# Emotion Reinforced Visual Storytelling

Nanxing Li*†
School of Software, Tsinghua University & Beijing National Research Center for Information Science and Technology (BNRist)
Beijing, China
linanxingthu@gmail.com

Bei Liu*
Microsoft Research Asia
Beijing, China
bei.liu@microsoft.com

Zhizhong Han
Department of Computer Science, University of Maryland
College Park, Maryland, USA
h312h@umd.edu

Yu-Shen Liu‡
School of Software, Tsinghua University & Beijing National Research Center for Information Science and Technology (BNRist)
Beijing, China
liuyushen@tsinghua.edu.cn

Jianlong Fu
Microsoft Research Asia
Beijing, China
jianf@microsoft.com

## ABSTRACT

Automatic story generation from a sequence of images, i.e., visual storytelling, has attracted extensive attention. The challenges mainly drive from modeling rich visually-inspired human emotions, which results in generating diverse yet realistic stories even from the same sequence of images. Existing works usually adopt sequence-based generative adversarial networks (GAN) by encoding deterministic image content (e.g., concept, attribute), while neglecting probabilistic inference from an image over emotion space. In this paper, we take one step further to create human-level stories by modeling image content with emotions, and generating textual paragraph via emotion reinforced adversarial learning. Firstly, we introduce the concept of emotion engaged in visual storytelling. The emotion feature is a representation of the emotional content of the generated story, which enables our model to capture human emotion. Secondly, stories are generated by recurrent neural network, and further optimized by emotion reinforced adversarial learning with three critics, in which visual relevance, language style, and emotion consistency can be ensured. Our model is able to generate stories based on not only emotions generated by our novel emotion generator, but also customized emotions. The introduction of emotion brings more variety and realistic to visual storytelling. We evaluate the proposed model on the largest visual storytelling dataset (VIST). The superior performance to state-of-the-art methods are shown with extensive experiments.

## CCS CONCEPTS

• **Computing methodologies** → *Natural language generation*; *Image representations*; *Neural networks*.

## KEYWORDS

Storytelling; Multi-Modal; Emotion; Reinforcement Learning

**ACM Reference Format:**
Nanxing Li, Bei Liu, Zhizhong Han, Yu-Shen Liu, and Jianlong Fu. 2019. Emotion Reinforced Visual Storytelling. In *International Conference on Multimedia Retrieval (ICMR '19), June 10–13, 2019, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3323873.3325050

## 1 INTRODUCTION

Recent years, we have seen a bursting number of researches in bridging the gap between vision and language. Driven by the availability of large scale of pairs consisting of images and natural language descriptions and successful adoption of recurrent neural network (RNN), encouraging progress has been made in language generation from images [16, 25, 35]. In this paper, we tackle the problem of generating a story that consists of several sentences from a sequence of images, i.e., visual storytelling. Compared with image captioning and image paragraphing that take one single image as input, visual storytelling is a more subjective task which requires an overall understanding on the connection among all images and aims to generate sentences with consistent semantics.

Visual storytelling has been explored by many researches in recent years [13, 19, 23, 26, 28, 30]. Most of them focus on modeling the embedding between images and sentences and they consider the sequence of image contents as the most important factor for story generation. In other words, they only leverage low level features that come from image content for the input of decoding sentences to fit the distribution of words in language level. However, the process

*Both authors contributed equally to this research.
†This work was conducted when Nanxing Li was a research intern at Microsoft Research
‡Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
*ICMR '19, June 10–13, 2019, Ottawa, ON, Canada*
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6765-3/19/06…$15.00
https://doi.org/10.1145/3323873.3325050

(a) (b) (c) (d)

User 1:
(a) Today we went to the zoo and were amazed by the wild life.
(b) Our little boy was really excited to see all those cute little animals.
(c) We also saw the elephants wandering around in the field.
(d) We ended the day relaxing with our very own animals at home.

User 2:
(a) We visited the zoo last weekend.
(b) The kid was a bit upset because the animals wouldn't come out from their caves.
(c) The elephants were fun to watch though.
(d) We were exhausted after the day and saw our dogs lying lazily on the couch.

**Figure 1: Example of stories annotated by different users for the same images. For image (b), user 1 reads excited from the kid while user 2 captures upset. For image (d), the day feels relaxing for user 1 while family is exhausted in the eye of user 2. Different interpretations from two users result in sentences that contain different contents for these two images.**

of making stories from a sequence of images, especially for human being, is rather subjective. Different interpretations to the same image sequence will end with different stories. Even for the same person, making a story for the same image sequence at different times will result in different stories if he/she holds various moods. Existing models that only take consideration of visual contents will result in stories with general sentences due to the fact that they tend to model the common features of different sentences for the same image while failing to modeling the specific features that lead to different sentences. To simulate the real process of storytelling by human, we consider emotion as an important factor that differentiates our interpretations to a given image sequence and guides to different stories. Given the example of Figure 1, different emotion interpretations for image (b) and image (d) result in different sentences for these images and also affect the trend of whole story.

To generate a human-level story from an image sequence, we are facing with the following challenges. First of all, emotion prediction from images is rather difficult, as it usually involves cross-modality inference. Besides, to simulate human process of making stories from image sequence, one image is possible to correspond to several emotions. Moreover, emotion of each image can be influenced by its contextual images in a time sequence. To address the above challenges, we propose a novel emotion reinforced visual story generator, which is the first to introduce emotion as an important feature into visual storytelling. We consider two means of obtaining the emotion features: 1) automatically generated from images and 2) manually customized by users.

