

Human-like Storyteller: A Hierarchical Network with Gated Memory for Visual Storytelling

Lu Zhang^{1,2*}[0000-0001-9693-1122], Yawei Kong^{1,2*}[0000-0001-8823-866X], Fang Fang², Cong Cao^{2**}[0000-0003-1881-1947], Yanan Cao²[0000-0003-3534-1094], Yanbing Liu²[0000-0002-9653-073X], and Can Ma²

¹ Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

² School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

{zhanglu0101,kongyawei,fangfang0703,caocong,caoyanan,
liuyanbing,macan}@iie.ac.cn

Abstract. Different from the visual captioning that describes an image concretely, the visual storytelling aims at generating an imaginative paragraph with a deep understanding of the given image stream. It is more challenging for the requirements of inferring contextual relationships among images. Intuitively, humans tend to tell the story around a central idea that is constantly expressed with the continuation of the storytelling. Therefore, we propose the Human-Like StoryTeller (HLST), a hierarchical neural network with a gated memory module, which imitates the storytelling process of human beings. First, we utilize the hierarchical decoder to integrate the context information effectively. Second, we introduce the memory module as the story’s central idea to enhance the coherence of generated stories. And the multi-head attention mechanism with a self adjust query is employed to initialize the memory module, which distills the salient information of the visual semantic features. Finally, we equip the memory module with a gated mechanism to guide the story generation dynamically. During the generation process, the expressed information contained in memory is erased with the control of the read and write gate. The experimental results indicate that our approach significantly outperforms all state-of-the-art (SOTA) methods.

Keywords: Visual Storytelling · Central Idea · Gated Memory.

1 Introduction

Recently, the tasks of combining vision and text have made a great stride, such as the high-profile visual captioning [3], whose purpose is to generate literal descriptions based on the images or the videos. To further investigate the model’s capabilities in generating structured paragraphs under more complicated scenarios, visual storytelling has been proposed by [9]. This task aims to generate a

* Equal contribution

** Corresponding author

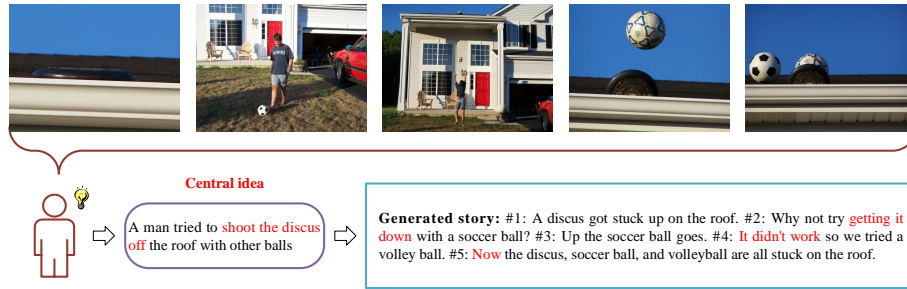


Fig. 1. An example of visual storytelling. “# i ” indicates that this is the i -th sentence.

coherent and expressive story with a given temporal image stream, which not only expects an intuitive understanding of image ground content but also requires plentiful emotion as well as imagination. It is more challenging since the model must have the capabilities of inferring contextual relationships that are not explicitly depicted in the images.

Encouraged by the success of the Seq2Seq model in visual captioning, most visual storytelling methods usually employ this typical framework that consists of a visual encoder and a sentence decoder. The visual encoder transforms the image stream to feature vectors and then the sentence decoder generates every storyline. Based on the Seq2Seq framework, Kim et al. [11] and Jung et al. [10] optimize the specific architecture with maximum likelihood estimation (MLE) method; Wang et al. [22] and Hu et al. [7] employ reinforcement learning or adversarial training strategies to improve performance; Yang et al. [23] and Li and Li [12] focus on generating more expressive results by drawing into external knowledge or performing additional processing on the dataset. Although progresses have been made, the generated story still lacks centrality and have lots of semantic repetition, which significantly reduces the coherence and readability.

Intuitively, the contextual description of a story will revolve around a central idea. As shown in Figure 1, all the contents of the five storylines are related to the central idea - “A man tried to shoot the discus off the roof with other balls”. If there is no guidance of it, the second sentence may be about “playing football” instead of “getting it down with a soccer ball”. The former just depicts the intuitive content of the image, resulting in the incoherence of the context. Therefore, it is critical to model the central idea during the storytelling process.

