# Imagine, Reason and Write: Visual Storytelling with Graph Knowledge and Relational Reasoning

**Chunpu Xu[1], Min Yang[1*], Chengming Li[1], Ying Shen[2], Xiang Ao[3], Ruifeng Xu[4]**

[1]Shenzhen Key Laboratory for High Performance Data Mining,
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[2]School of Intelligent Systems Engineering, Sun Yat-Sen University
[3]Institute of Computing Technology, Chinese Academy of Sciences
[4]Harbin Institute of Technology (Shenzhen)
chunpuxu19@gmail.com, {min.yang, cm.li}@siat.ac.cn, sheny76@mail.sysu.edu.cn, aoxiang@ict.ac.cn,
xuruifeng@hit.edu.cn

## Abstract

Visual storytelling is the task of generating a short story to describe an ordered image stream. Different from visual captions, stories contain not only factual descriptions, but also imaginary concepts that do not appear in the images. In this paper, we propose a novel imagine-reason-write generation framework (IRW) for visual storytelling, inspired by the logic of humans when they write a story. First, a multimodal imagining module is leveraged to learn the imaginative storyline explicitly, improving the coherence and reasonability of the generated story. Second, we employ a relational reasoning module to fully exploit the external knowledge (commonsense knowledge base) and task-specific knowledge (scene graph and event graph) with a relational reasoning method based on the storyline. In this way, we can effectively capture the most informative commonsense and visual relationships among objects in images, enhancing the diversity and informativeness of the generated story. Finally, we integrate the visual information and semantic (concept) information to generate human-like stories. Extensive experiments on a benchmark dataset (i.e., VIST) demonstrate that the proposed IRW framework substantially outperforms the state-of-the-art methods across multiple evaluation metrics.

## Introduction

Visual storytelling (VST), which aims to generate a sequence of coherent sentences to describe an ordered image stream, has gained increasing attention in the vision and language communities. Different from visual captions, stories have more diverse structures and include imaginary concepts that do not explicitly appear in the image sequence. VST is challenging because it requires machines not only to fully understand semantic meaning of each image in a steam and the relations among the images, but also to possess the linguistic intelligence to generate the fluent paragraph and imaginary concepts for storytelling.

Most recent VST methods employ the sequence-to-sequence (seq2seq) models to generate the stories based on the semantics of images (Wang et al. 2018a; Huang et al. 2019; Jung et al. 2020). The common idea of these methods is to use the convolutional neural network (CNN) as an encoder to extract the visual feature of each image and fuse them together to obtain the whole image stream representation, and then feed this representational vector into a hierarchical long short-term memory (LSTM) so as to generate the narrative story. Seq2seq models have brought great improvements for VST and almost dominate this field, since they have the capacity to generate stories with good flexibility and quality.

Despite the remarkable progress of previous methods, there are still several technical challenges for visual storytelling. First, one major challenge for storytelling is how to learn the storyline (a sequence of coherent concepts) for the image stream, which can serve as a guidance for generating a subjective and imaginative story. However, most previous methods learn the main concepts for each image without considering the previously generated sentences (i.e., semantic features), which may hurt the coherence of the generated story. Second, most existing studies merely learn fine-grained visual features for images, which do not explicitly detect the objects and reason about their visual relationship. Fortunately, with advanced progress in deep learning and image recognition, higher-level visual understanding tasks such as scene graph construction have achieved noticeable success in various computer vision tasks (Teney, Liu, and van Den Hengel 2017; Johnson, Gupta, and Fei-Fei 2018), which may provide complementary strengths with symbolic reasoning for visual storytelling. Third, previous Seq2Seq models generate the stories solely from the original images. Such information is insufficient for generating diversity and imaginative stories since a certain story could contain not only the concepts described in the image, but also the imaginary concepts that do not appear in the image.

In this study, we aim at resolving the aforementioned challenges in a unified framework. We explore the symbolic knowledge (graph knowledge) and relation reasoning for visual storytelling, aiming to benefit from the complementary strengths of both symbolic reasoning and end-to-end multimodal feature mapping. To this end, we propose a novel

*Min Yang is corresponding author (min.yang@siat.ac.cn).

imagine-reason-write generation framework (IRW), inspired by the logic of humans when they write the story. **First**, we propose a multimodal imagining module to predict the imaginary concept (topic) for each image based on the current visual feature extracted by an encoder and the previously generated sentence, aiming to improve the coherence and reasonability of the generated story. **Second**, we employ a relational reasoning module to fully exploit the symbolic knowledge from commonsense knowledge graph (KG) and task-specific KG (scene graph and event graph) for improving the diversity and informativeness of the generated story. In particular, we first retrieve a set of candidate sub-graphs from the three kinds of graphs. Then, we extract the most informative graph knowledge from the sub-graphs via graph convolutional networks (GCNs) (Johnson, Gupta, and Fei-Fei 2018) as the complementary semantic information for story generation. **Third**, we design a guiding unit to integrate the visual features from images and the semantic features from the three kinds of graphs to generate a human-like story for the input image stream.

We summarize our main contributions as follows:

- We propose a novel imagine-reason-write generation framework for visual storytelling, which benefits from the complementary strengths of graph knowledge, relational reasoning, and multi-modal feature mapping.
- We propose a retrieval-enhanced method to build the event graph from the training corpus. The event graph learns the high-level events from the stories of similar images, which can provide auxiliary knowledge for the story generation.
- Experiments on a benchmark dataset demonstrate that IRW outperforms the compared methods by a substantial margin, across multiple evaluation metrics.

## Related Work

### Visual Captioning

Automatic image captioning is the task of generating a sentence to describe the image content. Inspired by the development of neural machine translation, most recent image captioning methods adopt the encoder-decoder framework (Xu et al. 2015; Lu et al. 2017; Rennie et al. 2017; Anderson et al. 2018). Xu et al. (2015) proposed an attention-based model that incorporated the spatial attention on image features into the decoding process. Lu et al. (2017) proposed a spatial attention model to capture spatial image features and utilized a sentinel gate to decide whether to attend to image features or rely on the language model. Rennie et al. (2017) presented a self-critical sequence training (SCST) method to directly optimize the model on non-differentiable metrics. Anderson et al. (2018) proposed a combined bottom-up and top-down attention mechanism, which calculated the attention at the level of objects and other salient image regions.

