Video Storytelling: Textual Summaries for Events

Junnan Li, Yongkang Wong, *Member, IEEE,* Qi Zhao, *Member, IEEE,* Mohan S. Kankanhalli, *Fellow, IEEE*

***Abstract*—Bridging vision and natural language is a longstanding goal in computer vision and multimedia research. While earlier works focus on generating a single-sentence description for visual content, recent works have studied paragraph generation. In this work, we introduce the problem of video storytelling, which aims at generating coherent and succinct stories for long videos. Video storytelling introduces new challenges, mainly due to the diversity of the story and the length and complexity of the video. We propose novel methods to address the challenges. First, we propose a context-aware framework for multimodal embedding learning, where we design a Residual Bidirectional Recurrent Neural Network to leverage contextual information from past and future. The multimodal embedding is then used to retrieve sentences for video clips. Second, we propose a Narrator model to select clips that are representative of the underlying storyline. The Narrator is formulated as a reinforcement learning agent which is trained by directly optimizing the textual metric of the generated story. We evaluate our method on the Video Story dataset, a new dataset that we have collected to enable the study. We compare our method with multiple state-of-the-art baselines and show that our method achieves better performance, in terms of quantitative measures and user study.**

***Index Terms*—Video Storytelling, Video Captioning, Sentence Retrieval, Multimodal Embedding Learning**

I. Introduction

Generating natural language descriptions for visual content has become a major research problem for computer vision and multimedia research. Driven by the advent of large datasets pairing images and videos with natural language descriptions, encouraging progress has been made in both image and video captioning task. Building on the earlier works that describe images/videos with a single sentence, some recent works focus on visual paragraph generation which aims to provide detailed descriptions for images/videos [1], [2], [3], [4], [5], [6], [7], [8], [9], [10].

Existing literature on visual paragraph generation can be divided into two categories: a) fine-grained dense description of images or short videos [1], [2], [3], [4], [5], [6], [7], [8] and b) storytelling of photo streams [9], [10]. However, other than images and short videos clips, consumers often record long videos for important events, such as birthday party or wedding. In order to assist users to access long videos, previous works on video summarization focused on generating a shorter version of the video [11], [12], [13], [14], [15], [16], [17]. We argue that compared to visual summary, a textual story generated for a long video is more compact, semantically meaningful, and easier to search and organize. In this paper, we introduce a new problem: visual storytelling of long videos. Specifically, we aim to compose coherent and succinct stories for videos that cover entire events.

The proposed video storytelling problem is different from the aforementioned visual paragraph generation works. First, compared to dense-captioning [1], [2], [3], [4], [5], we focus on long videos with more complex event dynamics while not aiming to describe every detail presented in the video. Instead, we focus on extracting the important scenes and compose a story. Second, video storytelling is a more visually-grounded problem. This is in contrast with the storytelling of photo streams [9], [10], where the datasets [9], [18] only consist of a few photos per story. As a result, the annotated stories may depend on annotators' imagination and prior knowledge, and the challenge is to fill in the visual gap between photos. On the other hand, the challenge for video storytelling is not lack of visual data, but to compose a coherent and succinct story from abundant and complex visual data.

The task of video storytelling introduces two major challenges. First, compared to single-sentence descriptions, long stories contain more diverse sentences, where similar visual contents can be described very differently depending on the context. However, the widely-used Recurrent Neural Network (RNN) based sentence generation model tends to produce generic, repetitive and high-level descriptions [18], [19], [20]. Second, long video usually contains multiple actors, multiple locations and multiple activities. Hence it is difficult to discover the major storyline that is both coherent and succinct.

In this work, we address these two challenges with the following contributions. First, we propose a context-aware framework for multimodal semantic embedding learning. The embedding is learned through a two-step *local-to-global* process. The first step models individual clip-sentence pairs to learn a local embedding. The second step models the entire video as a sequence of clips. We design a Residual Bidirectional RNN (ResBRNN) that captures the temporal dynamics of the video and incorporates contextual information from past and future into the multimodal embedding space. The proposed ResBRNN preserves temporal coherence and increases diversity of the corresponding embeddings.

Second, we propose a Narrator model to generate stories. Given an input video, the Narrator extracts from it a sequence of important clips. Then a story is generated by retrieving a sequence of sentences that best matches the clips in the multimodal embedding space. Fig. 1 shows an example story using the proposed method. Discovering the important clips is difficult because there is no clear definition as to what visual aspects are important in terms of forming a good story. To this end, we formulate the Narrator as a reinforcement learning agent that sequentially observes an input video, and learns a

Junnan Li is with the Graduate School for Integrative Sciences and Engineering, National University of Singapore (email: lijunnan@u.nus.edu). Qi Zhao is with the Department of Computer Science and Engineering, University of Minnesota (email: qzhao@cs.umn.edu). Yongkang Wong and Mohan S. Kankanhalli is with the School of Computing, National University of Singapore (email: yongkang.wong@nus.edu.sg, mohan@comp.nus.edu.sg).

Manuscript received XXX, 201X; revised XXX, 201X.

Guests begin to walk into a church for the ceremony.
The wedding party makes their way down the aisle followed by the bride and her father.
The bride and groom face each other to recite their vows in front of the priest.
Everyone gathers for the reception as the bride and groom take the first dance
The bride and groom cut the wedding cake together and take the traditional first bite.

The wedding procession begins as groom and pastor go down the aisle.
The bridesmaids and groomsmen walk in two by two.
The bride and groom face each other and recite their wedding vow to one another
The bride and groom dance onto the dance floor.
The bride and the groom stand outside of the building and kiss as the video ends.

Fig. 1: The story written by human (left) and the proposed method (right). We show five sentences and their corresponding key frames uniformly sampled from each story.

policy to select clips that maximize the reward. We formulate the reward as the textual metric between the retrieved story and the reference stories written by human. By directly optimizing the textual metric, the Narrator can learn to discover important and diversified clips that form a good story.

Third, we have collected a new Video Story dataset (see Section IV for details) to enable this study, which is publicly available [1]. We evaluate our proposed method on this dataset, and quantitatively and qualitatively show that our method outperforms existing baselines.

The rest of the paper is organized as follows. Section II reviews the related work. Section III delineates the details of the proposed context-aware multimodal embedding and narrator network. The new Video Story dataset is described in Section IV, whereas the experimental results and discussion are presented in Section V. Section VI concludes the paper.

II. Related Work

A. Visual Paragraph Generation

Bridging vision and language is a longstanding goal in computer vision. Earlier works mainly target image captioning task, where a single-sentence factual description is generated for an image [21], [22], or a corresponding segment [23]. Recent works [1], [2] aim to provide more comprehensive and fine-grained image descriptions by generating multi-sentence paragraphs. Generating paragraphs for photo streams have also been explored. Park and Kim [9] first studied blog posts, and propose a method to map a sequence of blog images into a paragraph. Liu *et al.* [10] proposes a skip Gated Recurrent Unit (sGRU) to deal with visual variance among photos in a stream.

The progression from single-sentence to paragraph generation also appears in the video captioning task. Pioneering works [3], [4] first studied paragraph generation for cooking videos, and the more recent work [5] focuses on dense descriptions of activities. An encoder-decoder paradigm has been widely studied, where an encoder first encodes the visual content of a video, followed by a recurrent decoder to generate the sentence. Existing encoder approach include Convolutional Neural Network (CNN) with attention mechanism [24], [25], [26], mean-pooling [27], and RNN [5], [28], [29], [30], [31]. To capture the temporal structure in more detail, hierarchical RNN is proposed specifically for paragraph generation [4] where the top level RNN generates hidden vectors that are used to initialize low level RNNs to generate individual sentences.