Towards filling these gaps, we propose the Human-Like StoryTeller (HLST), a hierarchical neural network with the gated memory module. First, considering the importance of the context in generating coherent stories, we introduce a hierarchical decoder that includes the narration decoder and the sentence decoder. The narration decoder constructs a semantic concept for guiding the sentence decoder to generate every storyline, which makes the generating process in chronological order rather than parallel. Second, we utilize the multi-head attention mechanism with a self adjust query to obtain global memory as our central idea. The self adjust query questions the model “What is the story about?” and

the attention mechanism solves it by focusing on different pieces of salient information within the visual semantic features. Then, the model grasps the central idea to generate more coherent and relevant story. Finally, we equip the memory module with a gated mechanism to guide the story generation dynamically. The memory information is gradually erased under the control of the read and write gate, which improves the story’s diversity and informativeness. We conduct the experiments on the VIST dataset and the results show that HLST significantly outperforms all baseline models in terms of automatic evaluations and human judgments. Further qualitative analysis indicates that the generated stories by HLST are highly coherent with human understanding.

Our main contributions are summarized as follows:

- To our knowledge, we are the first one to introduce the concept of central idea to benefit the task of visual storytelling.
- We propose the memory module as the central idea to enhance the coherence of stories. It guides the generation process of the hierarchical decoder that integrates the contextual information effectively.
- We equip the memory module with a gated mechanism to dynamically express the central idea. The gated mechanism is conducive to generate more informative stories by removing redundant information.
- Our approach achieves state-of-the-art (SOTA) results, in terms of automatic metrics and human judgments. By introducing the central idea, the generated stories are more coherent and diverse.

2 Related work

In early visual to language tasks, visual captioning task achieves impressive results [20]. Generally, most visual captioning models utilize the CNN to extract the features of the image or video and send them to a decoder for generating a sentence caption. Take one step further, the visual storytelling is expected to generate an expressive and coherent paragraph with a temporal image stream instead of a single image. Notably, this task is more difficult because it not only focuses on the objective descriptions of the visual objects but also requires to consider the contextual coherence with a deeper understanding of the inputs.

Park and Kim [16] has made pioneering research to explore the visual storytelling task, which retrieves a sequence of natural sentences for an image stream. For the better development of this field, Huang et al. [9] releases a more compatible and sophisticated dataset, named VIST. The VIST is composed of visual-story pairs, in which each item contains five images and the corresponding sentences. In addition, they first employ the Seq2Seq framework to generate stories, which naturally extends the single-image captioning technique of [3] to multiple images. Hence, the subsequent endeavors are concentrating on improving the specific architectures. Kim et al. [11] and Gonzalez-Rico [4] aim at incorporating the contextual information. To alleviate the repetitiveness, Hsu [6] proposes the inter-sentence diverse beam search algorithm. Furthermore, some researchers [2, 7, 21] strive to incorporate reinforcement learning with rewards or

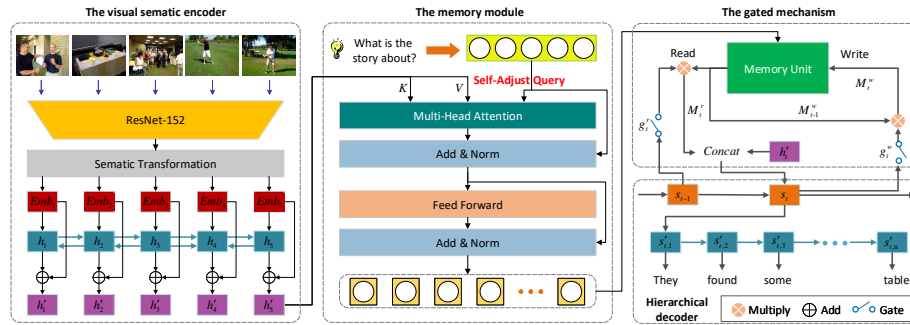


Fig. 2. Model overview. The detailed HLST includes three important modules: a visual semantic encoder, a gated memory module and a hierarchical decoder.

adversarial training strategies for generating more relevant stories. And other studies [8, 12, 23] are based on drawing into external knowledge or preprocessing data to improve performance. More specifically, Wang et al. [22] first implements a baseline model (XE-ss) as a policy model, which employs a Seq2Seq framework to generate storylines in parallel. They further propose the adversarial reward learning (AREL) to learn an implicit reward for optimizing the policy model.

