Knowledge-Enriched Visual Storytelling

Chao-Chun Hsu[1*] Zi-Yuan Chen[2*] Chi-Yang Hsu[3]
Chih-Chia Li[4] Tzu-Yuan Lin[5] Ting-Hao (Kenneth) Huang[3] Lun-Wei Ku[2,6]
[1]University of Colorado Boulder, [2]Academia Sinica, [3]Pennsylvania State University,
[4]National Chiao Tung University, [5]National Taiwan University,
[6]Most Joint Research Center for AI Technology and All Vista Healthcare
chao-chun.hsu@colorado.edu, {zychen, lwku}@iis.sinica.edu.tw, {cxh5437, txh710}@psu.edu

Abstract

Stories are diverse and highly personalized, resulting in a large possible output space for story generation. Existing end-to-end approaches produce monotonous stories because they are limited to the vocabulary and knowledge in a single training dataset. This paper introduces **KG-Story**, a three-stage framework that allows the story generation model to take advantage of external **K**nowledge **G**raphs to produce interesting stories. KG-Story distills a set of representative words from the input prompts, enriches the word set by using external knowledge graphs, and finally generates stories based on the enriched word set. This *distill-enrich-generate* framework allows the use of external resources not only for the enrichment phase, but also for the distillation and generation phases. In this paper, we show the superiority of KG-Story for visual storytelling, where the input prompt is a sequence of five photos and the output is a short story. Per the human ranking evaluation, stories generated by KG-Story are on average ranked better than that of the state-of-the-art systems. Our code and output stories are available at https://github.com/zychen423/KE-VIST.

Introduction

Stories are diverse and highly personalized. Namely, story generation models, whether using text or images as input prompts, aim at a large possible output space with rich information and vocabulary. One classic example is visual storytelling (VIST), an interdisciplinary task that takes a sequence of five photos as the input and generates a short story as the output (Huang et al. 2016). However, existing visual storytelling approaches produce monotonous stories with repetitive text and low lexical diversity (Hsu et al. 2019). We believe three major factors contribute to this problem. First, most prior work uses only a single training set (*i.e.*, the VIST dataset) in an end-to-end manner (Wang et al. 2018b; Kim et al. 2018; Wang et al. 2018a). Although this can result in legitimate stories, the end-to-end architecture makes it difficult to use external data. The generated stories are thus limited in terms of vocabulary and knowledge in VIST. Similar phenomena are observed for text story generation models using a single training set, where the output can be "fairly generic" (Fan, Lewis, and Dauphin 2018). Second, the VIST dataset is considerably smaller than that of other text generation tasks. The MSCOCO dataset for image captioning contains 995,684 captions (Lin et al. 2014); the VQA dataset for visual question answering contains ∼0.76M questions and ∼10M answers (Antol et al. 2015); and the ROC dataset for text story generation contains 98,159 stories (Mostafazadeh et al. 2016). The VIST dataset, in contrast, contains only 49,603 stories. Thus it is not surprising that stories generated using VIST show low lexical diversity. Finally, no existing work takes into account relations between photos. Visual storytelling was introduced as an artificial intelligence task that attempts to "interpret causal structure" and "make sense of visual input to tie disparate moments together" (Huang et al. 2016). However, existing approaches still treat it as a sequential image captioning problem and omit relations between images.

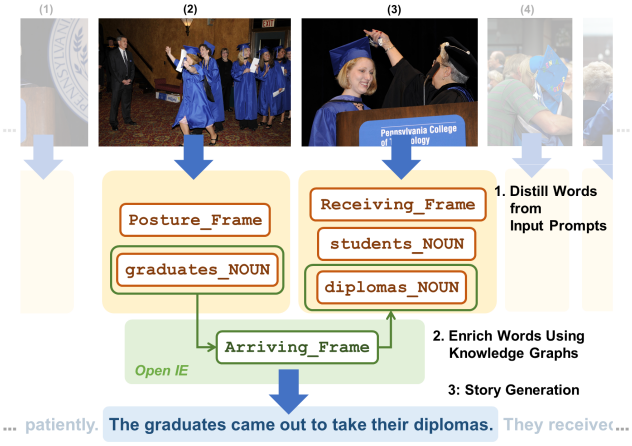

Figure 1: Overview of KG-Story. KG-Story first *(i)* distills a set of representative words from the input prompts and *(ii)* enriches the word set by using external knowledge graphs, and *(iii)* generates stories based on the enriched word set.

*denotes equal contribution

We introduce KG-Story, a three-stage framework that allows the story generation model to take advantage of external resources, especially knowledge graphs, to produce interesting stories. KG-Story first distills a set of representative words from the input prompts and enriches the word set using external knowledge graphs, and finally generates stories based on the enriched word set. This *distill-enrich-generate* framework allows the use of external resources not only for the enrichment phase, but also for the distillation and generation phases. Figure 1 overviews the workflow of KG-Story. KG-Story first distills a set of words from each image, respectively. For example, the words `Posture_Frame` and `graduates_NOUN` are extracted from image (2), and the words `Receiving_Frame`, `students_NOUN` and `diplomas_NOUN` are extracted from image (3). In stage 2, KG-Story then searches in a knowledge graph to find potential relations between word pairs across images. KG-Story uses a scoring function to rate potential relations when multiple relations are found. In Figure 1, the system finds the relation `Arriving_Frame` between `graduates_NOUN` and `diplomas_NOUN`, denoted as graduates_NOUN $\xrightarrow{\text{Arriving\_Frame}}$ diplomas_NOUN. Finally, in stage 3, KG-Story uses an story generator to produce the final story sentence using all ingredients. The `Arriving_Frame` is defined as "An object Theme moves in the direction of a Goal." in the knowledge graph (FrameNet), illustrating a graduates' common body motion in the commencement. As a result, the phrase "came out to take" in the final output sentence captures this knowledge extracted by KG-Story.

