Knowledgeable Storyteller: A Commonsense-Driven Generative Model for Visual Storytelling

Pengcheng Yang[1,2] , **Fuli Luo**[2] , **Peng Chen**[2] , **Lei Li**[2] , **Zhiyi Yin**[2] , **Xiaodong He**[3] , **Xu Sun**[1,2]

[1]Deep Learning Lab, Beijing Institute of Big Data Research, Peking University

[2]MOE Key Lab of Computational Linguistics, School of EECS, Peking University

[3]JD AI Research, China

{yang_pc,luofuli,chen.peng,yinzhiyi,xusun}@pku.edu.cn, tobiaslee@foxmail.com, xiaodong.he@jd.com

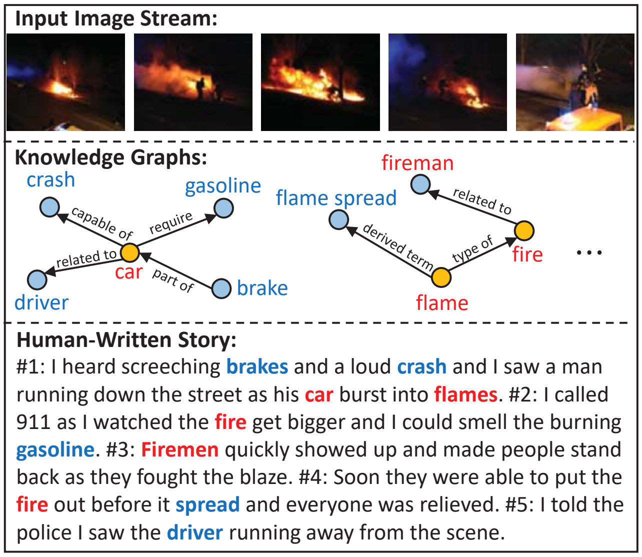
Abstract

The visual storytelling (VST) task aims at generating a reasonable and coherent paragraph-level story with the image stream as input. Different from caption that is a direct and literal description of image content, the story in the VST task tends to contain plenty of imaginary concepts that do not appear in the image. This requires the AI agent to reason and associate with the imaginary concepts based on implicit commonsense knowledge to generate a reasonable story describing the image stream. Therefore, in this work, we present a commonsense-driven generative model, which aims to introduce crucial commonsense from the external knowledge base for visual storytelling. Our approach first extracts a set of candidate knowledge graphs from the knowledge base. Then, an elaborately designed vision-aware directional encoding schema is adopted to effectively integrate the most informative commonsense. Besides, we strive to maximize the semantic similarity within the output during decoding to enhance the coherence of the generated text. Results show that our approach can outperform the state-of-the-art systems by a large margin, which achieves a 29% relative improvement of CIDEr score. With additional commonsense and semantic-relevance based objective, the generated stories are more diverse and coherent. [1]

1 Introduction

Automatic **v**isual **s**tory**t**elling (VST) aims to generate a reasonable and coherent story with a set of images as input [Huang *et al.*, 2016]. It not only can be applied in plenty of real-world scenarios, e.g., helping visually impaired people better understand the content of images on the web, but also reflects the advanced creativity of an intelligent system.

Despite its importance described above, the VST task has not been widely explored. One line of research [Gonzalez-Rico, 2018; Hsu *et al.*, 2018; Kim *et al.*, 2018] focuses on designing specific network architectures to improve results under the framework of maximum likelihood estimation (MLE) method, while another line [Chen *et al.*, 2017; Wang *et al.*, 2018a; Wang *et al.*, 2018b] strives to generate more expressive outputs via adversarial training or reinforcement learning. Although these researches have achieved promising results to some extent, they lack the capability of commonsense reasoning that is crucial for visual storytelling. The story on the VST task tends to contain plenty of imaginary concepts that do not appear in the image, and semantic association and logical reasoning based on implicit commonsense knowledge are of great help to generate these imaginary concepts. Figure 1 presents an example of visual storytelling. The human-written story contains imaginary concepts "*brake*" and "*crash*", which can be regarded as commonsense about "*car*" depicted in images. These three concepts are connected as a knowledge graph in the knowledge base (*ConceptNet*). Thus, with the help of commonsense, the model can easily generate these imaginary concepts.



#1: I heard screeching **brakes** and a loud **crash** and I saw a man running down the street as his **car** burst into **flames**. #2: I called 911 as I watched the **fire** get bigger and I could smell the burning **gasoline**. #3: **Firemen** quickly showed up and made people stand back as they fought the blaze. #4: Soon they were able to put the **fire** out before it **spread** and everyone was relieved. #5: I told the police I saw the **driver** running away from the scene.

Figure 1: An example of visual storytelling. Red words are concepts depicted in images, and blue words are imaginary concepts that do not appear in images. These concepts are connected as a graph structure in the knowledge base (*ConceptNet*). "#i" indicates that this is the i-th sentence.

