# Show, Reward and Tell: Automatic Generation of Narrative Paragraph from Photo Stream by Adversarial Training

**Jing Wang**[1*], **Jianlong Fu**[2], **Jinhui Tang**[1†], **Zechao Li**[1], **Tao Mei**[2]
[1]School of Computer Science and Engineering, Nanjing University of Science and Technology
[2]Microsoft Research, Beijing, China
{jwang,jinhuitang,zechao.li}@njust.edu.cn,{jianf,tmei}@microsoft.com

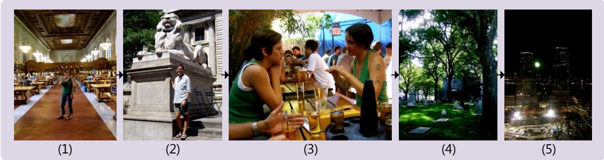
## Abstract

Impressive image captioning results (i.e., an objective description for an image) are achieved with plenty of training pairs. In this paper, we take one step further to investigate the creation of narrative paragraph for a photo stream. This task is even more challenging due to the difficulty in modeling an ordered photo sequence and in generating a relevant paragraph with expressive language style for storytelling. The difficulty can even be exacerbated by the limited training data, so that existing approaches almost focus on search-based solutions. To deal with these challenges, we propose a sequence-to-sequence modeling approach with reinforcement learning and adversarial training. First, to model the ordered photo stream, we propose a hierarchical recurrent neural network as story generator, which is optimized by reinforcement learning with rewards. Second, to generate relevant and story-style paragraphs, we design the rewards with two critic networks, including a multi-modal and a language-style discriminator. Third, we further consider the story generator and reward critics as adversaries. The generator aims to create indistinguishable paragraphs to human-level stories, whereas the critics aim at distinguishing them and further improving the generator by policy gradient. Experiments on three widely-used datasets show the effectiveness, against state-of-the-art methods with relative increase of 20.2% by METEOR. We also show the subjective preference for the proposed approach over the baselines through a user study with 30 human subjects.

## Introduction

Creating a narrative paragraph from an ordered photo stream poses a fundamental challenge to the research on both computer vision and natural language processing. In this paper, we refer to this task as "visual storytelling", which can specifically generate one natural language sentence for each photo. This is challenging because of the difficulty in fully understanding the visual clues and the relation from photo streams, and the difficulty in generating the paragraph with expressive language style for storytelling.

Existing researches have focused more on visual captioning (Vinyals et al. 2015; Xu et al. 2015; Venugopalan et al. 2015) and paragraphing (Krause et al. 2017) for a single image or a short video clip. These works often adopt the paradigm for integrating Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Although significant results have been achieved by training with plenty of image-caption pairs, these works can only generate objective descriptions for a single image or a video clip, which are far from narrative creations.



**Captions**
(1) The woman is standing in the middle of the room.
(2) Man standing near lion monument while posing for the camera.
(3) Two women are talking together in a beer garden.
(4) Sunlight shining through a group of trees onto a small cemetery.
(5) A city on the horizon at night taken from high up.

**Stories**
***Annotator 1:*** We had so much fun *being tourists while looking at colleges*. The statues were awe inspiring. We stopped for lunch at a great *restaurant with a beautiful patio*. We visited some old cemeteries to see the history. The view from our hotel room at the end of the night was amazing.
***Annotator 2:*** The *dance party at church had just finished*. We were standing in front of a nice statue. The *kids were sad* that they didn't get to play patty cake. The trail was very bright and lovely. At night ,The scenery was nice and it would be great if I can visit here again in the future.

Figure 1: Examples of image captioning and storytelling. Both the captions and stories are provided by human annotators. The blue emotional words and the red clauses from story paragraphs are more expressive than captions. [Best viewed in color]

In this paper, we take one step further to investigate the creation of narrative paragraph for a photo stream. Existing works on this task mainly focus on search-based solutions by cross-modality embedding and ranking, because of the limited training data (Park and Kim 2015; Liu et al. 2017). However, these works cannot generate appropriate and emotional stories for new photo streams from personal media collections (e.g., Flickr.com or albums on the phone), which makes storytelling still largely unexplored.

In particular, the challenges of visual storytelling can be summarized as follows. First, different from single image captioning/paragraphing, visual storytelling targets at modeling an ordered photo sequence and further decoding the

[*]This work was performed when Jing Wang was visiting Microsoft Research Asia as a research intern.

[†]Corresponding author.

