SODA: Story Oriented Dense Video Captioning Evaluation Framework

Soichiro Fujita[1], Tsutomu Hirao[2], Hidetaka Kamigaito[1], Manabu Okumura[1], and Masaaki Nagata[2]

[1] Institute of Innovative Research, Tokyo Institute of Technology
{fujiso, kamigaito, oku}@lr.pi.titech.ac.jp
[2] NTT Communication Science Laboratories, NTT Corporation
{tsutomu.hirao.kp, masaaki.nagata.et}@hco.ntt.co.jp

Abstract. Dense Video Captioning (DVC) is a challenging task that localizes all events in a short video and describes them with natural language sentences. The main goal of DVC is video story description, that is, to generate a concise video story that supports human video comprehension without watching it. In recent years, DVC has attracted increasing attention in the vision and language research community, and has been employed as a task of the workshop, ActivityNet Challenge. In the current research community, the official scorer provided by ActivityNet Challenge is the de-facto standard evaluation framework for DVC systems. It computes averaged METEOR scores for matched pairs between generated and reference captions whose Intersection over Union (IoU) exceeds a specific threshold value. However, the current framework does not take into account the story of the video or the ordering of captions. It also tends to give high scores to systems that generate several hundred redundant captions, that humans cannot read. This paper proposes a new evaluation framework, Story Oriented Dense video cAptioning evaluation framework (SODA), for measuring the performance of video story description systems. SODA first tries to find temporally optimal matching between generated and reference captions to capture the story of a video. Then, it computes METEOR scores for the matching and derives F-measure scores from the METEOR scores to penalize redundant captions. To demonstrate that SODA gives low scores for inadequate captions in terms of video story description, we evaluate two state-of-the-art systems with it, varying the number of captions. The results show that SODA gives low scores against too many or too few captions and high scores against captions whose number equals to that of a reference, while the current framework gives good scores for all the cases. Furthermore, we show that SODA tends to give lower scores than the current evaluation framework in evaluating captions in the incorrect order.

Keywords: Automatic Evaluation, Dense Video Captioning, Video Story Description

1 Introduction

Dense Video Captioning (DVC) [6] mainly involves two tasks: event detection to identify all events in a short video, and caption generation to describe the event proposals using natural language sentences. DVC is one of the major tasks in vision and language research and has attracted more attention in recent years. In fact, it has been adopted as a task of ActivityNet Challenge[3] since 2017. Its main goal is to generate concise captions that describe the story of a video to help humans understand it. Actually, humans describe the story of a video using 3-4 captions on average. Thus, the generated captions are utilized for grasping an overview of the video without having to watch the entire video [3].

However, the current de-facto standard evaluation framework for DVC systems, which is the official evaluation framework in ActivityNet Challenge, is inappropriate for measuring the performance of a video story description since it disregards the story of the video and the ordering of captions[4]. The framework first matches generated and reference captions when the Intersection over Union (IoU) between them exceeds a specific threshold value. Then, it computes METEOR [2] scores for all matched pairs between the generated and reference captions, and averages them by the number of the pairs. That is, the framework evaluates captions for events without considering the order of their proposals.

In addition, another problem with the current framework is that it often obtains a high score by producing several hundred captions that are inadequate as video story descriptions since the scores rely only on the number of matched pairs. As the result, as we will point out in Section 4.2, systems that produce more redundant captions are more advantageous. Most current DVC systems generate several hundred captions for a video, while the number of reference captions is only 3-4.

To appropriately and correctly evaluate video story description systems, we need a framework that can consider a video story, the ordering of captions, and can penalize redundant captions. This paper proposes a new evaluation framework, Story Oriented Dense video cAptioning evaluation framework (SODA), for measuring the performance of video story description systems. SODA first applies dynamic programming, that finds the optimal matching between generated and reference captions that maximizes the sum of the IoU by considering the temporal ordering of captions. Thus, it finds the best sequence of generated proposals that maximizes the sum of the IoU against reference proposals. Then, it computes METEOR scores for the matched pairs and derives precision and recall scores on the basis of the calculated METEOR scores. Finally, our framework evaluates generated captions with F-measure scores to consider both the numbers of generated and reference captions.

To demonstrate the effectiveness of our framework, we evaluate two state-of-the-art systems with it, varying the number of captions. Experimental results

[3] `http://activity-net.org/`

[4] In this paper, we follow the concept in [3] that the correct order of captions represents a story.

on the ActivityNet Captions dataset [6] show that our framework gives low scores to too many or too few captions, inadequate captions as video story description, and gives high scores to captions whose number equals to that of a reference, while the current framework gives almost the same scores to all the cases. Furthermore, we demonstrate that SODA gives lower scores to captions with incorrect order, inconsistent story description, than the current evaluation framework. In addition to the above automatic evaluation, our simple manual evaluation also shows that SODA is superior to the current framework.

