Knowledge-enriched Attention Network with Group-wise Semantic for Visual Storytelling

Tengpeng Li, Hanli Wang, *Senior Member, IEEE,* Bin He, Chang Wen Chen, *Fellow, IEEE*

Abstract—As a technically challenging topic, visual storytelling aims at generating an imaginary and coherent story with narrative multi-sentences from a group of relevant images. Existing methods often generate direct and rigid descriptions of apparent image-based contents, because they are not capable of exploring implicit information beyond images. Hence, these schemes could not capture consistent dependencies from holistic representation, impairing the generation of reasonable and fluent story. To address these problems, a novel knowledge-enriched attention network with group-wise semantic model is proposed. Three main novel components are designed and supported by substantial experiments to reveal practical advantages. First, a knowledge-enriched attention network is designed to extract implicit concepts from external knowledge system, and these concepts are followed by a cascade cross-modal attention mechanism to characterize imaginative and concrete representations. Second, a group-wise semantic module with second-order pooling is developed to explore the globally consistent guidance. Third, a unified one-stage story generation model with encoder-decoder structure is proposed to simultaneously train and infer the knowledge-enriched attention network, group-wise semantic module and multi-modal story generation decoder in an end-to-end fashion. Substantial experiments on the popular Visual Storytelling dataset with both objective and subjective evaluation metrics demonstrate the superior performance of the proposed scheme as compared with other state-of-the-art methods.

Index Terms—Visual Storytelling, Knowledge-enriched Attention, Group-wise Semantic, Multi-modal Decoder, Encoder-decoder.

✦

1 INTRODUCTION

Visual storytelling, which aims at producing a set of expressive and coherent sentences to depict the contents of a group of sequential images, has been an interesting and rapidly growing research topic in the fields of computer vision and multimedia computing. Different from visual captioning which devotes to describe the superficial contents in an image or a video, visual storytelling is expected not only to recognize the diverse semantical contexts and relations within one image and across images, but also to generate the storyline of image stream and express more implicit imaginations out of the images. Visual storytelling can be used in many real-world applications, such as helping the disabled to comprehend image contexts from social media, verifying advanced properties of intelligent devices, etc.

In visual storytelling, it is essential to learn the storyline and express with informative sentences. Therefore, valuable contextual information should be deduced for the target image stream. In general, a visual storytelling model intends to solve two main issues: (1) generating the abundant information of extracted features in single image, and (2) providing the precise storyline about the event occurred in the image sequence. On one hand, most visual captioning schemes focus on detecting visual features, where

convolutional features [1]–[3] and object features [4]–[6] have been widely used in these schemes. Nevertheless, regional-visual features can merely detect the intrinsic and superficial information, lacking the the capability to explore diverse and creative textures that were not apparent from images. Several recent approaches [7]–[10] introduced external knowledge by leveraging graph-based structures like the scene graph [11] and the commonsense graph [12] to strengthen symbolic creativity and achieve desired performances. Nonetheless, these approaches either did not establish the associations of cross-modal information or only learned the implicit external contents in two separated stages, leading to sub-optimal performance. We strongly believe that the attentive visual and textual representations are essential to produce concrete and imaginative descriptions.

On the other hand, a number of unified frameworks [1], [3], [13] have been developed recently to solve the problem of lacking global consistency in image sequence, where the recurrent neural network (RNN) [1], [3] or temporal convolutional network (TCN) [13] has been adopted to explore the temporal feature relations. However, both RNN and TCN encounter problems in their optimization [14] because of memory dilution along the longer feature sequence, failing to generate the topic-aware information of an image stream. Nevertheless, the storyline containing long-range dependencies is crucial to output the coherent multi-sentences. Furthermore, the most serious problem among existing approaches is that they are incapable of establishing a unified framework to simultaneously capture sufficient regional features and topic-aware global features for visual storytelling.

To address the aforementioned challenges, a knowledge-enriched attention network with group-wise semantic (KAGS) model is proposed in this research for visual

---

Corresponding author: Hanli Wang.

T. Li and H. Wang are with the Department of Computer Science & Technology, Key Laboratory of Embedded System and Service Computing (Ministry of Education), Tongji University, Shanghai 200092, P. R. China, and with Shanghai Research Institute for Intelligent Autonomous System, Shanghai 201210, P. R. China (e-mail: ltpfor1225@tongji.edu.cn, hanli-wang@tongji.edu.cn).

B. He is with Shanghai Research Institute for Intelligent Autonomous System, Shanghai 201210, P. R. China (e-mail: hebin@tongji.edu.cn).

C. W. Chen is with the Department of Computing, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China (e-mail: Chang-wen.chen@polyu.edu.hk).

storytelling. The proposed KAGS model will first leverage a CNN [15] and a Faster-RCNN [16] as encoder to extract convolutional features, semantic labels and regional object features from the input image stream. Then the semantic labels and regional features will be sent into the proposed knowledge-enriched attention network (KAN), where the semantic labels are processed with ConceptNet [12] and the regional features are dealt with the cascade cross-modal attention module. The proposed KAN can achieve sufficient feature representation to enable the establishment of cross-modal correlations of both textual and visual information. Meanwhile, the group-wise semantic module (GSM) with second order pooling (SOP) is introduced to transform the convolutional group features into global guided vector. Different from the sequential memory enhanced behaviour in RNN or TCN, GSM directly computes the higher-order interaction of any local or non-local pairwise convolutional vectors, in spite of their intra- or inter-spatial positions. The designed GSM can obtain the global feature guidance because it can capture the long-range dependencies of the sequential convolutional features. Finally, the optimized visual and textual features, combined with the global semantic vector, are sent into a multi-modal story decoder to generate the story. As a result, a unified one-stage framework with superior performance is established to optimize all proposed modules for attentive cross-modal features and global semantic guidance in an end-to-end manner. Major contributions of this work are summarized below.

- A knowledge-enriched attention network is designed to capture attentive enriched contexts and visual representations to address the problem in external information shortage and feature distraction. The contexts are generated from commonsense graphs and the cascade cross-modal attention is employed to highlight the valuable embedding of heterogeneous information.

- A group-wise semantic module is developed to capture the global consistency of an image stream to overcome the challenge about the incoherent descriptions in a story. This module is able to compute the higher-order interaction of any pairwise semantic vectors regardless of spatial distance restriction, thus contributing to the accurate guidance of the storyline.

- A unified one-stage visual storytelling framework with encoder-decoder structure is devised to simultaneously optimize the knowledge-enriched attention network, group-wise semantic module and multi-modal story decoder in an end-to-end fashion. It has been shown that the proposed KAGS scheme is both efficient and effective.

The rest of this paper is organized as follows. We introduce in Section 2 the related works in both image captioning and visual storytelling. The proposed knowledge-enriched attention network with group-wise semantic model is described in detail in Section 3. We present in Section 4 the statistic performances, ablative studies and visualization analyses. Finally, we conclude this paper in Section 5.

