

End-to-End Video Captioning with Multitask Reinforcement Learning

Lijun Li*
Beihang University

lilijun1990@buaa.edu.cn

Boqing Gong
Tencent AI LAB

boqinggo@outlook.com

Abstract

Although end-to-end (E2E) learning has led to impressive progress on a variety of visual understanding tasks, it is often impeded by hardware constraints (e.g., GPU memory) and is prone to overfitting. When it comes to video captioning, one of the most challenging benchmark tasks in computer vision, those limitations of E2E learning are especially amplified by the fact that both the input videos and output captions are lengthy sequences. Indeed, state-of-the-art methods for video captioning process video frames by convolutional neural networks and generate captions by unrolling recurrent neural networks. If we connect them in an E2E manner, the resulting model is both memory-consuming and data-hungry, making it extremely hard to train. In this paper, we propose a multitask reinforcement learning approach to training an E2E video captioning model. The main idea is to mine and construct as many effective tasks (e.g., attributes, rewards, and the captions) as possible from the human captioned videos such that they can jointly regulate the search space of the E2E neural network, from which an E2E video captioning model can be found and generalized to the testing phase. To the best of our knowledge, this is the first video captioning model that is trained end-to-end from the raw video input to the caption output. Experimental results show that such a model outperforms existing ones to a large margin on two benchmark video captioning datasets.¹

1. Introduction

Video captioning, i.e., to automatically describe videos by full sentences or phrases, not only serves as a challenging testbed in computer vision and machine learning but also benefits many real-world applications. The automatically generated video captions may enable fast video retrieval, assist the visually impaired, and engage users in a versatile chatbot, to name a few.

*Work done in Tencent.

¹Code available at <https://github.com/adwardlee/multitask-end-to-end-video-captioning>

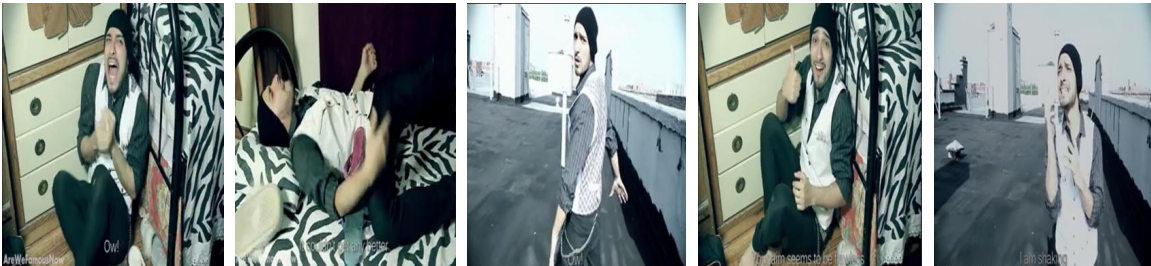
Most recent works [41, 2, 21, 5] that tackle this problem fall under an encoder-decoder framework which has been shown effective in speech recognition [8, 38], natural language translation [23, 13], and image captioning [48, 25]. The encoder extracts compact representations of the visual content. In the context of video captioning, the convolutional neural networks (CNNs) are usually used to encode the video frames followed by a temporal model [21, 40, 26, 51] or simply temporal pooling [42] and the decoder maps the codes to a sequence of words often by the recurrent neural networks (RNNs) [32, 44] (e.g., the long short-term memory (LSTM) [20] units are a popular choice). In order to train such networks, most existing works employ a cross-entropy loss at each decoding step. We refer the readers to the seminal work that spurs the resurging interests in video captioning, sequence to sequence - video to text (S2VT) [41], for a quick understanding about the backbone techniques.

Despite the impact of the encoder-decoder framework on video captioning, it inherently impedes the use of end-to-end (E2E) training which has led to very impressive results on a large variety of tasks. Indeed, both CNNs and RNNs are memory consuming, leaving little GPU space to the training data which are yet key to the training procedure. Besides, the input videos and output sentences are both sequences, making the encoder-decoder framework very lengthy and data-hungry. On the one hand, it is tempting to explore the E2E training strategy on the video captioning task. On the other hand, this seemingly straightforward idea is confined by the hardware and the relatively small size of existing video captioning datasets. Our experiments show that the conventional cross-entropy loss coupled with stochastic gradient descent cannot effectively exploit the E2E training.

In this paper, we propose a multitask reinforcement learning approach to training a video captioning model in an E2E manner. Our main idea is to mine and construct as many effective tasks as possible from the human captioned videos such that they can jointly regulate the search space of the encoder-decoder network, from which an E2E video captioning model can be found and generalized to the test-



Caption: a man is playing a guitar and singing
 Caption: a man is singing and playing guitar in an airport



Caption: a man is singing and doing funny act
 Caption: a man making a music video and having slippers thrown at him

Figure 1. Exemplar video frames and user captions in the MSVD and MSR-VTT datasets.

ing phase. The auxiliary tasks consist of two broad types: to predict the attributes extracted from the captions of the training videos and to maximize the rewards defined from the reinforcement learning perspective. When the training set is relatively small for the big encoder-decoder network, it is important to mine as much supervision as possible from the limited data so that it helps reduce the search space for the main task of interest.

