

# Dixit: Interactive Visual Storytelling via Term Manipulation

Chao-Chun Hsu\*

Academia Sinica

Taipei, Taiwan

joe32140@iis.sinica.edu.tw

Yu-Hua Chen\*

Academia Sinica

Taipei, Taiwan

cloud60138@iis.sinica.edu.tw

Zi-Yuan Chen\*

Academia Sinica

Taipei, Taiwan

zychen@iis.sinica.edu.tw

Hsin-Yu Lin

Academia Sinica

Taipei, Taiwan

swat230@iis.sinica.edu.tw

Ting-Hao (Kenneth) Huang

Pennsylvania State University

State College, PA, USA

txh710@psu.edu

Lun-Wei Ku

Academia Sinica

Most Joint Research Center for AI  
Technology and All Vista Healthcare

Taipei, Taiwan

lwku@iis.sinica.edu.tw

## ABSTRACT

In this paper, we introduce **Dixit**, an interactive visual storytelling system that the user interacts with iteratively to compose a short story for a photo sequence. The user initiates the process by uploading a sequence of photos. Dixit first extracts text *terms* from each photo which describe the objects (e.g., boy, bike) or actions (e.g., sleep) in the photo, and then allows the user to add new terms or remove existing terms. Dixit then generates a short story based on these terms. Behind the scenes, Dixit uses an LSTM-based model trained on image caption data and FrameNet to distill terms from each image, and utilizes a transformer decoder to compose a context-coherent story. Users change images or terms iteratively with Dixit to create the most ideal story. Dixit also allows users to manually edit and rate stories. The proposed procedure opens up possibilities for *interpretable* and *controllable* visual storytelling, allowing users to understand the story formation rationale and to intervene in the generation process.

## CCS CONCEPTS

• **Computing methodologies** → **Natural language generation**; Scene understanding; • **Human-centered computing** → **Interactive systems and tools**.

## KEYWORDS

Interactive System, Visual Storytelling, Web Application

### ACM Reference Format:

Chao-Chun Hsu\*, Yu-Hua Chen\*, Zi-Yuan Chen\*, Hsin-Yu Lin, Ting-Hao (Kenneth) Huang, and Lun-Wei Ku. 2019. Dixit: Interactive Visual Storytelling via Term Manipulation. In *Proceedings of the 2019 World Wide Web Conference (WWW '19)*, May 13–17, 2019, San Francisco, CA, USA. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3308558.3314131>

\* denotes equal contribution

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW '19, May 13–17, 2019, San Francisco, CA, USA

© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.

ACM ISBN 978-1-4503-6674-8/19/05.

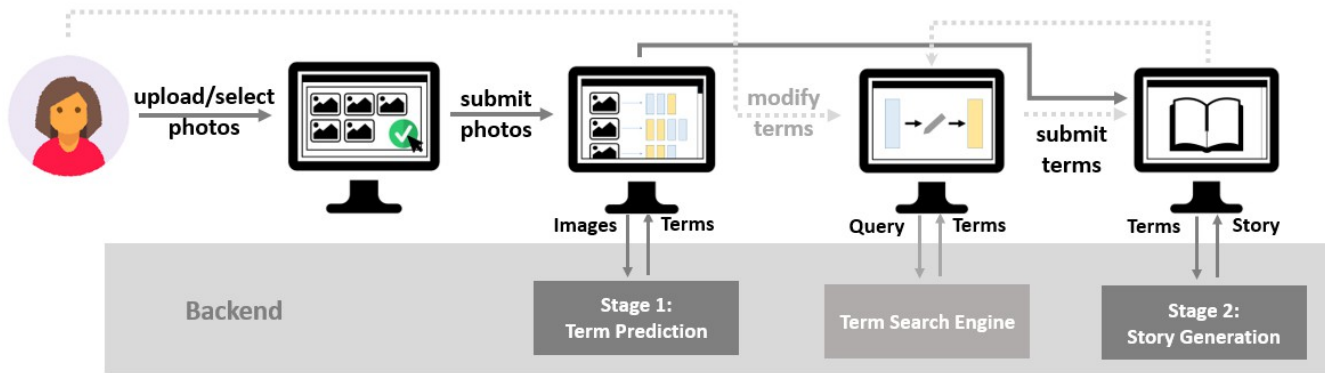
<https://doi.org/10.1145/3308558.3314131>

