Deep Reinforcement Polishing Network for Video Captioning

Wanru Xu, Jian Yu, Zhenjiang Miao, Lili Wan, Yi Tian, and Qiang Ji, *Fellow, IEEE*

***Abstract*—The video captioning task aims to describe video content using several natural-language sentences. Although one-step encoder-decoder models have achieved promising progress, the generations always involve many errors, which are mainly caused by the large semantic gap between the visual domain and the language domain and by the difficulty in long-sequence generation. The underlying challenge of video captioning, i.e., sequence-to-sequence mapping across different domains, is still not well handled. Inspired by the proofreading procedure of human beings, the generated caption can be gradually polished to improve its quality. In this paper, we propose a deep reinforcement polishing network (DRPN) to refine the caption candidates, which consists of a word-denoising network (WDN) to revise word errors and a grammar-checking network (GCN) to revise grammar errors. On the one hand, the long-term reward in deep reinforcement learning benefits the long-sequence generation, which takes the global quality of caption sentences into account. On the other hand, the caption candidate can be considered a bridge between visual and language domains, where the semantic gap is gradually reduced with better candidates generated by repeated revisions. In experiments, we present adequate evaluations to show that the proposed DRPN achieves comparable and even better performance than the state-of-the-art methods. Furthermore, the DRPN is model-irrelevant and can be integrated into any video captioning models to refine their generated caption sentences.**

***Index Terms*—Video captioning, deep reinforcement learning, word polishing, grammar polishing.**

Manuscript received December 17, 2019; revised March 28, 2020 and April 30, 2020; accepted June 1, 2020. Date of publication June 15, 2020; date of current version May 26, 2021. This work was supported in part by NSFC under Grants 61672089, 61703436, 61572064, 61906013, 61273274, 61876016, and 61632004, in part by CELFA, in part by the National Key R&D Program of China under Grant 2018AAA0100302, and in part by China Postdoctoral Science Foundation under Grant 2019M650469. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Wen-Huang Cheng. *(Corresponding author: Wanru Xu.)*

Wanru Xu is with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China, with the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China, and also with the Beijing Key Lab of Traffic Data Analysis and Mining, Beijing 100044, China (e-mail: xuwanru@bjtu.edu.cn).

Jian Yu and Yi Tian are with the School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China, and also with the Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China (e-mail: jianyu@bjtu.edu.cn; tianyi@bjtu.edu.cn).

Zhenjiang Miao and Lili Wan are with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China, and also with the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China (e-mail: zjmiao@bjtu.edu.cn; llwan@bjtu.edu.cn).

Qiang Ji is with the Department of Electrical and Computer Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180 USA (e-mail: qji@ecse.rpi.edu).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

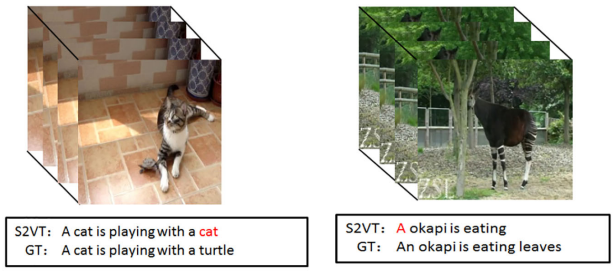
Digital Object Identifier 10.1109/TMM.2020.3002669

I. Introduction

THE goal of video captioning is to automatically generate natural-language descriptions of videos, which is a joint task of computer vision and natural-language processing. Video captioning plays a crucial role in many real-world applications, such as fast content-based video retrieval, video understanding, assist device for the visually impaired and automatic subtitle generation system.

The traditional encoder-decoder framework, e.g., sequence-to-sequence: video-to-text (S2VT) [1], has achieved promising performance on many sequence generation tasks, including machine translation, dialogue system, image and video question answering [2], and even image and video captioning. In such a framework, visual information is commonly encoded by the convolutional neural network (CNN) or recurrent neural network (RNN); then, RNN is used to decode caption sentences. However, the caption sentences generated by the one-step encoder-decoder based methods (e.g., S2VT [1] and S2VT+RL [3]), always involve many word errors and grammar errors. As shown in Fig. 1(a), the ground-truth word ("turtle") is incorrectly decoded as an error word ("cat"), so a noisy word is introduced by S2VT. As shown in Fig. 1(b), a grammar error occurs with the articles "a" and "an". These errors are mainly caused by two reasons: 1) The output of the video captioning task is a long sequence, but the global quality of the generated captions cannot be measured by the traditional training strategy with maximizing likelihood, which tends to select high-frequency words and introduces noisy words. 2) There is a large semantic gap between the visual domain and the language domain, and only considering visual information without language constraints would result in grammar errors. In summary, the underlying challenge of video captioning, i.e, sequence-to-sequence mapping across different domains, is still not well handled.

In such one-step framework, the generated sequence is directly considered as the final result without any polishing or revising. However, polishing or proofreading is a common behavior in the daily life of human beings. For example, when writing a paper, we usually first complete a "rough draft" and subsequently polish it again and again. Similarly, when translating a sentence, we often create an initial translation and incrementally refine it based on the global understanding of the entire text. To demonstrate the significance of word revision and grammar revision in video captioning, we show in Tables I and II the comparisons between the original result of S2VT+RL [3] and the optimal result that can be theoretically reached after the



(a) The generated caption sentence with word error.

(b) The generated caption sentence with grammar error.

Fig. 1. Examples of the generated caption sentences with word errors and grammar errors, where GT indicates the ground-truth caption sentences, and S2VT indicates the caption sentences generated by S2VT.

TABLE I
COMPARISON BETWEEN THE ORIGINAL RESULT OF S2VT+RL AND THE OPTIMAL RESULT THAT CAN BE REACHED AFTER REVISION IN THE IDEAL SITUATION ON MSVD

Models / Metrics	B@4	R	M	C
S2VT+RL	45.6	69.0	32.9	80.6
After Word Revision	68.8	85.3	43.2	116.9
After Grammar Revision	47.8	72.6	33.7	83.8
After Word+Grammar Revision	69.5	85.9	43.5	118.0

TABLE II
COMPARISON BETWEEN THE ORIGINAL RESULT OF S2VT+RL AND THE OPTIMAL RESULT THAT CAN BE REACHED AFTER REVISION IN THE IDEAL SITUATION ON MSR-VTT

Models / Metrics	B@4	R	M	C
S2VT+RL	39.2	60.3	26.7	44.8
After Word Revision	55.3	78.4	34.5	61.3
After Grammar Revision	41.3	63.4	27.6	48.6
After Word+Grammar Revision	55.7	79.9	35.6	62.5

revisions defined in this paper on MSVD and MSR-VTT, where BLEU4(B@4), ROUGE-L(R), METEOR(M) and CIDEr(C) are four widely employed metrics. The comparison clearly demonstrates that the original captions indeed have many errors, which can be theoretically tackled by word revision and grammar revision, since there is a large margin for improvement between the original result of S2VT+RL and the optimal result after revision, which will be detailed in the following sections. Motivated by this intuition, we propose a novel deep reinforcement polishing network, which gradually refines the output sentence based on the generated caption candidates and the grammar rules at each revision step. Since the semantics of a natural-language sentence is decided by both the words and grammar, the proposed DRPN consists of a word-denoising network and a grammar-checking network to revise word errors and grammar errors, respectively.

Introducing the polishing mechanism into video captioning has two benefits. First, the semantic gap between the visual domain and the language domain can be gradually reduced by the polishing mechanism via multi-step encoding-decoding, since better candidates are generated by repeated revisions that provide more valuable and correct evidences for video captioning. Thus, the original cues extracted from the visual domain and the intermediate cues obtained from the language domain are both encoded for the next decoding procedure. Second, compared to the one-pass decoding, which only generates the current word depending on the previous information, the polishing procedure provides a method of multiple-pass decoding to leverage the global information by looking into both previous words and future words in caption sentences. When polishing a word or a local part, we consider the whole picture of the caption sentence to evaluate how well the local revision fits into the global sentence by a long-term reward, which is defined in the deep reinforcement framework. Thus, the goal of each revision is to improve the global quality of the generations, which is beneficial for long-sequence generation.

In summary, our contributions are as follows: 1) We propose a novel video caption polishing problem to enable the model with the capacity of repeated polishing, which simulates the human cognitive behaviors. 2) We propose a novel deep reinforcement polishing network to gradually improve the generated captions by revising the word errors and grammar errors, which introduces a polishing mechanism into the video captioning framework via reinforcement learning. 3) The proposed polishing network is model-irrelevant and can be integrated into any video captioning models to refine their generated caption sentences. The experiment also demonstrates that we can achieve the best performance when we use the state-of-the-art method as the baseline to get better candidates.

II. RELATED WORK

In this section, we will discuss the existing video caption algorithms [4], [5], including the language-template-based methods [6]–[8], sequence-to-sequence (seq-to-seq) models, [1], [9]–[11], and dense video captioning methods [12], [13]. The seq-to-seq models can be further divided into the sequence-to-sequence model with attention mechanism, sequence-to-sequence model with deep reinforcement learning and sequence-to-sequence model with adversarial learning.

The language-template-based method aims to describe video content using a pre-defined language template, such as "subject-verb-object (SVO) Tuples". It commonly consists of two stages: content identification, which recognizes the main objects and actions in video sequence, and sentence generation, which fills the detected objects and actions into "subject," "verb" and "object" of the language template. In [7], a system is proposed for video captioning by the form of "who did what to whom, and where and how they did it," where object categories, properties, and spatial relations are considered as nouns, adjectival modifiers, and preposition; action categories and characteristics are considered as verbs and adverbial modifiers. In [14], a method is proposed to describe activities from video sequences based on hierarchical concepts of human actions, where the concepts with the appropriate syntactic component are first extracted from videos and subsequently translated into natural-language sentences. A language template of "subject-verb-object-place" is built in [6], and a factor graph is used to combine these detections to generate caption sentences, which are detected by the state-of-the-art objects, actions, and scenes recognition methods. To generate textual descriptions of action videos, a hybrid

method is presented in [15] to generate a caption sentence based on the detected verb, subject, and direct and indirect objects. In [16], the semantic representation of visual content including the object and action labels are first predicted by a CRF and subsequently translated to natural language by the machine translation techniques.

