Knowledge-Enriched Attention Network With Group-Wise Semantic for Visual Storytelling

Tengpeng Li, Hanli Wang, Senior Member, IEEE, Bin He, and 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 stories. 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 visual storytelling datasets with both objective and subjective evaluation metrics demonstrate the superior performance of the proposed scheme as compared with other state-of-the-art methods. The source code of this work can be found in https://mic.tongji.edu.cn.

Index Terms—Encoder-decoder, group-wise semantic, knowledge-enriched attention, multi-modal decoder, visual storytelling

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 research topic in the fields of computer vision and multimedia computing. Different from visual captioning which devotes to describe the 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 the imaginative expression out of the images.

Knowledge-Enriched Attention Network

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) exploiting the valuable and abundant information of extracted features in a single image, and (2) providing the precise storyline about the event that occurred in the image sequence. On one hand, most visual captioning schemes focus on detecting visual features, where convolutional features [1], [2], [3] and object features [4], [5], [6] have been widely used in these schemes. Nevertheless, regional-visual features can merely detect intrinsic and superficial information, lacking the capability to explore diverse and creative representations that were not apparent from images. Several approaches [7], [8], [9], [10] introduced external knowledge by using 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 association of cross-modal information or only learned the implicit external content 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 coherent multi-sentences. Furthermore, the most serious problem of 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 storytelling. The proposed KAGS model first employs a CNN [15] and a Faster-RCNN [16] as encoders to extract convolutional features, semantic labels and regional object features from the input image stream. Then the semantic labels and the regional features are 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 a proposed cascade cross-modal attention module. The proposed KAN can achieve sufficient feature representations 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 a 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. The 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 work of visual storytelling. The proposed knowledge-enriched attention network with group-wise semantic model is detailed in Section 3. We present in Section 4 the statistical performances, ablative studies and visualization analyses. Finally, we conclude this paper in Section 5.

## 2 Related Work

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.

### 2.1 End-To-End Framework

End-to-end framework is popular due to its efficiency for generating stories in a unified structure. Wang et al. [11] proposed a classical visual storytelling framework that has been the most popular base structure of following studies. 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. However, this work was not capable of exploiting the valuable semantic information and establishing the compact temporal interaction of a group of images, thus the generation lacked rich and topic-aware expressions. To this end, Huang et al. [17] 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. 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 characteristics of story, Hu et al. [18] designed three human-like criteria combined with a reinforcement learning structure and achieved superior performances on human evaluation metrics. In order to generate abundant descriptions, a commonsense-driven generator [19] was employed to caption essential external messages for abundant multi-sentence expressions. Nevertheless, these aforementioned methods have the difficulty to simultaneously determine the accurate storyline and capture the abundant semantic representations in one stage, resulting in unsatisfied descriptions.

### 2.2 Multi-Stage Approach

Many multi-stage approaches emerged which aimed to strengthen the diversity and informativeness of visual storytelling in a separated manner. Hsu et al. [20] merged various extracted concepts into the decoder for more diverse descriptions. Yao et al. [21] designed a hierarchical framework to plan the storyline in the first stage and wrote the topic-based story in the second stage. However, these methods ignored to introduce the textual information, which was crucial to enhance the diversity and imagination of expressions. Therefore, several works [7], [8] introduced the external commonsense knowledge from the bases like OpenIE [22], Visual
and the corresponding label set contains the $S_n$. For more diverse descriptions, the label set is enlarged from external enhanced images. ConceptNet can be acquired.

