Study of Video Captioning Problem

Jiaqi Su
Princeton University
jiaqis@princeton.edu

Abstract

With recent success of activity recognition on video, people have increasing interest in video captioning problem where a sentence (or more broadly speaking a paragraph) is generated to describe a video clip that captures its visual semantics. In this paper, we review various methods of video captioning in the literature, along with benchmark video captioning datasets and widely used evaluation metrics. At the end, we point out several possible future directions for video captioning problem.

1. Introduction

With recent success of activity recognition on video, people have increasing interest in advancing video understanding further by incorporating more details and more reasoning in understanding results. Video captioning problem arises naturally as the very next step where a sentence (or more broadly speaking a paragraph) is generated to describe a video clip that captures its visual semantics. Ultimately, we'd like a video captioning method that can selectively narrates what happens in a random video, either short or long, just like a human does. In this information age where exploding amount of visual data is generated every day, video captioning can have many real life applications. For example, automatic generation of captions for videos would greatly help users to filter what's interesting to them among the sheer number of videos on Youtube. Additionally, video captioning techniques will make videos accessible to the disabled.

While interesting, generating captions for videos from open domain is a very complex task for machine. Several challenges particular to video captioning problem have been noted in the literature.

First, there are limited video datasets that have accompanying text descriptions, because videos are significantly more difficult and expensive to collect and annotate than images and captioning is a much more complex and ambiguous task for human workers than, for example, recognition. Realizing the importance of having datasets right for video captioning task, some early efforts set off from constructing datasets for simple settings and closed domains, and further progress is made by leveraging internet and movie description services to collect data from a wide range of topics and to scale up. Especially recently several large-scale video captioning datasets have been proposed enabling deeper learning models, such as MSR Video-to-Text and ActivityNet Captions.

Second, there is lack of complex models that can capture the rich spatio-temporal information and dynamics of video. A video clip can consist of multiple activities which may or may not be interleaving with each other, and diverse dependencies between activities which may or may not be visually explicit. Such degree of complexity has rarely been explored in image domain and video activity recognition tasks. Most of early works treat video captioning as an analogy of image captioning task, where a single semantic representation is extracted from a video and is further sent to language model to generate sentence. This type of approaches only achieves limited success in short videos with single major activity, because all the dynamics of a video gets ignored. Therefore, recent works are paying more and more attention on ways to exploit temporal structures of video.

In the following sections, we will first discuss benchmark video captioning datasets that have been used in the literature. The commonly used evaluation metrics for video captioning task will also be briefly introduced. We will then dive into discussion of various video captioning methods, including classical works that approach video captioning from the perspective of template-based generation, video retrieval or machine translation, and more recent works that utilizes hierarchical recurrent neural networks to learn in varying temporal granularities and that pushes towards dense captioning .

2. Benchmark Datasets

While video captioning on the first sight poses similar problem as image captioning, one additional key challenge faced by video captioning is the scarcity of datasets that come with rich video descriptions, because videos are significantly more difficult and expensive to collect and annotate, not to say the increasing ambiguity of the task of describing visual content. There have been several classical benchmark datasets that are widely used in early works, which enable limited success of video caption generation via learning-based methods. However, they are usually small in scale in terms of number of videos and degree of narrations, and simple in semantics in terms of both visual and textual content. Until recently, several large-scale densely annotated video captioning datasets have been developed in order to fully exploit deep learning in generating human-level video description.

In the following, we will discuss these classical and new benchmark datasets for video captioning in details of their varying data sources, scales and annotations. Figure 1 shows detailed statistics of the datasets.

TACoS Dataset [10], as one of the earliest efforts, contains videos of different activities in the cooking domain in an indoor environment. The duration of video is preferably long, usually around magnitude of minutes. Each video is annotated with both fine-grained activity labels with temporal locations and descriptions with temporal locations by multiple Amazon Mechanical Turkers. It has a total of 18,227 video-sentence pairs on 7,206 unique time intervals. **TACoS-Multi Dataset** is an extension to the dataset with paragraph description per temporal segment, but the limitation is still the same that the setting is closed-domain and too simple for learning.

Microsoft Video Description Corpus (MSVD) [2], also referred as **Youtube Dataset** in early works, is one of the earliest open world dataset. It is a collection of Youtube clips collected on Mechanical Turk by requesting workers to pick short clips depicting a single activity. As a result, each clip lasts between 10 seconds to 25 seconds, with quite constant semantics and little temporal structure complexity. It has 1,970 videos clips in total and covers a wide range of topics such as sports, animals and music. Each clip comes with multiple parallel and independent sentences labeled by different Amazon Mechanical Turkers in a number of languages. Specifically for English, it has roughly 40 parallel sentences per video; resulting in a total number of 80k clip-description pairs. It has a vocabulary of 16k unique words; each sentence on average contains 8 words.