The sequence-to-sequence model is currently the state-of-the-art method for video captioning and usually adopts the encoder-decoder based framework [1], [17], [18], where the video encoding stage learns visual features and subsequently feeds them into the decoder for text generation, which is called the decoding stage. In [19], CNN is first utilized to encode the visual information of each frame, subsequently obtain the features of video sequence by mean pooling, and finally decode caption sentence by the long short-term memory (LSTM). To better capture the temporal dynamics of a video sequence, CNN is replaced by LSTM as the encoder in [1]. To improve the seq-to-seq model, several efforts are made as follows: 1) Seq-to-seq model with attention mechanism [20]–[22]. Since the traditional seq-to-seq model encodes the video sequence into a feature vector with a fixed length, where some detail visual information is lost, the attention mechanism is introduced to address this issue. Yao *et al.* [9] introduce a temporal attention mechanism into the 3DCNN-RNN framework, which considers both local and global temporal structures and can automatically select the most relevant temporal segments when decoding a certain part. A hierarchical recurrent neural network (h-RNN) is proposed in [23] to integrate both temporal attention and spatial attention. In [24], an LSTM with an attention model is the decoder, which enables the model to adaptively focus on the most correlated feature tubes to generate each word. 2) Seq-to-seq model with deep reinforcement learning [10], [11], [25]. Since the traditional seq-to-seq model adopts the log-likelihood as the objective function, which only focuses on maximizing the local similarities between sequences, deep reinforcement learning is introduced to allow the model to directly optimize the global similarities. A novel decision-making framework is proposed for video captioning in [10], which consists of a high-level manager module and a low-level worker module to design the sub-goals and recognize the primitive actions, respectively. 3) Seq-to-seq model with adversarial learning [26]. To improve the "naturalness" and "diversity" of the generated caption, the adversarial learning is recently introduced into video captioning [27], visual paragraph generation [28] and image captioning [29], [30]. An LSTM-GAN architecture is proposed in [27], where a generator generates textual sentences given the visual features of video, and a discriminator encourages the generations to be undistinguishable from the ground truth. In [31], MLADIC is proposed for two dual tasks: text-to-image synthesis and image captioning, where the multi-task learning helps to improve the performance of image captioning task by bridging the gap between language and vision domains.

Dense video captioning [12], [32], [33] can be considered a more complex video captioning task, and it aims to generate language descriptions for untrimmed videos, when multiple events occur at various timespans. Therefore, the common approach is first to predict the temporal intervals by action detection techniques, subsequently describe each detected event by captioning techniques, and obtain the final result by combining these short sentences to a long sentence. A weakly supervised dense video captioning model is introduced in [34], and it is trained only with the video-level sentences annotations, where a weakly supervised multi-instance multi-label learning method is required to first link the video regions with caption words. Currently, some end-to-end models are proposed for dense video captioning. In [13], a unified end-to-end transformer model is built to detect and describe events via a proposal decoder and a captioning decoder. Another end-to-end framework is proposed in [35], which integrates an event generation network to adaptively select several event proposals and a sequential video captioning network to generate caption sentences. In [36], a novel context-and-attribute grounded model is proposed for dense captioning, which combines a multi-level attribute generation network and a context mining network.

All above approaches adopt the one-step decoding process, where the generated sequence is directly considered as the final result without further polishing. Currently, some multiple-pass decoding based methods are proposed to improve the performance of machine translation and image captioning. In [37], two levels of decoders are first proposed for machine translation, where the hidden state of the first decoder is treated as the input of the second decoder to integrate some global information of the entire caption. Then, the multiple-pass decoding is used for image captioning [38], [39], and the model combines the output and the hidden state of the first decoder to feed into the second decoder. However, they fail to use the top-K candidates and only consider the top-1 candidate (i.e., raw caption) as an additional cue, which actually involves many errors, where we have demonstrated the significance of the top-K candidates in Tables I and II. Although they have achieved promising results, the progress is mainly caused by the stacking decoding layers instead of the polishing operation. Therefore, we will provide a clear problem formulation of caption polishing and introduce the true polishing mechanism into the video captioning task in this paper, which replaces the error word with the correct word (word revision) or selects an appropriate transformation for the error word (grammar revision).

III. Methodology

A. Overview

Following the real-word human cognitive processes, we improve the generated captions by repetitive revisions with an optimized polishing network in this paper. Given the visual representation $\mathbf{x}$ of a video and some sentences $\{w_{1,k}, w_{2,k}, w_{3,k}, \ldots, w_{L,k}\}_1^K$ as the caption candidates generated by any captioning models, the polishing process can be conducted as follows,

$$w_1^t, \ldots, w_L^t = \Psi_{A^t}(\mathbf{x}, \{w_{1,k}^{t-1}, w_{2,k}^{t-1}, \ldots, w_{L,k}^{t-1}\}_1^K) \quad (1)$$

Ψ is the candidate updater to change the caption sentence from $\{w_1^{t-1}, w_2^{t-1}, \ldots, w_L^{t-1}\}$ to $\{w_1^t, w_2^t, \ldots, w_L^t\}$ in terms of

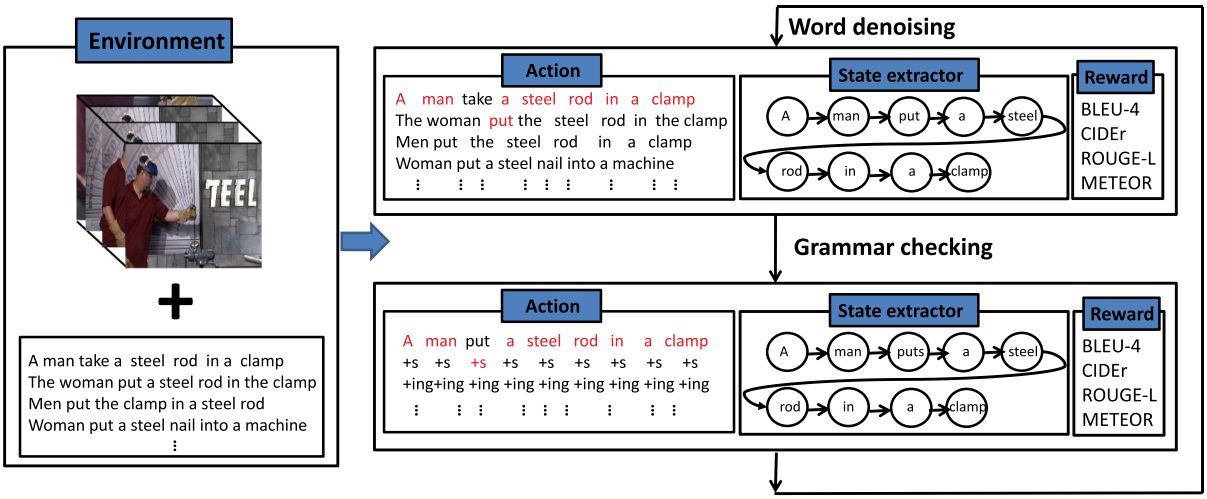


Fig. 2. Illustration of the proposed deep reinforcement polishing network, where the environment denotes the combination of the video sequence and several caption candidates; the action is defined as the way to revise the caption candidates; the state describes the current information and provides the evidences for the next revision; the reward function evaluates the global quality of caption candidates.

a series of polishing operations A^t, where $\{w_1^t, \ldots, w_L^t\} = \{w_{1,1}^t, w_{2,1}^t, \ldots, w_{L,1}^t\}$ is the top-1 caption candidate after t times of revision, and $\{w_{1,1}^0, \ldots, w_{L,1}^0\} = \{w_{1,1}, \ldots, w_{L,1}\}$ is the original top-1 caption candidate; L is the length of the caption sentence; and the size of caption candidates is K. The function of the candidate updater is to refine the caption sentences and update the caption candidates according to the selected polishing operations. It is considered an effective revision, only if this operation satisfies $R[w_1^t, w_2^t, \ldots, w_L^t] \geq R[w_1^{t-1}, w_2^{t-1}, \ldots, w_L^{t-1}]$, where R is a metric to measure the quality of caption sentences. Therefore, our goal is to find the optimal $\{A^1, \ldots, A^T\}$ to revise the generated caption sentence that maximizes R,

$$\max_{A^1,\ldots,A^T} R[\Psi_{\{A^1,\ldots,A^T\}}(\mathbf{x}, \{w_{k,1}, w_{k,2}, \ldots, w_{k,L}\}_1^K)] \quad (2)$$

Obviously, it is a sequential decision-making process, so we adopt reinforcement learning to address this problem.

The overall polishing procedure is shown in Fig. 2, including a word-denoising process and a grammar-checking process. In other words, our polishing network contains two types of polishing operations: $T = 2$, where A^1 is for word revision, and A^2 is for grammar revision. First, the environment in this framework is defined as the combination of video sequence and several caption candidates generated by any captioning models; Then, the word-denoising network selects a more appropriate word from the candidate pool to replace the incorrect word, such as "a man take a steel rod in a clamp" is revised as "a man put a steel rod in a clamp"; Finally, the grammar-checking network selects an appropriate transformation for each word in terms of grammar rules, such as "a man put a steel rod in a clamp" is revised as "a man puts a steel rod in a clamp".

