Curiosity-driven Reinforcement Learning for Diverse Visual Paragraph Generation

Yadan Luo[†], Zi Huang[†], Zheng Zhang[†], Ziwei Wang[†], Jingjing Li[‡], Yang Yang[‡]

[†]The University of Queensland
[‡]University of Electronic Science and Technology of China
lyadanluol@gmail.com,huang@itee.uq.edu.au,darrenzz219@gmail.com,ziwei.wang@uq.edu.au,lijin117@yeah.net,
dlyyang@gmail.com

ABSTRACT

Visual paragraph generation aims to automatically describe a given image from different perspectives and organize sentences in a coherent way. In this paper, we address three critical challenges for this task in a reinforcement learning setting: the mode collapse, the delayed feedback, and the time-consuming warm-up for policy networks. Generally, we propose a novel Curiosity-driven Reinforcement Learning (CRL) framework to jointly enhance the diversity and accuracy of the generated paragraphs. First, by modeling the paragraph captioning as a long-term decision-making process and measuring the prediction uncertainty of state transitions as intrinsic rewards, the model is incentivized to memorize precise but rarely spotted descriptions to context, rather than being biased towards frequent fragments and generic patterns. Second, since the extrinsic reward from evaluation is only available until the complete paragraph is generated, we estimate its expected value at each time step with temporal-difference learning, by considering the correlations between successive actions. Then the estimated extrinsic rewards are complemented by dense intrinsic rewards produced from the derived curiosity module, in order to encourage the policy to fully explore action space and find a global optimum. Third, discounted imitation learning is integrated for learning from human demonstrations, without separately performing the time-consuming warm-up in advance. Extensive experiments conducted on the Standford image-paragraph dataset demonstrate the effectiveness and efficiency of the proposed method, improving the performance by 38.4% compared with state-of-the-art.

CCS CONCEPTS

• **Computing methodologies → Natural language generation**; **Scene understanding**;

KEYWORDS

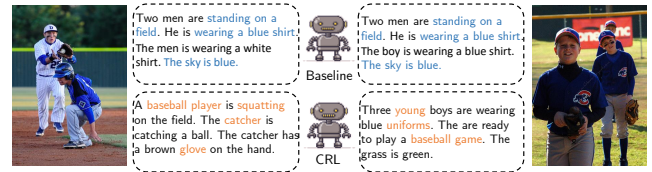Reinforcement Learning; Visual Paragraph Generation

Figure 1: An illustration of the mode collapse issue in paragraph captioning. Baseline model describes two distinct images with generic descriptions and similar patterns (as highlighted in blue), whereas the proposed CRL method generates diverse and more specific captions to context (as highlighted in orange).

1 INTRODUCTION

With a rapid growth of multimedia data [23, 30], understanding the visual content and interpreting it in natural language have been important yet challenging tasks, which could benefit a wide range of real-world applications, such as story telling [16, 36, 45], poetry creation [26, 27, 50, 51] and support of the disabled. While deep learning techniques have made remarkable progress in describing visual content via image captioning [10, 25, 39, 55], the obtained results are generally sentence-level, with fewer than twenty words. Generating such a short description could hardly convey adequate messages for all subtle objects and relationships, to say nothing of further pursuing coherence and diversity. Consequently, it is more natural to depict images in paragraph format [7, 21, 22, 24, 53], which has been investigated recently.

Generally, existing mainstream paragraph captioning models [21] follow the encoder-decoder architecture, where the language decoder is fully supervised to maximize the posterior probability of predicting each word, given the previous ground-truth sequence and image representations extracted by the visual encoder. The word-level cross-entropy objective, in this way, will encourage the use of exactly the same *n*-grams that appear in ground-truth samples, making the paragraph captions lack completeness and

consistency. Motivated to achieve more diverse and natural descriptions, an emerging line of work [7, 10, 24] combines supervised learning with generative adversarial models [10, 24] or autoencoders [7], aiming to capture inherent ambiguity of captions with a low-dimensional Gaussian manifold and model the structure of paragraphs in a hierarchical way. Nevertheless, existing paragraph captioning approaches are far from optimal due to two major issues, *i.e.,* the *mode collapse* and the *exposure bias*. First, the simple Gaussianity assumption is not sufficient enough to fully preserve the ground-truth distribution, many modes of which are underrepresented or missing. For instance, given two distinct pictures of baseball games as Figure 1 shows, the generated paragraphs by the baseline model only describe objects and actions with vague and general words (e.g., "two men", "standing") rather than specific noun entities (e.g., "baseball player", "catcher") or vivid verbs (e.g., "squatting"). Second, the language decoder predicts based on different inputs during training and testing, i.e., the ground-truth sub-sequences during training yet its own predictions during testing. This discrepancy as exposure bias severely hurts the model performance.

Recently, another line of work tackles the exposure bias and takes advantage of non-differential evaluation feedback by applying reinforcement learning, especially the REINFORCE [48] algorithm for the sentence-level captioning task [8, 25, 28, 39, 55]. This strategy reformulates the image captioning as the sequential decision-making process, where the language policy based on its previous decisions is directly optimized. Besides, it is more reasonable to achieve long-term vision and less greedy behaviors by optimizing the sentence-level evaluation metrics such as BLEU [32], METEOR [11] and CIDEr [43] instead of the cross entropy loss.

Unfortunately, it is challenging to extend the success to the paragraph captioning task: (1) **Mode Collapse:** optimizing evaluation metrics still does not alleviate the fixed-pattern issue, where the strategy could be easily tricked by repetition of frequent phrases and generic representations, yielding less variety of expressions; (2) **Delayed Feedback:** the language policy only receives feedback from the evaluator when the entire sequence is produced, resulting in high training variance especially with long sequence data like paragraphs; (3) **Time-consuming Warm-up:** current reinforcement learning suffers from low sample efficiency, which causes an unbearable time and computational cost for trial and error. Therefore, it usually requires long-term supervised pre-training for policy warm-up.

To address above-mentioned issues, in this paper, we propose a novel Curiosity-driven Reinforcement Learning (CRL) framework for diverse visual paragraph generation. Firstly, we design an intrinsic reward function (the curiosity module) that encourages the policy to explore uncertain behaviors and visit unfamiliar states, thereby increasing the expression variety rather than pursuing local phrase matching. In particular, the curiosity module consists of two sub-networks on top of language policy networks with self-supervision, *i.e.,* the Action Prediction Network (AP-Net) and the State Prediction Network (SP-Net). Moreover, the SP-Net measures state prediction error at each time steps as dense intrinsic rewards, which are complementary to the delayed extrinsic reward. Different from conventional reinforcement learning [25, 39] that simply averages the extrinsic reward to each state, we further adopt

the temporal difference learning [41] to correct the extrinsic reward estimation, considering the correlations of successive actions. Lastly, to avoid time-consuming warm-up for policy networks, our algorithm seamlessly integrates discounted imitation learning to stabilize learning for fast convergence, and gradually weakens supervision signal without shrinking action space.