The proposed KAGS 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 $I$, the model first employs the general object detection framework Faster-RCNN and ResNet backbones as the encoder to extract boxes of regional-visual features and high-level convolutional features, and then use a visual-language pretrained model to produce full stories. Nonetheless, the multi-stage methods were time-consuming and tended to generate incoherent sentences, which required to be further optimized. In this work, an end-to-end model is designed while considering efficiency, informativeness and coherence. Particularly, the proposed one-stage model can train and infer all modules in a unified fashion to promote its efficiency, and the attentive commonsense knowledge and global semantic are also introduced to improve 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 knowledge 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 $I = \{I_n\}_{n=1}^N$, as input, the task of visual storytelling aims to exploit the effective intra feature (image-to-image or text-to-text) and inter feature (image-to-text) representations of this image stream, producing a reasonable and coherent story with multiple descriptive sentences $S = \{S_n\}_{n=1}^N$. To tackle this issue, a novel KAGS model is elaborately designed to generate the story $S$ 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 $I$, the model first employs the general object detection framework Faster-RCNN and ResNet backbones as the encoder to extract boxes of regional-visual features $R = \{R_n\}_{n=1}^N$ and the corresponding label set $L = \{L_n\}_{n=1}^N$ with high confidence, and the high-level representations in the last convolutional layer $C = \{C_n\}_{n=1}^N$, respectively. Note that the semantic label set $L$ contains the semantic labels of all $N$ images, and the semantic label $L_n$ indicates the semantic labels of the $n$th image. Then, $L_n$ is fed into KAN to explore external knowledge. For the semantic label $L_n \in L$, the ConceptNet [12] is introduced to generate the knowledge concept $K_n$ from external enhanced knowledge base that can further boost the capability of absorbing imaginative and reasonable concepts, thus a group of knowledge concepts $K = \{K_n\}_{n=1}^N$ can be acquired. Moreover, to fully utilize the regional-visual features $R_n$ and the knowledge concepts $K_n$, a cascade cross-modal attention (CCA) module is designed to progressively model the dense semantic interactions of intra- and inter-features, outputting the enhanced knowledge concepts and attentive regional-visual features. The whole process is defined as...
\[ [K_p^n, R_p^n] = \mathcal{F}_{\text{cco}}(K^n, R^n), \] where \( \mathcal{F}_{\text{cco}}(\cdot) \) and \( P \) represent the function of CCA module and the number of cascade layers in CCA, respectively.

Moreover, the recent works [4], [24] 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 a global-visual aggregation \( \mathbf{A} = \mathcal{F}_{\text{pm}}(\mathcal{C}) \) is then produced to help capture the global guided semantic and avoid noisy interference. Finally, the model feeds \( K_p^n, R_p^n \) and \( \mathbf{A} \) into the multi-modal story decoder, generating the predicted sentence \( S^n \).

### 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 enhance the external priors from current knowledge repository and establish intra- and inter-density correlations of cross-modal features. In fact, several existing knowledge-based methods [8], [13], [25] for visual storytelling also utilized external implicit knowledge, but they only focused on the intra correspondence of textual concepts instead of considering the inter association of heterogeneous information that is crucial to visual storytelling. Different, 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, the enriched textual knowledge and the visual features can be assigned with higher attention weights in meaningful feature dimensions, facilitating to a more optimized visual storytelling estimation than the methods [8], [13], [25] only considering textual information.

#### 3.2.1 Knowledge Graph

To offer current storytelling datasets more imaginary and reasonable concepts, the proposed KAGS establishes commonsense knowledge graphs based on the semantic label set \( \mathcal{L} \) detected by Faster-RCNN [16]. Similar to [8], [19], KAGS adopts the generalized ConceptNet [12] as the knowledge extractor to collect numerous commonsense concepts with rich imagination, abundant emotions and objective facts. Specifically, the semantic labels of \( L^n \) are regarded as the seed concepts and each seed concept is used as a query to select the relative commonsense concepts in the ConceptNet. The number of the selected commonsense concepts of the \( n \)-th image is usually larger than 200, so the repetitive commonsense concepts are removed and the top-\( J \) candidates are selected according to the confident scores of ConceptNet. Finally, the knowledge concept \( \mathbf{K}^n = \{\mathbf{k}^{(n)}_j\}_1^J \) is constructed, where \( \mathbf{k}^{(n)}_j \) is composed of two entities and one edge relation.