Intuitively, people usually grasp global information as a central idea and tell the story around it. Hence, we propose HLST, which consists of a gated memory module and a hierarchical decoder. We utilize the multi-head attention [18] with a self adjust query to initialize the memory as the central idea. Inspired by [15, 17], we update the memory unit with the gated mechanism to dynamically express the central idea as human beings.

3 Our Approach

In this section, we introduce our Human-Like StoryTeller (HLST) model detailly. As shown in Figure 2, HLST is composed of three modules: a visual semantic encoder, a gated memory module and a hierarchical decoder. Given five images $V = (v_1, \dots, v_5)$ in order, the visual semantic encoder obtains the semantic vectors $H' = (h'_1, \dots, h'_5)$ by integrating the individual and contextual features. Then, we utilize a hierarchical decoder to strengthen the contextual relevance between sentences. It is composed of a narration decoder and a sentence decoder. Moreover, the narration decoder constructs the high-level sentence representations for $S = (s_1, \dots, s_5)$. And the sentence decoder generates a word sequence $W = (w_{t,1}, w_{t,2}, \dots, w_{t,n}), w_{t,j} \in \mathbb{V}$ in chronological order based on the corresponding sentence representation s_t . Here, \mathbb{V} is the vocabulary of all output tokens. Moreover, we first introduce the memory module M in this field, which acts as the central idea to enhance coherence in story generation. Specifically, we employ the multi-head attention with a self adjust query to distil the salient information within H' . To further dynamically express the central idea, we equip

the memory module with the gated mechanism that consists of a read gate and a write gate. With the continuation of the generation process, the information contained in M is eliminated gradually. In the following, we will describe these three parts details. For simplicity, we omit all bias in formulations.

3.1 Visual Semantic Encoder

The visual semantic encoder consists of a pre-trained CNN layer, a semantic transformation layer and a Bidirectional Gated Recurrent Units (BiGRU) layer. Given an image stream $V = (v_1, \dots, v_5)$, $v_i \in \mathbb{R}^{d_v}$, the CNN layer is responsible for extracting the visual features. Then, we map the visual representations into the semantic space to obtain semantic features $f_i \in \mathbb{R}^{d_f}$, $i \in [1, 5]$ with a linear transformation layer. Here, d_v and d_f is the dimension of the visual and semantic features, respectively. The BiGRU layer further encodes the semantic features as the context vectors $h_i = [\overleftarrow{h}_i; \overrightarrow{h}_i]$, $h_i \in \mathbb{R}^{d_h}$, which integrates the results of the forward and backward calculations. Here, d_h is the number of hidden units. Furthermore, since each sentence in the generated stories corresponds to the specific image, we strengthen the influence of the corresponding image features through a skip connection. The final semantic features h'_i at time-step i is computed as follows:

$$\begin{aligned} f_i &= W_1 \cdot \text{CNN}(v_i) \\ h_i &= \text{BiGRU}(h_{i-1}, f_i) \\ h'_i &= W_3(h_i \oplus W_2 f_i) \end{aligned} \quad (1)$$

where \oplus represents the vector concatenation and \cdot denotes the matrix multiplication. $W_1 \in \mathbb{R}^{d_f \times d_v}$, $W_2 \in \mathbb{R}^{d_h \times d_f}$ and $W_3 \in \mathbb{R}^{d_h \times 2d_h}$ are the learnable linear transformation matrices.

3.2 Hierarchical Decoder

Different from many existing works that only use a sentence decoder, we employ the hierarchical decoder to integrate the contextual information effectively. It is composed of a narration decoder and a sentence decoder. The narration decoder is an unidirectional GRU that constructs the sentence representations $s_t \in \mathbb{R}^{d_s}$, $t \in [1, 5]$. Here, d_s is the number of hidden state units. At each time-step t , the corresponding encoder output h'_t and previous sentence representation s_{t-1} are fed into the narration decoder for calculating s_t . Notably, s_t integrates the information of the generated sentences effectively for taking s_{t-1} as input. Meanwhile, the sentence decoder predicts the next word $w_{t,j}$ based on the s_t and the previous generated word $w_{t,j-1} \in \mathbb{R}^{d_e}$, where d_e is the dimension of the word embedding. The whole generation process can be described as follows:

$$\begin{aligned} s_t &= \text{GRU}_n(s_{t-1}, h'_t) \\ s'_{t,j} &= \text{GRU}_s(s'_{t,j-1}, W_4(s_t \oplus w_{t,j-1})) \end{aligned} \quad (2)$$