Overall, our contributions can be briefly summarized as follows:

- To our best knowledge, this is the first attempt to tackle the visual paragraph generation problem with pure reinforcement learning. Different from the conventional REINFORCE algorithm, our CRL learning motivates visual reasoning and language decoding both intrinsically and extrinsically.
- The intrinsic curiosity complements sparse and delayed extrinsic rewards with prediction error measurement, guiding the agent to fully explore and achieve a better policy.
- Instead of pre-training policy networks with supervised learning, we jointly stabilize reinforcement learning with discounted imitation learning for fast convergence.
- We show the effectiveness of the proposed strategy through extensive experiments on the Standford paragraph captioning benchmark and demonstrate the diversity of generated paragraphs with a visualization of semantic network graphs.

2 RELATED WORK

2.1 Sentence-level Captioning with Reinforcement Learning

Inspired by the recent advances in reinforcement learning, several attempts have been made to apply policy gradient algorithms to image captioning task [4, 52, 56], which could generally be categorized into two groups: policy based and actor-critic based. Policy based methods (e.g., DISC [10], SCST [39], PG-SPIDEr [28], CAVP [25], TD [8]) utilize the unbiased REINFORCE [48] algorithm which optimizes the gradient of the expected reward by sampling a complete sequence from the model during training. To suppress high variance of Monte-Carlo sampling, Self-critical Sequential Training (SCST) [39] utilizes a baseline subtracted from the return which is added to reduce the variance of gradient estimation. Rather than obtaining a single reward at the end of sampling, actor-critic based algorithms (e.g., Embedded Reward [38], Actor-Critic [55], Adapt [9], HAL [46]) learn both a policy and a state-value function ("critic"), which is used for bootstrapping, *i.e.,* updating a state from subsequent estimation, to reduce variance and accelerate learning [41]. Different from existing work, the proposed CRL algorithm learns about a critic from the inner environment, complementing the extrinsic reward from the perspective of agent learning.

2.2 Paragraph-level Captioning

While sentence-level captioning has been extensively studied, the problem of generating paragraph-level descriptions still remains under-explored. Existing solutions include (1) generating sentences individually with detected region proposals (DenseCap [17]), or with topic learning via the Latent Dirichlet Allocation (TOMS [31]); (2) preserving semantic content and linguistic order with hierarchical structure (Region-Hierarchical [21]). To further encourage the coherence and naturalness among successive sentences, this model was further extended by Liang et al. [24] and Dai et al. [10]
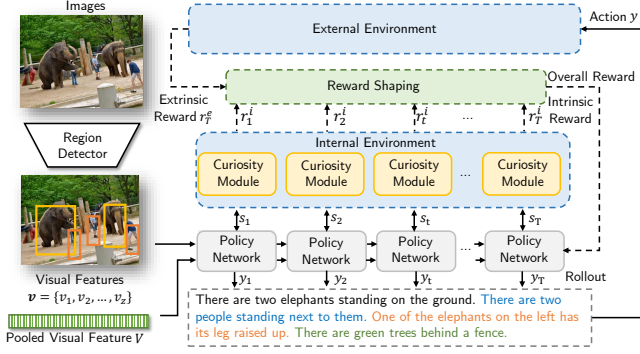
Figure 2: The general flowchart of the proposed image paragraph captioning model.

by adopting adversarial learning. To tackle the training difficulties of Generative Adversarial Networks (GANs), Chatterjee et al. [7] modeled the inherent ambiguity of paragraphs via a variational auto-encoder formulation. From another perspective, Wang et al. [47] leveraged depth estimation to discriminate objects at various depths and capture subtle interactions.

2.3 Intrinsically Motivated Reinforcement Learning

In reinforcement learning area, much theoretical work has been done on improving agent exploration and shaping sparse rewards via intrinsic motivation [40], where an information-theoretic critic measures the agent's surprisal [1, 5, 6, 34] (based on prediction error) or state novelty [3, 15, 29, 42] (based on counts of visited states), motivating the agent from the inner environment. Our proposed algorithm shares the same spirit with the former group. But instead of testing on simulated games, we, *for the first time*, adapt the intrinsic reward and validate its effectiveness and efficiency on a more practical task, *i.e.,* paragraph captioning.

3 CURIOSITY-DRIVEN LEARNING

The overview of the proposed paragraph captioning framework is illustrated in Figure 2. We firstly formulate the task of visual paragraph generation, followed by the introduction of the language policy network and policy learning. To enhance the agent exploration, two sub-networks of the curiosity module are trained in a self-supervised manner: the state prediction network (SP-Net) and the action prediction network (AP-Net). Then, we explain the detailed calculation for rewards with temporal-different learning, discounted imitation learning and collaborative optimization for all objectives.

3.1 Problem Formulation

Our target is to generate the paragraph caption $y = \{y_1, y_2, \ldots, y_T\} \in \{1, 0\}^{T \times D}$ for any given image I, where T denotes the length of the generated caption and D is the vocabulary size. The proposed framework follows the general encoder-decoder structure (see Figure 2), where E-dimensional visual features for m local regions $v = \{v_1, v_2, \ldots, v_m\} \in \mathbb{R}^{m \times E}$ are extracted by the Faster RCNN [37] at

the encoding stage. Language sequences are decoded by Recurrent Neural Networks (RNN) step by step. Different from traditional settings that directly force the language decoder to mimic ground truths by using cross entropy loss, our work casts the problem in reinforcement learning in order to optimize non-differential evaluation metrics (*e.g.,* CIDEr) and suppress the exposure bias. Playing as the "policy" π_θ in a finite Markov decision process (MDP), the language decoder model parameterized by θ predicts the next word y_t ("action" a_t) at time t based on the hidden "state" s_t. To suppress the sparsity and delay issues of the extrinsic reward r^e, the derived curiosity critic predicts the expected "intrinsic reward" r_t^i at each time step, which greatly complements and shapes the final reward.

