Keep it Consistent: Topic-Aware Storytelling from an Image Stream via Iterative Multi-agent Communication

Ruize Wang\textsuperscript{1}, Zhongyu Wei\textsuperscript{2*}, Piji Li\textsuperscript{3}, Haijun Shan\textsuperscript{4}, Ji Zhang\textsuperscript{4}, Qi Zhang\textsuperscript{5}, Xuanjing Huang\textsuperscript{5}

\textsuperscript{1}Academy for Engineering and Technology, Fudan University, China
\textsuperscript{2}School of Data Science, Fudan University, China
\textsuperscript{3}Tencent AI Lab, China
\textsuperscript{4}Zhejiang Lab, China
\textsuperscript{5}School of Computer Science, Fudan University, China

\{rzwang18,zywei,qz,xhuang\}@fudan.edu.cn; lipiji.pz@gmail.com; shanhaijun@zhejianglab.com; Ji.Zhang@usq.edu.au

Abstract

Visual storytelling aims to generate a narrative paragraph from a sequence of images automatically. Existing approaches construct text description independently for each image and roughly concatenate them as a story, which leads to the problem of generating semantically incoherent content. In this paper, we proposed a new way for visual storytelling by introducing a topic description task to detect the global semantic context of an image stream. A story is then constructed with the guidance of the topic description. In order to combine the two generation tasks, we propose a multi-agent communication framework that regards the topic description generator and the story generator as two agents and learn them simultaneously via iterative updating mechanism. We validate our approach on VIST, where quantitative results, ablations, and human evaluation demonstrate our method’s good ability in generating stories with higher quality compared to state-of-the-art methods.

1 Introduction

Image-to-text generation is an important topic in artificial intelligence (AI) which connects computer vision (CV) and natural language processing (NLP). Popular tasks include image captioning (Karpathy and Fei-Fei 2015) \cite{karpathy2015deep}, \cite{ren2017faster} \cite{vinyals2017show} \cite{antol2015vqa} \cite{yu2017image} \cite{fan2018image} \cite{fan2018ask}, aiming at generating a short sentence or a phrase conditioned on certain visual information. With the development of deep learning and reinforcement learning models, recent years witness promising improvement of these tasks for single-image-to-single-sentence generation.

Visual storytelling moves one step further extending the input and output dimension to a sequence of images and a sequence of sentences. It requires the model to understand the main idea of the image stream and generate coherent sentences. Most of existing methods \cite{huang2016image} \cite{liu2017image} \cite{yu2017image} \cite{bansal2017coherent} \cite{wang2018multi}, for visual storytelling extend approaches of image captioning without considering topic information of the image sequence, which causes the problem of generating semantically incoherent content.

\*Corresponding author

Figure 1: An example of visual storytelling from VIST dataset consisting of five images with the topic of car accident. Two stories presented are from an automatic approach and a human annotator respectively. Wavy lines highlight the inappropriate sentiment of the machine-generated sentence. Underlines indicate that the sentence provides little on-topic information about the image sequence.

An example of visual storytelling can be seen in Figure 1. An image stream with five images about a car accident is presented accompanied with two stories. One is constructed by a human annotator and the other is produced by an automatic storytelling approach. There are two problems with the machine generated story. First, the sentiment expressed in the text is inappropriate. In face of a terrible car accident, the model uses some words with positive emotion, i.e., “happy” and “excited”. Second, some sentence is uninformative. The sentence “we got to see a lot of different things at the event” provides little information about the car accident. This example shows that topic information about the image sequence is important for the story generator to produce an informative and semantically coherent story.

In this paper, we introduce a novel task of topic description generation to detect the global semantic context of an image sequence and generate a story with the guidance of such topic information. In practice, we propose a framework named Topic-Aware Visual Story Telling (TAVST)
to tackle the two generation tasks: (1) topic description generation for the given image stream; (2) story generation with the guidance of the topic information. To effectively combine these two tasks, we propose a multi-agent communication framework that regards the topic description generator and the story generator as two agents. In order to enable the interaction of these two agents for information sharing, an iterative updating (IU) module is incorporated into the framework. Extensive experiments on the VIST dataset (Huang et al. 2016) show that our framework achieves better performance compared to state-of-the-art methods. Human evaluation also demonstrates that the stories generated by our model are better in terms of relevance, expressiveness and topic consistency.

