What Makes A Good Story? Designing Composite Rewards for Visual Storytelling

Junjie Hu,[1] **Yu Cheng,**[2] **Zhe Gan,**[2] **Jingjing Liu,**[2] **Jianfeng Gao,**[3] **Graham Neubig**[1]
[1]Carnegie Mellon University, [2]Microsoft Dynamics 365 AI Research, [3]Microsoft Research
{junjieh, gneubig}@cs.cmu.edu, {yu.cheng, zhe.gan, jingjl, jfgao}@microsoft.com

Abstract

Previous storytelling approaches mostly focused on optimizing traditional metrics such as BLEU, ROUGE and CIDEr. In this paper, we re-examine this problem from a different angle, by looking deep into what defines a natural and topically-coherent story. To this end, we propose three assessment criteria: *relevance*, *coherence* and *expressiveness*, which we observe through empirical analysis could constitute a "high-quality" story to the human eye. We further propose a reinforcement learning framework, ReCo-RL, with reward functions designed to capture the essence of these quality criteria. Experiments on the Visual Storytelling Dataset (VIST) with both automatic and human evaluation demonstrate that our ReCo-RL model achieves better performance than state-of-the-art baselines on both traditional metrics and the proposed new criteria.

Introduction

There has been a recent surge of interest in enabling machines to understand the semantics of complex visual scenarios and depict visual objects/relations with natural language. One main line of research is grounding the visual concepts of a single image to textual descriptions, known as image captioning (Fang et al. 2015; Vinyals et al. 2015; You et al. 2016). Visual storytelling (Huang et al. 2016) takes one step further, aiming at understanding photo streams and generating a sequence of sentences to describe a coherent story.

Most existing visual storytelling methods focus on maximizing data likelihood (Yu, Bansal, and Berg 2017), topic consistency (Huang et al. 2019), or expected rewards by imitation learning (Wang et al. 2018b). However, maximizing data likelihood or implicit rewards does not necessarily optimize the quality of generated stories. In fact, we find that simply optimizing on standard automatic evaluation metrics may even hurt the performance of story generation according to other assessments that are more important to the human eye.

In this paper, we revisit the visual storytelling problem by asking ourselves the question: *what makes a good story?* Given a photo stream, the first and foremost goal should be telling a story that accurately describes the objects and the concepts that appear in the photos. This can be termed as the *"Relevance"* dimension. Secondly, the created story should read smoothly. In other words, the consecutive sentences should be semantically and logically coherent with each other, instead of being mutually-independent sentences describing each photo separately. This can be termed as the *"Coherence"* dimension. Lastly, to tell a compelling story that can vividly describe the visual scenes and actions in the photos, the language used for creating the story should contain a rich vocabulary and diverse style. We call this the *"Expressiveness"* dimension.

Most existing storytelling approaches that optimize on BLEU or CIDEr do not perform very well on these dimensions. As shown in Figure 1, compared with the model-generated story, the human-written one is more semantically relevant to the content of the photo stream (e.g., describing more fine-grained visual concepts such as "flower girls"), more structurally coherent across sentences, and more diversified in language style (e.g., less repetition in pattern such as "great time").

Motivated by this, we propose a reinforcement learning framework with composite reward functions designed to encourage the model to generate a relevant, expressive and coherent story given a photo stream. The proposed ReCo-RL (Relevance-Expressiveness-Coherence through Reinforcement Learning) framework consists of two layers: a high-level decoder (i.e., manager) and a low-level decoder (i.e., worker). The manager summarizes the visual information from each image into a goal vector, by taking into account the overall story flow, the visual context, and the sentences generated for previous images. Then it passes on the goal vector to each worker, which generates a word-by-word description for each image, guided by the manager's goal.

The proposed model consists of three quality evaluation components. The first *relevance function* gives a high reward to a generated description that mentions fine-grained concepts in an image. The second *coherence function* measures the fluency of a generated sentence given its preceding sentence, using a pre-trained language model. The third *expressiveness function* penalizes phrasal overlap between a generated sentence and its preceding sentences. The

BLEU-RL	it was a great time at the wedding. there was a lot of the wedding. it was a great time at the reception. it was a great time to the end of the day. at the end of the boat, we went to the boat.
Reference	a wedding is getting ready to happen . there is flower girls waiting the bride . father is bring her out . the bride and groom is getting married now . the bride and groom are kissing each other . this wedding is taking place on a boat .

Figure 1: Comparison between a story generated by the BLEU-RL model that is trained to optimize BLEU and human-written reference. Words in yellow indicate that there are more fine-grained concepts in the human-written reference than the model-generated one. The two segments in blue show an undesired repeating pattern in the output from the model.

framework aggregates these rewards and optimizes with the REINFORCE algorithm (Williams 1992). Empirical results demonstrate that ReCo-RL can achieve better performance than state-of-the-art baselines. Our main contributions can be summarized as follows:

- We propose three new criteria to assess the quality of text generation for the visual storytelling task.
- We propose a reinforcement learning framework, ReCo-RL, with composite rewards designed to align with the proposed criteria, using policy gradient for training.
- We provide quantitative analysis, qualitative analysis, and human evaluation to demonstrate the effectiveness of our proposed model.