where $W_4 \in \mathbb{R}^{d_s \times (d_s + d_e)}$ is a learnable projection matrix. Note that GRU_n and GRU_s represent the narration decoder and sentence decoder, respectively. Besides, the sequential structure GRU_n makes the whole generation process in chronological order. And the $s'_{t,j} \in \mathbb{R}^{d_s}$ represents the j^{th} hidden state of the sentence decoder at t^{th} sentence, which is utilized to compute the word probability distribution over the whole vocabulary:

$$p_\theta(w_{t,j}|w_{t,j-1}, s_t, \mathbf{v}) = \text{Softmax}(\text{MLP}(s'_{t,j})) \quad (3)$$

where $\text{MLP}(\cdot)$ represents the multi-layer perception that projects the $s'_{t,j}$ to the vocabulary size.

3.3 Multi-head attention with self adjust query

The multi-head attention mechanism distills the central idea by attending to the visual semantic features $H' = \{h'_1, \dots, h'_n\}$. However, different from the traditional method, we introduce a self adjust query that is a learnable variable trained with the model. As illustrated in Figure 2, the self adjust query is equivalent to put a question to the model - “*What is the story about?*”. Through the interaction between the self adjust question and visual semantic features, we get the question-aware memory $M_{init} \in \mathbb{R}^{d_h}$ as the central idea, which contains several pieces of salient semantic information. The process is formulated as follows:

$$\begin{aligned} M_{init} &= \text{MHA}(Q^A, H', H') \\ \text{MHA}(Q^A, H', H') &= \text{Concat}(\text{head}_1, \dots, \text{head}_N)W^O \\ \text{head}_j &= \text{Softmax}\left(\frac{Q^A W_j^Q (H' W_j^K)^T}{\sqrt{d_k}}\right) H' W_j^V \end{aligned} \quad (4)$$

where $W_j^Q, W_j^K \in \mathbb{R}^{d_h \times d_k}$, $W_j^V \in \mathbb{R}^{d_h \times d_v}$ and $W^O \in \mathbb{R}^{N d_v \times d_h}$ are learnable linear transformation matrices. N is the head number and $d_k = d_v = d_h/N$ in this paper. And $Q^A \in \mathbb{R}^{d_h}$ is the self adjust query vector, which is initialized randomly. Obviously, the attention mechanism grasps the global semantic information, which is utilized to guide the generation process and improves the coherent of generated stories.

3.4 Gated memory mechanism

To dynamically control the expression of the story’s central idea, we further equip the memory module with gated mechanism that includes a read and a write gate. At different decoding steps, each sentence requires various information to be expressed. Therefore, the read gate is responsible for reading the currently needed content from the memory unit and feeds it to the narration decoder. Then, to avoid elaborating the duplicated information, we employ the write gate to update the memory unit with the current hidden state of the narration

decoder, which leads to a decline of the information stored in memory as the decoding processes. We further conduct the ablation experiments to verify the effectiveness of each gate.

At the high-level decoding time t , the read gate $g_t^r \in \mathbb{R}^{d_h}$ is computed with the previous state s_{t-1} of the narration decoder, while the write gate $g_t^w \in \mathbb{R}^{d_h}$ is calculated with the current state s_t . The two gates can be formulated as follows:

$$g_t^r = \sigma(W_r s_{t-1}), \quad g_t^w = \sigma(W_w s_t) \quad (5)$$

where $W_r \in \mathbb{R}^{d_h \times d_s}$ and $W_w \in \mathbb{R}^{d_h \times d_s}$ are learnable parameters. The $\sigma(\cdot)$ is the sigmoid nonlinear activation function and the output value ranges from 0 to 1. Next, the read and write gates are used to update the memory unit as follows:

$$M_t^r = g_t^r \odot M_{t-1}^w, \quad M_t^w = g_t^w \odot M_{t-1}^w \quad (6)$$

Here, \odot denotes the element-wise multiplication and all $M \in \mathbb{R}^{d_h}$. Besides, M_{t-1}^w is the memory contents written back at the previous time-step, which is updated to M_t^w by the write gate g_t^w . And M_t^r is read from M_{t-1}^w with the read gate g_t^r and then is fed into the narration decoder with the encoder output h'_t to compute the current hidden state s_t . Finally, we modify the Equation (2) as follows:

$$s_t = \text{GRU}(s_{t-1}, W_m(h'_t \oplus M_t^r)) \quad (7)$$

where $W_m \in \mathbb{R}^{d_h \times 2d_h}$ is a linear transformation matrix. It is worth noting that the Equation (6) is executed several times, which is equivalent to continuously multiplying a matrix between $[0, 1]$. Therefore, the expressed information contained in memory vectors M is gradually decreasing in the decoding process, which is similar to the central idea expressed completely as human beings.

