IgSEG: Image-guided Story Ending Generation

Qingbao Huang[1,2,6], **Chuan Huang**[1], **Linzhang Mo**[1]
Jielong Wei[1], **Yi Cai**[2,5]*, **Ho-fung Leung**[3], **Qing Li**[4]
[1]School of Electrical Engineering, Guangxi University, Nanning, Guangxi, China
[2]School of Software Engineering, South China University of Technology, Guangzhou, China
[3]Dept. of Computer Sc. & Engin, The Chinese University of Hong Kong, Hong Kong, China
[4]Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
[5]Key Laboratory of Big Data and Intelligent Robot (SCUT), MOE of China
[6]Institute of Artificial Intelligence, Guangxi University, Nanning Guangxi, China
qbhuang@gxu.edu.cn, chuanhuangqy@gmail.com, mlz1997@163.com

Abstract

In this work, we propose a new task called Image-guided Story Ending Generation (IgSEG). Given a multi-sentence story plot and an ending-related image, IgSEG aims to generate a story ending that conforms to the contextual logic and the relevant visual concepts. In contrast to the story ending generation task, which generates open-ended endings, the major challenges of IgSEG are to comprehend the given context and image sufficiently, and mine the appropriate semantics from the image to make the generated story ending informative, reasonable, and coherent. To address the challenges, we propose a Multi-layer Graph convolution and Cascade-LSTM (MGCL) based model which mainly comprises of two collaborative modules: i) a multi-layer graph convolutional network to learn the dependency relations of sentences and the logical clue of the context; ii) a multiple context-image attention module to generate the story endings by gradually incorporating textual and visual semantic concepts. Our MGCL is thus capable of building logically consistent and semantically rich story endings. To evaluate the proposed model, we modify the existing VIST dataset to obtain the VIST-Ending dataset. Empirically, our MGCL outperforms all the strong baselines on both automatic and human evaluation.

1 Introduction

As two challenging subtasks of story generation, the story ending generation (SEG) and visual storytelling (Huang et al., 2016; Zhao et al., 2018) have attracted more attention recently. The former generates text-based story endings (Zhao et al., 2018; Li et al., 2018; Guan et al., 2019). While, the latter generates photo-streams-based stories (Huang et al., 2016; Wang et al., 2018; Hu et al., 2020) or one-image-based stories (Gaur, 2019). Distinctly,

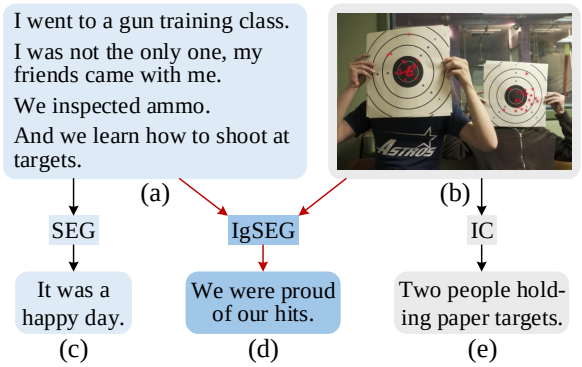

Figure 1: In SEG, existing methods tend to generate generic, safe, and inane story endings, e.g., (c). IgSEG is designed to generate specific, reasonable and informative endings induced by the given ending-related image. (d) is generated by the proposed MGCL model.

both of them are input with single-modal information merely. Actually, people often confront with demands to handle multi-modal inputs for generating a sentence or paragraph, e.g., *comments generation given a news story and an image* and *picture composition with a leading paragraph*. However, to our best knowledge, the SEG task incorporating a context and an image is still under-explored.

Furthermore, due to the limited textual information of the story context, the generated endings of SEG models remain tending to be generic, safe, and inane. To make the generation of story endings more coherent, specific, and informative, we consider introducing visual information to enrich the generation of story endings. For example (cf. Figure 1), the story context (a) mainly narrates that the experience of someone went for gun training. The story ending generated by SEG (c) just talks about the feeling (e.g., *happy*) of the day, which seems to be generic, safe, and unattractive for lack of interesting events, imaginative conception, and evocative plots. Meanwhile, Image Captioning (e) generates the description of a given image (b) with-

*Corresponding author: Cai Yi (ycai@scut.edu.cn).

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3114–3123

out any story context plot. Here, we introduce the image to induce the development of the story plot and guide the generation of the ending. The image-guided story ending (d) is associated with the senior semantic (e.g., *proud*) and events (e.g., *hits*) from the visual information. Obviously, this ending seems to be high-quality compared with the one generated by SEG.

We herein propose an Image-guided Story Ending Generation (IgSEG) task, which aims at generating a story ending with contextual plots and an ending-related image. Models need to comprehend the story plots and the image information, and grasp the visual semantic concepts strongly related to the story plots (e.g., *event*, *behavior*, and *emotion*). The main challenges of this task are three-fold: *(i)* How to accurately select and capture appropriate visual concepts matching the development trend of the story plot from the image. *(ii)* How to fuse the language and visual information and model inter- and intra-modality relations efficiently. *(iii)* How to make the utmost of high-level semantics mined from the image to write coherent, semantically-informative, and imaginative story endings.