Although many existing video captioning models [41, 27, 54] can literally be trained in the E2E fashion, none of them were probably due to the hardware constraint and concerns on overfitting. Indeed, our study reveals that, without the proposed multitask reinforcement learning strategy, E2E learning is easy to overfit the training set. This work is the first to end-to-end train a model for video captioning, to the best of our knowledge. It is nontrivial because the model becomes very large in order to take as input a raw video sequence and output a sequence of words, causing challenges to the computational resources and raising the need for large-scale well-labeled data. Our multitask reinforcement learning method is able to alleviate those challenges and gives rise to state-of-the-art results on MSVD [9] and MSR-VTT [47], two popular benchmark datasets for the video captioning task. Nonetheless, we believe that, supplied with larger-scale labeled data, the E2E training can further advance the video captioning results.

We summarize our contribution as the following. (1) We propose a multitask reinforcement training strategy which can effectively learn a video captioning model in an E2E fashion under the current constraints of hardware and data size. (2) We extract attributes from the captions of the training videos and define rewards upon the captions without using any external data. (3) Experiments show that our approach with a single model gives rise to state-of-the-art results on both the MSVD and MSR-VTT datasets.

2. Related works

Large amount of progress has been made in image and video captioning. A large part of it is due to the advances in machine translation. For example, the encoder-decoder framework and the attention mechanism were first introduced in machine translation [3, 12, 36] and then extended to captioning. Both image captioning approaches [48, 52, 10] and video captioning methods [50, 21, 53, 29] follow their pipeline and also apply attention mechanism in caption generation. Comparing with image captioning, video captioning describes dynamic scenes instead of static scenes. From Figure 1, we can clearly see that the video captioning is much more difficult with large variance in appearance. Baraldi et al. [5] propose boundary-aware LSTM cell to automatically detect the temporal video segments. Venugopalan et al. [40] integrate natural language knowledge

to their network by training language LSTM model on a large external text corpora. Zhu et al. [57] extend Gated Recurrent Unit (GRU) to multirate GRU to handle different video frame rates. Hendricks et al. [2] propose a deep compositional captioner to describe novel object with the help of lexical classifier training on external image description dataset.

In the recent years, maximum likelihood estimation algorithm has been widely used in video captioning which maximizes the probability of current words based on the previous ground truth words [17, 16, 33, 27, 42]. But they all have two major problems.

The first one is exposure bias which is the input mismatch in training and inference. In training, the output of decoder depends on ground truth words instead of model predictions. While in inference, the decoder only has access to the predictions. Bengio et al. [6] proposed scheduled sampling to mitigate the gap between the training and inference by selecting more often from the ground truth in the beginning but sampling more often from the model predictions in the end. However, it still optimizes at the word level.

The other problem is the objective mismatch between training and inference. In training, it optimizes the loss at the word level. While in inference, discrete metrics such as BLEU4 [28], METEOR [4], CIDEr [39], and ROUGE-L [24] are used for evaluation. A few image captioning works have been proposed to solve the problems and shown superior performance with the help of reinforcement learning. Ren et al. [30] introduce actor-critic method to image captioning and also propose a new lookahead inference algorithm which has better performance than beam search. Liu et al. [25] employ policy gradient method to optimize the SPIDER score. Dai et al. [15] combine a conditional generative adversarial network with policy gradient which can produce natural and diverse sentences. However, there are much less works using reinforcement learning in video captioning.

In this paper, we exploit the reinforcement learning in video captioning, especially for the jointly training of CNNs and LSTMs. Note that many video captioning models can actually be deployed in an end-to-end manner, such as [41, 27, 54], etc. Venugopalan et al. propose a stack of two LSTM networks [41]. Pan et al. propose a novel transfer unit to feed the semantic concept to LSTM [27]. Yu et al. develop a high-level word detector and semantic attention mechanism which combines the concept with caption decoder [54]. However, they actually treat CNN as feature extractor and do not train the CNN part of their framework. On the contrary, our method trains the CNN and the other part together.

Multitask learning is a kind of machine learning technique. During multitask learning, multiple tasks are solved

at the same time with a shared representation and is especially useful with limited number of original data. It has been widely utilized not only in computer vision [43, 55, 49, 18], but also in natural language processing [14]. It becomes a natural choice for us since the model capacity likely outweighs the existing datasets when we aim to update all its weights from the raw video input to the caption output. However, few works use multitask learning in video captioning. We explore the effectiveness and find the multitask learning can also be useful in video captioning.

3. An E2E trained video captioning model

We describe the end-to-end (E2E) trained video captioning model in this section. It is essentially a deepened version of the S2VT model [41]. Despite its simplicity in concept, it is very challenging to train the whole big model to reach a good generalization capability onto the test sets. Both our experiments and an earlier attempt by Yu et al. [54] indicate that the gain of jointly training the CNNs and LSTMs is only marginal over fixing the CNNs as feature extractors, if we do not have an effective training approach. To this end, one important contribution of this paper is the batch of techniques presented below which we find useful when they are combined for training the E2E video captioning model.