4 Experiments

4.1 Dataset

We conduct experiments on the VIST dataset, which is the most popular dataset for the visual storytelling task [9]. In detail, the dataset includes 10,117 Flickr albums with 210,819 unique images. Each story consists of five images and their corresponding descriptions. For the fair comparison, we follow the same experimental settings as described in [10, 22]. Finally, we obtain 40,098 training, 4,988 validation and 5,050 testing samples after filtering the broken images.

4.2 Evaluation Metrics

To evaluate our model comprehensively, we adopt both automatic evaluations and human judgments. Four different automatic metrics are utilized to measure our results, including BLEU [14], ROUGE [13], METEOR [1] and CIDEr [19]. For a fair comparison, we employ the open-source evaluation code³ as [22, 24].

³ https://github.com/lichengunc/vist_eval

Since automatic evaluation metrics may not be completely consistent with human judgments, we also invite 6 annotators with the corresponding linguistic background to conduct human evaluations as in [23]. We randomly sample 200 stories from the test dataset, of which each example consists of the image stream and the generated story of different models. We also evaluate the human-generated stories for comparison. Then, all annotators score the results from 1 to 5 in the four aspects: fluency, coherence, relevance and informativeness. In detail, **fluency** mainly evaluates whether the output is grammatically fluent, while **coherence** measures the semantic similarity between sentences. **Relevance** represents the correlation between the generated story and the images. **Informativeness** measures the diversity and richness of outputs.

4.3 Baseline Models

We mainly compare our model with the following representative and competitive frameworks.

Seq2Seq [9] introduces the first dataset for the visual storytelling task and proposes a simple Seq2Seq model. **HARAS** [24] first selects the most representative photos and then composes a story for the album. **GLAC Net** [11] aims to combine global-local (glocal) attention and context cascading mechanisms. **SRT** [21] utilizes reinforcement learning and adversarial training to train the model better. **XE-ss** [22] is a typical encoder-decoder framework that generates story sentences parallel. **GAN** [22] incorporates generative adversarial training based on the XE-ss model. **AREL** [22] learns an implicit reward function and then optimizes policy search with the learned reward function. **XE-TG** [12] exploits textual evidence from similar images to generate coherent and meaningful stories. **HSRL** [8] employs the hierarchical framework trained with the reinforcement learning. **Knowledge** [23] extracts a set of candidate knowledge graphs from the knowledge base and integrates in the attention model. **ReCo-RL** [7] introduces three assessment criteria as the reward function. **INet** [10] proposes a hide-and-tell model that can learn non-local relations across the photo streams.

4.4 Training Details

We extract the image features with a pre-trained ResNet-152 Network proposed by [5], which is widely used in visual storytelling. The feature vector of each image is obtained from the fully-connected (fc) layer that has 2,048 dimensions. Besides, the vocabulary contains the words that occur more than three times in the training set, of which the final size is 9,837. The head number of multi-head attention is set to 4. During training, we set the batch size to 64, and the dimensions of hidden units are all set to 512. We use the Adam optimizer with initial learning rate 10^{-4} to optimize the model. At test time, our stories are produced using the beam search algorithm with the beam size 4. We implement and run all models on a Tesla P4 GPU card with PyTorch⁴.