We cast the video caption polishing problem as a Markov decision process (MDP) to gradually revise the errors in caption candidates by a trained polishing agent. Typically, an MDP is defined as an $(S; A; R)$ tuple, including a set of states $s \in S$, a set of actions $a \in A$, and a reward function $R(s, a)$. Correspondingly, our WDN and GCN consist of three components: state extraction part, decision-making part, and reward measurement part. At each revision step, the decision-making part first selects a series of optimal actions $A^{t^*} = \{a_1^{t^*}, a_2^{t^*}, \ldots, a_L^{t^*}\}$ to polish the captions in terms of the current state $\mathbf{s}^t = \{\mathbf{s}_1^t, \ldots, \mathbf{s}_L^t\}$, which is considered as a policy mapping from the state set to the action set,

$$a_l^{t^*} = \arg\max \pi(a_l^t | \mathbf{s}_l^t; \Theta) \quad (3)$$

where $\pi(a_l^t | \mathbf{s}_l^t; \Theta)$ denotes the probability of selecting a_l^t at state $\mathbf{s}_l^t$. Then, each word in caption sentence is revised by $\{a_1^{t^*}, a_2^{t^*}, \ldots, a_L^{t^*}\}$, and the caption sentence is updated as $\{w_1^t, w_2^t, \ldots, w_L^t\}$. Correspondingly, the state is updated as $\{\mathbf{s}_1^{t+1}, \mathbf{s}_2^{t+1}, \ldots, \mathbf{s}_L^{t+1}\}$ by the state extraction part in terms of $\{w_1^t, \ldots, w_L^t\}$. Finally, the reward measurement part measures the improvement of caption's quality obtained via this revision, which is fed back to guide the learning of the decision-making part and state extraction part.

We optimize the parameters of WDN and GCN using reinforcement learning, i.e., the policy-gradient algorithm [40], which directly maximizes the expected long-term reward by,

$$\begin{aligned} J &= E_{(a_l^t, \mathbf{s}_l^t) \sim \pi(a_l^t | \mathbf{s}_l^t; \Theta)} R(w_1^t, \ldots, w_L^t | a_1^t \mathbf{s}_1^t, \ldots, a_L^t \mathbf{s}_L^t) \\ &\approx \sum_{a_1^t \mathbf{s}_1^t, \ldots, a_L^t \mathbf{s}_L^t} \prod_l \pi(a_l^t | \mathbf{s}_l^t; \Theta) R_L \end{aligned} \quad (4)$$

where $R_L = R(w_1^t, \ldots, w_L^t | a_1^t \mathbf{s}_1^t, \ldots, a_L^t \mathbf{s}_L^t)$, and it is a function of action a and state s, since different $a_1^t \mathbf{s}_1^t, \ldots, a_L^t \mathbf{s}_L^t$ result in different $w_1^t, \ldots, w_L^t$. For simplicity, this objective function is computed over several samples by sampling. To reduce the variance, a baseline b is introduced, where we will discuss the

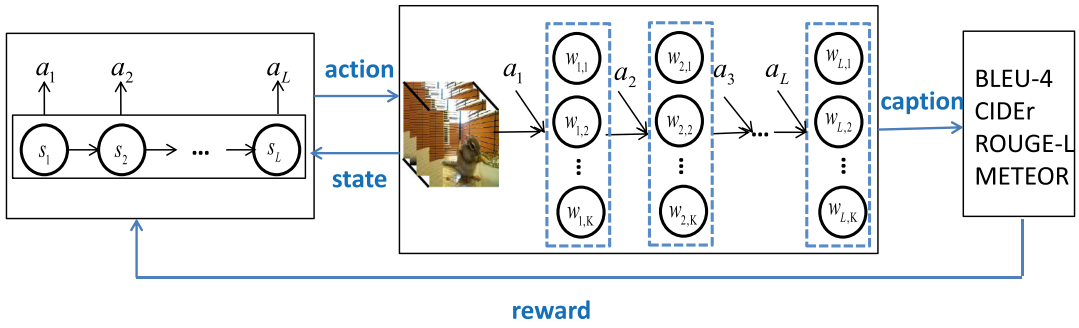


Fig. 3. Illustration of the word-denoising network, which is used to revise the word errors by word replacement in terms of the caption candidates.

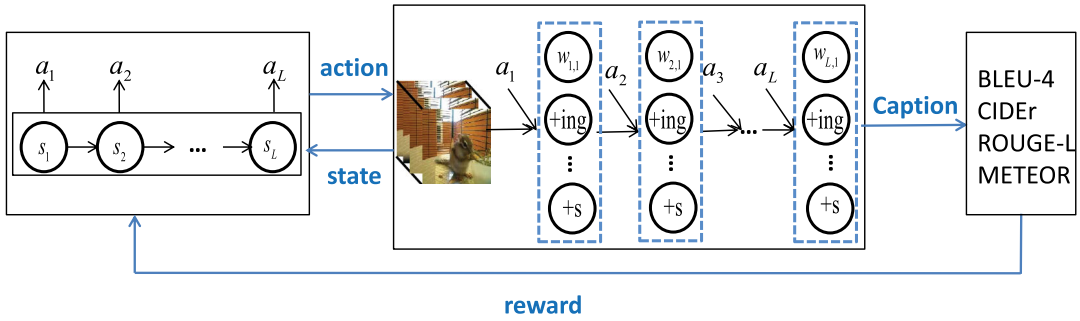


Fig. 4. Illustration of the grammar-checking network, which is used to revise the grammar errors by word transformation in terms of grammar rules.

selection of b in experiment; thus, the objective function is converted to Eq. (5).

$$J = \sum_{a_1^t \mathbf{s}_1^t, \ldots, a_L^t \mathbf{s}_L^t} \prod_l \pi(a_l^t | \mathbf{s}_l^t; \Theta)(R_L - b) \tag{5}$$

$R_L - b$ is used to scale the gradient, which can be considered as an estimation of the benefit of taking action a_l^t at state $\mathbf{s}_l^t$. Next, we describe the WDN and GCN in detail.

B. Word-Denoising Network

The goal of the WDN is to revise the word errors in caption candidates, as illustrated in Fig. 3. Now, we describe its structure and specific tuple $(S^1; A^1; R^1)$.

Action: The action of WDN is defined as the word replacement, and the action set is built by all caption candidates, i.e., $A^1 = \{A_1^1, \ldots, A_L^1\}$, where each A_l^1 is the top-K words with maximal possibilities at the $l-th$ position of the caption sentence generated by any video captioning models.

$$A_l^1 = \{w_{l,1}, w_{l,2}, \ldots, w_{l,K}\} = \underset{1 \le k \le K}{\operatorname{argmax}} P(w_l | \mathbf{x}, w_{l-1}; \theta) \tag{6}$$

In WDN, $a_l^1 = w_{l,k}$ indicates that the model selects the candidate word $w_{l,k}$ to replace the original word at position l.

State: We adopt two methods to define the state of the WDN, where $\mathbf{s}^1 = \{\mathbf{s}_1^1, \ldots, \mathbf{s}_l^1, \ldots, \mathbf{s}_L^1\}$ and each $\mathbf{s}_l^1 = \{\mathbf{s}_{l,1}^1, \ldots, \mathbf{s}_{l,K}^1\}$. In the first method, which is called the "concatenated state," the state is defined as the concatenation of the video's visual feature $\mathbf{x}$, previous word w_{l-1}^1 and current word candidates $\{w_{l,1}^0, w_{l,2}^0, \ldots, w_{l,K}^0\}$ by first mapping them into a unified space via f_w, f_{pre} and f_{can}, respectively,

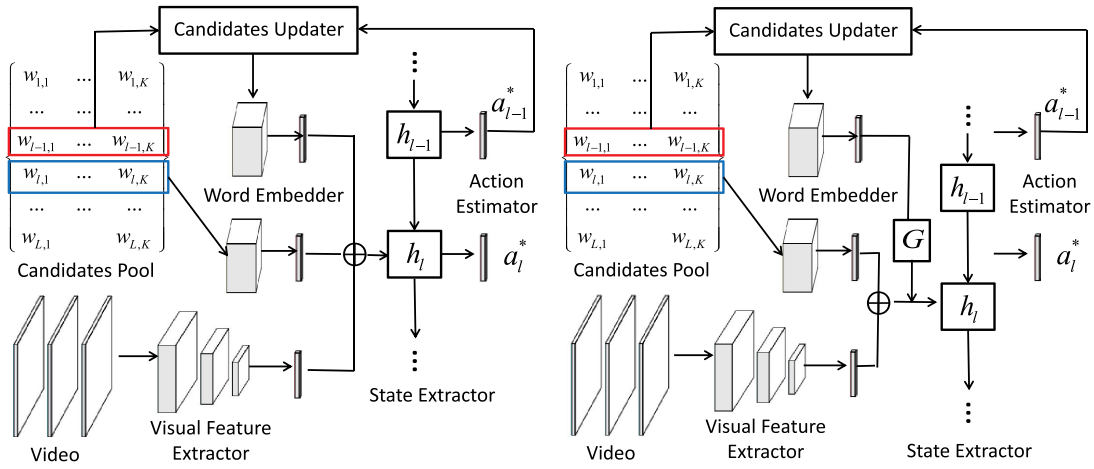
$$\mathbf{s}_l^1 = \begin{Bmatrix} f_w(\mathbf{x}) \oplus f_{pre}(w_{l-1}^1) \oplus f_{can}(w_{l,1}^0), \mathbf{h}_{l-1}^1 \\ f_w(\mathbf{x}) \oplus f_{pre}(w_{l-1}^1) \oplus f_{can}(w_{l,2}^0), \mathbf{h}_{l-1}^1 \\ \cdots \\ f_w(\mathbf{x}) \oplus f_{pre}(w_{l-1}^1) \oplus f_{can}(w_{l,K}^0), \mathbf{h}_{l-1}^1 \end{Bmatrix}^T \tag{7}$$

where $\oplus$ indicates the vector concatenation. To capture the temporal relationships between words, we adopt LSTM for state extraction, and $\mathbf{h}_{l-1}^1$ is the hidden state at time step $l-1$, which preserves the long-term information of the caption sentence before time step l. Another method is called the "gated state," where the current word candidate is considered as a gate, and we can obtain the candidate-related state by integrating the word candidate into the visual feature $\mathbf{x}$ and previous word w_{l-1}^1.

$$\mathbf{s}_l^1 = \begin{Bmatrix} F_{gate}(f_w(\mathbf{x}) \oplus f_{pre}(w_{l-1}^1), f_{gate}(w_{l,1}^0)), \mathbf{h}_{l-1}^1 \\ F_{gate}(f_w(\mathbf{x}) \oplus f_{pre}(w_{l-1}^1), f_{gate}(w_{l,2}^0)), \mathbf{h}_{l-1}^1 \\ \cdots \\ F_{gate}(f_w(\mathbf{x}) \oplus f_{pre}(w_{l-1}^1), f_{gate}(w_{l,K}^0)), \mathbf{h}_{l-1}^1 \end{Bmatrix}^T \tag{8}$$

where F_{gate} is a gating operation, and f_{gate} is a gating function that we will describe.