Our novel emotion generator is based on the conditional generative adversarial network (cGAN) [24] so that we are able to generate diverse but realistic emotions in a continuous space for the same image. In order to make the generated emotions coherent in one sequence, we connect cGANs of emotion generative model for each image in the image sequence in a recurrent way and sequentially update the generated emotions for each image. Furthermore, our model is capable of generating stories based on customized emotion features. We believe such feature is crucial to visual storytelling, since the same image sequence can be interpreted differently in terms of emotion. The introduction of emotion feature expands the possibilities of visual storytelling.

By leveraging emotion, stories are finally generated by a recurrent neural network and further optimized by policy gradient. Two discriminators and emotion affirmation are jointly used to provide rewards for story generation approximation. The two discriminator networks are designed to guarantee the generated sentences' relevance to the visual content of image sequence and accordance with story language style. While emotion affirmation is designed to measure the consistency of the generated story and input emotion.

We conduct experiments on the largest visual storytelling dataset (VIST) [13]. The generated stories are evaluated in both objective and subjective ways. We define automatic evaluation metrics in terms of visual relevance, emotion relevance and expressiveness. User studies are conducted concerning visual relevance, coherence, expressiveness, and emotiveness. Besides generated emotion, we also test stories generated by having customized emotions as input. The contribution of this paper can be concluded as follows:

- We introduce emotion as an important factor to generate story from image sequence. To the best of our knowledge, this is the first attempt to put forward emotion for visual story generation, which enables a machine to generate variable stories for the same image sequence.
- We propose an emotion reinforced visual story generator incorporating the emotion feature. It consists of an image encoder and a story generator with image and emotion sequence as input, in which two discriminators and an emotion affirmation measurement provide rewards for measuring image relevance, story style and emotion consistency.
- We conduct extensive experiments to demonstrate the effectiveness of our approach compared with several baseline methods in both objective and subjective evaluation metrics.

## 2 RELATED WORK

There are many studies conducted on generating sentence(s) from images. We will review them based on two categories: visual description generation and visual storytelling.

### 2.1 Visual Description Generation

Visual description generation (image captioning and paragraphing) aims to find or generate sentence(s) to describe one image. It was first considered as a retrieval problem so that to describe a given image, the algorithm returns several sentences with similar semantic meaning [9, 14]. The problem of search-based generation is that it cannot provide accurate sentences for images. Template filling method is thus proposed to overcome this problem [17]. Recently, with the development of deep neural network, integration of convolutional neural network (CNN) and recurrent neural network (RNN) is boosting the sentence generation research for readable human-level sentences [1, 7, 16, 20, 25, 31–34]. Later on, generative adversarial network (GAN) [11] is utilized to improve the quality of the generated sentences for different problem settings [6, 35]. Latest work [2, 5] strive to discover other structures for this task. However, as we have previously addressed, the target of image description generation is to use sentence(s) to describe factual visual

content while story is an combination of visual contents and human subjective perception (emotion).

## 2.2 Visual Storytelling

Visual storytelling is a rather new topic but has attracted many attentions. Generating several sentences for the purpose of storytelling is more challenging than visual description for one image. Relationship between different visual contents need to considered to form a good story and sentences for a story have to be coherent. Similar to visual description researches, early works mainly focus on search-based method to retrieve the most suitable sentence combination for an image sequence [23]. [19] proposes a skip Gated Recurrent Unit to deal with semantic relation between image sequence and generated sentences. Then methods leveraged by image captioning, especially CNN-RNN framework is extended for story generation [13]. Recently, we have seen some works that utilize reinforcement learning and generative network for better story generation [12, 26, 28]. Though topic is introduced in [12], existing works still lack of subjective perception of human when making stories, which we first introduce in this paper.

# 3 APPROACH

In this research, we aim to generate a story for a sequence of images. Existing researches on this task ignore the diversity of stories for the same image sequence in the training data, and it usually results in generating general and neutral sentences. In order to model the factor that guides to different stories for the same image sequence, we introduce emotion as an important factor. For this reason, our story generator generates stories considering both visual contents and emotions. The framework is shown in Figure 2.

## 3.1 Overview

We design our model as an encoder-decoder framework, implemented with a hierarchical recurrent neural network (RNN) structure. This is intuitive since our target is to generate a sentence sequence based on the input image sequence, and each sentence can be considered as a sequence of words. Given an image sequence, we first apply a convolutional neural network (CNN) to extract features of each image and use them as input of image encoder.

The image encoder is a RNN with gated recurrent unit (GRU) [8] as its cell. It is used to encode the story feature of each image. At $i$-th time step, we feed image feature of $i$-th image to the GRU as its input, and take the hidden state of the GRU cell as its story feature output. The story feature of the $i$-th image is as follows:

$$\{\mathbf{c}_i\} = E_I(\{\mathbf{x}_i\}), \tag{1}$$

where $E_I$ denotes image encoder, $\{\mathbf{x}_i\}$ denotes the sequence of input image features, and $\{\mathbf{c}_i\}$ denotes the sequence of story features. $i$ ranges from 1 to $N$, with $N$ being the total number of images in the sequence. Each story feature is corresponding to each input image. Coherence among images is enhanced with RNN, and it is crucial to story generation from several image inputs. Therefore we consider the output of the image encoder as the story feature of the current image.

In our encoder-decoder framework, story generator serves as the decoder. We use a language model that predicts the best possible sentence to generate story based on both story feature and emotion feature of the current image.

$$\{\mathbf{s}_i\} = Decoder(\{\mathbf{c}_i \oplus \mathbf{e}_i\}), \tag{2}$$

where $\mathbf{s}_i$ denotes the sentences decoded by the decoder given the story feature and emotion feature $\mathbf{e}_i$ of the input image, and $\oplus$ denotes concatenation.