To capture the text contextual plots and merge visual features effectively, we propose a Multi-layer Graph convolution and Cascade-LSTM (MGCL) model. A multi-layer graph convolution module is constructed to capture and encode the clues information (e.g., dependency relations (Zhang et al., 2018)) hidden in context. In detail, following (Huang et al., 2021), for each sentence, we construct a graph over the dependency parsing tree and conduct convolutional operations by Graph Convolution Networks (GCN). We then employ attention mechanism to compress each graph as one node and deliver the node from low layer to high layer for aggregation of inter-sentence information. Furthermore, inspired by the work (Anderson et al., 2018), we design a Multiple Context-Image Attention (MCIA) module to merge the contextual features and the image features. Specifically, we apply attention mechanism to weight sentence features and image features separately, then concatenate and feed them to the next Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) cell. We perform experiments on the VIST-Ending (VIST-E) dataset which is modified from the VIST (Huang et al., 2016).

Our contributions can be summarized as follows:

- We define a new task termed IgSEG to generate coherent, specific, and informative story endings guided by an ending-related image. To our best knowledge, this is the first story generation task with multi-modal inputs.
- We propose a model called MGCL, which employs multi-layer graph convolutional operations to capture story contextual plots and multiple context-image attentions to merge visual features effectively.
- Experiments show that our model outperforms several strong baselines on the VIST-E dataset. Human evaluations show that our model can generate story endings with better grammaticality, logicality, and relevance.

2 Related Work

The IgSEG task is related to *(i)* Story Ending Generation (SEG) and *(ii)* Visual Storytelling (VIST).

SEG (Zhao et al., 2018) is a subtask of story generation, which aims to understand the context and generate a coherent story ending. Many researchers have made great efforts on SEG. To improve the diversity and rationality of the generated story endings, (Li et al., 2018) tried to employ a Seq2Seq model based on adversarial training. Similarly, (Guan et al., 2019) made the model generate a reasonable ending by introducing external commonsense knowledge. Further, (Wang and Wan, 2019) adopted a transformer-based conditional autoencoder to capture contextual clues to improve coherence of story endings. (Guan et al., 2020) proposed a knowledge-enhanced pretraining approach for generating more reasonable stories. (Huang et al., 2021) proposed a multi-level GCN to capture the dependency relations of input sentences. Although previous studies have made great progress, due to the limitations of the SEG task itself, the generated endings tend to be generic and safe to some extent.

IgSEG is relevant to **VIST** as well. VIST aims to generate a coherent story according to an image stream. The main difficulty of VIST is how to generate image-relevant sentences. The previous works on VIST can be roughly divided into three categories. The first category focuses on designing specific model architectures to improve the quality of the generated stories (Kim et al., 2018; Wang et al., 2019). The second one generate more

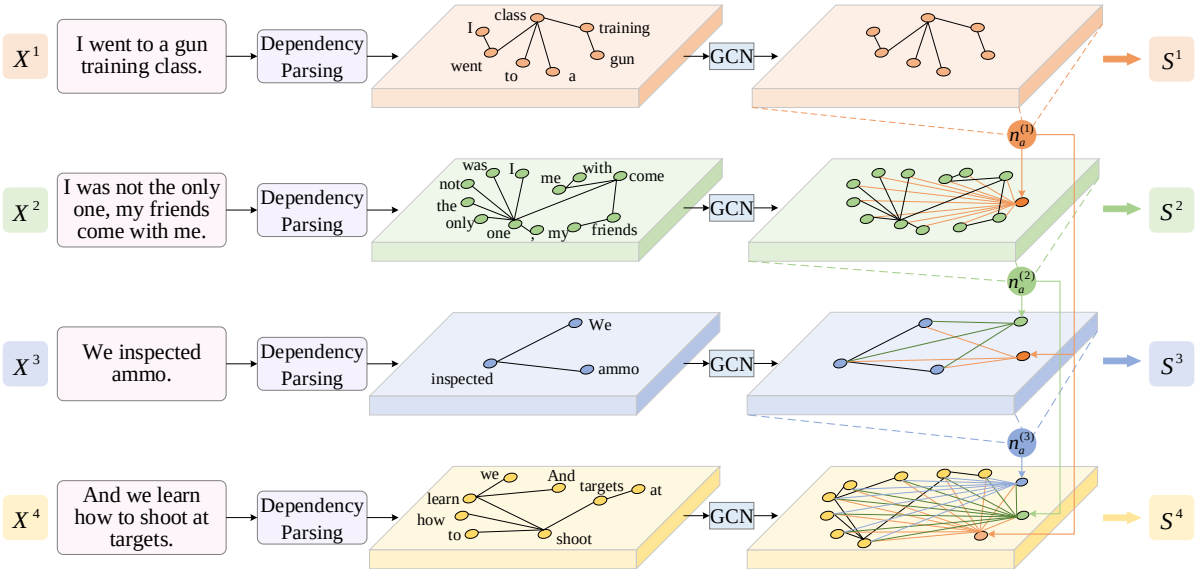

Figure 2: The Multi-layer Graph Convolutional Networks structure for context encoding. For intra-sentence information, each sentence is constructed as a graph over dependency parsing. For inter-sentence information, each graph is compressed as one node and is delivered to next layer. The output representation of each graph (i.e., S^1, S^2, S^3, S^4) are feed into decoder.