[1]https://zenodo.org/record/2383739

A common problem in using RNN sentence decoder is that it tends to generate 'safe' descriptions that are generic and repetitive. This is because the maximum likelihood training objective encourages the use of *n-grams* that frequently appear in training sentences [18], [19], [20]. In video storytelling task, this problem can be more severe, because a story should naturally contain diverse rather than repetitive sentences. Another line of work for visual descriptions generation pose the task as a retrieval problem, where the most compatible description in the training set is transferred to a query [9], [32], [33], [34], [35], [36], [37]. Ordonez *et al.* [33] and Park and Kim [9] select candidate sentences based on visual similarity, and re-rank them to find the best match. The other approaches jointly model visual content and text to build a common embedding space for sentence retrieval [34], [35], [36], [38], [39].

In this work, we draw inspiration from both approaches. We employ CNN+RNN encoders to learn a cross-modality context-aware embedding space, and generate stories by retrieving sentences. In this way, as shown in our experiment, we are able to generate natural, diverse, and semantically-meaningful stories.

B. Video Summarization

Video summarization algorithms aim to aggregate segments of a video that capture its essence. Earlier works mainly utilize low-level visual features such as interest points, color and motion [11], [12], [40]. Recent works incorporate high-level semantics information such as visual concept [17], vision-language semantics [41], human action [42], [43], human-object interaction [14] and video category [15], [16]. However, there has been no established standard as to what constitutes a good visual summary. Evaluating the visual similarity between a machine-generated summary and a user-generated summary is difficult, and researchers need to define objective functions to measure the quality of a visual summary. Vasudevan *et al.* [44] select keyframes based on their relevance to the query. Other researchers have manually defined different criteria such as interestingness, representativeness and uniformity [13], [41]. Those criteria suffer from subjectiveness, which also makes evaluation challenging.

Yeung *et al.* [45] show that textual distance is a better measurement of semantic similarity compared to visual distance. Therefore, they propose VideoSET where video segments are annotated with text, and evaluation is performed in text domain. Sah *et al.* [46] and Chen *et al.* [47] further propose methods to generate a textual summary for a video. However,

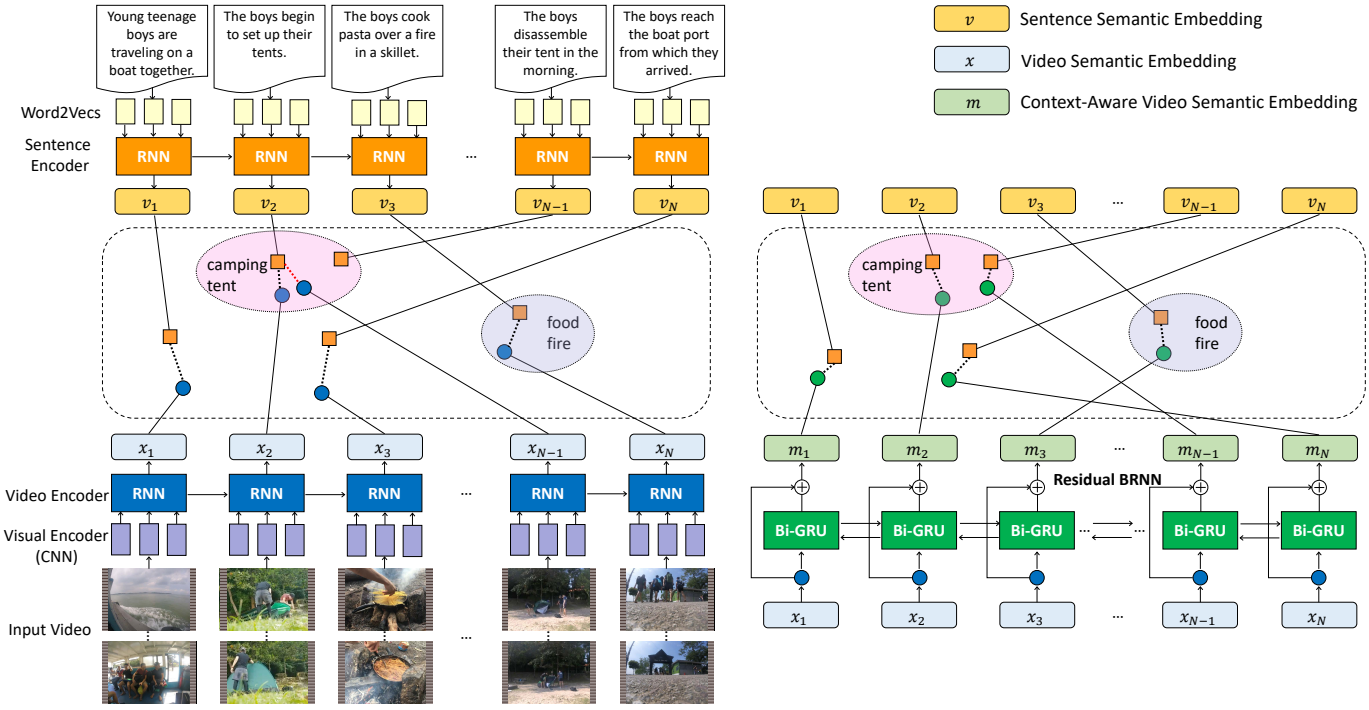

Fig. 2: *Local-to-global* multimodal embedding learning framework. Input is a sequence of video clips and sentences. Left side shows local embedding learning, where each clip is encoded by a CNN+RNN network to obtain video semantic embedding $\boldsymbol{x}$, and each sentence is encoded by a RNN to obtain sentence semantic embedding $\boldsymbol{v}$. Right side shows global embedding learning, where the entire sequence of $\boldsymbol{x}$ is fed to the ResBRNN to obtain the context-aware video semantic embedding $\boldsymbol{m}$. ResBRNN models the dynamic flow of the entire story, hence the penultimate clip can be correctly mapped to its corresponding sentence.

both methods generate disconnected captions for individual clips while ignoring the storyline of an event.

In this work, we propose a novel method to select the important video clips, which circumvents the need for predefined criteria. In contrast, we directly leverage the reference stories written by human for supervision, and learn a clip selection policy that maximizes the language metric of the generated stories. The advantage of our method is that we can train with a more well-defined and objective goal.

C. Learning Task-specific Policies

We draw inspiration from recent works that use REINFORCE [48] to learn task-specific policies. Ba *et al.* [49] and Mnih *et al.* [50] learn spatial attention policies for image classification, whereas other works apply REINFORCE for image captioning [22], [51], [52], [53]. In video domain, Yeung *et al.* [54] learn a policy for the task of action detection. Lan *et al.* [55] propose a reinforcement learning agent for video fast-forwarding.

III. Method

The proposed video storytelling method involves two subtasks: (a) discover the important clips from a long video, and (b) generate a story for the selected video clips with sentence retrieval. To address this, the proposed method includes two parts. In the first part, we propose a *context-aware multimodal embedding learning framework* where the learned embedding captures the event dynamics. In the second part, we propose a *Narrator model* that learns to discover important clips to form a succinct and diverse story.

A. Context-Aware Multimodal Embedding

We propose a two-step *local-to-global* multimodal embedding learning framework. An overview of the framework is shown in Fig. 2. In the first step, we follow the image-sentence ranking model proposed by [56] to learn a local clip-sentence embedding. In the second step, we propose ResBRNN that leverages video-story pairs to model the temporal dynamics of a video, and incorporates global contextual information into the embedding. We will first describe the clip-sentence embedding.