Towards filling this gap, we propose to introduce commonsense from the knowledge base for visual storytelling. In or-

[1] The code is available at https://github.com/lancopku/CVST

der to effectively integrate crucial commonsense from a large-scale knowledge base, we propose a commonsense-driven generative model, which consists of a vision-aware commonsense reasoning module $\mathbf{R}$ and a knowledge-augmented generation module $\mathbf{G}$. Given the input image stream, the module $\mathbf{R}$ first extracts a set of candidate knowledge graphs from the external knowledge base, and further integrate the most informative commonsense via vision-aware directional encoding. Finally, the module $\mathbf{G}$, which is implemented as a semantic-relevance based sentence-level decoder, strives to generate a reasonable and coherent story based on both original semantics of images and introduced commonsense knowledge.

The main contributions of this paper are listed as follows:

- We propose to introduce commonsense knowledge from the external knowledge base to benefit the task of visual storytelling.
- We propose a commonsense-driven generative model, in which the elaborately designed vision-aware directional encoding can effectively integrate the most informative commonsense and semantic-relevance based decoding enhances the coherence of the generated text.
- Experimental results show that our approach can outperform existing methods by a large margin. With the help of introduced commonsense knowledge, the generated stories are more diverse and coherent.

2 Background

The basic architecture of our approach is Seq2Seq model [Sutskever *et al.*, 2014], which consists of an encoder and a decoder. Given the input text $\boldsymbol{x} = (x_1, \cdots, x_m)$, the encoder computes the hidden representation of each word as follows:

$$h_i = \mathrm{GRU}\big(h_{i-1}, e(x_i)\big) \tag{1}$$

where $e(x_i)$ denotes the embedding of the word x_i and GRU refers to the gated recurrent unit [Cho *et al.*, 2014].

Given the hidden representations $(h_1, \cdots, h_m)$, the decoder generates words sequentially. In detail, the hidden state s_t of the decoder at time-step t is computed as follows:

$$s_t = \mathrm{GRU}\big(s_{t-1}, e(y_{t-1}) \oplus c_t\big) \tag{2}$$

where $\oplus$ denotes the vector concatenation, y_{t-1} is the word generated in the previous time step, and c_t is the context vector obtained by the attention mechanism. Readers can refer to [Bahdanau *et al.*, 2014] for more details. Finally, the decoder samples a word y_t from the output probability distribution:

$$y_t \sim \mathrm{softmax}(\mathbf{W}s_t) \tag{3}$$

where $\mathbf{W}$ is a learnable weight matrix.

3 Proposed Model

3.1 Overview

The VST task aims to generate a reasonable and coherent story $\boldsymbol{y} = (y_1, \cdots, y_5)$ with an image stream of 5 ordered images $\boldsymbol{v} = (v_1, \cdots, v_5)$ as input, where v_i and y_i represent the i-th image in the input and the i-th sentence in the output, respectively. Our proposed commonsense-driven generative model is composed of a vision-aware commonsense reasoning module $\mathbf{R}$ and a knowledge-augmented generation module $\mathbf{G}$, whose structures are presented in Figure 2 and Figure 3, respectively.

3.2 Vision-Aware Commonsense Reasoning

The vision-aware commonsense reasoning module $\mathbf{R}$ is responsible for integrating crucial commonsense from the external knowledge base to benefit visual storytelling. Its sketch is shown in Figure 2. For the input image stream, the module $\mathbf{R}$ first infers a set of key concepts depicted in each image. Since the commonsense knowledge of each inferred concept can be represented by its neighboring nodes in the knowledge base, we use each inferred concept as the query to extract all nodes that are directly connected to this concept. These extracted nodes and original inferred concepts form the candidate concept set, which is treated as additional commonsense knowledge to assist in the subsequent generation.

For implementation, we choose to apply a generic object detection model *Clarifai*[2] to extract key objects in the image as the inferred concepts. The external knowledge base is selected as *ConceptNet* [Speer and Havasi, 2012], a semantic network which consists of triples $\mathcal{R} = (\mathtt{h}; \mathtt{r}; \mathtt{t})$ meaning that head concept $\mathtt{h}$ has the relation $\mathtt{r}$ with tail concept $\mathtt{t}$.