2 RELATED WORK

2.1 Image Captioning

Image captioning aims at automatically generating a natural language sequence to depict the complex visual contents occurred in a single image, and it can be generally divided into two categories. First, benefiting from the rapid developing technology of natural language machine translation, most early approaches [17]–[23] attempted to establish the captioning framework with encoder-decoder structure and achieved satisfying performances. In these common approaches, CNN was usually regraded as encoder to extract image features, and RNN was often used to decode the integrated representations for sentence production. In [17], Mao *et al.* designed the m-RNN framework consisted of two sub-networks including a CNN-based image encoder and a RNN-based sentence decoder to accomplish sentence generation. Vinyals *et al.* [18] leveraged the CNN to extract visual representations and applied the long short-term memory (LSTM) [24] to output the final image description. Jia *et al.* [19] proposed a gLSTM model to add the extracted semantic contexts in each LSTM unit for guiding global image content generation. Second, a set of innovated methods with an attention mechanism [25]–[29] have been proposed to further improve image captioning performances by highlighting meaningful visual and textual information in recent years. Xu *et al.* [25] designed a LSTM-based decoder with soft attention and hard attention modules to focus on important image areas for generating accurate words in the decoding process. You *et al.* [26] presented a semantic attention model that integrates the extracted semantic visual feature proposals into the hidden states and RNN-based decoders for better language description. In [27], Lu *et al.* developed an adaptive attention structure to selectively choose image regions for obtaining meaningful features. Furthermore, Anderson *et al.* [30] proposed a bottom-up and top-down attention framework to explore the object-level salient regions and relate each region with one corresponding word for sentence generation. Li *et al.* [28] performed a scene graph strategy [11] to capture enriched structural information with semantic entities and pairwise relations. Yang *et al.* [29] proposed the CaptionNet model as an enhanced LSTM to focus on positive visual cues and absorb richer semantics for better feature encoding. In this work, the encoder-decoder structure is also developed by additionally merging attention mechanism and global guidance for robust feature representation.

2.2 Visual Storytelling

Visual storytelling is a challenging task in multimedia communities since the designed approaches should bridge an association between the group of visual messages and the sequential natural languages. As an emerging and promising topic, visual storytelling has attracted much attention of researchers and a number of elaborate innovations are proposed. Generally, visual storytelling models can be grouped into end-to-end framework and multi-stage based approach. First, end-to-end framework is popular due to its efficiency for generating stories in a unified structure. Wang *et al.* [1] proposed a classical visual storytelling framework that has been the most popular base structure of following studies.

Fig. 1. Pipeline of the proposed KAGS for visual storytelling. The framework contains four key components: (a) a Faster-RCNN network and a ResNet backbone to extract regional features and high-level convolutional features; (b) the proposed KAN to obtain the attentive heterogeneous representations by exploiting the intra- and inter- interactions of visual and knowledge concepts; (c) the proposed GSM to explore the global guided aggregation with a set of hierarchical second-order pooling algorithms in a convolutional feature group; and (d) a story generation that fuses the multi-modal information in a decoder to produce the final predicted sentences.

This framework designed an end-to-end structure to encode the sequential converted features jointly by bidirectional gated recurrent units (GRU) and decoded these processed features separately for the final story. Huang *et al.* [31] designed a hierarchical two-level decoder to produce the semantic topic and generate a sentence for each single image, and reinforcement learning was applied for optimization. In [32], a commonsense-driven generator was employed to caption essential external messages for abundant multi-sentence expressions. Jung *et al.* [3] proposed a hide and tell model to acquire the imaginative storyline by bridging the feature gap of image stream. To ensure the interesting and informative characteristics of story, Hu *et al.* [33] designed three human-like criteria combined with a reinforcement learning structure and achieved superior performances on human evaluation metrics. Second, many multi-stage approaches were also emerging which strengthened the diversity and informativeness of frameworks. Hsu *et al.* [34] merged various extracted concepts into decoder for more diverse descriptions. Yao *et al.* [35] designed a hierarchical framework to plan the storyline in the first stage and wrote the topic-based story in the second stage. Moreover, several works [7], [8] introduced the external commonsense knowledge from bases like OpenIE [36], Visual Genome [37] or ConceptNet [12] for more diverse descriptions, where Hsu *et al.* [7] proposed a three-stage framework to produce external knowledge to guide the decoder, Chen *et al.* [8] designed a concept selection module to select enriched concept candidates and then sent them in a visual-language pre-trained model to produce full stories. In this work, an

end-to-end model is designed while considering efficiency, informativeness and coherency. Particularly, the proposed one-stage model can train and inference all modules in a unified fashion to promote its efficiency, and the attentive commonsense knowledge and global semantic are also introduced to increase the feature representation for improving the informativeness and consistency of KAGS, respectively.

3 KNOWLEDGE-ENRICHED ATTENTION NETWORK WITH GROUP-WISE SEMANTIC

3.1 Framework Overview

The proposed KAGS is illustrated in Fig. 1. First, a knowledge-enriched attention network is designed to explore the intra- and inter- interactions of visual and textual features in Section 3.2. Meanwhile, a group-wise semantic module with a set of second-order pooling algorithms is developed to capture the global guided aggregation of sequential convolutional features in Section 3.3. Finally, the produced multi-modal features are sent into the multi-modal story decoder to generate the final reasonable and coherent story in Section 3.4.

With a group of N associated images $\mathcal{I} = \{I^n\}_{n=1}^N$ as input, the task of visual storytelling aims to exploit the effective intra- and inter-feature representations of this image stream, producing a reasonable and coherent story with multiple descriptive sentences $\mathcal{S} = \{\mathbf{S}^n\}_{n=1}^N$. To tackle this issue, a novel KAGS model is elaborately designed to generate the story $\mathcal{S}$ in an end-to-end manner.

The overall structure of the proposed KAGS model is illustrated in Fig. 1, which consists of four main components: (a) Encoder, (b) Knowledge-enriched Attention Network (KAN), (c) Group-wise Semantic Module (GSM), and (d) Story Generation. Specifically, given a group of relevant images $\mathcal{I}$, the model first leverages the general object detection framework Faster-RCNN [16] and ResNet [15] backbones as the encoder to extract boxes of regional-visual features $\mathcal{R} = \{\mathbf{R}^n\}_{n=1}^N$ and the corresponding labels $\mathcal{L} = \{\mathbf{L}^n\}_{n=1}^N$ with high confidence, and the high-level representations in the last convolutional layer $\mathcal{C} = \{\mathbf{C}^n\}_{n=1}^N$, respectively. Then, $\mathbf{L}^n$ of each image is fed into KAN to explore external knowledge. For the semantic label $\mathbf{L}^n \in \mathcal{L}$, the ConceptNet [12] is introduced to generate the knowledge concepts $\mathbf{K}^n$ from external enhanced knowledge base that can further boost the capability of absorbing imaginative and reasonable concepts, thus a group of knowledge concepts $\mathcal{K} = \{\mathbf{K}^n\}_{n=1}^N$ can be acquired. Moreover, to fully utilize the regional-visual features $\mathbf{R}^n$ and the knowledge concepts $\mathbf{K}^n$, a cascade cross-modal attention (CCA) module is designed to progressively model the dense semantic interactions of intra features (image-to-image or text-to-text) and inter features (image-to-text), outputting the enhanced knowledge concepts and attentive regional-visual features. The whole process is defined as $[\mathbf{K}_P^n, \mathbf{R}_P^n] = \mathcal{F}_{cca}(\mathbf{K}^n, \mathbf{R}^n)$, where $\mathcal{F}_{cca}(\cdot, \cdot)$ and P represent the function of CCA module and the number of cascade layers in CCA, respectively.

