Visual Storytelling

Ting-Hao (Kenneth) Huang¹*, Francis Ferraro²*, Nasrin Mostafazadeh³, Ishan Misra¹, Aishwarya Agrawal⁴, Jacob Devlin⁶, Ross Girshick⁵, Xiaodong He⁶, Pushmeet Kohli⁶, Dhruv Batra⁴, C. Lawrence Zitnick⁵, Devi Parikh⁴, Lucy Vanderwende⁶, Michel Galley⁶, Margaret Mitchell⁶

Microsoft Research
1 Carnegie Mellon University, 2 Johns Hopkins University, 3 University of Rochester, 4 Virginia Tech, 5 Facebook AI Research, 6 Corresponding authors: {jdevlin,lucyv,mgalley,memitc}@microsoft.com

Abstract

We introduce the first dataset for sequential vision-to-language, and explore how this data may be used for the task of visual storytelling. The dataset includes 81,743 unique photos in 20,211 sequences, aligned to descriptive and story language. We establish several strong baselines for the storytelling task, and motivate an automatic metric to benchmark progress. We argue that modelling figurative and social language, as provided for in this data and the storytelling task, has the potential to move artificial intelligence towards more human-like expression and understanding.

1 Introduction

Beyond understanding simple objects and concrete scenes lies interpreting causal structure; making sense of visual input to tie disparate moments together as they give rise to a cohesive narrative of events through time. This requires moving from reasoning about single images – static moments, devoid of context – to sequences of images that depict events as they occur and change. On the vision side, progressing from single images to images in context allows us to begin to create an artificial intelligence (AI) that can reason about a visual moment given what it has already seen. On the language side, progressing from literal description to narrative helps to learn more evaluative, conversational, and abstract language. This is the difference between, for example, “sitting next to each other” versus “having a good time”, or “sticking legs in the air” versus “falling over drunk” (see Figure 1). The first descriptions capture image content that is literal and concrete; the second requires further inference about what a good time may look like, and that when drinking one may then be more likely to fall over.

We introduce the first dataset of sequential images with corresponding descriptions, which captures some of these subtle but important differences, and advance the task of visual storytelling. We release the data in three tiers of language for the same images: (1) Descriptions of images-in-isolation (DII); (2) Descriptions of images-in-sequence (DIS); and (3) Stories for images-in-sequence (SIS). This tiered approach reveals the effect of temporal context and the effect of narrative language. As all the tiers are aligned to the same images, the dataset facilitates directly modeling the relationship between literal and more abstract visual concepts. We additionally propose an automatic evaluation metric which is best correlated with human judgments, and establish several strong baselines for the visual storytelling task.

---

* T.H. and F.F. contributed equally to this work.
beach (684) breaking up (350) Easter (259)
amusement park (525) carnival (331) church (243)
building a house (415) visit (321) graduation ceremony (236)
party (411) market (311) office (226)
birthday (399) outdoor activity (267) father’s day (221)

Table 1: The number of albums in our tiered dataset for the 15 most frequent kinds of stories.

2 Motivation and Related Work

Work in vision to language has exploded, with researchers examining image captioning (Lin et al., 2014; Karpathy and Fei-Fei, 2015; Vinyals et al., 2015; Xu et al., 2015; Chen et al., 2015; Young et al., 2014; Elliott and Keller, 2013), question answering (Antol et al., 2015; Ren et al., 2015; Gao et al., 2015; Malinowski and Fritz, 2014), visual phrases (Sadeghi and Farhadi, 2011), video understanding (Ramanathan et al., 2013), and visual concepts (Krishna et al., 2016; Fang et al., 2015).

Such work focuses on direct, literal description of image content. While this is an encouraging first step in connecting vision and language, it is far from the capabilities needed by intelligent agents for naturalistic interactions. There is a significant difference, yet unexplored, between remarking that a visual scene shows “sitting in a room” – typical of most image captioning work – and that the same visual scene shows “bonding.” The latter description is grounded in the visual signal, yet it brings to bear information about social relations and emotions that can be additionally inferred in context (Figure 1). Visually-grounded stories facilitate more evaluative and figurative language than has previously been seen in vision-to-language research: If a system can recognize that colleagues look bored, it can remark and act on this information directly.