The emotion features can be generated automatically or defined with customized emotions. To generate emotion automatically, we design a sequential conditional generative adversarial network based on conditional generative adversarial network (cGAN) [21] with an additional GRU layer. With the GRU layer, the generator of one image is affected by both the image and its contextual predicted emotions. To generate diverse yet realistic emotions, we incorporate this GAN structure instead of an classification network for generating emotion. Our model is also capable of handling customized emotion to guide the generation of stories with given emotion feature. Since the same sequence of images can be interpreted differently by different annotators in terms of emotion, we believe being able to customize emotion is crucial for visual storytelling.

Inspired by [26], we implemented reinforcement learning [29], where rewards $r$ consist of three parts: image-relevance $r_I$, story-likeness $r_S$ and emotion-consistency $r_E$. The first two parts, image-relevance and story-likeness are scored by two separate discriminators. Emotion-consistency is proposed to measure the consistency between emotion of generated story and input emotion (either generated or customized).

## 3.2 RNN as Image Encoder

We utilize GRU as the image encoder. Since the GRU generates the output based on the current input and the previous inputs, the output contains the contextual information of the current image. Therefore, coherence among images is enhanced, which is a crucial part of storytelling. We consider the output of the GRU $\mathbf{c}_i$ as the story feature of the input image:

$$\mathbf{c}_i, \mathbf{h}_i = GRU(\mathbf{x}_i, \mathbf{h}_{i-1}), \tag{3}$$

where $\mathbf{h}_i$ denotes the hidden state of the GRU. In order to further enhance the coherence of our generated story features, we feed the mean pooling of all the image features from the sequence to the GRU as the initial hidden-state $\mathbf{h}_{-1}$.

## 3.3 Emotion Feature

The emotion feature serves as one of the inputs to our story generator, which guides the latter to include emotion for the generated stories. Existing researches has shown that there is a pattern between visual content and human perceived emotions [3, 27]. Given an input sentence, we first predicts the emotions conveyed by the sentence through a probability distribution on multiple emotions.

$$\mathbf{e}_i(a) = p(a|\mathbf{s}_i, \theta_{DM}), (a \in \mathcal{F}), \tag{4}$$

where $\mathcal{F}$ denotes the event space (i.e., all the possible emotions), and $\theta_{DM}$ denotes the parameters of the emotion analysis model. $p(a|\mathbf{s}_i, \theta_{DM})$ denotes the probability of emotion $a$ in the distribution given by the analysis model. This probability distribution is then used as the emotion feature in our experiments for training. We use emotion features extracted from ground-truth sentences as input