Recently, some previous works have been proposed to consider the problem of generating a long, coherent, and detailed story for an image since a single sentence caption only describes the coarse details of image content (Krause et al. 2017; Chatterjee and Schwing 2018; Zha et al. 2019). Krause et al. (2017) integrated the encoder-decoder framework and a hierarchical recurrent neural network, where a sentence RNN was used to compress the extracted visual features into sentence topic vectors, and then a word RNN was employed to generate a sentence based on each topic vector. Chatterjee and Schwing (2018) augmented the paragraph generator by leveraging coherence vectors to ensure topic smoothness between successive sentences and global topic vectors to summarize information about the image. Zha et al. (2019) utilized previous visual attention to capture visual context for compositional visual reasoning and decided whether to use the context for the word generation.

### Visual Storytelling

The goal of visual storytelling is to generate a human-level narrative to describe the photo stream. Compared with visual captions, the generated stories may contain more complex structured expressions and imaginary content that are inferred similar to the performance of a human. Park and Kim (2015) used a local coherence model to resolve the patterns of local transitions of discourse entities. This is a pioneering research for visual storytelling. Afterwards, Huang et al. (2016) released a large-scale benchmark dataset (i.e., VIST) for visual storytelling task, which inspired many following works in this area. For example, Liu et al. (2017) proposed a novel joint learning model by leveraging the semantic coherence in a photo stream. There have been increasing interests in utilizing reinforcement learning architecture for visual storytelling (Wang et al. 2018a,b), where a reward model was designed to evaluate the quality of generated stories. However, the training process of these models is inherently unstable. Recently, Yang et al. (2019) proposed a commonsense-driven generative model, which used a commonsense knowledge base by self-attention mechanism to generate informative stories. Wang et al. (2020) introduced a graph-based method to enhance visual storytelling based on relationships of both with-image level and cross-images level. Different from previous works, our method reasons over the external knowledge (i.e., commonsense knowledge base) and task-specific knowledge (i.e., scene graph and event graph) based on the imaginary storyline to generate coherent, reasonable and imaginary stories.

## Proposed Method

### Overall Architecture

Given an image stream $I = \{I_1, ..., I_M\}$, the goal of visual storytelling is to understand the event flow in the stream and then generate a coherent, human-level story $y = \{y_1, .., y_M\}$, where $M$ indicates the number of images in $I$ and each sentence $y_m = \{w_{m,1}, .., w_{m,T}\}$ consists of $T$ words.

The overall architecture of the proposed model is illustrated in Figure 1. Our model adopts the encoder-decoder framework as the backbone for visual storytelling. The image stream encoder consists of a CNN and bidirectional Gated Recurrent Unit (BiGRU), which encodes the image features and the information from context images in the

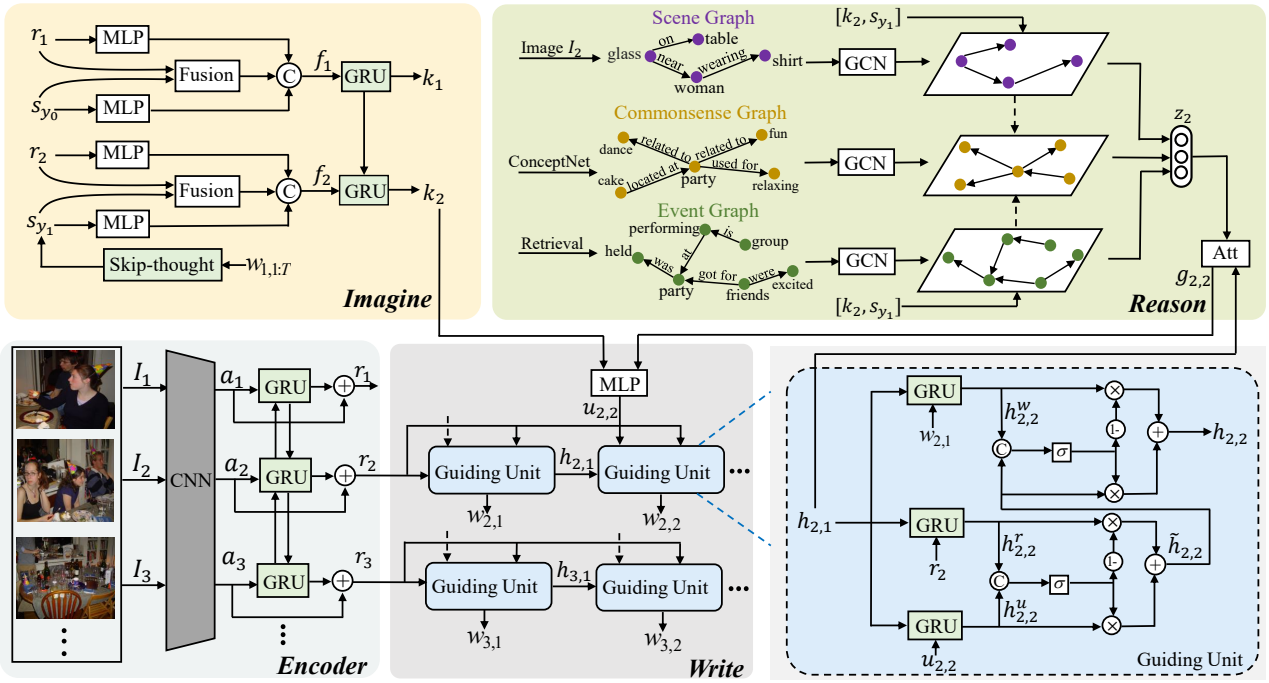


Figure 1: The overview of our model. The image stream encoder learns a visual cue vector for each image in the image steam via CNN and BiGRU, while the story decoder adopts an imagine-reason-write framework.

stream. The decoder aims to generate a coherent, informative and imaginative story word by word for the image stream. To this end, we propose a novel imagine-reason-write generation architecture. First, a multimodal imagining module is devised to generate an imaginative storyline. Second, we employ a relational reasoning module to fully exploit the external commonsense KG and task-specific knowledge (scene graph and event graph), and learn the complementary semantic features for storytelling. Finally, a story generation module with guiding units is devised to generates a coherent, informative, and imaginative story based on both visual and semantic features. Next, we describe the technical details of each component in the proposed model.