3.2 Visual-Language Policy Network

To adaptively control visual signals and generate context-aware descriptions, we adopt a double-layer LSTM structure coupled with attention mechanism as the policy networks (see Figure 2). The first LSTM layer serves as a top-down visual attention model, taking the input word y_{t-1}, concatenated with the mean-pooled image features V and previous Z-dimensional state of the language LSTM s_{t-1}^{lang} at each time step. Therefore, the state transition for the visual LSTM is,

$$s_t^{vis} = \text{LSTM}([s_{t-1}^{lang}, W_v^T V, W_e^T y_{t-1}], s_{t-1}^{vis}), \tag{1}$$

where W_e, W_v are the learnable weights and $w_t \in \{0, 1\}^D$ is one-hot embedding of the input word at the time step t. To further attend local visual features v based on language policy, the weighted visual features $\hat{v}_t \in \mathbb{R}^E$ can be calculated as,

$$\hat{v}_t = \sum_{i=1}^{m} \text{softmax}(W_\alpha^T \tanh(W_v v_i + W_h s_t^{vis})) v_i, \tag{2}$$

where W_v, W_h, and W_α are the weights to be learned. Obtaining weighted visual features and the hidden state s_t^{vis} from the attention LSTM, the top-level language LSTM gives the conditional distribution $\pi_\theta(y_t | \cdot)$ for the next word y_t prediction, *i.e.,*

$$\begin{aligned} \pi_\theta(y_t | s_t) &= \text{softmax}(W_p s_t^{lang}), \\ s_t^{lang} &= \text{LSTM}([\hat{v}_t, s_t^{vis}], s_{t-1}^{lang}), \end{aligned} \tag{3}$$

where W_p is the learnable weight. To encourage the agent to explore rare attended areas and words, we concatenate the state $s_t = [s_t^{vis}, s_t^{lang}]$, which will be used in policy learning and discounted imitation learning. To form the complete paragraph, the distribution can be calculated as the dot product of conditional distributions from previous steps,

$$\pi_\theta(y_{1:T}) = \prod_{t=1}^{T} \pi_\theta(y_t | s_t). \tag{4}$$

3.3 Policy Learning

In reinforcement learning, the policy network π_θ leverages the experiences obtained from interacting with an environment to learn behaviors that maximize a reward signal. Generally, the RL loss can be presented as,

$$\min_\theta \mathcal{L}_{RL} = -\mathbb{E}_{y \sim \pi_\theta}[A^{\pi_\theta}(s, y)], \tag{5}$$

where $A^{\pi_\theta}(s,y) = Q^{\pi_\theta}(s,y) + V^{\pi_\theta}(s)$ is the advantage function. $Q^{\pi_\theta}(s,y)$ stands for the state-action function estimating the long-term value instead of the instantaneous reward, and $V^{\pi_\theta}(s)$ indicates state value function, which serves as the inner critic. The core idea is to incentivize the policy to increase the probability of actions that are correct and rarely-seen. $s = \{s_1, s_2, \ldots, s_T\}$ denotes the concatenated hidden states of the policy network. Based on policy gradient [41], the gradient of non-differentiable reward-based loss function can be derived as,

$$
\begin{aligned}
\nabla_\theta \mathcal{L}_{RL} &= -\mathbb{E}_{y \sim \pi_\theta}[A^{\pi_\theta}(s,y)\nabla_\theta \log \pi_\theta(y|s)] \\
&= -\sum_{t=1}^{T} A^{\pi_\theta}(s_t, y_t)\nabla_\theta \log \pi_\theta(y_t|s_t).
\end{aligned}
\tag{6}
$$

3.4 Self-supervised State Prediction (SP-Net)

Before estimating the advantage function, we first detail two sub-networks for the state value function $V^{\pi_\theta}(s)$. The SP-Net is trained to predict the future state embedding $\phi(s_{t+1})$ based on the input action y_t, and $\phi(s_t)$, where $\phi(\cdot)$ indicates the state embedding layer and helps filter irrelevant memory for the prediction of next state. The mean-squared error is used as the objective function for SP-Net,

$$
\min_{\theta, \theta_{SP}, \theta_\phi} \mathcal{L}_{SP} = \frac{1}{2}\|\tilde{\phi}(s_{t+1}) - \phi(s_{t+1})\|_2^2
\tag{7}
$$
$$
\tilde{\phi}(s_{t+1}) = \mathcal{G}(\phi(s_t), y_t; \theta_{SP}),
$$

where $\mathcal{G}(\cdot; \theta_{SP})$ denotes the nonlinear transformation of SP-Net parameterized by θ_{SP}. In this way, the state value function can be obtained as,

$$
V^{\pi_\theta}(s) = \frac{\rho}{2}\sum_{t=1}^{T}\|\tilde{\phi}(s_t) - \phi(s_t)\|_2^2,
\tag{8}
$$

where ρ is the hyper-parameter. The prediction error quantifies the agent uncertainty towards the environment. The policy network trained to maximize state prediction error will explore transitions with less experience and high confusion, therefore rare attended areas and infrequent expressions can be well captured.

3.5 Self-supervised Action Prediction (AP-Net)

Given the transition tuple (s_t, s_{t+1}, y_t), the action prediction network targets at predicting the action y_t based on state transition. The objective of AP-Net can be defined as,

$$
\min_{\theta, \theta_{AP}, \theta_\phi} \mathcal{L}_{AP} = -\sum_{t=1}^{T} q(y_t)\log(\tilde{y}_t)
\tag{9}
$$
$$
\tilde{y}_t = \mathcal{F}(\phi(s_t), \phi(s_{t+1}); \theta_{AP}),
$$

where $\tilde{y}_t$ is the prediction of current action, shown as a softmax distribution among all possible words. $q(y_t)$ is the real distribution of action y_t and $\mathcal{F}(\cdot)$ denotes the nonlinear transformation of AP-Net parameterized by θ_{AP}. The intuition of AP-Net is to learn state embedding that corresponds to meaningful patterns of human writing behaviors, suppressing the impact of outliers.
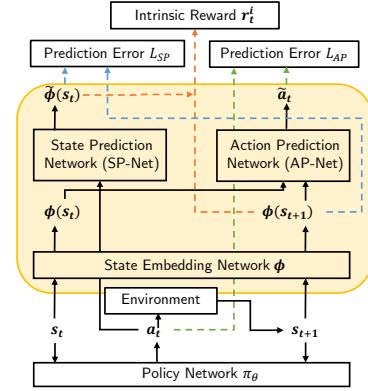

Figure 3: An illustration of the curiosity module. The blue and the green dotted line denote the loss calculation for $\mathcal{L}_{SP}$ and $\mathcal{L}_{AP}$, respectively. The orange line shows the calculation for the intrinsic reward r_t^i.

3.6 Reward Shaping

To encourage an agent to explore its environment for acquiring new knowledge and guide it to generate accurate and diverse paragraphs, the overall reward is generated by two parts, dense intrinsic curiosity reward r^i and sparse extrinsic reward r^e. The policy network is expected to maximize the weighted sum of two rewards.