3.1. Model architecture

Figure 2 sketches the model architecture which consists of three main components. On the top, five copies of the same Inception-Resnet-v2 [37] CNN are used to transform the raw video frames to high-level feature representations. Note that the last classification layer of the Inception-Resnet-v2 is replaced by a fully connected layer whose output dimension is 500. The LSTMs on the bottom first encode the video frames' feature representations and then decode a sentence to describe the content in the video. On the bottom left, there is a branch consisting of a temporal average pooling layer and an attribute prediction layer. We extract up to 400 attributes in our experiments. Accordingly, the attribute prediction layer's output dimension is 400 and the activation functions are sigmoid. This branch is introduced to assign relevant attributes to an input video. It is not used in the testing phase of the video captioning, but it generates informative gradients in the training phase for updating the weights of the CNNs in addition to those from the LSTMs. The design of the LSTMs (e.g., the number of hidden units, how to compute the input gates, etc.) is borrowed from S2VT [41].

3.2. The E2E training of the model

We train the model progressively in three steps. The first two steps aim to find a good initialization to the LSTMs (and the fully connected layer connecting the CNNs and

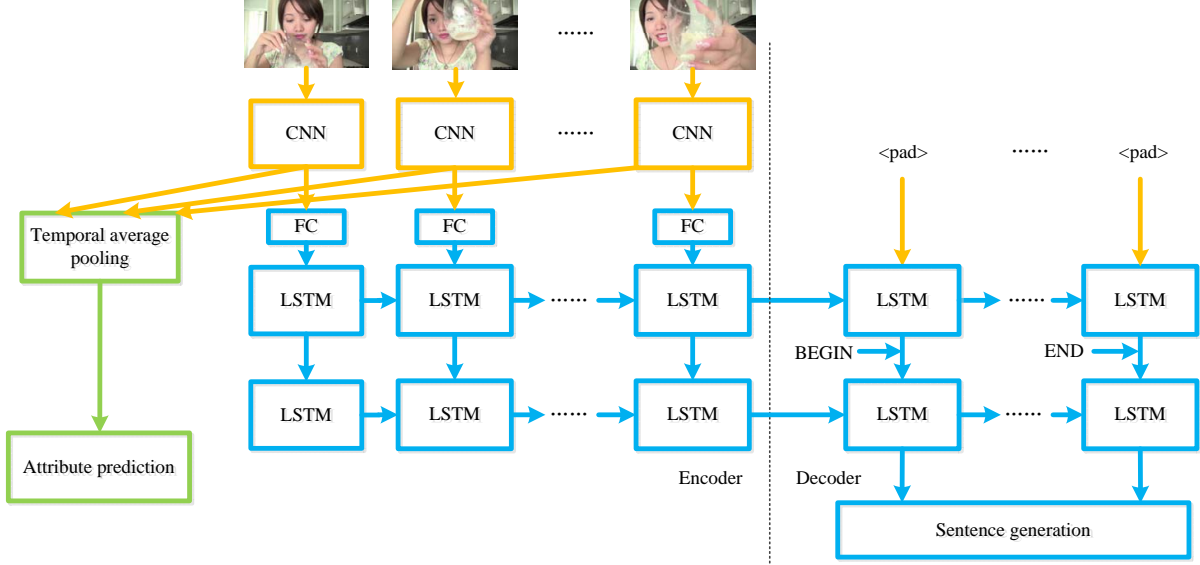


Figure 2. The multitask reinforcement learning framework for our E2E training of video captioning models.

the LSTMs) such that the last step, the E2E training of the whole model, can have a warm start. The weights of the CNNs are frozen until the third step.

Step 1 is the standard training approach to S2VT using the cross-entropy loss. For an input frame I_t at time step t , we encode it with the deep CNN and embed it with projection matrix W_I . Then for the projected feature representation x_t , the LSTM computes the hidden state h_t and cell state c_t . The details about the computation of hidden state and cell state are in the following:

$$\begin{aligned}
 i_t &= \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \\
 f_t &= \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f) \\
 o_t &= \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \\
 g_t &= \phi(W_{gx}x_t + W_{gh}h_{t-1} + b_g) \\
 c_t &= i_t \odot g_t + f_t \odot c_{t-1} \\
 h_t &= o_t \odot \phi(c_t)
 \end{aligned} \tag{1}$$

where σ is the sigmoid function, ϕ is the hyperbolic tangent function, \odot is element-wise multiplication. The second LSTM layer is similar to the first one, except that the input is the combination of first LSTM’s output and the word embeddings.

Given a “groundtruth” sentence $s^* = \{w_1^*, w_2^*, \dots, w_T^*\}$ describing an input video, we minimize the cross-entropy loss as follows,

$$L_x(\theta) := -\log p_\theta(s^*) = -\frac{1}{T} \sum_{t=1}^T \log p_\theta(w_t^* | w_1^*, \dots, w_{t-1}^*) \tag{2}$$

where θ denotes the model parameters.