Related Work

Visual Storytelling is a task where given a photo stream, the machine is trained to generate a coherent story in natural language to describe the photos. Compared with visual captioning tasks (Vinyals et al. 2015; Krishna et al. 2017; Rennie et al. 2017; Gan et al. 2017), visual storytelling requires capabilities in understanding more complex visual scenarios and generating more structured expressions. Pioneering work has used sequence-to-sequence model on this task (Park and Kim 2015). Huang et al. (2016) provided the benchmark dataset VIST for this task. Yu, Bansal, and Berg (2017) have shown promising results on VIST with a multi-task learning algorithm for both album summarization and sentence generation.

Recent efforts have explored REINFORCE training, by learning an implicit reward function (Wang et al. 2018b) to mimic human behavior or injecting a topic consistency constraint during training (Huang et al. 2019). Wang et al. (2018a) proposed a hierarchical generative model to create relevant and expressive narrative paragraphs. To improve the structure and diversity, Li et al. (2018) reconciled a traditional retrieval-based method with a modern learning-based method to form a hybrid agent. Notably, these studies did not directly (or explicitly) examine what accounts for a good story to the human eye, which is the main focus of our work.

Text Generation State-of-the-art text generation methods use encoder-decoder architectures for sequence-to-sequence learning (Rajendran et al. 2018; Sutskever, Vinyals, and Le 2014). To better model structured information, hierarchical models have been proposed (Li, Luong, and Jurafsky 2015). Follow-up work tried to overcome exposure bias resulting from MLE training (Bengio et al. 2015; Lamb et al. 2016). In recent years, reinforcement learning (RL) has gained popularity in many tasks (Ranzato et al. 2016), such as image captioning (Rennie et al. 2017), text summarization (Paulus, Xiong, and Socher 2018) and story plot generation (Tambwekar et al. 2019). Other techniques such as adversarial learning (Yu et al. 2017; Dai et al. 2017; Zhang et al. 2017), inverse reinforcement learning (IRL) (Ho and Ermon 2016) and pre-training (Chen et al. 2019) have also been applied. Compared with previous work, we define explicit rewards for the visual storytelling task and propose a reinforcement learning framework to optimize them.

Meanwhile, how to assess the quality of generated text still remains a major challenge. BLEU (Papineni et al. 2002) and METEOR (Banerjee and Lavie 2005) are widely used in machine translation. CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015) and SPICE (Anderson et al. 2016) are used for image captioning. ROUGE-L (Lin 2004) is used for evaluating text summarization. However, these metrics all have limitations in evaluating natural language output, as there exists a large gap between automatic metrics and assessment by humans. There have been some recent studies on more natural assessment for text generation tasks, such as evaluating on structuredness, diversity and readability (Yao et al. 2018; Dai et al. 2017; Chen and Bansal 2018; Wang et al. 2018b), although these studies do not explicitly consider relevance between a stream of images and a story for the task of visual storytelling. Similar to these studies, we argue that the aforementioned automatic metrics are not sufficient to evaluate the visual storytelling task, which requires high readability and naturalness in generated stories.

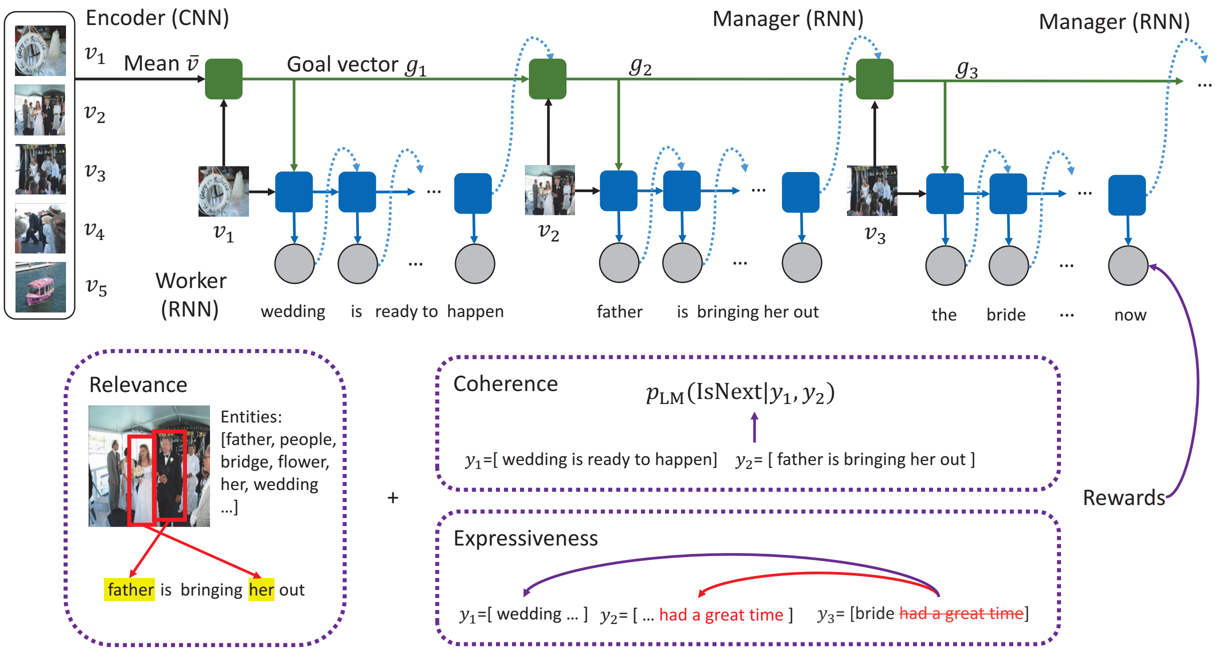
Encoder (CNN)
v_1
v_2
v_3
v_4
v_5
Mean $\bar{v}$
Goal vector g_1
g_2
g_3
Manager (RNN)
Manager (RNN)
...
v_1
v_2
v_3
Worker (RNN)
wedding is ready to happen
father is bringing her out
the bride ... now
Relevance
Entities: [father, people, bridge, flower, her, wedding ...]
father is bringing her out
+
Coherence
$p_{\mathrm{LM}}(\mathrm{IsNext}|y_1, y_2)$
y_1=[wedding is ready to happen] y_2= [father is bringing her out]
Expressiveness
y_1=[wedding ...] y_2= [... had a great time] y_3= [bride ~~had a great time~~]
Rewards