The contributions of this paper are twofold. First, the proposed framework opens up possibilities to using large external data sources to improve automated visual storytelling, instead of struggling with specialized small data sources. Second, the proposed model leverages external knowledge to explore the relations between images and increases the output text diversity. We show in this paper by human ranking evaluation that the stories generated by KG-Story are on average of higher quality than those from the state-of-the-art systems.

Related Work

As researchers have found that middle-level abstraction can result in more coherent stories, (Yao et al. 2018) show that more diverse stories are generated by using a pipeline involving first planning the overall storyline and then composing the story. They propose an event-based method which transforms the previous story into event representations, predicts successive events, and generates a story using the predicted events (Martin et al. 2018). Similarly, (Xu et al. 2018) propose a skeleton-based narrative story generation method which uses a skeleton as a middle-level abstraction layer. However, although multi-stage text story generation has been a focus of recent studies, limited effort has been expended on multi-stage visual storytelling. One of the few exceptions is work done by Huang et al. (2018). Leveraging the middle-level abstraction as well, they propose a two-level hierarchical decoder model for the VIST task. They first predict a topic (*e.g.*, indoors, kids, the baby) for each image in the sequence, after which the low-level decoder generates a sentence for each image conditioned on the given topic. The two decoders are jointly trained end-to-end using reinforcement learning. As the topics can be selected only from fixed clusters using a discriminative model, the ability of story planning is limited. Furthermore, under an end-to-end training process, extra plain text and knowledge base data cannot be utilized to better enrich the story. In a similar task, (Mathews, Xie, and He 2018) propose a two-stage style transfer model for image captioning. Their first stage consists of keyword prediction for the caption of each independent image, and then a stylish caption is generated using an RNN trained on story corpora. However, each such caption is a single story-like sentence and is independent of other captions; combined, the captions do not constitute a context-coherent story.

KG-Story

In this work, a three-stage approach is proposed for knowledge-enriched visual storytelling.

Stage 1: Word distillation from input prompts. Given an image, we train an image-to-term model to distill terms from each of the input images; this can be regarded as word-form conceptual representation.

Stage 2: Word enrichment using knowledge graphs. With the five sets of terms extracted from a sequence of images in the previous step, we utilize an external knowledge graph to identify possible links between sets, and generate the final enriched term path for story generation.

Stage 3: Story generation. We use a Transformer architecture to transform term paths into stories. A length-difference positional encoding and a repetition penalty are used in the proposed Transformer; we also apply a novel anaphoric expression generator.

In the following, we describe the three steps in detail.

Stage 1: Word Distillation from Input Prompts

In Stage 1, we extract terms from the image. To this end, we build a model that uses a pre-trained Faster R-CNN (Ren et al. 2015; Anderson et al. 2018) as the image feature extractor and a Transformer-GRU as term predictor.

The Faster R-CNN model was originally trained for the object detection task. Here we extract the features of its predicted object as the image representation. To reduce computational complexity, only the object features within the top 25 confidence scores are used.

As shown in Figure 2, the image representation is fed into a Transformer encoder (Vaswani et al. 2017) and a GRU (Chung et al. 2014) decoder with an attention mechanism (Bahdanau, Cho, and Bengio 2014) as our term prediction model. The only modification to the Transformer for distillation is the positional encoding. In contrast to the positional encoding in the original setting, object features are summed with trainable image-order embeddings as input, as

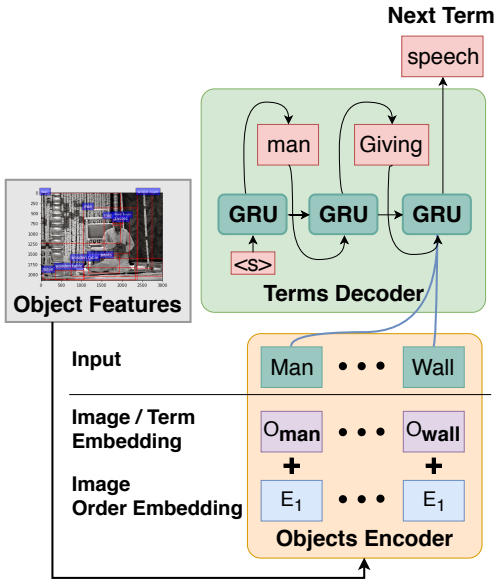

Figure 2: Stage 1: Image distillation model. In Stage 1, KG-Story extract terms from the image. We build a model that uses a pre-trained Faster R-CNN as the image feature extractor and a Transformer-GRU as the term predictor.

objects in the same image are sequentially independent:

$$x_i = Wo_i + \text{Order\_Embedding}(t)$$

where o_i is the i-th object and t is the order of the image. $W^{2048\times D}$ is a matrix that transforms the 2048-dimensional o_i into dimension D. Then object features are fed into the term prediction model to generate the terms for each photo. During decoding, we use an intra-sentence repetition penalty with beam search to reduce redundancy in the visual story (Hsu et al. 2018). The term score at each beam search step is computed as

$$\text{beam\_score}(x) = \log(p(x)) - 1e^{19} \cdot \mathbb{1}(x \in S),$$

where $p(x)$ is the model's output probability for each term, the last term denotes repetition penalty, and set S denotes terms that have already been predicted in an image.