⁴ <https://pytorch.org/>

5 Results and Discussion

5.1 Automatic Evaluation

Table 1. Automatic evaluation results on the VIST dataset. B: BLEU

Models	B1	B2	B3	B4	ROUGE	METEOR	CIDEr
Seq2Seq (Huang et al. 2016)†	52.2	28.4	14.5	8.1	28.5	31.1	6.4
HARAS (Yu et al. 2017)†	56.3	31.2	16.4	9.7	29.1	34.2	7.7
GLAC Net (Kim et al. 2018)†	52.3	28.4	14.8	8.1	28.4	32.4	8.4
SRT (Wang et al. 2018)†	60.5	36.7	20.8	12.5	28.9	33.1	8.5
XE-ss (Wang et al. 2018)	62.3	38.2	22.5	13.7	29.7	34.8	8.7
GAN (Wang et al. 2018)	62.8	38.8	23.0	14.0	29.5	35.0	9.0
AREL(Wang et al. 2018)	63.8	39.1	23.2	14.1	29.5	35.0	9.4
XE-TG (Li and Li 2019)	-	-	-	-	30.0	35.5	8.7
HSRL (Huang et al. 2019)	-	-	-	12.3	30.8	35.2	10.7
Knowledge (Yang et al. 2019)	66.4	39.2	23.1	12.8	29.9	35.2	12.1
ReCo-RL (Hu et al.2020)	-	-	-	14.4	30.1	35.2	6.7
INet (Jung et al. 2020)	64.4	40.1	23.9	14.7	29.7	35.6	10.0
Our Model (HLST)	67.7	42.6	25.3	15.2	30.8	36.4	11.3

Table 1 gives the automatic evaluation results on the VIST dataset. The results of all baselines are taken from the corresponding papers and the token “†” means that the results are achieved by [23]. The best performance is highlighted in bold and the results show that our approach significantly outperforms other methods. As shown in Table 1, HLST achieves the best performance on almost all automatic metrics. In detail, all BLEU scores have been improved significantly. Moreover, BLEU2 and BLEU3 exceed the highest score of baseline models by 2.5 and 1.4, respectively. The BLEU metric calculates the n-gram matching degree similarity of the reference and the candidate text and the higher-order BLEU can measure the sentence similarity to some extent. Therefore, the improvements in BLEU scores suggest that HLST is able to generate more coherent and informative stories as human beings. Besides, ROUGE-L and METEOR metrics both achieve state-of-the-art results. The ROUGE-L prefers to measure the recall rate between the ground truth and generated stories. And Huang et al. [9] turn out that the METEOR metric correlates best with human judgments on this task. Hence, the improvements indicate that our HLST can generate high-quality stories. In addition, HLST performs slightly worse than [23] on CIDEr since they equip their model with external knowledge base. And Wang et al. [22] empirically find that the references to the same image sequence are photostream different from each other while CIDEr measures the similarity of a sentence to the majority of the references. Hence, CIDEr may not be suitable for this task. In a word, without any external strategies, such as data preprocessing, knowl-

edge graphs and reinforcement learning, our HLST achieves the best results just by integrating the hierarchical decoder and the gated memory module.

5.2 Human Evaluation

Table 2. The human evaluation results.

Models	Fluency	Relevance	Coherence	Informativeness
XE-ss	4.19	3.72	2.88	2.94
AREL	4.23	4.19	3.25	3.18
HLST	4.46	4.58	4.17	4.24
Human	4.65	4.72	4.41	4.59

The human evaluation results are shown in Table 2, which are calculated by averaging all scores from 6 annotators. Obviously, our model significantly exceeds all baselines, especially in coherence and informativeness. For example, compared to the AREL, the coherence score increases from 3.25 to 4.17, and the informativeness score increases from 3.18 to 4.24. The coherence measures whether the output is semantically coherent, while the informativeness evaluates the diversity of the generated stories. Therefore, the high scores of coherence and informativeness suggest that the memory module promotes contextual coherence and the gated mechanism enhances the diversity by reducing semantic duplication. Furthermore, the results are very close to the human’s, which suggests that HLST can generate more informative and coherent stories as humans.

5.3 Ablation Study


Table 3. The automatic evaluation results of the ablation study. B: BLEU

Models	B1	B2	B3	B4	ROUGE	METEOR	CIDEr
Basic model	62.1	38.2	22.5	13.7	29.9	35.3	8.1
+ hierarchical decoder	63.4	39.3	23.3	14.2	30.1	35.5	8.7
+ memory	66.3	41.1	24.2	14.5	30	36	9.8
+ read gate	67.0	41.7	24.5	14.4	29.9	35.7	10
+ write gate (HLST)	67.7	42.6	25.3	15.2	30.8	36.4	11.3