Figure 2: Model architecture and three rewards. Words highlighted in yellow show relevant concepts in the image.

Approach

Notation

Given a stream of n images, we denote their features extracted by a pre-trained convolutional neural network as a sequence of vectors $V \equiv [\mathbf{v}_1, \cdots, \mathbf{v}_n]$. The reference descriptions are denoted as a sequence of sentences $Y \equiv [y_1^*, \cdots, y_n^*]$, where y_i^* is a sequence of word indices that depicts the i-th image. We define a dataset of input-output pairs as $\mathcal{D} = \{(V, Y)\}$. Based on the reference i-th image, our model generates the corresponding sentence y_i, where y_i^t denotes the t-th word in y_i. We denote $\mathbf{E}$ as the word embedding matrix, and $\mathbf{e}_i^t = \mathbf{E}[y_i^t]$ as the word embedding of y_i^t. We denote the hidden state of the manager and the worker as $\mathbf{h}_{M,i}$ and $\mathbf{h}_{W,i}^t$, respectively. We use a bold letter to denote a vector or matrix, and use a non-bold letter to denote a sequence or a set.

Model Architecture

The Encoder module consists of a pre-trained convolutional neural network which extracts deep visual features from each image, with ResNet-101 (He et al. 2016). The encoder obtains the overall summary of a photo stream by averaging the visual features of all the images, i.e., $\bar{\mathbf{v}} = \frac{1}{n}\sum_i \mathbf{v}_i$.

The Manager module in our model is a Long Short-term Memory (LSTM) network, which captures the consistency of the generated story at the sentence-level. When depicting one image of a photo stream, the manager should take into account three aspects: 1) the overall flow of the photo stream; 2) the context information in the current image; and 3) the sentences generated from previous images in the photo stream. To do so, for each image in the i-th step, the manager takes as input the features of the whole image sequence $\bar{\mathbf{v}}$, the features of the i-th image $\mathbf{v}_i$, and the worker's last hidden state $\mathbf{h}_{W,i-1}^T$ from the previous image. The manager then predicts a hidden state as the goal vector.

$$\mathbf{h}_{M,i} = \mathrm{LSTM}_M\left([\bar{\mathbf{v}}; \mathbf{v}_i; \mathbf{h}_{W,i-1}^T], \mathbf{h}_{M,i-1}\right) \qquad (1)$$

where $[;]$ denotes vector concatenation. The goal vector is then passed on to the worker, and the worker is responsible for completing the generation of word description based on the goal from the manager.

The Worker module is a fine-grained LSTM network, which predicts one word at a time and controls the fluency of one sentence. Intuitively, the worker is guided by the goal from the manager, and focuses more on fine-grained context information in the current image. More specifically, when predicting one word at the t-th step, the worker takes as input the features of the i-th image $\mathbf{v}_i$, the manager's goal vector $\mathbf{h}_{M,i}$, and the word embedding of the previously generated word $\mathbf{e}_i^{t-1}$. The worker then predicts a hidden state $\mathbf{h}_{W,i}^t$ and applies a linear layer f to approximate the probability of choosing the next word in Eq. (3).

$$\mathbf{h}_{W,i}^t = \mathrm{LSTM}_W\left([\mathbf{v}_i; \mathbf{h}_{M,i}; \mathbf{e}_i^{t-1}], \mathbf{h}_{W,i}^{t-1}\right) \qquad (2)$$

$$p_{\boldsymbol{\theta}}(y_i^t | y_i^{1:t-1}, \mathbf{v}_i, \bar{\mathbf{v}}) = \mathrm{softmax}(f(\mathbf{h}_{W,i}^t)) \qquad (3)$$

$$p_{\boldsymbol{\theta}}(y_i | \mathbf{v}_i, \bar{\mathbf{v}}) = \prod_t p_{\boldsymbol{\theta}}(y_i^t | y_i^{1:t-1}, \mathbf{v}_i, \bar{\mathbf{v}}) \qquad (4)$$

Composite Rewards Design

Relevance One way to measure the relevance between an image and its generated description is to ground the entities mentioned in the description to corresponding bounding boxes in the image. However, a straightforward way of comparing the n-gram overlap between the reference sentence and the generated sentence (e.g., BLEU or METEOR) treats each word in the sentence equally, without taking into account the semantic relevance of the words to the image.