Stage 2: Word Enrichment Using Knowledge Graphs

Given a sequence of images for story generation, we have observed that nearly irrelevant sequential images are common. In previous end-to-end models, such a phenomenon hurts story generation and causes models to create caption-like incoherent stories which are relatively boring and semantically disconnected. To take this into account while enriching the story, we introduce semantic terms as the intermediate and link terms in two adjacent images using the relations provided by the knowledge graph. In the knowledge graph, real world knowledge is encoded by entities and their relationship in the form of tuples *(head, rela, tail)*, indicating that *head* has a *rela* relationship with *tail*.

More specifically, given the terms of a sequence of image predicted in the previous step $\{m_1^1, .., m_i^t, .., m_{N_5}^5\}$, where m_i^t denotes the i-th term distilled from the t-th image, we pair the terms from two consecutive images and query the knowledge graph for all possible tuples $\{...., (m_i^t, r_k, m_j^{t+1}), ...\}$ where r_k is a one-hop relation that links m_i^t and m_j^{t+1} in the knowledge graph. Furthermore, we also consider two-hop relations, which can convey indirect relationships, to enrich the story by adding $(m_i^t, r_k, m_{middle}, r_q, m_j^{t+1})$ if such a relation exists in the knowledge graph.

After all tuples of one- and two-hop relations are extracted, we construct candidate terms from them and insert each relation into the term set for story generation, i.e., the extracted (m_i^l, r_k, m_j^{l+1}) or $(m_i^l, r_k, m_{middle}, r_q, m_j^{l+1})$, which contains the head term from image l and the tail term from image $l+1$, is inserted between images l and $l+1$ to generate an additional story sentence, as if these were the terms distilled in between the images.

With all possible term sets constructed, we must select the most reasonable one for the next term-to-story generation step. Hence we train an RNN-based language model on all the available textual stories. A language model estimates the probability distribution of a corpus

$$P(U) = \Sigma \text{logP}(u_i | u_1, ..., ui - 1)$$

where $U = u_1, ..., u_n$ is the corpus. Here we use it to compute the probability of a term conditioned on the other terms existing previously. Thus we obtain the perplexity of this term path, choose that term path with the lowest perplexity, and feed it to the next step.

Actually, Stage 2 mimics the way people generate a story based on two irrelevant images, that is, relating two images through imagination. Here the knowledge graph serves as the source of ideas connecting two images and the language model then ensures the coherence of the generated story when using the selected idea.

Stage 3: Story Generation

For the story generation step, we leverage the Transformer (Vaswani et al. 2017) shown in Figure 3 with the input, i.e., the term set, from Stage 2. We add to the original Transformer model three different modifications: *(i)* length-difference positional encoding for variable-length story generation, *(ii)* anaphoric expression generation for the unification of anaphor representation, and *(iii)* a repetition penalty for removing redundancy.

Length-Difference Positional Encoding The sinusoidal positional encoding method (Vaswani et al. 2017) inserts the absolute positions to a sinusoidal function to create a positional embedding. However, when the model is used to generate variable-length stories, this positional encoding method makes it difficult for the model to recognize sentence position in a story. For the visual storytelling task, all the samples contain five images, each of which is described generally in one sentence for the corresponding storyline; thus the generated stories always contain five sen-

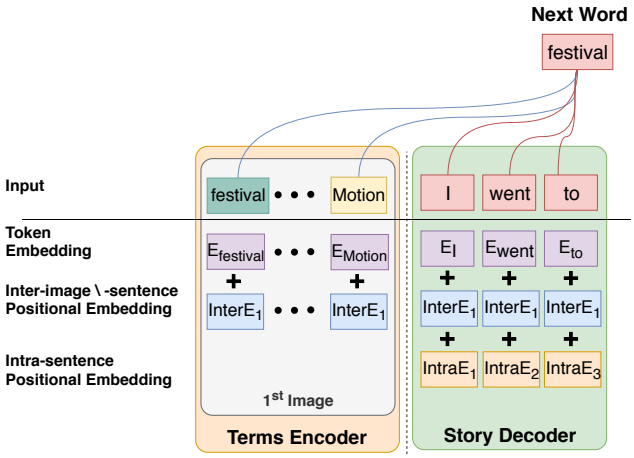

Figure 3: Stage 3: Story generation model. For the story generation step, we leverage the Transformer with the input (*i.e.*, the term set) from Stage 2.

tences, which prevents us from adding additional ingredients to enrich them. To tackle this problem, we adopt length-difference positional encoding (LDPE) (Takase and Okazaki 2019) as shown in Equations (1) and (2), where pos is the position, d is the embedding size, len is the length constraint, i is the dimension of the sinusoidal positional encoding, and $10000^{2i/d} \times 2$ is the period of a sinusoidal function. LDPE allows the Transformer to learn positional embeddings with variable-length stories. Unlike positional encoding that uses absolute positions, LDPE returns the remaining length of a story. As a result, the terminal position is identical for dissimilar story lengths. Consequently, the model learns to generate ending sentence for stories with different lengths.