## 2 Approach

Given an image stream \( x = (x_1, \ldots, x_N) \), where \( N \) is the number of images, we aim to output a topic description \( y_{\text{topic}} \) and \( N \) sub-stories to form a complete story \( y = (y_1, \ldots, y_N) \). The proposed framework mainly includes three stages, namely visual encoding, initial stage of generation and iterative updating (IU). The visual encoder is employed to extract image features as visual context vectors. In the initial stage, we have the initial version of the two generation agents. The initial topic description generator takes visual context vectors as input and generates a topic vector. The initial story generator combines the topic vector and visual context vectors via co-attention mechanism and construct the initial version of story. Considering that the generated story can also benefit the topic description generator, the two agents communicate with each other in the IU module via message passing mechanism as fine tuning. The overall architecture of our proposed model is shown in Figure 2. Each of these modules will be described in details in the following sections.

### Visual Encoder

Given an image stream \( x \) with \( N \) images, we first extract the high-level visual features \( f_i \) of each image \( x_i, (i \in 1, \ldots, N) \) through a pre-trained CNN based model - ResNet (He et al. 2016). Then for the whole image stream, following the previous works (Yu, Bansal, and Berg 2017), a bidirectional gated recurrent unit (bi-GRU) is employed as the visual encoder:

\[
\begin{align*}
    f_i &= \text{ResNet}(x_i) \\
    \overrightarrow{h_{i,t}} &= \text{GRU}(f_i, \overrightarrow{h_{i,t-1}}) \\
    \overleftarrow{h_{i,t}} &= \text{GRU}(f_i, \overleftarrow{h_{i,t+1}})
\end{align*}
\]

where \( \overrightarrow{h_{i,t}} \) is the forward hidden state at time step \( t \) of \( i \)-th visual feature \( f_i \), while the \( \overleftarrow{h_{i,t}} \) is the backward one. At each time step, the visual features are sequentially fed into the visual encoder to obtain visual context vector, which has integrated the visual information from all the images that have been observed.

At last, fused with the visual representation followed by ReLU layer, the final visual context vector \( h_v^i \) can be calculated as:

\[
    h_v^i = \text{ReLU}([\overrightarrow{h_i}; \overleftarrow{h_i}) + W_f f_i)
\]  

(1)

where \( W_f \) is a projection matrix.

### Initial Topic Description Generator

Given the visual context vector extracted from the image sequence, we first learn to generate the topic description. In practice, all visual context vectors \( h_v^i \) are concatenated and then fed into the initial topic description generator that employs a gated recurrent unit (GRU) decoder.

The output \( p_{\text{topic}}^{\text{init}} \) of this decoder is a sequence of probability distribution over the whole topic vocabulary \( V_t \). The training loss of initial topic description generator is the cross-entropy \( L_{\text{init}}^{\text{topic(mle)}} \) between the generated description \( p_{\text{topic}}^{\text{init}} \) and the ground-truth topic description \( p_{\text{topic}} \).

Note that at each time step, the decoder produces the a hidden state \( h_t^i \). Once the last topic hidden state \( h_t^M \) is obtained, we concatenate all topic hidden states \( h^i = [h_1^i, \ldots, h_M^i] \), \( M > 1 \) as the topic memory, which are fed into the story generation module.

### Initial Story Generator with Co-attention Network

The initial story generator is responsible for generating the story with the guidance of the topic description constructed by the initial topic description generator.

![Figure 2: Overall architecture of our TAVST model.](image-url)
In order to combine both visual information and topic information for story generation, we adopt a co-attention mechanism (Lu et al. 2016; Jing, Xie, and Xing 2018) for context information encoding. The structure of the co-attention encoding module can be seen in Figure 3. Specifically, given visual context vectors $h^v$ and topic vectors $h^t$, the affinity matrix $C$ is calculated by

$$C = \tanh(h^T W_v h^v)$$  \hspace{1cm} (2)

where $W_v$ is the weight parameter. After calculating this matrix, we compute attentions weights over the visual context vectors and the topic vectors via the following operations:

$$H^v = \tanh(W_v h^v + (W_t h^t) C)$$

$$H^t = \tanh(W_t h^t + (W_v h^v) C^T)$$

$$a^v = \text{softmax}(w^T h^v H^v)$$

$$a^t = \text{softmax}(w^T h^t H^t)$$  \hspace{1cm} (3)

where $W_v$, $W_t$, $w^v$, $w^t$ are the weight parameters. Then based on the above attention weights, the visual and semantic attentions are calculated as the weighted sum of the visual context vectors and the topic vectors:

$$a^v_{\text{att}} = \sum_{n=1}^{N} a^v_n h^v_n, a^t_{\text{att}} = \sum_{m=1}^{M} a^t_m h^t_m$$  \hspace{1cm} (4)

At last, we concatenate the visual and semantic attentions as $[a^v_{\text{att}}; a^t_{\text{att}}]$, and then use a fully connected layer $W_{fc}$ to obtain the joint context vector:

$$j = W_{fc}[a^v_{\text{att}}; a^t_{\text{att}}]$$  \hspace{1cm} (5)

In the story decoding stage, each joint context vector $j_i$ is fed into a GRU decoder to generate a sub-story sentence $y_i$ for the corresponding image. Formally, the generation process can be written as:

$$s^i_t = \text{GRU}(s^i_{t-1}, [w^i_{t-1}, j_i])$$  \hspace{1cm} (6)

$$p(w^i_t|w^i_{t-1}) = \text{softmax}(\text{MLP}(s^i_t))$$  \hspace{1cm} (7)

where $s^i_t$ denotes the $t$-th hidden state of $i$-th GRU. We concatenate the previous word token $w^i_{t-1}$ and the context vector $j_i$ as the input at each step. The output $p$ is a probability distribution over the whole story vocabulary $\mathbb{V}_s$.

### Loss Function for Training

We define two different loss functions including cross-entropy (MLE) and reinforce (RL). MLE loss is show in Equation (8):

$$L_{\text{story(mle)}}(\theta_1) = -\sum_{t=1}^{T} \log(p_{y_t}(y^*_t; ..., y^*_{t-1}))$$  \hspace{1cm} (8)

where $\theta_1$ is the parameter of story generator; $y^*$ is the ground-truth story and $y^*_t$ denotes the $t$-th word in $y^*$.

Recently, reinforcement learning has shown effectiveness for training text generation model via introducing automatic metrics (e.g. METEOR, etc.) to guide the training process (Wang et al. 2018b). We also explore the RL-based approach to train our generator. The reinforcement learning (RL) loss can be written as:

$$L_{\text{story(rl)}}(\theta_1) = -E_{y \sim p_{\theta_1}} \left( r(y; y^*) - b \right)^2$$  \hspace{1cm} (9)

where $r$ is a sentence-level metric for the sampled sentence $y$ and the ground-truth $y^*$; $b$ is the baseline which can be an arbitrary function but a linear layer in our experiments for simply. To stabilize the RL training process, a simple way is to linearly combine MLE and RL objectives as follows (Wu et al. 2018):

$$L_{\text{story(com)}} = \alpha L_{\text{story(rl)}} + (1 - \alpha) L_{\text{story(mle)}}$$  \hspace{1cm} (10)

where hyper-parameter $\alpha$ is employed to control the trade-off between MLE and RL objectives.

In the initial stage, a combined loss function of $L_{\text{story(com)}}$ and $L_{\text{topic}}$ is computed through:

$$L_{\text{init}} = \lambda_1 L_{\text{story(com)}} + (1 - \lambda_1) L_{\text{topic(mle)}}$$  \hspace{1cm} (11)

where hyper-parameter $\lambda_1$ is employed to balance these losses.

### Iterative Updating Module

Considering that the generated story would also be helpful for the generation of topic description, we design an iterative updating module for the two agents to interact with each other and update iteratively. In IU module, we generate the topic description from the previously generated story, and then use such topic information to further guide story generation. To distinguish the two agents from those of initial version, we call them the IU version.

#### IU Topic Description Generator

We envisage that the generated story is able to provide more accurate information for topic description generation than visual information.
Therefore, instead of using visual information as input, the IU version of topic description generator takes the generated story as input. Specifically, the last hidden states $s_{iter}$ of the IU story generator is used as input. Note that the IU version topic description generator is initialized as its initial version and keeps training with the same objective.