Our main contributions are as follows:

- We demonstrate that the current evaluation framework, utilized in ActivityNet Challenge, is insufficient for evaluating video story descriptions.
- We propose a new evaluation framework, Story Oriented Dense video cAptioning evaluation framework (SODA), for measuring the performance of video story description systems by considering the ordering of captions.
- We introduce F-measure into the evaluation metric to prevent redundant captions from obtaining good scores.
- Our source code will be available on `https://github.com/fujiso/SODA`.

2 Related Work

2.1 Dense Video Captioning

The goal of DVC is to obtain concise and coherent description of all events in a video. It requires understanding the entire video contents and contextual reasoning of individual events. Recent researches [6, 17, 20, 8] handled this challenge by dividing it into two subtasks: event proposal detection and caption generation for the events. For example, Wang *et al.* [17] proposed a bidirectional LSTM-based encoder-decoder model with a context gating mechanism. The mechanism reflects both past and future contexts to the event proposals and the captions. Zhou *et al.* [20] proposed a self-attention [14] based end-to-end model. The end-to-end architecture could bridge the event detection and the captioning modules, hence it tended to generate a consistent caption for each individual event. However, these models did not explicitly consider the dependency or relationship among the individual events. Mun *et al.* [10] challenged to generate brief and consistent captions by reducing the number of event proposals with pointer networks [16].

There are several existing datasets for video-to-text generation other than ActivityNet Captions [6]: Youcook II [19], VideoStory [3], TACoS [12], and TACoS Multi-Sentence [13]. Youcook II, TACoS, and TACoS Multi-Sentence datasets were constructed to evaluate the captioning of cooking videos. As these types of captions temporally depend on each other, their order is an important factor to evaluate the systems. However, since Youcook II employed the same evaluation framework as ActivityNet Challenge, and TaCoS and TaCoS Multi-Sentence employed BLEU, the systems might not be evaluated correctly on the datasets. The

VideoStory dataset was constructed to evaluate video story description systems for short videos on an social networking service. However, the systems are also evaluated on the dataset with the same framework as ActivityNet Challenge, which is insufficient to evaluate the story of a video. We believe that SODA is useful not only for the ActivityNet Captions dataset but also for the other datasets constructed to evaluate system captions that convey the story.

2.2 Dense Caption Evaluation

The automatic evaluation of video description/captioning is a long term and unsolved problem. The evaluation of DVC is required to measure two aspects: 1) the accuracy of localized events, and 2) that of generated captions for each event. The current evaluation framework of DVC is inspired by that of dense image captioning (DIC) [4], which generates captions that describe localized objects in an image comprehensively. In this evaluation framework, each generated caption is separately evaluated using some metrics (See Section 3 for details.) because the captions independently describe each localized object.

Thus, there is a significant difference between DVC and DIC in whether generated captions should consist of a story or not. However, the current evaluation framework of DVC, which is a simple extension of that of DIC, does not consider the temporal dependency between captions explicitly, which causes the potential risk of overestimation (See Section 4.2 for details.).

In contrast, SODA solves this problem through optimal matching with ground-truth events and penalizing redundant events, as we will explain in Section 5. It would be more difficult to obtain a factitiously high score with SODA compared with the current evaluation framework because SODA requires systems to detect the exact number of events and captions, that we believe will lead to further progress of DVC tasks.

The research community of DVC has mainly used the following six different evaluation metrics for caption sentences: ROUGE-L [9], METEOR[2], BLEU [11], CIDEr [15], SPICE [1], and WMD [7]. These metrics were originated from text generation tasks in natural language processing such as machine translation, text summarization, and image captioning. There have been several experiments to make clear which metrics are better for caption evaluation [18, 5] because of too many metrics. They showed that evaluation metrics being relatively less sensitive to word order and synonym changes in a sentence, like CIDEr and METEOR, can provide a high correlation with human judgments. Therefore, METEOR was adopted as the main evaluation metric in DVC.

3 Current Evaluation Framework

The automatic evaluation framework proposed for ActivityNet Captions [6] has been widely utilized for the DVC task. Let $\mathcal{G}$ be a set of manually-generated reference captions for a video and $\mathcal{P}$ be a set of captions generated by a system. We denote g as a reference caption and p as a caption generated by the system.

	p_1	p_2	p_3	p_4	p_5
g_1	0.7	0.1	0.4	0.9	0.1
g_2	0.2	0.3	0.5	0.4	0.5
g_3	0.4	1.0	0.3	0.7	0.8
g_4	0.8	0.7	0.6	1.0	0.1

Fig. 1. An example of IoUs between generated and reference captions.