To tackle this limitation, we propose to measure the semantic similarity between entities mentioned in the reference and generated sentences. More specifically, we are given a set of K reference sentences $Y_i^* = \{y_{i,k}^*\}_{k=1}^K$ for the i-th image. We then extract a set of entities $O^{Y_i^*}$ mentioned in its reference sentences Y_i^* with a Part-Of-Speech (POS) tagger, and count the frequency of the entities in its reference sentences as $C(o, Y_i^*), \forall o \in O^{Y_i^*}$. The normalized frequency of an entity is computed by dividing by the sum of the frequency of all entities of in $O^{Y_i^*}$ in Eq. (5).

$$F(o, Y_i^*) = \frac{C(o, Y_i^*)}{\sum_{o' \in O^{Y_i^*}} C(o', Y_i^*)} \tag{5}$$

Similarly, we extract all the entities mentioned in an n-gram of a hypothesis y_i sampled by the model, and denote the hypothesis n-gram as $\mathfrak{N}$ and its entity set as $O^{\mathfrak{N}}$. To measure the relevance of each hypothesis n-gram with respect to the key concepts in an image, we compute the relevance weight of an n-gram in Eq. (6).

$$W^{\mathfrak{N}} = 1 + \beta \sum_{o \in O^{\mathfrak{N}} \cap O^{Y_i^*}} F(o, Y_i^*) \tag{6}$$

If a hypothesis n-gram contains any key entities in $O^{Y_i^*}$, $W^{\mathfrak{N}}$ is greater than 1, which distinguishes it from other n-grams that do not ground to any bounding objects in the image. Notice that the weight is proportional to the number of key entities in $O^{Y_i^*}$ and the entity frequency in the reference sentences Y_i^*. Intuitively, the more entities an n-gram contains, the more bounding objects in the image this n-gram grounds to. If an entity is mentioned by multiple annotators in the reference sentences, the weight of mentioning this entity in the hypothesis should be high.

Inspired by the modified n-gram precision in the BLEU score calculation, we aim to avoid rewarding multiple identical n-grams in the hypothesis. To this end, we count the maximum number of times an n-gram exists in any single reference sentence in Eq. (7), and clip the count of each hypothesis n-gram by its maximum reference count in Eq. (8). We then compute the weighted precision of all the n-grams in the hypothesis y_i in Eq. (9).

$$C_{\text{max}}(\mathfrak{N}, Y_i^*) = \max_{y_{i,k}^* \in Y_i^*} C(\mathfrak{N}, y_{i,k}^*) \tag{7}$$

$$C_{\text{clip}}(\mathfrak{N}, y_i) = \min\{C(\mathfrak{N}, y_i), C_{\text{max}}(\mathfrak{N}, Y_i)\} \tag{8}$$

$$P_n = \frac{\sum_{\mathfrak{N} \in y_i} C_{\text{clip}}(\mathfrak{N}, y_i) \cdot W^{\mathfrak{N}}}{\sum_{\mathfrak{N}' \in y_i} C(\mathfrak{N}', y_i) \cdot W^{\mathfrak{N}'}} \tag{9}$$

The relevance score of a sampled hypothesis with respect to the key concepts of an image is computed as the product of a brevity penalty and the geometric mean of the weighted n-gram precision in Eq. (10). In our implementation, we consider unigram and bigram, i.e., $n = 2$, since most entities only contain one or two words.

$$\mathcal{R}(y_i) = \text{BP}\left(\prod_{i=0}^{n} P_n\right)^{\frac{1}{n}} \tag{10}$$

$$\text{BP} = \exp\left(\min\left(1 - \frac{r}{|y_i|}, 0\right)\right) \tag{11}$$

Coherence A coherent story should organize its sentences in a correct sequential order and preserve the same topic among adjacent sentences. One way to measure coherence between two sentences is a sentence coherence discriminator that models the probability of two sentences y_{i-1} and y_i being continuous in a correct sequential order as well as containing the same topic.