**IU Story Generator**

The IU story generator shares the structure and parameters with the initial story generator. It takes both topic vector and visual context vector as input. In the decoding process, the story $y$ is the concatenation of the sub-stories $y_i$ generated by IU story generator, and the last hidden states $s_{iter} = [s_{last}^{1}, \ldots, s_{last}^{N}]$ of the IU story generator will be passed to IU topic description generator for iterative updating.

**Loss Function for Training**

The training losses of the IU story generator are similar to Eq. (9)-(10), including the combination loss $L_{iter}^{story(com)}$ and $L_{iter}^{story(mle)}$.

At each iteration stage $n$, the IU module loss $L_{iter}^{n}$ is the weighted sum of IU topic description generation loss $L_{iter}^{topic(mle)}$ and IU story generation loss $L_{iter}^{story(mle)}$:

$$L_{iter}^{n} = \lambda L_{iter}^{topic(mle)} + (1 - \lambda)L_{iter}^{story(mle)}$$

where hyper-parameter $\lambda$ is employed to balance these losses.

**Multi-Agent Training**

The IU topic description generator and IU story generator communicate with each other iteratively in the IU module until it reaches the given iteration number $N_{iter}$. The loss for IU module is:

$$L_{iter} = \sum_{n=1}^{N_{iter}} L_{iter}^{n}$$

Therefore, to train the whole multi-agent learning framework, we introduce a combined loss $L$ which consists of the initial loss $L^{init}$ and IU module loss $L^{iter}$:

$$L = \beta L^{init} + (1 - \beta)L^{iter}$$

where $\beta$ is a hyper-parameter to balance these losses. During training, our goal is minimizing $L$ using stochastic gradient descent.

### 3 Experiments

**Datasets**

The VIST dataset (Huang et al. 2016) is the benchmark for the evaluation of visual storytelling. It includes 10,117 Flicker albums with 210,819 images. We evaluate our method on VIST dataset and use the same split settings as previous works (Huang et al. 2016; Yu, Bansal, and Berg 2017; Wang et al. 2018b). The samples have been split into three parts, 40,998 for training, 4,988 for validation and 5,050 for testing, respectively. Each sample (album) contains five images and a story consisting of five sentences. We use the title of each album as the ground-truth topic description.

**Implementation Details**

We use the pre-trained ResNet-152 (He et al. 2016) model to extract image features. The vocabulary for story and topic include words appearing no less than three times in corresponding parts (i.e., story and title) from the training set. And all the rest of words are represented as UNK. We adopt GRU models for both visual encoder and other decoders, and the hidden size of them is 512. Expect the encoder is bidirectional, the other decoders are unidirectional.

The batch size is set as 64 during the training. We use Adam (Kingma and Ba 2015) to optimize our models with the initial learning rate of 0.0002. We first pre-train the initial topic description generator using MLE. Then we pre-train both the topic description generator and the story generator jointly using MLE. The number of iteration $N_{iter}$ is set to 2, the weight of RL is $\alpha = 0$, and hyper-parameters in loss optimization are set as $\lambda_1 = 0.7$, $\lambda_2 = 0.7$ and $\beta = 0.3$, which are selected based on validation set (the details about hyper-parameter sensitivity analysis are shown in Supplementary Material). After warm-up pre-training, $\alpha$ and learning-rate are set to 0.8 and 0.00002 to fine-tune using RL. Here we use METEOR scores as the reward. We select the best model which achieves the highest METEOR score on the validation set. The reason is that METEOR is proved to correlate better with human judgment than CIDEr-D in the small references case and superior to BLEU@N all the time (Vedantam, Lawrence Zitnick, and Parikh 2015; Wang et al. 2018a). During the test stage, we generate the stories by performing a beam-search with a beam size of 3.

**Models for Comparison**

We compare our proposed methods with several baselines for visual storytelling, which are detailed as follows.

**seq2seq** (Huang et al. 2016): It generates caption for each single model via classic sequence-to-sequence model and concatenate all captions to form the final story.

**h-att-rank** (Yu, Bansal, and Berg 2017): On top of the classic sequence-to-sequence model, it adds an additional RNN to select photos for story generation.