## Image Steam Encoder

First, we employ the ResNet-152 (He et al. 2016) pre-trained on ImageNet (Russakovsky et al. 2015) as our CNN encoder to learn the visual feature vector $\mathbf{a}_m$ of each image $I_m$. Then, BiGRU is deployed to process the visual features of a sequence of images sequentially and keep its hidden state through time. In this way, we can obtain the visual cue vectors $\mathbf{r} = \{\mathbf{r}_1, ..., \mathbf{r}_M\}$ that capture the temporal relations among images, where each cue vector $\mathbf{r}_m$ captures both the visual feature of image $I_m$ and the information from context images in the stream. Formally, the visual cue vector $\mathbf{r}_m$ for image $I_m$ is computed as:

$$\mathbf{a}_m = \text{CNN}(I_m) \tag{1}$$

$$\mathbf{h}^I_m = \text{BiGRU}(\mathbf{h}^I_{m-1}, \mathbf{a}_m) \tag{2}$$

$$\mathbf{r}_m = \text{ReLU}(\mathbf{h}^I_m, \mathbf{a}_m) \tag{3}$$

## Multimodal Imagining Module

Given the visual cue vectors $\mathbf{r}$, our first goal is to extract a key concept for each image in the photo stream, and an imaginative storyline is explicitly constructed based on the predicted concepts. However, it is difficult to learn imaginary concepts and coherent storyline by merely relying on the visual features extracted from images. Therefore, we design a *multimodal imagining module* to integrate visual cue vectors and previous generated sentence to imagine the abstract concept for each image. Formally, given the previously generated sentence $w_{m-1,1:T}$ (or $y_{m-1}$) for image $I_{m-1}$, a skip-thought model (Kiros et al. 2015), which maps a full sentence to a dense sentence vector, is utilized to calculate a skip-thought vector $\mathbf{s}_{y_{m-1}}$ as the sentence representation for $y_{m-1}$. Then, a multimodal fusion vector $\mathbf{f}_m$ is computed by:

$$\mathbf{f}_m = \varphi([\mathbf{W}_1\mathbf{r}_m, \mathbf{W}_2\mathbf{s}_{y_{m-1}}, g(\mathbf{W}_3\mathbf{r}_m + \mathbf{W}_4\mathbf{s}_{y_{m-1}})]) \tag{4}$$

where $\mathbf{W}_1$, $\mathbf{W}_2$ , $\mathbf{W}_3$ and $\mathbf{W}_4$ are the parameters to be learned. Both $\varphi$ and $g$ are feed-forward neural networks.

We use a GRU to produce a stroyline by generating one imaginary concept for each image in the image stream. Taking as input the multimodal fusion vector $\mathbf{f}_m$, the hidden state of GRU at time step $m$ is calculated by:

$$\mathbf{h}_m = \text{GRU}(\mathbf{f}_m, \mathbf{h}_{m-1}) \tag{5}$$

where $\mathbf{h}_m$ is the hidden state of GRU when generating the concept for the image $I_m$. The generation probabilities of imaginary key concept for image $I_m$ is then computed by:

$$k_m \sim p^c_m = \text{softmax}(\mathbf{W}_k\mathbf{h}_m + \mathbf{b}_k) \tag{6}$$

where $\mathbf{W}_k$ and $\mathbf{b}_k$ are learned parameters. The word with maximum probability are selected as the final imaginary concept $k_m$ for image $I_m$.

## Relational Reasoning Module

We propose a *relational reasoning module* to fully exploit the commonsense knowledge and task-specific knowledge (scene graph and event graph) with a relational reasoning method based on the storyline.

**Commonsense Knowledge Graph** To bridge the gap between the predicted imaginary concepts and imaginary sentences, we exploit the commonsense knowledge graph (KG) to obtain the supporting knowledge corresponding to the imaginary concepts. Similar to (Yang et al. 2019), we choose ConceptNet (Speer and Havasi 2012) as our commonsense KG, which contains a large-scale commonsense facts. Each fact in KG is composed of two entities, one relation, and a relation weight, which can be formulated as a triple (subject entity, relation, object entity). Formally, given the generated imaginary key concept $k_m$ of image $I_m$, we conduct entity mention detection by performing exact matching between $k_m$ and entities, and obtain top-$L$ candidate relationships from KG for $k_m$ based on the relation weights. Hence, a sub-graph $\mathcal{G}_m^C = (\mathcal{V}_m^C, \mathcal{E}_m^C)$ is built based on the retrieved candidate relationships for image $I_m$. $\mathcal{V}_m^C = \{v_{m,i}^C\}_{i=1}^{N_1}$ indicates the entity node set and each edge $e_{m,ij}^C \in \mathcal{E}_m^C$ represents the relationship between the entity pair. $N_1$ represents the number of entity nodes in the retrieved sub-graph.

**Scene Graph** Scene graph generation aims at automatically mapping an image into a structured graph representation, which requires detecting salient objects and their relationships in the images. Recently, various studies (Zellers et al. 2018; Tang et al. 2019) work on scene graph networks and achieve remarkable progress. Inspired by these works, we employ the Faster R-CNN detector to generate a set of object proposals for each image, and then compute dynamic tree structures (Tang et al. 2019) to encode the object proposals into visual context for predicting the relationship between each object pair. The scene graph generation model is pre-trained on the Visual Genome dataset, which is then utilized directly to generate a scene graph $\mathcal{G}_m^S = (\mathcal{V}_m^S, \mathcal{E}_m^S)$ for each image $I_m$, where $\mathcal{V}_m^S = \{v_{m,i}^S\}_{i=1}^{N_2}$ represents the visual node set and $e_{m,ij}^S \in \mathcal{E}_m^S$ represents the relation edge between node $v_{m,i}^S$ and $v_{m,j}^S$. $N_2$ indicates the number of entity nodes in the retrieved scene sub-graph.