However, a tricky problem is that the candidate concept set introduced from *ConceptNet* is miscellaneous, so that it may contain some noise that impairs the model performance. Therefore, we elaborately design a vision-aware directional encoding schema to integrate the most informative commonsense from the candidate concept set. In more detail, for each image v_i, we first apply a convolutional neural network to extract its visual features f_i. Then, a GRU model encodes visual features of all images into dense representations capturing the temporal relationship. The final semantic representation h_i^v of the i-th image is calculated as:

$$f_i = \mathrm{CNN}(v_i) \tag{4}$$

$$h_i^v = \mathrm{GRU}(h_{i-1}^v, f_i) \tag{5}$$

Given the candidate concept set $\{c_{i,1}, \cdots, c_{i,m_i}\}$ corresponding to the i-th image v_i, where m_i represents the total number of candidate concepts corresponding to v_i, we apply the self-attention mechanism to obtain the initial representation $h_{i,t}^c$ of each concept $c_{i,t}$. Formally,

$$h_{i,t}^c = \mathrm{self\_attention}\big(e(c_{i,t}), \boldsymbol{e}_i\big) \tag{6}$$

The above equation indicates that $e(c_{i,t})$ is used as the query to attend to all representations $\boldsymbol{e}_i = \big(e(c_{i,1}), \cdots, e(c_{i,m_i})\big)$, where $e(c_{i,t})$ is the embedding of $c_{i,t}$. Readers can refer to [Vaswani *et al.*, 2017] for the details of self-attention.

For the i-th image, in order to select the most informative commonsense knowledge for generation, we design a vision-aware directional attention to integrate the representations $\{h_{i,t}^c\}_{i=1}^{m_i}$. Since the semantic representation h_i^v of the i-th image v_i characterizes the main semantic content of v_i and also contains the plot information implied in the whole input image stream, we use h_i^v as the query to attentively read

[2] https://clarifai.com/

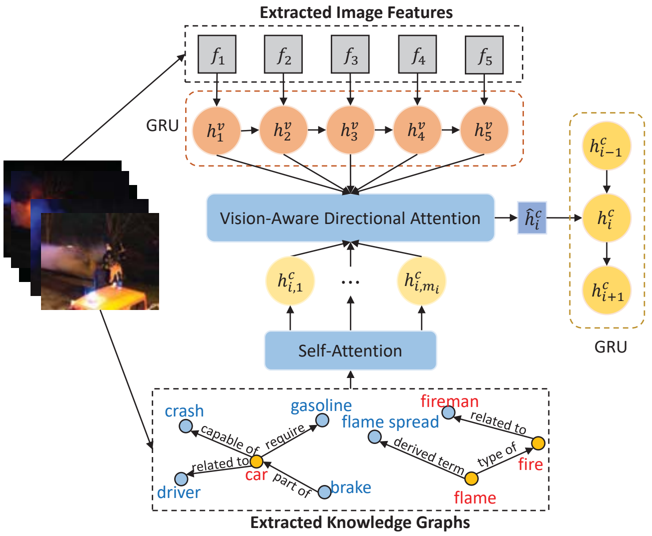


Figure 2: The overview of the commonsense reasoning module $\mathbf{R}$ equipped with the vision-aware directional encoding schema.

the dense representations of all candidate concepts:

$$\beta_{i,t} = (h^v_i)^{\mathrm{T}}\mathbf{W}h^c_{i,t} \tag{7}$$

$$\alpha_{i,t} = \frac{1}{1+e^{-\beta_{i,t}}} \tag{8}$$

$$\hat{h}^c_i = \sum_t \alpha_{i,t}h^c_{i,t} \tag{9}$$

where $\mathbf{W}$ is a learnable weight matrix. It is worth noting that we normalize the attention weights via Eq. (8), which is the explicit formula of *sigmoid* normalization. The reason is that the sum of attention weights of the standard *softmax* normalization is 1, which tends to cause the model to focus only on a few of the most relevant concepts [Kim *et al.*, 2017]. For the paragraph-level target output on the VST task, there are often multiple crucial concepts. Therefore, we utilize *sigmoid* normalization here to more fully extract multiple crucial commonsense from the candidate concept set.

With the vision-aware directional attention, $\hat{h}^c_i$ contains commonsense knowledge that is most relevant to the semantics of the image as well as the plot in the input. Considering the temporal relationship implied in the input image stream, we apply another GRU model to output the final commonsense representation h^c_i of the i-th image:

$$h^c_i = \mathrm{GRU}(h^c_{i-1}, \hat{h}^c_i) \tag{10}$$

3.3 Knowledge-Augmented Generation

The knowledge-augmented generation module $\mathbf{G}$ aims to generate a reasonable and coherent story through semantic-relevance based sentence-level decoding. Figure 3 visually shows its decoding process. Specifically, when $\mathbf{G}$ is generating the i-th sentence, the source information includes three parts: the semantic representation h^v_i and commonsense representation h^c_i of the i-th image, and the previously generated $i-1$ sentences that are concatenated into a word sequence $g_i = (g_{i,1}, \cdots, g_{i,l_i})$. Here l_i is the number of words in the generated $i-1$ sentences. For the first sentence, we mark the generated text as "*null*". In order to integrate the information of the generated text, we apply another GRU to obtain the hidden representation $h^g_{i,t}$ of the t-th word $g_{i,t}$:

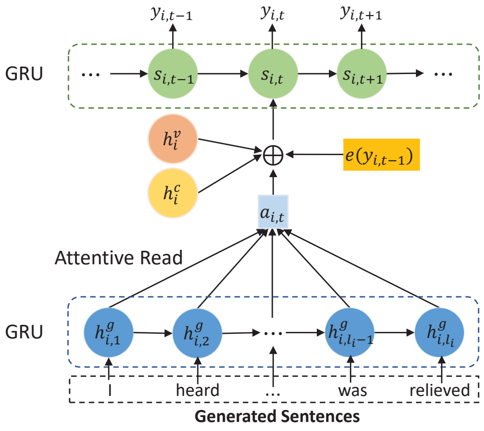


Figure 3: The illustration of the knowledge-augmented generation module $\mathbf{G}$ during the decoding process.

$$h^g_{i,t} = \mathrm{GRU}\big(h^g_{i,t-1}, e(g_{i,t})\big) \tag{11}$$

At time-step t, the decoder implemented as a GRU model takes these three sources of information and the embedding $e(y_{i,t-1})$ of the previously generated word $y_{i,t-1}$ as the input to update its state $s_{i,t}$, which is computed as follows:

$$s_{i,t} = \mathrm{GRU}\big(s_{i,t-1}, h^v_i \oplus h^c_i \oplus a_{i,t} \oplus e(y_{i,t-1})\big) \tag{12}$$

where $\oplus$ denotes the vector concatenation and $a_{i,t}$ is the context vector to allow the decoder to pay different attention [Bahdanau *et al.*, 2014] to different parts of the generated text g_i. Finally, the decoder generates the word $y_{i,t}$ by sampling from the output probability as follows:

$$y_{i,t} \sim \mathrm{softmax}(\mathbf{U}s_{i,t}) \tag{13}$$

where $\mathbf{U}$ is a trainable weight matrix.

We encourage the generated $i-1$ sentences to have higher semantic relevance to the currently generated i-th sentence, which can enhance the coherence of the generated story. We adopt the last hidden representation h^g_{i,l_i} of the previous $i-1$ sentences as the semantic vector v^s_i of the previously generated text, and the last hidden state s_{i,n_i} of the decoder when generating the i-th sentence as the semantic vector v^t_i of the currently generated text. Here n_i is the length of the generated i-th sentence. Following [Ma *et al.*, 2017], we calculate the similarity score between two semantic vectors as follows:

$$\mathcal{S}(v^s_i, v^t_i) = \frac{v^s_i \cdot (v^t_i - v^s_i)}{\|v^s_i\|\|v^t_i - v^s_i\|} \tag{14}$$

Our training objective is to maximize the semantic similarity within the story in addition to maximize the log-likelihood of true parallel data. The final loss function is formulated as:

$$\mathcal{L} = -\sum_{i=1}^{5}\Big(\log\big(p(y_i|h^v_i, h^c_i, \boldsymbol{y}_{<i})\big) + \lambda\mathcal{S}(v^s_i, v^t_i)\Big) \tag{15}$$

where $\boldsymbol{y}_{<i} = (y_1, \cdots, y_{i-1})$ denotes the sequence that consists of the previous $i-1$ sentences. When generating the first sentence, since the generated text marked as "*null*" has no valid meaning, we forcefully constrain $\mathcal{S}(v^s_1, v^t_1) = 0$.

Models	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**ROUGE**	**METEOR**	**CIDEr**
[Huang *et al.*, 2016]	52.2	28.4	14.5	8.1	28.5	31.1	6.4
[Yu *et al.*, 2017]	56.3	31.2	16.4	9.7	29.1	34.2	7.7
[Hsu *et al.*, 2018]	51.9	27.5	14.3	8.3	28.2	30.7	6.3
[Kim *et al.*, 2018]	52.3	28.4	14.8	8.1	28.4	32.4	8.4
[Gonzalez-Rico, 2018]	60.1	36.5	21.1	12.7	29.2	34.4	7.1
[Wang *et al.*, 2018a]	60.5	36.7	20.8	12.5	28.9	33.1	8.5
[Wang *et al.*, 2018b] (XE-ss)	62.3	38.2	22.5	13.7	29.7*	34.8	8.7
[Wang *et al.*, 2018b] (GAN)	62.8	38.8	23.0	14.0	29.5	35.0*	9.0
[Wang *et al.*, 2018b] (AERL)	63.8*	39.1*	**23.2***	**14.1***	29.5	35.0*	9.4*
Proposal	**66.4**	**39.2**	23.1	12.8	**29.9**	**35.2**	**12.1**

Table 1: Automatic evaluation results. The best performance is highlighted in bold and "*" indicates the best result achieved by the baselines.

4 Experiments

4.1 Dataset

We conduct experiments on the VIST dataset [Huang *et al.*, 2016], which consists of 10,117 Flickr albums and 210,819 unique photos. Each sample contains 5 images, each paired with a sentence in the story. We follow the standard split [Wang *et al.*, 2018b] for a fair comparison.