$$LDPE_{(pos,len,2i)} = \sin\left(\frac{len - pos}{10000^{\frac{2i}{d}}}\right) \quad (1)$$

$$LDPE_{(pos,len,2i+1)} = \cos\left(\frac{len - pos}{10000^{\frac{2i}{d}}}\right) \quad (2)$$

Anaphoric Expression Generation To enable the use of pronouns in the enhanced KG-Story model for anaphoric expression generation, we adopt a coreference replacement strategy. To replace coreferences, we first use a coreference resolution tool[1] on the stories to find the original mention of each pronoun and replace the pronouns with their root entities. The entity types are the ten defined in the OntoNotes dataset (Weischedel et al. 2013), including PERSON, PRODUCT, ORG, and so on. Open-SESAME then extracts terms again from the stories in which the coreferences have been replaced; these extracted terms are then used as the input – with the original stories as the output – to train the story generator in Stage 2. As such, the story generator learns to generate pronouns given two successive identical noun terms. Table 1 shows a short visual story of two successive images where both contain the same object (a dog), and the corresponding gold story reads "*The dog is ready to go. He is playing on the ground.*" In this example, the term predictor predicts *dog* for both images, but the term extractor cannot extract *dog* from the second gold sentence. To bridge this gap, coreference replacement replaces the pronoun with its original mention "*the dog*" in the second gold sentence, which enables the story generator to learn to generate anaphoric expressions when seeing multiple mentions of "*the dog*".

[1]NeuralCoref 4.0: Coreference Resolution in spaCy with Neural Networks. https://github.com/huggingface/neuralcoref


Original story	The dog is ready to go.	He is playing on the ground.
Terms	"Dog_Noun", "Motion_Frame"	"Performers_and_roles_Frame", "Ground_Noun"
Story w/ CR	The dog is ready to go.	The dog is playing on the ground.
Terms	"Dog_Noun", "Motion_Frame"	"Dog_Noun", "Performers_and_roles_Frame", "Ground_Noun"

Table 1: Term extraction and coreference replacement (CR). Image order is from left to right.

Repetition Penalty in Story Generation Instead of using a repetition penalty for predicted terms only, we use inter- and intra-sentence repetition penalties with beam search to reduce redundancy in the visual story (Hsu et al. 2018). The word score at each beam search step is

$$\text{beam\_score}(x) = \log(p(x)) - \alpha \cdot \mathbb{1}(x \in S) - (\gamma/l) \cdot \mathbb{1}(x \in R),$$

where α and γ are hyperparameters empirically set to 20 and 5. Set S denotes words that appear in the current sentence, and set R denotes words in previous sentences. The current story's length l is a regularization term which reduces the penalty to prevent the story generator from refusing to repeat words when generating grammatically correct sentences as the length of the sentence or story increases.

Experimental Setup

In this paper, KG-Story leverages the object detection model to find reliable candidates for story casting, which provides robust grounding. We also use the image knowledge graphs such as Visual Genome (Krishna et al. 2016) to extract activities to correlate the roles selected by the system. FrameNet (Baker, Fillmore, and Lowe 1998) terms are used to present the semantic concepts of selected roles and their extracted correlated activities for the story plot, which later guides the model to compose the visual story from a panoramic point of view. We describe the experimental setup in detail in this section.

Data Preparation

Four datasets were used in this paper: Visual Genome, OpenIE, ROCStories Corpora, and VIST Dataset. VIST Dataset provides image-to-term materials for learning in Stage 1. For Stage 2, the object relations from Visual Genome or the term relations from OpenIE are the materials for Stage 2. ROCStories Corpora supplies a large quantity of pure textual stories for generation in Stage 3, and the VIST Dataset, the sole end-to-end visual storytelling dataset, is used to fine-tune the model. Note that when we here refer to all the available textual stories, these are from the VIST Dataset and the ROCStories Corpora.

Visual Genome Visual Genome contains labels for tasks in language and vision fields such as object detection and object relation detection. Visual Genome has 108,077 images, 3.8 million object instances, and 2.3 million relationships. We use it to pretrain the object detection model used in Stage 1. In addition, we utilize the noun and verb relations provided by the scene graph of Visual Genome to brainstorm activities which could link two photos. Relations of nouns and verbs in each image have been labeled as *(subject, verb, object)*. The reasons we use verb relations are twofold. First, compared to unseen objects, activities are better ingredients to add in visual stories for correlation as they do less harm to grounding. Second, as most image-to-text datasets focus on objects, stories generated based on them are relatively static. Hence adding activities makes our stories vivid.

OpenIE OpenIE is an information extractor that finds relationships in sentences (Pal and others 2016; Christensen et al. 2011; Saha, Pal, and others 2017; Saha and others 2018). Here we use its relations extracted from various corpora as another knowledge base.[2] Note that we only extract one-hop relations from OpenIE.

ROCStories Corpora We use ROCStories, which contains 98,159 stories, to train the story generator (Mostafazadeh et al. 2016). As the annotators of ROCStories were asked to write a five-sentence story given a prompt, stories focus on a specific topic with strong logical inference. Similar to VIST, we extract terms from sentences of stories which are used as the input data of the story generation model.

VIST Dataset This is a collection of 20,211 human-written visual stories and 81,743 unique photos. Each story in VIST contains five sentences and five corresponding photo images. To train our image-to-term prediction model, we first extract terms from sentences of stories in the VIST dataset as the gold labels. Inspired by Semstyle (Mathews, Xie, and He 2018), we consider the key components of the sentence to be its noun (object) and verb (action) terms; therefore, given a sentence, we first detect nouns and verbs using a part-of-speech tagger[3]. Then, we extract verb frames to replace original verbs using OpenSESAME (Swayamdipta et al. 2017), a frame-semantic parser which automatically detects FrameNet (Baker, Fillmore, and Lowe 1998) frames. For example, in Table 1, the terms extracted from *"The dog is ready to go"* are "Dog_Noun" and "Motion_Frame".