Reward: The reward of WDN measures the quality of the caption sentence, which can be defined as any formulations. It is a long-term delayed reward $R_L^1 = R^1(w_1^1, \ldots, w_L^1 | a_1^1 \mathbf{s}_1^1, \ldots, a_L^1 \mathbf{s}_L^1)$, which calculates the global similarity between the generation and the ground truth after taking the action sequence of $\{a_1^1, a_2^1, \ldots, a_L^1\}$ to polish the entire



(a) Model with the concatenated state. (b) Model with the gated state.

Fig. 5. Architecture of the word-denoising network (WDN) and grammar-checking network (GCN), where they have identical model structures. We propose two methods to define the state: model with the concatenated state and model with the gated state.

caption sentence. In this paper, we choose ROUGE-L as the reward function based on experience.

Architecture: As mentioned, the WDN consists of a state extraction part (including a visual feature extractor, a word embedder and a state extractor), a decision-making part (an action estimator) and a reward measurement part (a metric function), as shown in Fig. 5.

1) For the model with the concatenated state as shown in Fig. 5(a). In the visual feature extractor, first, a 2D feature with 2048 dimensions and a 3D feature with 1536 dimensions are obtained from ResNeXt [41], where the former is the average pooling feature of the conv5/block3 output, and the latter is the global pooling feature in ECO. Then, they are compressed into a 256-dimensional vector by a fully connected layer f_w. Next, in the word embedder, each word vector is first gained by a pre-trained vocabulary, and the resulting 500-dimensional vector is further fed into a fully connected layer (i.e., f_{pre} and f_{can}, respectively) to obtain a 512-dimensional vector for the current word and a 256-dimensional vector for its previous word. Afterwards, we concatenate the three vectors to feed into the state extractor, which is implemented by an LSTM of 1024 neurons, and the action estimator is a fully connected layer. The detail process is as follows,

$$\begin{aligned}
&\mathbf{c}^1_{l,k} = \varphi_{\tanh}(\varphi_{emb}(w^1_{l-1}) \cdot W_p) \oplus \varphi_{\tanh}(\mathbf{x} \cdot W_i) \\
&\oplus \varphi_{\tanh}(\varphi_{emb}(w^0_{l,k}) \cdot W_c) \\
&\mathbf{h}^1_{l,k} = \varphi_{\tanh}(\mathbf{h}^1_{l-1} \cdot W_{hh} + \mathbf{c}^1_{l,k} \cdot W_{ih}) \\
&\mathbf{h}^1_l = \{\mathbf{h}^1_{l,k}, \mathbf{h}^1_{l,2}, \ldots, \mathbf{h}^1_{l,K}\} \\
&\pi(a^1_l|\mathbf{s}^1_l;\Theta) = \varphi_{sigm}(W_f * \mathbf{h}^1_l + \mathbf{b})
\end{aligned} \tag{9}$$

where φ_{sigm} and $\varphi_{\tanh}$ denote the sigmoid and tanh function, respectively; each word is converted to a word vector by an embedding function φ_{emb}; and the model is parameterized by $\Theta = \{W_f, \mathbf{b}, W_i, W_p, W_c, W_{hh}, W_{ih}\}$.

2) For the model with the gated state as shown in Fig. 5(b). In the visual feature extractor, the 2D and 3D features are compressed into a 512-dimensional vector by a fully connected layer. Next, in the word embedder, each 500-dimensional word vector is further fed into a fully connected layer to obtain a 1024-dimensional vector for the current word via f_{gate} as a gating vector and a 512-dimensional vector for its previous word via f_{pre}. Afterwards, we concatenate the visual feature vector and word vector of the previous word and obtain the candidate-related state by gating them with the gating vector via gating operator F_{gate}. The following process is similar to that of the model with the concatenated state.

$$\begin{aligned}
&\mathbf{c}^1_{l,k} = (\varphi_{\tanh}(\varphi_{emb}(w^1_{l-1}) \cdot W_p) \oplus \varphi_{\tanh}(\mathbf{x} \cdot W_i)) \\
&\odot (\varphi_{\tanh}(\varphi_{emb}(w^0_{l,k}) \cdot W_g)) \\
&\mathbf{h}^1_{l,k} = \varphi_{\tanh}(\mathbf{h}^1_{l-1} \cdot W_{hh} + \mathbf{c}^1_{l,k} \cdot W_{ih}) \\
&\mathbf{h}^1_l = \{\mathbf{h}^1_{l,k}, \mathbf{h}^1_{l,2}, \ldots, \mathbf{h}^1_{l,K}\} \\
&\pi(a^1_l|\mathbf{s}^1_l;\Theta) = \varphi_{sigm}(W_f * \mathbf{h}^1_l + \mathbf{b})
\end{aligned} \tag{10}$$

where gating operator F_{gate} is implemented by $\odot$, i.e., Hadamard operator (element-wise multiply); the model is parameterized by $\Theta = \{W_f, \mathbf{b}, W_i, W_p, W_g, W_{hh}, W_{ih}\}$.

C. Grammar-Checking Network

The goal of GCN is to revise the grammar errors in the caption candidates, as illustrated in Fig. 4. The structure and definition of tuple $(S^2; A^2; R^2)$ of GCN are similar to those of WDN. The only difference is the definition of action, where the action of GCN is defined as the word transformation in terms of the grammar rules, and the action of WDN is defined as the word replacement based on the caption candidates. In this setting, $\mathbf{a}^2 = \{a^2_1, a^2_2, \ldots, a^2_L\}$, and each action a^2_l at the $l-th$ position

of the caption sentence is selected from an action set A_l^2.

$$A_l^2 = \begin{Bmatrix} w_l^1 \\ w_l^1 + \text{“}ing\text{”}/w_l^1 - \text{“}ing\text{”} \\ w_l^1 + \text{“}s\text{”}/w_l^1 - \text{“}s\text{”}/w_l^1 + \text{“}es\text{”}/w_l^1 - \text{“}es\text{”} \\ \text{“}a\text{”} \leftrightarrow \text{“}the\text{”}/\text{“}a\text{”} \leftrightarrow \text{“}an\text{”}/\text{“}an\text{”} \leftrightarrow \text{“}the\text{”} \end{Bmatrix} \quad (11)$$

Specifically, w_l^1 is the word at the $l-th$ position of the caption sentence after revising the word error by WDN; $w_l^1 +$ “*ing*” $/w_l^1 -$ “*ing*” indicates the transformation of the present progressive on w_l^1; $w_l^1 +$ “*s*” $/w_l^1 -$ “*s*” $/w_l^1 +$ “*es*” $/w_l^1 -$ “*es*” indicates the transformation of the singular and plural forms for noun w_l^1, which also contains special cases (e.g., “woman” to “women,” “man” to “men”, “is” to “are” and vice versa), and the transformation of the third person singular form for verb w_l^1; “*a*” $\leftrightarrow$ “*the*” / “*a*” $\leftrightarrow$ “*an*” / “*an*” $\leftrightarrow$ “*the*” is the transformation for articles. Since it usually does not occur in video caption sentences, we do not employ the transformation of past tense for verbs. Note that we adopt the part-of-speech-irrelevant (POS-irrelevant) polishing operation, and it may transform w_l^1 to a word that does not exist in the vocabulary and even in the world, so we add a tag of “NULL” to the vocabulary.

IV. Experiments

We evaluate the performance of our proposed deep reinforcement polishing network on two commonly used video captioning datasets. In this section, we first introduce the datasets, evaluation metrics and implementation details. Then, we compare to the state-of-the-art methods. Next, we conduct comprehensive evaluations and ablation studies to evaluate the effectiveness of the components and parameters in DRPN. Finally, we present the qualitative results.

A. Dataset and Experiment Setup

In this paper, we evaluate our DRPN on MSVD and MSR-VTT for video captioning.

MSVD [42]: Microsoft Video Description dataset consists of 1,970 YouTube video clips with human annotated sentences, which contains 70,028 captions collected by Amazon Mechanical Turk (AMT) workers. On average, the duration of each video in MSVD is 10-25 seconds and mainly contains a single activity; there are 41 caption sentences per clip, where each sentence contains approximately 8 words. This dataset is commonly split into training, validation and testing partitions of 1200, 100 and 670 videos, respectively.

MSR-VTT [43]: Currently, MSR-Video To Text dataset is the largest public video captioning dataset presented in 2016. It consists of 10,000 video clips and 200,000 caption sentences, which are derived from a wide variety of videos in 20 general categories. The common split is provided as follows: 6,513 videos for training, 497 for validation, and 2,990 for testing. On average, each video in MSR-VTT is annotated with 20 reference captions by AMT workers, and its average duration is approximately 20 seconds.