To this end, we leverage a language model with a next-sentence-prediction objective, as was explored in Devlin et al. (2019). We first construct a sequence by concatenating two sentences y_{i-1} and y_i decoded by our model, and get the sequence representation using a pre-trained language model. Then, we apply a linear layer to the sequence representation followed by a $\tanh$ function and a softmax function to predict a binary label, which indicates whether the second sentence is the sentence that follows the first one.

$$\mathbf{u}_{i-1,i} = \text{LM}(y_{i-1}, y_i) \tag{12}$$

$$p_{\text{LM}}(s|y_{i-1}, y_i) = \text{softmax}\left(\tanh\left(\mathbf{W}\mathbf{u}_{i-1,i} + \mathbf{b}\right)\right) \tag{13}$$

$$\mathcal{C}(y_i) = p_{\text{LM}}(s = 0|y_{i-1}, y_i) \tag{14}$$

where $s = 0$ indicates y_i is the sentence that follows y_{i-1}.

Expressiveness An expressive story should contain diverse phrases to depict the rich content of a photo stream, rather than repeatedly using the same words. To capture this expressiveness, we keep track of already-generated n-grams, and punish the model when it generates repeated n-grams.

To this end, we propose a diversity reward which measures the n-gram overlap between the current sentence y_i and previously decoded sentences $\{y_1, \cdots, y_{i-1}\}$. More specifically, we first regard all the preceding sentences $\{y_1, \cdots, y_{i-1}\}$ as the reference sentences to the current sentence y_i, and compute the BLEU score of the current sentence compared to the reference sentences. Finally we substract this value from 1 as the expressiveness reward in Eq. (15). Intuitively, if the current sentence contains more identical n-grams as any one of preceding decoding sentences, the BLEU score of the current sentence with respect to that already-generated sentence would be high, thus the story is lack of expressiveness when adding the current decoding sentence. In our implementation of BLEU in Eq. (15), we only consider the precision of bigram, trigram and 4-gram, since we want to focus on repeated phrases that have more than one word.

$$\mathcal{E}(y_i) = 1 - \text{BLEU}(y_i, \{y_1, \cdots, y_{i-1}\}) \tag{15}$$

Training

We first train our proposed model using maximum likelihood estimation (MLE), and then continue training the

Method	METEOR	ROUGE	CIDEr	BLEU-4	SPICE
AREL	**35.2**	29.3	**9.1**	13.6	**8.9**
HSRL	30.1	25.1	5.9	9.8	7.5
MLE	34.8	30.0	7.2	14.3	8.5
BLEU-RL	**35.2**	**30.1**	6.7	**14.4**	8.3
ReCo-RL	33.9	29.9	8.6	12.4	8.3

Table 1: Comparison between different models on METEOR, ROUGE-L, CIDEr, BLEU-4 and SPICE.

model using REINFORCE algorithm together with an MLE objective.

Maximum Likelihood Estimation seeks an optimal solution $\boldsymbol{\theta}^*$ by minimizing the negative log-likelihood of predicting the next word over batches of training observations in Eq. (16). We apply stochastic gradient descent to update the model parameters on each mini-batch of data $\mathcal{D}'$ in Eq. (17).

$$J_{\text{MLE}}(\boldsymbol{\theta}, \mathcal{D}') = \sum_{Y,V \in \mathcal{D}'} \sum_{i=1}^{n} -\log p_{\boldsymbol{\theta}}(y_i^* | \mathbf{v}_i, \bar{\mathbf{v}}) \tag{16}$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \eta \frac{\partial J_{\text{MLE}}(\boldsymbol{\theta}, \mathcal{D}')}{\partial \boldsymbol{\theta}} \tag{17}$$

where η is the learning rate.

REINFORCE (Williams 1992) is able to learns a policy by maximizing an arbitrary expected reward in Eq. (18). This makes it possible to design reward functions specifically for the visual storytelling task. We compute the weighted sum of the aforementioned three reward functions, to encourage the model to focus on those key aspects of a good story and control the generation quality of the sentences.

$$J_{\text{RL}}(\boldsymbol{\theta}) = \sum_{Y,V \in \mathcal{D}'} \mathbb{E}_{y_i \sim \pi_i} \left[(b - r(y_i)) \log \pi_i \right] \tag{18}$$

$$r(y_i) = \lambda_R \mathcal{R}(y_i) + \lambda_C \mathcal{C}(y_i) + \lambda_E \mathcal{E}(y_i) \tag{19}$$

where $\pi_i \equiv p_{\boldsymbol{\theta}}(y_i | \mathbf{v}_i, \bar{\mathbf{v}})$ is the policy, and b is a baseline that reduces the variance of the expected rewards, λ_R, λ_C and λ_E are the weights of the three designed rewards. In our implementation, we sample H hypotheses generated by the current policy π_i for the i-th image, and approximate the expected rewards with respect to the empirical distribution π_i. We compute the baseline by using the average reward of all the sampled hypotheses, i.e., $b = \frac{1}{H} \sum_{y_i \sim \pi_i} r(y_i)$.

Rather than starting from a random policy model, we start from a model pre-trained by the MLE objective, and continue training the model jointly with MLE and REINFORCE objectives on each mini-batch $\mathcal{D}'$ in Eq. (20), following Ranzato et al. (2016).