As for the term-to-story generation model, the Transformer generator used in Stage 3 is first trained on the ROCStories dataset with the textual stories and their extracted terms, and then fine-tuned on both the ROCStories and VIST datasets.

Human Evaluation (Story Displayed **with** Images)

	GLAC (Kim et al. 2018)	No KG	OpenIE	Visual Genome	Human
Avg. Rank (1 to 5)	3.053	3.152	**2.975***	**2.975***	2.846

Table 2: Direct comparison evaluation of KG-Story model. Numbers indicate average rank given to stories (from 1 to 5, lower is better.) Stories generated by KG-Story using either OpenIE or Visual Genome are on average ranked significantly better (lower) than that of GLAC (unpaired t-test, $p < 0.05$, N=2500).

Hyperparameter Configuration

In all of our experiments, we used the same hyperparameters to train our model. The hidden size of the term prediction and story generation models was set to 512. The head and layer number of the Transformer encoder were 2 and 4. Both models were trained with the Adam optimizer with an initial learning rate of 1e-3, which decayed with the growth of training steps. During decoding, the beam size was set to 3 for both modules.

Results and Discussion

We conducted evaluations using both automatic metrics and human ranking. Automatic metrics were evaluated by shallow features such as the word coverage or semantics for the quality stories (Wang et al. 2018b), whereas human evaluation provided solid results due to their complete comprehension. For human evaluations, we adopted the ranking of stories from all models and the stories written by humans. All human evaluations were conducted on the Amazon Mechanical Turk; five different workers were assigned for each comparison. Through this direct comparison, we were able to provide a baseline for turkers, yielding reliable results.

Human Evaluation

Ranking Stories with Images We conducted a human evaluation using crowd workers recruited from Amazon Mechanical Turk. Each task displayed one sequence of photos and several generated stories based on the sequence. Five workers were recruited for each task to rank the story quality, "from the Best Story to the Worst Story." The compensation was $0.10 per task. Four models were included in this

[2] https://openie.allenai.org/

[3] SpaCy: https://spacy.io/

No KG: everyone was having a great time at the party . all of my friends were dressed up . some girls even had funny costumes . they sang and danced . we celebrated with them that night .

Visual Genome: i had a great time at the party . everyone was dressed up in their costumes . all of my friends were there . ***one of the girls sat with a man and his friends .*** the girls were very happy . at the end of the night , everyone was happy .

OpenIE: i had a great time at the party . ***everyone had a great time .*** they were dressed up in funny costumes . all of my friends were very impressed . the girls were very grateful . by the end of the night everyone was tired .

GLAC: the halloween party was a lot of fun . there were a lot of people dressed up . some of the costumes were very scary . i had a great time . everyone had a great time .

Human: the year 2000 was on its way , so a group of friends decided to welcome it with a party . they filled their cups with beer and put on some new year gear . excitement built as midnight approached and the friends talked and danced . at 11:59 pm it was time to gather around and count down to the new year together . after midnight , everybody felt tired and giddy . the year 2000 was off to a good start !

Visual Genome: the city at night was beautiful . this building had a lot of fun . ***there was a bus on the bridge next to an old building .*** the bridge was very old and beautiful . this is a very nice place to explore . the streets were very narrow .

GLAC: the city was lit up at night . the buildings were tall and bright . the skyline was beautiful . the streets were busy with people . the streets were empty .

Human: the skyscrapers are some of the tallest buildings across the country . at night , the city hosted a nightly carnival . the bridge is much more convenient at night . we decided to use the bridge to get to the city carnival in record breaking time . many vendors had great food to offer at the carnival . the carnival had many inner city people show up .

OpenIE: the wedding reception was very special . it was a beautiful house . there were so many trees . everyone had a great time . ***even the dog had a great time !*** the dog was very well behaved .

GLAC: the family was having a great time at the christmas party . the tree was covered in snow . the trees were beautiful . the kids were very excited . the baby was happy to be there .

Human: we visited family for christmas . they live out in the country far from the city . the trees lost their leaves because it is so cold outside . they were so happy that we had arrived . even the dog had a marry christmas .

Figure 4: Example stories generated by visual storytelling models. The knowledge-enriched sentences are highlighted.

Human Evaluation (Story Displayed **without** Images)

	GLAC (Kim et al. 2018)	No KG	OpenIE	Visual Genome	Human
Avg. Rank (1 to 5)	3.054	3.285	**3.049**	**2.990**	2.621

Table 3: Direct comparison evaluation without given photos. Numbers indicate average rank given to stories (from 1 to 5, lower is better).

evaluation: *(i)* KG-Story using OpenIE as the knowledge graph, *(ii)* KG-Story using VisualGenome as the knowledge graph, *(iii)* KG-Story without any knowledge graphs, and *(iv)* GLAC (Kim et al. 2018), the current state-of-the-art visual storytelling model. Note that in the first two models, only the knowledge graph for exploring inter-image relations in Stage 2 is different, while Stage 1 and Stage 3 remain identical. As we also included *(v)* human-written stories for comparison, the ranking number is from 1 (the best) to 5 (the worst). Table 2 shows the evaluation results, showing that KG-Story benefits from the enrichment of knowledge. Stories generated by KG-Story using either OpenIE or Visual Genome are on average ranked significantly better (lower) than that of GLAC (unpaired t-test, $p < 0.05$, N = 2500), whereas stories generated by KG-Story without using any knowledge graphs ranked significantly worst (higher) (unpaired t-test, $p < 0.01$, N = 2500).