#### 3.2.2 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 accuracy of the generated story. Recently, the method [9] investigates the visual-textual guided encoding pattern to 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. Furthermore, the X-Linear attention block [4] is utilized to selectively enhance the fine-grained visual and cross-modal representations. 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 enhanced output \( \mathbf{V} \in \mathbb{R}^{m \times d} \) of the X-Linear attention block is formulated as

\[ \mathbf{M} = \text{X-Linear}(\mathbf{M}_q, \mathbf{M}_k, \mathbf{M}_v). \]

In this schema, the X-Linear attention block is applied to both of the SA and CA units, followed by the function \( \text{LS}(\cdot) \) 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

\[ \text{SA}(\mathbf{F}_v) = \text{LS}(\text{X-Linear}(\mathbf{F}_v, \mathbf{F}_v, \mathbf{F}_v)), \]
\[ \text{SA}(\mathbf{F}_t) = \text{LS}(\text{X-Linear}(\mathbf{F}_t, \mathbf{F}_t, \mathbf{F}_t)). \]

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

\[ \text{CA}(\mathbf{F}_v, \mathbf{F}_t) = \text{LS}(\text{X-Linear}(\mathbf{F}_v, \mathbf{F}_v, \mathbf{F}_t)), \]

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

\[ [K_p^n, R_p^n] = \mathcal{F}_{\text{cco}}^{(p)}(K_p^n, R_p^n) \]
\[ = \text{CA}(\text{SA}(K_p^n), \text{SA}(R_p^n)), \text{SA}(R_p^n)), \]

![Fig. 2. The schematic diagram of the two attention units employed by the proposed CCA module, where the left unit is self-attention and the right unit is cross-attention.](image-url)
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 $p$th layer, respectively. For $F_{tea}^{(1)}$, we set the original input features $\mathbf{R}_p^n = \mathbf{R}^n$ and $\mathbf{K}_p^n = \mathbf{K}^n$. Finally, the outputs $[\mathbf{K}_p^n, \mathbf{R}_p^n] = F_{tea}^{(p-1)}([\mathbf{K}_{p-1}^n, \mathbf{R}_{p-1}^n])$ with Eq. (4) 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 a 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 the visual storytelling task is the lack of storyline, leading to the incoherent expressions of multiple sentences. To this end, a group-wise semantic module composed of several second order pooling algorithms is developed to capture the global consistent guidance.

#### 3.3.1 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, the width and the channel dimension of the feature tensor, respectively. SOP first introduces a $1 \times 1$ 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 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 representation by introducing a quadratic operator and thus can enable the model with the capacity of non-linear representation. As a consequence, SOP can strengthen the non-linear feature capability by learning higher-order dependencies of holistic representation [24], and GSM can capture the global consistent representation of group-wise features along the channel-wise dimension, facilitating the acquisition of topic-aware information for coherent and narrative descriptions.

#### 3.3.2 Group-Wise Semantic

In Fig. 1c, the GSM module first inputs every feature representation $\mathbf{C}_n \in \mathbb{R}^{h \times w \times d}$ into SOP with Eq. (5), and then the SOP outputs the processed tensor $\mathbf{C}_n \in \mathbb{R}^{1 \times 1 \times d}$. Afterwards, all processed tensors are sequentially concatenated into $\mathbf{A} = \{[\mathbf{C}_n]_{n=1}^N \} \in \mathbb{R}^{N \times 1 \times d}$, producing an initial group-wise semantic representation. Similarly, GSM again sends $\mathbf{A}$ into SOP with Eq. (5) to capture the long-range semantic associations along the channel-wise dimension, generating the global-visual aggregation $\mathbf{\hat{A}} \in \mathbb{R}^{1 \times 1 \times d}$ that can contribute to the subsequent multi-modal story decoder in Section 3.4, which can be formulated as