3.6.1 Extrinsic Reward. To improve the fidelity and interpretability of the learned paragraph, the extrinsic reward r_t^e is refined as the linear combination of linguistic measures. Specifically, we select the most representative and commonly used metrics, *i.e.,* BLEU-4 [32] and CIDEr [43],

$$
r_t^e = \begin{cases} a \cdot \text{BLEU-4}(y_{1:T}) + b \cdot \text{CIDEr}(y_{1:T}), & \text{if } t = T \\ 0, & \text{otherwise.} \end{cases}
\tag{10}
$$

In our case, the hyper-parameters a and b are empirically set to 1 and 2, respectively. Under such a reward setting, we adopt the temporal-difference learning TD(λ) [41] to estimate the action-state function $Q^{\pi_\theta}(s_t, y_t)$ for each time step,

$$
G_{t:t+j} = r_t^e + \gamma r_{t+1}^e + \gamma^2 r_{t+2}^e + \ldots + \gamma^j r_{t+j}^e
$$
$$
Q^{\pi_\theta}(s_t, y_t) = (1-\lambda)\sum_{j=0}^{T-t}\gamma^j G_{t:t+j} + \lambda^{T-t}G_t,
\tag{11}
$$

where the j-step expected return $G_{t:t+j}$ is defined as the sum of expected future rewards from the next j steps. The λ indicates the trade-off parameter between the future estimation and the current estimation. The discounted factor γ enables variance reduce by down-weighting extrinsic rewards. For simplicity, we set λ to 1, and the overall Q function can be formulated as,

$$
Q^{\pi_\theta}(s,y) = \sum_{t=1}^{T}\sum_{j=0}^{T-t}\gamma^j r_{t+j}^e = \sum_{t=1}^{T}\gamma^{T-t}r_T^e,
\tag{12}
$$

3.6.2 Intrinsic Reward. As discussed in Section 3.4, we train a state prediction network and calculate the prediction error as intrinsic reward $r^i = V^{\pi_\theta}(s)$ (see Equation (8)). Therefore, the

gradient of policy network can be rewritten as,

$$\nabla_\theta \mathcal{L}_{RL} = -\sum_{t=1}^{T}(Q^{\pi_\theta}(s,y) + V^{\pi_\theta}(s))\nabla_\theta \log \pi_\theta(y_t|s_t) \quad (13)$$

$$= -\sum_{t=1}^{T}(\gamma^{T-t}r_T^e + \frac{\rho}{2}\|\tilde{\phi}(s_t) - \phi(s_t)\|_2^2)\nabla_\theta \log \pi_\theta(y_t|s_t).$$

3.7 Discounted Imitation Learning

A major challenge of the reinforced agent to have a good convergence property is that the agent must start with a good policy at the beginning stage. The low sample efficiency issue [54] causes a huge amount of time and computational cost for trial and error. Existing sentence-level captioning methods [25, 28, 39, 55] with reinforcement learning apply the cross entropy loss to the language decoder for warm up, which is defined as,

$$\min_\theta \mathcal{L}_{XE} = -\sum_{t=1}^{T} \log \pi_\theta(y_t^*|y_{t-1}^*,\ldots,y_1^*,V), \quad (14)$$

where $y^* = \{y_1^*, y_2^*, \ldots, y_T^*\}$ is the human-labeled ground-truth. Despite that supervised learning is essential to initialize the policy network, it usually consumes a long period of time (e.g., 40 epochs on Standford dataset), and highly restricts the agent search space, which probably leads to a local minimum. Therefore, we introduce the discounted imitation learning at the first epoch of training, then gradually decrease the loss coefficient η to weaken the supervision.

3.8 Collaborative Optimization

To collaboratively optimize the objectives of reinforcement learning, the curiosity modules and discounted imitation learning, the overall learning loss function can be formulated as,

$$\min_{\theta, \theta_{AP}, \theta_{SP}, \theta_\phi} \mathcal{L}_{RL} + \alpha\mathcal{L}_{AP} + \beta\mathcal{L}_{SP} + \eta\mathcal{L}_{XE}, \quad (15)$$

where α and β are constant loss coefficients, η is a dynamic scaling factor that gradually reduces with ten percent decay every epoch. Notably, we dynamically estimate the intrinsic reward of agent behavior to shape reward signal, which avoids additional baseline calculation in the advantage function. The overall algorithm is shown in the Algorithm 1.

4 EXPERIMENTS

4.1 Settings

4.1.1 Dataset. All state-of-the-art methods and our proposed method are evaluated on the *Stanford image-paragraph* dataset [21], where 14,579 image-paragraph pairs from the Visual Genome and MS COCO dataset are used for training, 2,490 for validation and 2,492 for testing. The number of unique words in its vocabulary is 12,186. All images are annotated with human-labeled paragraphs of 67.5 words on average.

4.1.2 Evaluation Metrics. We report the performance of all models on six widely used automatic evaluation metrics, i.e., **BLEU-{1,2,3,4}** [32], **METEOR** [11] and **CIDEr** [43]. BLEU-n is defined as the geometric mean of n-gram precision scores and CIDEr measures n-gram accuracy by term-frequency inverse-document-frequency (TF-IDF). METEOR is defined as the harmonic mean of precision and

---

Algorithm 1 Pseudo-code of the Proposed CRL Learning.

1: **Inputs:**
    Annotated image-paragraph set
    $\mathcal{D} = \{I_{(n)}, y_{(n)}^*\}_{n=1}^N$;
2: **Outputs:**
    Visual-language policy network π_θ;
3: **Initialize:**
    Hyper-parameters: $\rho, \gamma, \alpha, \beta, \eta, a, b$;
    Visual features v from m regions;
    Minibatch size m and learning rate μ;
4: **for** k epochs **do**
5:     **for** T time steps **do**
6:         Sample action a_t based on policy π_θ
7:     **end for**
8:     Calculate intrinsic rewards $\{r_t^i\}_{t=1}^T$ and state-value function $V^{\pi_\theta}(s)$ in Equation (8);
9:     Calculate extrinsic rewards $\{r_t^e\}_{t=1}^T$ and action-state function $Q^{\pi_\theta}(s,y)$ in Equation (12);
10:     Update parameters $\theta, \theta_{SP}, \theta_{AP}$ and θ_ϕ by descending stochastic gradients:
11:     $\theta \leftarrow \theta - \mu \cdot \nabla_\theta \frac{1}{m}(\mathcal{L}_{RL} + \alpha\mathcal{L}_{AP} + \beta\mathcal{L}_{SP} + \eta\mathcal{L}_{XE})$;
12:     $\theta_{AP} \leftarrow \theta_{AP} - \mu \cdot \nabla_{\theta_{AP}} \frac{1}{m}\mathcal{L}_{AP}$;
13:     $\theta_{SP} \leftarrow \theta_{SP} - \mu \cdot \nabla_{\theta_{SP}} \frac{1}{m}\mathcal{L}_{RL}$;
14:     $\theta_\phi \leftarrow \theta_\phi - \mu \cdot \nabla_{\theta_\phi} \frac{1}{m}(\alpha\mathcal{L}_{AP} + \beta\mathcal{L}_{SP})$;
15:     update the dynamic factor $\eta \leftarrow \delta\eta$;
16: **end for**

---

recall of exact, stem, synonym, and paraphrase matches between paragraphs.

4.2 Baselines

We compare our approach with several state-of-the-art paragraph captioning methods and one RL-based method.