Encoder The model by [56] consists of two branches: a Long Short-Term Memory (LSTM) network to encode a sentence into a fixed-length vector, and a Convolutional Neural Network (CNN) followed by linear mapping to encode an image. Similarly, we construct two encoders for sentence and video, respectively. The sentence encoder is a RNN that takes the Word2Vecs [57] of a word sequence as input, and represents the sentence with its last hidden state $\boldsymbol{v}$. For the video encoder, inspired by the encoder-decoder model [28], [29], we use a RNN that take in the output of a CNN applied to each input frame, and encodes the video as its last hidden state $\boldsymbol{x}$. In terms of the recurrent unit, we use Gated Recurrent Unit (GRU) [58] instead of LSTM, because GRUs have been shown to achieve comparable performance to LSTM on several sequential modeling tasks while being simpler [59]. The GRU

uses gating units to modulate the flow of information, specified by the following operation:

$$\begin{aligned}\boldsymbol{r}_t &= \sigma(\boldsymbol{W}_{rx}\boldsymbol{x}_t + \boldsymbol{W}_{rh}\boldsymbol{h}_{t-1} + \boldsymbol{b}_r)\\ \boldsymbol{z}_t &= \sigma(\boldsymbol{W}_{zx}\boldsymbol{x}_t + \boldsymbol{W}_{zh}\boldsymbol{h}_{t-1} + \boldsymbol{b}_z)\\ \tilde{\boldsymbol{h}} &= tanh(\boldsymbol{W}_{hx}\boldsymbol{x}_t + \boldsymbol{W}_{hh}\boldsymbol{r}_t \odot \boldsymbol{h}_{t-1} + \boldsymbol{b}_h)\\ \boldsymbol{h}_t &= \boldsymbol{z}_t \odot \boldsymbol{h}_{t-1} + (1 - \boldsymbol{z}_t) \odot \tilde{\boldsymbol{h}}\end{aligned} \tag{1}$$

where $\odot$ denotes element-wise multiplication, σ are Sigmoid functions, t is the current time, $\boldsymbol{x}_t$ is the input, $\tilde{\boldsymbol{h}}$ is the current hidden state, $\boldsymbol{h}_t$ is the output. $\boldsymbol{r}_t$ and $\boldsymbol{z}_t$ are *reset gate* and *update gate*, respectively.

In this work, the dimensionality of the hidden state of both encoder RNNs are set to be 300. The Word2Vecs are 300-dimensional vectors. The CNN we use on video frame is the 101-layer ResNet [60] that outputs 2048-dimensional vectors. Following [28], [27], a video is sampled on every tenth frame.

Clip-Sentence Ranking Loss We aim to learn a joint embedding space where distance between embeddings reflects semantic relation. We assume that paired video clip and sentence share the same semantics, hence they should be closer in the embedding space. We define a similarity scoring function $s(\boldsymbol{v}, \boldsymbol{x}) = \boldsymbol{v} \cdot \boldsymbol{x}$ for sentence embedding $\boldsymbol{v}$ and clip embedding $\boldsymbol{x}$, where $\boldsymbol{v}$ and $\boldsymbol{x}$ are first scaled to have unit norm (making s equivalent to cosine similarity).

We then minimize the following pairwise ranking loss:

$$\begin{aligned}\min_{\theta_{en}} &\sum_x \sum_k \max\{0, \alpha - s(\boldsymbol{x}, \boldsymbol{v}) + s(\boldsymbol{x}, \boldsymbol{v}_k)\}\\ +\lambda &\sum_v \sum_k \max\{0, \alpha - s(\boldsymbol{v}, \boldsymbol{x}) + s(\boldsymbol{v}, \boldsymbol{x}_k)\}\end{aligned} \tag{2}$$

where θ_{en} denote all the parameters to be learned (weights of the two encoder RNNs), $\boldsymbol{v}_k$ is the negative paired (non-descriptive) sentence for clip embedding $\boldsymbol{x}$, and vice-versa with $\boldsymbol{x}_k$. The negative samples are randomly chosen from training set and re-sampled every epoch. α denotes the margin and is set to 0.1 in our experiment. The weight λ balances the strengths of the two ranking terms. We found $\lambda = 0.5$ produces the best results in our experiment.

Temporal Dynamics with Residual Bidirectional RNN One major challenge for video storytelling is that visually similar video clips could have different semantic meanings depending on the global context. For example, in Fig. 2, the second and the penultimate clip both show some people around a tent. However, depending on the sequence of events that happen before and after, their descriptions differ from 'set up their tent' to 'disassemble their tent'. It is therefore very important to model the entire flow of the story. To this end, we propose a ResBRNN model, that builds upon the embedding learned in the first step, and leverages video-story pairs to incorporate global contextual information into the embedding.

ResBRNN takes a sequence of clip embedding vectors $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$, and outputs a sequence of context-aware clip embedding vectors $\boldsymbol{M} = \{\boldsymbol{m}_1, \boldsymbol{m}_2, ..., \boldsymbol{m}_N\}$ of the same size. The role of ResBRNN is to refine the embeddings $\boldsymbol{X}$ by incorporating past and future information. Inspired by the residual mapping [60], we add a shortcut connection from the input directly to the output. Intuitively, to refine the embedding, it is much easier to add in fine-scale details via residual $\boldsymbol{M} - \boldsymbol{X}$, where the shortcut connection enables an identity mapping from $\boldsymbol{X}$ to $\boldsymbol{M}$. Compared with regular BRNN, the identity mapping provides a good initialization in the output space that is much more likely to be closer to the optimal solution. Therefore learning would be more effective. Our experiment in Section V proves that ResBRNN achieves significant performance improvement. To enable this, we re-write the GRU formula in (1) into a compact form: $\boldsymbol{h}_t = GRU(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}; \boldsymbol{W})$, and define the operation of the proposed ResBRNN as:

$$\begin{aligned}\boldsymbol{h}_t^f &= GRU(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}^f; \boldsymbol{W})\\ \boldsymbol{h}_t^b &= GRU(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}^b; \boldsymbol{W})\\ \boldsymbol{m}_t &= \boldsymbol{x}_t + \boldsymbol{h}_t^f + \boldsymbol{h}_t^b\end{aligned} \tag{3}$$

where f denotes the forward pass and b denotes the backward pass. The parameters $\boldsymbol{W}$ for the two passes are shared to reduce the number of parameters.

During training, we minimize a video-story pairwise ranking loss:

$$\begin{aligned}\min_{\theta} &\sum_m \sum_k \max\{0, \beta - s(\boldsymbol{m}, \boldsymbol{v}) + s(\boldsymbol{m}, \boldsymbol{v}_k)\}\\ +\lambda &\sum_v \sum_m \max\{0, \beta - s(\boldsymbol{v}, \boldsymbol{m}) + s(\boldsymbol{v}, \boldsymbol{m}_k)\}\end{aligned} \tag{4}$$

$\boldsymbol{v}$ and $\boldsymbol{m}$ are paired sentence and clip embeddings from the same story (or video), while $\boldsymbol{v}_k$ (or $\boldsymbol{m}_k$) are the negative paired sample to $\boldsymbol{m}$ (or $\boldsymbol{v}$). θ include the parameters of ResBRNN and the parameters of the encoders θ_{en}. β denotes the margin. We set this margin to be larger than α ($\beta = 0.2$) to apply harder constraints on the embeddings.

Optimization We first optimize the clip-sentence ranking loss (2) until validation error stops decreasing. Then we optimize the video-story ranking loss (4). We use ADAM [61] optimizer with a first momentum coefficient of 0.8 and a second momentum coefficient of 0.999.