expressive output with reinforcement learning and adversarial training (Wang et al., 2018; Huang et al., 2019; Mo et al., 2019; Hu et al., 2020). The third one generates more common-sense stories by incorporating external knowledge. (Yang et al., 2019; Li et al., 2019; Wang et al., 2020; Jung et al., 2020).

However, the inputs of both SEG and VIST are single-modal information. The work on generating story endings given a textual sequence and an image simultaneously is unexplored. Therefore we propose the IgSEG task.

3 Methodology

3.1 Overview

The proposed IgSEG task aims to generate a story ending conforming the given contextual and visual information. Given a story context $X = \{X^1, X^2, \cdots, X^\mu\}$ and an ending-related image V, where $X^\mu = x_1^\mu x_2^\mu \cdots x_c^\mu$ represents the μ-th sentence with c words, IgSEG aims at generating a story ending $E = y_1 y_2 \cdots y_m$ with m words.

To generate the contextual-consistent and image-related story endings, we propose a Multi-layer Graph convolutional networks and Cascade-LSTM (MGCL) model based on the encoder-decoder framework. In the encoder, we propose a Multi-layer Graph Convolutional Networks (MGCN) over dependency trees to learn the context representation (cf. Figure 2), and we extract the image features with ResNet-152 (He et al., 2016). When

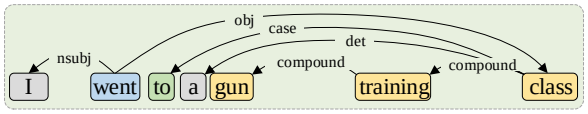

Figure 3: Dependency parsing of a sentence.

decoding, we generate the story ending with the cascaded LSTM framework. Specifically, we employ Top-Down LSTM to joint the context features and image features. And we devise a Multiple Context-Image Attention (MCIA) module to grasp the image-related context and contextual-relevant information of image for text generation. We will introduce each part of MGCL below.

3.2 Story Context Representation

Graph Construction We parse the sentences with Stanford Dependency tool (De Marneffe et al., 2014) (cf. Figure 3). To capture the relations of words in a sentence, we construct a graph G over the dependency parsing tree for each sentence. Regarding the words x as nodes $\mathcal{O}^k$, the word representation n_i as node feature, and the corresponding relations on the dependency parsing tree as edges ξ^k, the graph G^k of k-th sentence (k=1,2,3,4) can be constructed:

$$G^k = (\mathcal{O}^k, \xi^k). \quad (1)$$

Multi-Layer GCN To deliver inter-sentence information, we utilize attention mechanism to weight

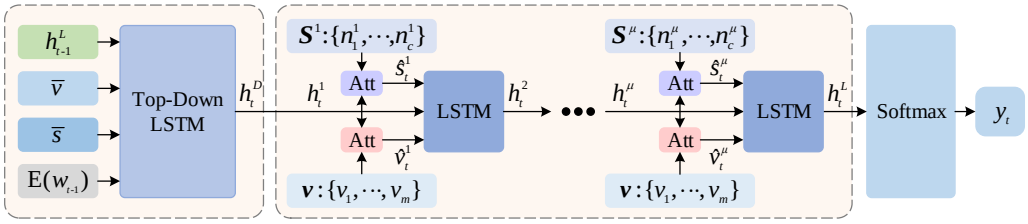

Figure 4: Illustration of the Cascade-LSTM framework. The Cascade-LSTM consists of TD and MCIA modules. The MCIA module consists of μ LSTM. The $\bar{v}$ denotes the mean-pooled image features, the $\bar{s}$ denotes the mean-pooled context features. The symbol "$\cdots$" denotes the omitted parts of MCIA module.

each node and sum them together as a new node $n_a^{(k)}$ for the $(k+1)$-th layer GCN (cf. Figure 2):

$$\boldsymbol{S}^k = [n_1^k \cdots n_c^k], \tag{2}$$

$$\theta = softmax(W_0^k \boldsymbol{S}^k + b_0^k), \tag{3}$$

$$n_a^k = \sum_{i=1}^{n} \theta n_i^k, \tag{4}$$

where n_i^k denotes the features of the i-th word of the k-th sentence, W_0 and b_0 are trainable parameters.