Sentence-Concat: Two sentence-level captioning models (**Neuraltalk** [18] and **NIC** [44]) pre-trained on the MS COCO dataset are adopted to predict five sentences for each given image, which are further concatenated into a paragraph.

Image-Flat: Different from sentence-concat group, **Image-Flat** [44] method directly generates a paragraph word by word, with the ResNet-152 network [13] for visual encoding and a single LSTM layer to decode language. **DAM-Att** [47] couples the encoder-decoder architecture with attention mechanism, and additionally introduces depth information to enhance recognition to spatial object-object relationships. **TOMS** [31] learns topic-transition among multiple sentences with Latent Dirichlet Allocation (LDA).

Hierarchical: **Region-Hierarchical** [21] leverages a hierarchical recurrent network to learn sentence topic transition and decode language sentence by sentence. **RTT-GAN** [24] implements the hierarchical learning in a GAN [12] setting, where the generator mimics the human-annotated paragraphs and tries to fool the discriminator. **VAE** [7] models the paragraph distribution with variational auto-encoder [20], which preserves the coherence and global topics of paragraphs. Notably, Liang et al. took advantage of the local phrases that are predicted by the dense-captioning model [17], which additionally used training data from the MS-COCO dataset.

Table 1: Performance comparisons using BLEU-{1,2,3,4}, METEOR and CIDEr on Standford Image-paragraph dataset. The human performance is provided for reference. In our setting, the proposed CRL only optimizes CIDEr and BLEU-4 and achieves the highest scores compared with state-of-the-art.

Methods	Language Decoder	Beam Search	METEOR	CIDEr	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Sentence-Concat (Neuraltalk[18])	1*LSTM	2-beam	12.05	6.82	31.11	15.10	7.56	3.98
Sentence-Concat (NIC[44])	1*LSTM	2-beam	9.27	7.09	22.31	10.72	4.91	2.32
Image-Flat (NIC[44])	1*LSTM	2-beam	13.44	14.71	34.80	19.42	10.91	6.03
DAM-Att ([47])	2*LSTM	Greedy	13.91	17.32	35.02	20.24	11.68	6.57
TOMS ([31])	1*LSTM	3-beam	18.60	20.80	43.10	25.80	14.30	8.40
Region-Hierarchical ([21])	2*LSTM	2-beam	13.85	10.64	35.58	17.94	9.08	4.49
RTT-GAN ([24])	2*LSTM	2-beam	17.12	16.87	41.99	24.86	14.89	9.03
VAE ([7])	2*GRU	Greedy	**18.62**	20.93	42.38	25.52	15.15	9.43
SCST ([39])	1*LSTM	Greedy	16.01	22.74	40.89	23.71	14.43	8.38
CRL	1*LSTM	Greedy	17.71	25.03	43.10	26.93	16.65	9.91
CRL	1*LSTM	2-beam	17.42	**31.47**	**43.12**	**27.03**	**16.72**	**9.95**
Humans (as in [21])	-	-	19.22	28.55	42.88	25.68	15.55	9.66

REINFORCE: For fair comparison, we compare the proposed framework with the RL-based image captioning method **SCST** [39]. The model shares the same backbone encoder-decoder structure but a different reinforcement learning strategy and reward functions. As it requires supervised warm-up for policy networks, we pre-train the SCST with the cross-entropy (XE) objective using ADAM optimizer [19] with the learning rate of 5×10^{-4}.

4.3 Implementation Details

Our source code is based on PyTorch [33] and all experiments are conducted on a server with two GeForce GTX 1080 Ti GPUs.

4.3.1 Data Pre-processing. For textual pre-processing, we first tokenize all annotated paragraphs, and replace words that appear less than five times with the unknown <unk> token for the vocabulary. For **ResNet Features** extraction, we encode each image with Resnet-101 [13] with a 2048-D vector, while we select top $m = 50$ salient regions for **Region Features** with Faster R-CNN [37].

4.3.2 Module Architecture. The AP-Net maps input state $s_t \in \mathbb{R}^{2 \times 512}$ into a state embedding $\phi(s_t) \in \mathbb{R}^{512}$ with one fully connected layer and one LeakyReLu layer. The SP-Net takes $\phi(s_t)$ and 512-D embedding for y_t as input, then passes it into a sequence of two fully connected layers with 512 units and 12,186 units.

4.3.3 Parameter Settings. The hidden size, all embedding size for images and words are fixed to 512. The batch size for non-attention based models is 32, but 16 for attention-based model. The learning rate μ is initiated as 6×10^{-4} then decayed by a factor of 0.8 every three epochs. The discounted factor δ for discounted imitation learning is set to 0.9. The hyper-parameter ρ and discounted coefficient γ are set to 1 and 0.9, respectively. The loss coefficients α and β are fixed at 0.2 and 0.8. For compared models, the embedding size of topic vector is set to 100.

4.4 Comparisons with State-of-The-Art

4.4.1 Quantitative Analysis. In this section, we quantitatively evaluate various paragraph captioning methods using the standard metrics on the Standford image-paragraph dataset. Here we report the best performance for every model, along with the specification of the language model and the search method at inference stage.

Greedy denotes greedy search (equals to 1-beam search) and *n*-**beam** indicates the beam search with *n* most probable sub-sequence [35]. Generally, more beams used for inference will lead to better performance but higher time-cost. From Table 1, we can observe that our **CRL** is superior to all the compared paragraph-based and sentence-based image captioning methods in most cases, especially improving CIDEr [43] by 38.4% (from 22.74% to 31.47%). With only a single layer language decoder, we achieve a significant performance boost over hierarchical methods. Since we select metrics to optimize paragraph-level quality (e.g., CIDEr), the proposed CRL achieves relatively lower performance on the uni-gram metric with synonymous substitution (e.g., METEOR). Regarding observations on compared methods, Non-hierarchical methods (e.g., **Image-Flat** [44], **DAM-Att** [47] and **TOMS** [31]) perform much better than simple concatenation of sentence-level outputs (e.g., **Neuraltalk** [18], **NIC** [44]), yet they fail to capture the overall structure and topic transition of paragraphs, thus obtaining a lower performance than hierarchical approaches (e.g., **Region-Hierarchical** [21], **RTT-GAN** [24] and **VAE** [7]). Different from 'Region-Hierarchical' that simply concatenates sentences from the bottom LSTM, 'RTT-GAN' and 'VAE' preserve a better consistency among sentences. The RL-based method, i.e., **SCST** [39] with single-layer language decoder achieves competitive outcomes compared with the hierarchical model [21], which demonstrates the power of policy optimization. **Humans**, as reported in [21], show the results by collecting additional paragraphs for 500 randomly chosen images. We can see that the results show a large gap between automatic synthetic captions and natural language, whereas our proposed CRL with 2-beam search mitigates the gap and achieves competitive outcomes. Besides, experimental results verify that CIDEr metric align better with human judgment than any other evaluation metrics.