$$\mathbf{A} = F_{gan}(\mathbf{C}) = SOP([SOP(\mathbf{C}^n)]_n^{N}).$$

### 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 these multi-modal features, generating reasonable and coherent sentences of the final story. Fig. 4 illustrates the diagram of the proposed multi-modal story decoder. Specifically, in order to generate the $n$th sentence with various representations including the attentive regional-visual features $\mathbf{R}_p^n$, the enhanced knowledge concepts $\mathbf{K}_p^n$, and the global-visual aggregation $\mathbf{\hat{A}}$, the model first flattens $\mathbf{R}_p^n \in \mathbb{R}^{h \times w \times d}$ to $\mathbf{R}_p^n \in \mathbb{R}^{1 \times \frac{h \times w}{d}}$, $\mathbf{K}_p^n \in \mathbb{R}^{j \times d}$ to $\mathbf{K}_p^n \in \mathbb{R}^{1 \times \frac{j \times d}{d}}$ with the designed flattening layer composed of two linear layers and one softmax layer, resulting in the regional-visual indicator vector $\mathbf{R}_p^n$ and the knowledge indicator vector $\mathbf{K}_p^n$, where $M, J$ and $d$ denote the number of detected regional boxes, the number of graph relations and the number of feature channels, respectively.

To further exploit compact interactions of visual features, enriched contexts and word embeddings, a multi-modal story decoder is designed by combining the CA unit and LSTM to accomplish the inference of regional and global representations. Particularly, for the regional-visual information reasoning of the $n$th image at the time step $t$ (see the left side of Fig. 4), the decoder sends the previous regional hidden state $\mathbf{h}_n^{t-1}$, the knowledge indicator vector $\mathbf{K}_p^n$, the previous word embedding $\mathbf{w}_n^{t-1}$ and the regional-visual indicator vector $\mathbf{R}_p^n$ into LSTM, and outputs the current regional hidden state $\mathbf{h}_n^t$. Afterwards, the decoder considers $\mathbf{h}_n^t$ as the query of the CA unit, and $\mathbf{R}_p^n$ 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 $\mathbf{v}_n^t$ by encouraging the cross-
Fig. 4. Illustration of the proposed multi-modal story decoder. For the knowledge indicator vector $K^n$, the regional-visual indicator vector $R^n$, the global-visual indicator vector $A^n$, the previous regional hidden state $hr^n_{t-1}$, the previous global hidden state $ha^n_{t-1}$, and the previous word embedding $w^n_{t-1}$ as inputs, the decoder feeds these vectors into a two-stream structure by combining the CA unit and LSTM to obtain a set of vectors $vr^n_t$, $hr^n_t$, $va^n_t$, and $ha^n_t$. Finally, these vectors are concatenated and sent into the following layers to obtain the current word prediction $w^n_t$.

modal correlation between $R^n$ and $hr^n_t$, which can be formulated as

$$hr^n_t = \text{LSTM}(K^n \oplus w^n_{t-1} \oplus R^n, hr^n_{t-1}),$$

$$vr^n_t = \text{Embed}(CA(hr^n_t, R^n)),$$

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

$$ha^n_t = \text{LSTM}(K^n \oplus w^n_{t-1} \oplus A^n, ha^n_{t-1}),$$

$$va^n_t = \text{Embed}(CA(ha^n_t, A^n)).$$

Next, the contextual vector $v^n_t$ is calculated by concatenating $vr^n_t$, $hr^n_t$, $va^n_t$, and $ha^n_t$, followed with a GLU [26] and a linear layer, respectively. Finally, the contextual vector $v^n_t$ is fed into a softmax layer to generate the current word embedding $w^n_t$. The word generation probability can be formulated as

$$p(w^n_t | w^n_{t-1}) = \text{softmax}(v^n_t),$$

where the prediction $p$ is a probability distribution over the dataset’s vocabulary $V_s$. Finally, the word embedding $w^n_t$ is transformed into the word $w^n_t$ obtaining the sub-story $S^n = \{w^n_1, \cdots, w^n_T\}$ of story $S$, where $T$ represents the length of sub-story $S^n$.