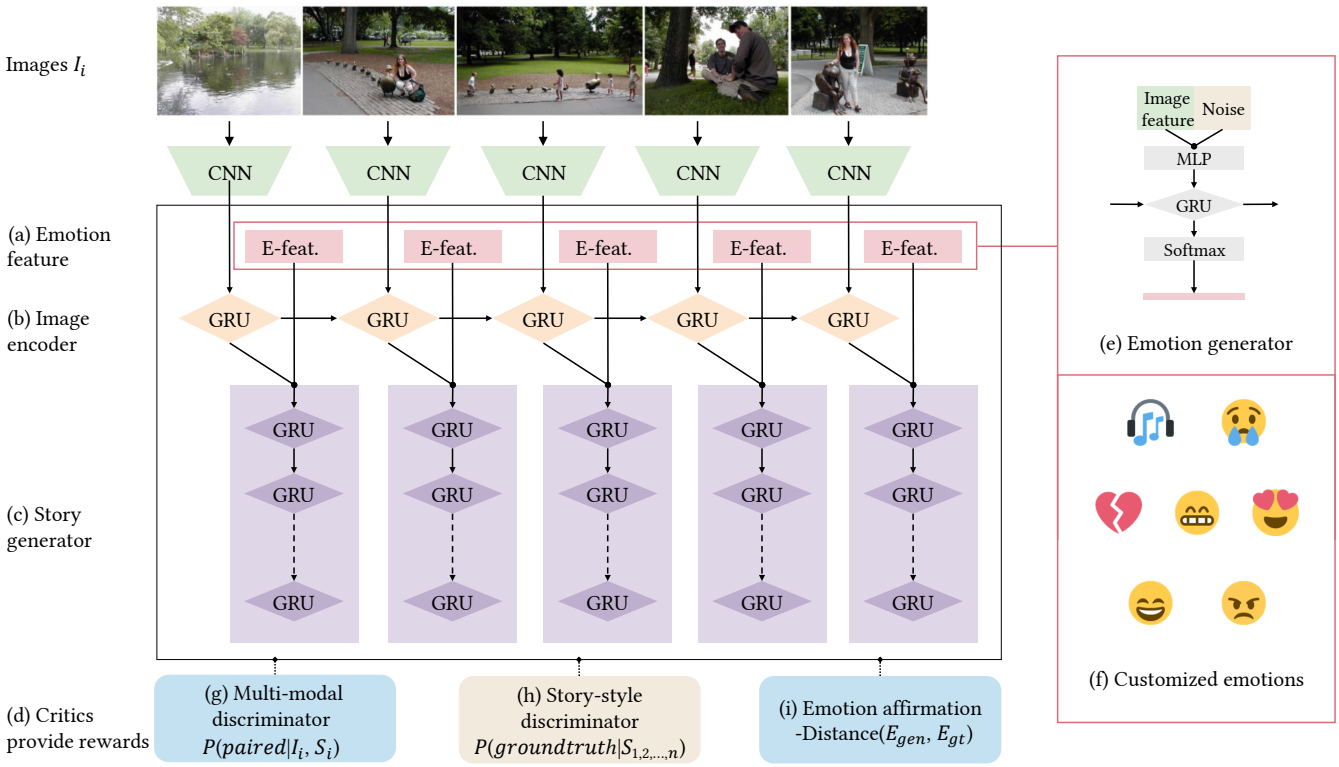


**Figure 2: The framework of emotion reinforced visual story generator. The model can be considered as an encoder-decoder with a hierarchical recurrent neural network structure. Sequential image encoder (b) encodes the images with GRU as story features. Emotion representations are either generated by our proposed emotion generator conditioned by images (e) or predetermined with customized emotions (f). Both story features and emotion features serve as input of story generator (c). In addition, we apply reinforcement learning with multi-modal discriminator (g), story-style discriminator (h) and emotion affirmation (i) to provide rewards for the optimization of the story generator.**

of the decoder during training. The extracted emotion features are also utilized to train our emotion generator.

The emotion generator is designed to generate emotion from images automatically. Since we consider the emotion generation process as a creative process rather than a predictive process, we design the generator as a generative model, which has shown its creativity in recent work [4, 15]. Our emotion generator is based on cGAN [21] and follows the generator-discriminator structure. The generator is a multi-layer perceptron followed by a GRU. Since the emotion features we need are probability distributions, we add a softmax function to the end of the generator.

The emotion generator takes image features $\{\mathbf{x}_i\}$ of an image sequence and a sequence of random noise vectors $\{\mathbf{z}_i\}$ as input and learns to generate a sequence of creative yet plausible emotion vectors. The GRU in the generator enables it to generate context-aware emotions. The emotion generator can be denoted as:

$$\mathbf{y}_i = MLP(\mathbf{x}_i, \mathbf{z}_i), \tag{5}$$

$$\hat{\mathbf{e}}_i, \mathbf{h_i} = GRU(\mathbf{y}_i, \mathbf{h_{i-1}}), \tag{6}$$

$$\mathbf{e}_i = softmax(\hat{\mathbf{e}}_i), \tag{7}$$

where $\hat{\mathbf{e}}_i$ and $\mathbf{h_i}$ denotes the output and the hidden state of the GRU.

We design discriminators in two levels. An instance-level discriminator measures the image-relevance of the generated emotions and a sequence-level discriminator measures the consistency of the generated story sequence. The generator and the discriminators are jointly trained in the same way as cGAN [21].

The story generator is also capable of generating stories based on customized emotion features, which we believe is crucial to storytelling since the same image sequence can be interpreted with different emotions. There are many ways to customize emotions. How customized emotion features are obtained in our experiments will be further discussed in the experiment section.

## 3.4 RNN Decoder as Story Generator

Given the predicted story feature and emotion feature, the decoder predicts the best possible sentence. We use a RNN language model as decoder, which predicts sentences by predicting each word in a sequence according to the story and emotion feature, as well as all the previously predicted words.

## 3.5 Reinforcement Learning

We incorporate reinforcement learning [29] in our approach by considering the story generator as the agent, and the process of

picking up each word as an action given the situation. The generator is guided with a reward that consists of three parts: image-relevance $r_I$, story-likeness $r_S$ and emotion-consistency $r_E$. The first two measurements are judged by two discriminators, an instance-level discriminator $D_I$ that measures the image-relevance and a sequence-level discriminator $D_S$ that measures the story-likeness, as described by [26]. The instance-level discriminator is trained to discriminate paired sentences and images from randomly selected sentences and generated sentences, while the story-level discriminator is trained to discriminate real stories picked from the dataset from stories formed with randomly selected sentences and generated stories.

For the first two rewards, we simply use the probability predicted by the discriminators that measures how likely our generated sentences are ground-truth sentences as the reward functions:

$$r_I(\mathbf{s}_i|\mathbf{x}_i) = P_{D_I}(groundtruth|\mathbf{s}_i, \mathbf{x}_i), \tag{8}$$

$$r_S(\{\mathbf{s}_i\}) = P_{D_S}(groundtruth|\{\mathbf{s}_i\}). \tag{9}$$

Emotion-consistency is computed by the distance between and input emotion and emotion feature of generated sentences. Since we use probability distribution as emotion feature, total variation distance is applied as distance between two emotion features. The negative distance is used as the emotion-consistency reward:

$$r_E(\mathbf{s}_i|\mathbf{e}_i) = -\sup_{a\in\mathcal{F}} |\mathbf{e}_i(a) - p(a|\mathbf{s}_i)|, \tag{10}$$

where $\mathbf{e}_i(a)$ and $p(a|\mathbf{s}_i)$ denote the probability of emotion $a$ in the input emotion feature $\mathbf{e}_i$ and emotion feature extracted from the generated story $\mathbf{s}_i$.

The final reward is the weighted sum of the aforementioned three parts of reward:

$$\begin{aligned} r(\{\mathbf{s}_i\}|\{\mathbf{x}_i\}, \{\mathbf{e}_i\}) = \gamma_1\{r_I(\mathbf{s}_i|\mathbf{x}_i)\} + \gamma_2 r_S + \\ (1-\gamma_1-\gamma_2)\{r_E(\mathbf{s}_i|\mathbf{e}_i)\}. \end{aligned} \tag{11}$$

With this reward, we incorporate policy gradient to train our network. Since we consider story generator as the agent, and each word picked as an action, we have the policy defined as:

$$p_\theta(w_{i,t}|\mathbf{x}_i, \mathbf{e}_i, w_{i,t-1:0}), \tag{12}$$

where $w_{i,t}$ denotes the $t$-th word picked in the $i$-th sentence.