Step 2: REINFORCE+ training of the LSTMs. After Step 1, we introduce the self-critical REINFORCE algorithm [45, 31] to the video captioning to seek better weights for the LSTMs in terms of their generalization performance on the validation and test sets.

It is well known that the cross-entropy loss exposes the recurrent LSTMs under different data distributions in the training and test stages because it feeds the model groundtruth words which are only available in training [6, 31]. Moreover, the loss function is not necessarily a good proxy for the evaluation metrics. To address these challenges, we opt to directly optimize the captioning system by REINFORCE learning as in [31]. In reinforcement learning, the goal is to train an agent to complete tasks by executing a series of actions in an environment. In the context of video captioning, the goal of the captioning model is to generate a proper sentence upon observing the video input. The captioning model corresponds to the agent and the action is to predict the next word at each time step. We can consider the input video with user annotated captions as the environment. We define the reward for the agent’s action as the actual evaluation metric used in the test stage. In particular, we use the CIDEr score as reward in this paper. Here is a brief summary of the reinforcement learning pipeline for video captioning: an agent receives an observation about the environment which contains the visual features and groundtruth words up to current step, as well as a reward (the CIDEr score) from the environment; the agent then takes an action to predict a word; the environment provides another state (revealing one more groundtruth word) and reward in response to the agent’s action.

The objective function of reinforcement learning is:

$$L_r(\theta) = -\mathbb{E}(r(w^s)) \quad (3)$$

where w^s is the sentence consisting of (w_1, w_2, \dots, w) sampled from the network and r is the reward function.

In order to solve the above problem, as in [31], we also use the REINFORCE algorithm [45]. The general updates of the parameter θ can be written as:

$$\nabla_{\theta} L_r(\theta) = -\mathbb{E}[r(w) \nabla \log p(w^s)], \quad (4)$$

where $p(w^s)$ is basically determined by the video captioning model $p_{\theta}(w^s)$ (cf eq. (2)). In practice, the expectation is approximated by a sample mean which incurs variance to the gradients. To reduce the variance, the reward r is often calibrated by a baseline b :

$$\nabla_{\theta} L_r(\theta) = -\mathbb{E}[(r(w^s) - b) \nabla_{\theta} \log p_{\theta}(w^s)], \quad (5)$$

where it is obvious that the gradient remains unchanged since the baseline b does not depend on the sampled words w^s . How to choose the baseline b can affect the performance of the REINFORCE algorithm. We choose the reward of the greedily inferred words as our baseline. Denoting by $\hat{w}_t := \arg \max p_{\theta}(w_t | h_t)$, the baseline is $r(\hat{w}^s)$.

We are now ready to describe the practical algorithm for solving eq. (3). A one-sample approximation to Eq. (5) is:

$$\nabla_{\theta} L_r(\theta) \approx -(r(w^s) - b) \nabla_{\theta} \log p_{\theta}(w^s) \quad (6)$$

which further be seen as the following cost function. At the beginning of each iteration, we sample up to M trajectories (i.e., sentences) from the current model. Denoting them by s_1, \dots, s_M , we can then write down the cost function for generating the gradients of this iteration,

$$L_r(\theta) \approx -\frac{1}{M} \sum_{m=1}^M (r(s_m) - b) \log p_{\theta}(s_m) \quad (7)$$

where $r(s_m)$ is the reward assigned to the trajectory s_m . We denote this algorithm as REINFORCE+ or RFC+ in the following.

It is interesting to note that Eq. (7) acts as a running loss over the full course of the training. It changes at different iterations, being realized by the sampled trajectories as opposed to the constant groundtruth captions in the cross-entropy loss L_x across different iterations. Moreover, the rewards offset by the baseline dynamically weigh the contributions of the trajectories to the gradients. Jointly, they push the model trained in Step 1 further to the point that generalizes better to the unseen data.

Step 3: Multitask training of the full model. We jointly tune the full model in this step, freeing the weights of the CNNs. As the starting point, it might seem natural to repeat Step 1 and/or Step 2 for the E2E optimization. However, this only gives rise to marginal gain over freezing the CNNs weights in our experiments. Such quick saturation of accuracy is actually common for very deep neural networks and may be alleviated by the skip connections between different layers of feedforward networks [19, 35]. Our model, however, heterogeneously mixes LSTMs and CNNs, leaving it unclear how to apply the skip connections.

Instead, we propose to supply extra and informative gradients directly to the CNNs, so as to complement those reached to the CNNs indirectly through the LSTMs. Such direct gradients are provided by the attribute prediction branch (cf. Figure 2).