**HPSR** (Wang et al. 2019a): It introduces an additional RNN stacked on the RNN-based photo encoder to detect the scene change. Information from both RNNs are fed into an RNN for story generation.

**AREL** (Wang et al. 2018b): It is based on the framework of reinforcement learning and the generation of a single word is treated as the policy. The reward model learns the reward function from human demonstrations.

**HSRL** (Huang et al. 2019): It is based on the framework of hierarchical reinforcement learning. The higher level agent is responsible for generating a local concept for each image as the guidance to the lower level agent for sentence generation.

**VST**: This is the baseline version of our model without using topic information as guidance.

**TAVST w/o IU**: This is our proposed TAVST method without IU module, which only equipped with initial topic description generator.

**TAVST**: This is our full model. TAVST (MLE) is trained using MLE while TAVST (RL) is trained via RL loss.
Table 1: Overall performance of story generation on VIST dataset for different models in terms of BLEU (B), METEOR (M), ROUGE-L (R-L), and CIDEr-D (C). **Bolded** numbers are the best performance in each column.

| Methods                  | B-1 | B-2 | B-3 | B-4 | R-L | C   | M   |
|--------------------------|-----|-----|-----|-----|-----|-----|-----|
| seq2seq (Huang et al. 2016) | −   | −   | −   | 3.5 | 6.8 | 31.4|
| h-attn-rank (Yu, Bansal, and Berg 2017) | −   | −   | 21.0 | −   | 29.5 | 7.5 | 34.1|
| HPSR (Wang et al. 2019a)  | 61.9| 37.8| 21.5| 12.2| **31.2** | 8.0 | 34.4|
| VST (MLE)                | 62.3| 38.0| 21.8| 12.7| 29.7| 7.8 | 34.3|
| TAVST w/o IU (MLE)       | 63.1| 38.6| 22.9| 14.0| 29.7| 8.5 | 35.1|
| TAVST (MLE)              | **63.6**| **39.3**| **23.4**| **14.2**| **30.3**| **8.7** | **35.4**|
| RL                       |      |     |     |     |     |     |     |
| AREL (Wang et al. 2018b) | 63.7| 39.0| 23.1| 14.0| 29.6| 9.5 | 35.0|
| HSRL (Huang et al. 2019) | −   | −   | −   | 12.3| 30.8| **10.7** | 35.2|
| TAVST w/o IU (RL)        | 63.5| 39.2| 23.2| 14.3| 30.0| 8.7 | 35.3|
| TAVST (RL)               | **64.2**| **39.6**| **23.7**| **14.6**| **31.0**| **9.2** | **35.7**|

Table 2: Performance of topic description generation in terms of automatic metrics.

| Methods                  | B-1 | B-2 | R   | C   | M   |
|--------------------------|-----|-----|-----|-----|-----|
| TAVST w/o IU             | 5.1 | 2.2 | 12.6| 10.9| 4.8 |
| TAVST                    | **6.0**| **2.7**| **13.4**| **12.1**| **5.6**|

Table 3: Human evaluation results on Turing test.

| Method | Win (%) | Lose (%) | Unsure (%) |
|--------|---------|----------|------------|
| VST    | 22.4    | 71.7     | 5.9        |
| AREL   | 38.4    | 54.2     | 7.4        |
| TAVST  | **42.7**| **52.2** | **7.1**    |

**Automatic Evaluation Results**

We evaluate our model on two generation tasks i.e., story generation and topic description generation, in terms of four automatic metrics: BLEU (Papineni et al. 2002), ROUGE-L (Lin and Och 2004), METEOR (Banerjee and Lavie 2005), and CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015).

**Story Generation** The overall experimental results are shown in Table 1. **TAVST (MLE)** outperforms all of the baseline models trained with MLE. This confirms the effectiveness of topic information for generating better stories. Noticeably, compared with the RL-based models, our **TAVST (MLE)** has already achieved a competitive performance and outperforms other RL models (i.e., **AREL** and **HSRL**). In terms of METEOR and BLEU@[2-4] metrics. After equipped with RL, our **TAVST (RL)** model is able to further improve the performance, outperforming the two RL models in terms of all metrics except CIDEr-D. Our full model **TAVST** (both MLE and RL versions) outperforms **TAVST w/o IU**, which directly demonstrates the effectiveness of the IU module. **TAVST w/o IU** achieves better performance than VST, which proves that topic description generator can provide guidance for story generation.