**Event Graph** We first describe the event extraction process for each image, and then introduce the operation of constructing the event graph. First, we develop a retrieval-enhanced method to retrieve top-$R$ visually similar images from the training set for each image in the image stream by exhaustively computing the cosine similarity between the query image and the training images. Then, the description sentences of the retrieved similar images are concatenated to form a guidance story, which is exploited to construct the event graph. Specifically, given the retrieved guidance story, we apply the Stanford Open IE approach (Angeli, Premkumar, and Manning 2015) to extract an event for each sentence. Each event can be represented as a relational triple $(e_1, r, e_2)$, where $e_1$ is the subject entity, $e_2$ is the object entity, and $r$ is the relation between $e_1$ and $e_2$. After obtaining all the events for the image $I_m$, we extract the consensus events that are representative in the event set $D$. In particular, we first compute the semantic similarity between each event $d_i$ and the other event $d_j$ in $D$:

$$score_i = \frac{1}{\mid D \mid} \sum_{d_j \in D} \text{cosine}\left(\mathbf{s}_{d_i}, \mathbf{s}_{d_j}\right) \tag{7}$$

where $score_i$ is the consensus score for event $d_i$. $\mathbf{s}_{d_i}$ and $\mathbf{s}_{d_j}$ are skip-thought vectors of event $d_i$ and $d_j$, respectively. The top-$K$ events (relational triples) with highest consensus scores are collected to form the event graph $\mathcal{G}_m^E = (\mathcal{V}_m^E, \mathcal{E}_m^E)$ for the image $I_m$, where $\mathcal{V}_m^E = \{v_{m,i}^E\}_{i=1}^{N_3}$ is the entity node set, and $e_{m,ij}^E \in \mathcal{E}_m^E$ indicates the relation between $v_{m,i}^E$ and $v_{m,j}^E$. $N_3$ is the number of entity nodes in $\mathcal{V}_m^E$.

**Relational Reasoning from Graphs** We reason over the retrieved commonsense KG $\mathcal{G}_m^C = (\mathcal{V}_m^C, \mathcal{E}_m^C)$, the scene graph $\mathcal{G}_m^S = (\mathcal{V}_m^S, \mathcal{E}_m^S)$, and the event graph $\mathcal{G}_m^E = (\mathcal{V}_m^E, \mathcal{E}_m^E)$, to capture relevant and appropriate knowledge, which help to improve the diversity and informativeness of the generated knowledge. The three kinds of graphs share the same relational reasoning process. Due to limited space, we merely introduce the details of relational reasoning over the event graph $\mathcal{G}_m^E$ for conciseness.

Specifically, we utilize a GCN to aggregate information along edges of the knowledge. Given a triple $(v_{m,i}^E, e_{m,ij}^E, v_{m,j}^E)$, each node and relation entity are first transformed into vector representations by an embedding layer, and the triple is represented as $(\mathbf{v}_{m,i}^E, \mathbf{e}_{m,ij}^E, \mathbf{v}_{m,j}^E)$. Since the node entity $\mathbf{v}_{m,i}^E$ can act as subject or object in different relationship triples, we encode the entity $\mathbf{v}_{m,i}^E$ by considering all the neighbor nodes of $\mathbf{v}_{m,i}^E$. Similar to (Johnson, Gupta, and Fei-Fei 2018), we compute the subject entity vector $\overrightarrow{\mathbf{x}_{m,i}^E}$ for the entity $\mathbf{v}_{m,i}^E$ by encoding over the edges which start at $\mathbf{v}_{m,i}^E$. Similarly, the object entity vector $\overleftarrow{\mathbf{x}_{m,i}^E}$ for the entity $\mathbf{v}_{m,i}^E$ is calculated by encoding over the edges which end at $\mathbf{v}_{m,i}^E$. Then, the encoded entity vector $\mathbf{x}_{m,i}^E$ is the average of $\overrightarrow{\mathbf{x}_{m,i}^E}$ and $\overleftarrow{\mathbf{x}_{m,i}^E}$:

$$\overrightarrow{\mathbf{x}_{m,i}^E} = \sum_{\mathbf{v}_{m,i}^E \in \mathcal{S}_{m,i}^E} \rho(\mathbf{W}_s[\mathbf{W}_5\mathbf{v}_{m,i}^E, \mathbf{W}_6\mathbf{e}_{m,ij}^E, \mathbf{W}_5\mathbf{v}_{m,j}^E]) \tag{8}$$

$$\overleftarrow{\mathbf{x}_{m,i}^E} = \sum_{\mathbf{v}_{m,i}^E \in \mathcal{O}_{m,i}^E} \rho(\mathbf{W}_o[\mathbf{W}_5\mathbf{v}_{m,l}^E, \mathbf{W}_6\mathbf{e}_{m,li}^E, \mathbf{W}_5\mathbf{v}_{m,i}^E]) \tag{9}$$

$$\mathbf{x}_{m,i}^E = \frac{1}{\mathcal{N}_{m,i}^E}\left(\overrightarrow{\mathbf{x}_{m,i}^E} + \overleftarrow{\mathbf{x}_{m,i}^E}\right) \tag{10}$$

where $\mathbf{W}_5$,$\mathbf{W}_6$,$\mathbf{W}_s$,$\mathbf{W}_o$ are the learned parameters, and $\rho$ is an ReLU activation function. $\mathcal{S}_{m,i}^E$ indicates the triple set where $\mathbf{v}_{m,i}^E$ plays a subject role while $\mathcal{O}_{m,i}^E$ indicates the triple set where $\mathbf{v}_{m,i}^E$ plays a object role. $\mathcal{N}_{m,i}^E = |\mathcal{S}_{m,i}^E| + |\mathcal{O}_{m,i}^E|$ represents the number of knowledge triples associated with $\mathbf{v}_{m,i}^E$. Hence, after processing the whole event knowledge graph via GCN, the original node entities $\{\mathbf{v}_{m,i}^E\}_{i=1}^{N_3}$ are converted into event graph embeddings $\{\mathbf{x}_{m,i}^E\}_{i=1}^{N_3}$. Similarly, we obtain the relational commonsense

KG embeddings $\{\mathbf{x}^C_{m,i}\}^{N_1}_{i=1}$ and scene graph embeddings $\{\mathbf{x}^S_{m,i}\}^{N_2}_{i=1}$ by reasoning over commonsense KG and scene graph, respectively.