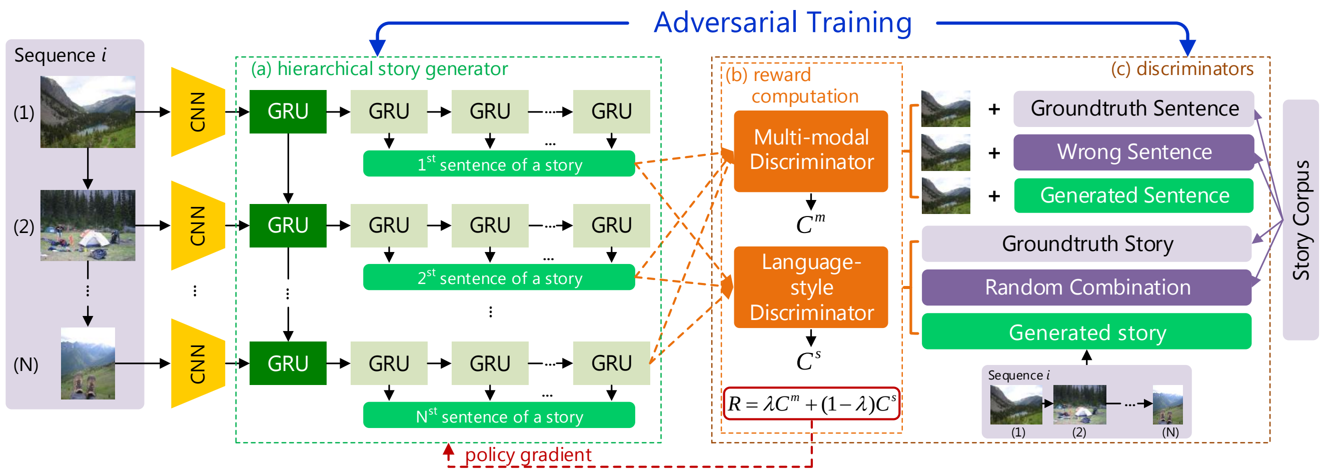


Figure 2: The overview of the proposed method. For storytelling, CNN features for the photo stream with N photos are sequentially input to the dark green GRU encoder in (a) to obtain embeded features for the story generator, which are further fed into the light green GRU decoder to generate sentences. The expected reward $R$ is computed by jointly considering the score $C^m$ from multi-modal discriminator and the score $C^s$ from language-style discriminator in (b). The generator is updated by policy gradient with the reward $R$. The two discriminators in (c) are trained with three types of samples (marked with different colors), respectively. We iteratively update the generator and the discriminators by adversarial training. [Best viewed in color]

joint visual embedding into consistent sentences. Second, as an aphorism goes "a thousand Hamlets in a thousand people's eyes", stories in existing datasets (Huang et al. 2016) can be relevant but diverse to a photo stream. Existing search-based works assume neighborhood preserving property in training (Hadsell, Chopra, and LeCun 2006), i.e., the same image should share similar semantic descriptions, which are hard to converge in storytelling. As in Figure 1, although the paragraphs from two human annotators are relevant to the photo stream, these stories have different meanings from each other. Third, storytelling prefers expressive language style, instead of objective descriptions, which poses grand challenges for the learning of story generators. For example, the emotional words like "had so much fun" and the clauses like "it would be great if" often appear in vivid paragraphs. Even worse, this challenge can further be exacerbated by the limited training data in storytelling.

To generate a narrative paragraph for a photo stream, we propose a sequence-to-sequence modeling approach with reinforcement learning and adversarial training as in Figure 2. First, to model the ordered photo stream, we propose a hierarchical recurrent neural network as story generator, which is optimized by reinforcement learning with rewards. Such a design ensures that the complex visual clues can be captured and a coherent story can be further produced. Note that the hierarchical RNN can sequentially generate sentences by feeding the ordered visual content which has been observed from a photo stream. Second, to generate relevant and story-style paragraphs, we design the rewards with two critic networks, including a multi-modal discriminator and a language-style discriminator. The multi-modal critic determines whether a photo stream and its generated narrative paragraph are semantically relevant. The language-style critic determines whether the generated paragraph is compliant with a human-level story. As the discriminators are designed as triple classifiers, the optimization can be readily to converge with limited training data. Third, to further enhance the capability for storytelling, we propose to use an adversarial training strategy to update the generator and the discriminators in an iterative way. Specifically, the generator aims to create relevant and expressive paragraphs which can be indistinguishable to human-level stories, whereas critics aim at distinguishing them and providing sustainable rewards to further improve the generator by policy gradient.

To the best of our knowledge, this is the first attempt on storytelling by generative models with adversarial training. The main contributions can be summarized as follows:

- We propose a novel reinforcement learning framework with two discriminators as rewards for visual storytelling, which enables the hierarchical generative model to generate relevant and story-style narrative paragraphs.
- We propose to use an adversarial training strategy to further improve the capability for storytelling, which makes the learning of the story generator and the discriminators in a mutually reinforced way.
- We conduct extensive experiments on the three storytelling datasets (Park and Kim 2015; Huang et al. 2016) and achieve superior performance over state-of-the-art methods for all the metrics.

## Related Work

In this section, we review related works along three main dimensions: visual captioning, paragraphing, and storytelling.

**Visual Captioning** In early works (Farhadi et al. 2010; Hodosh, Young, and Hockenmaier 2013; Karpathy, Joulin, and Li 2014), the visual captioning problem is treated as a ranking task, which retrieves existing captions from the database for the given images and hence cannot provide suitable descriptions for new images. Later works try to overcome the problem by template filling (Kulkarni et al. 2011) or adopting the paradigm for integrating Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) (Chen and Lawrence Zitnick 2015; Donahue et al. 2015; You et al. 2016). Multiple instance learning has been introduce to help improve captioning in recent works (Yao et al. 2017) for its outstanding performance in tasks like classification (Qi et al. 2007; Li and Tang 2017). Similar models have been successfully applied to video captioning (Donahue et al. 2015; Yao et al. 2015), which can also be improved by incorporating video concept detection (Zha et al. 2007).