Moreover, the recent works [4], [38] have shown that an outstanding non-linear feature capability of second-order pooling is achieved by exploiting both channel-wise and spatial-wise interactions. Thereby, the GSM with hierarchical second-order pooling is designed to capture the topic-aware consistency of group convolutional features $\mathcal{C} = \{\mathbf{C}^n\}_{n=1}^N$, and then produces a global-visual aggregation $\tilde{\mathbf{A}} = \mathcal{F}_{gsm}(\mathcal{C})$, which can help to capture the global guided semantic and avoid noisy interference. Finally, the model feds $\mathbf{K}_P^n$, $\mathbf{R}_P^n$ and $\tilde{\mathbf{A}}$ into the multi-modal story decoder, generating the predicted sentence $\mathbf{S}^n$.

3.2 Knowledge-enriched Attention Network

As aforementioned, to overcome the problem of insufficient external information and distracted features, the knowledge-enriched attention network (KAN) is designed to increase the external priors from current knowledge repository and establish intra- and inter- dense correlations of cross-modal features. In fact, several existing knowledge-based methods [8], [13], [39] for visual storytelling also devote to leverage external implicit knowledge for better model performance, but they only focus on the intra correspondence of textual concepts instead of considering the inter association of heterogeneous information that is crucial to visual storytelling, resulting in sub-optimal representation capability. Differently, the proposed KAN constructs the interactions of both enriched knowledge and visual concepts based on CCA, which establishes the long-range dependencies of homogeneous and heterogeneous features between any pairwise feature vectors. Therefore, enriched textual knowledge and visual features can be assigned with higher attention weights in meaningful feature dimensions, facilitating to a more optimized visual storytelling estima-

tion than the methods only considering textual information [8], [13], [39].

Fig. 2. The schematic diagram of two attention units employed by the proposed CCA module, where the left unit is self-attention and the right unit is cross-attention.

Knowledge Graph. To offer current storytelling datasets more imaginary and reasonable concepts, the proposed KAGS establishes commonsense knowledge graphs based on the semantic labels $\mathcal{L}$ detected by Faster-RCNN [16]. Similar to [8], [32], KAGS adopts the generalized ConceptNet [12] as the knowledge extractor to collect numerous commonsense words with rich imagination, abundant emotions and objective facts. Specifically, the knowledge concepts $\mathbf{K}^n = \{\mathbf{K}_k^n\}_{k=1}^K$ is constructed for the given semantic label $\mathbf{L}^n$, where K indicates to employ the the top-K candidates of the nth image based on their scores of confidence, and each $\mathbf{K}_k^n$ is composed of two entities and one edge relation.

Cascade Cross-modal Attention. Given the extracted rich knowledge, a tricky challenge is that many selected concepts are irrelevant to the visual information, thus introducing many interferences that reduce the story description accuracy. Recently, the method [9] investigates the visual-textual guided encoding pattern to selectively highlight the positive information and suppress the negative message. Motivated by this and the self-attention mechanism in [14], the CCA module is designed through stacking self-attention (SA) and cross-attention (CA) as shown in Fig. 2 to progressively explore and optimize cross-modal interactions. In detail, having the query matrix $\mathbf{M}_q \in \mathbb{R}^{m \times d}$, the key matrix $\mathbf{M}_k \in \mathbb{R}^{m \times d}$ and the value matrix $\mathbf{M}_v \in \mathbb{R}^{m \times d}$, the attentive feature $\mathbf{F} \in \mathbb{R}^{m \times d}$ can be obtained by summing all values of $\mathbf{M}_v$ with the corresponding matrix weights learned from $\mathbf{M}_q$ and $\mathbf{M}_k$, and the dot-product attention is defined as

$$\mathbf{F} = Attention(\mathbf{M}_q, \mathbf{M}_k, \mathbf{M}_v) = softmax(\frac{\mathbf{M}_q \mathbf{M}_k^\top}{\sqrt{d}})\mathbf{M}_v, \quad (1)$$

where $\frac{1}{\sqrt{d}}$, m and d represent scale factor, vector number and feature dimension, respectively.

In order to enhance the feature capacity of different subspaces, a multi-head attention mechanism [14] is also leveraged, which consists of h parallel subspaces. The attentive feature $\mathbf{F}$ is formulated as

$$
\begin{aligned}
\mathbf{F} &= MultiHead(\mathbf{M}_q, \mathbf{M}_k, \mathbf{M}_v) \\
&= [head^1, head^2, \cdots, head^h]\mathbf{W}_o,
\end{aligned}
\tag{2}
$$

$$
head^i = Attention(\mathbf{M}_q\mathbf{W}_q^i, \mathbf{M}_k\mathbf{W}_k^i, \mathbf{M}_v\mathbf{W}_v^i), \tag{3}
$$

where $\mathbf{W}_q^i \in \mathbb{R}^{d \times d_q}$, $\mathbf{W}_k^i \in \mathbb{R}^{d \times d_k}$ and $\mathbf{W}_v^i \in \mathbb{R}^{d \times d_v}$ are the learnable projection matrices of the ith head, and $\mathbf{W}_o \in \mathbb{R}^{(h \times d_v) \times d}$. In this schema, the multi-head attention is applied to both of the SA and CA units, followed by the function $LS(\cdot)$ consisting of a point-wise addition, a linear layer and a BatchNorm layer. In Fig. 2, given the visual features $\mathbf{F}_v$ or the textual features $\mathbf{F}_t$ of each image, the SA unit outputs the self-attentive representation as

$$
\begin{aligned}
SA(\mathbf{F}_v) &= LS(MultiHead(\mathbf{F}_v, \mathbf{F}_v, \mathbf{F}_v)), \\
SA(\mathbf{F}_t) &= LS(MultiHead(\mathbf{F}_t, \mathbf{F}_t, \mathbf{F}_t)).
\end{aligned}
\tag{4}
$$

Similarly, both visual features $\mathbf{F}_v$ and textual features $\mathbf{F}_t$ can be fed into CA unit, generating the cross-attentive representation as

$$
CA(\mathbf{F}_t, \mathbf{F}_v) = LS(MultiHead(\mathbf{F}_t, \mathbf{F}_v, \mathbf{F}_v)), \tag{5}
$$

Now, the proposed CCA can be constructed by cascading $P-1$ layers as shown in Fig. 1 (b), which is represented as $\mathcal{F}_{cca} = [\mathcal{F}_{cca}^{(1)}, \mathcal{F}_{cca}^{(2)}, \cdots, \mathcal{F}_{cca}^{(P-1)}]$. Specifically, the pth cascade layer of $\mathcal{F}_{cca}$ including two SA units and one CA unit can be defined as

$$
\begin{aligned}
[\mathbf{K}_{p+1}^n, \mathbf{R}_{p+1}^n] &= \mathcal{F}_{cca}^{(p)}(\mathbf{K}_p^n, \mathbf{R}_p^n) \\
&= [CA(SA(\mathbf{K}_p^n), SA(\mathbf{R}_p^n)), SA(\mathbf{R}_p^n)],
\end{aligned}
\tag{6}
$$

where $\mathbf{K}_p^n$, $\mathbf{R}_p^n$, $\mathbf{K}_{p+1}^n$ and $\mathbf{R}_{p+1}^n$ represent input knowledge concepts, input regional-visual features, output knowledge concepts and output regional-visual features at the pth cascade layer, respectively. For $\mathcal{F}_{cca}^{(1)}$, we set original input features $\mathbf{R}_1^n = \mathbf{R}^n$ and $\mathbf{K}_1^n = \mathbf{K}^n$. Finally, the outputs $[\mathbf{K}_P^n, \mathbf{R}_P^n] = \mathcal{F}_{cca}^{(P-1)}(\mathbf{K}_{P-1}^n, \mathbf{R}_{P-1}^n)$ with Eq. (6) are regarded as the enhanced knowledge concepts and attentive regional-visual features of CCA, respectively.