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 dataset for story prediction. The cross-entropy loss is employed in the training stage as

$$L(\theta) = -\sum_{n}^{N} \sum_{t}^{T} \log(p^n_t (g^n_t | g^n_1, \cdots, g^n_{t-1})),$$

where $\theta$ indicates the set of optimized parameters during training, $g^n_t$ represents the $t$th word embedding in the ground-truth sub-story $g^n$. Eventually, the goal is to minimize the loss $L(\theta)$. In the inference stage, the model predicts the story using the beam search method.

4 EXPERIMENT

4.1 Implementation Detail

Following the previous works [1], [17], [27], the proposed KAGS model adopts the ResNet-152 [15] pretrained on the ImageNet [28] dataset for convolutional feature extraction and utilizes the Faster-RCNN [16] pretrained on the ImageNet [28] dataset and the Visual Genome [23] 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 $1 \times 2048$ 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 commonsense concepts is set as 20. Moreover, the number of detected regional boxes is set as 36, the dimension of word embedding is set as 1024, the feature dimension in the convolutional feature and the regional feature is 2048, and the number of candidate 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 [29] is employed with the initial weight decay $5 \times 10^{-4}$ and the learning rate $4 \times 10^{-4}$. The model is converged in only 30 epochs with the batch size equal to 50. Note that the model does not use any post processing such as reinforcement learning [30]. The words appearing more than 3 times in the training dataset are selected to build a storytelling vocabulary. During inference, the beam search strategy is applied with the beam size of 3 for visual storytelling prediction. The model is implemented with PyTorch\(^1\) using a Tesla V100 for acceleration.

4.2 Dataset and Automatic Metric Evaluation

Dataset. The Visual Storytelling (VIST) dataset [31] 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], three broken photos which do not contain any visual information are removed, and thus 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 stories annotated by humans as the reference. Besides VIST, in order to further evaluate the effectiveness and generality of the proposed method, the LSMDC [32] dataset is introduced to the visual storytelling task, which is

\(^1\) [Online]. Available: https://pytorch.org/
originally designed for the movie description task to generate coherent sentences according to the visual contents of multiple video clips. Specifically, given a group of five video clips, a visual storytelling model is required to generate a consecutive and expressive story. The LSMDC dataset has 128,085 clips from 200 movies, where 20,283 training groups, 1,486 validation groups and 2,018 public testing groups are selected for visual storytelling.

Automatic Evaluation Metric. Comprehensive experiments are conducted in terms of four automatic metrics including BLEU [33], METEOR [34], ROUGE_L [35] and CIDEr [36]. These metrics calculate the similarities and relevances between the predicted story and reference. Concluding in [31], 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 a number of state-of-the-art approaches including (1) seq2seq [31], an original model with RNN-based structure; (2) BARNN [37], a relational attended model with designed GRU; (3) h-attn-rank [38], 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 [39], a hierarchical image encoder-decoder model; (7) HSRL [17], a hierarchical reinforcement learning framework; (8) VSCMR [40], a conceptual exploration network; (9) ReCO-RL [18], a relevant context reinforcement learning method; (10) INet [3], an imaginative concept reasoning network; (11) SGVST [13], a scene-graph knowledge enhanced model; (12) IRW [27], a multi-graph knowledge reasoning framework; and (13) TAPM [41], a transitional adaptation pretrained model. For fair comparisons, we directly present the statistical results provided by the authors or conduct the experiments by the official source codes of these competing approaches.

4.3.1 Qualitative Results

Fig. 5 presents several visual comparisons between the proposed KAGS model and the methods AREL, VSCMR and ground-truth on the VIST dataset. Only parts of the extracted commonsense knowledge graphs are visualized due to the space limit, where the words highlighted in red and blue represent the semantic seeds and the commonsense words, respectively.