**Visual Paragraphing** Hierarchical neural networks are utilized to produce more detailed and coherent paragraph descriptions for images and short video clips (Yu et al. 2016; Krause et al. 2017). The method proposed by (Yu et al. 2016) leverages the strong temporal dependencies to generate multi-sentence descriptions for cooking videos. A hierarchical recurrent neural network proposed by (Krause et al. 2017) is more related to our work. It generates paragraphs for the images by combining the sentences generated for specific regions. Our method differs from this in the encoding phase and the direct transmission of the encoded feature in the decoding phase.

**Visual Storytelling** The pioneering work was done by (Park and Kim 2015) to explore the description task for image streams. It chooses the most suitable sentence combinations for image streams, considering both the fitness of image-sentence pairs and the sentence coherence. An attention based RNN with a skip gated recurrent unit (Liu et al. 2017) is designed to leverage the semantic relation between photo streams and narrative paragraphs. These works focus on search-based solutions hence are restricted to the scale of the database. The generative model proposed by (Huang et al. 2016) is the naturally extension of image captioning model and regards the whole story as a long caption. Due to vanishing gradients, such a model lack the ability to model the complex paragraph structures. The proposed model applies the hierarchical architecture to solve the problem and leverages two discriminators to ensure the capability of the hierarchical generative model to generate relevant and expressive narrative paragraphs.

## Approach

In order to generate relevant and story-style narrative paragraphs, we propose a sequence-to-sequence modeling approach with reinforcement learning and adversarial training as in Figure 2. The proposed method incorporates a hierarchical story generator and two discriminators into a unified framework. The story generator is regarded as an agent which is responsible for taking actions (i.e., generating words) and is further updated by policy gradient with the guidance of rewards. The discriminators conduct evaluation on the actions and provide rewards for the generator.

### Hierarchical Story Generator as an Agent

**Hierarchical Story Generator** Given a photo stream, the story generator aims at generating a relevant and story-style narrative paragraph. In order to reduce the difficulty for RNNs to learn from paragraph structures (i.e. stories), we propose a hierarchical sequence-to-sequence generative model to decompose the image streams and the long paragraphs into single images and shorter sentences. The generative model includes two RNNs, one of which acts as the image encoder and the other is the sentence decoder. The image encoder sequently embeds image features and integrates all the visual clues observed, whereas the sentence decoder generates sequences of words.

We build both encoder and decoder on Gated Recurrent Units (GRUs) (Cho et al. 2014). Given an image stream $\mathbf{x}$, we first extract the CNN feature $\mathbf{x}_n$ $(n \in 1, \dots, N)$ of each image. At each time step, the image encoder orderly takes $\mathbf{x}_n$ as input and sequently produces the hidden state $\mathbf{h}_n$, which is regarded as the embedded image feature and has integrated the visual information from all the images that have been observed. Once the hidden state $\mathbf{h}_n$ is obtained, the sentence decoder takes it as the initial state and generates a sequence of words $\mathbf{y}_n = [y_{n,1}, \dots, y_{n,T}]$ with length $T$. Finally, after all the corresponding sentences to the images in the stream are produced, the complete story $\mathbf{y} = [\mathbf{y}_1, \dots, \mathbf{y}_N]$ is obtained by concatenating the generated sentences.

**Generator as an Agent** When the hierarchical story generator is viewed as an agent, all the parameters $\theta$ of the generator define a policy $p_\theta(y_{n,t}|\mathbf{x}_{1:n}; \mathbf{y}_{n,1:t-1})$, according to which sequences of actions are taken, where $\mathbf{x}_{1:n} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$ represents the observed images and $\mathbf{y}_{n,1:t-1} = [y_{n,1}, \dots, y_{n,t-1}]$ is the partial description generated in the past time steps for the current image $\mathbf{x}_n$. In the reinforcement learning framework, the generator is guided with rewards and the rewards will not be computed until the generation of the end token or the last word of the maximum sequence length. The calculation of the reward will be introduced in the next section. Given the reward $R(\mathbf{y}_n)$, our goal is to minimize the negative loss function:

$$\begin{aligned} L(\theta) &= -\sum_{n=1}^{N}\sum_{t=1}^{T} p_\theta(y_{n,t}|\mathbf{x}_{1:n}; \mathbf{y}_{n,1:t-1}) R(\mathbf{y}_n) \\ &= -\sum_{n=1}^{N} \mathbb{E}_{\mathbf{y}_n \sim p_\theta}[R(\mathbf{y}_n)]. \end{aligned} \tag{1}$$

We use the expression with $K$ sample paths $\mathbf{y}^k \sim p_\theta$ to approximate the above expectation of the reward:

$$L(\theta) \approx \frac{1}{k}\sum_{k=1}^{K} L_k(\theta), \tag{2}$$

$$L_k(\theta) = -\sum_{n=1}^{N} R(\mathbf{y}_n^k),\ \mathbf{y}_n^k \sim p_\theta. \tag{3}$$

In practice, we simply set $K = 1$, which means that a single sample is taken.