Ranking Stories without Images We also explored the relation between images and the generated stories in visual storytelling. To determine whether the generated stories are well matched to the photos, we removed the photos and conducted another overall story comparison. Results in Table 3 show that when photos were provided in human evaluation, visual stories from the KG-Story were ranked higher than those from the state-of-the-art GLAC (2.97 vs. 3.05, significant with p-value < 0.05); they become comparable (3.04 vs. 3.05) when the photos were not provided. The ranking of visual stories from the state-of-the-art model remains the same. This suggests that visual stories from KG-Story better fit the photos, where the images and texts support each other. The third story in Figure 4 illustrates an example. The visual story from the state-of-the-art GLAC clearly is not grounded: there is no snow in the second photo, no kids in the fourth, and no baby in the fifth, though only considering the text it is a reasonable Christmas story. In contrast, given the photos, the proposed KG-Story generates an obviously better story. This example also illustrates that the given images are both hints and constraints in visual storytelling, which makes it a more challenging task than general story generation. Besides, even when photos are not provided, the average rank of visual stories from KG-Story is comparable to that from the state-of-the-art model. This demonstrates that our stories are both closely aligned to the images and also stand without the images, which we attribute to the use of the additional Virtual Genome dataset and ROCStories Corpora for the model learning.

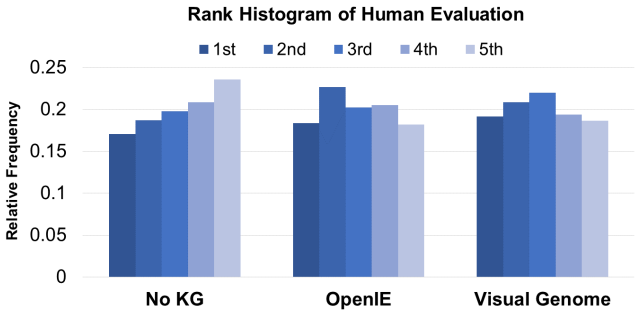

Figure 5: The distribution of the ranks received by each model. From the darkest blue to the lightest blue is the best rank to the worst.

The Effect of Using Knowledge Graphs We also investigate the effect of using a knowledge graph on story quality. Figure 5 shows the distribution of ranks assigned to models with and without the enrichment of the knowledge graph. Overall we note a clearly increasing number of stories ranked as the first, second, and third place when considering the knowledge graph, which confirms its benefit to the quality of stories. Specifically, stories from the model with the Visual Genome receive more first ranks and also more worst ranks than those from the model with OpenIE. In other words, compared to stories enriched by Virtual Genome, those from OpenIE are relatively moderate. This may result from the existence of the two-hop relations in Visual Genome, whereas OpenIE contains only one-hop relations. As a result, the model with Visual Genome tends to be more imaginative and unrestrained, leading to stories that are either extremely good or extremely bad.

Observing Automatic Metrics As prior work has clearly shown that the classic automatic evaluation metrics are weak or even negative quality indicators for visual storytelling (Hsu et al. 2019; Wang et al. 2018b), we do not use these for evaluation. However, they still have something to say about token-based textual analysis. Interestingly, we observe that the use of knowledge graphs increases the scores in precision-based metrics (i.e., BLUE1, BLEU2, and CIDEr) with the cost of slightly reduced scores in recall-based metrics (Table 4).

	BLEU1	BLEU4	METEOR	ROUGE	CIDEr
GLAC (Kim et al. 2018)	.374	.050	**.297**	**.250**	.049
No KG	.376	.050	.295	.237	.055
OpenIE	.446	**.056**	**.297**	.241	.094
Visual Genome	**.451**	**.056**	.296	.241	**.096**

Table 4: Automatic evaluation results. Prior work has clearly shown that the classic automatic evaluation metrics are weak or even negative quality indicators for visual storytelling (Hsu et al. 2019; Wang et al. 2018b).

Discussion

In this section, we discuss some of our observations throughout the experiments.

A New Relation Bridges Two Images We observe that often the newly added relation creates bridges between two images. For example, for the first photo sequence in Figure 4, KG-Story with Visual Genome distills *friends* from the third image and *girls* from the fourth image. Then the system connects the two terms with the two-hop relation *(friends, posture, man)* and *(man, posture, girls)*. Similarly, KG-Story with OpenIE links *time* from the first image and *everyone* from the second image with the relation *(time, cause to experience, everyone)*. We also find that after adding KG terms to the term set, the entire generated story is different from the No-KG stories, even though the terms for the original images are not changed. This shows that KG-Story indeed considers all terms together when generating stories, and hence added terms influence its context.

Low-ranked Human Generated Stories From Table 2 and Table 3 we notice that human generated stories are not ranked prominently better comparing the the ranked given to machine generated stories. In Table 2 it is 2.85 compare to 2.97 and in Table 3 it is 2.62 compare to 2.99. In addition to the good quality of our generated stories, we found that this is also because in the evaluation we allow crowd workers to rank stories directly by comparison without explicitly giving pre-defined rubrics. Since people may have different definition for a good story, the machine-generated stories were not always ranked lower than human-written ones. For example, some people may value an interesting plot more than a focused or detailed description.