We devise an attention mechanism to selectively attend to the event graph embeddings $\{\mathbf{x}^E_{m,i}\}^{N_3}_{i=1}$ by using the imaginary concept $k_m$ and previous generated sentence representation $\mathbf{s_{y_{m-1}}}$ as attention source. Formally, the attended event graph vector $\mathbf{q}^E_m$ is computed over $\{\mathbf{x}^E_{m,i}\}^{N_3}_{i=1}$ as follows:

$$\mathbf{q}^E_m = \sum_{i=1}^{N_3} \alpha^E_{m,i} \mathbf{x}^E_{m,i} \tag{11}$$

$$\alpha^E_{m,i} = \frac{\exp\left(\mu^E_{m,i}\right)}{\sum_{i=1}^{N_3} \exp\left(\mu^E_{m,j}\right)},\ \mu^E_{m,i} = \zeta([\mathbf{W}_c k_m, \mathbf{s_{y_{m-1}}}], \mathbf{x}^E_{m,i}) \tag{12}$$

where $\alpha^E_{m,i}$ indicates the attention weight assigned to $\mathbf{x}^E_{m,i}$. $\zeta(\cdot)$ is a feed-forward neural network. $\mathbf{W}_c$ is the word embedding matrix for each input word. Similarly, we can obtain an attended scene graph vector $\mathbf{q}^S_m$ from $\{\mathbf{x}^E_{m,i}\}^{N_2}_{i=1}$.

Then, the combination of attended scene graph vector $\mathbf{q}^S_m$ and event graph vector $\mathbf{q}^E_m$ is utilized to reason over commonsense knowledge graph. We also devise an attention mechanism to selectively attend to the commonsense graph embeddings $\{\mathbf{x}^C_{m,i}\}^{N_1}_{i=1}$ by using $[\mathbf{q}^S_m;\mathbf{q}^E_m]$ as attention source. Formally, the attended commonsense graph vector $\mathbf{q}^C_m$ is computed as follows:

$$\mathbf{q}^C_m = \sum_{i=1}^{N_1} \alpha^C_{m,i} \mathbf{x}^C_{m,i} \tag{13}$$

$$\alpha^C_{m,i} = \frac{\exp\left(\mu^C_{m,i}\right)}{\sum_{i=1}^{N_1} \exp\left(\mu^C_{m,j}\right)},\ \mu^C_{m,i} = \delta([\mathbf{q}^S_m, \mathbf{q}^E_m], \mathbf{x}^C_{m,i}) \tag{14}$$

where $\alpha^C_{m,i}$ is the attention weight assigned to $\mathbf{x}^C_{m,i}$, $\delta(\cdot)$ is a feed-forward neural network.

Finally, we can obtain the relational knowledge vector $\mathbf{z}_m$ by concating the scene graph vector $\mathbf{q}^S_m$, commonsense graph vector $\mathbf{q}^C_m$, and event graph vector $\mathbf{q}^E_m$:

$$\mathbf{z_m} = [\mathbf{q}^S_m, \mathbf{q}^C_m, \mathbf{q}^E_m] \tag{15}$$

## Story Generation Module

Our IRW model aims to generate a coherent, reasonable, and imaginative story for an image stream. Each sentence in the story corresponds to an image in the photo stream. In this section, we elaborate on the decoding process for generating the sentence $y_m$ for each image $I_m$ in the image stream.

First, we employ attention mechanism to calculate an attended graph vector $\mathbf{g}_{m,t}$ based on previous hidden state $\mathbf{h}_{m,t-1}$ and the relational knowledge vector $\mathbf{z}_m$:

$$\mathbf{g}_{m,t} = \sum_{i=1}^{3} \gamma_{m,i} \mathbf{z}_{m,i} \tag{16}$$

$$\gamma_{m,i} = \frac{\exp\left(\xi_{m,i}\right)}{\sum_{i=1}^{3} \exp\left(\xi_{m,j}\right)},\ \xi_{m,i} = \phi(\mathbf{h}_{m,t-1}, \mathbf{z}_{m,i}) \tag{17}$$

where $\phi(\cdot)$ is a two-layer neural network that projects a vector into a scalar value. In addition, we can obtain the semantic cue vectors $\mathbf{u}_{m,t}$ via a multilayer perceptron (MLP) by taking as input the combination of the attended graph vector $\mathbf{g}_{m,t}$ and the embedding of generated concept $k_m$:

$$\mathbf{u}_{m,t} = \mathrm{MLP}([\mathbf{g}_{m,t}, \mathbf{W}_c k_m]) \tag{18}$$

where $\mathbf{W}_c$ is defined in Eq. (12).

Then, to automatically integrate the visual cue vectors $\mathbf{r}_m$ and semantic cue vectors $\mathbf{u}_{m,t}$ based on the previous generated word, we propose a *guiding unit* to generate the story by integrating the visual cue vectors $\mathbf{r}_m$ and semantic vectors $\mathbf{u}_m$ deeply. Specifically, the guiding unit extends the single GRU decoder to three GRUs (i.e., word GRU, vision GRU and semantics GRU) that are designed to model the previously generated word $w_{m,t-1}$, the visual cue vector $\mathbf{r}_m$, and the semantic cue vector $\mathbf{u}_{m,t}$ respectively, and outputs corresponding hidden states $\mathbf{h}^w_{m,t}$, $\mathbf{h}^r_{m,t}$, $\mathbf{h}^u_{m,t}$.

Inspired by the update gate in GRU, we then summarize the knowledge of visual clue and semantic cue by integrating the hidden states $\mathbf{h}^r_{m,t}$ and $\mathbf{h}^u_{m,t}$ into a final cue hidden state $\tilde{\mathbf{h}}_{m,t}$:

$$\tilde{\mathbf{h}}_{m,t} = \boldsymbol{\beta}_{m,t} \mathbf{h}^r_{m,t} + (1 - \boldsymbol{\beta}_{m,t}) \mathbf{h}^u_{m,t} \tag{19}$$

$$\boldsymbol{\beta}_{m,t} = \sigma(\mathbf{W}_7[\tanh(\mathbf{W}_8 \mathbf{h}^r_{m,t}), \tanh(\mathbf{W}_9 \mathbf{h}^u_{m,t})] \tag{20}$$

where $\mathbf{W}_7, \mathbf{W}_8, \mathbf{W}_9$ are learnable parameter, $\sigma$ is the sigmoid function. Additionally, we also integrate the final cue hidden state $\tilde{\mathbf{h}}_{m,t}$ and word hidden state $h^w_{m,t}$ into a final hidden state $\mathbf{h}_{m,t}$.