We mine the attributes in the video captions following the previous practice on image captioning [46]. Among the words in the sentences of the training set, we extract the most frequent words including nouns, verbs and adjectives as the attributes. Accordingly, the attribute prediction branch is equipped by sigmoid functions in order to each predict the existence or not (y_i) of an attribute in the input video. We define a binary cross entropy loss for this network branch, denoted by $L_a(\theta) = -\frac{1}{N} \sum_i [y_i \log q_{\theta}(i) + (1 - y_i) \log(1 - q_{\theta}(i))]$, where N is the number of attributes in total and $q_{\theta}(i)$ is the network output for the i -th attribute.

The overall cost function we use in Step 3 is a convex combination of the attribute loss and the REINFORCE loss:

$$L(\theta) = \alpha L_r(\theta) + (1 - \alpha) L_a(\theta) \quad (8)$$

where $\alpha = 0.95$ is selected by the validation set.

4. Comparison experiments

We present extensive experimental results and ablation studies in this section.

4.1. Datasets and experiment details

In this section, we report the results of our E2E trained model and compare with other state-of-the-art methods on two popular video captioning datasets. One is the MSVD dataset [9]. MSVD consists of 1,970 video clips and 70,028 captions collected via Amazon Mechanical Turk (www.mturk.com) which covers a lot of topics. On average, the video duration is about 10 seconds and each sentence contains about 8 words. A common split of the videos is provided by [41] and maintained by the existing works as well as this paper: 1,200 videos for training, 100 for validation, and 670 for testing. The other is the MSR-VTT dataset which contains 10,000 video clips and 200,000 captions. We use the data split defined in [47] in our experiments: 6,513 videos for training, 497 for validation, and

Table 1. Comparison with state-of-the-art methods on the MSVD dataset.

Models/Metrics	BLEU4	ROUGE-L	METEOR	CIDEr
h-RNN [53]	0.499	–	0.326	0.658
Attention fusion [21]	0.524	–	0.320	0.688
BA encoder [5]	0.425	–	0.324	0.635
SCN [17]	0.502	–	<u>0.334</u>	0.770
TDDF [56]	0.458	<u>0.697</u>	<u>0.333</u>	0.730
LSTM-TSA [27]	<u>0.528</u>	–	<u>0.335</u>	0.740
MVRM [57]	0.538	–	0.344	0.812
S2VT (our Step 1) [41]	0.428	0.687	0.325	0.750
REINFORCE (our Step 2) [31]	0.456	0.690	0.329	0.806
REINFORCE+ (our Step 2) [31]	0.466	<u>0.694</u>	0.330	0.816
E2E (ours, greedy search)	0.480	0.705	<u>0.336</u>	<u>0.865</u>
E2E (ours, beam search)	0.503	0.708	0.341	0.875

Table 2. Comparison with state-of-the-art methods on the MSR-VTT dataset.

Models	BLEU4	ROUGE-L	METEOR	CIDEr
TDDF [56]	0.372	0.586	<u>0.277</u>	0.441
v2t_navigator[22]	0.408	<u>0.609</u>	0.282	0.448
Aalto [34]	<u>0.398</u>	0.598	0.269	0.457
Attention fusion [21]	<u>0.394</u>	–	0.257	0.404
S2VT (our Step 1) [41]	0.353	0.578	0.266	0.407
REINFORCE (our Step 2) [31]	0.392	0.603	0.267	0.448
REINFORCE+ (our Step 2) [31]	<u>0.398</u>	<u>0.609</u>	0.271	<u>0.468</u>
E2E (ours, greedy search)	0.404	0.610	0.270	0.483
E2E (ours, beam search)	0.404	0.610	0.270	0.483

Table 3. Ablation study: video captioning results on MSVD with greedy decoding.

Models	BLEU4	ROUGE-L	METEOR	CIDEr
S2VT (Step 1) [41]	0.428	0.687	0.325	0.750
RFC (Step 2) [31]	0.456	<u>0.690</u>	<u>0.329</u>	0.806
RFC+ (Step 2) [31]	<u>0.466</u>	<u>0.694</u>	<u>0.330</u>	<u>0.816</u>
E2E (xentropy)	0.439	<u>0.690</u>	<u>0.328</u>	0.767
E2E (att+xentropy)	0.453	<u>0.694</u>	<u>0.331</u>	0.790
E2E w/o attribute prediction	<u>0.466</u>	0.696	<u>0.332</u>	<u>0.824</u>
E2E w/o reinforcement or attribute or Step 1	0.424	0.684	0.325	0.719
E2E (ours)	0.480	0.705	0.336	0.865

Table 4. Ablation study: video captioning results on MSR-VTT dataset with greedy decoding.

Models	BLEU4	ROUGE-L	METEOR	CIDEr
S2VT (Step 1) [41]	0.353	0.578	<u>0.266</u>	0.407
RFC (Step 2) [31]	0.392	<u>0.603</u>	0.267	0.448
RFC+ (Step 2) [31]	<u>0.398</u>	0.609	0.271	<u>0.468</u>
E2E (ours)	0.404	0.610	0.270	0.483

2,990 for testing. It is the largest publicly available video captioning dataset in terms of the number of sentences. The average duration of the videos is 20 seconds.