Repetition Penalty Reduces Repetition We observe that there are repetitions in GLAC's stories. For example, in story "*the new office is a very interesting place. the kids love the toy. they also have a lot of books. they have a nice collection of books. they even have a stuffed animal*" and "*the kids were having a great time at the party. they had a lot of fun. they had a lot of fun. they had a lot of fun. they had a great time.*", the pattern "*they have*" and "*they had a lot of fun*" both respectively appear three times, which is generally less human and not ideal. We find that this phenomenon is reduced by the repetition penalty in KG-Story.

When Nonexistent Entities Are Added Adding terms that refer to entities or people that are not actually in the image is a potential risk when enriching stories with information from the KG. We find this problem to be especially severe for two-hop relations. The second photo sequence in Figure 4 is an example in which the KG fails to help. After *building* is distilled from the second image and *bridge* is distilled from the third image, KG-Story selects the link for enrichment: $building \xrightarrow{\text{obscurity}} bus \xrightarrow{\text{travel}} bridge$. While this link itself as well as the generated sentence "*there was a bus on the bridge next to an old building*" are both quite valid, it does not align to the images as the bus, the building, and the bridge are not in the described relative positions. In this case, the model generates unsatisfactory stories.

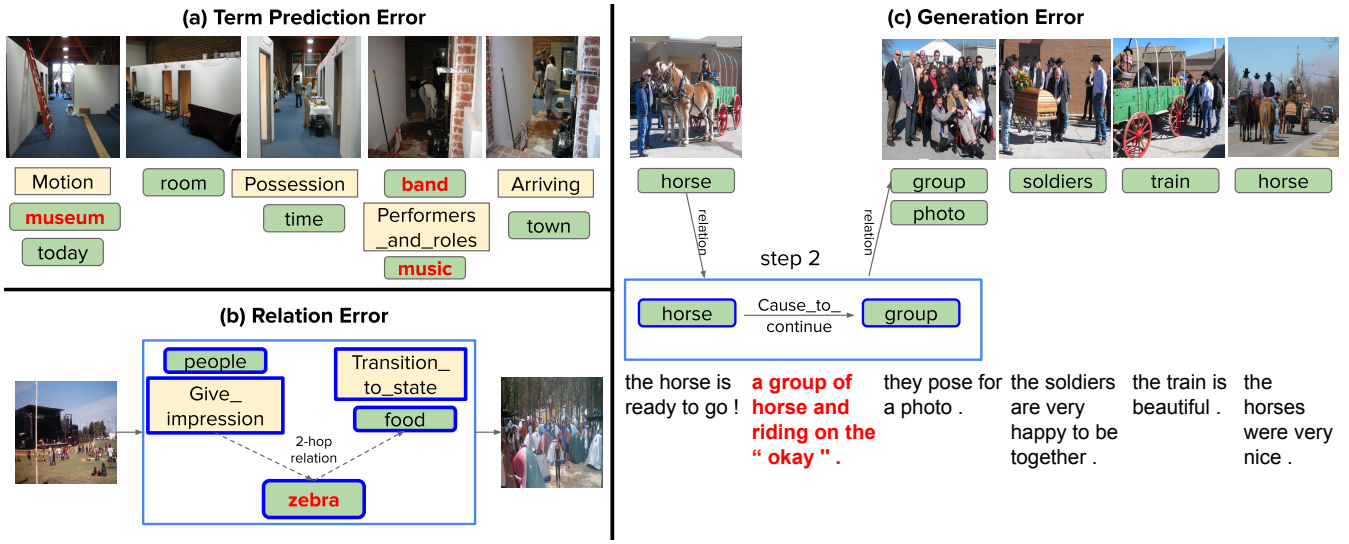

Figure 6: Error analysis examples. Green cells are noun terms. Yellow cells are verb terms. Blue border cells are terms enriched by step 2. Red text cells are the identified errors. (a) shows a term prediction error where the model predicted unrelated terms museum, band and music for pictures about construction; in (b), the model selected a 2-hop relation which contained an unrelated object zebra from the knowledge graph to add in, causing a relation error; in (c) a generation error was identified due to the incomprehensible sentence marked in red.

Error Analysis To understand the behavior of KG-Story, we conducted a small-scale error analysis on 50 stories. We randomly selected stories for examination from 1,999 stories with their corresponding photo sequences, and kept those stories containing errors until we reached the number 50. Then these 50 stories were manually examined. Each of the 50 stories exhibit one of the following errors: term prediction error, relation error, and generation error. Figure 6 shows the example of each error type. The following describes a distribution of the errors from our analysis:

1. **Term Prediction Error (Stage 1): 48%.** The term prediction error occurs when the story or terms are redundant or unrelated to the images content regardless the additional sentence. The term that were unrelated to the images caused the model to generate irrelevant story, and thus, deteriorate the overall performance.
2. **Relation Error (Stage 2): 26%.** The relation error means the extracted relations are redundant with existing terms or does not match with visual content in the images. Bounding the added relations with images may further mitigate their impact.
3. **Generation Error (Stage 3): 26%.** Generation error refers to the low-quality generated sentence under the condition that the given terms are valid. This may result from some rarely seen combination of terms for which the generation model does not know how to describe them. Collecting a large quantity of diverse stories for learning could be a solution.