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \eta_1 \frac{\partial J_{\text{MLE}}(\boldsymbol{\theta}, \mathcal{D}')}{\partial \boldsymbol{\theta}} + \eta_2 \frac{\partial J_{\text{RL}}(\boldsymbol{\theta}, \mathcal{D}')}{\partial \boldsymbol{\theta}} \tag{20}$$

Experiment

Dataset and Baseline

Dataset: The VIST dataset (Huang et al. 2016) used in our evaluation consists of 10,117 Flickr albums with 210,819 unique photos. Each sample contains one story that describes 5 selected images from a photo stream, and the same album is paired with 5 different stories as references. The split is similar to previous work, with 40,098 samples for training, 4,988 for validation and 5,050 for testing. The vocabulary size of VIST is 12,977. The released data was processed by a name entity recognition (NER) tagger to solve the sparsity issue of low-frequence words. The name of a person, a location and an organization are replaced by [male]/[female], [location], and [organization], respectively.

Implementation Details: The visual features are extracted from the last fully-connected layer of ResNet152 pretrained on ImageNet (He et al. 2016). The word embeddings of size 300 are uniformly initialized within $[-0.1, 0.1]$. We use a 512-hidden-unit LSTM layer for both the manager and the worker modules. We apply dropout to the embedding layer and every LSTM layer with the rate of 0.3. We set the hyper-parameters $\lambda_R = \lambda_C = \lambda_E = 1$ to assign equal weights to all the three aspects of the reward functions, and set $\eta_1 = \eta_2 = 1$ to balance both MLE and REINFORCE objectives during training. We use BERT (Devlin et al. 2019) as our next sentence predictor and fine-tune the predictor on sentence pairs in the correct and random order in the VIST dataset. For negative sentence pairs, we randomly concatenate two sentences in two different albums to make sure that the topics of these sentences are different.

Baseline: We compare our method with the following baselines: (1) *AREL* (Wang et al. 2018b)[1], an approach to learn an implicit reward with imitation learning; (2) *HSRL* (Huang et al. 2019)[2], a hierarchical RL approach that injects a topic consistency constraint during training. These two approaches achieved state-of-the-art results on VIST, and we follow the same parameter settings in the original papers.

In addition, we also compare three variants of our model: (1) *MLE* that uses MLE training in Eq. (17); (2) *BLEU-RL* that is jointly trained by MLE and REINFORCE, using sentence-level BLEU as a reward; and (3) *ReCo-RL* that is jointly trained by MLE and REINFORCE, using the designed rewards in Eq. (20). The decoding outputs generated are evaluated by the same scripts as Wang et al. (2018b).

Quantitative Evaluation

Automatic metrics, including METEOR, CIDEr, BLEU-4, ROUGE-L and SPICE are used for quantitative evaluation. Table 1 summarizes the results of all the methods in comparison. Our models (MLE, BLEU-RL and ReCo-RL) achieve competitive or better performance over the baselines on most metrics except CIDEr. Specially, BLEU-RL achieves better performance in METEOR, ROUGE-L and BLEU-4, while ReCo-RL improves the CIDEr score.

In addition to the automatic metrics, we can also use the designed reward functions to score each story generated by

[1] https://github.com/eric-xw/AREL.git

[2] Codes are provided by the authors.

Aspects	AREL	ReCo-RL	Tie	Agree	HSRL	ReCo-RL	Tie	Agree	MLE	ReCo-RL	Tie	Agree	BLEU-RL	ReCo-RL	Tie	Agree
R	27.6%	62.2%	10.2%	0.72	36.1%	53.8%	10.1%	0.74	27.0%	64.1%	8.9%	0.49	17.6%	74.5%	7.9%	0.78
C	31.3%	58.7%	10.0%	0.78	38.0%	51.9%	10.1%	0.80	34.3%	57.7%	8.0%	0.53	18.9%	72.3%	8.8%	0.71
E	32.4%	58.6%	9.0%	0.68	38.6%	53.3%	8.1%	0.72	30.5%	61.0%	8.5%	0.55	19.5%	71.5%	9.0%	0.62

Table 2: Pairwise human comparison between ReCo-RL and three methods on three quality aspects (**R**: Relevance, **C**: Coherence, **E**: Expressiveness). For each pairwise comparison, the first three columns indicate the percentage that turkers prefer one system outputs over the other one, and turkers think both stories are of equal quality. The last column is the Fleiss' kappa (Fleiss and others 1971) which is a statistical measure of inter-rater consistency. Agreement scores in the range of $[0.6, 0.8]$ show substantial agreement between multiple turkers.

Method	Relevance	Coherence	Expressiveness
HSRL	1.95	7.21	33.27
AREL	3.27	9.90	34.98
MLE	5.46	7.92	30.76
BLEU-RL	2.17	12.40	30.41
ReCo-RL	**10.39**	**12.74**	**39.37**

Table 3: Comparison between different models on three rewards, i.e., Relevance, Coherence and Expressiveness.

different methods. To evaluate the overall performance of one method at the corpus level, we average the reward scores of all stories generated by the method on the test set. Similar to the automatic evaluation metrics, we multiply the average reward scores by 100 and report the scaled results of all the methods on the test set in Table 3. Our proposed ReCo-RL method outperforms all the start-of-the-art methods and our variants (BLEU-RL, MLE) on all three quality aspects.