4.2 Settings

We set the batch size to 64 and the vocabulary size is 30,000. The 512-dim word embeddings are learned from scratch. We apply the ResNet-152 [He *et al.*, 2016] pre-trained on the ImageNet to extract visual features. All GRU models are set to two layers, and the hidden size is 512. Except that the decoder is unidirectional, the other GRU models are bidirectional. The parameter λ is set to 0.05. We use the Adam [Kingma and Ba, 2014] optimizer with the initial learning rate 10^{-3}.

4.3 Evaluation Metrics

Automatic evaluation. The automatic evaluation of visual storytelling remains an open and tricky question since this task is highly flexible and stories are very subjective. Therefore, we adopt a combination of multiple evaluation metrics, including BLEU, ROUGE, METEOR, and CIDEr.

Human evaluation. We also conduct human evaluation to more accurately assess the quality of the output. We hire three annotators with the linguistic background to score 200 items, each consisting of the input image stream and stories generated by different systems. The evaluation criteria include the following four aspects: **Fluency** evaluates whether the output is fluent and **relevance** measures how relevant the generated story and input images are. **Informativeness** evaluates whether the output is diverse and valuable and **coherence** assesses whether the output is semantically coherent. We stipulate the score to be an integer from 1 to 5 and the average of scores given by three annotators is reported as the final result.

5 Results and Discussion

5.1 Experiment Results

Table 1 presents the automatic evaluation results, illustrating that our proposed model can outperform the baselines by a large margin and achieves the best performance in almost all metrics. For instance, our approach achieves a 29% relative improvement over the best baseline on the CIDEr score. It demonstrates that external commonsense knowledge is conducive to generating high-quality outputs. The proposed model is able to extract crucial commonsense from the external knowledge base according to the semantics of the image stream. This facilitates the generation of imaginary concepts to make the generated story more human-like.

Models	**Flue.**	**Rele.**	**Cohe.**	**Info.**
[Huang *et al.*, 2016]	3.1	3.8	3.3	3.5
[Yu *et al.*, 2017]	3.2	4.1	3.2	3.7
[Gonzalez-Rico, 2018]	3.5	4.0	3.7*	3.6
[Wang *et al.*, 2018a]	4.2*	4.3	3.1	3.6
[Wang *et al.*, 2018b] (AREL)	4.1	4.4*	3.5	3.8*
Proposal	**4.4**	**4.6**	**4.1**	**4.3**
Human	4.8	4.7	4.3	4.6

Table 2: Human evaluations of different systems. **Flue.**, **Rele.**, **Cohe.**, and **Info.** denotes fluency, relevance, coherence, and informativeness, respectively. For each metric, the averaged Kappa coefficient is greater than 0.5, which ensures inter-annotator agreement. The best performance is highlighted in bold and "*" indicates the best result achieved by the baselines. We select several representative and well-performing baselines to perform comparion.

The human evaluations are shown in Table 2, which shows that our approach can substantially outperform the baselines, especially in terms of coherence and informativeness. For example, compared to the best baseline model, the coherence score increases from 3.7 to 4.1 and the informativeness score increases from 3.8 to 4.3. Our model can perform semantic association based on the introduced commonsense knowledge. This promotes the generation of imaginary concepts that do not appear in images, resulting in the increment in the informativeness score. Besides, the semantic-relevance based sentence-level decoding schema is able to enhance the semantic dependency within the output, thus improving the coherence of the generated story.

5.2 Ablation Study

Here we perform an ablation study to explore the importance of different components. Table 3 shows the relevant results.

Encoding ablation. The encoding ablation aims to explore the importance of different input information. Table 3 shows that "*w/o concept inference*" results in the largest decline in

Models	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**ROUGE**	**METEOR**	**CIDEr**
Full model	66.4	39.2	23.1	12.8	29.9	35.2	12.1
w/o image features	61.8	37.3	21.6	12.6	29.7	34.8	11.3
w/o external commonsense	60.3	35.2	20.6	12.3	29.4	34.4	9.1
w/o concept inference	58.2	32.9	18.1	10.5	29.3	33.6	8.6
w/o semantic similarity	65.9	39.1	22.9	12.8	29.8	34.7	11.9
w/o sentence-level decoding	64.7	38.9	22.6	12.7	29.6	34.3	11.8

Table 3: The automatic evaluation results of ablation study. Encoding ablation includes: "*w/o image features*" meaning that we remove the semantic representation h_i^v of the image in the input, "*w/o external commonsense*" meaning that we use the inferred concepts without extracting other knowledge graphs from the knowledge base, and "*w/o concept inference*" meaning that the inferred concepts and extracted knowledge graphs are all removed. Decoding ablation includes: "*w/o semantic similarity*" meaning that the similarity score $\mathcal{S}$ in Eq. (15) is removed and "*w/o sentence-level decoding*" meaning that the model generates the entire story at once and does not use similarity score.