4.4.2 Qualitative Analysis. In order to intuitively understand the performance of the proposed CRL training, we showcase some outputs with greedy search for randomly selected images in Figure 5, *i.e.,* the paragraphs generated by canonical paragraph captioning method 'Region-Hierarchical' [21], the proposed CRL method and the RL-based method 'SCST' [39]. With comparisons with counterparts, our proposed CRL model generates the paragraph in a coherent order: the first sentence (in red) tends to cover a global

Table 2: Ablative performance comparisons on Standford image-paragraph dataset. "w/o" indicates without. The best performances are shown in boldface.

| Methods	Policy	ResNet Features						Region Features					
		METEOR	CIDEr	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	BLEU-1	BLEU-2	BLEU-3	BLEU-4
CRL w/o RL	FC	13.10	11.18	36.19	17.67	8.41	3.93	13.33	11.83	37.01	18.32	8.77	4.35
	Att	13.77	12.34	37.40	20.94	11.48	6.10	13.35	12.23	36.50	19.11	9.28	4.39
	Up-Down	14.02	11.46	37.68	19.17	9.34	4.34	14.28	14.10	38.07	20.42	10.57	5.26
CRL w/o intrinsic	FC	15.12	18.31	39.38	21.81	11.84	6.23	15.67	20.14	40.98	23.04	13.44	7.58
	Att	15.89	19.07	41.32	24.72	14.04	7.96	15.92	19.12	40.31	24.82	14.36	8.27
	Up-Down	15.91	20.45	41.41	24.77	14.40	8.14	16.01	22.74	40.89	23.71	14.43	8.38
CRL	FC	15.53	20.31	39.68	22.66	12.69	6.89	15.87	21.13	40.98	24.30	14.12	7.96
	Att	16.13	19.67	41.17	24.18	14.92	8.92	16.11	19.21	41.17	25.00	15.03	8.82
	Up-Down	**16.71**	**24.99**	**41.88**	**25.24**	**15.25**	**9.03**	**17.71**	**25.03**	**43.10**	**26.93**	**16.65**	**9.91**
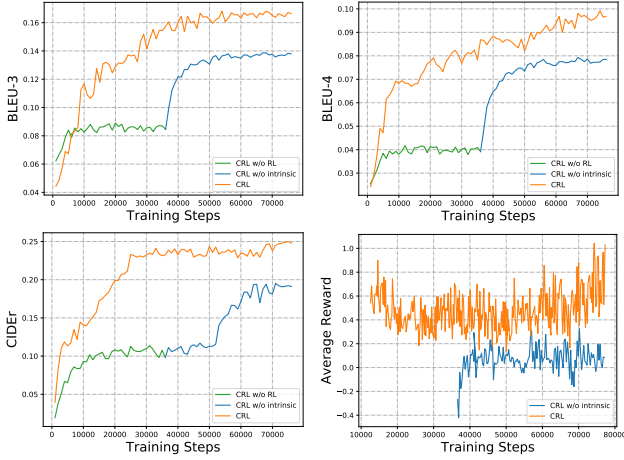
Figure 4: The curves of language measures on validation set and average rewards on Standford image-paragraph dataset.

topic or major actions in visual content, followed by several sentences (in blue) to describe the details of the scene. Generally, the last sentence gives descriptions about objects or environment in the background, which exactly matches human writing styles. Notably, our synthetic paragraphs capture more subtle and accurate words and relationships, such as 'platform' and 'standing behind the man'. In contrast, both 'Region-Hierarchical' and 'SCST' could barely guarantee the completeness and richness of the generated paragraphs.

4.5 Ablation Study

In this section, we study the impact of the RL strategy, the policy network architecture and visual features, respectively. The major experimental results are shown in Table 2 and the detailed curve of evaluation metrics and average rewards are illustrated in Figure 4.

4.5.1 RL Training Strategy. By comparing the performance of each policy used in different methods shown in Table 2, we can observe that pure supervised learning (**CRL w/o RL**) drops greatly on aggregative metrics like CIDEr and multi-gram metrics like BLEU-{3,4}. Removing the intrinsic reward, **CRL w/o intrinsic** tightly follows the learned policy and lacks essential exploration, thus leading to the sub-optimal performance. Moreover, in Figure 4, we show detailed curves of BLEU-{3, 4}, CIDEr per training step and

curves of average rewards on validation set based on the 'Up-Down' decoder and the 'Region Features'. From Figure 4, we could observe that 'CRL w/o intrinsic' (as the blue line shows) needs a long-period warm-up by 'CRL w/o RL' (as the green line shows) and sharply boosts evaluation metrics of the evaluation score after 40 epochs, and gradually converges afterwards. While our strategy **CRL** (as the orange line shows) gains a smooth increase during training, since it benefits from the power of combining discounted imitation learning and policy gradient training. Different from 'CRL w/o intrinsic', our CRL method avoids tedious and time-consuming initialization, and obtains a full exploration and a better policy network. With respect to the average reward curve, the extrinsic reward achieved by 'CRL w/o intrinsic' climbs fast after pre-training, yet our reward signal moves a downhill then slowing uphill. This phenomenon is probably caused by the intrinsic reward, that decreases at very first beginning as the agent learns to control state transition and linguistic patterns. Regarding to variance shown in Figure 4, we infer the variance is mainly introduced by iterative action optimization where the average reward fluctuates correspondingly. The variance can be alleviated by gradient clipping or adjusting learning rate. With the extrinsic rewards that gradually accumulate, the overall reward slowly increases until the model converges. Moreover, it is clearly observed that 'CRL' achieves faster convergence (around 25 epochs) compared with 'CRL w/o intrinsic' and 'CRL w/o RL'.

4.5.2 Policy Network Architecture. Regarding to the backbone language decoder, we switch the policy networks from **FC** (vanilla LSTM [14]), to **Att** (attention-based LSTM [49]), to **Up-Down** (attention-based LSTM + language LSTM [2]) for comparison. By comparing each method under different training policies, it is clear that the 'Up-Down' model achieves a higher performance among all metrics, as it dynamically attends on local areas of images and captures more visual details. In particular, the 'Up-Down' model trained with the proposed curiosity-driven RL averagely increase the CIDEr score by 16.36% and 97.89%, the BLEU-4 score by 14.59% and 98.23% compared with the 'CRL w/o instrinsic' and 'CRL w/o RL', respectively.