Fig. 5. Visualization of the comparison between the proposed KAGS and other state-of-the-art methods including AREL, VSCMR and ground-truth on the VIST dataset. Only parts of the extracted commonsense knowledge graphs are visualized due to the space limit, where the words highlighted in red and blue represent the semantic seeds and the commonsense words, respectively.
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 diversity of this story. Regarding the proposed KAGS, it can avoid this problem and generate sentences with different formats and styles. The adjective “beautiful” in the second sentence and the noun “picture” in the fifth sentence also benefit from knowledge graphs. 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 “the whole family posed for a picture .”, 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 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 presented in Table 1, where it can be observed that the statistical results of KAGS show better performances than the competing approaches by a large margin. Generally, on the VIST [31] dataset, 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 runner-up performance on BLEU-4. Specifically, KAGS achieves the BLEU-1 score of 70.3, the BLEU-2 score of 44.4, the BLEU-3 score of 25.5, the METEOR score of 36.3, the ROUGE_L score of 31.6 and the CIDEr score of 11.4, surpassing the scene graph based method SGVST [13] by 5.2%, 4.3%, 1.7%, 0.5%, 1.7% and 1.6%, respectively. Moreover, compared with the runner-up method IRW [27] that employs a number of external knowledge including scene graph, commonsense graph and event graph, the proposed KAGS model can achieve higher scores on most metrics. On the LSMDC [32] dataset, KAGS outperforms four competing methods in terms of all metrics and the scores are higher than the runner-up method TAPM [41] by 2.3%, 0.6%, 0.7%, 0.4%, 0.7%, 1.0% and 0.8%, respectively.

### 4.4 Human Evaluation

The previous works [1], [40] have concluded that automatic evaluation metrics cannot reflect the semantic properties of many stories (e.g., coherence and expressiveness), therefore human evaluation metrics [40] are further adopted for comparison in a pairwise manner. Specifically, five volunteers with professional English certificates and skills participate in the human evaluation. For every photo album, each volunteer needs to finish five comparison tests (i.e., XE-ss [1] versus KAGS, AREL [1] versus KAGS, VSCMR [40] versus KAGS, IRW [27] versus KAGS and GT versus KAGS). Since there are 150 photo albums in the VIST test dataset for evaluation, each volunteer needs to finish 150 × 5 = 750 comparisons. Each comparison contains two stories generated by KAGS and another competing method, and the volunteers choose a better story according to the metrics of relevance, expressiveness and concreteness. Note that the optional orders in each comparison are shuffled for fairness.

The 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 2 lists the results of the comparison tests. As seen from the results, it is obvious that the statistical results of KAGS are better than the other four competing methods on...
4.5 Experimental Analysis

4.5.1 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 statistical results are presented in Table 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 70.3 to 68.9 by 1.4%, from 44.4 to 42.8 by 1.6%, 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 statistical results prove the positive effects of GSM to capture the long-range dependencies for global guidance.

Second, without CCA, the statistical results of KAGS-C also show obvious performance drop on most metrics, especially on the metrics of METEOR and ROUGE_L, the former score reduces from 36.3 to 35.5 by 0.8% and the latter score reduces from 31.6 to 30.5 by 1.1%, respectively. The ablative results verify the effectiveness of the designed CCA to establish the cross-modal interactions for visual and textural information enhancement.

Third, without KAN, all the metrics obtained by KAGS-K present 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 70.3 versus 66.7, 44.4 versus 41.7 and 11.4 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 statistical performance of KAGS-KG has the extreme decline in terms of all metrics, further demonstrating the superiority of the designed KAN and GSM to learn the attentive multi-modal representation and the global semantic for visual storytelling.