The designed KAN has proved its superior potential to collect external commonsense facts and capture long-range pairwise correlations of cross-modal features, so as to better discriminate the valuable heterogeneous representations from imaginative corpus and visual contexts. Nevertheless, KAN only establishes multiple interactions of single image, neglecting to explore the topic-aware global consistency that is necessary for visual storytelling. To tackle this problem, the group-wise semantic module (GSM) is further developed to exploit the global guided aggregation as presented in the following Section 3.3.

3.3 Group-wise Semantic Module

One major difficulty in visual storytelling task is the lack of storyline, leading to the incoherent expressions of multiple sentences. To this end, a group-wise semantic module

Fig. 3. The schematic diagram of SOP. Given an input feature tensor with size $h \times w \times d$, it is fed into SOP, which consists of two 1×1 convolutions, one transpose multiplication operator and one row-wise convolution, generating a $1 \times 1 \times d$ global guided aggregation.

composed of several second order pooling algorithms is developed to capture the global consistent guidance.

Second Order Pooling (SOP). Given the convolutional feature tensor $\mathbf{X} \in \mathbb{R}^{h \times w \times d}$ as shown in Fig. 3, where h, w and d represent the height, width and channel dimension of feature tensor, respectively. SOP first introduces a 1×1 convolution to reduce the channel number from d to c, thus projecting the convolutional feature from high to low dimension while alleviating the computation cost. Then SOP converts a $h \times w \times c$ feature tensor to a $c \times c$ covariance matrix by computing dense semantic interactions regardless of positional distance. Each element in the covariance matrix indicates the similarity of any pairwise vectors in the feature tensor, which formulates the high-order property of significant holistic representation by introducing a quadratic operator and thus can enable the model with the capacity of non-linear feature discrimination. Finally, a row-wise convolutional layer and a 1×1 convolutional layer are leveraged to convert the $c \times c$ covariance matrix to a $1 \times 1 \times d$ tensor to highlight the meaningful feature channels. Specifically, the process of SOP can be described as

$$
\begin{aligned}
\tilde{\mathbf{X}} &= SOP(\mathbf{X}) \\
&= f^{1 \times 1}(f^{row}([\mathcal{R}(f^{1 \times 1}(\mathbf{X}))]^{\top} * [\mathcal{R}(f^{1 \times 1}(\mathbf{X}))])),
\end{aligned}
\tag{7}
$$

where $\tilde{\mathbf{X}} \in \mathbb{R}^{1 \times 1 \times d}$, $*$ indicates matrix multiplication, $\mathcal{R}$ is a reshaping operator that flattens a tensor from size $h \times w \times c$ to $(hw) \times c$, $f^{1 \times 1}$ and f^{row} represent a 1×1 convolution and a row-wise convolution, respectively.

Group-wise Semantic. In Fig. 1(c), the GSM module first inputs every feature representation $\mathbf{C}^n \in \mathbb{R}^{h \times w \times d}$ into SOP with Eq. (7), and then the SOP outputs the processed tensor $\tilde{\mathbf{C}}^n \in \mathbb{R}^{1 \times 1 \times d}$. Afterwards, all processed tensors are sequentially concatenated into $\mathbf{A} = [\{\tilde{\mathbf{C}}^n\}_{n=1}^N] \in \mathbb{R}^{N \times 1 \times d}$, producing an initial group-wise semantic representation. Similarly, the GSM again sends $\mathbf{A}$ into SOP with Eq. (7) to capture the long-range semantic associations along the channel-wise dimension, generating the global-visual aggregation $\tilde{\mathbf{A}} \in \mathbb{R}^{1 \times 1 \times d}$ that can contribute to the subsequent multi-modal story decoder in Section 3.4, which can be formulated as

$$
\begin{aligned}
\tilde{\mathbf{A}} &= \mathcal{F}_{gsm}(\mathcal{C}) \\
&= SOP([\{SOP(\mathbf{C}^n)\}_{n=1}^N]).
\end{aligned}
\tag{8}
$$

As a consequence, the SOP can strengthen the non-linear feature capability by learning higher-order statistic dependencies of holistic representation [38], and the GSM can capture the global consistent representation of group-wise features along the channel-wise dimension as shown in

Fig. 4. Illustration of the proposed multi-modal story decoder. For the knowledge indicator vector $\bar{\mathbf{K}}^n$, the regional-visual indicator vector $\bar{\mathbf{R}}^n$, the global-visual indicator vector $\tilde{\mathbf{A}}$, the previous regional hidden state $\mathbf{hr}_{t-1}^n$, the previous global hidden state $\mathbf{ha}_{t-1}^n$ and the previous word embedding $\mathbf{w}_{t-1}^n$ as inputs, the decoder feds these vectors into a two-stream structure by combining the CA unit and LSTM to obtain a set of vectors $\mathbf{vr}_t^n$, $\mathbf{hr}_t^n$, $\mathbf{va}_t^n$ and $\mathbf{ha}_t^n$. Finally, these vectors are concatenated and sent into following layers to obtain the current word prediction $\mathbf{w}_t^n$.

Fig. 6, facilitating to acquire topic-aware information for coherent and narrative descriptions.

3.4 Multi-modal Story Decoder

To fully utilize the produced attentive local-visual features, enhanced knowledge concepts and global-visual aggregation, a multi-modal story decoder is designed to explore the final contextual representation with above multi-modal features, generating reasonable and coherent sentences of the final story. Figure 4 illustrates the diagram of the proposed multi-modal story decoder. Specifically, in order to generate the nth sentence with various representations including attentive regional-visual features $\mathbf{R}_P^n$, enhanced knowledge concepts $\mathbf{K}_P^n$ and global-visual aggregation $\tilde{\mathbf{A}}$, the model first flattens $\mathbf{R}_P^n \in \mathbb{R}^{M \times d}$ to $\bar{\mathbf{R}}^n \in \mathbb{R}^{1 \times d}$, $\mathbf{K}_P^n \in \mathbb{R}^{K \times d}$ to $\bar{\mathbf{K}}^n \in \mathbb{R}^{1 \times d}$ with designed flatten layer composed of two linear layers and one softmax layer, resulting in the regional-visual indicator vector $\bar{\mathbf{R}}^n$ and the knowledge indicator vector $\bar{\mathbf{K}}^n$, where M, K and d denote the number of detected regional boxes, graph relations and feature channels, respectively.