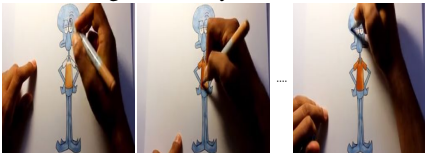


Implementation details. We implement our algorithm with Tensorflow [1]. In our end-to-end trained model, we keep the layers of Inception-Resnet-v2 [37] until the last

pooling layer whose dimension is 1,536. After that, we add a fully connected layer whose output dimension is 500. The dimension of the LSTM hidden layers is 1000. A dropout layer is attached to each LSTM unit during training with dropout rate of 0.2. Each word is represented as one-hot vector. The image embedding dimension and word embedding dimension are both 500. We fix the encoder step size to 5 and decoder step size to 35. All the trainable parameters are initialized by drawing from the uniform distribution $[-0.1, 0.1]$. The ADAM optimizer is used in our experiments. The learning rate is $1e-4$ to train S2VT. For other methods, it is $1e-6$. The hyperparameter α is 0.95 in Eq. (8). For both datasets, we resize the video frames to 224×224 . For inference, we use beam search to keep multiple generated words at current time step and select the best sentence with the beam size 3 in the end. All the free parameters are chosen by the validation sets. For the evaluation metrics, we choose four types of widely used caption metrics: BLEU4 [28], METEOR [4], CIDEr [39], and ROUGE-L [24]. The scores are calculated using the MS COCO evaluation code [11].

Table 5. Qualitative results of video captioning on MSVD dataset. Baseline is the sentence generated by our baseline model, MR stands for sentence generated by our multitask reinforce model and GT represents Ground Truth captions

		
<p>Captions:</p> <p>S2VT: a man is giving a woman</p> <p>E2E: a man is talk</p> <p>GT: a man is talking</p>	<p>Captions:</p> <p>S2VT: a woman is putting some meat in a pan</p> <p>E2E: a woman is frying meat</p> <p>GT: a woman is frying meat</p>	<p>Captions:</p> <p>S2VT: a soccer player is kicking a soccer ball</p> <p>E2E: men are playing soccer</p> <p>GT: the men are playing soccer</p>

Table 6. Qualitative results of video captioning on MSR-VTT dataset. Baseline is the sentence generated by our baseline model, MR stands for sentence generated by our multitask reinforce model and GT represents Ground Truth captions

		
<p>Captions:</p> <p>S2VT: a person is talking about a computer</p> <p>E2E: a person is drawing a cartoon</p> <p>GT: a person is drawing a cartoon</p>	<p>Captions:</p> <p>S2VT: a man is singing</p> <p>E2E: a man is jumping on a trampoline</p> <p>GT: a man is jumping on a trampoline</p>	<p>Captions:</p> <p>S2VT: a man is talking about a guitar</p> <p>E2E: a man is playing a guitar</p> <p>GT: a man is playing a guitar</p>

4.2. Baseline methods

Table 1 and 2 show the comparison results with several recently proposed methods on the two datasets, respectively. On the MSVD dataset, we compare our approach with following seven recent methods.

h-RNN [53] proposes a hierarchical-RNN framework and designs an attention scheme over both temporal and spatial dimensions to focus on the visual elements.

Attention fusion [21] develops a modality-dependent attention mechanism together with temporal attention to combines the cues of multiple modalities, which can attend not only time but also the modalities.

BA encoder [5] presents a new boundary-aware LSTM cell to detect the discontinuity of consecutive frames. Then the cell is used to build a hierarchical encoder and makes its structure adapt to the inputs.

SCN [17] detects semantic concepts from videos and proposes a tag-dependent LSTM whose weights matrix depends on the semantic concepts.

TDDF [56] combines motion feature and appearance feature, and automatically determines which feature should be focused according to the word.

LSTM-TSA [27] presents a transfer unit to control and fuse the attribute, motion, and visual features for the video representations.

MVRM [57] learns a multirate video representation which can adaptively fit the motion speeds in videos.

On the MSR-VTT dataset, we include four methods in the comparison: TDFF [56], v2t_navigator [22], Aalto [34], and Attention fusion [21].

V2t_navigator [22] represents the videos by their visual, aural, speech, and category cues, while we only employ the raw video frames in our approach.

Aalto [34] trains an evaluator network to drive the captioning model towards semantically interesting sentences.

During the experiments, REINFORCE (RFC) denotes the vanilla self-critical REINFORCE algorithm extended to video captioning, and REINFORCE+ (RFC+) represents our REINFORCE algorithm with multi-sampling trajectories. We denote by E2E our final multitask reinforcement learning approach (cf. Eq. (8)).