4.5.2 Visualization Analysis

In order to better verify the effectiveness of GSM and KAN, the class activation map [42] of each image and the attention distributions of each image region during word generation are visualized in Figs. 6 and 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 $M’ = CA^T$ referenced from [42]. In the second line of Fig. 6, the model fails to discriminate the consistency among group images and suffers from the background

| Methods | XE-ss versus KAGS | AREL versus KAGS | VSCMR versus KAGS | IRW versus KAGS | GT versus KAGS |
|---------|------------------|-----------------|-------------------|----------------|--------|
| Choice  | XE-ss | KAGS | Tie | AREL | KAGS | Tie | VSCMR | KAGS | Tie | IRW | KAGS | Tie | GT | KAGS | Tie |
| Relevance | 33.7% | 58.9% | 7.4% | 34.6% | 53.1% | 12.3% | 30.8% | 49.4% | 19.8% | 34.6% | 41.2% | 24.2% | 46.3% | 31.2% | 22.5% |
| Expressiveness | 28.4% | 65.8% | 5.8% | 30.9% | 57.7% | 11.4% | 32.6% | 43.9% | 23.5% | 35.2% | 37.7% | 27.1% | 48.5% | 23.6% | 27.9% |
| Concreteness | 32.6% | 61.3% | 6.1% | 32.3% | 51.8% | 15.9% | 31.0% | 42.2% | 26.8% | 31.5% | 38.9% | 29.6% | 44.1% | 26.7% | 29.2% |
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 advantage 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 the surrounding environment. 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 the legs and bodies of the runners. 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.

### 4.5.3 Extension on Transformer

Many image captioning methods [4], [5], [6] have achieved superior performances by applying the transformer technology [14], which can establish long-term relationships of extracted features. In order to further improve the performance of KAGS for visual storytelling, a variant of KAGS (named KAGS-T) is proposed by replacing the LSTM-based decoder with a transformer-based decoder. Table 4 lists the comparative results of KAGS and KAGS-T on the VIST and LSMDC datasets, which shows that KAGS-T outperforms KAGS in terms of most metrics on these two datasets. For example, on the metrics of BLEU-1, BLEU-2 and CIDEr, the evaluation scores on the VIST dataset increase from 70.3 to 70.9 by 0.6%, from 44.4 to 44.9 by 0.5%, from 11.4 to 11.7 by 0.3%, respectively. The statistical comparison on these two datasets indicates the advantage of transformer structure to establish long-term relationships of multi-modal features. In the future work, we will further consider to apply the transformer technology in both the encoder and the decoder for visual storytelling.

### 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 utilize 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 and LSMDC datasets 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 are 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 Proc. Annu. Meeting Assoc. Comput. Linguistics. Hum. Lang. Technol., 2018, pp. 899–909.
[2] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, “Attention on attention for image captioning,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 4634–4643.
[3] Y. Jung, D. Kim, S. Woo, K. Kim, S. Kim, and J. S. Kweon, “Hide-and-tell: Learning to bridge photo streams for visual storytelling,” in Proc. AAAI Conf. Artif. Intell., 2020, pp. 11 213–11 220.
[4] Y. Pan, T. Yao, Y. Li, and T. Mei, “X-linear attention networks for image captioning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 10 578–10 587.
[5] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, “Meshed-memory transformer for image captioning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 10 588–10 596.
[6] Y. Luo et al., “Dual-level collaborative transformer for image captioning,” in Proc. AAAI Conf. Artif. Intell., 2021, pp. 2286–2293.
[7] C.-C. Hsu et al., “Knowledge-enhanced visual storytelling,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 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 Proc. AAAI Conf. Artif. Intell., 2021, pp. 999–1008.
[9] X. Yang, K. Tang, H. Zhang, and J. Cai, “Auto-encoding scene graphs for image captioning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 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 Proc. Eur. Conf. Comput. Vis., 2020, pp. 211–229.
[11] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, “Scene graph generation by iterative message passing,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 5410–5419.
[12] R. Spetter, J. Chin, and C. Havasi, “ConceptNet 5.5: An open multilingual graph of general knowledge,” in Proc. AAAI Conf. Artif. Intell., 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 Proc. AAAI Conf. Artif. Intell., 2020, pp. 9185–9192.
[14] A. Vasswani et al., “Attention is all you need,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2017, pp. 6000–6010.
[15] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–777.
[16] S. Ren, K. He, R. Girshick, and J. Sun, “ Faster R-CNN: Towards real-time object detection with region proposal networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 2921–2929.