To further exploit compact interactions of visual features, enriched contexts and word embedding, a regional-visual and global-visual based story decoder is designed by combining the CA unit and LSTM to accomplish multi-modal inference. Particularly, for regional-visual information reasoning of the nth image at the time step t (see the left side of Fig. 4), the decoder sends the previous regional hidden state $\mathbf{hr}_{t-1}^n$, the knowledge indicator vector $\bar{\mathbf{K}}^n$, the previous word embedding $\mathbf{w}_{t-1}^n$ and the regional-visual indicator vector $\bar{\mathbf{R}}^n$ into LSTM, outputting the current regional

hidden state $\mathbf{hr}_t^n$. Afterwards, the decoder considers $\mathbf{hr}_t^n$ as the query of the CA unit, and $\bar{\mathbf{R}}^n$ is set as the key or value of the CA unit. As a result, the output of the CA unit followed with an embedded layer obtains the attended regional representation $\mathbf{vr}_t^n$ by encouraging the cross-modal correlations between $\bar{\mathbf{R}}^n$ and $\mathbf{hr}_t^n$, which can be formulated as

$$\mathbf{hr}_t^n = LSTM(\bar{\mathbf{K}}^n \oplus \mathbf{w}_{t-1}^n \oplus \bar{\mathbf{R}}^n, \mathbf{hr}_{t-1}^n), \qquad (9)$$

$$\mathbf{vr}_t^n = Embed(CA(\mathbf{hr}_t^n, \bar{\mathbf{R}}^n)), \qquad (10)$$

where $Embed(\cdot)$ represents a fully-connected layer and $\oplus$ denotes the concatenation operator. Similarly, with the input of the previous global hidden state $\mathbf{ha}_{t-1}^n$, the knowledge indicator vector $\bar{\mathbf{K}}^n$, the previous word embedding $\mathbf{w}_{t-1}^n$ and the global-visual aggregation $\tilde{\mathbf{A}}$, the global-visual information reasoning (see the right side of Fig. 4) can also generate the current global hidden state $\mathbf{ha}_t^n$ and attended global representation $\mathbf{va}_t^n$, which can be formulated as

$$\mathbf{ha}_t^n = LSTM(\bar{\mathbf{K}}^n \oplus \mathbf{w}_{t-1}^n \oplus \tilde{\mathbf{A}}, \mathbf{ha}_{t-1}^n), \qquad (11)$$

$$\mathbf{va}_t^n = Embed(CA(\mathbf{ha}_t^n, \tilde{\mathbf{A}})). \qquad (12)$$

Next, the contextual vector $\mathbf{v}_t^n$ is calculated by concatenating $\mathbf{vr}_t^n$, $\mathbf{hr}_t^n$, $\mathbf{va}_t^n$ and $\mathbf{ha}_t^n$, followed with a GLU [40] and a linear layer, respectively. Finally, the contextual vector $\mathbf{v}_t^n$ is fed into a softmax layer to generate the current word embedding $\mathbf{w}_t^n$. Definitely, the word generation probability can be formulated as

$$p(\mathbf{w}_t^n | \mathbf{w}_{1:t-1}^n) = softmax(\mathbf{v}_t^n), \qquad (13)$$

where the prediction p is a probability distribution over the Visual Storytelling (VIST) dataset [41] vocabulary $\mathbb{V}_s$. Finally, the word embedding $\mathbf{w}_t^n$ is transformed into word w_t^n, obtaining the sub-story $\mathbf{S}^n = \{w_1^n, \cdots, w_T^n\}$ of story $\mathcal{S}$, where T represents the length of sub-story $\mathbf{S}^n$.

3.5 Training and Inference Procedure

In the training stage, given a group of N images, all the key components of the proposed model in Fig. 1 are jointly trained on the VIST dataset [41] for story prediction. The cross-entropy loss is employed in the training stage as

$$L(\theta) = -\sum_{n}^{N} \sum_{t}^{T} log(p_{\theta}^n(\mathbf{g}_t^n | \mathbf{g}_1^n, \cdots, \mathbf{g}_{t-1}^n)), \qquad (14)$$

where θ indicates the set of optimized parameters during training, $\mathbf{g}_t^n$ represents the tth word embedding in the ground-truth sub-story $\mathbf{g}^n$. Eventually, the goal is to minimize the loss $L(\theta)$. In the inference stage, the model predicts the story using the beam search method with the beam size equal to 3.

4 EXPERIMENTS

4.1 Implementation Details

Following the previous works [1], [31], [42], the proposed KAGS model adopts the ResNet-152 [15] pretrained on the ImageNet [43] dataset for convolutional feature extraction and utilizes the Faster-RCNN [16] pretrained on the ImageNet [43] dataset and the Visual Genome [37] dataset for

regional-level feature extraction, where the original convolutional feature and the regional feature are a $7 \times 7 \times 2048$ tensor and a 1×2048 tensor, respectively. Then these features are transposed into tensors with the channel dimension equal to 1024. The number of images in an album is set as 5. For each commonsense graph, the max number of relations is set as 20. Moreover, the number of the detected regional boxes is set as 36, the dimension of word embedding is set as 1024, the feature dimension in the hidden layer of LSTM is set to 512, and the number of cascade layers (*i.e.*, $P - 1$) in CCA is set as 6. In the current work, the cross-entropy loss is used to train the whole model and the Adam optimizer [44] is employed with the initial weight decay $5 \times e^{-4}$ and the learning rate $4 \times e^{-4}$. The model is converged in only 21 epochs with the batch size equal to 50, note that the model does not leverage any post processing such as reinforcement learning [22]. The words appearing more than 3 times in the training dataset are selected to build a storytelling vocabulary with a size of $9,837$. Then the vocabulary size is extended to $12,322$ with external knowledge base. During inference, the beam search strategy is leveraged with the beam size of 3 for visual storytelling prediction. The model is implemented with PyTorch[1] with a Tesla V100 for acceleration.

4.2 Dataset and Automatic Metric Evaluation

VIST Dataset. The VIST dataset [41] is a customized dataset for visual storytelling, which contains $210,819$ specific images and $10,117$ interesting Flicker albums. It is challenging to employ VIST for visual storytelling, because the story descriptions are more subjective and need emotional and imaginative concepts that do not appear explicitly in images. Following the previous work [1], the broken photos are removed and $40,098$ training groups, $4,988$ validation groups and $5,050$ testing groups are constructed. Each group consists of 5 images collected from one photo album and each image usually corresponds to one sentence. Every album has 5 differentiate stories as reference.

Automatic Metric Evaluation. Comprehensive experiments are conducted on the VIST dataset in terms of four automatic metrics including BLEU [45], METEOR [46], ROUGE_L [47] and CIDEr [48]. These metrics calculate the similarities and relevances between the predicted story and reference. Concluding in [41], the METEOR score is chosen as the key performance indicator for its high correlation with human evaluation standards.

4.3 Comparison with State-of-the-art Methods on Automatic Metrics

The proposed KAGS model is compared with other twelve state-of-the-art visual storytelling approaches including (1) seq2seq [41], an original model with RNN-based structure; (2) BARNN [49], a relational attended model with designed GRU; (3) h-attn-rank [50], a hierarchical attentive recurrent network; (4) XE-ss [1], a LSTM-based encoder-decoder model; (5) AREL [1], an adversarial reward optimizing framework; (6) HPSR [51], a hierarchical image

encoder-decoder model; (7) HSRL [31], a hierarchical reinforcement learning framework; (8) VSCMR [52], a conceptual exploration network; (9) ReCO-RL [33], a relevant context reinforcement learning method; (10) INet [3], an imaginative concept reasoning network; (11) SGVST [13], a scene-graph knowledge enhanced model; and (12) IRW [42], a multi-graph knowledge reasoning framework. For fair comparisons, this paper directly presents the statistic results provided by the authors or conducts the experiments by the official source codes of these competing approaches.