Each caption has a proposal that indicates a time span of an event that appears in a video. Here, the IoU between g and p is defined as follows:

$$\text{IoU}(g,p) = \max\left(0, \frac{\min(e(g),e(p)) - \max(s(g),s(p))}{\max(e(g),e(p)) - \min(s(g),s(p))}\right), \tag{1}$$

where functions $s(\cdot)$ and $e(\cdot)$ return the start and end time of the proposal, respectively. Here, a set of reference captions whose IoU exceeds a specific threshold, τ, for p is defined as follows:

$$G_{p,\tau} = \{g \in \mathcal{G} | \text{IoU}(g,p) \geq \tau\}. \tag{2}$$

When $G_{p,\tau} = \phi$, i.e., p does not have any g that exceeds a specific threshold, we add a random string to $G_{p,\tau}$ as a member instead of the caption as a penalty. Finally, a set of generated captions, $\mathcal{P}$, is evaluated on the basis of a set of reference captions, $\mathcal{G}$, by the following equation:

$$E(\mathcal{G},\mathcal{P},\tau) = \frac{\sum_{p\in\mathcal{P}} \sum_{g\in G_{p,\tau}} f(g,p)}{\sum_{p\in\mathcal{P}} |G_{p,\tau}|}, \tag{3}$$

where function $f(\cdot,\cdot)$ denotes an evaluation metric such as METEOR [2], BLEU [11], or CIDEr [15]. In this paper, we use METEOR as $f(\cdot,\cdot)$ since ActivityNet Challenge use it as its official metric. In most cases, the final evaluation score was computed as the average for $\tau = 0.9, 0.7, 0.5, 0.3$.

Consider, for example, that IoUs between $\mathcal{G}$ and $\mathcal{P}$ are given as in Figure 1. When we set τ to 0.5, we obtain the following: $G_{p_1,0.5} = \{g_1, g_4\}$, $G_{p_2,0.5} = \{g_3, g_4\}$, $G_{p_3,0.5} = \{g_2, g_4\}$, $G_{p_4,0.5} = \{g_1, g_3, g_4\}$, $G_{p_5,0.5} = \{g_2, g_3\}$. Then, we compute METEOR scores for the eleven matched pairs between g and p, (p_1, g_1), (p_1, g_4), (p_2, g_3), (p_2, g_4), (p_3, g_2), (p_3, g_4), (p_4, g_1), (p_4, g_3), (p_4, g_4), (p_5, g_2), and (p_5, g_3), and average the scores.

4 Problems of Current Framework

4.1 Loose Matching

As we explained in the previous section, the current evaluation framework determines the correspondence between g and p only with the IoU threshold, τ. Thus,

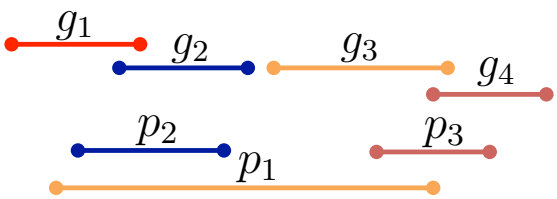

Fig. 2. Example of system and reference proposals that produce loose matching.

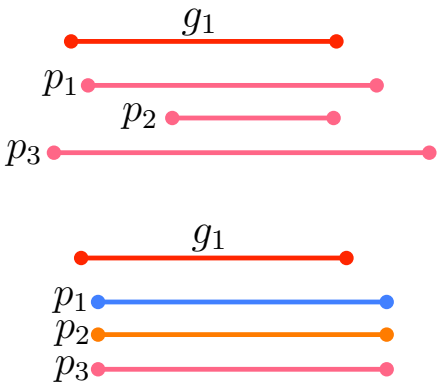

Fig. 3. Examples of system and reference proposals that produce redundant captions.

it causes loose matching; a generated caption is matched with many reference captions or a reference caption is matched with many generated captions. The order of generated captions produced by the matching does not correspond to the order of reference captions, i.e., the loose matching disregards the story of the video. For example, when reference and generated captions are given as shown in Figure 2, the current evaluation framework produces the following matches, (g_1, p_1), (g_1, p_2), (g_2, p_1), (g_2, p_2), (g_3, p_1), (g_3, p_3), and (g_4, p_3) for small τ. Thus, the order of the generated captions corresponding to the reference captions is $p_1, p_2, p_1, p_2, p_1, p_3, p_3$, which is invalid because it contains many duplicates, i.e., the captions do not represent the story of the video. The best order of the generated captions that represents the story can be p_2, p_1, p_3.