## 1 INTRODUCTION

In visual storytelling, as introduced by Huang et al. [5], we take a sequence of photos as input and attempt to generate a short story narrating the sequence. This is an interdisciplinary task comprising image semantic reasoning and structural text generation. In contrast to image captioning or video captioning, which have less annotation variation across annotators, authors of visual stories can set down totally different stories for a given image sequence based on their preference and imagination. Instead of generating one story and leaving it *to* the user, we believe that working on a story draft together and iterating *with* the user results in much better story quality. In this paper, we introduce **Dixit**<sup>1</sup>, an *interactive* visual storytelling system. The purpose of Dixit is to enable users to control the story generation process and create their preferred stories. This system is inspired by our observation that people often start their short visual story with the key objects or actions in the pictures, and then build connections between these terms to form the final story. However, recent studies using deep neural networks (DNNs) are generally designed in an end-to-end fashion without intermediate human-recognizable representations [5, 11]. This closed process prevents users from adjusting intermediate variables and thus makes it impossible to create diverse stories from a given image sequence. In addition, as story generation is a subjective process depending on one's imagination, there is no ground truth given identical inputs, and automatic metrics fail to evaluate the quality of the story with an absolute score. In sum, we have the following four main issues:

First, **lack of interpretability in intermediate DNN Layers**. Because of the end-to-end DNN model's powerful ability as a universal approximator, in recent years it has achieved significant improvements in many fields, including computer vision and natural language processing. However, though some studies have attempted to get inside the DNN black box, it is difficult to clearly explain what information propagates to the next layer in a DNN model. For example, in visual storytelling, how exactly an end-to-end framework selects entities from images remains a mystery and the model's behavior is difficult to interpret. Second, **inability to modify generated stories**. With an end-to-end model, another problem is that the same input always yields the same result. That is,

<sup>1</sup>Dixit is available at <https://dixitdemo.github.io/index.html>. The web page will be responsive and can be used via mobile devices upon demo.



**Figure 1: System architecture.** The user initiates the process by uploading a sequence of photos. Dixit first extracts text *terms* from each photo which describe the objects or actions in the photo, and then allows the user to add new terms or remove existing terms. Dixit then generates a short story based on these terms.

we can obtain only a single story given the same image sequence – but this goes against the creative nature of storytelling. Although some generative models such as the variational autoencoder and the generative adversarial network produce more results by sampling from a prior distribution, they cannot generate storylines in the desired direction. Third, **difficulty in automatic evaluation**. Automatic metrics such as BLEU and METEOR have been shown to be insufficient to evaluate the quality of a visual story generation system [11]. Due to its subjective characteristics, human evaluation, which takes money and time, has become an essential part of the storytelling task. Finally, **data scarcity**. The quantity of image-story pair data is relatively insufficient because visual storytelling is an complex task; thus data annotation is expensive and time consuming.

To address these four issues, we introduce a two-stage interactive visual story generation system with which (i) users can understand what model sees in the images as predicted terms represent the objects and actions in the images, and also observe how these terms influence story generation, (ii) users can modify terms in the first stage to control the story generation, (iii) we can obtain human evaluation by asking users to score the quality of the generated story and the whole user-system interaction record in a turn is saved for future usage, and (iv) the modular nature of the two-stage framework, which utilizes image caption data and story corpora separately, allows us to train the visual storytelling model without image-story pair data.