4.3.1 Qualitative Results

Figure 5 presents several visual comparisons between the proposed KAGS model and the methods AREL [1] and VSCMR [52], together with the human-annotated referenced stories (GT). Generally, compared with the other two approaches (*i.e.*, AREL and VSCMR), KAGS can better generate emotional, imaginative, coherent and accurate descriptions by jointly exploring the knowledge enriched cross-modal interactions and global semantic guidance.

Specifically, the left album of the five images in Fig. 5 is related to a graduation activity with various scenes, it is apparent that the predicted sentences obtained by KAGS show promising performances. For the second sentence of VSCMR, it simply produces the sentence "there was a lot of people there" and neglects to record the detailed visual and implicit contexts in this picture, leading to suboptimal results. Notwithstanding, the second sentence of KAGS shows the description "the crowd was excited for the graduation ceremony", where the word "excited" properly depicts the emotions of people and the phrase "graduation ceremony" accurately illustrates the social activity using the information from knowledge graphs, confirming the capability of KAGS to capture rich emotions and external contexts according to visual and textual information. For the fourth sentence of AREL (*i.e.*, "he was so proud of him"), it only characterizes the emotions of people and is irrelevant to the precise visual context in this photo, which is ambiguous to understand. However, the fourth sentence of KAGS outputs the sentence "the students were happy to finally graduate", which highly corresponds to the graduation topic of this photo album.

Moreover, the story generation of the right photo album in Fig. 5 is also challenging due to its numerous characters and various semantic objects in different scenarios. In the estimated story obtained by AREL, the third and fourth sentences show the repetitive phrase "had a great time", which impairs the abundant descriptions of this story. Notwithstanding, the proposed KAGS can avoid this problem and generate sentences with different formats and styles (*i.e.*, the third and fourth sentences generated by KAGS). In addition, regarding the fifth sentence of story, the VSCMR method predicts the sentence of "everyone had a great time at the reception", which generally introduces the event happened in this scene. And the proposed KAGS generates the sentence of "after the wedding they all posed for pictures", which shows that the generated sentence is associated with the visual information in the fourth image, further validating the long-range dependency capacity of the proposed KAGS model. Totally, the experimental results demonstrate that the designed model is able to obtain favorable story estimations

---

1. [Online]. Available: https://pytorch.org/

Images	Images
Knowledge Graphs	Knowledge Graphs
Generated Stories	Generated Stories

KAGS: the stadium was packed for the big time . the crowd was excited for the graduation ceremony . my son posed for a picture with the mascot to support the festivities . the students were happy to finally graduate . he was proud of his team .

AREL: the game was a lot of fun . it was a great day to see a lot of people there . i had a great time at the fair yesterday . he was so proud of him . he was so happy to see him .

VSCMR: today was the day of the graduation ceremony . there was a lot of people there . the students were very happy to be there . the students were very proud of their accomplishments . this is my best friend [male] . i am so proud of .

GT: the stage is set for a wonderful graduation . all of the students wait attentively to receive their degree . he pauses in the excitement to take a picture with the school mascot . the teachers pose with one of their favorite students . after its all said and done , they take a nice family picture .

KAGS: the bride and [female] were ready for the reception . the flowers were beautiful . the guests arrived at the wedding . [male] and [female] also had a night of guests . after the wedding they all posed for pictures .

AREL: the bride and groom cut the cake . this is a picture of a table . we had a great time at the party . i had a great time at the party yesterday . the bride and groom were so happy to be there .

VSCMR: the wedding party was a lot of fun . the flowers were beautiful . everyone was having a great time . my friends and i went to the reception and had a great time . everyone had a great time at the reception .

GT: this is [male] and [female] 's beautiful wedding day . this has to be the most beautiful flower arrangement for a wedding ever made . here is the grooms family having a good time at the reception . [male] and marv our trying to out do each other in getting that perfect picture . here is a great picture of the wedding party .

Fig. 5. Visualization of the comparison between the proposed KAGS and other state-of-the-art methods including AREL, VSCMR and ground-truth. It only visualizes parts of the extracted commonsense knowledge graphs due to space limit.

TABLE 1
Statistic comparisons of KAGS with other state-of-the-art approaches, where the bold font indicates the best performance.

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr
seq2seq [41] (NAACL2016)	-	-	-	3.5	31.4	-	6.8
BARNN [49] (AAAI2017)	-	-	-	-	33.3	-	-
h-attn-rank [50] (EMNLP2017)	-	-	21.0	-	34.1	29.5	7.5
XE-ss [1] (ACL2018)	62.3	38.2	22.5	13.7	34.8	29.7	8.7
AREL [1] (ACL2018)	63.7	39.0	23.1	14.0	35.0	29.6	9.5
HPSR [51] (AAAI2019)	61.9	37.8	21.5	12.2	34.4	31.2	8.0
HSRL [31] (AAAI2019)	-	-	-	12.3	35.2	30.8	10.7
VSCMR [52] (ACMMM2019)	63.8	39.5	23.5	14.3	35.5	30.2	9.0
ReCO-RL [33] (AAAI2020)	-	-	-	12.4	33.9	29.9	8.6
INet [3] (AAAI2020)	64.4	40.1	23.9	14.7	35.6	29.0	10.0
SGVST [13] (AAAI2020)	65.1	40.1	23.8	14.7	35.8	29.9	9.8
IRW [42] (AAAI2021)	66.7	41.6	25.0	**15.4**	35.6	29.6	11.0
KAGS	**70.1**	**43.5**	**25.2**	14.7	**36.2**	**31.4**	**11.3**

in several challenging conditions, confirming the superior performance of the proposed KAGS model.

4.3.2 Quantitative Results

The comparison of the proposed KAGS model with other state-of-the-art approaches is also presented in Table 1, where it can be observed that the statistic results of KAGS show better performances than the competing approaches by a large margin. Generally, the proposed KAGS achieves the best scores in terms of six metrics including BLEU-1, BLEU-2, BLEU-3, METEOR, ROUGE_L and CIDEr, and obtains the second best performance on BLEU-4. Specifically, KAGS achieves the BLEU-1 score of 70.1, the BLEU-2 score of 43.5, the BLEU-3 score of 25.2, the METEOR score of 36.2, the ROUGE_L score of 31.4 and the CIDEr score of 11.3, significantly surpassing the scene graph based method SGVST [13] by 5.0%, 3.4%, 1.4%, 0.4%, 1.5% and 1.5%, respectively. Moreover, compared with the second best method IRW [42] that leverages many external knowledge including scene graph, commonsense graph and event graph, the proposed KAGS model can achieve higher scores on most metrics. Particularly, 70.1 versus 66.7 on BLEU-1, 43.5 versus 41.6 on BLEU-2, 25.2 versus 25.0 on BLEU-3, 36.2 versus 35.6 on METEOR, 31.4 versus 29.6 on ROUGE_L, 11.3 versus 11.0 on CIDEr. In summary, the quantitative results confirm that the proposed modules can boost the performance of visual storytelling by enhancing interactions of heterogeneous information and capturing the global guidance of storyline.

4.4 Experimental Analysis

4.4.1 Ablation Study

To investigate the effectiveness of the proposed modules, ablative experiments are conducted in absence of KAN & GSM (KAGS-KG), KAN (KAGS-K), CCA (KAGS-C) and GSM (KAGS-G), respectively. The statistic results are presented in Table 2.