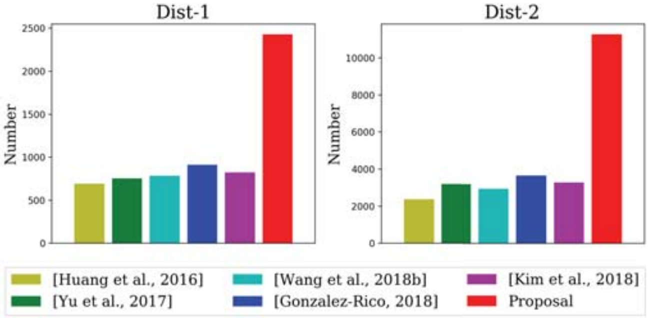


Figure 4: The comparison of diversity among different systems. "Dist-1" and "Dist-2" denote the number of distinct unigrams and bigrams, respectively.

the model performance. This illustrates that the commonsense reasoning module $\mathbf{R}$ is the core component of our model since it equips the model with the capability of commonsense reasoning and semantic association to promote the generation of imaginary concepts. Besides, the semantic representation of the image plays an important role in the generation because it contains the overall information of input and captures the temporal relationship within the image stream.

Decoding ablation. We also perform decoding ablation to explore the role of semantic-relevance scoring and sentence-level decoding. Table 3 shows that both of them can have a positive impact on improving results. However, it is worth noting that the encoding ablation brings a larger decline in the model performance than the decoding ablation in general. This further demonstrates that commonsense knowledge is crucial for visual storytelling, which can do a great favour to generating stories that are more human-like.

5.3 Effectiveness of Improving Diversity

In the experiment, we also find that our model can greatly improve the diversity of the outputs, thus alleviating the problem of duplicate phrases that the VST task is vulnerable to. Figure 4 presents the number of distinct unigrams and bigrams contained in the output of different systems. Results illustrate that our approach can substantially outperform baselines on both metrics. Compared to the baselines, our approach enriches the source information by introducing external commonsense knowledge, which enables the model to generate more diverse and novel expressions.

5.4 Visualization of Directional Attention

Here we visualize the attention weight of each candidate concept to demonstrate the effectiveness of our vision-aware directional attention in the commonsense reasoning module $\mathbf{R}$. Figure 5 presents the attention heatmap, illustrating that our directional attention can effectively extract multiple crucial commonsense concepts that closely surrounds the semantics of the input image stream. Take the first image as an example, our approach first infers the concept "*drink*" depicted in the image, and then extracts related knowledge graphs from *ConceptNet* to form the candidate concept set including "*buy*", "*beer*", and so on. Then, the directional attention is able to integrate the most informative commonsense by automatically assigning larger weights to more important concepts like "*drink*" and "*beer*". Besides, compared to the *softmax* normalization that focuses on only a few concepts, *sigmoid* normalization can pay attention to more concepts that are equally crucial, e.g., "*tea*" and "*cup*", yielding representation containing more useful commonsense.

5.5 Case Study

Table 4 presents outputs of different systems with the image stream in Figure 5 as input. Here we compare our approach with the most representative baseline [Huang *et al.*, 2016] and the state-of-the-art system [Wang *et al.*, 2018b]. As shown in Table 4, [Huang *et al.*, 2016] generates the output that is not fluent and contains duplicate words, e.g., "*We had a fire pit fire*". Although [Wang *et al.*, 2018b] improves fluency to some extent, it still tends to generate sentences that are too simple. In contrast, our approach not only improves the fluency, but also generates a more diverse and expressive story that contains plenty of novel imaginary concepts like "*oven*", "*porch*", and so on. According to Figure 5, "*porch*" can be regarded as commonsense about "*light*" since these two concepts are connected in a knowledge graph. With the help of the extracted commonsense knowledge, our model is capable of generating these imaginary concepts during decoding, resulting in the output that is more novel and diverse.

6 Related Work

Visual storytelling. [Huang *et al.*, 2016] is the first to propose this task and constructs a large-scale dataset. Then, the subsequent endeavors are mainly divided into two categories. One line of research focuses on elaborately designing