Tengpeng Li received the BS and ME degrees from the School of Information and Control, Nanjing University of Information Science and Technology, Nanjing, China, in 2017 and 2020, respectively. He is currently working toward the PhD degree with the Department of Computer Science and Technology, Tongji University, Shanghai, China. His research interests include computer vision, multimedia computing, and machine learning.

Hanli Wang (Senior Member, IEEE) received the BE and ME degrees in electrical engineering from Zhejiang University, Hangzhou, China, in 2001 and 2004, respectively, and the PhD degree in computer science from the City University of Hong Kong, Kowloon, Hong Kong, in 2007. From 2007 to 2008, he was a research fellow with the Department of Computer Science, City University of Hong Kong. From 2007 to 2008, he was also a visiting scholar with Stanford University, Palo Alto, California. From 2008 to 2009, he was a research engineer with Precoad, Inc., Menlo Park, California. From 2009 to 2010, he was an Alexander von Humboldt research fellow with the University of Hagen, Hagen, Germany. Since 2010, he has been a full professor with the Department of Computer Science and Technology, Tongji University, Shanghai, China. His research interests include computer vision, multimedia signal processing, and machine learning.

Bin He received the BS degree in engineering machinery from Jilin University, Changchun, China, in 1996, and the PhD degree in mechanical and electronic control engineering from Zhejiang University, Hangzhou, China, in 2001, where he held a postdoctoral research appointment with the State Key Lab of Fluid Power Transmission and Control, from 2001 and 2003. He joined Tongji University in 2003. He is currently a professor with the Department of Control Science and Engineering, Tongji University, Shanghai, China, and with the Frontiers Science Center for Intelligent Autonomous Systems, Shanghai. His current research interests include intelligent robots, autonomous systems, intelligent perception, and wireless networks.

Chang Wen Chen (Fellow, IEEE) received the BS degree from the University of Science and Technology of China, in 1983, the MS degree from the University of Southern California, in 1986, and the PhD degree from the University of Illinois at Urbana-Champaign, in 1992. He is currently chair professor of visual computing with the Hong Kong Polytechnic University. He has previously served as an empire innovation professor of computer science and engineering with the University at Buffalo, State University of New York from 2008 to 2021. From 2017 to 2020, he served as dean with the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen. He was Allen Henry Endow chair professor with the Florida Institute of Technology from 2003 to 2007. He was on the faculty of Electrical and Computer Engineering, University of Rochester from 1992 to 1996 and on the faculty of Electrical and Computer Engineering, University of Missouri-Columbia from 1996 to 2003. He has been the editor-in-chief for IEEE Transactions on Multimedia from 2014 to 2016. He has also served as the editor-in-chief for IEEE Transactions and Systems for Video Technology from 2006 to 2009. He has been an editor for several other major IEEE Transactions and Journals, including the Proceedings of IEEE and IEEE Journal of Selected Areas in Communications. He has served as Conference Chair for several major IEEE, ACM and SPIE conferences related to multimedia, video communications and signal processing, including ACM MM2020 and ACM MM2018 in recent years. His research expands a broad range of topics in multimedia communication, Internet of Video Things, multimedia systems, image/video processing, machine learning and multimedia signal processing. He and his students have received nine (9) best paper awards or best student paper awards. He has also received several research and professional achievement awards, including Sigma Xi Excellence in Graduate Research Mentoring Award in 2003, Alexander von Humboldt Research Award in 2009, the University at Buffalo Exceptional Scholar C Sustained Achievement Award in 2012, the State University of New York System Chancellor's Award for Excellence in Scholarship and Creative Activities in 2016, and the Distinguished ECE Alumni Award from University of Illinois at Urbana-Champaign in 2019. He is an SPIE fellow since 2007.

For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/csdl.