## 2 RELATED WORK



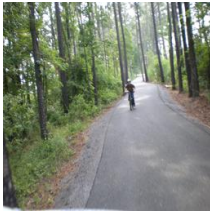
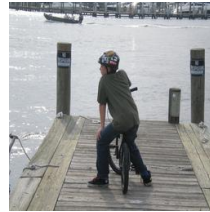
Whereas end-to-end text generation from visual contents has been widely explored for tasks such as image caption, video caption, and visual storytelling, recent work has also been done on two-stage frameworks. Alexander et al. propose two-stage style transfer for image captioning. They first extract objects and verbs from an image and generate a stylish caption using an RNN trained on story corpora [7]. However, each such caption is only a single story-like sentence and is independent of other captions; put together, the captions do not constitute a context-coherent story. Xu et al. propose

a skeleton-based narrative story generation method which in the first stage extracts the skeleton of the input sentence after which they generate the next sentence according to the given skeleton, one-by-one [12]. Although this framework shares with our system the insight that key phrases are important for generating a coherent story, its skeleton is extracted from pure text and does not serve as the grounding for the story. Moreover, as the framework generates the skeleton after generating the sentence, it lacks the ability to consider all important terms together to create the best fit.

## 3 SYSTEM

**Workflow.** Users of our system either upload images from their devices or choose from an image pool. After five images are uploaded or chosen, the first stage uses a term prediction model that captures terms – i.e., the objects and actions in images – to help users understand what the model sees in the current photos. As explained in Section 4, the actions are represented by verb frames. In the second stage, the predicted terms are then concatenated and fed into a story generator to construct a story with appropriate and coherent context. Note that the terms can be modified arbitrarily to control the story generation. If users want a story that depicts objects other than those detected from the images, they can remove the terms and add desired terms; to this end the system also includes a term search function. After the story is generated, the users are asked to rate the story; the complete interaction record (including image selection and term modification) is saved in the database. We expect the routine of modifying terms and generating a story to be repeated many times, as the user controls the story generation process and gradually gets better at creating good stories. For instance, in Table 1, the generated story in the second row is not perfect because some predicted terms are, although reasonable, not preferred. However, after removing these terms and adding better ones, users create a better story. This auto-labeling is an efficient way to provide samples for the training of future story models.

**User Interface.** A screenshot of the user interface is shown in Figure 2. The user’s work flow starts at the left-top block, where the

					
<b>Predicted Terms</b>	"man", "Placing", "bike"	"motorcycle", "rider", "Emptying", "bike"	"trees", "sign", "Preventing_or_letting",	"man", "Placing", "bench"	"boy", "seat", "Placing"
<b>Story</b>	the man was sitting on his bike.	he was riding his bike with a rider on his motorcycle.	he stopped at a stop sign from a trees.	the man sat on the bench.	the boy sat in his seat.
<b>Modified Terms</b>	"boy", "Placing", "bike"	"rider", "bike"	"Seeking", "trees", "forest"	"bike", "dock"	"Sleep", "seat"
<b>New Story</b>	the boy sat down at his bike.	he was riding his bike like a rider.	he went into the forest looking for trees.	he threw his bike at the dock.	he went back to his seat to sleep.

**Table 1: Via term adjustment, story generation is focused on specific objects or actions**

user is asked to upload their own images or choose from the image pool. After the users submit images to the system, the predicted terms are shown in the left-bottom block. We meanwhile provide a term searching area (right bottom) using which the user searches for and modifies predicted terms. Term search is used for nouns and verb frames. For nouns, the search function retrieves nouns containing the query sub-string. For instance, for the query *wood*, the search engine returns *wood*, *woods*, *firewood*, *woodworker*, *wood-working*, and *plywood*. While searching for verb frames, instead of directly querying by the frame name, we search for lexical units of the verb frame. Results for the query *watch* would include *seeking*, of which *watch* is a lexical unit of *seeking*. After clicking the magic wand icon in the left-bottom area, the selected terms are sent to our story generation module, after which the generated story is retrieved and shown in the left middle area. The user is then asked to rate the story via the star icons. Since the concept of terms can be unfamiliar to users, we provide a term description area in the right-top corner. When user hovers the pointer over a term, the description of that term is displayed in the term description area. For noun terms, the descriptions are related images illustrating that

noun, whereas for verb terms, the descriptions are textual explanations of the verbs. In the middle, on the right, a storage area is provided for the user to temporarily store unused terms.