Thus, given the reward $r$, the loss function to minimize can be denoted as:

$$L(\theta) = -\sum_{i=1}^{N}\sum_{t=1}^{T} p_\theta\left(w_{i,t}|\mathbf{x}_i, \mathbf{e}_i; w_{i,t-1:0}\right) r(\mathbf{s}_i|\mathbf{x}_i, \mathbf{e}_i). \tag{13}$$

Following [26], We utilize policy gradient to minimize the loss function and approximate it with Monte-Carlo sample by sampling each $w_{n,t}$. The approximated gradient can be similarly denoted as:

$$\begin{aligned} \nabla_\theta L(\theta) = -\sum_{i=1}^{N}\sum_{t=1}^{T} r(\mathbf{s}_i|\mathbf{x}_i, \mathbf{e}_i)\times \\ \nabla_\theta p_\theta\left(w_{i,t}|\mathbf{x}_i, \mathbf{e}_i; w_{i,t-1:0}\right). \end{aligned} \tag{14}$$

As shown by [26], reinforcement learning with both the instance-level and sequence-level discriminator greatly improves the quality of the generated stories in terms of language. We take one step further by introducing the emotion-consistency reward, which helps our model to generate stories that are more emotive.

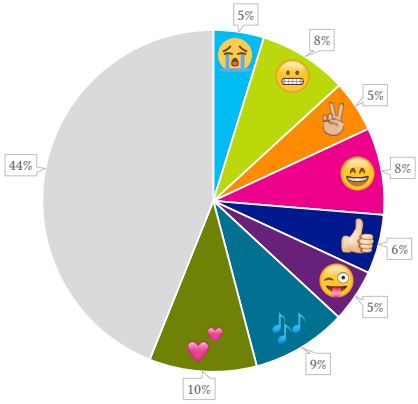


**Figure 3: Distribution of emotions in the training data of VIST. Emotions are analyzed by DeepMoji [10] and clustered into 9 categories. For each category, a typical emoji is shown, as DeepMoji analyzes emotions in the form of emojis. The grey section represents stories with neutral emotion, i.e., stories not displaying strong emotion.**

# 4 EXPERIMENT

## 4.1 Dataset and Emotion Analysis

We conduct our experiments on the VIST dataset created by [13]. The VIST dataset is the current largest dataset created specifically for the task of visual storytelling. It contains 81,743 images obtained from Flickr website with 20,211 image sequences arranged. Each image sequence is annotated with corresponding stories through AMT (Amazon's Mechanical Turk). Each sequence contains 5 images and most sequences has multiple annotations.

For pre-processing, we filtered out sequences with images that are no longer available in the dataset, with 40,143 images and 26,890 sequences remaining for training set, 5,055 images and 1,011 sequences for testing set. In addition, we tokenized the sentences and filtered out words with occurrence less than 4, creating a vocabulary with 10,698 words.

To investigate the importance of emotions for stories, we make an analysis to sentences in the VIST dataset in terms of emotion diversity. We check the diversity of emotions among annotated sentences. We utilize a state-of-the-art method for emotion analysis from text, DeepMoji[10], which is trained on more than 1 billion sentences and the corresponding emotions, achieving as high as 82.4% on human agreement[10]. Figure 3 shows the emotion (corresponding to each emoji respectively) distribution in the training data. We can see that more than half of the sentences infer obvious emotions different and emotions are equally distributed among these emotive sentences.

## 4.2 Implementation Details

In our experiments, we use the outputs from the fc7 layer of a pre-trained VGG16 model, which has 4096 dimensions, as our image features. The sizes of the hidden states of the story encoder RNN and the language decoder RNN are 1,000 and 1,025 respectively. We utilized DeepMoji [10] to extract emotion features for sentences.

The emotion features are of 64 dimensions and are embedded into a space with 25 dimensions.

Before training our model with reinforcement learning, we first pre-trained our generator without the discriminators. The weights of three rewards used for reinforcement learning are empirically set to 0.72 for image-relevance, 0.18 for story-likeness and 0.1 for emotion-consistency, respectively.

## 4.3 Stories with Auto-Generated Emotions

We first conducted experiment with automatically generated emotions with images as the only input. In this way, we can compare fairly with previous researches and demonstrate our model's ability of generating diverse and emotive stories. We conducted both objective and subjective evaluation on the results.

*4.3.1 Compared Methods.* We compare the results of our models with four baseline methods. We include several image/video captioning models and the previous state-of-the-art model on story generation. The models are:

- **Sentence-Concat** [25]: a classic method to incorporate the basic CNN-RNN framework on the problem of image captioning. For story generation, we simply concatenate individual outputs for each image together for the complete story.
- **Regions-Hierarchical** [16]: a hierarchical recurrent neural network that generates several sentences for one image based on regions and bidirectional retrieval. We use images instead of regions to generate corresponding sentences.
- **SRT** [26]: a state-of-the-art visual storytelling model which is the first to incorporate reinforcement learning in the task of storytelling. We test their methods with two settings: **SRT w/o D** and **SRT w/ D** for fair comparison with our model with and without discriminators.
- **Our Model**: to examine the effectiveness of three critics as rewards, we train our model with five settings. Pre-trained model without critics (**Ours w/o critics**), with discriminator only (**Ours w/ $D_m$&$D_s$**), with emotion affirmation only (**Ours w/ E**) and with all critics (**Ours**).