B. Narrator Network

The goal of the Narrator is to take in a long video and output a sequence of important video clips that form a story. The model is formulated as a reinforcement learning agent that interacts with a video over time. The video is first processed by the video encoder CNN+RNN that has been trained following the method described in Section III-A. The video encoder produces a sequence of features $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$ (300-d hidden state) for N frames, where the frames are sampled on every tenth frame from the video. At timestep n, $\boldsymbol{x}_n$ contains semantic information of the current and all previous frames. The Narrator sequentially observes $\boldsymbol{X}$, and decides both when to sample a clip and the length of the sampled clip. An overview of the model is shown in Fig. 3. We now describe the Narrator in detail.

Candidate Gate At each timestep n, a binary gate R_n is used to decide whether the current position is a candidate to sample a clip. The gate is defined as

$$R_n(\boldsymbol{x}_n, \boldsymbol{x}_p) = \begin{cases}1 & \text{if } s(\boldsymbol{x}_n, \boldsymbol{x}_p) < \tau\\ 0 & \text{otherwise}\end{cases} \tag{5}$$

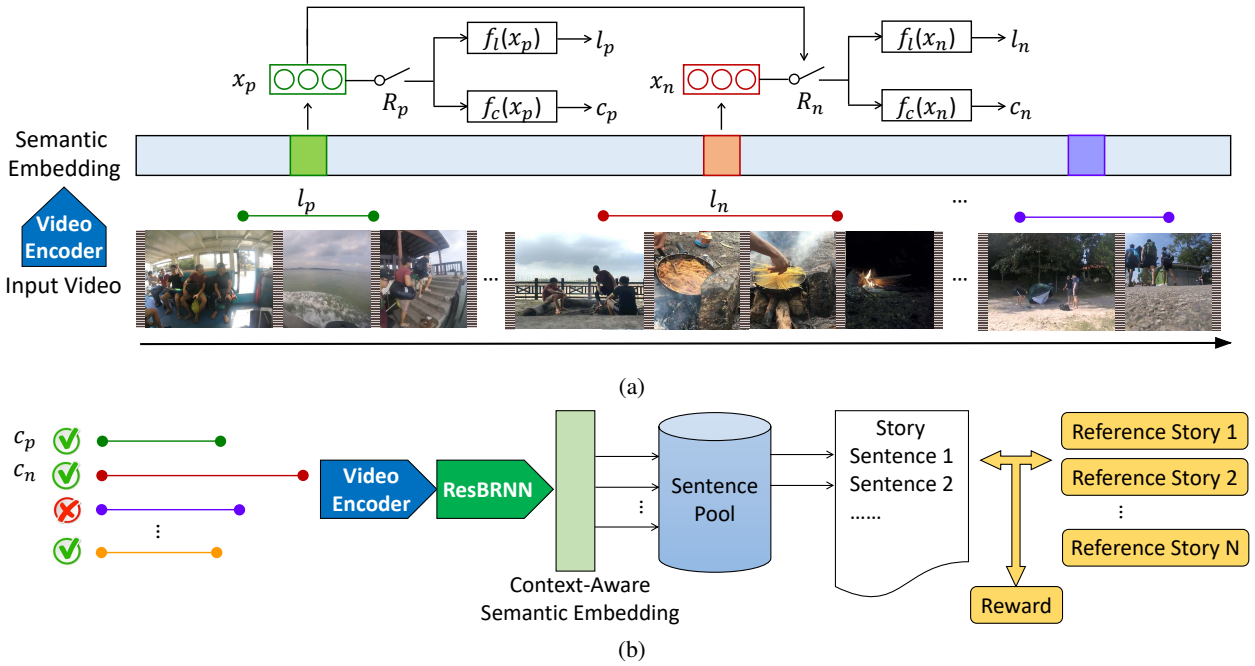

Fig. 3: (a) Narrator Network. Input is a sequence of encoded video frame features and output is a sequence of clip proposals. Here, we illustrate an example of forward pass. At timestep n, the agent observes the frame feature $\boldsymbol{x}_n$ (red), and decides the current position is a candidate ($R_n = 1$) because $\boldsymbol{x}_n$ is semantically different from the previous sampled frame (green). Then the agent produces a clip indicator c_n that indicates whether the current frame is important. If $c_n = 1$, a clip with length l_n is sampled at the current position. The agent continues proceeding until it reaches the end of the video. (b) A story is generated by retrieving sentences for the sampled clips. The textual metric of the story is computed as the reward to train the Narrator.

where $s(\boldsymbol{x}_n, \boldsymbol{x}_p) = \boldsymbol{x}_n \cdot \boldsymbol{x}_p$ is the similarity score for the normalized frame embeddings of the current position n and the previous sample position p. The candidate gate serves as an attention mechanism. It reject the current frame if it is semantically similar with the previous sampled frame. It enforces succinctness and diversity to the story, and also makes training easier by reducing the size of the agent's possible action space. In this work, we set the threshold τ to be 0.7.

Clip Indicator If $R_n = 1$ at timestep n, the current frame embedding $\boldsymbol{x}_n$ is processed by a function $f_c(\boldsymbol{x}_n; \theta_c)$ to get the clip indicator c_n, which is a binary value that signals whether a clip should be sampled at the current position. $f_c(\boldsymbol{x}_n; \theta_c)$ is a function learned from data that decides how important the current frame is for a story. During training, c_n is sampled from a Bernoulli distribution parameterized by $f_c(\boldsymbol{x}_n; \theta_c)$ defined as

$$f_c(\boldsymbol{x}_n; \theta_c) = \max(0, \sigma(\boldsymbol{W}_c \boldsymbol{x}_n) + b_c) \tag{6}$$

where $\theta_c = \{\boldsymbol{W}_c, b_c\}$, σ denotes the Sigmoid function, $\boldsymbol{W}_c$ denotes the weights for a fully-connected layer, and b_c is a scalar value to offset the probability distribution. Since our goal is to generate succinct rather than dense descriptions, the clips should be sampled sparsely from the video. Hence we initialize b_c to be -0.4, so that $f_c(\boldsymbol{x}_n; \theta_c)$ is more biased towards zero.

At test time, c_n is computed as

$$c_n = \begin{cases} 1 & \text{if } f_c(\boldsymbol{x}_n; \theta_c) > \epsilon \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where ϵ is a threshold that controls the succinctness of the story. In our experiment, we set $\epsilon = 0.2$ based on performance on the validation set.

Clip Length If c_n is set to 1, the agent will sample a clip centered at the current position with length l_n that refers to the number of frames (note that the frames are sampled on every tenth frame from the video). l_n is computed by a function $f_l(\boldsymbol{x}_n; \theta_l) = \kappa \cdot \sigma(\boldsymbol{W}_l \boldsymbol{x}_n + \boldsymbol{b}_l)$, where κ is a scaling constant and is set to 40. During training, l_n is stochastically sampled from a Gaussian distribution with a mean of $f_l(\boldsymbol{x}_n; \theta_l)$ and a fixed variance. At test time, $l_n = f_l(\boldsymbol{x}_n; \theta_l)$.

Storytelling As shown in Fig. 3b, we take all the clips sampled from a video and compute their context-aware semantic embeddings m following the method in Section III-A. Meanwhile, we compute the sentence semantic embeddings v for all candidate sentences. Then we generate the story by retrieving a set of sentences that best match the sampled clips in the embedding space. Note that the candidate pool consists of all sentences from the training set.