Furthermore, loose matching produces overestimations of METEOR scores. When we have the same sentences with different length proposals, any of them would match with a reference caption for any τ, even though the redundant captions make no sense. Consider IoUs and METEOR scores are given as follows: $\mathrm{IoU}(g_1, p_1) = 0.9$, $\mathrm{IoU}(g_1, p_2) = 0.4$, $\mathrm{IoU}(g_1, p_3) = 0.6$, and $\mathrm{METEOR}(g_1, p_*) = 0.6$, as shown in the top of Figure 3. When we set τ to 0.9, only g_1 matches to p_1, and the average METEOR score is 0.2, while p_2 and p_3 are eliminated in the

matching. However, when we have only p_2, g_1 is not matched, and the average METEOR score is 0.0. It indicates that the current evaluation sometimes gives higher scores for low-confidence proposals than a single high-confidence proposal. Thus, redundant caption sentences, such as identical sentences with proposals of different lengths, may obtain good METEOR scores.

Another problematic case is when we generate multiple different sentences for a proposal, the average METEOR score tends to be good. Consider, for example, METEOR scores are given as follows: $\mathrm{METEOR}(g_1, p_1) = 0$, $\mathrm{METEOR}(g_1, p_2) = 0.3$, and $\mathrm{METEOR}(g_1, p_3) = 0.6$, as shown in the bottom of Figure 3. In this example, the average METEOR score is 0.3, while it is 0 when generating only p_1. Thus, generating multiple different caption sentences for a proposal tends to prevent a zero Meteor score.

4.2 Averaging METEOR Scores

As shown in Equation (3), the sum of METEOR scores is averaged based on the number of matched pairs between g and p. That is, the number of captions generated by a system, $|\mathcal{P}|$, and the number of reference captions, $|\mathcal{G}|$, are disregarded in calculating the evaluation metric. Thus, it cannot take into account the coverage of generated captions (recall) and the accuracy of the captions (precision) either.

As we mentioned above, a better score is obtained by generating more different sentences for a proposal and more identical sentences for different proposals. Most of current DVC systems generate many and redundant caption sentences for a video. The average number of generated caption sentences is around several hundred. Thus, the current evaluation framework is inadequate since we cannot read several hundred sentences in a short time, while, of course, it may be reasonable to assess DVC systems in terms of video indexing, which does not require human reading. This is a critical problem of the current evaluation framework for video story description systems.

Furthermore, too few caption sentences are also inappropriate because such captions cannot represent the whole story of a video. To penalize such inappropriate captions, we can derive precision and recall by replacing the denominator of Equation (3) with $|\mathcal{P}|$ and $|\mathcal{G}|$, respectively. However, they are invalid since the scores might exceed 1.0.

5 Story Oriented Dense video cAptioning evaluation framework (SODA)

5.1 Optimal Matching Using Dynamic Programming

To determine the matching between generated and reference captions, we regard the matching as a combinatorial optimization problem: finding one-to-one matching between the captions that maximizes the sum of the IoU by considering temporal ordering. Following the current evaluation framework, we also use

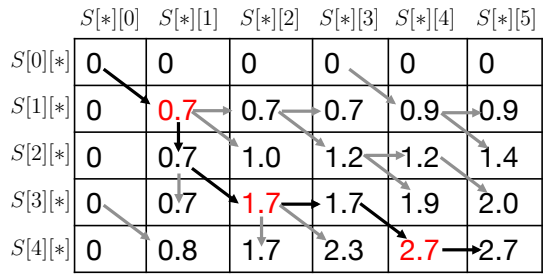

Fig. 4. Illustration of a dynamic programming table.

the threshold τ for the matching; we define cost $C_{i,j}$ between a reference caption g_i and a generated caption p_j based on the IoU as follows:

$$C_{i,j} = \begin{cases} \text{IoU}(g_i, p_j) \text{ if } \text{IoU}(g_i, p_j) \geq \tau, \\ 0 \qquad\qquad \text{otherwise.} \end{cases} \tag{4}$$

Then, we sort the captions based on temporal ordering, that is, in the order of the beginning time of their proposals, by utilizing function $s(\cdot)$, and define $S[i][j]$, which stores the maximum score of optimal matching between 1st to i-th generated captions and the 1st to j-th reference truth captions, as follows:

– Initialization

$$S[i][0] = 0 \ (0 \leq i \leq |\mathcal{P}|), S[0][j] = 0 \ (0 \leq j \leq |\mathcal{G}|), \tag{5}$$

– Recurrence $(1 \leq i \leq |\mathcal{P}|, 1 \leq k \leq |\mathcal{G}|)$

$$S[i][j] = \max \begin{cases} S[i-1][j], \\ S[i-1][j-1] + C_{i,j}, \\ S[i][j-1]. \end{cases} \tag{6}$$

Figure 4 shows an example process to obtain the optimal matching for the example given in Figure 1, with $\tau = 0$. After filling out table S by dynamic programming, $S[4][5]$ stores the optimal matching score, 2.7. Thus, we can obtain the optimal matching between g_k and p_ℓ by tracing the path, from [4,5] to [0,0]. In the example, the optimal matching is (g_1, p_1), (g_3, p_2), (g_4, p_4). The pseudo code of the algorithm is shown in the supplementary material.