4.3. Comparison results

Table 1 and 2 present the results evaluated by BLEU4 [28], METEOR [4], CIDEr [39], and ROUGE-L [24] on MSVD and MSR-VTT, respectively. The CIDEr scores on MSVD dataset are much higher than those on MSR-VTT dataset. The reason may due to the much complex scenes, actions, and large variance in the MSR-VTT dataset. Our approach is denoted by E2E and two decoding results are reported, one by the greedy search and the other by the beam search of a window size of three. We can see that our approach outperforms the existing ones to large margins under the CIDEr metric, which is taken as the reward func-

tion in our training procedure. Under the other metrics, ours is also among the top performing methods while we notice that one can conveniently replace the CIDEr reward by the other metrics as the reward functions. On MSVD dataset, our E2E method can achieve 0.865 in CIDEr. Comparing with the baseline method, S2VT, our E2E model can make a relative improvement by 15.3% in CIDEr with greedy decoding. On the MSR-VTT dataset, our E2E method can reach 0.483 in CIDEr. It can make a relative improvement by 18.6%.

Several factors may have contributed to the superior performance of our model. First, we fine-tune the CNNs such that the extracted features of the video frames are purposely tailored for the video captioning task, as opposed to the generic features pre-trained from the ImageNet. Besides, the LSTM architecture inherently exploits the temporal nature of the videos. At last but not the least, the performance attributes to the progressive and multitask techniques of training the model. Next, we provide in-depth analyses about the last point by some ablation studies.

4.4. Ablation study

Due to the time and computation resource constraint, we mainly run the ablation experiments on the MSVD dataset. See Table 3 for the results. Additionally, we also report some key results on MSR-VTT in Table 4.

First of all, we note that Step 2 is able to significantly improve the results of Step 1, reinforcing the effectiveness of the REINFORCE algorithm [45]. Besides, by sampling multiple trajectories (cf. row RFC+) we can boost the original REINFORCE by 1% to 2%.

If we skip Step 2 and directly fine-tune the CNNs using the cross-entropy loss L_x in Step 3, the results (cf. row E2E (xentropy)) are only marginal better than those of freezing the CNNs. This observation is not surprising, given that the full model is actually both deep in CNNs and long in terms of the unrolled LSTMs, making it very hard to train.

If we skip Step 2 and instead minimize the convex combination of the attribute prediction loss L_a and the cross entropy loss L_x of the video captions, the results are much better than those of Step 1 and yet still worse than Step 2's. Hence, we conclude that 1) the attribute prediction branch helps the video captioning task and 2) the REINFORCE training is inevitable for eliminating the exposure mismatch [6] of the LSTMs between the training and testing stages.

If we remove the attribute prediction branch from our model and only use the REINFORCE+ to fine-tune the CNNs in Step 3, the results can only be very slightly improved which can be seen from E2E w/o attribute. This verifies the necessity of the attribute prediction branch. Indeed, this branch back-propagates the gradients from an albeit different attribute prediction task directly to the CNNs,

being able to complement the gradients coming through the LSTM branch. If we do not follow the steps and directly train the model end-to-end, the result is even lower than S2VT. It confirms our E2E progressive training pipeline is effective.

5. Qualitative results

In Table 5 and Table 6, a few video caption instances are shown of the MSVD dataset and the MSR-VTT dataset. The captions are generated by the S2VT model and our E2E model, respectively. We compare the sentences with the ground truth sentences in the Tables. Generally, our E2E model can generate relevant sentences. The sentences generated by our E2E model can reflect the visual content more faithfully with less grammar errors. For instance, our multi-task model generates “a woman is frying meat” and it shows exactly what the woman is doing in the middle image of Table 5. It is more reasonable and relevant to the video content than “a woman is putting some meat in a pan” generated by the baseline model. Our E2E model also describe the event correctly, it recognizes there are a group of players playing soccer instead of one player in the right image of Table 5. On the other dataset, the examples also illustrate the correctness and faithfulness of our method. It can correctly detect drawing a cartoon compared to talking about a computer, the action of jumping on a trampoline and playing instead of talking about guitar. Under most test cases, our model is descriptive and more accurate than the baseline.

6. Conclusion

We propose a novel method which combines the reinforcement learning with attribute prediction to train the whole framework end-to-end for video captioning. For our E2E model, it is a multitask end-to-end network and combines multisampling reinforce algorithm to generate captions. It is the first time that the CNNs are learned together with RNNs in video captioning and show much improvement, to best of our knowledge. The experiments on two standard video captioning datasets show our model can outperform the current methods. It also shows that the domain adopted video representation is more powerful than the generic image features. In the future, we will explore more representative video representations. As the 3D convolution methods are effective in the video classification, e.g I3D [7], we believe our model can also benefit from employing the effective video representation in video classification field. We may also explore other multitasks to better fine-tune the video representation.