After updating the nodes of $(k+1)$-th layer GCN, the graph G^{k+1} structure is represented by a $(\lambda+k)\times(\lambda+k)$ adjacency matrix $\boldsymbol{A}^{k+1} = \{A_{ij}, (i,j) \in (\lambda+k)\}$. The corresponding value A_{ij} is 1 if the relation exists between node i and node j, otherwise it is 0. The representations of node i and its neighbor node $j \in \phi(i)$ are n_i^{k+1} and n_j^{k+1}, respectively. To obtain the correlation score w_{ij}^{k+1} between node i and node j, we learn a connected layer over concatenation of nodes features:

$$w_{ij}^{k+1} = w_{k+1}^T \sigma\left(W_1^{k+1}[n_i^{k+1}; n_j^{k+1}] + b_1^{k+1}\right), \tag{5}$$

where w_{k+1}, W_1^{k+1}, and b_1^{k+1} are trainable parameters, σ is the non-linear activation function, $(\cdot)^T$ denotes transpose operation, and $[;]$ denotes the concatenation operation.

We apply the softmax function over the correlation score w_{ij} to obtain the weight α_{ij}:

$$\alpha_{ij}^{k+1} = \frac{\exp(w_{ij}^{k+1})}{\sum_{j\in\phi(i)} \exp(w_{ij}^{k+1})}. \tag{6}$$

In the adjacency matrix $\boldsymbol{A}^{k+1}$, the value is α_{ij}^{k+1} if the relation exists between node i and node j, otherwise is 0. The A_{ij}^{k+1} can be denoted as:

$$A_{ij}^{k+1} = \begin{cases} \alpha_{ij}^{k+1} & \text{nodes } i,j \text{ are related} \\ 0 & \text{nodes } i,j \text{ are unrelated} \end{cases}. \tag{7}$$

For each node of the $(k+1)$-th GCN layer, we update the $(h+1)$-th representation of node n_i^{h+1} with aggregating the representations of h-th neighboring nodes n_j^h. This procedure is denoted as:

$$n_i^{h+1} = \sigma(A_{ii} n_i^h + \sum_{j\in\phi(i)} A_{ij}(W_2^h n_j^h + b_2^h)), \tag{8}$$

where W_2^h and b_2^h are trainable parameters. By l updates, the output S^{k+1} of GCN is denoted as:

$$\boldsymbol{S}^{k+1} = [n_1^{k+1} \cdots n_c^{k+1}]. \tag{9}$$

3.3 Decoder

The inputs of the decoder are the context features $\boldsymbol{S} = \{\boldsymbol{S}^k\}_{k=1}^4$ and the image features $\boldsymbol{v}$ extracted with the pre-trained model ResNet152 (He et al., 2016), as shown in Figure 4.

Top-Down LSTM (TD) Following previous work (Anderson et al., 2018), we employ Top-Down LSTM to incorporate the visual information (cf. Figure 4(a)). We operate LSTM over a single time step in the decoder with the following notation:

$$h_t = \text{LSTM}(x_t, h_{t-1}) \tag{10}$$

where x_t is the input vector of LSTM and h_t is the output vector. The inputs of TD module x_t^D consists of the previous output h_{t-1}^L of MCIA module, the mean-pooled image features $\bar{v}$, and the embedding of the previously generated word $E(w_{t-1})$, where t denotes the current time step.

To incorporate the context information, we modify the original inputs of TD. Firstly, we calculate the mean-pooled context features $\bar{s}$:

$$\bar{s} = \frac{1}{\mu}\frac{1}{c}\sum_{k=1}^{\mu}\sum_{i=1}^{c} n_i^k, \tag{11}$$

where μ denotes the number of sentences, c denotes the number of words of each sentence, n_i^k denotes the representation of the i-th word of the k-th sentence. Then, the vector x_t^D is denoted as:

$$x_t^D = [h_{t-1}^L; \bar{s}; \bar{v}; E(w_{t-1})]. \quad (12)$$

Multiple Context-Image Attention (MCIA) To merge the context features and image features, we devise the MCIA module. The MCIA module consists of four LSTM layers (cf. Figure 4(b)), which share all the parameters. The output h_t^D of TD is input to the MCIA module. Given the output h_t^k of $(k-1)$-th LSTM layer in MCIA module ($h_t^1 = h_t^D$), at each time step, we calculate the normalized attention weight $a_{i,t}^k$ for each of word representations n_i^k of the k-th sentence:

$$a_{i,t}^k = (w_a^k)^T \tanh(W_a^k n_i^k + W_a^{hk} h_t^k), \quad (13)$$

where w_a^k, W_a^k, and W_a^{hk} are trainable parameters. The convex combination of all words $\hat{s}_t^k$ can be calculated by n_i^k:

$$\beta_{i,t}^k = softmax(\boldsymbol{a}_t^k), \quad (14)$$

$$\hat{s}_t^k = \sum_{i=1}^{c} \beta_{i,t}^k n_i^k. \quad (15)$$

Likewise, we calculate the normalized weight $b_{i,t}^k$ for features v_i of each region of the image:

$$b_{i,t}^k = (w_b^k)^T \tanh(W_b^k v_i + W_b^{hk} h_t^k), \quad (16)$$

where w_b^k, W_b^k, and W_b^{hk} are trainable parameters. The convex combination of the image $\hat{v}_t^k$ can be calculated by the image features v_i:

$$\gamma_{i,t}^k = softmax(\boldsymbol{b}_t^k), \quad (17)$$

$$\hat{v}_t^k = \sum_{i=1}^{M} \gamma_{i,t}^k v_i, \quad (18)$$

where M denotes the the number of region of the image. We concatenate $\hat{s}_t^k$, $\hat{v}_t^k$, and h_t^k as inputs of the next LSTM layer:

$$x_t^{k+1} = [\hat{s}_t^k; h_t^k; \hat{v}_t^k]. \quad (19)$$

Given the output h_t^L of MCIA module, we calculate the conditional distribution over possible output words at each time step as follows:

$$p(y_t|y_{1:t-1}) = softmax(W_p h_t^L + b_p), \quad (20)$$

where W_p and b_p trainable parameters, and $y_{1:m}$ is the notation to refer to a sequence of words $(y_1, \cdots, y_m)$. Finally, the product of conditional distributions can be obtained by:

$$p(y_{1:m}) = \prod_{t=1}^{T} p(y_t|y_{1:t-1}). \quad (21)$$

4 Experiments

4.1 Dataset

To serve the IgSEG task, we modify the VIST dataset (Huang et al., 2016) to obtain a VIST-Ending (VIST-E) dataset, as shown in Table 1. Specifically, we keep the first four sentences, the ending sentence, and the last image of the photo stream of the VIST dataset. We have removed the stories which have corrupted images and rigmarole sentences over 40 words.

Dataset	Total	Training	Validation	Test
VIST	50,200	40,155	4,990	5,055
VIST-E	49,913	39,920	4,963	5,030

Table 1: Statistics of VIST and VIST-E.

4.2 Baselines

We compare our model with following models. **Seq2Seq** is a stack RNN-based model (Luong et al., 2015) with attention mechanisms. **Transformer** is a parallel model based solely on attention mechanisms (Vaswani et al., 2017). **IE+MSA** incorporates external knowledge with incremental encoding model for story ending generation(Guan et al., 2019). **T-CVAE** is a transformer-based conditional variational autoencoder for missing story plots generation (Wang and Wan, 2019). To adapt the IgSEG task, we rebuild above baselines by concatenating visual features as inputs. For fair comparison and testing our model, two variants of MGCL are created with the same inputs of the baselines. **MG+CIA**: We keep one Context-Image-Attention (CIA) unit in the decoder of MGCL. **MG+Trans**: We replace the decoder of MGCL with Transformer.

4.3 Evaluation Metrics

Automatic Evaluation We adopt four automatic metrics: **BLEU (B)** (Papineni et al., 2002) evaluates n-gram overlap between generated ending and a reference. **METEOR (M)** (Banerjee and Lavie, 2005) evaluates a generated sentence with

Model	B1	B2	B3	B4	M	C	R-L	Gram.	Logic.	Rele.
Seq2Seq† (Luong et al., 2015)	13.96	5.57	2.94	1.69	4.54	12.04	16.84	1.59	1.61	1.65
Transformer† (Vaswani et al., 2017)	17.18	6.29	3.07	2.01	6.91	12.75	18.23	3.01	2.15	1.96
IE+MSA† (Guan et al., 2019)	19.15	5.74	2.73	1.63	6.59	15.56	20.62	3.41	2.09	1.52
T-CVAE† (Wang and Wan, 2019)	14.34	5.06	2.01	1.13	4.23	11.49	15.51	1.89	1.76	1.25
MG+Trans† (ours)	19.43	7.47	3.92	2.46	7.63	14.42	19.62	3.46	2.77	2.60
MG+CIA† (ours)	20.91	7.46	3.88	2.35	7.29	19.88	21.12	2.80	2.35	1.97
MGCL (our full model)	**22.57**	**8.16**	**4.23**	**2.49**	**7.84**	**21.46**	**21.66**	**3.51**	**3.17**	**2.75**

Table 2: Experiments on the VIST-E dataset for the IgSEG task (p-value < 0.01). The **bold** / underline denotes the best and the second performance, respectively. † denotes the image features are directly concatenated.

Model	B2	B4	M	C	R-L
MGCL	**8.16**	**2.49**	**7.84**	**21.46**	**21.66**
w/o MGCN	7.11	2.03	7.23	20.26	19.74
w/o TD	5.18	1.04	6.74	10.33	18.93
w/o MCIA	7.17	2.17	7.13	17.62	19.75
w/o TD, MCIA	3.96	0.77	5.49	8.66	17.64

Table 3: Ablation studies. "w/o" means "without".

direct word-ordering. **CIDEr (C)** (Vedantam et al., 2015) evaluates the similarity of a generated sentence against the references by human consensus. **ROUGE-L (R-L)** (Lin, 2004) is applied to find the length of the longest common subsequence.