5.2 F-measure for Evaluating Video Story Description

To give a low score for too many or too few captions, the sum of METEOR scores should be normalized by considering the number of generated and reference captions. Thus, we propose an evaluation metric based on F-measure as follows:

$$\text{F-measure}(\mathcal{G}, \mathcal{P}) = \frac{2 \times \text{Precision}(\mathcal{G}, \mathcal{P}) \times \text{Recall}(\mathcal{G}, \mathcal{P})}{\text{Precision}(\mathcal{G}, \mathcal{P}) + \text{Recall}(\mathcal{G}, \mathcal{P})}. \tag{7}$$

Here, Precision$(\mathcal{G}, \mathcal{P})$ and Recall$(\mathcal{G}, \mathcal{P})$ are defined on the basis of the optimal matching as follows:

$$\text{Precision}(\mathcal{G}, \mathcal{P}) = \frac{\sum_{g \in \mathcal{G}} f(g, p_{a(g)})}{|\mathcal{P}|}, \tag{8}$$

$$\text{Recall}(\mathcal{G}, \mathcal{P}) = \frac{\sum_{g \in \mathcal{G}} f(g, p_{a(g)})}{|\mathcal{G}|}. \tag{9}$$

When systems generate too many captions, Precision scores tend to be low, while Recall scores tend to be high. Thus, the systems cannot obtain good F-measure scores. When systems generate too few captions, they also cannot obtain good F-measure scores since they tend to receive good Precision scores but poor Recall scores.

5.3 Evaluation Scores Directly Dependent on IoU

In evaluating video story descriptions, the IoU plays an important role. Even if METEOR scores between generated and reference captions are perfect, they make no sense if the IoU between the captions is zero. However, in the current evaluation framework, the IoU is utilized only for determining the matching between the captions. Thus, the IoU does not directly affect the sum of METEOR scores. In fact, METEOR scores with larger IoUs and those with smaller ones cannot be distinguished when computing them. To reflect the IoU more directly to evaluation scores, we propose an alternative of the cost in Equation (4), which is utilized to solve dynamic programming as follows:

$$C_{i,j} = \text{IoU}(g_i, p_j) f(g_i, p_j). \tag{10}$$

By utilizing this cost, even if the METEOR score is high, the evaluation score can be lowered when the IoU score is low.

6 Experiments

6.1 Experimental Settings

We used the ActivityNet Captions dataset [6][5], which contains 20k YouTube videos. The dataset consists of 10,024, 4,915 and 5,044 videos for training, validation and test data, respectively. We evaluated our evaluation framework only on the validation data because the test data is not publicly available[6]. Each video in the validation data has on average 3.52 human-written captions with start/end time annotations. The average number of words in a caption is 13.54.

[5] Using the VideoStory dataset [3] would have been more effective as it was constructed to evaluate video story description systems. However, unfortunately, it has not been publicly available.

[6] The test data is only available on the ActivityNet evaluation server.

Because it is known that METEOR has sufficient correlation against human evaluation, we did not evaluate our framework by calculating the correlation against manual evaluation. To demonstrate the effectiveness of the optimal proposal detection and F-measure, we simply examined whether our framework could give low scores for inadequate captions and high scores for adequate captions in terms of video story description. Thus, we evaluated the following two state-of-the-art DVC systems with two settings below:

- End-to-end transformer-based system [20]: The end-to-end transformer-based models could detect events by considering the whole video information and generate consistent captions for the events simultaneously. The number of output captions per video is 228.21 on average.
- LSTM-based system [17]: The bidirectional LSTM-based encoder-decoder models with a context gating mechanism. The context gating mechanism makes it possible to generate captions by filtering both past and future contexts. The number of output captions per video is 97.10 on average.

Note that we obtained captions generated by the end-to-end transformer-based model by running their code, available at the github repository,[7] and those by the LSTM-based model were provided by the authors of the paper.