## 4 DATA PREPARATION

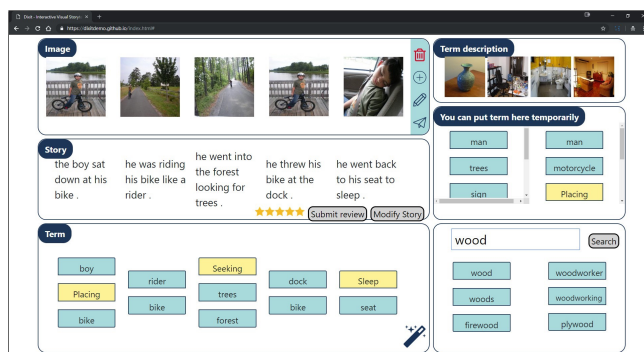
Due to the modular nature of the proposed system, we train the model in two stages<sup>4</sup> separately, without parallel data.

To train the term prediction model, we extract terms from captions in the COCO dataset [6] as the gold labels for the first stage to identity terms in images. The selection of terms is inspired by Semstyle [7]. Each sentence is represented as a combination of several noun (objects) and verb (actions) terms, which preserves the most crucial information needed to generate stories. Given a sentence, we extract verb terms by using Open-SESAME [9] to predict verb frames and leave nouns as noun terms. Frames are defined in FrameNet [1], a lexical database of semantic frame. The basic idea of FrameNet is straightforward: each frame is a description of a type of event, relation, or entity, and the participants in it.

To train the story generator, we use ROCStories Corpora [8], which contains stories with five sentences. For each story, we use a technique similar to back-translation to build our term-to-story corpora. That is, we transform stories to their corresponding terms, and then use these term-story pairs to train the story generator.

## 5 IMAGE-TO-TERM PREDICTION

As shown in Figure 3, the image-to-term prediction process can be considered a simplified image captioning problem which instead predicts the terms of an image. Given an image, we first convert it to the RGB format and resize it to 3x224x224 (channel \* width \* height) and then utilize a pre-trained Resnet-152 [3] model to extract image features. Resnet is a very deep neural network model with residual links between layers, originally pre-trained for image classification. The literature has suggested that it captures image features from low-level components like arcs to high-level concepts. With the image features produced in the previous step, we feed image features into an LSTM [4] to generate terms. Note that as



**Figure 2: System user interface**

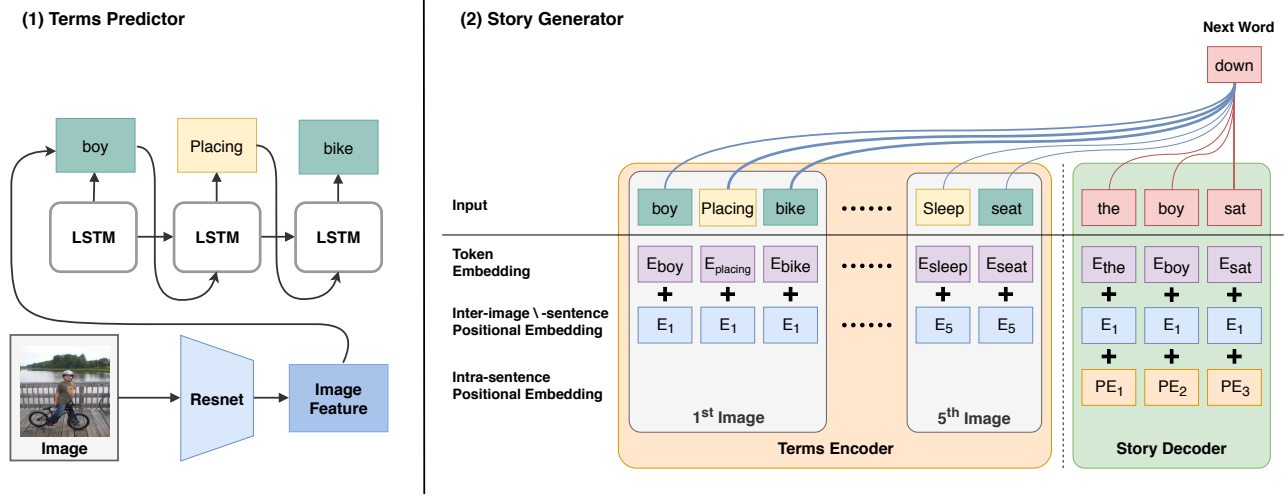


Figure 3: Architecture of (1) term prediction model and (2) story generation model. Predicted terms in green boxes denote noun terms, and terms in yellow boxes denote verb frames. In the training process, which uses the a single vocabulary, we treat both noun terms and verb frames equally. Term box colors are used here to emphasize their different origins.

described in Section 4, the LSTM model is trained with images and terms extracted from their corresponding captions in the COCO dataset; this pipeline is similar to that depicted in [7].