**Policy Gradient** To compute the gradient $\nabla_\theta L(\theta)$, we refer to (Williams 1992; Zaremba and Sutskever 2015; Rennie et al. 2016) to apply the REINFORCE algorithm and get the gradient:

$$\begin{aligned}\nabla_\theta L(\theta) = -&\sum_{n=1}^{N}\sum_{t=1}^{T} p_\theta(y_{n,t}|\mathbf{x}_{1:n};\mathbf{y}_{n,1:t-1}) \\ &\times R(\mathbf{y}_n)\nabla_\theta \log p_\theta(y_{n,t}|\mathbf{x}_{1:n};\mathbf{y}_{n,1:t-1}).\end{aligned} \tag{4}$$

The approximate gradient with K sample paths can be written as:

$$\begin{aligned}\nabla_\theta L(\theta) \approx& \frac{1}{k}\sum_{k=1}^{K}\nabla_\theta L_k(\theta)) \\ =& -\frac{1}{k}\sum_{k=1}^{K}\sum_{n=1}^{N}\sum_{t=1}^{T} R(\mathbf{y}_n^k) \\ &\times \nabla_\theta \log p_\theta(y_{n,t}^k|\mathbf{x}_{1:n};\mathbf{y}_{n,1:t-1}^k).\end{aligned} \tag{5}$$

Since we use a single Monte-Carlo sample $\mathbf{y}^k \sim p_\theta$ in practice, the gradient estimator may suffer from large variance. To decrease the variance, follow (Zaremba and Sutskever 2015), we apply the reward baseline and estimate it with the expectation reward of the current generator. The approximate gradient with baseline b is:

$$\begin{aligned}\nabla_\theta L(\theta) \approx -&\sum_{n=1}^{N}\sum_{t=1}^{T}(R(\mathbf{y}_n) - b) \\ &\times \nabla_\theta \log p_\theta(y_{n,t}|\mathbf{x}_{1:n};\mathbf{y}_{n,1:t-1}).\end{aligned} \tag{6}$$

We have dropped the superscript $k$ for clarity from here on. Using the chain rule, we have

$$\nabla_\theta L(\theta) = \sum_{n=1}^{N}\sum_{t=1}^{T}\frac{\partial L(\theta)}{\partial s_{n,t}}\frac{\partial s_{n,t}}{\partial \theta}, \tag{7}$$

where $s_{n,t}$ is the input of softmax. According to (Zaremba and Sutskever 2015), the approximate partial derivative with a baseline b is given by

$$\frac{\partial L(\theta)}{\partial s_{n,t}} \approx (R(\mathbf{y}_n) - b)(p_\theta(y_{n,t}|h_{n,t}) - 1_{y_{n,t}}). \tag{8}$$

## Discriminators Provide Rewards

A good narrative paragraph has two significant factors: (1) the narrative paragraph should be relevant to the photo stream. (2) the narrative paragraph should resemble human-level stories in language style. According to the two factors, the multi-modal discriminator and the language-style discriminator are designed to provide rewards for the separate sentences and the whole paragraph, respectively.

**Multi-modal Discriminator** In order to estimate the relevance between image $\mathbf{x}_n$ and the generated sentence $\mathbf{y}_n$, the Multi-modal Discriminator $D^m$ is trained to classify ($\mathbf{x}_n$, $\mathbf{y}_n$) into three classes, denoted as *paired*, *unpaired* and *generated*, respectively. The $D^m$ model includes a fusion mechanism similar to (Antol et al. 2015; Li et al. 2015) and a classifier that has a simple structure of a fully connected layer followed by a softmax layer as follows,

$$\mathbf{v}_{x_n} = W_x \cdot \mathbf{x}_n + b_x, \tag{9}$$

$$\mathbf{v}_{y_n} = W_y \cdot \mathrm{LSTM}_\eta(\mathbf{y}_n) + b_y, \tag{10}$$

$$f_n = \tanh(\mathbf{v}_{x_n}) \odot \tanh(\mathbf{v}_{y_n}), \tag{11}$$

$$C_n^m = softmax(W_m \cdot f_n + b_m), \tag{12}$$

where $W_x$,$b_x$,$\eta$,$W_y$,$b_y$,$W_m$,$b_m$ are parameters to be learned. First, in Eqn. 9 and Eqn. 10, the single image $\mathbf{x}_n$ is embeded by a linear layer and the sentence $\mathbf{y}_n$ is put into a Long-short Term Memory network (LSTM) word by word to get the sentence vector. Then, in Eqn. 11, the embeded image feature $\mathbf{v}_{x_n}$ and the sentence vector $\mathbf{v}_{y_n}$ are fused by the fusion mechanism via an element-wise multiply operation. Finally, the classifier takes the fused vector $f_n$ as input and produces probability $C_n^m(c|\mathbf{x}_n, \mathbf{y}_n)$, where $c \in \{paired, unpaired, generated\}$. The probability $C_n^m(paired|\mathbf{x}_n, \mathbf{y}_n)$ indicates how likely the image and the generated sentence are related to each other.