Detecting inappropriate captions: To demonstrate whether SODA gives low scores to inadequate captions, we first performed experiments by varying the number of captions. Since both systems generate around hundred or more captions for each video, we randomly selected $\texttt{int}(m \times |\mathcal{G}|)$ captions[8] without duplication and evaluated the captions by the current evaluation framework,[9] and the following metrics in our framework:

- SODA (a): averaging F-measure scores with $\tau = 0.9, 0.7, 0.5, 0.3$,
- SODA (b): F-measure, where τ is set to 0,
- SODA (c): F-measure, utilizing the cost in Equation (11).

We examined $m = 0.1, 0.5, 1.0, 2.0, 10$, and "all," where $m = 1.0$ indicates a case of outputting the same number of captions as a reference. We report the average scores with five times of the randomized procedure.

Detecting incorrect ordering: To demonstrate whether SODA gives low scores to captions with incorrect ordering, we evaluated original captions with correct ordering and two types of captions with incorrect ordering: (a) Swap, swapping the order of two adjacent captions for a randomly selected pair in the

[7] `https://github.com/salesforce/densecap`

[8] m is a parameter for controlling the number of captions in the experiments.

[9] We utilized the official scorer provided by ActivityNet Challenge. The code is available at `https://github.com/ranjaykrishna/densevid_eval`. We used the revised version from November 2017 that fixed an overestimation bug; the number of pairs (the denominator in Equation (3)) was not counted correctly.

original, and (b) Shuffle, randomly shuffling the order of captions in the original. Since systems generate a huge number of captions, we cannot evaluate them as they are, and we need to perform the experiments with a reasonable number of captions. Therefore, we assessed the systems by their potential, the upper bound performances; we examined captions that were the closest to the corresponding references. That is, we used oracle captions as the original captions with correct ordering. Here, an oracle caption indicates a caption of the same length as a reference caption that receives the maximum METEOR score. We created the oracles by selecting those generated captions that receive the maximum METEOR score for each reference caption. Then, we created the Swap and Shuffle captions from the oracle captions by randomly varying their proposals so that keeping IoUs exceed 0.7.

6.2 Results

Detecting inappropriate captions: Table 1 shows the results. From the results, the scores of the current evaluation framework do not change significantly even when the number of captions changes. We observe a similar tendency in both results obtained from the two different systems. In particular, there is only a slight difference between $m = 1$, the appropriate number of captions, and "all," the inappropriate number of captions for humans to read. As long as DVC systems are evaluated with the current evaluation framework, they would continue to generate many and redundant captions. As we mentioned before, the results are caused by (1) loose matching between generated and reference captions, and (2) averaging METEOR scores by the number of the matched pairs. The results reveal that the framework has a critical problem, i.e., it cannot distinguish good captions from bad captions in terms of video story description. In contrast, SODA gave low scores for too many and too few captions. When we utilized a small m, precision became high, while recall became low. Precision became low and recall became high when we utilized a large m. Thus, SODA can penalize inadequate captions.

Before we compare the performance of the two systems, we need to address the question: "Which is better: E2E Transformer or LSTM in terms of video story description?". Therefore, we compared their oracle performances and their diversities. We created the oracles as above and averaged their METEOR scores. Then, we computed Self-BLEU [21][10] as a measure to assess the quality of oracle captions in terms of diversity. The results shown in Table 2 demonstrate that E2E Transformer outperformed LSTM in terms of both METEOR and Self-BLEU scores with significant differences. Thus, we can conclude that the potential of E2E Transformer to describe the video story is superior to that of LSTM. Although LSTM outperformed E2E Transformer with the current evaluation framework in Table 1, the result does not agree with Table 2. However, the

[10] Self-BLEU has been used to assess the diversity of a set of generated sentences in text generation tasks. A lower score indicates a higher diversity.

Table 1. Changes of scores (%) obtained from the current evaluation framework and SODA when varying the number of captions. Prec., Rec., and F_1 indicate precision, recall, and F-measure, respectively.

		E2E Transformer [20]						LSTM [17]					
	m	0.1	0.5	1.0	2.0	10	All	0.1	0.5	1.0	2.0	10	All
	Current	3.78	4.04	4.10	4.14	4.18	4.19	4.72	4.88	4.92	4.94	4.96	4.97
SODA (a)	Prec.	10.06	5.22	3.86	2.59	0.77	0.27	9.97	4.77	3.14	1.85	0.47	0.37
	Rec.	1.43	2.79	3.86	5.18	7.65	8.76	1.42	2.54	3.14	3.70	4.53	4.66
	F_1	2.51	3.63	3.86	3.45	1.40	0.52	2.44	3.32	3.14	2.46	0.86	0.63
SODA (b)	Prec.	10.50	9.32	7.55	4.68	1.06	0.34	9.94	8.05	6.07	3.50	0.81	0.57
	Rec.	1.50	4.97	7.55	9.36	10.52	11.1	1.42	4.29	6.07	7.00	7.75	7.82
	F_1	2.62	6.48	7.55	6.24	1.92	0.66	2.43	5.60	6.07	4.67	1.47	1.05
SODA (c)	Prec.	5.93	4.89	4.02	2.87	0.94	0.33	7.01	4.49	3.15	1.94	0.53	0.37
	Rec.	0.84	2.61	4.02	5.74	9.30	10.62	1.00	2.40	3.15	3.87	5.06	5.25
	F_1	1.47	3.41	4.02	3.83	1.70	0.63	1.75	3.12	3.15	2.58	0.96	0.71