C. Narrator Training

There are two challenges for learning effective clip selection policy. First, it is difficult to visually evaluate the clip proposals [45]. Second, the sentences retrieved from training set can introduce bias to the story. To address those challenges, we learn the clip selection policy by directly optimizing the generated story in text domain. Inspired by [52], [53], we evaluate the story with NLP metrics such as BLEU [63],

TABLE I: Comparison of Video Story with existing video description&visual story datasets. Note that SIND [18] is an image dataset, and * marks the number of albums

Dataset	Domain	Num. videos	Avg. video length	Avg. text length	Avg. sent length	Vocabulary
MSR-VTT [62]	open	7180	20.65s	9.6	9.6	24,549
TaCos M-L [3]	cooking	185	5m 5s	115.9	8.2	2,833
ActivityNet Captions [5]	open	20k	180s	52.5	13.5	-
SIND [18]	Flickr album	10,117*	5 images	51.0	10.2	18,200
VideoSet [45]	TV & egocentric	11	3h 41m	376.2	8.6	1,195
Video Story	open	105	12m 35s	162.6	12.1	4,045

METEOR [64] or CIDEr [65]. Optimizing on NLP metrics has two advantages: (1) it is more practical, well-defined and semantically-meaningful compared with optimizing on the visual content. (2) the Narrator takes account of the bias from the candidate sentences.

The training of Narrator is a non-differentiable process because (1) the NLP metrics are discrete and non-differentiable, and (2) the clip indicator and clip length outputs are non-differentiable components. We take advantage of the REINFORCE algorithm [48] that enables learning of non-differentiable models. We first briefly describe REINFORCE below. Then we introduce a variance-reduced reward to learn effective clip selection policy.

REINFORCE Given a space of possible action sequences A, the policy of the agent, in our case f_c and f_l, induces a distribution $p_\theta(a)$ over $a \in A$ parameterized by $\theta = \{\theta_c, \theta_l\}$. The objective of REINFORCE is to maximize the expected reward $J(\theta)$ defined as

$$J(\theta) = \sum_{a \in A} p_\theta(a) r(a) \tag{8}$$

where $r(a)$ is a reward assigned to each individual action sequence. The gradient of the objective is

$$\nabla J(\theta) = \sum_{a \in A} p_\theta(a) \nabla \log p_\theta(a) r(a) \tag{9}$$

Maximizing J is non-trivial due to the high-dimensional space of possible action sequences. REINFORCE addresses this by approximating the gradient equation with Monte Carlo sampling:

$$\nabla J(\theta) \approx \frac{1}{M} \sum_{i=1}^{M} \sum_{n=1}^{N} \nabla \log \pi_\theta(a_n^i | x_{1:n}^i, a_{1:n-1}^i) R^i \tag{10}$$

where π_θ is the agent's current policy. At time step n, a_n is the policy's current action (i.e. clip indicator c_n and clip length l_n), $x_{1:n}$ are the past states of the environment including the current (frame embeddings), and $a_{1:n-1}$ are the past actions. R^i is the reward received by running current action sequence to the end. The approximate gradient is computed by running the agent's current policy for M episodes.

In this work, we use the CIDEr score for the generated story as the reward. Empirically, we find that optimizing over CIDEr achieves the best overall performance compared with other metrics, which has also been observed by [51], [53].

Variance-reduced reward The gradient estimate in (10) may have high variance. Therefore it is common to use a baseline reward b to reduce the variance, and so the gradient equation becomes:

$$\nabla J(\theta) \approx \frac{1}{M} \sum_{i=1}^{M} \sum_{n=1}^{N} \nabla \log \pi_\theta(a_n^i | x_{1:n}^i, a_{1:n-1}^i)(R^i - b) \tag{11}$$

The baseline reward is often estimated with a separate network [50], [51], [52], [54]. For our model, there are two drawbacks of using a separate network for baseline estimation. First, it introduces extra parameters to be learned. Second, a baseline estimator that only operates on the visual input x does not account for the bias introduced by the candidate sentences. Therefore, we propose a greedy approach that computes a different baseline for each video. For each video, we first randomly sample a set of clips, and retrieval a story using those clips. Then we repeat the above procedure K times, and take the average of the K stories' CIDEr score as the baseline reward for that video ($K = 10$). In this way, a policy receives a positive reward if it performs better than a random policy, and negative otherwise.

IV. Dataset

Since video storytelling is a new problem, we have collected a Video Story dataset to enable research in this direction. We choose four types of common and complex events (i.e. birthday, camping, Christmas and wedding), and use keyword search to download videos from Youtube. Then we manually selected 105 videos with sufficient inter-event and intra-event variation. The stories are collected via Amazon Mechanical Turk. For each video we asked crowd workers to write their stories following three rules: (1) The story should have at least 8 sentences. (2) Each sentence should have at least 6 words. (3) The story should be coherent and relevant to the event. We then asked workers to label the start and end time in the video where each sentence occurred. Each video is annotated by at least 5 different workers. In total, we collected 529 stories.

Table I shows the statistics of the Video Story dataset in comparison with existing datasets. Compared with video captioning datasets (i.e. MSR-VTT [62], TaCos M-L [3], ActivityNet Captions [5]) and album story dataset (i.e. SIND [18]), Video Story dataset has longer videos (12 min 35 sec in average) and longer descriptions (162.6 words in average). Moreover, the sentences in Video Story are more sparsely distributed across the video (55.77 sec per sentence). Compared with VideoSet [45] for video summarization study that only has 11 videos, Video Story dataset has more videos in open domain, the sentences are more diverse, and the vocabulary size is also larger.

V. EXPERIMENT

We compare the proposed approach with state-of-the-art methods and variations of the proposed model using quantitative measures and user study. Experiments are performed on the Video Story dataset. We randomly split 70% (73 videos) as training set, 15% (16 videos) as validation set and the others (16 videos with 4 per event category) as test set. We perform different tasks to validate the proposed multimodal embedding learning framework and the Narrator model.

A. Multimodal Embedding Evaluation

1) Sentence retrieval task: The goal of this task is to retrieve a sequence of sentences given a sequence of clips as query. We consider a withheld test set, and evaluate the median rank of the closest ground truth (GT) sentence and Recall@K, which measures the fraction of times a GT sentence was found among the top K results. Since the temporal location of clips are sparse in the Video Story dataset, overlapping clips usually share similar description. Therefore for a clip, we also include the paired sentences for its overlapping clips into the GT sentences. In average, each clip has 4.44 GT sentences.

In this work, we compare with the following baseline methods to fully evaluate our model.

- **Random:** A naive baseline where sentences are randomly ranked.
- **Category Random:** Only sentences from the same category as the test clip are considered candidates in the random ranking.
- **CCA:** Canonical Correlation Analysis (CCA) has been used to build an image-text space [66]. We use CCA to learn a joint video-sentence space. The video feature is the average of frame-level ResNet outputs, and the sentence is represented by the average of Word2Vecs. Then we project clips and sentences into the joint space for ranking.
- **m-RNN [21]:** m-RNN consists of two sub-networks that encode sentence and image respectively, which interact with each other in a multimodal layer. To apply m-RNN for video data, we replace the original CNN image encoder with mean-pooling of frame-level ResNet outputs.
- **Xu *et al.* [35]:** This is a sentence-video joint embedding model, where sentences are embedded with a dependency-tree structure model. We replace the original CNN with ResNet.
- **EMB:** This is the first local step of our embedding learning framework. Sentences are ranked based on their distance from the video in the embedding space. Note that EMB is similar to the image-sentence ranking model in [10], [56] and the movie-text ranking model in [36].
- **BRNN:** We use a BRNN in the second step to model the context, which makes it a variant to our model without the residual mapping.
- **ResBRNN:** The proposed context-aware multimodal embedding learning framework.