Implementation details and evaluation metrics: We follow the standard training and testing split scheme for each dataset. We implement our model with Tensorflow [44]. The parameters are optimized by the Adam optimizer [45] with a learning rate of 1e-5 and a small decaying rate, and the dropout regularization [46] is used to avoid overfitting. We initialize all trainable parameters by drawing from a uniform distribution $[-0.1, 0.1]$. The number of caption candidates is set to $K = 7$ for MSVD and $K = 20$ for MSR-VTT by cross validation, which will be evaluated in the following section. For the evaluation metrics, we adopt four widely employed metrics in video captioning or image captioning: BLEU4 [47], ROUGE-L [48], METEOR [49] and CIDEr [50], whose scores can be calculated utilizing the MSCOCO evaluation.

B. Training Methods Considered

Policy-gradient approaches in reinforcement learning (RL) have two common procedures: warm-start training and sample variance reduction. From the two perspectives, we train our model with four strategies, which result in four variants of our model: model-XE, model-XE-RL-SCST, model-XE-RL-Max and model-XE-RL-Avg.

Video captioning is basically a sequence generation problem, and the difficulty of aligning the model output distribution with the reward distribution over the large search space of possible sequences makes RL training slow and inefficient. As a result, like other RL methods, we require a warm-start phase (e.g., the first 50 epoches), where supervised learning is used to pre-train the model by a cross-entropy objective (XE), followed by a model-refinement phase, where reinforcement learning is used to refine the model. The traditional supervised learning method [1] cannot be directly used here, since the “best caption” with respect to the original caption after revision is not given, which causes the ground-truth word to be missing at each position. Therefore, the supervised learning with a reward-metric based objective is used in the warm-start phase, where the ground-truth word at each position is considered the word to maximize the current intermediate reward in this paper.

$$w_l^* = argmax_{w_l \in \{w_{l,1},\ldots,w_{l,K}\}} R(w_1, \ldots w_l, \ldots w_L) \quad (12)$$

We fix the previous words $w_1, w_2, \ldots w_{l-1}$ and subsequent words $w_{l+1}, w_{l+2}, \ldots w_L$ and select a word w_l^* from the candidates $\{w_{l,1}, \ldots, w_{l,K}\}$ as the ground-truth word by maximizing the reward function R. In the model-refinement phase, the pre-trained model is considered an initialization; then, reinforcement learning is utilized to continue training the model by optimizing Eq. (5). In summary, if the model is only trained by the warm-start phase, it is model-XE. If the model is trained by both phases, using different baselines in Eq. (5) results in three variants of our model: the model using the average reward across all possible actions as the baseline is model-XE-RL-Avg; the model using the maximal reward across all possible actions as the baseline is model-XE-RL-Max; the model using the greedy sampling value (e.g., SCST [3]) as the baseline is model-XE-RL-SCST. In the following experiment, we evaluate our model with these different training strategies.

TABLE III
COMPARISONS TO THE EXISTING STATE-OF-THE-ART VIDEO CAPTIONING METHODS ON MSVD

Methods	B@4	R	M	C
S2VT [1]	42.8	68.7	32.5	75.0
SA [9]	41.9	-	29.6	51.7
LSTM-E [51]	45.3	-	31.0	-
S2VT+RL [3]	45.6	69.0	32.9	80.6
aLSTMs [52]	50.8	-	33.3	74.8
MA-LSTM [53]	52.3	-	33.6	70.4
AF [21]	52.4	-	32.0	68.8
Song *et al*. [54]	53.0	-	33.6	73.8
Wei Li *et al*. [55]	48.0	-	31.6	68.8
TM-P-HRNE [56]	52.8	70.5	33.4	68.9
PickNet [11]	46.1	69.2	33.1	76.0
V-ShaWei-GA [57]	47.9	-	30.9	-
M^3 [58]	52.0	-	32.2	-
hLSTMat [20]	53.0	70.3	33.5	73.8
TSA-ED [59]	51.7	-	34.0	74.9
RecNet [17]	52.3	69.8	34.1	80.3
TDConvED [18]	53.3	-	33.8	76.4
FCVC-CF-IA [22]	53.1	71.8	**34.8**	79.8
S2VT + DRPN	45.3	69.4	32.6	75.8
S2VT+RL + DRPN	49.2	71.5	34.2	**86.4**
hLSTMat + DRPN	**57.3**	**72.0**	34.3	78.3

TABLE IV
COMPARISONS TO THE EXISTING STATE-OF-THE-ART VIDEO CAPTIONING METHODS ON MSR-VTT

Methods	B@4	R	M	C
S2VT [1]	35.3	57.8	26.6	40.7
SA [9]	36.6	-	25.9	-
LSTM-E [51]	36.1	58.6	25.8	38.5
MA-LSTM [53]	36.5	59.8	26.5	41.0
Song *et al*. [54]	38.3	-	26.3	-
S2VT+RL [3]	39.2	60.3	26.7	44.8
aLSTMs [52]	38.0	-	26.1	43.2
AF [21]	39.4	-	25.7	40.4
Wei Li *et al*. [55]	37.5	-	26.4	-
M^3 [58]	38.1	-	26.6	-
PickNet [11]	38.9	59.5	27.2	42.1
V-ShaWei-GA [57]	37.9	-	25.9	-
hLSTMat [20]	39.1	59.3	26.6	42.7
RecNet [17]	39.1	60.3	27.5	48.7
TDConvED [18]	39.5	-	27.5	42.8
TM-P-HRNE [56]	39.2	60.1	26.9	44.6
Ruminant Decoding [39]	38.8	60.5	27.1	49.0
S2VT+RL + DRPN	**40.8**	**61.5**	27.5	48.0
[39] + DRPN	39.5	61.0	**27.7**	**49.2**

C. Comparison to the State-of-the-Art Methods

We start the evaluation with a comparison to the state-of-the-art methods in Tables III and IV.

There are several comparison methods: 1) S2VT [1] is the basic encoder-decoder based method. 2) S2VT+RL [3] is an improved method of S2VT by reinforcement learning. 3) SA [9] introduces a temporal attention mechanism into 3D CNN-RNN framework, which considers both local and global temporal structures. 4) M^3 [58] proposes a multi-modal memory model (M^3) for video captioning, where a visual and textual shared memory is built to model the long-term visual-textual dependency. 5) PickNet [11] proposes a plug-and-play to select the most informative frame in video captioning. 6) AF [21] proposes a modality-dependent attention mechanism with temporal attention to integrate the cues from multiple modalities. 7) MA-LSTM [53] presents a novel long-short term memory network with multimodal attention to boost video captioning by fully exploiting both multi-modal streams and temporal attention. 8) LSTM-E [51] proposes a novel long short-term memory network with visual-semantic embedding, which simultaneously explores the learning of LSTM and visual-semantic embedding. 9) TDConvED [18] presents a temporal deformable convolutional encoder-decoder network to conduct convolutions in both encoder and decoder. 10) RecNet [17] proposes a reconstruction network, which integrates an encoder-decoder-reconstructor architecture to leverage the forward flow (i.e., video to sentence) and backward flow (i.e., sentence to video) for video captioning. 11) hLSTMat [20] proposes a novel hierarchical LSTM with adjusted spatial-temporal attention, which decides when to depend on the language context information or the visual information. 12) V-ShaWei-GA [57] proposes several multimodal deep fusion strategies to take full advantage of the visual-audio information. 13) FCVC-CF-IA [22] exploits a novel architecture, i.e., the fully convolutional network with coarse-to-fine and inherited attention. 14) Wei Li *et al.* [55] proposes a multimodal framework combined with the attention mechanism and memory networks together. 15) TSA-ED [59] is a trajectory-structured attentional encoder-decoder-based model, which integrates the spatial-temporal representation at the trajectory level via the structured attention mechanism. 16) Ruminant Decoding [39] contains an image encoder, a base decoder, and a ruminant decoder for image captioning, which performs global planning with the output of the base decoder. 17) TM-P-HRNE [56] is proposed to jointly leverage several sorts of visual features and semantic attributes.

From Tables III and IV, we obtain the following observations and conclusions. 1) Our DRPN can indeed revise the errors in caption sequences and improve the quality of caption sequences. Take S2VT+RL as an example, when we use it to generate the caption candidates, for MSVD, we achieve a gain of 3.6%, 2.5%, 1.3% and 5.8% on BLEU4, ROUGE-L, METEOR and CIDEr, respectively; for MSR-VTT, we achieve a gain of 1.6%, 1.2%, 0.8% and 3.2% on the four metrics. The same case occurs when we use other captioning models as the baselines and conduct polishing on them. The performances of S2VT, S2VT+RL and hLSTMat are clearly improved by appending our DRPN. 2) Our DRPN can achieve comparable and even better performance than the state-of-the-art video captioning methods. Because no metric can perfectly measure the quality of the generated caption sentences, no method can obtain all highest scores on the four metrics. For example, FCVC-CF-IA [22] obtains the highest score of METEOR, but its CIDEr score is lower than S2VT+RL [3] and RecNet [17] on MSVD; RecNet [17] and Guo *et al*. [39] obtain the highest scores of METEOR and CIDEr, respectively, but their BLEU4 scores are lower than TDConvED [18] on MSR-VTT. Even so, our DRPN outperforms the state-of-the-art methods on BLEU4, ROUGE-L and CIDEr and is on par with

TABLE V
COMPARISON OF WDN, GCN AND DRPN ON MSVD TO INDIVIDUALLY EVALUATE THE EFFECTIVENESS OF WORD POLISHING AND GRAMMAR POLISHING

Methods/Metrics	B@4	R	M	C
S2VT+RL [3]	45.6	69.0	32.9	80.6
S2VT+RL+WDN	49.1	71.2	34.1	86.0
S2VT+RL+GCN	46.5	70.2	33.3	82.2
S2VT+RL+DRPN	**49.2**	**71.5**	**34.2**	**86.4**

TABLE VI
COMPARISON BETWEEN THE MODEL WITH THE CONCATENATED STATE AND THE MODEL WITH THE GATED STATE ON MSVD

Methods/Metrics	B@4	R	M	C
S2VT+RL+WDN (concatenated)	47.4	71.0	33.7	83.6
S2VT+RL+WDN (gated)	**49.1**	**71.2**	**34.1**	**86.0**

FCVC-CF-IA [22] on METEOR for MSVD. For MSR-VTT, our DRPN achieves the highest scores of BLEU4, ROUGE-L and METEOR. When we use better candidates generated by Guo *et al.* [39], our DRPN also obtains the highest CIDEr score. Both being multiple-pass decoding-based methods, our model outperforms Ruminant Decoding [39] on most metrics using less information. In addition to the caption candidates, the hidden state of the previous decoder is employed in [39], so Ruminant Decoding [39] is model-relevant. Our model can be further integrated into [39] to improve its performance. One advantage of our DRPN is that the long-term reward is adopted, so we obtain better performance on most metrics, where both local and global similarities are measured. Since we consider ROUGE-L as the reward function that is directly optimized during training, the highest score of ROUGE-L is obtained by our DRPN.