Table 2. Average METEOR and Self-BLEU scores (%) for oracle captions.

	E2E Transformer	LSTM
METEOR	21.3	13.43
Self-BLEU	79.5	90.6

evaluation results with SODAs in Table 1 can agree with Table 2 in that E2E Transformer outperformed LTSM.

Comparing the variants of SODA, the fluctuation of SODA (a) is smaller than that of the other metrics since it received low scores when utilizing a big τ. On the other hand, the scores of SODA (b) and (c) sensitively change with m, and the fluctuation is large. To assess the performance of video story description, the metric should be sensitive to the number of captions to be evaluated. Thus, SODA (b) and (c) are more appropriate for measuring the performance of video story description systems. Since SODA (c) involves the IoU in the evaluation score, that is, the score depends both on METEOR and IoU, we believe that SODA (c) is the most appropriate.

Detecting incorrect ordering: Table 3 shows the scores obtained with the current evaluation framework and SODA (c) (F-measure) for correctly- and incorrectly-ordered captions. With both metrics, the scores for captions with incorrect ordering degraded properly. The percentage decreases for Shuffle are larger than those for Swap because Shuffle tends to be worse in the ordering. In comparing SODA with the current evaluation framework, the percentage de-

Table 3. Evaluation scores (%) for captions with correct and incorrect order. The number in parentheses indicates the percentage decrease from the score for Correct.

	E2E Transformer			LSTM		
	Correct	Swap	Shuffle	Correct	Swap	Shuffle
Current	16.1	14.5 (*-10.2*)	10.8 (*-33.1*)	10.5	9.92 (*-5.6*)	8.58 (*-18.4*)
SODA (c)	17.8	14.5 (*-18.9*)	7.66 (*-57.0*)	10.7	8.89 (*-17.0*)	5.60 (*-47.7*)

creases for Shuffle with SODA (c) are in range of 47–57%, while those with Current are in range of 18–33%. Therefore, SODA (c) can evaluate the incorrectly-ordered captions more severely than the current framework. While the percentage decreases for Swap are smaller than those for Shuffle, we can find a similar tendency as Shuffle that SODA (c) can evaluate the incorrectly ordered captions more severely. These results indicate that SODA is more sensitive to the incorrect ordering of captions than Current, i.e., SODA is more suitable to evaluate the story of a video than the current evaluation framework.

In summary, our experiments revealed that SODA finds inappropriate captions in terms of video story description. That is, SODA gives low scores to too many or too few captions and incorrectly ordered captions. These are significant advantages of SODA against the current evaluation framework.

6.3 Manual Evaluation

To investigate whether SODA agrees with human judgment, we computed the accuracies of SODA (c) and the current evaluation framework against the results obtained from human judgment, (1) that compared E2E Transformer with LSTM oracles, and (2) that compared `Shuffle` with `Swap` for gold standard captions.[11] We randomly selected 50 videos with less than 6 captions, whose length is from 90 to 180 seconds, from the validation data. Then, we showed the video and the two captions to 12 crowdsourced workers and asked them to compare the captions A and B and to select an integer score from -2 to 2, where the score -2 indicates A is better and the score 2 indicates B is better. We asked them to judge whether the captions correctly describe the entire video, and the events are described in the correct order. We employed only faithful workers who correctly answered test questions, which cannot be answered without watching the video.

In the former human judgment, E2E Transformer obtained better results for 80% of the 50 videos. Thus, the results demonstrate that the potential of E2E Transformer is superior to LSTM. The results agree with those in Table 2. The accuracies of SODA and the current evaluation framework against the human judgment are 0.76 and 0.66, respectively. The results imply that SODA is superior to the current evaluation framework.

[11] In order to prevent `Shuffle` from being the same as `Swap`, we employed only captions with reverse ordering of the gold standard as `Shuffle` in the human judgment.