In addition to random selection, we also identified 12 additional low-ranked KG-stories and their errors to find the cause of low ranking. We checked stories that were ranked 4 and 5 in the previous evaluation. After the analysis, the error distribution was 33%, 17%, and 50% for term prediction error, relation error, and generation error, respectively. Compared to the errors in the randomly selected stories, the decrease percentage of the grounding error and the increase percentage of the sentence generation error suggest that people tend to give relatively low rating when seeing simple, easily detected errors in story sentences.

Conclusion

In this paper, we seek to compose high-quality, visual stories that are enriched by the knowledge graph. We propose KG-Story, a novel three-stage visual storytelling model which leverages additional non-end-to-end data. In the generation process we address positional information, anaphors, and repetitions. Human evaluation shows that KG-Story outperforms the state of the art in both automatic evaluation and human evaluation for direct comparison. In addition, we show that even when evaluating without the corresponding images, the generated stories are still better than those from the state-of-the-art model, which demonstrates the effectiveness of improving coherence by the knowledge graph and learning from additional uni-modal data. Given these encouraging results, composing arbitrary-length stories from the same input hints is our next goal.

Acknowledgements

This research is partially supported by Ministry of Science and Technology, Taiwan under the project contract 108-2221-E-001-012-MY3, 108-2634-F-001-004-, and the Seed Grant (2019) from the College of Information Sciences and Technology (IST), Pennsylvania State University.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6077–6086.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

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

Christensen, J.; Soderland, S.; Etzioni, O.; et al. 2011. An analysis of open information extraction based on semantic role labeling. In *Proceedings of the sixth international conference on Knowledge capture*, 113–120. ACM.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Fan, A.; Lewis, M.; and Dauphin, Y. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

Hsu, C.-C.; Chen, S.-M.; Hsieh, M.-H.; and Ku, L.-W. 2018. Using inter-sentence diverse beam search to reduce redundancy in visual storytelling. *arXiv preprint arXiv:1805.11867*.

Hsu, T.-Y.; Huang, C.-Y.; Hsu, Y.-C.; and Huang, T.-H. 2019. Visual story post-editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6581–6586. Florence, Italy: Association for Computational Linguistics.

Huang, T.-H. K.; Ferraro, F.; Mostafazadeh, N.; Misra, I.; Devlin, J.; Agrawal, A.; Girshick, R.; He, X.; Kohli, P.; Batra, D.; et al. 2016. Visual storytelling. In *Proceedings of NAACL 2016*.

Huang, Q.; Gan, Z.; Çelikyilmaz, A.; Wu, D. O.; Wang, J.; and He, X. 2018. Hierarchically structured reinforcement learning for topically coherent visual story generation. *CoRR* abs/1805.08191.

Kim, T.; Heo, M.-O.; Son, S.; Park, K.-W.; and Zhang, B.-T. 2018. Glac net: Glocal attention cascading networks for multi-image cued story generation. *CoRR* abs/1805.10973.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; Bernstein, M.; and Fei-Fei, L. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Martin, L. J.; Ammanabrolu, P.; Wang, X.; Hancock, W.; Singh, S.; Harrison, B.; and Riedl, M. O. 2018. Event representations for automated story generation with deep neural nets. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Mathews, A. P.; Xie, L.; and He, X. 2018. Semstyle: Learning to generate stylised image captions using unaligned text. *CoRR* abs/1805.07030.

Mostafazadeh, N.; Chambers, N.; He, X.; Parikh, D.; Batra, D.; Vanderwende, L.; Kohli, P.; and Allen, J. F. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of NAACL 2016*.

Pal, H., et al. 2016. Demonyms and compound relational nouns in nominal open ie. In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction*, 35–39.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

Saha, S., et al. 2018. Open information extraction from conjunctive sentences. In *Proceedings of the 27th International Conference on Computational Linguistics*, 2288–2299.

Saha, S.; Pal, H.; et al. 2017. Bootstrapping for numerical open ie. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 317–323.

Swayamdipta, S.; Thomson, S.; Dyer, C.; and Smith, N. A. 2017. Frame-Semantic Parsing with Softmax-Margin Segmental RNNs and a Syntactic Scaffold. *arXiv preprint arXiv:1706.09528*.

Takase, S., and Okazaki, N. 2019. Positional encoding to control output sequence length. *Proceedings of the 2019 Conference of the North*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008.

Wang, J.; Fu, J.; Tang, J.; Li, Z.; and Mei, T. 2018a. Show, reward and tell: Automatic generation of narrative paragraph from photo stream by adversarial training. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Wang, W. Y.; Wang, X.; Chen, W.; and Wang, Y. 2018b. No metrics are perfect: Adversarial reward learning for visual storytelling. In *Proceedings of NAACL 2018*.

Weischedel, R.; Palmer, M.; Marcus, M.; Hovy, E.; Pradhan, S.; Ramshaw, L.; Xue, N.; Taylor, A.; Kaufman, J.; Franchini, M.; et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*.

Xu, J.; Zhang, Y.; Zeng, Q.; Ren, X.; Cai, X.; and Sun, X. 2018. A skeleton-based model for promoting coherence among sentences in narrative story generation. *arXiv preprint arXiv:1808.06945*.

Yao, L.; Peng, N.; Weischedel, R. M.; Knight, K.; Zhao, D.; and Yan, R. 2018. Plan-and-write: Towards better automatic storytelling. *CoRR* abs/1811.05701.