Human Evaluation Considering the limitation of automatic evaluation and the complexity of the IgSEG task, it is necessary to conduct human evaluation. The criteria of human evaluation includes three aspects: **Grammaticality (Gram.)** (Wang and Wan, 2019) evaluates correct, natural, and fluent of the generated story endings. **Logicality (Logic.)** (Wang and Wan, 2019) evaluates whether the story endings are reasonable and coherent. **Relevance (Rele.)** (Yang et al., 2019) measures how relevant the generated story endings and the input images are. We randomly pick 100 generated story endings from test-set for each model and employ three professional annotators skills to make evaluation. Following (Yang et al., 2019), we apply a 5-grade marking system, with 5 as the maximum grade and 1 as the worst. The final results are the average of the scores given by the three annotators.

4.4 Experimental Settings

The dimension of word embedding is 300 from GloVe.6B (Pennington et al., 2014). The update times of each GCN is 5, the maximum number of nodes in GCN is 43. The hidden layer dimension of all LSTM is 512. The number of LSTM layer is 4 in MCIA module. The dimension of image features is $7 \times 7 \times 2048$ from ResNet-152 (He et al., 2016). During training on the VIST-E dataset, the epoch is set to 30 and the batch size is 128. The optimizer is Adam with an initial learning rate of 4e-4. All baselines keep their own default settings. The dropout rate is 0.5. Specially, inputs of Seq2Seq, Transformer, IE+MSA, and T-CVAE are concatenated with context representations and image features.

4.5 Result Analysis

Automatic and Manual Evaluation

We perform experiments on the VIST-E dataset comparing with several strong baselines, i.e., Seq2Seq, Transformer, IE+MSA, and T-CVAE.

The results of automatic and manual evaluation are shown in Table 2. We have done significant test comparing our model with these baselines by running all these models ten times. The results shows that our model significantly outperforms them with all p-values < 0.01. Specifically, our model implements an improvement of 8.66 / 5.39 / 3.42 / 8.23 / 3.14 / 1.66 over the Seq2Seq / Transformer / IE-MAS / T-CVAE / MG+Trans / MG+CIA on B1. As for B4, our model achieves an improvement of 0.8 / 0.48 / 0.86 / 1.36 / 0.03 / 0.14 over the Seq2Seq / Transformer / IE-MAS / T-CVAE / MG+Trans/ MG+CIA. With respect to M, our model outperforms the Seq2Seq / Transformer / IE-MAS / T-CVAE / MG+Trans / MG+CIA by 3.3 / 0.93 / 1.25 / 3.61 / 0.21 / 0.55. And for R-L, our model implements an improvement of 4.82 / 3.43 / 1.04 / 6.15 / 2.04 / 0.54 over the Seq2Seq / Transformer / IE-MAS / T-CVAE / MG+Trans / MG+CIA. The results show that our MGCL model can comprehend the context better with the MGCN module, and merge the context features and image features effectively with the MCIA module.

Our MGCL model also outperforms baselines on all manual evaluation. Compared with the best baseline, Gram. increases from 3.46 to 3.51, Logic.

Model	B1	B2	B4	M	R-L
Seq2Seq	14.27	4.27	1.05	6.02	16.32
Transformer	17.06	6.18	1.57	6.55	18.69
IE+MSA	20.11	6.62	1.68	6.87	**21.27**
T-CVAE	20.36	6.63	1.88	6.74	20.98
Plan&Write	<u>20.92</u>	5.88	1.44	7.10	20.17
KE-GPT2	**21.92**	**7.40**	<u>1.90</u>	**7.41**	20.58
MG+Trans (ours)	18.55	<u>6.76</u>	**2.33**	<u>7.31</u>	19.02
MGCL (ours)	20.27	6.26	1.81	6.91	<u>21.01</u>

Table 4: Experiments on the VIST-E dataset (plain text) for the SEG task. The **bold** / <u>underline</u> denotes the best and the second performance, respectively.

increases from 2.77 to 3.17, and Rele. increases from 2.60 to 3.75. It shows that our model can generate the more coherent and reasonable story endings than other baselines. Notably, our model has a good performance on Rele., which shows that MCIA module is helpful for enhancing the link of the generated story endings with the images.

Ablation Study To explore the effectiveness of our MGCL, we perform the ablation experiments on VIST-E (cf. Table 3). When removing the MGCN module and using the hidden features of the previous LSTM directly, the performance of our model drops 0.46 on B4, 0.61 on M, 1.2 on C, and 1.92 on R-L, respectively. When removing the MCIA and using hidden features of TD directly, the performance drops 0.32 on B4, 0.71 on M, 3.84 on C, and 1.91 on R-L, respectively. When removing TD and MCIA and using a LSTM unit to decoder, the performance drops 1.72 on B4, 2.35 on M, 12.8 on C, and 4.02 on R-L. All of the these show that the MGCN module and MCIA module can help to generate the more contextual-consistent and image-related story endings.