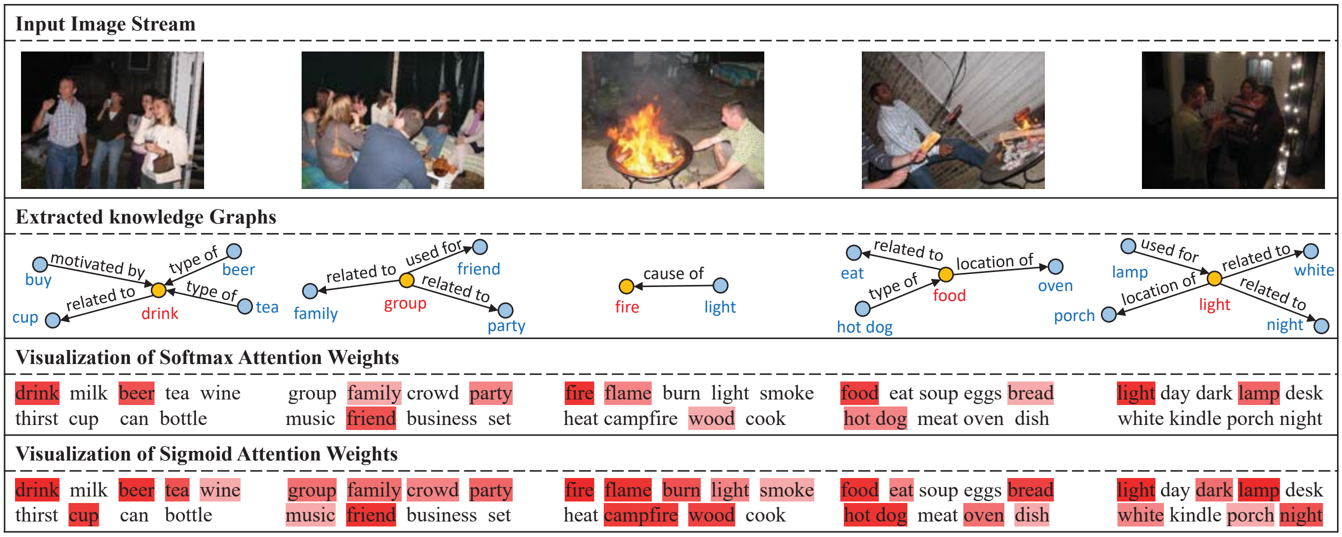


Figure 5: Visualization of vision-aware directional attention weight of each candidate concept. Darker color means larger weight. Due to the space limitation, here we only show part of the candidate concepts.

[Huang *et al.*, 2016]: *#1: The family had a great time at the house. #2: We had a lot of fun at the party. #3: We had a fire pit fire. #4: There were many different types of food. #5: [male] was very happy to be there.*

[Wang *et al.*, 2018b] (AREL): *#1: There were a lot of people there. #2: The whole family was there to celebrate. #3: The man is getting ready to go fire. #4: They had a lot of food for the party. #5: At the end of the night, everyone was very happy.*

Proposal: *#1: A group of friends decided to have a nice dinner together and bought a lot of drinks. #2: They talked about things and enjoyed the beer and the tea in the cup. #3: They lit a fire in the house. #4: They put hot dogs on the oven and enjoyed food. #5: The lamp on the porch had white light and they had a great time that night.*

Human-Written: *#1: It was our first big backyard barbeque of summer and we invited all friends. #2: We all sat around and caught up with each others' lives. #3: Dave started the fire pit, look at those flames! #4: Everyone put hot dogs on skewers and roasted them over the fire. #5: We all had a great time hanging out until very late in the night and it was a great party!*

Table 4: The output of different systems with the image stream in Figure 5 as input. Underlined words are concepts that are not clearly expressed in images. Red words are concepts inferred from images and blue words are connected concepts in the knowledge base.

specific architectures to improve results. For instance, [Hsu *et al.*, 2018] presents an inter-sentence diverse beam search approach and [Kim *et al.*, 2018] combines global-local attention and context cascading mechanism. [Gonzalez-Rico, 2018] set separate decoders for different images to more differentiated visual information. However, these approaches are trained by the MLE method, which tends to result in pattern-stiff outputs. Another line strives to generate more expressive outputs via adversarial training or reinforcement learning. For example, both [Chen *et al.*, 2017] and [Wang *et al.*, 2018a] adopts the adversarial training, while [Wang *et al.*, 2018b] utilizes inverse reinforcement learning to learn the implicit reward function. However, the training of these methods is unstable and sensitive to hyper-parameters.

Narrative story generation. Our work is also related to story generation, which aims to generate a story based on the text description of an event. [Jain *et al.*, 2017] explores story generation via statistical machine translation models. Furthermore, a hierarchical generation model is presented in [Lewis *et al.*, 2018] to generate stories from prompts. To improve the coherence, [Xu *et al.*, 2018] applies reinforcement learning to extract a skeleton of the story while [Yao *et al.*, 2018] presents two planning strategies to fully leverage storyline. However, different from plain text-based narrative story generation, visual storytelling involves the understanding of image and thereby faces more serious challenges.

7 Conclusion

This work presents a commonsense-driven generative model, which aims at introducing commonsense from external knowledge base to benefit visual storytelling. The proposed model employs vision-aware directional encoding to effectively integrate the most informative commonsense and enhances the coherence of the output via semantic-relevance based sentence-level decoding. The experiments show that our approach can outperform existing methods by a large margin. With the help of additional commonsense and semantic-relevance based decoding, the generated stories are more diverse and coherent.

