limited—and these limitations persist in automatic captioning systems. We have offered a discourse perspective on these limitations, fol-

[1] The annotations are available at https://github.com/malihealikhani/Captions

lowing Hobbs (1990): a distinctive coherence relation governs the inferential and intentional relationships between images and caption text.

This is no slight to captions—they may well be challenging to model and useful to produce. However, this seems not to be the only kind of move that authors use to connect images and text. Broader corpora also feature play-by-play narrative, reactions and comments, illustrations, and perhaps other coherence relations between images and text. These relations deserve further study, but the preliminary evidence we have provided already suggests that these relations can accommodate a very different range of verbs than what's found in captions.

For now, the diversity of verb usage (and, perhaps, coherence relations) found in naturalistic image–text corpora like the Recipe dataset suggests some drawbacks for applying captioning models for novel applications. For example, consider using text as a cue for image retrieval: caption models might have good coverage for descriptions of extended activities that are clearly cued by people's pose, but they won't be very helpful for descriptions that characterize ongoing events in terms of their ultimate goal or outcome. This is not because those pictures are missing, because people aren't interested in seeing or describing those events, or because of the inherent limits of computer vision or semantic modeling techniques, but simply because the relevant descriptions happen to be missing from caption datasets, because of the conventions for writing coherent captions. We might well get better models by training on a broader range of data, including corpora where texts are accompanied by illustrations. Similarly, we can expect caption models to have limited utility in generating illustrated documents, as reported in one case by Ravi et al. (2018), because the vocabulary of events we might want to illustrate diverges so much from the vocabulary of captions.

We therefore recommend that future image–text corpora should explicitly look to explore and characterize the different ways text and imagery can be coherently related, including using the kinds of semantic and pragmatic analyses that we have presented here. A more inclusive collection effort should have the effect of increasing the diversity of event descriptions observed in image–text corpora, while laying the groundwork for more systematic coverage of applications. At the same time, our explorations have also revealed a clear need to improve theoretical and computational resources for verb classification to better characterize perceptual and temporal inference. So such efforts promise to refine theories of coherence and verb meaning in linguistics and cognitive science.

7 Acknowledgement

The research presented in this paper is supported by NSF Award IIS-1526723 and in part through a fellowship from the Rutgers Discovery Informatics Institute. Thanks to Divya Appasamy for helping us with the annotations. We are grateful for discussions and comments from Doug DeCarlo and Margaret Mitchell.

References

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

Raffaella Bernardi, Ruket Çakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2017. Automatic description generation from images: A survey of models, datasets, and evaluation measures (extended abstract). In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4970–4974.

Neil Cohn. 2013. Visual narrative structure. *Cognitive science*, 37(3):413–452.

Bernard Comrie. 1976. *Aspect: An introduction to the study of verbal aspect and related problems*, volume 2. Cambridge university press.

Samuel Cumming, Gabriel Greenberg, and Rory Kelly. 2017. Conventions of viewpoint coherence in film. *Philosophers' Imprint*, 17(1):1–29.

Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. 2017. Towards diverse and natural image descriptions via a conditional gan. In *The IEEE International Conference on Computer Vision (ICCV)*.

David R Dowty. 1986. The effects of aspectual class on the temporal structure of discourse: semantics or pragmatics? *Linguistics and philosophy*, 9(1):37–61.

F Faul, E Erdfelder, AG Lang, and A Buchner. 2014. G* power: statistical power analyses for windows and mac.

Francis Ferraro, Nasrin Mostafazadeh, Lucy Vanderwende, Jacob Devlin, Michel Galley, Margaret Mitchell, et al. 2015. A survey of current datasets for vision and language research. *arXiv preprint arXiv:1506.06833*.

Spandana Gella, Frank Keller, and Mirella Lapata. 2019. Disambiguating visual verbs. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):311–322.

Friedrich Hamm and Oliver Bott. 2018. Tense and aspect. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, fall 2018 edition. Metaphysics Research Lab, Stanford University.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

Jerry R Hobbs. 1990. *Literature and cognition*. 21. Center for the Study of Language (CSLI).

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239.

Harold V King. 1969. Punctual versus durative as covert categories 1. *Language Learning*, 19(3-4):185–190.

Manfred Krifka. 1998. The origins of telicity. In *Events and grammar*, pages 197–235. Springer.

Geoffrey Leech, Paul Rayson, et al. 2014. *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Chang Liu, Fuchun Sun, Changhu Wang, Feng Wang, and Alan Yuille. 2017. Mat: A multimodal attentive translator for image captioning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4033–4039.

Scott McCloud. 1993. *Understanding comics: The invisible art*. William Morrow.

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2018a. Talking about other people: an endless range of possibilities. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 415–420, Tilburg University, The Netherlands. Association for Computational Linguistics.

Emiel van Miltenburg, Ruud Koolen, and Emiel Krahmer. 2018b. Varying image description tasks: spoken versus written descriptions. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 88–100. Association for Computational Linguistics.

Marc Moens. 1987. Tense, aspect and temporal reference.

Jonghwan Mun, Minsu Cho, and Bohyung Han. 2017. Text-guided attention model for image captioning. In *AAAI Conference on Artificial Intelligence*.

Geoffrey K. Pullum, Rodney Huddleston, L. Bauer, B. Birner, T. Briscoe, P. Collins, D. Denison, D. Lee, A. Mittwoch, G. Nunberg, F. Palmer, J. Payne, P. Peterson, L. Stirling, and Gregory Ward. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press.

Hareesh Ravi, Lezi Wang, Carlos Muniz, Leonid Sigal, Dimitris Metaxas, and Mubbasir Kapadia. 2018. Show me a story: Towards coherent neural story illustration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7613–7621.

Karin Kipper Schuler. 2005. Verbnet: A broad-coverage, comprehensive verb lexicon.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2556–2565.

Hamed R Tavakoli, Rakshith Shetty, Ali Borji, and Jorma Laaksonen. 2017. Paying attention to descriptions generated by image captioning models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2487–2496.

Zeno Vendler. 1957. Verbs and times. *The philosophical review*, 66(2):143–160.

Henk J Verkuyl. 2005. Aspectual composition: Surveying the ingredients. In *Perspectives on aspect*, pages 19–39. Springer.

Anthony J Viera, Joanne M Garrett, et al. 2005. Understanding interobserver agreement: the kappa statistic. *Fam Med*, 37(5):360–363.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes. *arXiv preprint arXiv:1809.00812*.

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.