**Language-style Discriminator** In order to determine whether the generated paragraph is compliant with a human-level story, the Language-style Discriminator $D^s$ is trained to differentiate three classes: ground truth stories ($gt$), random combinations of ground truth sentences ($random$) and the generated narrative paragraphs ($generated$). The $D^s$ model is composed of an encoder and a classifier as follows,

$$\mathbf{v}_p = \mathrm{LSTM}_\phi(\bar{\mathbf{p}}), \tag{13}$$

$$C^s = softmax(W_p \cdot \mathbf{v}_p + b_p), \tag{14}$$

where $\bar{\mathbf{p}} = [\bar{\mathbf{p}}_1, \ldots, \bar{\mathbf{p}}_N]$ denotes the paragraph embedding and $\phi, W_p, b_p$ are parameters to be learned. In Eqn. 13, the encoder is a single layer LSTM that recurrently takes the sentence embeddings $\bar{\mathbf{p}}_n$ of a paragraph as input and produces the last hidden state $\mathbf{v}_p$ as the encoded paragraph vector. The sentence embedding $\bar{\mathbf{p}}_n$ here is the average of the embeddings of all sentence words. The encoded paragraph vector $\mathbf{v}_p$ is then put into a classifier as in Eqn. 14 to compute the probability $C^s(gt|\mathbf{y})$, which indicates how likely the paragraph is a real story. All the sentences in the paragraph share the same probability score, namely $C_n^s(gt|\mathbf{y}) = C^s(gt|\mathbf{y})$.

**Reward Function** Based on the above two probability scores, the reward function for sentence $\mathbf{y}_n$ is defined as

$$R(\mathbf{y}_n|\cdot) = \lambda C_n^m(paired|\mathbf{x}_n, \mathbf{y}_n) + (1-\lambda)C_n^s(gt|\mathbf{y}). \tag{15}$$

The contribution of the two discriminators is controlled by the tradeoff parameter $\lambda$.

## Adversarial Training

To further enhance the capability for storytelling, we use an adversarial training strategy to iteratively update the generator and the discriminators. Before adversarial training, we pre-train the generator with cross entropy loss on the training data. This operation is important because it provides a much better policy for us.

Given an image stream $\mathbf{x}$, the generator generates a narrative paragraph $\mathbf{y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$ which is expected to have large reward $R(\mathbf{y}|\cdot) = \sum_{n=1}^{N} R(\mathbf{y}_n|\cdot)$, namely large $C_n^m(paired|\mathbf{x}_n, \mathbf{y}_n)$ for N sentences and large $C^s(gt|\mathbf{y})$ for the whole story. At the same time, the discriminators try to distinguish between the generated narrative paragraphs and human-level stories. That is to say, when the generator provides a paragraph, the discriminators attempt to increase the probabilities $C_n^m(generated|\mathbf{x}_n, \mathbf{y}_n)$ and the probability $C^s(gt|\mathbf{y})$. The adversarial training strategy further improves the capability for the generative model to generate relevant and story-style narrative paragraphs.

# Experiment

In this section, we conduct experiments on three widely-used visual storytelling datasets and present the comparisons with some state-of-the-art methods.

## Datasets

**SIND** SIND (Huang et al. 2016) is the first dataset which is created for the sequential vision-to-language task and includes 20057 sequences with 50200 stories. The image sequences are collected from Flickr with several event titles as searching keywords and then annotated by Amazon's Mechanical Turk (AMT). Each story has 5 images and 5 corresponding descriptions. The 50200 stories have been split into three parts, 40155 for training, 4990 for validation and 5055 for testing, respectively.

**Disney** Disney (Park and Kim 2015) is collected from blog posts with "Disneyland" as searching topic. There are 11863 blog posts and 78467 images in total. Following (Park and Kim 2015), we take 80% for training, 10% for validation and 10% for testing, respectively.

**NYC** NYC (Park and Kim 2015) is collected from blog posts with "NYC" as searching topic. In total, there are 11863 blog posts and 78467 images, in which 80% are used for training, 10% for validation and 10% for testing as in (Park and Kim 2015).

## Metrics

To quantitatively evaluate all the models, we adopt three popular metrics in vision-to-language tasks: BLEU@N (Papineni et al. 2002), CIDEr-D (Vedantam, Lawrence Zitnick, and Parikh 2015) and METEOR (Lavie 2014). We use the codes [1] released by Microsoft COCO Evaluation Server to compute all the metrics (Chen et al. 2015).

## Compared Methods

We compare the proposed storytelling model with four related approaches to evaluate the quantitative results. One of them is the narrative paragraph generation model and the others are three of the most popular methods for image/video captioning and paragraphing. We also compare the proposed method with the generative model and two search-based methods for user study.

**Sentence-Concat (Vinyals et al. 2015):** Sentence-Concat is a popular image caption model which leverages CNN-RNN framework to generate captions for single images.

**Story-Flat (Huang et al. 2016):** Story-Flat is a basic sequence-to-sequence model, which resembles the encoder-decoder model in machine translation. Story-Flat has two RNNs, one of which acts as an encoder to encode the image sequences and the other acts as a decoder to generate the whole story word by word.

**S2VT (Venugopalan et al. 2015):** S2VT, which is known as Sequence to Sequence-Video to Text model, utilizes stacked LSTM to generate descriptions for video clips.

**Regions-Hierarchical (Krause et al. 2017):** Regions-Hierarchical is a hierarchical recurrent neural network which generates paragraphs for single images by producing sentences for each region in images.

**CRCN (Park and Kim 2015):** CRCN leverages CNN, RNN and an entity-based local coherence model to learn the semantic relations between long streams of images and texts.