D. Effectiveness of Components and Parameters

In this section, we perform ablation studies to evaluate the components and parameters in DRPN.

GCN and WDN: We evaluate the WDN and GCN to estimate the effectiveness of word revision and grammar revision. In Table V, we compare the GCN, WDN and DRPN using the caption candidates generated by S2VT+RL on MSVD. The results show that the word-denoising model and grammar-checking model are both important for caption polishing, and we achieve the best performance by incorporating both models. As shown in the table, the word revision results in an improvement of 3.5%, 2.2%, 1.2% and 5.4%; the grammar revision results in an improvement of 0.9%, 1.2%, 0.4% and 1.6%; after revising both word errors and grammar errors, we obtain the most significant improvement on MSVD.

Concatenated state and gated state: We evaluate the model with different states to verify which state is better suited for caption polishing. The comparison between WDN with concatenated state and WDN with gated state on MSVD is shown in Table VI, where the number of hidden state is set to 1024 for both. The model with the gated state clearly performs better, i.e., a gain of 1.7%, 0.2%, 0.4% and 2.4% is achieved on the four metrics by using the gated state instead of the concatenated state. The reason is that the gating operation is beneficial for fusing one information into other information [60], so we can obtain the candidate-related state with 1024 dimensions in this paper. However, for the model with the concatenated state, the simple concatenation operation results in a 1024-dimensional fusion state, where only the latter 512-dimensional vector is related to the current candidate, and the former 512-dimensional vector is related to the visual information and information of the previous word.

Model with different training strategies: To measure the impact of training strategies on the final captioning performance, we report the result in Table VIII when choosing different training strategies to optimize the parameters of WDN. As expected, all three models trained by both the warm-start phase and model-refinement phase outperform the model that is only trained by the warm-start phase (i.e.,WDN-XE). The reason is that WDN-XE maximizes the current intermediate reward, where the model distribution achieves a local maximum with respect to the cross-entropy objective, while other RL-based methods aim at maximizing the long-term delayed reward. Among the three RL-based methods, our proposed WDN-XE-RL-Avg performs best on most metrics. The reason is that the caption sentences are diverse, and there may be more than one right words at each position, where WDN-XE-RL-Max trends to select "the most correct" word that maximizes the reward, and WDN-XE-RL-SCST somewhat alleviates the problem by greedy sampling, while WDN-XE-RL-Avg encourages to improve the gradient of all the actions that improve the reward. For this reason, our model is optimized by RL with average baseline (i.e., WDN-XE-RL-Avg) by default.

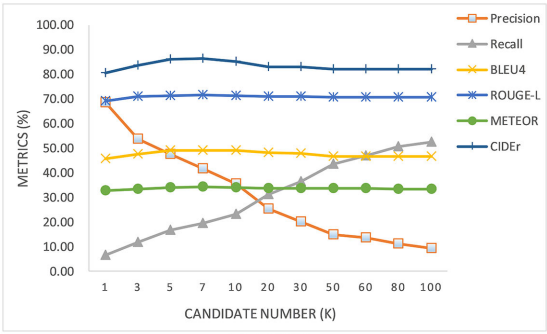
Training on different metrics: As aforementioned, we can use any metrics as the reward function to train the model. To measure the impact of reward function on the final captioning performance, we report the result in Table IX when choosing different metrics as the reward function to train the WDN. In general, we can see that optimizing for a given metric during training leads to the best performance on that metric in testing, where METEOR is an exception, but it is very close to the corresponding optimal value. Among the four widely used metrics, optimizing ROUGE-L can achieve the best result, which considerably lifts the performance of all other metrics. Thus, we select ROUGE-L as the reward function by default in this paper.

Impact of the revision time: We measure the impact of revision time on the final captioning performance by running DRPN with different number of polishing operations on MSVD as shown in Table VII, where each iteration consists of one word revision and one grammar revision. The revision time has a very slight influence on captioning performance, and after more than 2 times polishing, the result no longer improves. The reason is that the first grammar revision adds several new words into the candidate by word transformation for the next word revision; then, the candidate is almost fixed. Therefore, we only conduct one word revision and one grammar revision (i.e., iter=1) by default in this paper.

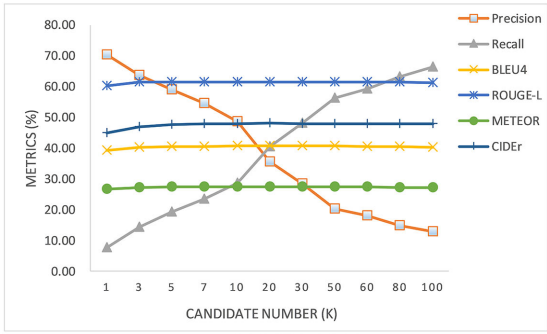
Number of caption candidates: Finally, we evaluate the parameters in DRPN, i.e., the number of caption candidates. As

TABLE VII
EVALUATION OF THE PERFORMANCE WHEN CONDUCTING DIFFERENT NUMBERS OF POLISHING OPERATIONS ON MSVD

Metrics / Iterations	Iter=1	Iter=2	Iter=3	Iter=4	Iter=5	Iter=6	Iter=7	Iter=8	Iter=9	Iter=10
BLEU4	49.2	49.3	49.3	49.3	49.3	49.3	49.3	49.3	49.3	49.3
ROUGE-L	71.5	71.5	71.5	71.5	71.5	71.5	71.5	71.5	71.5	71.5
METEOR	34.2	34.2	34.2	34.2	34.2	34.2	34.2	34.2	34.2	34.2
CIDEr	86.4	86.9	86.9	86.9	86.9	86.9	86.9	86.9	86.9	86.9



(a) MSVD



(a) MSR- VTT

Fig. 6. Evaluation of our DRPN using different numbers of candidates for caption polishing on MSVD and MSR-VTT.

TABLE VIII
COMPARISON OF THE MODEL WITH DIFFERENT TRAINING STRATEGIES ON MSVD

Methods/Metrics	B@4	R	M	C
WDN-XE	47.3	70.7	33.7	83.9
WDN-XE-RL-SCST	**49.3**	71.1	**34.1**	85.8
WDN-XE-RL-Max	48.4	70.9	34.0	85.4
WDN-XE-RL-Avg	49.1	**71.2**	**34.1**	**86.0**

TABLE IX
COMPARISON OF WDN WITH DIFFERENT METRICS AS THE REWARD FUNCTION ON MSVD

Training Metric	BLEU-4	ROUGE-L	METEOR	CIDEr
BLEU-4	**50.5**	70.6	33.6	83.3
ROUGE-L	49.1	**71.2**	**34.1**	86.0
METEOR	48.1	71.0	34.0	84.6
CIDEr	48.3	71.1	33.8	**86.8**

shown in Fig. 6, we measure the impact of the number of candidates on the captioning performance by running DRPN with different values of K. Additionally, we report the precision of word and the recall of word, where the precision estimate is the percentage of word candidates A that also occur in the ground-truth captions G, and the recall estimate is the percent of the ground-truth words G that also occur in the word candidates A.

$$\text{precision} = \frac{|A \cap G|}{|A|}, \text{recall} = \frac{|A \cap G|}{|\text{G}|} \tag{13}$$

The results show the following: 1) When the number of candidates increases, the precision of word decreases, and the recall of word increases correspondingly. Extracting more candidates clearly results in covering more words and simultaneously introducing more noises. In generally, the high precision contributes to generate accurate captions and high recall tends to generate long and diverse captions, so we should find a suitable value of K to balance them. 2) The performance is relatively insensitive to the changes in number of candidates. Within a certain range, regardless of the change in value of K, all results are improved by the polishing operations, compared to the original captions (i.e., $K = 1$). For MSVD, the change range is [46.5, 49.2] on BLEU4, [33.5, 34.2] on METEOR, [70.6, 71.5] on ROUGE-L, and [81.9, 86.4] on CIDEr. For MSR-VTT, the change range is [40.2, 40.8] on BLEU4, [27.3, 27.5] on METEOR, [61.2, 61.5] on ROUGE-L, and [47.0, 48.0] on CIDEr, respectively. Even so, it is still hopeful to find a suitable value of K, since an excessively small K results in a small action space, while an excessively large K introduces too much noise. To balance the four metrics, we employ the caption candidates with the size of 7 for MSVD and 20 for MSR-VTT by default in this paper.

E. DRPN is Model-Irrelevant

In addition, another advantage of our proposed model is that DRPN is model-irrelevant, and we adopt two methods to better demonstrate this result. On the one hand, we employ several captioning models [1], [3], [20], [39], [60]–[62] as the baselines and use our model to polish their generated caption sentences for video captioning. We even conduct polishing on [63], [64] for image captioning. On the other hand, we artificially design some caption candidates as the baselines by adding noises to the ground-truth captions and subsequently use our DRPN to polish them. These results are shown in Table X, Table XI and Table XII, where "GT+30% noises" indicates that we add 30% noises into the ground-truth captions by randomly selecting noisy words to replace the ground-truth words. The score of CIDEr is quite low when we add too many noises because the random replacement breaks down the dependency in caption sentences.