**Topic Description Generation** The overall results for topic description generation can be seen in Table 2. **TAVST** achieves higher performance compared to **TAVST w/o IU**, indicating that the generated story is able to provide assistance for better topic description generation. In general, the description generator obtains low scores in terms of automatic metrics. observations on the dataset reveal that the length of titles for each album is relatively short, ranging from 2 to 6 words mostly. Given such a short reference, it is difficult for models to obtain high scores in terms of automatic metrics. We further look into the generated descriptions and some of them are actually semantically correct. For example, the reference is “happy birthday party at my home” and the generated topic description is “the birthday gathering”. Another example is that the reference is “family feast” and the generated topic is “dinner party”. We believe such kind of topic description with similar meaning can still provide positive guidance for the story generator.

**Human Evaluation**

We perform two kinds of human evaluation through Amazon Mechanical Turk (AMT), namely the Turing test and the pairwise comparison. Since we only find one previous work (Wang et al. 2018b) which published the sampled results of their model, we chose it for comparison. In specific, we re-collect human labels for their sample results and stories generated by our models on the same sub-set of albums. A total of 150 stories (750 images) are used, and each of them is evaluated by 3 human evaluators.

**Turing Test** For the Turing test, we design a survey (as shown in Appendices) that contains an image stream, a generated story by our **TAVST** model and a story written by a human. Evaluators are required to choose the story that is more likely written by a human. The experimental result (Table 3) shows that 47.8% of evaluators think the stories generated by our model are written by a human (v.s 38.4% win rate from **AREL**).
Table 4: Human evaluation results on pairwise comparison.

| Choice (%) | TAVST vs VST | TAVST vs AREL | TAVST vs GT |
|------------|--------------|---------------|-------------|
| Relevance  | 56.00        | 51.56         | 41.44       |
| Expression | 60.00        | 46.22         | 42.12       |
| Topic Consistency | 66.44 | 58.67 | 43.24 |

Table 5: Sentiment scores corresponding to different types of events. Note that higher score indicates more positive polarity.

| Method | Positive | Negative |
|--------|----------|----------|
|        | new year’s eve | sporting event | breaking up | car accident |
| GT     | 1.20     | 1.25     | 0.48       | 0.38       |
| TAVST  | 1.31     | 1.17     | 0.53       | 0.42       |
| AREL   | 1.22     | 1.15     | 0.95       | 0.81       |
| VST    | 1.65     | 1.82     | 1.66       | 1.84       |

Pairwise Comparison A good story for an image stream should have three significant factors: (1) Relevance: the story should be relevant to the image stream. (2) Expressiveness: the story should be concrete and coherent, and have a human-like language style. (3) Topic Consistency: the story should be consistent to the topic. We compare our model with three other models in terms of these three metrics: TAVST vs VST, TAVST vs AREL and TAVST vs ground-truth. In this annotation task, AMT evaluators need to compare two given stories according to these three factors and choose which story is better in terms of a certain factor. Results are shown in Table 4. Our model performs better than the other two models in terms of relevance and topic consistency. The advantage of topic consistency is more promising. This proves that the topic description generator can help the story generation agent construct a more consistent story.

Further Analysis on Topic Consistency

We further evaluate the quality of the generated story in terms of topic consistency from the perspective of sentiment. Specifically, we employ a lexicon-based approach using a subjectivity lexicon (Wilson, Wiebe, and Hoffmann 2005). We count the number of sentiment words in each sentence for the polarity evaluation. The score will be 1, 0, -1 if a sentence is positive, neutral and negative, respectively. Based on the score for each sentence, two qualitative experiments are designed to measure the in-story sentiment consistency and topic-story sentiment consistency.

In-story Sentiment Consistency We argue that the sentiment of sentences in a story should be consistent given the album is related to a certain topic. For each story, we obtain a vector with 5 sentiment scores in correspondence to 5 sentences. We then calculate the standard deviation for the vector to represent the divergence score of a story. For each model, we average the divergence scores of all stories generated as its final score. Figure 4 presents the results from different models. Results illustrate that our method can generate stories with higher in-story sentiment consistency.