Finally, the generation probability of the $t$-th word in the $m$-th sentence is computed by:

$$p^w_{m,t} = \mathrm{softmax}(\mathbf{W}\mathbf{h}_{m,t} + \mathbf{b}) \tag{21}$$

where $\mathbf{W}$ and $\mathbf{b}$ are parameters to be learned.

## Training Procedure

Overall, given the input photo stream with $M$ images, our training objective is to (i) maximize the log-likelihood of storyline concepts for the story and (ii) maximize the log-likelihood of ground-truth story. For the story generation, we have the parallel training data. However, the storyline information is not provided in the corpus. In this paper, we employ the graph-based ranking RAKE method (Rose et al. 2010) to extract the abstract concepts (storyline) from each story in the training set. Specifically, the RAKE method evaluates the importance of each word in the story and give each word a score based on the word frequency and word degree. We extract the word $c_m$ with the highest score from each sentence $y_m$ and treat it as the gold abstract concept for $y_m$ to construct the storyline. Therefore, the final loss function is formulated as:

$$\mathcal{L}_{overall} = -\left(\lambda_c \sum_{m=1}^{M} \log p^c_m + \lambda_w \sum_{m=1}^{M} \sum_{t=1}^{T} \log p^w_{m,t}\right) \tag{22}$$

where $\lambda_c$ and $\lambda_w$ are hyperparameters that control the relative importance of the two loss functions.

# Experimental Setup

## Experimental Dataset

In order to evaluate the effectiveness of the proposed method, we conduct experiments on the widely used benchmark dataset VIST (Huang et al. 2016). The dataset consists of 10,117 Flicker albums with 210,819 unique images. Following previous works (Jung et al. 2020; Wang et al. 2020),

| Method | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE | CIDEr |
|---|---|---|---|---|---|---|---|
| Seq2Seq (Huang et al. 2016) | - | - | - | 3.5 | 31.4 | - | 6.8 |
| h-attn-rank (Yu, Bansal, and Berg 2017) | - | - | 21.0 | - | 34.1 | 29.5 | 7.5 |
| XE-ss (Wang et al. 2018b) | 62.3 | 38.2 | 22.5 | 13.7 | 34.8 | 29.7 | 8.7 |
| AREL (Wang et al. 2018b) | 63.8 | 39.1 | 23.2 | 14.1 | 35.0 | 29.5 | 9.4 |
| HRSL (Huang et al. 2019) | - | - | - | 12.3 | 35.2 | **30.8** | 10.7 |
| ReCo-RL (Hu et al. 2020) | - | - | - | 12.4 | 33.9 | 29.9 | 8.6 |
| INet (Jung et al. 2020) | 64.4 | 40.1 | 23.9 | 14.7 | 35.6 | 29.7 | 10.0 |
| SGVST (Wang et al. 2020) | 65.1 | 40.1 | 23.8 | 14.7 | **35.8** | 29.9 | 9.8 |
| IRW | **66.7** | **41.6** | **25.0** | **15.4** | 35.6 | 29.6 | **11.0** |
| w/o reasoning | 64.3 | 39.5 | 23.2 | 13.5 | 35.0 | 29.5 | 9.4 |
| w/o event | 65.2 | 40.5 | 24.2 | 14.7 | 35.2 | 29.7 | 10.3 |
| w/o scene | 65.9 | 41.1 | 24.4 | 15.0 | 35.6 | 29.7 | 10.5 |
| w/o commonsense | 65.5 | 40.8 | 24.1 | 14.8 | 35.2 | 29.6 | 10.2 |
| w/o guiding unit | 66.1 | 40.9 | 24.6 | 14.6 | 35.3 | 29.7 | 10.3 |
| w/o concept loss | 63.8 | 39.0 | 22.7 | 13.1 | 34.8 | 29.5 | 9.0 |

Table 1: Comparisons of the proposed model and the state-of-the-art baseline methods on VIST dataset.

the VIST dataset is split into training/validation/testing sets with 40,098/4,988/5,050 samples. Each album has five images and a corresponding story with five sentences.

### Baseline Methods

We evaluate and compare the proposed method with several strong visual storytelling methods, including (1) Seq2Seq (Huang et al. 2016); (2) h-attn-rank (Yu, Bansal, and Berg 2017), a hierarchically-attentive RNN; (3) XE-ss/AREL (Wang et al. 2018b), an adversarial reward learning framework; (4) HSRL (Huang et al. 2019), a hierarchically structured reinforcement learning model; (5) ReCo-RL (Hu et al. 2020), a reinforcement learning framework with relevance, coherence and expressiveness assessment criteria; (6) INet (Jung et al. 2020), an imagination network; (7) SGVST (Wang et al. 2020).

### Evaluation Metrics

We quantitatively evaluate our model with six automatic evaluation metrics that are widely used in previous work (Jung et al. 2020; Wang et al. 2020), including BLEU-N (N=1,2,3,4) (Papineni et al. 2002), ROUGE (Lin and Och 2004), METEOR (Denkowski and Lavie 2014), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015). These metrics measure the consistency between the n-gram overlap between generated stories and reference stories.

### Implementation Details

Following previous works (Wang et al. 2018b; Huang et al. 2019; Jung et al. 2020), we utilize the pretrained ResNet-152 (He et al. 2016) to learn image features. The dimension of each generated image feature vector is set as 2048. The max number of relationships in each scene graph, event graph and commonsense graph is set to 25, 25, 40, respectively. We set the size of the word embedding to 512. The hidden sizes of all GRUs are set to 512. The dimension of each skip-thought vector is set to 2400. The maximum length of each sentence is set to 30 via padding operation. We train the model with cross-entropy loss, which consists of concept loss and word loss, and use the Adam optimizer with an initial learning rate $5 \times 10^{-4}$ to optimize the model. The batch size is set to 80 for all experiments.