The results are shown in Table II. Our local embedding learning method (EMB) is a strong baseline that outperforms existing methods. The proposed context-aware embedding learning framework (ResBRNN) further improves the performance by a large margin. The improvement suggests that by incorporating information from contextual clips, the model can learn temporal coherent embeddings that are more representative of the underlying storyline. BRNN without residual mapping performs worse than ResBRNN and EMB, which demonstrates the importance of residual mapping.

TABLE II: Sentence retrieval results. **R@K** is Recall@K (higher is better). **Medr** is the median rank (lower is better)

Method	R@1	R@5	R@10	Medr
Random	0.31	1.38	3.98	215
Category Random	1.18	5.47	16.03	57
CCA	2.37	12.83	23.25	46
Xu *et al.* [35]	4.72	19.85	35.46	31
m-RNN [21]	5.34	21.23	39.02	29
EMB	5.50	22.02	40.52	27
BRNN	5.50	20.26	36.39	29
ResBRNN	**7.44**	**25.77**	**46.41**	**22**

2) Story generation task: In this task, we generate stories for test videos with different methods. In order to evaluate our embedding learning framework, we need to first fix on a good sequence of clips to represent the story visually. We employ a 'Pseudo-GT' clip selection scheme. For each test video, we compare a human annotator's selected clips with other annotators', and choose the sequence of clips that have the largest overlap with others, where overlap is computed as intersection-over-union (IoU). For quantitative measures, we compare the generated story with reference stories and compute NLP metrics of language similarity (i.e. BLEU-N [63], ROUGE-L [67], METEOR [64] and CIDEr [65]) using MSCOCO evaluation code [68]. We would like to first clarify that each of the above metrics has its own strengths and weaknesses [69]. For example, BLEU metric computes an *n-gram* based precision, thus it favors short and generic descriptions, and may not be a good measure for fully evaluating the semantic similarity of long paragraphs. CIDEr and METEOR have shown reasonable correlation with human judgments in image captioning [65], [69]. Therefore, we consider those two metrics as primary indicators of a model's performance. Nonetheless, we report results for all metrics.

For this task, we compare against the baselines in *sentence retrieval* task as well as two state-of-the-art generative models for video captioning and video paragraph generation. The additional baselines are as follow.

- **S2VT [28]**: This model encodes each clip with a CNN+RNN model and decodes the sentence with another RNN.
- **H-RNN [4]:** This model is designed for generating multiple sentences, where it uses previous sentences to initialize the hidden state for next sentence generation.
- **ResBRNN-kNN:** ResBRNN retrieves the best sentence for each clip in a sequence, which may result in duplicate sentences. To increase diversity, we apply k-nearest search. For each clip, we first find its k-nearest sentences. Then we consider the entire sequence, and find the best sequence of *non-duplicate* sentences that have the

TABLE III: Evaluation of story generation on Video Story test set with fixed Pseudo-GT clips.

Method	Model Type	CIDEr	METEOR	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
S2VT* [28]	generative	64.0	14.3	28.6	63.3	40.6	24.6	15.4
H-RNN* [4]	generative	64.6	15.5	28.8	61.6	41.4	**26.3**	**16.1**
Random	retrieval	30.2	13.1	21.4	43.1	23.1	10.0	4.8
CCA	retrieval	71.8	16.5	26.7	60.1	34.7	11.8	10.1
Xu *et al.* [35]	retrieval	79.5	17.7	28.0	61.7	36.4	20.2	11.5
m-RNN [21]	retrieval	81.3	18.0	28.5	61.9	37.0	21.1	11.8
EMB	retrieval	88.8	19.1	28.9	64.5	39.3	22.7	13.4
BRNN	retrieval	81.0	18.1	28.3	61.4	36.6	20.3	11.3
ResBRNN	retrieval	94.3	19.6	29.7	66.0	41.7	24.3	14.7
ResBRNN-kNN	retrieval	**103.6**	**20.1**	**29.9**	**69.1**	**43.5**	26.1	15.6

H-RNN	A boy is in the car and is opening a gift. The boy and the little girl is shown. The family is in the car. The children are playing with their gifts. The children are playing with the children. A young boy and the woman talks to the camera. The children are playing with the children. The family is in the living room and a girl is talking to the camera. The boy is playing with his gifts. The birthday girl is in the car.
EMB	The dog are all playing around with the children. There are new types of piñata that the children are enjoying using. All of the little boys are playing with the dog. An adult dressed as a clown making balloons animal for children at the party. The little boy sits down and opens the bag full of various toy. The person gets out of the car and looks at the surrounding and balloons. People at the party are playing going under the stick. The girls continue to join in the activities. The children play the game hit the piñata. All women happy and enjoying themselves they had a great party.
Proposed	The camera shows the birthday cake and snack to be served at the party. An adult dressed as a clown sits with balloons. The children gather around the picnic table for birthday cake. Now everyone gets to try out the fun new present. The children play with a themed piñata. The children play the game and hit the piñata. The girl runs around and play with the balloons. The family is outside and watch the balloons to fly into sky. The girls continue to join in the activities.
GT	The cartoon characters show up with gifts and the little boy lets them in. The kids and the cartoon characters eat cake. The little boy opens up his birthday presents. The kids hit the piñata and the boy chases a cartoon character around and beats him with the piñata stick. Everyone gets in the pool after one of the cartoon characters gets pushed in. The kids and the cartoon characters dance around with balloons and let one go in the sky. The little kids hug the cartoon character and they dance again. Everyone waves goodbye and they dance around some more.

H-RNN	The girls are in the living room with a girl is talking to the camera. The camera shows the dog. They are all in the living room. People are walking down to the party. Now they are in the car and is talking to the camera. The girls are in the car and the children are shown. The girls are in front of the wedding party. Now the dog is shown. The camera goes to the bride and groom. The girls are in the living room with the dog. The man and a woman is playing.
EMB	The man and his dog are to take out their camping gear. We arrive at the campsite and get in. The couple begin to take down the tent and pack to leave the campsite. The boy and father are inside the tent and camping for the night. They explore and hike the trail around the camping area and lake. They get on a boat and go through the river. Lets talk to the camera about our camping trip. The man makes breakfast in a pan over the fire. Back at the campers a young mother holds her young kid on her and places kid sunglasses on the kid. We arrive at the campsite and get in.
Proposed	The natural of the forest is shown while the hiker and dog are on the hiking trail. A girl talks to the camera while laying down her tent. They are preparing the campsite and starting up the trail on a hike. The campers explore the woods and hike up some rocks. The couple hike along a forest trail with their dog. The campers take turns swinging over the water. The friends start a campfire. A man is cooking his s'mores on the fire. The campers are exploring the area around their campsite and the surrounding woods. They get back in the car and drive again
GT	They are driving down a one lane road then go under a bridge. A man and woman lay out a tarp then a group of people put up a tent. A woman throws stuff in the water and the dog swims in the water. A dog is walking on the rocks and jumps over the water to another rock. A man is swimming in the water and the dog is on a rock beside him. People are at a table playing cards and someone shows the camera their hand. A dog comes from under a blanket and walks to someone that pets him. People are at a table outside a restaurant talking.

Fig. 4: Example stories generated by the proposed method (ResBRNN-kNN+Narrator), two baselines (H-RNN and EMD), and a ground truth. The frames are handpicked from the set of clips proposed by Narrator and the ground truth example. Green boxes highlight the frames from GT, orange boxes highlight the frames from Narrator proposals, while red boxes highlight frames shared by both.