Storytelling itself is one of the oldest known human activities (Wiessner, 2014), providing a way to educate, preserve culture, instill morals, and share advice; focusing AI research towards this task therefore has the potential to bring about more human-like intelligence and understanding.

3 Dataset Construction

Extracting Photos We begin by generating a list of “storyable” event types. We leverage the idea that “storyable” events tend to involve some form of possession, e.g., “John’s birthday party,” or “Shabnam’s visit.” Using the Flickr data release (Thomee et al., 2015), we aggregate 5-grams of photo titles and descriptions, using Stanford CoreNLP (Manning et al., 2014) to extract possessive dependency patterns. We keep the heads of possessive phrases if they can be classified as an EVENT in WordNet3.0, relying on manual winnowing to target our collection efforts.1 These terms are then used to collect albums using the Flickr API.2 We only include albums with 10 to 50 photos where all album photos are taken within a 48-hour span and CC-licensed. See Table 1 for the query terms with the most albums returned.

The photos returned from this stage are then presented to crowd workers using Amazon’s Mechanical Turk to collect the corresponding stories and descriptions. The crowdsourcing workflow of developing the complete dataset is shown in Figure 2.

Crowdsourcing Stories In Sequence We develop a 2-stage crowdsourcing workflow to collect naturalistic stories with text aligned to images. The first stage is storytelling, where the crowd worker selects a subset of photos from a given album to form a photo sequence and writes a story about it (see Figure 3). The second stage is re-telling, in which the worker writes a story based on one photo sequence generated by workers in the first stage.

In both stages, all album photos are displayed in the order of the time that the photos were taken.

1 We simultaneously supplemented this data-driven effort by a small hand-constructed gazetteer.
2 https://www.flickr.com/services/api/
Crowdsourcing Descriptions of Images In Isolation & Images In Sequence We also use crowdsourcing to collect descriptions of images-in-isolation (DII) and descriptions of images-in-sequence (DIS), for the photo sequences with stories from a majority of workers in the first task (as Figure 2). In both DII and DIS tasks, workers are asked to follow the instructions for image captioning proposed in MS COCO (Lin et al., 2014) such as describe all the important parts. In DII, we use the MS COCO image captioning interface. In DIS, we use the storyboard and story cards of our storytelling interface to display a photo sequence, with a “storyboard” underneath. In storytelling, by clicking a photo in the album, a “story card” of the photo appears on the storyboard. The worker is instructed to pick at least five photos, arrange the order of selected photos, and then write a sentence or a phrase on each card to form a story; this appears as a full story underneath the text aligned to each image. Additionally, this interface captures the alignments between text and photos. Workers may skip an album if it does not seem storyable (e.g., a collection of coins). Albums skipped by two workers are discarded. The interface of re-telling is similar, but it displays the two photo sequences already created in the first stage, which the worker chooses from to write the story. For each album, 2 workers perform storytelling (at $0.3/HIT), and 3 workers perform re-telling (at $0.25/HIT), yielding a total of 1,907 workers. All HITs use quality controls to ensure varied text at least 15 words long.

Data Analysis

Our dataset includes 10,117 Flickr albums with 210,819 unique photos. Each album on average has 20.8 photos ($\sigma = 9.0$). The average time span of each album is 7.9 hours ($\sigma = 11.4$). Further details of each tier of the dataset are shown in Table 2.\footnote{https://github.com/tylin/coco-ui}

<table>
<thead>
<tr>
<th>Data Set</th>
<th>#(Txt, Img)</th>
<th>Vocab Size (k)</th>
<th>Avg. #Tok</th>
<th>Abs Frazier Yonge Ppl</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown</td>
<td>52.1</td>
<td>47.7</td>
<td>20.8</td>
<td>15.2%</td>
</tr>
<tr>
<td>DII</td>
<td>151.8</td>
<td>13.8</td>
<td>11.0</td>
<td>21.3%</td>
</tr>
<tr>
<td>DIS\footnote{The DIS columns VocabSize-Ppl estimated based on 17,425 Txt,Img pairs. Full set will be updated shortly.}</td>
<td>151.8</td>
<td>5.0</td>
<td>9.8</td>
<td>24.8%</td>
</tr>
<tr>
<td>SIS</td>
<td>252.9</td>
<td>18.2</td>
<td>10.2</td>
<td>22.1%</td>
</tr>
</tbody>
</table>

Table 2: A summary of our dataset,\footnote{We use those names occurring at least 10,000 times. \url{https://ssa.gov/oact/babynames/names.zip}} following the proposed analyses of Ferraro et al. (2015), including the Frazier and Yonge measures of syntactic complexity. The balanced Brown corpus (Marcus et al., 1999), provided for comparison, contains only text. Perplexity (Ppl) is calculated against a 5-gram language model learned on a generic 30B English words dataset scraped from the web.