**BARNN (Liu et al. 2017):** BARNN is an attention based RNN with a skip gated recurrent unit which leverages the semantic relation between photo streams and stories.

Note that as the models from image/video captioning and paragraphing cannot be directly applied to our task, we make minor changes to adapt these methods to our task. For Sentence-Concat, the sentences corresponding to the images in a story are concatenate to be a paragraph. For S2VT, We keep the stacked LSTM structure, treat the images in a sequence as frames in a video clip and generate the whole story for the photo stream. As the features of different regions are pooled and mapped as the topic vector for the sentence RNN in Regions-Hierarchical, we similarly average the image features in the sequences.

## Experimental settings

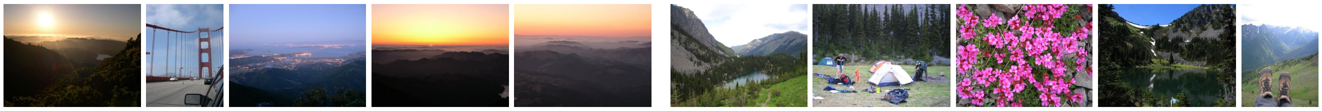
For SIND dataset, we evaluate all the methods on the whole validation set for a fair comparison. For Disney and NYC, we process the blog data according to the procedures stated as follows. Given an image, we aim to represent the corresponding description with the most relevant sentence. First, three most relevant labels obtained from VGGNet (Simonyan and Zisserman 2015) are concatenated as the representation of the image content. Then, the pre-trained skip-thought model (Kiros et al. 2015) is utilized to extract the 4800-dim vectors for the image content representation and the blog sentences, respectively. Finally, the cosine distance is computed and the nearest sentence to the image is selected to form an image-sentence pair together with the corresponding image. Similar to (Huang et al. 2016), we fix the image sequence length $N = 5$. Those sequences with the length less than 5 are dropped and longer sequences are decomposed to one or more. When performing evaluation on Disney and NYC, the only metric METEOR is taken for the reason that METEOR is proved to be better than CIDEr-D in the small references case and superior to BLEU@N all the time (Vedantam, Lawrence Zitnick, and Parikh 2015).

For all the compared methods, same VGG16 fc7 features are taken as representations of images. We have applied finetuning in the experiments on SIND. Single-layer Gated Recurrent Units (GRUs) with 1000 dimensions are ap-

[1] https://github.com/tylin/coco-caption

Table 1: The compared results (%) on SIND in terms of six language metrics.

| Method | METEOR | CIDEr | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
|---|---|---|---|---|---|---|
| Sentence-Concate (Vinyals et al. 2015) | 10.52 | 8.45 | 38.28 | 18.20 | 8.73 | 4.17 |
| Story-Flat (Huang et al. 2016) | 10.25 | 6.84 | 36.53 | 16.52 | 7.50 | 3.50 |
| S2VT (Venugopalan et al. 2015) | 9.81 | 6.76 | 34.44 | 15.99 | 7.54 | 3.64 |
| Regions-Hierarchical (Krause et al. 2017) | 9.97 | 6.51 | 34.92 | 16.00 | 7.69 | 3.70 |
| Ours (w/o two D) | 11.03 | 9.79 | 40.88 | 19.50 | 9.31 | 4.49 |
| Ours (w/o language-style D) | 12.19 | 10.99 | 43.04 | 21.07 | 10.22 | 4.98 |
| Ours (w/o multi-modal D) | 11.15 | 10.69 | 41.07 | 19.77 | 9.67 | 4.81 |
| Ours | **12.32** | **11.35** | **43.40** | **21.35** | **10.40** | **5.16** |



**Ground truth:** The view of the sunset was beautiful. We decided to get back on the road before it got too dark. When we hopped on the plane we could see the entire landscape. The view of the sunset was beautiful from the plane. I had a great time.
**Proposed:** We went on vacation to a beautiful country. We drove through a tunnel that we saw a lot of traffic on our way through the city. We took a helicopter ride to the top of the mountain to see the beautiful view. The sunset was the best part of the day. The sun was setting and it was a beautiful day for a hike.

**Ground truth:** We traveled to the mountains for a camp out last weekend. First we set up our tent and camping area. Then we went on a hike, the mountain flowers were beautiful. We then passed by the lake, it was so calm. After our hike we just relaxed and enjoyed the beautiful view.
**Proposed:** We went on a hike in the mountains of location. We are camping and we are going to go camping. There are some beautiful flowers growing in the garden. We took a hike through the forest and took a picture of the beautiful scenery. We had a great view of the mountains from our trip.

Figure 3: Examples for narratives generated by the proposed model and stories from ground truth. The words in same colors indicate the correct semantic matches between the generated narratives and those from ground truth. [Best viewed in color]

plied for all methods except Regions-Hierarchical (Krause et al. 2017). We keep the two-layer LSTM for word RNN in Regions-Hierarchical and set the hidden size $H = 500$. The two discriminators are both implemented as a single-layer LSTM with 512 dimensions.

Before adversarial training, we first pre-train the generator on the whole training set with the cross entropy loss. We use ADAM optimizer with the initial learning rate set to $5 \times 10^{-4}$. Schedule sampling (Bengio et al. 2015) and batch normalization (Ioffe and Szegedy 2015) are applied for mitigating the exposure bias and stable training respectively. We select the best model according to the validation performance and use it as the initial model for adversarial training.