*4.3.2 Objective Evaluation Metrics.* For objective evaluation, similar to other visual storytelling researches, we compare the generated stories with reference stories and compute the language similarity with NLP metric (BLEU [22]). As discussed similarly by [28], sentence translating metrics are not ideal metrics for the task of visual storytelling, as it relies heavily on correlation between predicted sentences and ground-truth sentences, which is heavily biased. Thus, we define two more objective evaluation metrics to measure the emotiveness and novelty of sentences. The metrics are as follows:

- **Relevance:** BLEU is an evaluation metric for machine translation. This metric calculate scores based on the correlation between the generated stories and the ground-truth stories. Note that the relevance here only indicates the relevance between generated stories and reference stories and cannot reflect whether the generated stories are really relevant to image sequence.
- **Emotiveness:** Since we introduce emotion to story generation, we define emotiveness as *to what extent the sentences can express emotions*. We use the confidence score, i.e. , the sum of the probabilities of the top-5 candidate emotions, predicted by DeepMoji [10] for the measurement. Therefore the score ranges from 0 to 1, with a higher score indicates a more emotive story. Similar to inception score in GAN, since DeepMoji is a classification model, the confidence of its prediction can be considered as the clearness and strength of emotion in the sentences.
- **Novelty:** With the introduction of emotion, our generated stories is expected to inclue more diverse and novel words. Therefore we evaluate our method with novelty. Novelty for stories is proposed in [28] for their subjective evaluation. We try to quantify it by computing the less frequent words used in the sentences. Following the novelty definition in [18], we compute the proportion of N-grams that occur in the training dataset except the most 10% frequent ones. We use both bi-gram and tri-gram as measurement here due to the fact that expressive expressions usually reside in phrases rather than words.

*4.3.3 Subjective Evaluation.* We conducted human evaluation since objective metrics are not capable of perfectly evaluate the performance of generated stories, as the evaluation of stories should be considered a subjective task. To better illustrate the performance of stories from human perception, we further conduct extensive user studies. We conducted user studies with 10 volunteers, who are students specialized in English language (4 females and 6 males). Their age distribution is: 20-25 (40%), 25-30 (30%), 30-35 (30%).

This task aims to compare stories generated by our method and baseline methods for same image sequences from different aspects. Given an image sequence, the users were asked to give ratings to a story on a 0-10 scale with respect to four criteria:

- **Relevance:** how much relevant the story is to the given image sequence;
- **Coherence:** how coherent sentences are so that they can easily make a story;
- **Expressiveness:** how expressive the language in the story is, in terms of words, phrases and sentence structure, etc.;
- **Emotiveness:** how strong emotion the story contain.

*4.3.4 Results and Analysis.* **Objective evaluation**: Results of objective evaluation are shown in Table 1. From this result, we can see that our proposed approach is comparable with previous state-of-the-art in terms of BLEU metric which emphasizes the similarity between generated sentences and ground-truth sentences in a translative way. However, this metric is not our main focus due to the fact that the ground-truth sentences are limited and that such metric is not capable of emotion evaluation. On the other hand. our proposed approach achieves much higher score in terms of emotiveness and novelty, with improvements of 6.7% and 24.1% on *Emotiveness* and *Novelty-2*. This shows that our model is able to generate sentences that can express much strong emotion and include more novel words.

**Subjective evaluation**: Results of human evaluation are shown in Table 2. While our proposed approach shows similar performance in terms of relevance and coherence compared with previous research, our approach yields marginally superior results in terms of

**Groundtruth**: We all got together for my parents anniversary. We had amazing steak to celebrate. We had a few making speeches and be funny. There was a few sentimental moments thrown in. Then it quickly turned funny again.
**SRT**: Everyone gathered for a dinner dinner and had a great time at the reception. The food was the main course of the food that was prepared. The bride and groom are relaxing together and are having a great time. And they were happy. the bride and groom are relaxing together and having a great time.
**Ours**: A lovely couple on a small town. We watched so we took our photos in the venue. Friends and family dinner at the restaurant. The bride and groom laughed with their faces and drinking. Later we sat at a bar and ordered some tasty wine before the sun went down.

**Groundtruth**: He loved his boat and spending time on the water. No one had loaded the ski 's so they had to settle for the tube. He gave each of them a turn on the tube. When it was his best friends turn, he pulled him over the wake on purpose. The water was the perfect temperature and the sky was clear and blue.
**SRT**: The family decided to go camping to location for the summer. The pool was beautiful and the scenery was beautiful. The water was still so nice. The water was still so nice. The view from the top was amazing and I loved it so much.
**Ours**: There were a lot of friends at a beach. Then we are headed to to shore. The water was beautiful and the sun was shining bright so we couldn't see the perfect. We d never forget to live in this awesome boat. The proud part of the event was well.

**Figure 4: Example of ground-truth stories annotated by users, generated by state-of-art method (SRT) [26] and by our approach. Stories generated by our proposed approach with our novel emotion generator contain more emotion and are generally more expressive, compared to previous approach without said emotion generator. Expressive and emotive phrases are colored in green, while repetitive phrases are colored in red.**

😊 It is his graduation day he is smiling and having a lot of fun. Graduates are smiling. The students were very loud and smiling. The smiles professor had always really exciting this photo of their son's new hat. The bride was really fun.

😍 The woman smiles at the graduation ceremony. Smiling for the audience to see the graduates. Getting ready for the graduation is what a day. Excited to see the graduation of the graduation. Excited to see the graduation of the new smiles.

😭 Today was my sad day at my [male]'s graduation ceremony. All of our families were there to say goodbye. Everyone was sad to watch the ceremony in a distance. The children are so sad. I was really happy to be able to see my son again.