References

- [1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensor-

- flow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.
- [2] L. Anne Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2016.
- [3] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [4] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, volume 29, pages 65–72, 2005.
- [5] L. Baraldi, C. Grana, and R. Cucchiara. Hierarchical boundary-aware neural encoder for video captioning. *arXiv preprint arXiv:1611.09312*, 2016.
- [6] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1171–1179, 2015.
- [7] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733. IEEE, 2017.
- [8] W. Chan, N. Jaitly, Q. Le, and O. Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 4960–4964. IEEE, 2016.
- [9] D. L. Chen and W. B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 190–200. Association for Computational Linguistics, 2011.
- [10] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, and T.-S. Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. *arXiv preprint arXiv:1611.05594*, 2016.
- [11] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.
- [12] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- [13] J. Chung, K. Cho, and Y. Bengio. A character-level decoder without explicit segmentation for neural machine translation. *arXiv preprint arXiv:1603.06147*, 2016.
- [14] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.
- [15] B. Dai, D. Lin, R. Urtasun, and S. Fidler. Towards diverse and natural image descriptions via a conditional gan. *arXiv preprint arXiv:1703.06029*, 2017.
- [16] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.
- [17] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5630–5639, 2017.
- [18] T. Gebu, J. Hoffman, and L. Fei-Fei. Fine-grained recognition in the wild: A multi-task domain adaptation approach. *arXiv preprint arXiv:1709.02476*, 2017.
- [19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [20] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [21] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi. Attention-based multimodal fusion for video description. In *Proceedings of the IEEE international conference on computer vision*, pages 4193–4202, 2017.
- [22] Q. Jin, J. Chen, S. Chen, Y. Xiong, and A. Hauptmann. Describing videos using multi-modal fusion. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 1087–1091. ACM, 2016.
- [23] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, et al. Google’s multilingual neural machine translation system: enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*, 2016.
- [24] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain, 2004.
- [25] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy. Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE international conference on computer vision*, pages 873–881, 2017.
- [26] P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang. Hierarchical recurrent neural encoder for video representation with application to captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1029–1038, 2016.
- [27] Y. Pan, T. Yao, H. Li, and T. Mei. Video captioning with transferred semantic attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6504–6512, 2017.
- [28] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

- [29] Y. Pu, M. R. Min, Z. Gan, and L. Carin. Adaptive feature abstraction for translating video to language. *arXiv preprint arXiv:1611.07837*, 2016.
- [30] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li. Deep reinforcement learning-based image captioning with embedding reward. *arXiv preprint arXiv:1704.03899*, 2017.
- [31] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [32] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.
- [33] Z. Shen, J. Li, Z. Su, M. Li, Y. Chen, Y.-G. Jiang, and X. Xue. Weakly supervised dense video captioning. *arXiv preprint arXiv:1704.01502*, 2017.
- [34] R. Shetty and J. Laaksonen. Frame-and segment-level features and candidate pool evaluation for video caption generation. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 1073–1076. ACM, 2016.
- [35] R. K. Srivastava, K. Greff, and J. Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.
- [36] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.
- [37] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, pages 4278–4284, 2017.
- [38] S. Toshniwal, H. Tang, L. Lu, and K. Livescu. Multitask learning with low-level auxiliary tasks for encoder-decoder based speech recognition. *arXiv preprint arXiv:1704.01631*, 2017.
- [39] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.
- [40] S. Venugopalan, L. A. Hendricks, R. Mooney, and K. Saenko. Improving lstm-based video description with linguistic knowledge mined from text. *arXiv preprint arXiv:1604.01729*, 2016.
- [41] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542, 2015.
- [42] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*, 2014.
- [43] X. Wang, C. Zhang, and Z. Zhang. Boosted multi-task learning for face verification with applications to web image and video search. In *computer vision and pattern recognition, 2009. CVPR 2009. IEEE conference on*, pages 142–149. IEEE, 2009.
- [44] P. J. Werbos. Generalization of backpropagation with application to a recurrent gas market model. *Neural networks*, 1(4):339–356, 1988.
- [45] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.
- [46] Q. Wu, C. Shen, L. Liu, A. Dick, and A. van den Hengel. What value do explicit high level concepts have in vision to language problems? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 203–212, 2016.
- [47] J. Xu, T. Mei, T. Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5288–5296, 2016.
- [48] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.
- [49] Y. Yan, E. Ricci, R. Subramanian, G. Liu, O. Lanz, and N. Sebe. A multi-task learning framework for head pose estimation under target motion. *IEEE transactions on pattern analysis and machine intelligence*, 38(6):1070–1083, 2016.
- [50] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*, pages 4507–4515, 2015.
- [51] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei. Boosting image captioning with attributes. *OpenReview*, 2(5):8, 2016.
- [52] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4651–4659, 2016.
- [53] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4584–4593, 2016.
- [54] Y. Yu, H. Ko, J. Choi, and G. Kim. End-to-end concept word detection for video captioning, retrieval, and question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3165–3173, 2017.
- [55] X.-T. Yuan, X. Liu, and S. Yan. Visual classification with multitask joint sparse representation. *IEEE Transactions on Image Processing*, 21(10):4349–4360, 2012.
- [56] X. Zhang, K. Gao, Y. Zhang, D. Zhang, J. Li, and Q. Tian. Task-driven dynamic fusion: Reducing ambiguity in video description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3713–3721, 2017.
- [57] L. Zhu, Z. Xu, and Y. Yang. Bidirectional multirate reconstruction for temporal modeling in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2653–2662, 2017.