In the adversarial training phase, we take ADAM optimizer with the learning rate being reduced to $5 \times 10^{-5}$. The two discriminators are both trained from scratch and do not share parameters with each other.

## Main Results

**Comparison with Baseline methods** Experiment results for storytelling on the three datasets are shown in Table 1, 2 and 3. The results on all metrics show that our proposed model greatly outperforms all the baselines. As expected, the proposed method which applies hierarchical RNN exhibits better performance than Story-Flat and S2VT, which use a flat structure and treat the story as a long caption. This indicates the effectiveness of the hierarchical model that learns from the long paragraph structures. Further more, the proposed method using not only hierarchical RNNs but also discriminators leads to a performance boost against **Ours (w/o two D)**, which denotes that no discriminators are used. This clearly demonstrates the effectiveness of the discriminators that impose the sequence-to-sequence relevance and the story-style to the paragraph generator. There is a performance gap between Regions-Hierarchical and the proposed method. Despite the similar hierarchical architecture, the Regions-Hierarchical model performs not good. This is because the pooling operation for image features has destroyed the complex relation between images.

Figure 3 shows some qualitative results from the proposed model and example stories from groundtruth. The promising results provide further evidence that the proposed storytelling approach has the capability to generate relevant, expressive and flexible stories.

**Study of Discriminators** In order to evaluate the two discriminators, we decompose our design of discriminators and conduct experiments respectively utilizing only one of them. Results are shown in table 1, with **Ours (w/o language-style D)** denoting using only multi-modal discriminator and **Ours (w/o multi-modal D)** denoting using only language-style discriminator. Using only multi-modal discriminator brings significant improvement, which verifies that the stronger correlation between photo streams and generated narrative paragraphs is obtained by the multi-modal discriminator. Though using only language-style discriminator achieves lower values, the relative increases compared to **Ours (w/o two D)** are 1.1 % by METEOR and 9.2 % by CIDEr. In future, we will use more unpaired story data from book corpus (e.g., (Zhu et al. 2015)) to improve the language-style discriminator. Finally, the two discriminators together achieve the best performance across all metrics.

**Parameter Sensitivity Analysis** Figure 4 shows the effect of the tradeoff parameter $\lambda$. All metric values are normalized

Table 2: Disney dataset (METEOR in %, higher is better).

| Method | METEOR |
|---|---|
| Sentence-Concate (Vinyals et al. 2015) | 8.01 |
| Story-Flat (Huang et al. 2016) | 7.61 |
| S2VT (Venugopalan et al. 2015) | 6.34 |
| Regions-Hierarchical (Krause et al. 2017) | 7.72 |
| Ours (w/o two D) | 7.82 |
| Ours | **9.90** |

Table 3: NYC dataset (METEOR in %, higher is better).

| Method | METEOR |
|---|---|
| Sentence-Concate (Vinyals et al. 2015) | 6.97 |
| Story-Flat (Huang et al. 2016) | 7.37 |
| S2VT (Venugopalan et al. 2015) | 7.38 |
| Regions-Hierarchical (Krause et al. 2017) | 6.07 |
| Ours (w/o two D) | 7.39 |
| Ours | **8.39** |

as follows,

$$m'_\lambda = \frac{m_\lambda - \min\limits_\lambda \{m_\lambda\}}{\min\limits_\lambda \{m_\lambda\}}, \tag{16}$$

where $m_\lambda$ and $m'_\lambda$ are the metric values before and after normalized, respectively.

We observe that when $\lambda$ varies from 0.1 to 0.9, all the curves are subject to unimodal distribution and the best value is achieved when $\lambda$ is around 0.7. This further demonstrates the rationality of the combination of the two discriminators.

## User Study

For the reason that search-based methods often achieve satisfying results, we perform a human study to compare our method against the generative Story-Flat method and two search-based methods, i.e., CRCN (Park and Kim 2015) and BARNN (Liu et al. 2017), to further verify the effectiveness. Though search-based methods share the same vocabulary with the proposed method, they can only retrieve sentences from the training data, whereas the proposed method is capable of generating new sentences. A real-world personal photo set of 150 photo streams (5 photos each) has been collected from 30 volunteers (15 females and 15 males). We will share the data (150 photo streams) in future. All of the volunteers are experienced bloggers. The education background distribution is: economics (16.7%), computer science (33.3%), linguistics (6.7%), engineering (13.3%), biology (20%) and art (10%). The age distribution is: 21-25 (23.3%), 26-30 (33.3%), 31-35 (20%), 36-40 (13.3%) and 41-45 (10%). After reading the paragraphs produced by the four methods, the volunteers are asked to give subjective scores (1-10, 10 means best) according to the two criteria:

- Relevance (C1): whether the story is relevant to the photo stream?
- Story-style (C2): whether the story has expressive and story-style language?