4.5.3 Visual Features. In addition to the evaluation considering the impact of visual features, we also list the performance based on **ResNet Features** and **Region Features** (see Section 4.3.1 for details). From the Table 2, we can draw the observation that 'Region Features' contribute positively to the captioning model, which enriches the visual recognition and representations. Compared
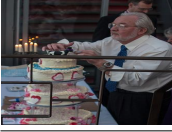
	Region - Hierarchical	SCST	Ours
	a picture of a woman in <unk> in there is ski tracks. a woman is wearing skis and standing on top of the snow in blue and are all around. a black pole is on the skis in blue and white clouds and there are poles on them. there is a black pole under the pole in black and is behind it on with look around. the pole is black and black in and white with	A snowboarder is skiing on a ski slope. The man is wearing a red and white shirt and black and to the snow. There is a small ski in Behind The man is wearing a black and a black and has in a The of them is in the snow. There are snow on the snow on the ground. There is a black on the	A woman is standing in on ski slope. She is wearing a red and white jacket. The woman is holding a black pole, on the snow. The woman is wearing a red and white. The skier is standing on the snow. The trees are covered in snow. There are trees in the snow.
	there are a couple of people standing around a table in blue and is outside. he is wearing a white shirt with a black scarf, and there are no other items on the fridge. he is almost happy face in blue and are behind him in each right.	A young men is sitting on blue table at a table. The baby is holding the table. The man is wearing a white shirt and has The boy is sitting on a chair on the table. The little is sitting on a black of the table. The woman is wearing a black and brown brown table There is a man sitting on of a table in front of the table.	A man is standing in front of a cake. He is wearing a white shirt. The man is wearing a blue tie. The man has a white cake. There is a person sitting on the table. There is a white cake. The cake is sitting on the table. The man is standing behind the man.
	a black and gold train car is parked on a train corner of a car in gray and there is a pole. the train cars are black and have white frames on them on and there are poles next to them. the front of the red stop sign a white shirt and it is silver in blue is. the train is white and red and there are two lights on the back of this car and there is a pole behind the	A silver and black train is on the road. There is a <unk> on the side of the train. The train is a green and There is a large in the side of the train. There is a green and yellow behind the There are black and white on the side of the train. The clock is on a red The front on the blue in	A model train is on the platform. The train is on the tracks. There is a blue and white train. A green and yellow train is besides. There are trees on the tracks. There is a black on the side of the train. There are lights above the tracks. There is a clear light on it.

Figure 5: Paragraphs generated for the images from the Standford image-paragraph dataset. It is observed that our generated paragraphs associated with logic and coherence, *i.e.,* starting with the global sentence marked in red, followed with the details in blue and the background description in green.
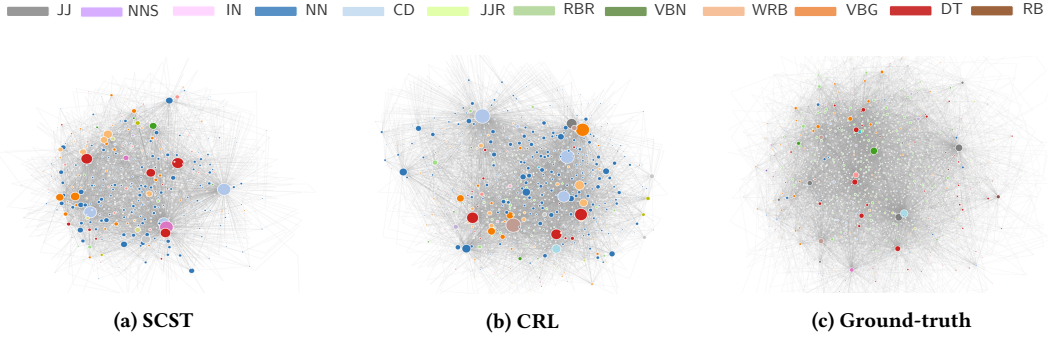

(a) SCST   (b) CRL   (c) Ground-truth

Figure 6: Visualization of diversity of the generated paragraphs by (a) RL-based method SCST [39] (b) our proposed CRL method (c) human-beings with semantic network graphs. Each node with different color indicates the unique token with various part-of-speech (POS) tags. Edges show the proximity relationship between tokens.

with other language networks, the 'Up-Down' architecture is more sensitive to the selection of visual features.

4.6 Diversity Analysis

To shed a quantitative light on the linguistic property of generated paragraphs, we randomly select 500 images from the test set of the Stanford image-paragraph dataset, and show the statistics of the paragraphs produced by a representative spread of methods in Figure 6. We visualize the semantic graphs of language distribution with a d3-force package in JavaScript. In each semantic graph, each node indicates a unique token from the vocabulary, with different colors to show the associated part-of-speech (POS) tagging. For instance, blue demonstrates Noun in singular form (NN), red shows Determiner (DT) and orange indicates Verb, Gerund or Present Participles (VGB). The edge between two nodes presents the proximity relationship of two words. It is worth noting that the massiveness of semantic graphs imply the diversity and richness of generated paragraphs in an intuitive way. The **Ground-truth** graph (in Figure 6(c)), annotated by human beings, contains the most comprehensive relationships and extensive object entities. Even though there is

still a gap between synthetic paragraphs and real natural language, our generated paragraphs by **CRL** (in Figure 6(b)) have a much wider vocabulary compared with the one generated by RL-based method **SCST** (in Figure 6(a)).

5 CONCLUSION

In this work, we propose an intrinsically motivated reinforcement learning model for visual paragraph generation. Towards generating diverse sentences with coherence, the proposed CRL mines the human writing patterns behind long narratives and well captures precise expressions by modeling the agent's uncertainty of the environment. Distinguishing our work from conventional policy-based and actor-critic based reinforcement learning methods, it alleviates the sparse reward and low exploration issues and thus encourages the agent to fully explore rare states and obtain a better policy.

6 ACKNOWLEDGEMENT

REFERENCES

[1] Joshua Achiam and Shankar Sastry. 2017. Surprise-Based Intrinsic Motivation for Deep Reinforcement Learning. *CoRR* abs/1703.01732 (2017). arXiv:1703.01732 http://arxiv.org/abs/1703.01732

[2] 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 *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. 6077–6086.

[3] Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Rémi Munos. 2016. Unifying Count-Based Exploration and Intrinsic Motivation. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. 1471–1479.

[4] Yi Bin, Yang Yang, Jie Zhou, Zi Huang, and Heng Tao Shen. 2017. Adaptively Attending to Visual Attributes and Linguistic Knowledge for Captioning. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*. 1345–1353. https://doi.org/10.1145/3123266.3123391

[5] Yuri Burda, Harrison Edwards, Deepak Pathak, Amos J. Storkey, Trevor Darrell, and Alexei A. Efros. 2018. Large-Scale Study of Curiosity-Driven Learning. *CoRR* abs/1808.04355 (2018). arXiv:1808.04355 http://arxiv.org/abs/1808.04355

[6] Yuri Burda, Harrison Edwards, Amos J. Storkey, and Oleg Klimov. 2018. Exploration by Random Network Distillation. *CoRR* abs/1810.12894 (2018). arXiv:1810.12894 http://arxiv.org/abs/1810.12894

[7] Moitreya Chatterjee and Alexander G. Schwing. 2018. Diverse and Coherent Paragraph Generation from Images. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II.* 747–763.