😄 The little girl was very funny in the the cap she took a picture with her dad. The principal made his best friend posing by the audience with his diploma. The principal congratulated him as the bride and groom give jokes. The bride's brother is hilarious as the high school classmates made funny faces to laugh the ceremony. The bride is amused with her diploma.

😊 It is snowing outside today and it is very nice. It was very happy to see her family on the farm. It was really nice to see all the old shops. They are smiling and having a great time. It is very exciting to see how excited they are is happy.

😍 On my trip today I was excited to see my friends. I have a lot of fun on the road today. Here I am going to see how excited the bike tour is is exciting. Time to see the bike and the boy is so excited. Time to smile on his phone for the first day of the day.

😭 [Male] took the first time of his family trip to location his city and planned to visit his hometown. It was sad seeing the bike ride down to the home I hope that melts soon. It was a long day and the road were sad to see my car. He was going to miss a train after his trip last week and he passed his head in. [Male] was sad to leave the village by the station and we headed back home because he was exhausted.

😄 He was very funny at school when his friends got his first bicycle for him to take a bike. He was on a bike ride that captured him so funny to watch the park. The little girl riding the bike ride very funny. He was very funny at the bike race he was having a blast. Grandpa and son are very funny when he tries to play his bike.

**Figure 5: Example of stories generated with customized emotions. These example show that our model is capable of generating reasonable stories of different emotions based on the same image sequences.**

expressiveness, with 32.2% improvement, and emotiveness, with 25.8% improvement. This agrees with similar improvements seen in the objective evaluation, and demonstrates our model's capability of generating stories with richer emotion and greater diversity, which is the main focus and advantage of our emotion feature and emotion generator.

**Case study**: Examples of stories generated with generated emotions are shown in Figure 4. As can be seen from the examples, stories generated by our model generally contains much more emotions and diversity, while previous models often generates stories that are general and repetitive. This is due to the emotion generator, which enables our model to generate stories that are much more expressive and diverse in terms of words, phrases, etc.

## 4.4 Stories with Customized Emotions

With the introduction of the emotion feature, our model is capable of generating stories with customized emotions. To explore such ability, we feed our model with customized emotion features and images as input at the same time, bypassing the emotion generator.

We use four categories of emotions that are among the most common emotions in the dataset, which are **happy**, **sad**, **excited** and

| Method | B | E | N-2 | N-3 |
|---|---|---|---|---|
| Sentence-Concat[25] | 33.6 | 53.6 | 49.4 | 72.6 |
| Regions-Hierarchical[16] | 37.7 | 60.9 | 46.4 | 78.0 |
| SRT w/o D[26] | 44.5 | 62.9 | 43.5 | 74.2 |
| SRT w/ D[26] | **44.8** | 65.2 | 41.1 | 72.6 |
| Ours w/o critics | 43.2 | 68.4 | 49.5 | 81.9 |
| Ours w/ $D_m$&$D_s$ | 42.5 | 68.9 | 49.1 | 81.9 |
| Ours w/ E | 40.9 | 69.1 | 50.3 | 82.1 |
| Ours | 41.3 | **69.6** | **51.0** | **82.7** |

**Table 1: Automatic evaluation with generated emotions. B, E and N stands for BLEU, emotiveness and novelty, respectively. N-2 and N-3 stands for bi-gram novelty and tri-gram novelty. Note that BLEU scores are computed in comparison with human-annotated ground-truth stories. All scores are reported as percentage (%).**

| Method | Rel | Coh | Exp | Emo |
|---|---|---|---|---|
| Sentence-Concat | 4.80 | 4.40 | 3.90 | 5.20 |
| Regions-Hierarchical | 1.11 | 2.55 | 3.78 | 3.78 |
| SRT | 6.62 | **6.85** | 5.77 | 5.92 |
| Ours | **6.81** | 6.63 | **7.63** | **7.45** |

**Table 2: Human evaluation results of methods on four criteria: relevance (Rel), coherence (Coh), expressiveness (Exp) and emotiveness (Emo). All criteria are evaluated on 0-10 scale (0-bad, 10-good).**

**funny**. To obtain emotion features in each category, we first manually set an emotion feature of that category, and then retrieve with that emotion feature from the emotions extracted from training text. We thus obtain multiple emotion features from the training set of that category. In practice, we use 20 different emotion features for each emotion category. For each emotion category, we assign each input image with one emotion feature in the 20 emotion features in that category randomly. We conduct extensive user studies to evaluate the performance of our approach with customized emotions in a subjective way.

*4.4.1 Human Evaluation.* We conducted human evaluation with subjective metrics. This task aims to evaluate the generated stories, in terms of the emotion represented in the stories and the quality of the stories themselves. The scores are collected from the same 10 students described in section 4.3.3, who are each given 20 stories randomly picked from the results of our testing experiment and their corresponding images. The users are asked to give ratings to the generated stories on a 0-10 scale with respect to the following three criteria:

- **Correspondence**: does the story generated show the emotion it is conditioned with?
- **Reasonableness**: are the emotion represented in a reasonable way?
- **Relevance**: how relevant are the story generated to the corresponding images?

*4.4.2 Results and Analysis.* **User study**: Results of human evaluation are shown in Table 3. The results given are convincing that

| Method | Correspondence | Reasonableness | Relevance |
|---|---|---|---|
| Happy | 8.40 | 7.98 | 6.78 |
| Excited | 8.74 | 8.02 | 7.18 |
| Sad | 7.66 | 6.84 | 6.64 |
| Funny | 8.36 | 7.32 | 6.10 |
| Avg. | 8.29 | 7.54 | 6.68 |

**Table 3: Human evaluation results with customized emotions on three criteria. Each emotion category are evaluated separately while avg. stands for the overall average score.**

our model is capable of generating reasonable stories based on different customized emotions given, as the overall average scores on all three criteria achieve satisfying scores. Especially the criteria regarding emotion, which is our main focus in this work, correspondence and reasonableness, achieve 82.9% and 75.4% respectively. This ability to generate stories based on customized emotion results from our introduction of emotion feature.