TABLE 2
Ablation study of the proposed model on the VIST dataset, here KAGS-KG, KAGS-K, KAGS-C and KAGS-G represent the model without KAN & GSM, KAN, CCA, and GSM, respectively. The bold font represents the best performance.

Metrics	KAGS-KG	KAGS-K	KAGS-C	KAGS-G	**KAGS**
BLEU-1	62.6	66.7	68.5	68.4	**70.1**
BLEU-2	37.7	41.7	42.5	42.0	**43.5**
BLEU-3	21.6	24.4	24.8	24.7	**25.2**
BLEU-4	12.7	14.4	14.6	14.5	**14.7**
METEOR	34.3	36.0	35.5	35.4	**36.2**
ROUGE_L	28.5	31.2	30.5	31.1	**31.4**
CIDEr	7.8	9.5	11.0	10.7	**11.3**

First, without GSM, the KAGS-G presents apparent performance degradation on the VIST dataset, particularly on the metrics of BLEU-1 and BLEU-2 with the evaluation scores being declined from 70.1 to 68.4 by 1.7%, from 43.5 to 42.0 by 1.5%, respectively. In addition, the visualized activation maps obtained by GSM are illustrated in Fig. 6, which proves that GSM can focus more attention on the global consistent regions while removing the semantic foreground and background interferences. Therefore, the statistic results prove the positive effects of GSM to capture the long-range dependencies for global guidance.

Fig. 6. Effectiveness of GSM to capture topic-aware global consistency. Top to down: input images, activation maps without GSM, activation maps with GSM.

Second, without CCA, the statistic results of KAGS-C also show obvious performance drop on all metrics, especially on the metrics of METEOR and ROUGE_L, the former score reduces from 36.2 to 35.5 by 0.7% and the latter score reduces from 31.4 to 30.5 by 0.9%, respectively. The ablative results verify the effectiveness of the designed CCA to establish the cross-modal interactions for visual and textual information enhancement.

Third, without KAN, all the metrics obtained by KAGS-K present significant decrease on the VIST dataset. Especially, KAGS outperforms KAGS-K by a large margin in terms of BLEU-1, BLEU-2 and CIDEr, with the scores being 70.1 versus 66.7, 43.5 versus 41.7 and 11.3 versus 9.5, respectively. It is worth noting that KAN can capture the external rich knowledge and explore the correlation of heterogeneous information, facilitating to more abundant and reasonable descriptions.

Finally, without KAN and GSM, the statistic performance of KAGS-KG has extreme decline in terms of all metrics, further demonstrating the superiority of the designed KAN and GSM to learn attentive multi-modal representation and global semantic tailored to the visual storytelling task.

4.4.2 Visualization Analysis

In order to better verify the effectiveness of GSM and KAN, the class activation map [53] of each image and the attention distributions of each image region during word generation are visualized in Fig. 6 and Fig. 7, respectively.

First, as aforementioned, the class activation map of each image is visualized in Fig. 6, where the class activation map is computed by $\mathbf{M}^n = \mathbf{C}^n \tilde{\mathbf{A}}^\top$ referenced from [53]. In the second line of Fig. 6, the model fails to discriminate the consistency among group images and suffers from the background clutters, such as wrongly localizing the people under the stage in the second image and introducing the background interferences in the third image. Nevertheless, the designed GSM can well capture the consistent characteristics of bride and groom in this image sequence and suppress the background clutters, thus again confirming the advantages of triggering the global semantic of group-wise features.

Second, several generated sentences of differentiate images are presented in Fig. 7 to illustrate the effectiveness of KAN, where the whiter the color of image regions are, the higher attention weights are given to these regions. When referring to generate the nouns (e.g., 'runners', 'street', 'flowers', 'woman', 'mountain'), the module prefers to assign higher weights to the relevant areas; when predicting the verbs, KAN often gives more valuable attention weights to both of the local and non-local areas of relative action. Moreover, the imaginative words can be assigned with higher attention scores by KAN according to surrounding environments. For example, in the first line of Fig. 7, the region corresponding to the noun 'runners' is highlighted by assigning higher attention weights, when generating the verb 'running', the module pays more attention on runners' legs as well as their whole bodies. In the second line of Fig. 7, the noun 'garden' doesn't significantly appear in this image, but higher weights are correctly assigned to the surrounding areas of the flower. The visualized examples further verify the merit of KAN of paying attention to important regions, meaningful actions and abstract areas.

4.5 Human Evaluation

The previous works [1], [52] have concluded that automatic evaluation metrics can not reflect the semantic properties of many stories (e.g., coherence and expressiveness), therefore

Fig. 7. Visualization of KAN, where the whiter color of an image area represents that higher attention weights are given to that area.

TABLE 3
Statistic results of human evaluation metrics, here the percentage numbers represent the confident scores of the tester believe that a model surpasses its opponent, and Tie means the tester can not choose the better story.

Methods	XE-ss vs KAGS			AREL vs KAGS			VSCMR vs KAGS			IRW vs KAGS		
Choice	XE-ss	KAGS	Tie	AREL	KAGS	Tie	VSCMR	KAGS	Tie	IRW	KAGS	Tie
Relevance	35.9%	**59.5%**	4.6%	38.2%	**51.0%**	10.8%	32.1%	**47.6%**	20.3%	36.0%	**42.9%**	21.1%
Expressiveness	27.1%	**66.4%**	6.5%	31.5%	**58.8%**	9.7%	33.5%	**45.2%**	21.3%	34.3%	**39.6%**	26.1%
Concreteness	32.8%	**60.9%**	6.3%	37.9%	**49.4%**	12.7%	30.8%	**44.3%**	24.9%	31.7%	**37.2%**	31.1%

human evaluation metrics [52] are further adopted for comparison in pairwise manner. Specifically, 150 photo albums with a total of 750 images from the VIST test dataset are randomly selected and two stories generated by KAGS and another competing method are presented for every volunteer, noting that the optional orders in each item are shuffled for fairness. Then each volunteer needs to choose a better story according to the metrics of relevance, expressiveness and concreteness. The detailed illustrations of these three criteria are defined as follows.

- **Relevance** describing the precise topic of happened activity in image sequence.
- **Expressiveness** generating the grammatical, imaginary, coherent and abundant sentences.
- **Concreteness** providing the narrative and concrete descriptions of image contexts.

Table 3 lists four comparison tests: XE-ss [1] *vs* KAGS, AREL [1] *vs* KAGS, VSCMR [52] *vs* KAGS, and IRW [42] *vs* KAGS. As seen from the results, it is obvious that the statistic results of KAGS are better than other competing methods in all the three metrics. Especially, the scores of KAGS are much higher than XE-ss by 23.6%, 39.3% and 28.1% in terms of relevance, expressiveness and concreteness, respectively. Compared with newest method IRW, the

proposed KAGS model also shows superior performances and achieves more significant advantages than the scores on automatic evaluation metrics. Thus, it can empirically prove that the generated stories of KAGS can better obtain the storyline of image sequence, produce the imaginative words and generate concrete descriptions, which can not be obviously revealed by automatic metrics.