[8] Hui Chen, Guiguang Ding, Sicheng Zhao, and Jungong Han. 2018. Temporal-Difference Learning With Sampling Baseline for Image Captioning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. 6706–6713.

[9] 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 *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. 521–530.

[10] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. 2017. Towards Diverse and Natural Image Descriptions via a Conditional GAN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. 2989–2998.

[11] Michael J. Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation, WMT@ACL 2014, June 26-27, 2014, Baltimore, Maryland, USA*. 376–380.

[12] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. 2672–2680.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 770–778.

[14] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[15] Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. 2016. VIME: Variational Information Maximizing Exploration. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. 1109–1117.

[16] Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross B. Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual Storytelling. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*. 1233–1239.

[17] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. DenseCap: Fully Convolutional Localization Networks for Dense Captioning. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 4565–4574.

[18] Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. 3128–3137.

[19] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[20] Diederik P. Kingma and Max Welling. 2013. Auto-Encoding Variational Bayes. *CoRR* abs/1312.6114 (2013). arXiv:1312.6114 http://arxiv.org/abs/1312.6114

[21] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A Hierarchical Approach for Generating Descriptive Image Paragraphs. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. 3337–3345.

[22] Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. 2015. A Hierarchical Neural Autoencoder for Paragraphs and Documents. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers.* 1106–1115.

[23] Yang Li, Yadan Luo, Zheng Zhang, Shazia Sadiq, and Peng Cui. 2019. Context-Aware Attention-Based Data Augmentation for POI Recommendation. In *35th IEEE International Conference on Data Engineering Workshops, ICDE Workshops 2019, Macao, China, April 8-12, 2019*. 177–184.

[24] Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P. Xing. 2017. Recurrent Topic-Transition GAN for Visual Paragraph Generation. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. 3382–3391.

[25] Daqing Liu, Zheng-Jun Zha, Hanwang Zhang, Yongdong Zhang, and Feng Wu. 2018. Context-Aware Visual Policy Network for Sequence-Level Image Captioning. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*. 1416–1424.

[26] Lixin Liu, Xiaojun Wan, and Zongming Guo. 2018. Images2Poem: Generating Chinese Poetry from Image Streams. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*. 1967–1975.

[27] Lixin Liu, Xiaojun Wan, and Zongming Guo. 2018. Images2Poem: Generating Chinese Poetry from Image Streams. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*. 1967–1975.

[28] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. 2017. Improved Image Captioning via Policy Gradient optimization of SPIDEr. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. 873–881.

[29] Manuel Lopes, Tobias Lang, Marc Toussaint, and Pierre-Yves Oudeyer. 2012. Exploration in Model-based Reinforcement Learning by Empirically Estimating Learning Progress. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.* 206–214.

[30] Yadan Luo, Ziwei Wang, Zi Huang, Yang Yang, and Cong Zhao. 2018. Coarse-to-Fine Annotation Enrichment for Semantic Segmentation Learning. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*. 237–246.

[31] Yuzhao Mao, Chang Zhou, Xiaojie Wang, and Ruifan Li. 2018. Show and Tell More: Topic-Oriented Multi-Sentence Image Captioning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.* 4258–4264.

[32] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*. 311–318.

[33] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. (2017).

[34] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. 2017. Curiosity-driven Exploration by Self-supervised Prediction. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*. 2778–2787.

[35] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence Level Training with Recurrent Neural Networks. In *ICLR*. http://arxiv.org/abs/1511.06732

[36] Hareesh Ravi, Lezi Wang, Carlos Muñiz, Leonid Sigal, Dimitris N. Metaxas, and Mubbasir Kapadia. 2018. Show Me a Story: Towards Coherent Neural Story Illustration. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. 7613–7621.

[37] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*. 91–99.

[38] Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. 2017. Deep Reinforcement Learning-Based Image Captioning with Embedding Reward. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. 1151–1159.

[39] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2017. Self-Critical Sequence Training for Image Captioning. In *2017 IEEE*

Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. 1179–1195.

[40] Satinder P. Singh, Andrew G. Barto, and Nuttapong Chentanez. 2004. Intrinsically Motivated Reinforcement Learning. In *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*. 1281–1288.

[41] Richard S Sutton, Andrew G Barto, Francis Bach, et al. 2017. *Reinforcement learning: An introduction (2nd Edition)*. MIT press.

[42] Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. 2017. #Exploration: A Study of Count-Based Exploration for Deep Reinforcement Learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*. 2750–2759.

[43] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. 4566–4575.

[44] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. 3156–3164.

[45] Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. 2018. No Metrics Are Perfect: Adversarial Reward Learning for Visual Storytelling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*. 899–909.

[46] Xin Wang, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, and William Yang Wang. 2018. Video Captioning via Hierarchical Reinforcement Learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. 4213–4222.

[47] Ziwei Wang, Yadan Luo, Yang Li, Zi Huang, and Hongzhi Yin. 2018. Look Deeper See Richer: Depth-aware Image Paragraph Captioning. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*. 672–680.

[48] Ronald J. Williams. 1992. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Machine Learning* 8 (1992), 229–256.

[49] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*. 2048–2057.

[50] Linli Xu, Liang Jiang, Chuan Qin, Zhe Wang, and Dongfang Du. 2018. How Images Inspire Poems: Generating Classical Chinese Poetry from Images with Memory Networks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. 5618–5625.

[51] Xiaopeng Yang, Xiaowen Lin, Shunda Suo, and Ming Li. 2018. Generating Thematic Chinese Poetry using Conditional Variational Autoencoders with Hybrid Decoders. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*. 4539–4545. https://doi.org/10.24963/ijcai.2018/631

[52] Yang Yang, Jie Zhou, Jiangbo Ai, Yi Bin, Alan Hanjalic, Heng Tao Shen, and Yanli Ji. 2018. Video Captioning by Adversarial LSTM. *IEEE Transactions on Image Processing* 27, 11 (2018), 5600–5611.

[53] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. 2016. Video Paragraph Captioning Using Hierarchical Recurrent Neural Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 4584–4593.

[54] Yang Yu. 2018. Towards Sample Efficient Reinforcement Learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*. 5739–5743.

[55] Li Zhang, Flood Sung, Feng Liu, Tao Xiang, Shaogang Gong, Yongxin Yang, and Timothy M Hospedales. 2017. Actor-Critic Sequence Training for Image Captioning. In *NIPS Workshop on Visually-Grounded Interaction and Language*.

[56] Mingxing Zhang, Yang Yang, Hanwang Zhang, Yanli Ji, Heng Tao Shen, and Tat-Seng Chua. 2019. More is Better: Precise and Detailed Image Captioning using Online Positive Recall and Missing Concepts Mining. *IEEE Transactions on Image Processing* 28, 1 (2019), 32–44.