It can be noticed that our model performs worse in certain categories, particularly in the **sad** category. This can be explained by the fact that, as can be observed in 3, most images in the VIST dataset are of positive emotion and some images are difficult to be interpreted with **sad** emotion.

**Case study**: Examples of stories generated with customized emotions by our approach are shown in Figure 5. As can be seen from the examples, the generated stories comply with both the given emotion and the images themselves in a reasonable way. This shows our model is able to generate stories based on customized emotions, which is one of its greatest strengths. We believe such capability is crucial in storytelling, while neglected by previous works. We can also notice that stories generated with customized emotion have some grammar problems compared with automatically generated emotion stories. This is due to the large variances of simply defined customized emotion and the image content, and the relatively limited and biased dataset.

## 5 CONCLUSION

We present the first approach in visual storytelling to incorporate emotion as a key factor for the generation of story given a sequence of images. We propose a novel framework with emotion feature to model human emotion in the generation of stories, which enables our model to generate stories of great diversity and expressiveness. We incorporated reinforcement learning, with our novel reward that rewards emotional consistency. Extensive experiments show that our proposed approach is capable of not only generating stories with greater diversity and contains richer emotions, but also generating reasonable stories based on different customized emotions, which greatly extends the variety of visual storytelling.

## ACKNOWLEDGMENTS

This work was supported by National Key R&D Program of China (2018YFB0505400), the National Natural Science Foundation of China (61472202). We thank all volunteers for their effort in the user study.

## REFERENCES


[1] 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 *CVPR*.

[2] Jyoti Aneja, Aditya Deshpande, and Alexander G. Schwing. 2018. Convolutional Image Captioning. In *CVPR*.

[3] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM MM*. 223–232.

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096* (2018).

[5] Fuhai Chen, Rongrong Ji, Xiaoshuai Sun, Yongjian Wu, and Jinsong Su. 2018. GroupCap: Group-Based Image Captioning With Structured Relevance and Diversity Constraints. In *CVPR*.

[6] Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan-Ting Hsu, Jianlong Fu, and Min Sun. 2017. Show, Adapt and Tell: Adversarial Training of Cross-domain Image Captioner. In *ICCV*. 521–530.

[7] Xinlei Chen and C Lawrence Zitnick. 2015. Mind's eye: A recurrent visual representation for image caption generation. In *CVPR*. 2422–2431.

[8] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*. 1724–1734.

[9] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *ECCV*. 15–29.

[10] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *EMNLP*.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*. 2672–2680.

[12] 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 *AAAI*.

[13] Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *NAACL HLT*. 1233–1239.

[14] Andrej Karpathy, Armand Joulin, and Fei Fei F Li. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*. 1889–1897.

[15] Tero Karras, Samuli Laine, and Timo Aila. 2018. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948* (2018).

[16] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *CVPR*. 3337–3345.

[17] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2011. Baby talk: Understanding and generating image descriptions. In *CVPR*. 1601–1608.

[18] Bei Liu, Jianlong Fu, Makoto P Kato, and Masatoshi Yoshikawa. 2018. Beyond Narrative Description: Generating Poetry from Images by Multi-Adversarial Training. In *ACM MM*. 783–791.

[19] Yu Liu, Jianlong Fu, Tao Mei, and Chang Wen Chen. 2017. Let Your Photos Talk: Generating Narrative Paragraph for Photo Stream via Bidirectional Attention Recurrent Neural Networks.. In *AAAI*. 1445–1452.

[20] Alexander Mathews, Lexing Xie, and Xuming He. 2018. SemStyle: Learning to Generate Stylised Image Captions Using Unaligned Text. In *CVPR*.

[21] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *Computer Science* (2014), 2672–2680.

[22] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*. 311–318.

[23] Cesc C Park and Gunhee Kim. 2015. Expressing an image stream with a sequence of natural sentences. In *NIPS*. 73–81.

[24] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *NIPS*. 3483–3491.

[25] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*. 3156–3164.

[26] Jing Wang, Jianlong Fu, Jinhui Tang, Zechao Li, and Tao Mei. 2018. Show, reward and tell: Automatic generation of narrative paragraph from photo stream by adversarial training. AAAI.

[27] Jingwen Wang, Jianlong Fu, Yong Xu, and Tao Mei. 2016. Beyond Object Recognition: Visual Sentiment Analysis with Deep Coupled Adjective and Noun Neural Networks.. In *IJCAI*. 3484–3490.

[28] Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. 2018. No metrics are perfect: Adversarial reward learning for visual storytelling. In *ACL*. 899–909.

[29] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8, 3-4 (1992), 229–256.

[30] Jingjing Xu, Yi Zhang, Qi Zeng, Xuancheng Ren, Xiaoyan Cai, and Xu Sun. 2018. A skeleton-based model for promoting coherence among sentences in narrative story generation. *EMNLP*, 4306–4315.

[31] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*. 2048–2057.

[32] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting image captioning with attributes. In *ICCV*. 22–29.

[33] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *CVPR*. 4651–4659.

[34] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image Captioning With Semantic Attention. In *CVPR*.

[35] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient.. In *AAAI*. 2852–2858.