To investigate the correctness and effectiveness of different modules, we conduct the ablation study and the results are listed in Table 3. The token “+” indicates that we add the corresponding module to the model. Note that our

basic model is XE-ss, on which we add the hierarchical decoder, the memory module and the gated mechanism sequentially. We first verify the correctness of the hierarchical decoder. As shown in Table 3, the model with the hierarchical decoder achieves better results. Taking the BLEU1 for an example, the score exceeds the basic model by 1.3 points. Therefore, the model is able to generate more fluent and relevant stories by introducing the hierarchical decoder. We further analyze the effectiveness of the memory module by employing the multi-head attention with self adjust query. And the model with the memory module outperforms the former significantly. The score of BLEU1 is improved to 66.3, which is 2.9 points higher than the model only with a hierarchical decoder. And the results of METEOR and CIDEr are both greatly improved. It suggests that the memory module is important to generate fluent, relevant and coherent stories for integrating the central idea. Furthermore, we conduct experiments to explore the influence of the gated mechanism by sequentially adding the read gate and write gate. The read gate enables the model to read the currently needed information by removing the expressed message. And the improvement of automatic evaluations verifies its effectiveness. Finally, we equip the memory module with both the read and write gate to further filter out the expressed information. Compared with the Basic model, all BLEU scores have increased by about 10%, and CIDEr has increased by 39.5%. This further demonstrates that the central idea is crucial for visual storytelling, which can do a great favour to generating coherent and diverse stories.

5.4 Qualitative Analysis



AREL	It was a beautiful day for the wedding.	It was a beautiful day.	The bride and groom cut the cake.	The bride and groom were very happy to be married.	The bride and groom were so happy to be married.
HLST	It was a beautiful day for the bride and groom.	The bride and groom walked down the aisle to the wedding.	The bride and groom cut the cake, the wedding cake.	They were very happy to be married.	After the ceremony, everyone got together for a picture.
Human	[male] and [female] 's wedding.	All the bridal maids gather for a picture.	After they tie the knot, it 's time to cut some cake.	The lovely couple together.	The dance party starts, what a great event.

Fig. 3. An example of qualitative comparison. We mainly compare our model with the competitive baseline AREL and the human annotations.

We take out an image stream from the test dataset and compare the story’s qualities from AREL, HLST and the ground truth for qualitative analysis. As shown in Figure 3, the image stream depicts a story about the wedding. Clearly,

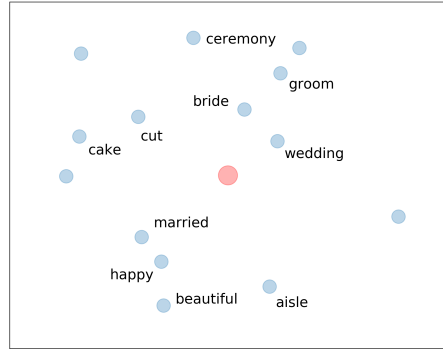


Fig. 4. Memory visualization.

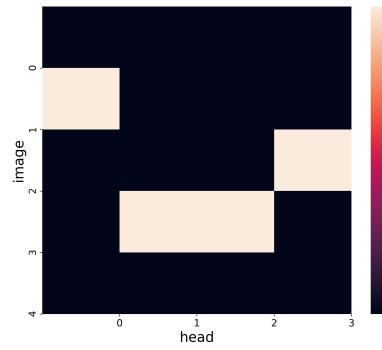


Fig. 5. Attention visualization.

the story of HLST is more diverse and expressive than the AREL. In this example, the story of AREL contains a lot of duplicate phrases and sentences, e.g., “It was a beautiful day” and “The bride and groom were very/so happy to be married”. The semantic repetition significantly reduces the readability and informativeness of the whole story. However, by introducing the multi-head attention with self adjust query, our model grasps the central idea “wedding” and generates the story that contains corresponding keywords, i.e., “wedding cake” and “ceremony”. And the semantic repetition is also improved for equipping the memory module with the gated mechanism. Furthermore, HLST has the capability of capturing the chronological relationship with the hierarchical decoder, such as the phrase “After the ceremony”.

To verify whether the performance improvements are owing to the central idea, we further conduct data analysis, including memory-aware words distribution and attention visualization. Specifically, we compute the cosine similarity between the memory and the generated word embeddings. As shown in Figure 4, the red dot represents the memory, and the blue dots are generated words in the story. For simplicity, we remove the common words, such as “a”, “the” and “for” etc. The distance between the blue dot and red dot measures the correlation between the corresponding words and the central idea. Hence, the words “bride”, “groom” and “wedding” are more relevant than the words “aisle”. It indicates that the memory module grasps the central idea and improves the informative of the generated story. We also visualize the attention of this example in Figure 5. The x-coordinate indicates the number of head and the y-coordinate indicates the input image stream. The lighter the color is, the more important the image is. We can see that the HLST pays more attention to the salient images, i.e. image2, image3 and image4, which is coherent with human understanding. The image3 and image4 explicitly demonstrate that “wedding” is the central idea of the story. Besides, although the scenes in the first two pictures are similar, the bride in image2 is more prominent than the groom in image1. Therefore, the image2 is more relevant with the central idea “wedding” than image1. That is why HLST generates the coherent and diverse stories as human beings.