TABLE X
EVALUATION OF DRPN USING DIFFERENT CAPTIONING MODELS TO GENERATE CANDIDATES FOR THE VIDEO CAPTIONING TASK ON MSVD

Methods / Metrics	B@4	R	M	C
S2VT [1]	42.8	68.7	32.5	75.0
S2VT + DRPN	45.3	69.4	32.6	75.8
S2VT+RL [3]	45.6	69.0	32.9	80.6
S2VT+RL+ DRPN	49.2	71.5	34.2	**86.4**
hLSTMat [20]	53.0	70.3	33.5	73.8
hLSTMat + DRPN	**57.3**	**72.0**	**34.3**	78.3
SCN [60]	51.1	70.6	33.5	77.7
SCN [60] + DRPN	56.3	71.7	34.1	81.6
E2E [61]	48.0	70.3	33.4	83.7
E2E [61] + DRPN	48.6	71.0	33.9	85.2
GT+40% noises	27.1	60.3	28.3	44.5
GT+40% noises + DRPN	59.4	78.6	37.7	110.4
GT+30% noises	43.0	71.3	35.4	67.0
GT+30% noises + DRPN	63.1	80.6	39.9	121.8
GT+20% noises	54.3	77.6	40.5	83.5
GT+20% noises + DRPN	68.8	82.8	42.2	133.6
GT+10% noises	66.7	83.6	44.9	108.7
GT+10% noises + DRPN	**77.0**	**85.3**	**45.2**	**150.1**

TABLE XI
EVALUATION OF DRPN USING DIFFERENT CAPTIONING MODELS TO GENERATE CANDIDATES FOR THE VIDEO CAPTIONING TASK ON MSR-VTT

Methods / Metrics	B@4	R	M	C
S2VT [1]	35.3	57.8	26.6	40.7
S2VT + DRPN	35.5	58.5	26.9	42.2
S2VT+RL [3]	39.2	60.3	26.7	44.8
S2VT+RL + DRPN	**40.8**	**61.5**	27.5	48.0
RecNet$_{global}$ [62]	37.4	58.0	25.5	40.0
RecNet$_{global}$ [62] + DRPN	38.1	59.1	25.7	40.6
Ruminant Decoding [39]	38.8	60.5	27.1	49.0
Ruminant Decoding [39]+DRPN	39.5	61.0	**27.7**	**49.2**

TABLE XII
EVALUATION OF DRPN USING DIFFERENT CAPTIONING MODELS TO GENERATE CANDIDATES FOR THE IMAGE CAPTIONING TASK ON MSCOCO

Methods / Metrics	B@4	R	M	C
Hard-Attention [63]	25.0	46.9	23.0	69.6
Hard-Attention [63] + DRPN	26.2	48.0	23.9	72.9
Up-Down [64]	35.9	56.2	26.9	111.5
Up-Down [64] + DRPN	37.1	58.9	27.6	115.0

The comparisons demonstrate two points: 1) DRPN is model-irrelevant and can be integrated into any video captioning networks and even image captioning networks to improve their performance. The tables clearly show that the performances of S2VT [3], S2VT+RL [1], SCN [60], E2E [61], RecNet$_{global}$ [62], Ruminant Decoding [39] and hLSTMat [20] are improved by appending our DRPN for video captioning. In addition to video captioning, our DRPN can be integrated into image captioning models to refine their generated caption sentences, such as Hard-Attention [63] and Up-Down [64]. 2) The performance of DRPN has an obviously positive correlation with the quality of caption candidates. Intuitively, better caption candidates correspond to better captioning performance. For example, S2VT+RL [3] and E2E [61] perform better than S2VT [1], and S2VT+RL [3] and E2E [61] also have better results after the polishing procedure. Therefore, we achieve better performance when we use a better baseline model to obtain better candidates, and we can achieve the state-of-the-art performance simply by using our DRPN to refine the caption sentences generated by the state-of-the-art methods. In addition, the improvement between S2VT+RL+DRPN and S2VT+RL is larger than that of S2VT+DRPN and S2VT. The reason is that S2VT+RL can generate better candidates with more correct cues for further polishing and provide a larger margin for improvement for our DRPN. As shown in Table XIII, we compare the original result, revised result and optimal result that can be reached after revision in the ideal situation when we use S2VT and S2VT+RL to generate candidates on MSVD. Thus, the optimal result of S2VT+RL+DRPN is clearly better, and the margin between S2VT+RL and S2VT+RL+DRPN (optimal) for improvement is also larger. Another reason is that we use the candidates generated by S2VT+RL to train DRPN in this paper, and we do not fine-turn the model with other candidates; thus, DRPN can best use the information provided by S2VT+RL.

TABLE XIII
COMPARISON OF THE ORIGINAL RESULT, REVISED RESULT, AND OPTIMAL RESULT WHEN USING S2VT AND S2VT+RL TO GENERATE CANDIDATES ON MSVD

Models / Metrics	B@4	R	M	C
S2VT [1]	42.8	68.7	32.5	75.0
S2VT+DRPN	45.3	69.4	32.6	75.8
S2VT+DRPN (optimal)	64.8	84.4	42.4	105.8
S2VT+RL [3]	45.6	69.0	32.9	80.6
S2VT+RL+DRPN	49.2	71.5	34.2	86.4
S2VT+RL+DRPN (optimal)	69.5	85.9	43.5	118.0

F. Qualitative Results

Fig. 7 shows several captioning examples generated by S2VT+RL, our S2VT+RL+DRPN and human annotation. The top shows examples of word revision, and the bottom shows examples of grammar revision. From these examples, both S2VT+RL and S2VT+RL+DRPN can clearly generate somewhat relevant caption sentences for videos. After the polishing procedure, our S2VT+RL+DRPN can generate sentences with more accurate keywords by incorporating the word-denoising network to revise the word errors and sentences with more accurate grammar by incorporating the grammar-checking network to revise the grammar errors. For example, our DRPN revises the word error in the first video by replacing the incorrect noun "fish" with "shrimp". Similarly, the noun "cat" instead of "kitten" is used with more precision in the fifth video. Compared to the verb "cleaning" generated by S2VT+RL, the verb "vacuuming" generated by our S2VT+RL+DRPN is more detailed and precise to describe the video content in the third video. For example, our DRPN revises the grammar errors in the last three videos by replacing the incorrect term "a" with "an," "shoe" with "shoes" and "men" with "man". Certainly, the captions annotated by human more comprehensively and naturally describe the video, which shows that we need further works to reach the goal of automatic video captioning.

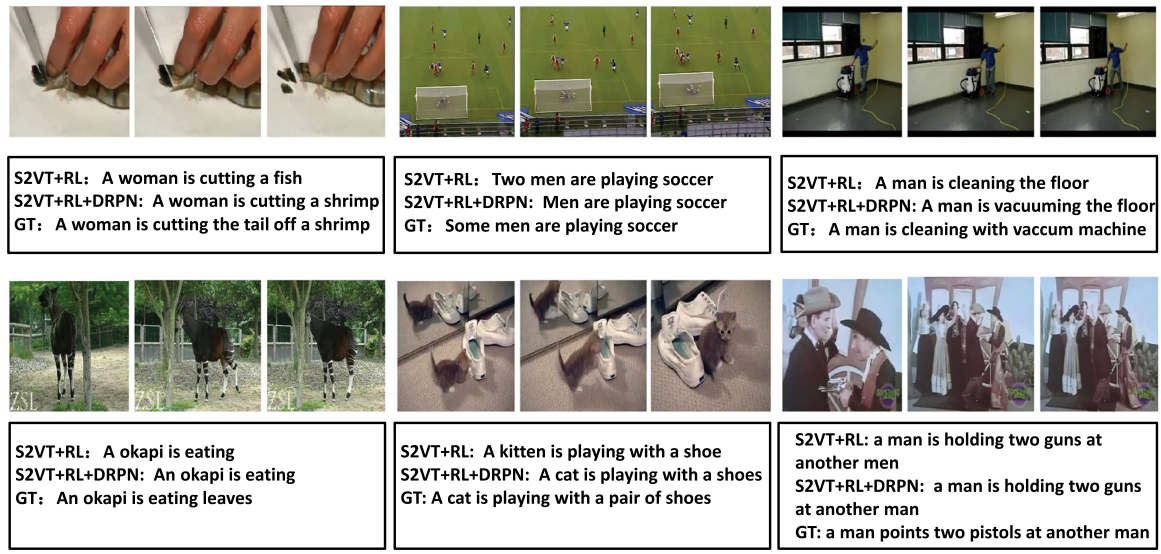


Fig. 7. Visualization of some video captioning examples on MSVD, including the examples generated by the original S2VT+RL, our S2VT+RL+DRPN, and human annotation. The top shows some examples of word revision, and the bottom shows some examples of grammar revision.

V. Conclusion

In this paper, by simulating the proofreading procedure of human beings, we propose a novel deep reinforcement polishing network to improve the performance of video captioning by iteratively polishing the generated caption sentences. For better long-sequence generation, the long-term reward in deep reinforcement learning is adopted to directly optimize the global quality of caption sentences. To reduce the semantic gap between the visual domain and the language domain, the caption candidate is considered an additional cue for video captioning, which is gradually updated by revising the word errors and grammar errors. The experiments on MSVD and MSR-VTT demonstrate that our DRPN can indeed improve the result of video captioning to achieve comparable and even better performance than the state-of-the-art methods. Our DRPN is model-irrelevant, which can be integrated into any captioning models to refine their generations. Certainly, some limitations remain for future works. For example, we will attempt to conduct more high-level polishing instead of the word-level polishing and conduct grammar polishing with POS tagging instead of the POS-irrelevant polishing.