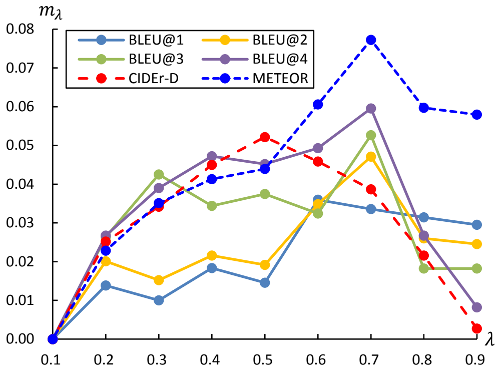


Figure 4: The effect of the tradeoff parameter $\lambda$ in the proposed method on SIND.

Table 4: User study results on real-world photo streams.

| Method | C1 | C2 |
|---|---|---|
| CRCN (Park and Kim 2015) | 2.05 | 3.95 |
| BARNN (Liu et al. 2017) | 3.49 | 6.23 |
| Story-Flat (Huang et al. 2016) | 5.75 | 5.42 |
| Ours | **6.63** | **6.68** |

Table 4 lists the results. We observe that the proposed method achieves highest scores across both criteria. In particular, the proposed method achieves 6.63 and 6.68 in terms of C1 and C2, higher than the retrieval method BARNN (Liu et al. 2017) by 2.14 and 0.45 improvements, respectively. The low values of C2 for the two retrieval methods demonstrate their weakness in generating relevant narratives for novel data. All the results indicate that our method is superior to all the baselines and works well on new user photos.

## Conclusion

In this paper, we propose a hierarchical generative model to generate narrative paragraphs for photo streams. Thanks to the hierarchical structure and the reinforcement learning framework with two discriminators, the proposed model is capable of creating relevant and expressive narrative paragraphs. The story generator leverages the hierarchical recurrent neural network to learn from the paragraph structure. The two discriminators act as critics and ensure the relevance and the story-style for the generated paragraphs. Extensive experiments demonstrate the effectiveness of the proposed approach for storytelling. In the future, we will collect more image-stream-to-sentence-sequence pairs and leverage more unpaired data to further improve the capability of the model for storytelling.

## Acknowledgement

This work was partially supported by the 973 Program (Project No. 2014CB347600), the National Natural Science Foundation of China (Grant No. 61522203, 61772275 and 61732007), the Natural Science Foundation of Jiangsu Province (Grant BK20140058 and BK20170033) and CCF-Tencent Open Research Fund.

## References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. VQA: Visual question answering. In *ICCV*.

Bengio, S.; Vinyals, O.; Jaitly, N.; and Shazeer, N. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *NIPS*.

Chen, X., and Lawrence Zitnick, C. 2015. Mind's eye: A recurrent visual representation for image caption generation. In *CVPR*.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv:1504.00325*.

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *EMNLP*.

Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; and Darrell, T. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*.

Farhadi, A.; Hejrati, M.; Sadeghi, M. A.; Young, P.; Rashtchian, C.; Hockenmaier, J.; and Forsyth, D. 2010. Every picture tells a story: Generating sentences from images. In *ECCV*.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *CVPR*.

Hodosh, M.; Young, P.; and Hockenmaier, J. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR* 47:853–899.

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

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*.

Karpathy, A.; Joulin, A.; and Li, F. F. F. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*.

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

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

Kulkarni, G.; Premraj, V.; Dhar, S.; Li, S.; Choi, Y.; Berg, A. C.; and Berg, T. L. 2011. Baby talk: Understanding and generating image descriptions. In *CVPR*.

Lavie, M. D. A. 2014. Meteor universal: Language specific translation evaluation for any target language. *ACL*.

Li, Z., and Tang, J. 2017. Weakly supervised deep matrix factorization for social image understanding. *TIP* 26(1):276–288.

Li, Z.; Liu, J.; Tang, J.; and Lu, H. 2015. Robust structured subspace learning for data representation. *TPAMI* 37(10):2085–2098.

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*.

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

Park, C. C., and Kim, G. 2015. Expressing an image stream with a sequence of natural sentences. In *NIPS*.

Qi, G. J.; Hua, X. S.; Rui, Y.; Mei, T.; Tang, J.; and Zhang, H. J. 2007. Concurrent multiple instance learning for image categorization. In *CVPR*.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2016. Self-critical sequence training for image captioning. In *CVPR*.

Simonyan, K., and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. *ICLR*.

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

Venugopalan, S.; Rohrbach, M.; Donahue, J.; Mooney, R.; Darrell, T.; and Saenko, K. 2015. Sequence to sequence-video to text. In *ICCV*.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *CVPR*.

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

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*.

Yao, L.; Torabi, A.; Cho, K.; Ballas, N.; Pal, C.; Larochelle, H.; and Courville, A. 2015. Describing videos by exploiting temporal structure. In *ICCV*.

Yao, T.; Pan, Y.; Li, Y.; Qiu, Z.; and Mei, T. 2017. Boosting image captioning with attributes. In *CVPR*.

You, Q.; Jin, H.; Wang, Z.; Fang, C.; and Luo, J. 2016. Image captioning with semantic attention. In *CVPR*.

Yu, H.; Wang, J.; Huang, Z.; Yang, Y.; and Xu, W. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In *CVPR*.

Zaremba, W., and Sutskever, I. 2015. Reinforcement learning neural turing machines. *Technical report* 419.

Zha, Z. J.; Mei, T.; Wang, Z.; and Hua, X. S. 2007. Building a comprehensive ontology to refine video concept detection. In *CVPR*.

Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*.