MS COCO instructions adapted for sequences. We recruit 3 workers for DII (at $0.05/HIT) and 3 workers for DIS (at $0.07/HIT).

Data Post-processing We tokenize all storylets and descriptions with the CoreNLP tokenizer, and replace all people names with generic MALE/FEMALE tokens,\footnote{We use those names occurring at least 10,000 times. \url{https://ssa.gov/oact/babynames/names.zip}} and all identified named entities with their entity type (e.g., location). The data is released as training, validation, and test following an 80%/10%/10% split on the stories-in-sequence albums. Example language from each tier is shown in Figure 4.
Table 4: Correlations of automatic scores against human judgements, with p-values in parentheses.

<table>
<thead>
<tr>
<th></th>
<th>METEOR</th>
<th>BLEU</th>
<th>Skip-Thoughts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(r)</td>
<td>(\rho)</td>
<td>(\tau)</td>
</tr>
<tr>
<td></td>
<td>0.22 (2.8e-28)</td>
<td>0.20 (3.0e-31)</td>
<td>0.14 (1.0e-33)</td>
</tr>
<tr>
<td></td>
<td>0.08 (1.0e-06)</td>
<td>0.08 (8.9e-06)</td>
<td>0.06 (8.7e-08)</td>
</tr>
<tr>
<td></td>
<td>0.18 (5.0e-27)</td>
<td>0.16 (6.4e-22)</td>
<td>0.11 (7.7e-24)</td>
</tr>
</tbody>
</table>

Table 5: Baselines with associated METEOR scores.

<table>
<thead>
<tr>
<th>Viterbi</th>
<th>Greedy</th>
<th>-Dups</th>
<th>+Grounded</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.23</td>
<td>0.28</td>
<td>0.30</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Figure 5: Example story generated by +Grounded.

The family got together for a cookout. They had a lot of delicious food. The dog was happy to be there. They had a great time on the beach. They even had a swim in the water.

Table 5: Baselines with associated METEOR scores.

7 Scale presented ranged from “Strongly disagree” to “Strongly agree”, which we convert to a scale of 1 to 5.

8 We use METEOR version 1.5 with hter weights.

The detailed study of automatic evaluation of stories is an area of interest for a future work.

6 Baseline Experiments

We report baseline experiments on the storytelling task in Table 5, training on the SIS tier and testing on half the validation set. We use a sequence-to-sequence recurrent neural net (RNN) approach, which naturally extends the single-image captioning technique of Devlin et al. (2015) and Vinyals et al. (2014) to multiple images. Here, we encode an image sequence by running an RNN over the \( fc7 \) vectors of each image, in reverse order. We use Gated Recurrent Units (GRUs) (Cho et al., 2014) for both the image encoder and story decoder, and compare Viterbi decoding with a Greedy approach taking the most likely word at each step.

The baseline models learn to prefer generic language, which follows from maximum likelihood training. Both the Viterbi and Greedy baselines have no mechanisms to distinguish between visual and non-visual language, which further causes these models to ‘hallucinate’ visual objects.

A simple way to handle these issues is to (1) add constraints for lexical variation and length; and (2) force grounding. For (1), we prohibit repeating a content word more than twice, and require all stories to be at least 40 words long. We refer to this method as -Dups. For (2), we train a model on the DII tier; high frequency caption words are only permitted in the story if also generated by the DII model. We refer to this as +Grounded. An example of this top-performing baseline is shown in Figure 5.

7 Conclusion and Future Work

We have introduced the first dataset for sequential vision-to-language, which incrementally moves
from images-in-isolation to stories-in-sequence. We argue that modelling the more figurative and social language captured in this dataset is essential for evolving AI towards more human-like understanding. We have established several strong baselines for the task of visual storytelling, and have motivated METEOR as an automatic metric to evaluate progress on this task moving forward.

References


Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.


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 Meet-
ing of the Association for Computational Linguistics: System Demonstrations, pages 55–60.