References

[1] S. Venugopalan *et al.*, "Sequence to sequence-video to text," in *Proc. IEEE Int. Conf. Comput. Vision*, Santiago, Chile, Dec. 2015, pp. 4534–4542.
[2] X. Li *et al.*, "Learnable aggregating net with diversity learning for video question answering," in *Proc. ACM Int. Conf. Multimed.*, Nice, France, Oct. 2019, pp. 1166–1174.
[3] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 7008–7024.
[4] N. Aafaq, A. Mian, W. Liu, S. Z. Gilani, and M. Shah, "Video description: A survey of methods, datasets, and evaluation metrics," *ACM Comput. Surv.*, vol. 52, no. 6, pp. 1–37, Oct. 2019.
[5] S. Chen, T. Yao, and Y.-G. Jiang, "Deep learning for video captioning: A review," in *Proc. Int. Joint Conf. Artif. Intell.*, Macao, China, Aug. 2019, pp. 6283–6290.
[6] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R. Mooney, "Integrating language and vision to generate natural language descriptions of videos in the wild," in *Proc. Int. Conf. Comput. Linguist.*, Dublin, Ireland, Aug. 2014, pp. 1218–1227.
[7] A. Barbu *et al.*, "Video in sentences out," 2012, *arXiv:1204.2742*.
[8] M. U. G. Khan, L. Zhang, and Y. Gotoh, "Human focused video description," in *Proc. IEEE Int. Conf. Comput. Vision*, Barcelona, Spain, Nov. 2011, pp. 1480–1487.
[9] L. Yao *et al.*, "Describing videos by exploiting temporal structure," in *Proc. IEEE Int. Conf. Comput. Vision*, Santiago, Chile, Dec. 2015, pp. 4507–4515.
[10] X. Wang *et al.*, "Video captioning via hierarchical reinforcement learning," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 4213–4222.
[11] Y. Chen, S. Wang, W. Zhang, and Q. Huang, "Less is more: Picking informative frames for video captioning," in *Proc. Eur. Conf. Comput. Vision*, Munich, Germany, Sep. 2018, pp. 358–373.
[12] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles, "Dense-captioning events in videos," in *Proc. IEEE Int. Conf. Comput. Vision*, Venice, Italy, Oct. 2017, pp. 706–715.
[13] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Salt Lake City, UT, USA, Jul. 2018, pp. 8739–8748.
[14] A. Kojima, T. Tamura, and K. Fukunaga, "Natural language description of human activities from video images based on concept hierarchy of actions," *Int. J. Comput. Vision*, vol. 50, no. 2, pp. 171–184, Nov. 2002.
[15] P. Hanckmann, K. Schutte, and G. J. Burghouts, "Automated textual descriptions for a wide range of video events with 48 human actions," in *Proc. Eur. Conf. Comput. Vision*, Florence, Italy, Oct. 2012, pp. 372–380.
[16] M. Rohrbach *et al.*, "Translating video content to natural language descriptions," in *Proc. IEEE Int. Conf. Comput. Vision*, Sydney, NSW, Australia, Dec. 2013, pp. 433–440.
[17] W. Zhang, B. Wang, L. Ma, and W. Liu, "Reconstruct and represent video contents for captioning via reinforcement learning," 2019, *arXiv:1906.01452*.
[18] J. Chen *et al.*, "Temporal deformable convolutional encoder-decoder networks for video captioning," 2019, *arXiv:1905.01077*.
[19] S. Venugopalan *et al.*, "Translating videos to natural language using deep recurrent neural networks," 2014, *arXiv:1412.4729*.

[20] L. Gao, X. Li, J. Song, and H. T. Shen, "Hierarchical LSTMs with adaptive attention for visual captioning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1112–1131, May 2020.

[21] C. Hori *et al.*, "Attention-based multimodal fusion for video description," in *Proc. IEEE Int. Conf. Comput. Vision*, Venice, Italy, Oct. 2017, pp. 4193–4202.

[22] K. Fang *et al.*, "Fully convolutional video captioning with coarse-to-fine and inherited attention," in *Proc. 33th AAAI Conf. Artif. Intell.*, Honolulu, HI, USA, Jan. 2019, pp. 8271–8278.

[23] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, "Video paragraph captioning using hierarchical recurrent neural networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 4584–4593.

[24] B. Zhao *et al.*, "Video captioning with tube features," in *Proc. Int. Joint Conf. Artif. Intell.*, Stockholm, Sweden, Jul. 2018, pp. 1177–1183.

[25] R. Pasunuru and M. Bansal, "Reinforced video captioning with entailment rewards," 2017, *arXiv:1708.02300*.

[26] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng, "StyleNet: Generating attractive visual captions with styles," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 3137–3146.

[27] Y. Yang *et al.*, "Video captioning by adversarial LSTM," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5600–5611, Nov. 2018.

[28] X. Liang, Z. Hu, H. Zhang, C. Gan, and E. P. Xing, "Recurrent topic-transition GAN for visual paragraph generation," in *Proc. IEEE Int. Conf. Comput. Vision*, Venice, Italy, Oct. 2017, pp. 3362–3371.

[29] C. Chen *et al.*, "Improving image captioning with conditional generative adversarial nets," 2018, *arXiv:1805.07112*.

[30] B. Dai, S. Fidler, R. Urtasun, and D. Lin, "Towards diverse and natural image descriptions via a conditional GAN," in *Proc. IEEE Int. Conf. Comput. Vision*, Venice, Italy, Oct. 2017, pp. 2970–2979.

[31] M. Yang *et al.*, "Multitask learning for cross-domain image captioning," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 1047–1061, Apr. 2018.

[32] X. Duan *et al.*, "Weakly supervised dense event captioning in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, PQ, Canada, Dec. 2018, pp. 3059–3069.

[33] Y. Li, T. Yao, Y. Pan, H. Chao, and T. Mei, "Jointly localizing and describing events for dense video captioning," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Salt Lake City, UT, USA, Jul. 2018, pp. 7492–7500.

[34] Z. Shen *et al.*, "Weakly supervised dense video captioning," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 1916–1924.

[35] J. Mun, L. Yang, Z. Ren, N. Xu, and B. Han, "Streamlined dense video captioning," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 6588–6597.

[36] G. Yin *et al.*, "Context and attribute grounded dense captioning," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 6241–6250.

[37] Y. Xia *et al.*, "Deliberation networks: Sequence generation beyond one-pass decoding," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 1784–1794.

[38] Z. Zhu, Z. Xue, and Z. Yuan, "Think and tell: Preview network for image captioning," in *Proc. Brit. Mach. Vis. Conf.*, Newcastle, U.K., Sep. 2018.

[39] L. Guo, J. Liu, S. Lu, and H. Lu, "Show, tell and polish: Ruminant decoding for image captioning," *IEEE Trans. Multimedia*, to be published.

[40] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. Adv. Neural Inf. Process. Syst.*, Denver, CO, USA, Nov. 2000, pp. 1057–1063.

[41] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. 31th AAAI Conf. Artif. Intell.*, San Francisco, CA, USA, Feb. 2017, pp. 4278–4284.

[42] D. L. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proc. Annu. Meeting Assoc. Comput. Linguistics: Human Lang. Technol.*, Portland, USA, Jun. 2011, pp. 190–200.

[43] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Las Vegas, USA, Jun. 2016, pp. 5288–5296.

[44] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. USENIX Symp. Oper. Syst. Des. Implement.*, Savannah, GA, USA, Nov. 2016, pp. 265–283.

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[46] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jun. 2014.

[47] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, Philadelphia, CA, USA, Jul. 2002, pp. 311–318.

[48] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. ACL Workshop Text Summarization Branches Out*, Baltimore, MD, USA, Jun. 2004, pp. 74–81.

[49] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, Beijing, China, Jul. 2005, pp. 65–72.

[50] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 4566–4575.

[51] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly modeling embedding and translation to bridge video and language," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Las Vegas, CA, USA, Jun. 2016, pp. 4594–4602.

[52] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based LSTM and semantic consistency," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 2045–2055, Sep. 2017.

[53] J. Xu, T. Yao, Y. Zhang, and T. Mei, "Learning multimodal attention LSTM networks for video captioning," in *Proc. ACM Int. Conf. Multimed.*, Mountain View, CA, USA, Oct. 2017, pp. 537–545.

[54] J. Song *et al.*, "Hierarchical LSTM with adjusted temporal attention for video captioning," in *Proc. Int. Joint Conf. Artif. Intell.*, Melbourne, FL, Australia, Aug. 2017, pp. 2737–2743.

[55] W. Li, D. Guo, and X. Fang, "Multimodal architecture for video captioning with memory networks and an attention mechanism," *Pattern Recogn. Lett.*, vol. 105, pp. 23–29, Apr. 2018.

[56] X. Long, C. Gan, and G. de Melo, "Video captioning with multi-faceted attention," *IEEE Trans. Assoc. Comput. Linguistics*, vol. 6, pp. 173–184, Mar. 2018.

[57] W. Hao, Z. Zhang, H. Guan, and H. Guan, "Integrating both visual and audio cues for enhanced video caption," in *Proc. 32th AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, Feb. 2018, pp. 6894–6901.

[58] J. Wang, W. Wang, Y. Huang, L. Wang, and T. Tan, "M3: Multimodal memory modelling for video captioning," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7512–7520.

[59] X. Wu, G. Li, Q. Cao, Q. Ji, and L. Lin, "Interpretable video captioning via trajectory structured localization," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Salt Lake City, UT, USA, Jul. 2018, pp. 6829–6837.

[60] Z. Gan *et al.*, "Semantic compositional networks for visual captioning," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 5630–5639.

[61] L. Li and B. Gong, "End-to-end video captioning with multitask reinforcement learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Waikoloa Village, HI, USA, Jan. 2019, pp. 339–348.

[62] B. Wang, L. Ma, W. Zhang, and W. Liu, "Reconstruction network for video captioning," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7622–7631.

[63] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, Lile, France, Jul. 2015, pp. 2048–2057.

[64] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Salt Lake City, UT, USA, Jul. 2018, pp. 6077–6086.