Acknowledgements

We thank the anonymous reviewers for their thoughtful comments. This work was supported in part by National Natural Science Foundation of China (No. 61673028). Xu Sun is the corresponding author of this paper.

References

[Bahdanau *et al.*, 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[Chen *et al.*, 2017] Zhiqian Chen, Xuchao Zhang, Arnold P. Boedihardjo, Jing Dai, and Chang-Tien Lu. Multimodal storytelling via generative adversarial imitation learning. In *IJCAI*, pages 3967–3973, 2017.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, pages 1724–1734, 2014.

[Gonzalez-Rico, 2018] Diana Gonzalez-Rico. Contextualize, show and tell: A neural visual storyteller. *arXiv preprint arXiv:1806.00738*, 2018.

[Guan *et al.*, 2018] Jian Guan, Yansen Wang, and Minlie Huang. Story ending generation with incremental encoding and commonsense knowledge. *arXiv preprint arXiv:1808.10113*, 2018.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Hsu *et al.*, 2018] Chao-Chun Hsu, Szu-Min Chen, Ming-Hsun Hsieh, and Lun-Wei Ku. Using inter-sentence diverse beam search to reduce redundancy in visual storytelling. *arXiv preprint arXiv:1805.11867*, 2018.

[Huang *et al.*, 2016] Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross B. Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. Visual storytelling. In *NAACL-HIT*, pages 1233–1239, 2016.

[Huang *et al.*, 2018] Qiuyuan Huang, Zhe Gan, Asli Celikyilmaz, Dapeng Wu, Jianfeng Wang, and Xiaodong He. Hierarchically structured reinforcement learning for topically coherent visual story generation. *arXiv preprint arXiv:1805.08191*, 2018.

[Jain *et al.*, 2017] Parag Jain, Priyanka Agrawal, Abhijit Mishra, Mohak Sukhwani, Anirban Laha, and Karthik Sankaranarayanan. Story generation from sequence of independent short descriptions. *arXiv preprint arXiv:1707.05501*, 2017.

[Kim *et al.*, 2017] Yoon Kim, Carl Denton, Luong Hoang, and Alexander M Rush. Structured attention networks. *arXiv preprint arXiv:1702.00887*, 2017.

[Kim *et al.*, 2018] Taehyeong Kim, Min-Oh Heo, Seonil Son, Kyoung-Wha Park, and Byoung-Tak Zhang. Glac net: Glocal attention cascading networks for multi-image cued story generation. *arXiv preprint arXiv:1805.10973*, 2018.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Lewis *et al.*, 2018] Mike Lewis, Yann Dauphin, and Angela Fan. Hierarchical neural story generation. In *ACL*, pages 889–898, 2018.

[Ma *et al.*, 2017] Shuming Ma, Xu Sun, Jingjing Xu, Houfeng Wang, Wenjie Li, and Qi Su. Improving semantic relevance for sequence-to-sequence learning of chinese social media text summarization. In *ACL*, pages 635–640, 2017.

[Mostafazadeh *et al.*, 2016] Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. Generating natural questions about an image. In *ACL*, 2016.

[Speer and Havasi, 2012] Robert Speer and Catherine Havasi. Representing general relational knowledge in conceptnet 5. In *LREC*, pages 3679–3686, 2012.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112, 2014.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 6000–6010, 2017.

[Wang *et al.*, 2018a] Jing Wang, Jianlong Fu, Jinhui Tang, Zechao Li, and Tao Mei. Show, reward and tell: Automatic generation of narrative paragraph from photo stream by adversarial training. In *AAAI*, pages 7396–7403, 2018.

[Wang *et al.*, 2018b] Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. No metrics are perfect: Adversarial reward learning for visual storytelling. In *ACL*, pages 899–909, 2018.

[Xu *et al.*, 2018] Jingjing Xu, Xuancheng Ren, Yi Zhang, Qi Zeng, Xiaoyan Cai, and Xu Sun. A skeleton-based model for promoting coherence among sentences in narrative story generation. In *EMNLP*, pages 4306–4315, 2018.

[Yang *et al.*, 2018] Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. SGM: sequence generation model for multi-label classification. In *COLING*, pages 3915–3926, 2018.

[Yao *et al.*, 2018] Lili Yao, Nanyun Peng, Weischedel Ralph, Kevin Knight, Dongyan Zhao, and Rui Yan. Plan-and-write: Towards better automatic storytelling. *arXiv preprint arXiv:1811.05701*, 2018.

[Yu *et al.*, 2017] Licheng Yu, Mohit Bansal, and Tamara L. Berg. Hierarchically-attentive RNN for album summarization and storytelling. In *EMNLP*, pages 966–971, 2017.