Human Evaluation

Due to the subjective nature of the storytelling task, we further conduct human evaluation to explicitly examine the quality of the stories generated by all the models, through crowdsourcing using Amazon Mechanical Turk (AMT). Specifically, we randomly sampled 500 stories generated by all the models for the same photo streams. Given one photo stream and the stories generated by two models, three turkers were asked to perform a pairwise comparison and select the better one from the two stories based on three criteria: *relevance*, *coherence* and *expressiveness*. The user interface of the evaluation tool also provides a neutral option, which can be selected if the turker thinks both outputs are equally good on one particular criterion. The order of the outputs for each assignment is randomly shuffled for fair comparison. Notice that in the pairwise human evaluation, each pair of system outputs for one photo stream was judged by a different group of three people. The total number of turkers for all photo streams is 862.

Table 2 reports the pairwise comparison between ReCo-RL and three other methods. Based on human judgment, the quality of the stories generated by ReCo-RL are significantly better than the BLEU-RL variant on all dimensions, even though BLEU-RL is fine-tuned to obtain comparative scores on existing automatic metrics. Comparing with two strong baselines, AREL and HSRL, ReCo-RL can still achieve better performance. For each pairwise comparison between two model outputs, we also scored each story based on the number of votes from three turkers, and performed the Student's paired t-test between the scores of two systems. Our ReCo-RL is significantly better than all baseline methods with $\rho < 0.05$.

Qualitative Analysis

In Figure 3, we show two image streams and the stories generated by four models. For the second image stream on the right, BLEU-RL repeatedly generates uninformative segments, such as "we had a lot of people there", even though BLEU-RL achieves high scores on automatic metrics. The same problem exists in the stories generated by HSRL in the first and second examples such as "i had a great time.". From our observation, when the images are similar across an image stream, the three baseline methods are not able to discover the different content between subsequent images, thus generating repeated sentences with redundant information.

With regards to the relevance between the visual concepts in the image stream and the stories, ReCo-RL consistently generates more specific concepts highly correlated to the appearing objects in the image stream. In Figure 3, words highlighted in yellow represent the entities that can be grounded in the images. In the second example, our ReCo-RL is encouraged to generate rare entities such as "sign" and "flags" in addition to frequent entities such as "people".

In the first example of Figure 3, sentence pairs that are not semantically coherent are highlighted with an underline. The forth sentence generated by AREL mentions "the president of the company" that is quite different from the previously-described entity "military officer", showing that AREL forgets the content in previous images when it generates the next sentence. Similarly the second sentence generated by HSRL suddenly changes the subject of the story from "i" to "he", and mentions the "new professor" that is quite different from the previously-described entity "new team". From our observation, this type of disconnection is quite common in stories generated by the three baseline methods. The stories generated by ReCo-RL are a lot more coherent in content.

Moreover, we further compare the stories generated by our proposed ReCo-RL and our variant BLEU-RL. These two methods use different sentence-level reward functions during the reinforcement training. In Figure 4, we find that ReCo-RL generates more related entities such as "meal" and "drink". We also observe that the key entities generated by ReCo-RL make the story more consistent in the topic,

Methods		
AREL	the officers of the military officers are in charge of the military. he was very proud of his speech. the meeting was a great success. the president of the company gave a speech to the audience. we had a great time.	the wedding was held at a church. the wedding was beautiful. the bride and groom cut the cake. the bride and groom were very happy. the whole family was there to celebrate.
HSRL	i was so excited to see my new team. he was very happy to see the new professor. i had a lot of time to talk about. i had a great time. i had a great time.	the wedding was a beautiful wedding. the bride and groom cut the dance together. the bride and groom were very happy to be married. the bride and groom were very happy. at the end of the night, the bride and groom were happy to be married.
BLEU-RL	at the end of the day, the men were very proud of the military. they had a lot of people there. this is a picture of the meeting. a group of people had a great time. after the end of the day , we all had a lot of questions.	it was a beautiful day for the wedding. at the end of the night, the bride and groom were very happy. the bride and groom were very happy. she was so happy to be married. the bride and groom pose for pictures.
ReCo-RL	today was a picture of the military officer, he was ready to go to the organization. they were very happy to see the awards ceremony. the speaker was very excited to be able to talk about the meeting. everyone was having a great time to get together for the event after the ceremony. we all had a lot of people there.	it was a beautiful day at the wedding party. the bride and groom were so happy to be married. [female] was happy and she was so excited to celebrate. she had a great time to take a picture of her wedding. all of the girls posed for pictures.

Figure 3: Example stories generated by our model and the baselines. Words in yellow indicate entities appearing in the image, and words in blue show repetitive patterns. Pairs of sentences that describe different topics are annotated by an underline.