Topic-Story Sentiment Consistency Considering that albums related to some events might express a tendency to a certain polarity. For example, the sentiment of stories about new year’s eve are more likely to be positive while the sentiment of stories about breaking up are more likely to be negative. We enumerate albums with different event types to see if the model has the ability to generate stories with sentiment consistent with the type of events. For each story, we add all the sentiment scores of sentences as its final score. The higher score a story obtain, the more positive the story is. Four types of events are considered. Results is shown in Table 5. In general, all automatic models tend to generate stories with higher sentiment scores compared to human-written stories. This is because a large portion of albums in the dataset are related to positive events. Both VST and AREL generate stories with similar sentiment scores for both types of events. This indicates that they are not able to distinguish positive and negative events. With the guidance of topic description, our model TAVST is able to distinguish events with different sentiment tendency.

Case Study

Figure 5 shows an example of the ground-truth story and stories generated automatically by different models. The words in red, blue and yellow color represent the topic, subject, and emotion, respectively. Our model shows promising results according to topic consistency, which further confirms that our model can extract appropriate topic which serves as the guidance of generating a topic-consistent story.
4 Related work

Our research is related to image captioning, visual storytelling and multi-task learning. In early works (Yang et al. 2011; Kulkarni et al. 2011; Elliott and Keller 2013), image captioning task is treated as a ranking problem, which is based on retrieval models to identify similar captions from the database. Later, the end-to-end framework based on the CNN and RNN is proposed by researchers (Xu et al. 2015; Karpathy and Fei-Fei 2015; Vinyals et al. 2017; Dai et al. 2017; Fan et al. 2019). Such work focuses on the literal description of image content. Although some encouraging results have been generated in connecting vision and language, the generated text is limited in a single sentence.

Visual storytelling is the task of generating a narrative paragraph given an image stream. The pioneering work was done by (Park and Kim 2015) that retrieves a sequence of sentences for an image stream. Huang et al. (2016) introduces the first dataset (VIST) for visual storytelling and establishes some baseline approaches. An attention-based RNN with a skip gated recurrent unit (Liu et al. 2017) is designed to maintain longer range information. Yu, Bansal, and Berg (2017) designs a hierarchically-attentive RNN structure. Recently, a reinforcement learning framework with two discriminators is proposed (Wang et al. 2018a) for this task. Due to the bias can be brought by the hand-coded evaluation metrics, Wang et al. (2018b) proposes an adversarial reward learning framework to uncover a robust reward function from human demonstrations.

The most similar work to ours is from Huang et al. (2019). They propose to generate a local semantic concept for each image in the sequence and generate a sentence for each image using a semantic compositional network in a fashion of hierarchical reinforcement learning. Although both of us consider topic information to facilitate the story generation. Our model is different from three aspects. First, the concepts of topic are different. We treat topic as the global semantic context of the album while topic represents local semantic information in their case adhering to each single image. Second, our modeling topic is more interpretable. We generate topic description directly instead of producing latent representation and this provides more insights for further improving the performance. Third, the communication framework is compatible with any RL based training methods. Experiment results also show that with RL, our framework can outperform their model.

Our research is related to visual storytelling, multi-task learning and topic modeling. We treat topic as the global semantic context of the album while topic represents local semantic information in their case adhering to each single image. Second, our modeling topic is more interpretable. We generate topic description directly instead of producing latent representation and this provides more insights for further improving the performance. Third, the communication framework is compatible with any RL based training methods. Experiment results also show that with RL, our framework can outperform their model.