Comparison on SEG To verify the effectiveness of image guidance, we conduct experiments on VIST-E dataset removing the image. The automatic evaluation results are shown in Table 4. Compared with the corresponding results in Table 2, the Seq2Seq, Transformer, and our models have poor performance overall, which indicates the image is helpful for generating better endings. But for IE+MSA and T-CVAE model, they have poor performance when adding the image. One possible reason is that they are designed for the textual story generation specially, so it is hard to change to generate better story endings with an image. Further, we also conduct the SEG experiments with another two recent models, the Plan&Write model (Yao et al., 2019) and the pretrained language KE-GPT2 model (Guan et al., 2020). The results show that KE-GPT2 achieves the best performance on the plain text dataset, while our models are close to it.

Task	Noun	Verb	Adjective
IgSEG	11,659	11,050	6,508
SEG	9,808	9,027	5,835

Table 5: Statistics of Noun, Verb, and Adjective in the generated sentences of our MGCL model.

To research the quantity transformation of part of speech, we count the number of Noun, Verb, and Adjective in the generated story endings (cf. Table 5). With the guidance of image, the story endings achieves an improvement of 18.87%, 22.41%, and 11.53% on Noun, Verb, and Adjective, respectively. The results indicates that the MGCL model can enrich story endings on the IgSEG task.

4.6 Visualization and Case Study

To explicitly demonstrate our model, we present the visualization and case study (cf. Figure 5).

The context (Figure 5(a)) is mainly about people going to the seaside for a holiday. The context is encoded by the MGCN module, and fed with the image features together into the MCIA module. Key words are marked in red by our model, the darker means more important. The entities, events, and emotion (e.g., *We*, *go*, *trip*, and *exciting*) are assigned more attention weights by our model. It shows that our model can understand the semantic information of context sufficiently. Similarly, we present visualization for image (Figure 5(b)). The regions in darker red are where our model focuses on. Our model pays more attention to the regions with the important objects (e.g., *bridge*, *city*, and *sky*, which can be regarded as *view*). It shows that our model can also capture the vital visual concepts in the image. As shown in Figure 5(c), we present the cases of baselines on SEG and IgSEG, respectively. On SEG, the Seq2Seq model generates a long ending with repeat words , but it is still reasonable. The Transformer and T-CVAE generate the generic endings with positive sentiment. Our model generates the sentence describing how "we" ended the day, which is more interesting. On IgSEG, all the models generate the more specific endings where the visual concepts are mentioned. These cases show that our model can capture context-relevant image concepts and generate informative story plots.

Model	SEG	IgSEG
Seq2Seq	At the end of the day, we went back to location location and view what a beautiful sky, it was a trip home from many times.	We went to the bridge.
Tramsformer	At the end of the day, we had a great time.	After that we decided to go back to the beach.
IE+MSA	We were all gone, but we had fun all the time.	We have the great time at the beach
T-CVAE	We were glad to be going to the future.	We these the tired enjoy my day in a nap.
MG+Trans(Ours)	We ended the day with a beautiful sunset.	It was a great day at the beach.
Ours†	We ended the day with a great view.	We have the great time at the beach.
Ours	We ended the day with a beautiful view of the water.	We ended the day by watching the **bridge** and enjoy the **view**.

(c)

Figure 5: Visualization and case study. The regions where are dark in color of sentences and images mean that model pays more attention to. GT denotes Ground True. (Best viewed in color)

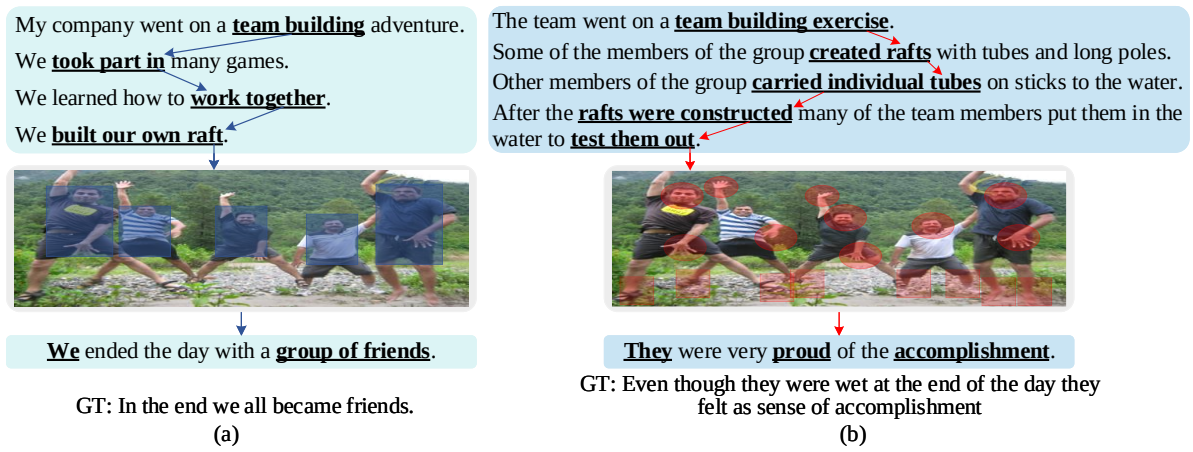

Figure 6: Different story context guided by the same ending-related image for IgSEG.