TABLE IV: Evaluation of story generation on Video Story test set with different clip selection methods and fixed story retrieval method

Method	CIDEr	METEOR	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Uniform	89.9	18.3	28.0	65.7	40.6	23.2	13.2
SeqDPP [70]	91.6	18.3	28.3	66.3	41.0	23.6	13.1
Submodular [13]	92.0	18.4	28.1	66.4	41.0	23.8	13.3
vsLSTM [71]	92.4	18.2	28.2	66.6	41.5	24.1	13.6
Narrator IoU	93.7	18.6	28.2	67.9	42.1	24.7	14.1
Narrator w/o $\boldsymbol{x}$	93.4	18.5	28.3	67.3	41.6	24.5	14.2
Narrator w/o f_l	96.1	19.0	29.1	68.6	42.9	25.2	14.5
Narrator	**98.4**	**19.6**	**29.5**	**69.1**	**43.0**	**25.3**	**15.0**

minimum *total* distance with the clips. We use $k = 4$ in our experiment.

Table III shows the results for story generation with fixed clip selection. Retrieval based methods generally outperform generative methods, because generative methods tend to give short, repetitive and low-diversity descriptions. Among the retrieval methods, we observe similar trend of performance improvement as in Table II. ResBRNN-kNN further improves upon ResBRNN by increasing story diversity. H-RNN has higher BLEU-3 and BLEU-4 score because the BLEU metric favors generic descriptions. Qualitative examples are shown in Fig. 4.

B. Narrator Evaluation

Here we evaluate the Narrator model for clip selection. Since our end goal is video storytelling, we directly evaluate the NLP metrics of the story retrieved from different clip proposals, which is a more well-defined and objective evaluation compared with evaluating the overlap of visual content. We fix the story retrieval method with ResBRNN-kNN, and compare Narrator with multiple baseline methods for clip selection, as described below.

- **Uniform:** As a naive baseline, we uniformly sample 12 clips per test video, while 12 approximates the average number of sentences per story.
- **SeqDPP [70]:** A probabilistic model for diverse subshot selection that learns from human-created summaries.
- **Submodular [13]:** A supervised approach for video summarization, which jointly optimizes for multiple objectives including *Interestingness, Representativeness* and *Uniformity*.
- **vsLSTM [71]:** vsLSTM is a bi-directional LSTM network proposed for video summarization. The inputs are frame-level CNN features, in our case ResNet outputs. The outputs are frame-level importance scores, which can be used to select clips following the approach in [71]. We compute the GT frame-level importance scores using the same approach as [71], where a frame is considered more important if it has been selected by more human annotators.

In addition to the above baseline methods, we evaluate several variants of the proposed Narrator as ablation study to understand the efficacy of each components.

- **Narrator IoU:** Instead of using the CIDEr score of the story as reward, we compute the overlap (IoU) between the Narrator's clip proposal and the annotators' selected clips as reward.
- **Narrator w/o $\boldsymbol{x}$:** A variant of the proposed method. Inputs are frame-level CNN features instead of the semantic embedding $\boldsymbol{x}$.
- **Narrator w/o f_l:** A variant of the proposed method without the component f_l to decide clip length. Each clip has a fixed length of 20.

Table IV shows the evaluation of the retrieved story with various clip selection methods. The proposed Narrator model outperforms all other methods. The improvement over Narrator w/o $\boldsymbol{x}$ indicates the representative power of the video semantic embedding, and the improvement over Narrator w/o f_l suggests that it is useful to have a parameterized clip length compared with fixed. The state-of-the-art video summarization methods (i.e. Submodular [13], SeqDPP [70] and vsLSTM [71]) only perform slightly better than uniform sampling. This is because the human-selected clips among different annotators have high variance, which makes frame-level supervision a weak supervision to learn from. On the other hand, the stories written by annotators are more congruent, because different set of clips can still convey similar stories, thus it is stronger supervision. For the same reason, using IoU as reward for the Narrator does not perform well.

Note that the performance gaps between clip selection methods are relatively small compared with that of Table III. This is because the ResBRNN-kNN method used in this task can already retrieve relevant and meaningful sentences for a video, hence the clip selection methods would make more subtle improvements on the quality of the story.

C. User Study

We perform user studies using Amazon Mechanical Turk to observe general users' preference on stories generated by four methods: (a) the proposed ResBRNN-kNN with Narrator (ResNarrator), (b) EMB with Pseudo-GT clips , (c) H-RNN [4] with Pseudo-GT clips and (d) GT story. The results are shown in Table V. Please see Fig. 4 for qualitative examples of the stories.

First, we compare stories generated by the ResNarrator with ones generated by the two baseline methods EMB and H-RNN. For each of the 16 videos in the test set, we ask 10 users to watch the video, read the three stories presented in random order, and rank the stories based on how well they describe the video. On average, 80.0% users prefer ResNarrator over EMB,

TABLE V: User study results. Numbers indicate the percentage of pairwise preference

Compared methods	Percentage of preference
EMB > H-RNN	70.0%
ResNarrator > H-RNN	88.8%
ResNarrator > EMB	80.0%
ResNarrator > GT	38.1%

and 88.8% prefer ResNarrator over H-RNN. which validate that the proposed method generate better stories compared with the baselines.

Then, we ask the user to do a pairwise selection between stories from ResNarrator and GT. 38.1% users prefer stories generated by ResNarrator over GT stories, which further shows its efficacy.

VI. Conclusion

In this work, we have studied the problem of video storytelling. To address the challenges posed by the diversity of the story and the length and complexity of the video, we propose a ResBRNN model for context-aware multimodal embedding learning, and a Narrator model directly optimized on the story for clip selection. We evaluate our method on a new Video Story dataset, and demonstrate the efficacy of the proposed method both quantitatively and qualitatively. One limitation of our method is that the story is limited by sentences in the training set. For future work, we intend to utilize sentences in-the-wild to further improve the diversity of the story. Furthermore, we intend to explore NLP based methods to refine the story with smoother transition between sentences.

Acknowledgment

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Strategic Capability Research Centres Funding Initiative.

References

[1] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, "A hierarchical approach for generating descriptive image paragraphs," in *CVPR*, 2017, pp. 317–325.

[2] X. Liang, Z. Hu, H. Zhang, C. Gan, and E. P. Xing, "Recurrent topic-transition GAN for visual paragraph generation," in *ICCV*, 2017, pp. 3362–3371.

[3] A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, and B. Schiele, "Coherent multi-sentence video description with variable level of detail," in *German Conference on Pattern Recognition*, 2014, pp. 184–195.

[4] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, "Video paragraph captioning using hierarchical recurrent neural networks," in *CVPR*, 2016, pp. 4584–4593.

[5] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos," in *ICCV*, 2017, pp. 706–715.

[6] N. Xu, A. Liu, Y. Wong, Y. Zhang, W. N. Y. Su, and M. S. Kankanhalli, "Dual-stream recurrent neural network for video captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[7] A. Liu, Y. Qiu, Y. Wong, Y. Su, and M. S. Kankanhalli, "A fine-grained spatial-temporal attention model for video captioning," *IEEE Access*, vol. 6, pp. 68 463–68 471, 2018.

[8] A. Liu, N. Xu, Y. Wong, J. Li, Y. Su, and M. S. Kankanhalli, "Hierarchical & multimodal video captioning: Discovering and transferring multimodal knowledge for vision to language," *Computer Vision and Image Understanding*, vol. 163, pp. 113–125, 2017.

[9] C. C. Park and G. Kim, "Expressing an image stream with a sequence of natural sentences," in *Advances in Neural Information Processing Systems*, 2015, pp. 73–81.

[10] Y. Liu, J. Fu, T. Mei, and C. W. Chen, "Let your photos talk: Generating narrative paragraph for photo stream via bidirectional attention recurrent neural networks," in *AAAI*, 2017, pp. 1445–1452.