5 Conclusions and Future Work

In this paper, we introduce a topic-aware visual storytelling task, which identifies the global semantic context of a given image sequence and then generate the story with the help of such topic information. We propose a multi-agent communication framework that combines two generation tasks namely topic description generation and story generation effectively. In future, we will explore to model topic generation as a keyword extraction task. Besides, we will study to identify the interactions among images for better visual information encoding.

| Ground-truth | VST | AREL | TAVST |
|--------------|-----|------|-------|
| i went to the award ceremony yesterday. there were a lot of people there. everyone received an award for their effort. they had a great time. i really enjoyed being there. some of the soldiers started singing. | we were a lot of people there. he was very happy to be there. he gave a speech to the audience. he was very happy to be there. we had a great time at the end of the day. | we had a lot of fun at the meeting. he was very proud of his accomplishments. the president of the class gave a speech to the audience. the president of the school was very proud of his accomplishments. at the end of the ceremony, the family was happy to be together. | yesterday was the day of the military graduation ceremony. the speaker gave a speech to the audience. they were very happy to be there. the soldier was very proud of his award. we had a great time at the end of the day. |

Figure 5: Story samples generated by different models and the ground truth. The words in red, blue and yellow color represent the topic, subject, and emotion, respectively. (Best viewed in color)
[Xu et al. 2015] 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 et al. 2011] Yang, Y.; Teo, C. L.; Daumé, H.; and Aloi-minos, Y. 2011. Corpus-guided sentence generation of natural images. In EMNLP, 444–454.

[Yu, Bansal, and Berg 2017] Yu, L.; Bansal, M.; and Berg, T. 2017. Hierarchically-attentive rnn for album summarization and storytelling. In EMNLP, 966–971.

[Yu et al. 2017] Yu, Z.; Yu, J.; Fan, J.; and Tao, D. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In ICCV.
Supplementary Material

A Hyper-Parameter Sensitivity Analysis

Impact of $\lambda_1$ and $\lambda_2$ In our experiments, we try many different values of $\lambda_1$ and $\lambda_2$ on the validation set. We found that $\lambda_1 = 0.7$ and $\lambda_2 = 0.7$ best balances the topic extractor and the story generator. As we know, the overall performance of our model depends more on the generated story rather than the generated topic. The topic guides story generation, which plays an auxiliary role. So the weight for the topic extractor should be lower than the weight of story generator.

Impact of $N_{iter}$ and $\beta$ The choice of the hyper-parameter $N_{iter}$ affects the performance of the model. In our experiments, we find that when $N_{iter} = 1$ or $2$, the performance of the model is better than $N_{iter} = 0$, and $N_{iter} = 2$ performs best on the validation set. But when $N_{iter} = 3$, the performance declines. In addition, we observe that $\beta = 0.3$ plays a very good regulating effect.

Impact of $\alpha$ The hyper-parameter $\alpha$ controls the trade-off between MLE and RL objectives. For comparison, we set $\alpha$ to be [0, 0.5, 0.8, 1] in our experiments. The results show that, when $\alpha = [0.5, 0.8, 1]$, the model achieves a better performance than $\alpha = 0$; and $\alpha = 0.8$ best balances the RL loss and the MLE loss.
Survey Instructions (Click to expand)

In this task, you will see 5 photos and 2 stories.

Read the following stories, and choose the one that you think is more likely wrote by human.

Pleas look at the 5 photos and 2 stories below.

A. the baby is getting ready for the big day. they had a great time playing with the kids. the kids were very happy to see them. the kids had a great time. i got to eat a delicious meal.

B. he enjoyed playing the game to pass time. he preferred puzzle pieces. dad teaches her how to work the computer. the baby is interested in the game in his hand. dessert was almost gone.

Question: Which story is more likely wrote by human?

- [ ] A
- [ ] B
- [ ] Not sure

Figure 6: Turing Test Survey
In this task, you will see 5 photos and 2 stories.

Read the following stories, and compare them in these three factors: (1) Relevance (2) Expression (3) Topic Consistency

(1) Relevance: whether the story is relevant to the photo stream.
(2) Expression: whether the story is concrete and coherent, and has a human-like language style.
(3) Topic Consistency: whether the story is consistent in topic.

Please look at the 5 photos and 2 stories below.

A. we went on a hike yesterday. there were a lot of strange plants there. i had a great time. we drank a lot of water while we were hiking. the view was spectacular.

B. we went to a camping trip. there were a lot of different kinds of flowers. we had a great time. here is a picture of a man. the sky was beautiful.

Q1: Which story is more relevant to the photo stream?

- A
- B
- Not sure

Q2: Which story is more concrete and coherent, and have a human-like language style?

- A
- B
- Not sure

Q3: Which story is more consistent in topic?

- A
- B
- Not sure

Figure 7: Pairwise Comparison Survey