To vividly demonstrate the impact of image on IgSEG task, we show the generation cases (cf. Figure 6) which are offered the same ending-related image but the different story context. The content of image is that five people are very happy and jump in front of camera. The generated endings (a) and (b) are both coherent with their corresponding context. The context (a) has the logic chain (e.g., *team building* → *took part in games* → *work together* → *built raft*), and the context (b) has the logic chain (e.g., *team building* → *created raft* → *carried tubes* → *rafts test*). According to different logic chains, our model may focus on different regions of image and generate the story endings with the various semantics. The context (a) merely links to number of people in the image (e.g., *friends*), while the context (b) may be associated with people's postures and emotions (e.g., *dump* and *laugh* mean *proud*). To some extent, our MGCL model is able to capture some latent high-level semantics (e.g., *pride* and *celebration*) hidden in the image.

5 Conclusion

We propose a new task termed Image-guided Story Ending Generation. We transform the VIST dataset to VIST-Ending for IgSEG. We propose a MGCL model which uses a multi-layer graph convolutional networks to capture intra- and inter-sentence relations, a multiple context-image attention module to merge the context features and image features. Results on automatic and manual evaluation show that our model outperforms all the baselines.

Acknowledgements

We thank the anonymous reviewers for valuable comments and thoughtful suggestions.

This work was supported by National Natural Science Foundation of China (62076100, 51767005), and the collaborative research grants from the Fundamental Research Funds for the Central Universities, SCUT (D2210010, D2200150, and D2201300), the Science and Technology Planning Project of Guangdong Province (2017B050506004), the Science and Technology Programs of Guangzhou (201704030076, 201707010223, 201802010027, 201902010046), and the Hong Kong Research Grants Council, China (PolyU1121417, C1031-18G), and an internal research grant from the Hong Kong Polytechnic University, China (1.9B0V).

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. 2014. Universal stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4585–4592.

Shivam Gaur. 2019. Generation of a short narrative caption for an image using the suggested hashtag. In *2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW)*, pages 331–337. IEEE.

Jian Guan, Fei Huang, Minlie Huang, Zhihao Zhao, and Xiaoyan Zhu. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8(0):93–108.

Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6473–6480.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Junjie Hu, Yu Cheng, Zhe Gan, Jingjing Liu, Jianfeng Gao, and Graham Neubig. 2020. What makes a good story? designing composite rewards for visual storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7969–7976.

Qingbao Huang, Linzhang Mo, Pijian Li, Yi Cai, Qingguang Liu, Jielong Wei, Qing Li, and Ho-fung Leung. 2021. Story ending generation with multi-level graph convolutional networks over dependency trees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1–9.

Qiuyuan Huang, Zhe Gan, Asli Celikyilmaz, Dapeng Wu, Jianfeng Wang, and Xiaodong He. 2019. Hierarchically structured reinforcement learning for topically coherent visual story generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8465–8472.

Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross B. Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239.

Yunjae Jung, Dahun Kim, Sanghyun Woo, Kyungsu Kim, Sungjin Kim, and In So Kweon. 2020. Hide-and-tell: Learning to bridge photo streams for visual storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11213–11220.

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

Nanxing Li, Bei Liu, Zhizhong Han, Yu-Shen Liu, and Jianlong Fu. 2019. Emotion reinforced visual storytelling. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 297–305.

Zhongyang Li, Xiao Ding, and Ting Liu. 2018. Generating reasonable and diversified story ending using sequence to sequence model with adversarial training. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1033–1043.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Lingbo Mo, Chunhong Zhang, Yang Ji, and Zheng Hu. 2019. Adversarial learning for visual storytelling with sense group partition. In *Computer Vision – ACCV 2018*, pages 175–190.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Bairui Wang, Lin Ma, Wei Zhang, Wenhao Jiang, and Feng Zhang. 2019. Hierarchical photo-scene encoder for album storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8909–8916.

Ruize Wang, Zhongyu Wei, Piji Li, Qi Zhang, and Xuanjing Huang. 2020. Storytelling from an image stream using scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9185–9192.

Tianming Wang and Xiaojun Wan. 2019. T-cvae: Transformer-based conditioned variational autoencoder for story completion. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5233–5239.

Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. 2018. No metrics are perfect: Adversarial reward learning for visual storytelling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 899–909.

Pengcheng Yang, Fuli Luo, Peng Chen, Lei Li, Zhiyi Yin, Xiaodong He, and Xu Sun. 2019. Knowledgeable storyteller: A commonsense-driven generative model for visual storytelling. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5356–5362.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7378–7385.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215.

Yan Zhao, Lu Liu, Chunhua Liu, Ruoyao Yang, and Dong Yu. 2018. From plots to endings: A reinforced pointer generator for story ending generation. In *Natural Language Processing and Chinese Computing*, pages 51–63.