Method		Quality Metrics			
		R	C	E	B
BLEU-RL	a group of friends gathered together for dinner. the turkey was delicious. the guests were having a great time. at the end of the night, we had a great time. at the end of the night, we had a great time.	2.47	11.06	37.10	73.57
ReCo-RL	a group of friends gathered together for a party . the turkey was delicious . it was a delicious meal . everyone was having a great time . after the party , we all sat down and talked about the night . my friend and [female] were very happy to drink .	3.32	16.99	41.71	78.51

Figure 4: Example stories generated by our model and BLEU-RL. Words in yellow indicate entities appearing in the image, and words in blue show repetitive patterns. Quality metrics including our proposed reward scores and BLEU-4 (R: Relevance, C: Coherence, E: Expressiveness, B: BLEU-4) are shown on the right.

while BLEU-RL forgets the previous context when generating the last two sentences. Our proposed ReCo-RL also obtains higher scores of our proposed rewards than BLEU-RL .

Conclusion

In this paper, we propose ReCo-RL, a novel approach to visual storytelling, which directly optimizes story generation quality on three dimensions natural to human eye: relevance, coherence, and expressiveness. Experiments demonstrate that our model outperforms state-of-the-art methods on both the existing automatic metrics and the proposed assessment criteria. In future work, we will extend the proposed model to other text-generation tasks, such as storytelling based on some writing prompts (Dianqi et al. 2019) and table-to-text generation (Wiseman, Shieber, and Rush 2018).

Acknowledgments

This work was partially done while the first author was an intern at Microsoft Dynamics 365 AI Research.

References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. SPICE: Semantic propositional image caption evaluation. In *ECCV*.

Banerjee, S., and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop*.

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

Chen, Y.-C., and Bansal, M. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *ACL*.

Chen, Y.-C.; Li, L.; Yu, L.; Kholy, A. E.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2019. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*.

Dai, B.; Fidler, S.; Urtasun, R.; and Lin, D. 2017. Towards diverse and natural image descriptions via a conditional GAN. In *ICCV*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Dianqi, L.; Yizhe, Z.; Zhe, G.; Yu, C.; Chris, B.; Ming-Ting, S.; and Bill, D. 2019. Domain adaptive text style transfer. In *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Fang, H.; Gupta, S.; Iandola, F.; Srivastava, R. K.; Deng, L.; Dollár, P.; Gao, J.; He, X.; Mitchell, M.; Platt, J. C.; et al. 2015. From captions to visual concepts and back. In *CVPR*.

Fleiss, J., et al. 1971. Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin*.

Gan, Z.; Gan, C.; He, X.; Pu, Y.; Tran, K.; Gao, J.; Carin, L.; and Deng, L. 2017. Semantic compositional networks for visual captioning. In *CVPR*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Ho, J., and Ermon, S. 2016. Generative adversarial imitation learning. In *NeurIPS*.

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

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

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Niebles, J. C. 2017. Dense-captioning events in videos. In *ICCV*.

Lamb, A. M.; Goyal, A. G. A. P.; Zhang, Y.; Zhang, S.; Courville, A. C.; and Bengio, Y. 2016. Professor forcing: A new algorithm for training recurrent networks. In *NeurIPS*.

Li, Y.; Liang, X.; Hu, Z.; and Xing, E. P. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. In *NeurIPS*.

Li, J.; Luong, M.-T.; and Jurafsky, D. 2015. A hierarchical neural autoencoder for paragraphs and documents. In *ACL*.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*.

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 *NeurIPS*.

Paulus, R.; Xiong, C.; and Socher, R. 2018. A deep reinforced model for abstractive summarization. In *ICLR*.

Rajendran, J.; Ganhotra, J.; Singh, S.; and Polymenakos, L. 2018. Learning end-to-end goal-oriented dialog with multiple answers. In *EMNLP*.

Ranzato, M.; Chopra, S.; Auli, M.; and Zaremba, W. 2016. Sequence level training with recurrent neural networks. In *ICLR*.

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

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *NeurIPS*.

Tambwekar, P.; Dhuliawala, M.; Martin, L. J.; Mehta, A.; Harrison, B.; and Riedl, M. O. 2019. Controllable neural story plot generation via reward shaping. In *IJCAI*.

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

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

Wang, J.; Fu, J.; Tang, J.; Li, Z.; and Mei, T. 2018a. Show, reward and tell: Automatic generation of narrative paragraph from photo stream by adversarial training. In *AAAI*.

Wang, X.; Chen, W.; Wang, Y.-F.; and Wang, W. Y. 2018b. No metrics are perfect: Adversarial reward learning for visual storytelling. In *ACL*.

Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*.

Wiseman, S.; Shieber, S.; and Rush, A. 2018. Learning neural templates for text generation. In *EMNLP*.

Yao, L.; Peng, N.; Weischedel, R. M.; Knight, K.; Zhao, D.; and Yan, R. 2018. Plan-and-write: Towards better automatic storytelling. In *AAAI*.

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

Yu, L.; Bansal, M.; and Berg, T. 2017. Hierarchically-attentive RNN for album summarization and storytelling. In *EMNLP*.

Yu, L.; Zhang, W.; Wang, J.; and Yu, Y. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*.

Zhang, Y.; Gan, Z.; Fan, K.; Chen, Z.; Henao, R.; Shen, D.; and Carin, L. 2017. Adversarial feature matching for text generation. In *ICML*.