6 Conclusion

In this paper, we propose a hierarchical neural network with the gated memory module to imitate the process of human storytelling. We utilize a hierarchical decoder to integrate the contextual information, including a narration decoder and a sentence decoder. Besides, the multi-head attention with self adjust query is employed to capture the salient information to initialize the memory unit as the central idea. Furthermore, we equip the memory module with the gated mechanism that includes a read gate and a write gate. The automatic evaluation results and human judgments show that our HLST outperforms all state-of-the-art methods significantly.

Acknowledgement

This research is supported by the National Key R&D Program of China (No.2017YFC0820700, No.2018YFB1004700).

References

1. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
2. Chen, Z., Zhang, X., Boedihardjo, A.P., Dai, J., Lu, C.T.: Multimodal storytelling via generative adversarial imitation learning. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17. pp. 3967–3973 (2017)
3. Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., Zweig, G., Mitchell, M.: Language models for image captioning: The quirks and what works. arXiv preprint arXiv:1505.01809 (2015)
4. Gonzalez-Rico, D., Pineda, G.F.: Contextualize, show and tell: A neural visual storyteller. CoRR **abs/1806.00738** (2018)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
6. Hsu, C., Chen, S., Hsieh, M., Ku, L.: Using inter-sentence diverse beam search to reduce redundancy in visual storytelling. CoRR **abs/1805.11867** (2018)
7. Hu, J., Cheng, Y., Gan, Z., Liu, J., Gao, J., Neubig, G.: What makes a good story? designing composite rewards for visual storytelling. Proceedings of the AAAI Conference on Artificial Intelligence **34**(05), 7969–7976 (Apr 2020)
8. Huang, Q., Gan, Z., Celikyilmaz, A., Wu, D., Wang, J., He, X.: Hierarchically structured reinforcement learning for topically coherent visual story generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8465–8472 (2019)
9. Huang, T.H.K., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., et al.: Visual storytelling. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1233–1239 (2016)

10. Jung, Y., Kim, D., Woo, S., Kim, K., Kim, S., Kweon, I.S.: Hide-and-tell: Learning to bridge photo streams for visual storytelling. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11213–11220 (2020)
11. Kim, T., Heo, M., Son, S., Park, K., Zhang, B.: GLAC net: Glocal attention cascading networks for multi-image cued story generation. CoRR [abs/1805.10973](https://arxiv.org/abs/1805.10973) (2018)
12. Li, T., Li, S.: Incorporating textual evidence in visual storytelling. In: Proceedings of the 1st Workshop on Discourse Structure in Neural NLG. pp. 13–17. Association for Computational Linguistics, Tokyo, Japan (Nov 2019)
13. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004)
14. Lin, C.Y., Och, F.J.: Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04). pp. 605–612. Barcelona, Spain (Jul 2004)
15. Liu, F., Perez, J.: Gated end-to-end memory networks. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. pp. 1–10 (2017)
16. Park, C.C., Kim, G.: Expressing an image stream with a sequence of natural sentences. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 28, pp. 73–81. Curran Associates, Inc. (2015)
17. Sukhbaatar, S., szlam, a., Weston, J., Fergus, R.: End-to-end memory networks. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 28, pp. 2440–2448. Curran Associates, Inc. (2015)
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 6000–6010 (2017)
19. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4566–4575 (2015)
20. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3156–3164 (2015)
21. Wang, J., Fu, J., Tang, J., Li, Z., Mei, T.: Show, reward and tell: Automatic generation of narrative paragraph from photo stream by adversarial training. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
22. Wang, X., Chen, W., Wang, Y.F., Wang, W.Y.: 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). pp. 899–909 (2018)
23. Yang, P., Luo, F., Chen, P., Li, L., Yin, Z., He, X., Sun, X.: Knowledgeable storyteller: a commonsense-driven generative model for visual storytelling. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence. pp. 5356–5362. AAAI Press (2019)
24. Yu, L., Bansal, M., Berg, T.: Hierarchically-attentive RNN for album summarization and storytelling. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 966–971. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017)