## Experimental Results

### Quantitative Evaluation

We firstly report the model comparison from the quantitative perspective. The experimental results on VIST dataset are summarized in Table 1. We observe that IRW model achieves noticeably better performance than the state-of-the-art methods on most of the automatic evaluation measures. Concretely, our IRW model makes the relative improvement over the best existing score by 4.8% on BLEU-4 and 3.7% on CIDEr. In addition, our model also substantially outperforms the AREL, HRSL and ReCo-RL methods, which all employ reinforcement learning paradigms to optimize the model. The performance of IRW could be further improved by deploying reinforcement learning.

### Ablation Study

To analyze the effect of each component of the IRW model, we also perform the ablation test of IRW in terms of discarding the reasoning module (denoted as w/o reasoning), the event knowledge (denoted as w/o event knowledge), the guiding unit (denoted as w/o guiding unit), and the storyline concept loss (denoted as w/o concept loss). As illustrated in Table 1, we can observe that both reasoning module and storyline concept loss contribute greatly to our model. Benefiting by relational reasoning from graphs to exploit the commonsense knowledge, scene knowledge and event knowledge, the IRW model has the capability of utilizing both explicit objects and implicit perception. Meanwhile, that the IRW (w/o concept loss) model obtains poor performance over all metrics is within our expectation since the model would reason from the three kinds of graphs and obtain irrelevant information which might be usefulness even harmful to the decoding process. On the contrast, the model could learn to automatically generate coherent and reasonable storylines by calculating the concept loss. In addition, we can also observe that all three kinds of graphs greatly influence the performance of the IRW model. This may be

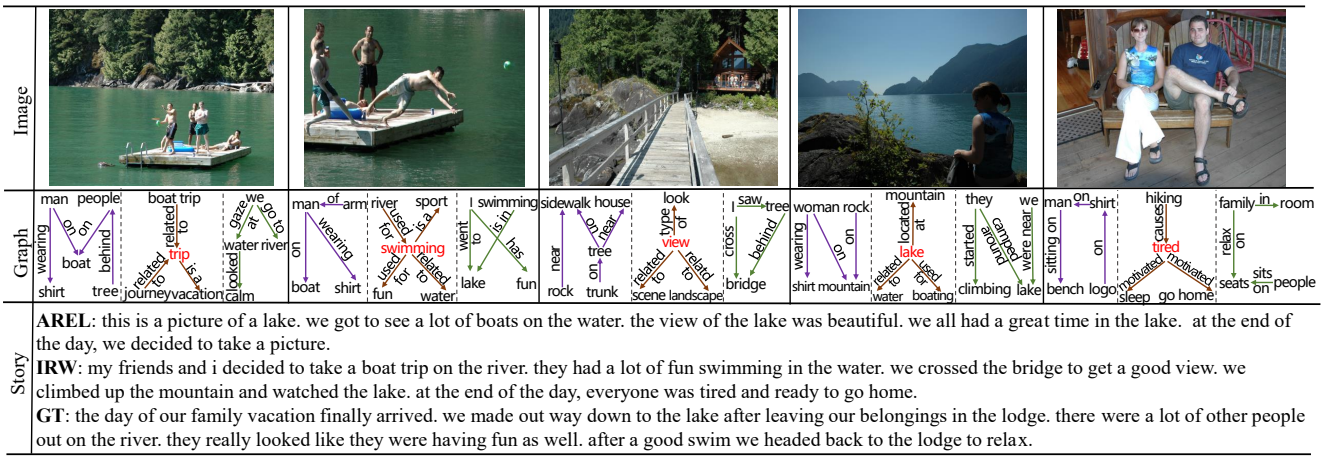


Figure 2: Example visual stories generated by AREL and IRW, and human-annotated ground-truth descriptions. The graphs in the left, middle and right for each image indicate the scene graph, commonsense graph and evnet graph, respectively. And the word with red indicate the predicted concept. Limited by the space, we just show part of the graphs.

| Model | Rele | Cohe | Info |
|---|---|---|---|
| Seq2Seq | 0.73 | 0.81 | 0.69 |
| AREL | 0.91 | 0.94 | 0.80 |
| HSRL | 0.89 | 0.97 | 0.85 |
| **IRW** (Ours) | **1.17** | **1.20** | **1.13** |
| w/o reasoning | 1.05 | 1.07 | 0.98 |
| w/o concept loss | 0.99 | 1.02 | 0.95 |

Table 2: Human evaluation results. Here, Rele, Cohe and Info are short for relevance, coherence and informativeness, respectively.

the reason that the scene graph and event graph capture the task-specific knowledge, providing essential knowledge of the salient objects and their relationships. The IRW model can also capture the external commonsense knowledge from the commonsense KB to generate informative and imaginary stories based on the predicted storyline.

## Human Evaluation

We also use human evaluation to verify the proposed model quantitatively. Specifically, we randomly select 200 image streams from the VIST test set and invited five well-educated volunteers to judge the quality of the generated stories generated by different models based on their relevance (whether the story is relevant with the image stream), coherence (whether the sentence in the story is coherent with the other sentences), and informativeness (whether the story contains diverse and rich content). The volunteers are asked to give each story a score of 0 (poor), 1 (satisfactory), 2 (good) for relevance, coherence and informativeness, respectively. Table 2 reports the results of human evaluation. Consistent with the results of automatic evaluation metrics, IRW can generate more relevant, coherent, and informative captions than other models.

## Qualitative Evaluation

To evaluate the proposed IRW model qualitatively, we show one visual story generated by IRW and AREL model in Figure 2. IRW can generate coherent and reasonable stories by utilizing the imagine-reason-write architecture. On the contrary, the AREL model, which just relies on visual features to generate descriptions, is prone to generate simple and noncoherent stories. Taking the first generated sentence in Figure 2 as an example, the phrase "a picture of a lake" generated by AREL is curiously short on detail, while the phrase "take a boat trip" generated by IRW is informative and imaginative. In addition, IRW can capture the change between images to generate reasonable and coherent stories. For example, the content of the last image is totally different from previous images in the image stream. The imagine module in our model first predicts the imaginary concept "tired" based on the image feature and previous generated sentence. Then, the reasoning module utilizes the concept "tired" to reason from the three kinds of graphs, and generates the sentence "everyone was tired and ready to go home", which is coherent with previous generated sentence and conforms to human commonsense. However, the phrase generated by AREL "take a picture" is just reasonable for the last image, but not coherent to the whole story.