## 6 TERM-TO-STORY GENERATION

Transformer [10] is used as the basic model for story generation. In the Transformer paper, Vaswani et al. propose positional embedding as sequential information to replace the recurrent neural network. Here, we extend Transformer to be sentence-aware with the addition of an inter-sentence positional embedding to indicate the order of sentences within a story. This is shown in Figure 3. We do not consider the order of the input terms because terms represent only what appears in the images and need no sequential signal. At decoding time, an input embedding is the summation of the token embedding, the inter-image/-sentence positional embedding (InterPE), and the intra-sentence positional embedding (IntraPE). In particular,

- The InterPE is randomly initialized and updated during the training procedure. Similar to the segment embedding in BERT [2], we differentiate sentences/images by adding  $E_s$  to token embeddings of the  $s$ -th sentence/image;  $s$  denotes the order of a sentence/image in a story.  $s \in \{1, 2, 3, 4, 5\}$ .
- For IntraPE, we follow the Transformer implementation:

$$\text{IntraPE}_{(pos, 2i)} = \sin(pos/10000^{2i/d_{model}})$$

$$\text{IntraPE}_{(pos, 2i+1)} = \cos(pos/10000^{2i/d_{model}})$$

where  $pos$  is the position and  $i$  is the dimension.  $d_{model}$  denotes the dimension of the input and output token. The IntraPE parameters are fixed, and the sinusoidal representation allows the model to extrapolate the sentence length that is longer than the training instances. The IntraPE  $PE_s$  is added

to the  $s$ -th token embedding in a sentence.  $s \in \{1, 2, \dots, n\}$  and  $n$  is the max sentence length.

Based on the terms predicted for each image in the first step, we concatenate terms predicted for the five images as the input to the story generator. While decoding the story at every time step, the sentence-aware Transformer generates the context-coherent story conditioned on all input terms and the prior words of the story.

## 7 CONCLUSION AND FUTURE WORK

In this paper, we propose a two-stage interactive visual storytelling system that extracts terms (nouns and verb frames) and generates stories based on these terms. Users interactively adjust the terms to produce the desired story, and then record and rate the story quality for later model improvement. By splitting the task into two steps and their middle semantic term layer, the training data are not limited to the pair of images and their story, but include more accessible images with captions, and the text stories themselves. In addition, the process of visual story generation becomes interpretable. We believe this is a worthy direction for the advance of visual storytelling.

## ACKNOWLEDGEMENT

This research is partially supported by Ministry of Science and Technology, Taiwan, under Grant no. MOST108-2634-F-001-004-, MOST108-2634-F-002-008- and MOST107-2218-E-002-009-.

## REFERENCES

- [1] 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*. Association for Computational Linguistics, 86–90.
- [2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

- [3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [4] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [5] 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*. 1233–1239.
- [6] 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*. Springer, 740–755.
- [7] Alexander Mathews, Lexing Xie, and Xuming He. 2018. Semstyle: Learning to generate stylised image captions using unaligned text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8591–8600.
- [8] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. In *Proceedings of NAACL-HLT*. 839–849.
- [9] Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A Smith. 2017. Frame-semantic parsing with softmax-margin segmental rnns and a syntactic scaffold. *arXiv preprint arXiv:1706.09528* (2017).
- [10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.
- [11] Xin Wang, Wenhui Chen, Yuan-Fang Wang, and William Yang Wang. 2018. No Metrics Are Perfect: Adversarial Reward Learning for Visual Storytelling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 899–909.
- [12] Jingjing Xu, Xuancheng Ren, Yi Zhang, Qi Zeng, Xiaoyan Cai, and Xu Sun. 2018. A Skeleton-Based Model for Promoting Coherence Among Sentences in Narrative Story Generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 4306–4315.