[11] F. Dufaux, "Key frame selection to represent a video," in *ICIP*, 2000, pp. 275–278.

[12] D. B. Goldman, B. Curless, D. Salesin, and S. M. Seitz, "Schematic storyboarding for video visualization and editing," *ACM Transactions on Graphics*, vol. 25, no. 3, pp. 862–871, 2006.

[13] M. Gygli, H. Grabner, and L. J. V. Gool, "Video summarization by learning submodular mixtures of objectives," in *CVPR*, 2015, pp. 3090–3098.

[14] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video," in *CVPR*, 2013, pp. 2714–2721.

[15] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *ECCV*, ser. Lecture Notes in Computer Science, vol. 8694, 2014, pp. 540–555.

[16] K. Zhang, W. Chao, F. Sha, and K. Grauman, "Summary transfer: Exemplar-based subset selection for video summarization," in *CVPR*, 2016, pp. 1059–1067.

[17] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "Tvsum: Summarizing web videos using titles," in *CVPR*, 2015, pp. 5179–5187.

[18] F. Huang, Ting-Hao K.and Ferraro, N. Mostafazadeh, I. Misra, J. Devlin, A. Agrawal, R. Girshick, X. He, P. Kohli, D. Batra *et al.*, "Visual storytelling," in *NAACL*, 2016.

[19] B. Dai, D. Lin, R. Urtasun, and S. Fidler, "Towards diverse and natural image descriptions via a conditional GAN," in *ICCV*, 2017, pp. 2970–2979.

[20] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," in *NAACL HLT*, 2016, pp. 110–119.

[21] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Deep captioning with multimodal recurrent neural networks (m-RNN)," in *ICLR*, 2015.

[22] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015, pp. 2048–2057.

[23] A. Karpathy and F. Li, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015, pp. 3128–3137.

[24] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Attention transfer from web images for video recognition," in *ACM Multimedia*, 2017, pp. 1–9.

[25] L. Yao, A. Torabi, K. Cho, N. Ballas, C. J. Pal, H. Larochelle, and A. C. Courville, "Describing videos by exploiting temporal structure," in *ICCV*, 2015, pp. 4507–4515.

[26] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Dual-glance model for deciphering social relationships," in *ICCV*, 2017, pp. 2669–2678.

[27] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. J. Mooney, and K. Saenko, "Translating videos to natural language using deep recurrent neural networks," in *NAACL HLT*, 2015, pp. 1494–1504.

[28] S. Venugopalan, M. Rohrbach, J. Donahue, R. J. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence - video to text," in *ICCV*, 2015, pp. 4534–4542.

[29] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Transactions on Pattern Analysis and Machine Learning*, vol. 39, no. 4, pp. 677–691, 2017.

[30] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Unsupervised learning of view-invariant action representations," in *Advances in Neural Information Processing Systems*, 2018, pp. 1260–1270.

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

[32] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.

[33] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2text: Describing images using 1 million captioned photographs," in *Advances in Neural Information Processing Systems*, 2011, pp. 1143–1151.

[34] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," *TACL*, vol. 2, pp. 207–218, 2014.

[35] R. Xu, C. Xiong, W. Chen, and J. J. Corso, "Jointly modeling deep video and compositional text to bridge vision and language in a unified framework," in *AAAI*, 2015, pp. 2346–2352.

[36] Y. Zhu, R. Kiros, R. S. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *ICCV*, 2015, pp. 19–27.
[37] J. Dong, X. Li, and C. G. M. Snoek, "Predicting visual features from text for image and video caption retrieval," *IEEE Transactions on Multimedia*, vol. 20, no. 12, pp. 3377–3388, 2018.
[38] Y. H. Tsai, L. Huang, and R. Salakhutdinov, "Learning robust visual-semantic embeddings," in *ICCV*, 2017, pp. 3591–3600.
[39] F. Faghri, D. J. Fleet, J. Kiros, and S. Fidler, "VSE++: improving visual-semantic embeddings with hard negatives," in *BMVC*, 2018, p. 12.
[40] S. Lu, Z. Wang, T. Mei, G. Guan, and D. D. Feng, "A bag-of-importance model with locality-constrained coding based feature learning for video summarization," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1497–1509, 2014.
[41] B. A. Plummer, M. Brown, and S. Lazebnik, "Enhancing video summarization via vision-language embedding," in *CVPR*, 2017, pp. 1052–1060.
[42] P. Varini, G. Serra, and R. Cucchiara, "Personalized egocentric video summarization of cultural tour on user preferences input," *IEEE Transactions on Multimedia*, vol. 19, no. 12, pp. 2832–2845, 2017.
[43] A. T. de Pablos, Y. Nakashima, T. Sato, N. Yokoya, M. Linna, and E. Rahtu, "Summarization of user-generated sports video by using deep action recognition features," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2000–2011, 2018.
[44] A. B. Vasudevan, M. Gygli, A. Volokitin, and L. V. Gool, "Query-adaptive video summarization via quality-aware relevance estimation," in *ACM Multimedia*, 2017, pp. 582–590.
[45] S. Yeung, A. Fathi, and L. Fei-Fei, "VideoSET: Video summary evaluation through text," in *CVPR Workshop*, 2014.
[46] S. Sah, S. Kulhare, A. Gray, S. Venugopalan, E. Prud'hommeaux, and R. W. Ptucha, "Semantic text summarization of long videos," in *WACV*, 2017, pp. 989–997.
[47] B. Chen, Y. Chen, and F. Chen, "Video to text summary: Joint video summarization and captioning with recurrent neural networks," in *BMVC*, 2017.
[48] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, vol. 8, pp. 229–256, 1992.
[49] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," in *ICLR*, 2015.
[50] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Advances in Neural Information Processing Systems*, 2014, pp. 2204–2212.
[51] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of spider," in *ICCV*, 2017, pp. 873–881.
[52] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," in *ICLR*, 2016.
[53] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *CVPR*, 2016, pp. 7008–7024.
[54] S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei, "End-to-end learning of action detection from frame glimpses in videos," in *CVPR*, 2016, pp. 2678–2687.
[55] S. Lan, R. Panda, Q. Zhu, and A. K. Roy-Chowdhury, "Ffnet: Video fast-forwarding via reinforcement learning," in *CVPR*, 2018, pp. 6771–6780.
[56] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," in *NIPS Deep Learning Workshop*, 2014.
[57] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119.
[58] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical smachine translation," in *EMNLP*, 2014, pp. 1724–1734.
[59] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *NIPS Deep Learning Workshop*, 2014.
[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
[61] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
[62] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *CVPR*, 2016, pp. 5288–5296.
[63] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLUE: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.
[64] S. Banerjee and A. Lavie, "METEOR: An automatic metric for mt evaluation with improved correlation with human judgments," in *ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.
[65] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *CVPR*, 2015, pp. 4566–4575.
[66] R. Socher and F. Li, "Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora," in *CVPR*, 2010, pp. 966–973.
[67] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *ACL workshop*, 2004.
[68] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft COCO captions: Data collection and evaluation server," *CoRR*, vol. abs/1504.00325, 2015.
[69] M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, and E. Erdem, "Re-evaluating automatic metrics for image captioning," in *EACL*, 2017, pp. 199–209.
[70] B. Gong, W. Chao, K. Grauman, and F. Sha, "Diverse sequential subset selection for supervised video summarization," in *Advances in Neural Information Processing Systems*, 2014, pp. 2069–2077.
[71] K. Zhang, W. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *ECCV*, ser. Lecture Notes in Computer Science, vol. 9911, 2016, pp. 766–782.