5 CONCLUSION

A knowledge-enriched attention network with group-wise semantic for visual storytelling has been developed, which consists of two main novel designs: KAN and GSM. The proposed KAN is designed to leverage the external knowledge and visual information extracted to characterize the cross-modal interactions with attention mechanism. In order to obtain the storyline with global feature guidance, a novel GSM is devised to explore the group-wise semantic with second-order pooling. All these extracted multi-modal representations are then fed into the decoder for story generation. Finally, a one-stage encoder-decoder framework is established to optimize all these designed modules in an end-to-end manner. Extensive experiments on the VIST dataset have been carried out to demonstrate the superior performance of the proposed KAGS model as compared

with other state-of-the-art methods. The proposed KAGS scheme is capable of learning robust feature representations at regional and global levels to achieve superior performances. However, there is still some gaps between the storyline generated by KAGS and that of human storytellers who are trained to generate narrative stories with human language styles. We are working on taking this KAGS to its next level by considering the following three aspects: (1) investigating reinforcement learning rewards correlated with human evaluation to enhance natural expression, (2) studying more effective frameworks to accomplish visual storytelling in more sophisticated and realistic scenarios which contain much interference, and (3) generating dense visual storytelling under a complex scenario where the target image sequence contains multiple storylines.

REFERENCES

[1] X. Wang, W. Chen, Y.-F. Wang, and W. Y. Wang, "No metrics are perfect: Adversarial reward learning for visual storytelling," in *ACL*, 2018, pp. 899–909.

[2] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *ICCV*, 2019, pp. 4634–4643.

[3] Y. Jung, D. Kim, S. Woo, K. Kim, S. Kim, and I. S. Kweon, "Hide-and-tell: Learning to bridge photo streams for visual storytelling," in *AAAI*, 2020, pp. 11 213–11 220.

[4] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *CVPR*, 2020, pp. 10 971–10 980.

[5] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *CVPR*, 2020, pp. 10 578–10 587.

[6] Y. Luo, J. Ji, X. Sun, L. Cao, Y. Wu, F. Huang, C.-W. Lin, and R. Ji, "Dual-level collaborative transformer for image captioning," in *AAAI*, 2021.

[7] C.-C. Hsu, Z.-Y. Chen, C.-Y. Hsu, C.-C. Li, T.-Y. Lin, T.-H. Huang, and L.-W. Ku, "Knowledge-enriched visual storytelling," in *AAAI*, 2020, pp. 7952–7960.

[8] H. Chen, Y. Huang, H. Takamura, and H. Nakayama, "Commonsense knowledge aware concept selection for diverse and informative visual storytelling," in *AAAI*, 2021.

[9] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *CVPR*, 2019, pp. 10 685–10 694.

[10] Y. Zhong, L. Wang, J. Chen, D. Yu, and Y. Li, "Comprehensive image captioning via scene graph decomposition," in *ECCV*. Springer, 2020, pp. 211–229.

[11] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *CVPR*, 2017, pp. 5410–5419.

[12] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," in *AAAI*, 2017, pp. 4444–4451.

[13] R. Wang, Z. Wei, P. Li, Q. Zhang, and X. Huang, "Storytelling from an image stream using scene graphs," in *AAAI*, 2020, pp. 9185–9192.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 6000–6010.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[17] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Explain images with multimodal recurrent neural networks," in *NeurIPS Workshop*, 2014.

[18] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015, pp. 3156–3164.

[19] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the long-short term memory model for image caption generation," in *ICCV*, 2015, pp. 2407–2415.

[20] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015, pp. 3128–3137.

[21] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *CVPR*, 2015, pp. 2625–2634.

[22] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *CVPR*, 2017, pp. 7008–7024.

[23] F. Liu, X. Ren, Y. Liu, H. Wang, and X. Sun, "simnet: Stepwise image-topic merging network for generating detailed and comprehensive image captions," in *EMNLP*, 2018, pp. 137–149.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[25] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015, pp. 2048–2057.

[26] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *CVPR*, 2016, pp. 4651–4659.

[27] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *CVPR*, 2017, pp. 375–383.

[28] X. Li and S. Jiang, "Know more say less: Image captioning based on scene graphs," *IEEE Transactions on Multimedia*, vol. 21, no. 8, pp. 2117–2130, 2019.

[29] L. Yang, H. Wang, P. Tang, and Q. Li, "Captionnet: A tailor-made recurrent neural network for generating image descriptions," *IEEE Transactions on Multimedia*, vol. 23, pp. 835–845, 2020.

[30] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *CVPR*, 2018, pp. 6077–6086.

[31] Q. Huang, Z. Gan, A. Celikyilmaz, D. Wu, J. Wang, and X. He, "Hierarchically structured reinforcement learning for topically coherent visual story generation," in *AAAI*, 2019, pp. 8465–8472.

[32] P. Yang, F. Luo, P. Chen, L. Li, Z. Yin, X. He, and X. Sun, "Knowledgeable storyteller: A commonsense-driven generative model for visual storytelling." in *IJCAI*, 2019, pp. 5356–5362.

[33] J. Hu, Y. Cheng, Z. Gan, J. Liu, J. Gao, and G. Neubig, "What makes a good story? designing composite rewards for visual storytelling," in *AAAI*, 2020, pp. 7969–7976.

[34] T.-Y. Hsu, C.-Y. Huang, Y.-C. Hsu, and T.-H. Huang, "Visual story post-editing," in *ACL*, 2019, pp. 6581–6586.

[35] L. Yao, N. Peng, R. Weischedel, K. Knight, D. Zhao, and R. Yan, "Plan-and-write: Towards better automatic storytelling," in *AAAI*, 2019, pp. 7378–7385.

[36] G. Angeli, M. J. J. Premkumar, and C. D. Manning, "Leveraging linguistic structure for open domain information extraction," in *ACL*, 2015, pp. 344–354.

[37] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.

[38] Z. Gao, J. Xie, Q. Wang, and P. Li, "Global second-order pooling convolutional networks," in *CVPR*, 2019, pp. 3024–3033.

[39] X. Yang and I. Tiddi, "Creative storytelling with language models and knowledge graphs," in *CIKM Workshop*, 2020.

[40] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *ICML*, 2017, pp. 933–941.

[41] T.-H. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra *et al.*, "Visual storytelling," in *NAACL*, 2016, pp. 1233–1239.

[42] C. Xu, M. Yang, C. Li, Y. Shen, X. Ao, and R. Xu, "Imagine, reason and write: Visual storytelling with graph knowledge and relational reasoning," in *AAAI*, 2021.

[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[45] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.

[46] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *ACL Workshop*, 2005, pp. 65–72.

[47] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[48] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *CVPR*, 2015, pp. 4566–4575.

[49] Y. Liu, J. Fu, T. Mei, and C. W. Chen, "Let your photos talk: Generating narrative paragraph for photo stream via bidirectional attention recurrent neural networks," in *AAAI*, 2017, pp. 1445–1452.

[50] L. Yu, M. Bansal, and T. L. Berg, "Hierarchically-attentive rnn for album summarization and storytelling," in *EMNLP*, 2017, pp. 966–971.

[51] B. Wang, L. Ma, W. Zhang, W. Jiang, and F. Zhang, "Hierarchical photo-scene encoder for album storytelling," in *AAAI*, 2019, pp. 8909–8916.

[52] J. Li, H. Shi, S. Tang, F. Wu, and Y. Zhuang, "Informative visual storytelling with cross-modal rules," in *ACM MM*, 2019, pp. 2314–2322.

[53] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2016, pp. 2921–2929.