In the latter human judgment, `Swap` obtained better results for 94% of the 50 videos. The results are reasonable and agree with our intuition since `Swap` keeps better temporal ordering than `Shuffle`. The results indicate that humans give higher scores when captions meet correct ordering. The accuracies of SODA and the current evaluation framework are 0.94 and 0.72, respectively. The results also show that SODA is superior to the current evaluation framework.

7 Conclusion

In this paper, we demonstrated that the current evaluation framework, which is the official evaluation framework utilized in ActivityNet Challenge, is inadequate for evaluating the performance of video story description systems. Then, we proposed a new evaluation framework, Story Oriented Dense video cAptioning evaluation framework (SODA), to perform better evaluations. To match generated and reference captions considering temporal ordering, SODA first finds the optimal matching that maximizes the sum of the IoU by using dynamic programming. Then, it computes F-measure based on the METEOR scores for the matched pairs.

Evaluation results obtained on the ActivityNet Captions dataset showed that we can detect inadequate captions and too many or too few captions by utilizing SODA, which cannot be detected by using the current evaluation framework. Furthermore, we demonstrated that SODA gives lower scores to captions with incorrect ordering and inconsistent story descriptions, than the current evaluation framework. We also showed that SODA is superior to the current framework in detecting appropriate captions and in detecting captions with incorrect temporal order via manual evaluation.

References

1. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: European Conference on Computer Vision. pp. 382–398. Springer (2016)
2. Banerjee, S., Lavie, A.: 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. pp. 65–72 (2005)
3. Gella, S., Lewis, M., Rohrbach, M.: A dataset for telling the stories of social media videos. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 968–974 (2018), `https://www.aclweb.org/anthology/D18-1117.pdf`
4. Johnson, J., Karpathy, A., Fei-Fei, L.: Densecap: Fully convolutional localization networks for dense captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
5. Kilickaya, M., Erdem, A., Ikizler-Cinbis, N., Erdem, E.: Re-evaluating automatic metrics for image captioning. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long

Papers. pp. 199–209. Association for Computational Linguistics, Valencia, Spain (Apr 2017), https://www.aclweb.org/anthology/E17-1019
6. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 706–715 (2017), http://openaccess.thecvf.com/content_ICCV_2017/papers/Krishna_Dense-Captioning_Events_in_ICCV_2017_paper.pdf
7. Kusner, M., Sun, Y., Kolkin, N., Weinberger, K.: From word embeddings to document distances. In: International conference on machine learning. pp. 957–966 (2015)
8. Li, Y., Yao, T., Pan, Y., Chao, H., Mei, T.: Jointly localizing and describing events for dense video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7492–7500 (2018), http://openaccess.thecvf.com/content_cvpr_2018/papers/Li_Jointly_Localizing_and_CVPR_2018_paper.pdf
9. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://www.aclweb.org/anthology/W04-1013
10. Mun, J., Yang, L., Ren, Z., Xu, N., Han, B.: Streamlined dense video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6588–6597 (2019), http://openaccess.thecvf.com/content_CVPR_2019/papers/Mun_Streamlined_Dense_Video_Captioning_CVPR_2019_paper.pdf
11. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 311–318 (2002)
12. Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., Pinkal, M.: Grounding action descriptions in videos. Transactions of the Association for Computational Linguistics **1**, 25–36 (2013), https://www.aclweb.org/anthology/Q13-1003.pdf
13. Rohrbach, A., Rohrbach, M., Qiu, W., Friedrich, A., Pinkal, M., Schiele, B.: Coherent multi-sentence video description with variable level of detail. In: German conference on pattern recognition. pp. 184–195. Springer (2014)
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
15. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4566–4575 (2015)
16. Vinyals, O., Fortunato, M., Jaitly, N.: Pointer networks. In: Advances in neural information processing systems. pp. 2692–2700 (2015)
17. Wang, J., Jiang, W., Ma, L., Liu, W., Xu, Y.: Bidirectional attentive fusion with context gating for dense video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7190–7198 (2018)
18. Wang, J., Gaizauskas, R.: Cross-validating image description datasets and evaluation metrics. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 3059–3066. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016), https://www.aclweb.org/anthology/L16-1489
19. Zhou, L., Xu, C., Corso, J.J.: Towards automatic learning of procedures from web instructional videos. In: Thirty-Second AAAI Conference on Artificial In-

telligence (2018), https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17344/16367
20. Zhou, L., Zhou, Y., Corso, J.J., Socher, R., Xiong, C.: End-to-end dense video captioning with masked transformer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8739–8748 (2018), http://openaccess.thecvf.com/content_cvpr_2018/papers/Zhou_End-to-End_Dense_Video_CVPR_2018_paper.pdf
21. Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J., Yu, Y.: Texygen: A benchmarking platform for text generation models. In: Proceedings of SIGIR 18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 1097–1100 (2018)