## Conclusion

In this paper, we proposed a novel imagine-reason-write generation framework (IRW) for visual storytelling, inspired by the logic of humans when they write the story. We leveraged an imagining module to learn the imaginative storyline, which could improve the coherence and reasonability of the generated story. Then, we proposed a reasoning module to fully exploit the external commonsense knowledge and task-specific knowledge (scene graph and event graph) with relational reasoning method. In this way, the diversity and informativeness of the generated story could be enhanced substantially. Finally, we devised a guiding unit to integrate the visual and semantic knowledge to generate human-like stories. Extensive experiments on a benchmark dataset demonstrated that IRW achieved competitive results when compared to strong baselines.

## Acknowledgements

Min Yang was partially supported by the National Natural Science Foundation of China (NSFC) (No. 61906185), Youth Innovation Promotion Association of CAS China (No. 2020357), Shenzhen Basic Research Foundation (No. JCYJ20200109113441941). Xiang Ao was partially supported by the National Natural Science Foundation of China (No. 61976204, U1811461).

## References

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

Angeli, G.; Premkumar, M. J. J.; and Manning, C. D. 2015. Leveraging linguistic structure for open domain information extraction. In *ACL*, 344–354.

Chatterjee, M.; and Schwing, A. G. 2018. Diverse and coherent paragraph generation from images. In *ECCV*.

Denkowski, M.; and Lavie, A. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, 376–380.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*, 770–778.

Hu, J.; Cheng, Y.; Gan, Z.; Liu, J.; Gao, J.; and Neubig, G. 2020. What Makes A Good Story? Designing Composite Rewards for Visual Storytelling. In *AAAI*, 7969–7976.

Huang, Q.; Gan, Z.; Celikyilmaz, A.; Wu, D.; Wang, J.; and He, X. 2019. Hierarchically structured reinforcement learning for topically coherent visual story generation. In *AAAI*.

Huang, T.-H.; Ferraro, F.; Mostafazadeh, N.; Misra, I.; Agrawal, A.; Devlin, J.; Girshick, R.; He, X.; Kohli, P.; Batra, D.; et al. 2016. Visual storytelling. In *NAACL*.

Johnson, J.; Gupta, A.; and Fei-Fei, L. 2018. Image generation from scene graphs. In *CVPR*, 1219–1228.

Jung, Y.; Kim, D.; Woo, S.; Kim, K.; Kim, S.; and Kweon, I. S. 2020. Hide-and-Tell: Learning to Bridge Photo Streams for Visual Storytelling. *arXiv preprint arXiv:2002.00774* .

Kiros, R.; Zhu, Y.; Salakhutdinov, R. R.; Zemel, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Skip-thought vectors. In *NeurIPS*, 3294–3302.

Krause, J.; Johnson, J.; Krishna, R.; and Fei-Fei, L. 2017. A hierarchical approach for generating descriptive image paragraphs. In *CVPR*, 317–325.

Lin, C.-Y.; and Och, F. J. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *ACL*, 605–612.

Liu, Y.; Fu, J.; Mei, T.; and Chen, C. W. 2017. Let your photos talk: Generating narrative paragraph for photo stream via bidirectional attention recurrent neural networks. In *AAAI*.

Lu, J.; Xiong, C.; Parikh, D.; and Socher, R. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, 375–383.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*, 311–318.

Park, C. C.; and Kim, G. 2015. Expressing an image stream with a sequence of natural sentences. In *NeurIPS*, 73–81.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-critical sequence training for image captioning. In *CVPR*, 7008–7024.

Rose, S.; Engel, D.; Cramer, N.; and Cowley, W. 2010. Automatic keyword extraction from individual documents. *Text mining: applications and theory* 1: 1–20.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. S.; et al. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 211–252.

Speer, R.; and Havasi, C. 2012. Representing General Relational Knowledge in ConceptNet 5. In *LREC*, 3679–3686.

Tang, K.; Zhang, H.; Wu, B.; Luo, W.; and Liu, W. 2019. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6619–6628.

Teney, D.; Liu, L.; and van Den Hengel, A. 2017. Graph-structured representations for visual question answering. In *CVPR*, 1–9.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *CVPR*, 4566–4575.

Wang, J.; Fu, J.; Tang, J.; Li, Z.; and Mei, T. 2018a. Show, Reward and Tell: Automatic Generation of Narrative Paragraph From Photo Stream by Adversarial Training. In *AAAI*.

Wang, R.; Wei, Z.; Li, P.; Zhang, Q.; and Huang, X. 2020. Storytelling from an Image Stream Using Scene Graphs. In *AAAI*, 9185–9192.

Wang, X.; Chen, W.; Wang, Y.; and Wang, W. Y. 2018b. No Metrics Are Perfect: Adversarial Reward Learning for Visual Storytelling. In Gurevych, I.; and Miyao, Y., eds., *ACL*.

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

Yang, P.; Luo, F.; Chen, P.; Li, L.; Yin, Z.; He, X.; and Sun, X. 2019. Knowledgeable Storyteller: A Commonsense-Driven Generative Model for Visual Storytelling. In Kraus, S., ed., *IJCAI*, 5356–5362.

Yu, L.; Bansal, M.; and Berg, T. L. 2017. Hierarchically-attentive rnn for album summarization and storytelling. *arXiv preprint arXiv:1708.02977* .

Zellers, R.; Yatskar, M.; Thomson, S.; and Choi, Y. 2018. Neural motifs: Scene graph parsing with global context. In *CVPR*, 5831–5840.

Zha, Z.-J.; Liu, D.; Zhang, H.; Zhang, Y.; and Wu, F. 2019. Context-aware visual policy network for fine-grained image captioning. *TPAMI* .