Montreal Video Annotation Dataset (M-VAD) [17] is a large-scale movie description dataset from the DVD descriptive video service (DVS) narrations. DVS are audio tracks describing the visual elements of a movie, produced to help visually impaired people. The dataset has 49k video clips extracted from 92 DVD movies. Each clip is accompanied with a single sentence narration from semi-automatically transcribed DVS narrations. The vocabulary usage varies according to genre of the respective movie. It's a particularly challenging dataset for video captioning task due to the high diversity of visual and textual content of movies.

MPII Movie Description Corpus (MPII-MD) [12] is another recent large-scale movie description dataset built in a way similar to M-VAD. It contains around 37,000 movie clips from 55 audio descriptions (ADs) available movies and about 31,000 movie clips of 49 Hollywood movies. Each video clip is equipped with one sentence from movie scripts and one sentence from DVD descriptive video service (DVS). Annotations of dataset are semi-automatically segmented and manually aligned with clips. Since it has been manually corrected, the alignment between video snippets and descriptions is more correct in this case than in M-VAD.

MSR Video-to-Text (MSR-VTT) [20] is a recently released large-scale video captioning benchmark, and is by far the largest video captioning dataset in terms of the number of sentences and the size of the vocabulary. It contains 10k video clips crawled from a video search engine from 20 most representative categories of video search, including news, sports etc. The duration of each clip is between 10 and 30 seconds, while the total duration is 41.2 hours. Each video clip is annotated with 20 parallel and independent sentences by multiple Amazon Mechanical Turkers, which provide a good coverage of the semantics of a video clip. There are in total 200K clip-sentence pairs with a vocabulary of 29,316 unique words.

ActivityNet Captions [5] is a recently released large-scale benchmark dataset specific for dense-captioning events. It contains 20k videos amounting to 849 video hours. The videos are collected from video search engine, covering a wide range of categories. On average, each video contains 3.65 temporally localized sentences, resulting in a total of 100k sentences. Each sentence covers an unique segment of the video and describes an event that occurs over varying span of time. On average, each sentence has length of 13.48 words, and describes 36 seconds and 31% of its respective video. The rich annotation enables explicit exploration of temporal structures.

3. Evaluation Metrics

Video captioning result is evaluated based on correctness as natural language and relevance of semantics to its respective video. The following are widely used evaluation metrics that concern the aspects.

SVO Accuracy [19] is used in early works to measure whether the generated SVO (Subject, Verb, Object) triplets coheres with ground truth. The purpose of this evaluation metrics is to focus on matching of broad semantics and ignore visual and language details.

BLEU [9] is one of the most popular metrics in the field of machine translation. The idea is measuring a numerical translation closeness between two sentences by computing

Dataset	Content	Caption Source	#Videos	#Clips	#Sentence per video	Vocabulary	Duration per video	Localization
TACoS [10]	Cooking	AMTurker	127	7K	2	-	360s	✓
MSVD [2]	Youtube	AMTurker	-	2K	40	16K	10s	
M-VAD [17]	Movie	DVS	92	49K	1	18K	6s	
MPII-MD [12]	Movie	DVS + Script	94	68K	1	25K	4s	
MSR-VTT [20]	Open	AMTurker	7,180	10K	20	29K	20s	
ActivityNet Captions [5]	Open	AMTurker	-	20k	5	-	180s	✓

Figure 1. Statistics of video captioning datasets. Specifically, ”per video” statistics refers to average numbers; ”localization” refers to whether the dataset provides temporal localization of caption annotations. For video captioning task, the preferred characteristics of datasets are large scale, open domain, large vocabulary and long sentence to ensure semantic complexity, long video to ensure visual complexity, and multiple annotations per video clip so that it is robust to occasional poor annotations.

geometric mean of n-gram match counts. As a result, it is sensitive to position mismatching of words. Also, it may favor shorter sentences, which makes it hard to adapt to complex contents.

ROUGE [9] is similar to BLEU score in the sense that they measure the n-gram overlapped sequences between the reference sentences and the generated ones. The difference is that ROUGE considers the n-gram occurrences in the total sum of the number of reference sentences while BLEU considers the occurrences in the sum of candidates. Since ROUGE metric relies highly on recall, it favors long sentences.

CIDEr [9] is originally a metric to evaluate a set of descriptive sentences for a image, which measures the consensus between candidate captioning and the reference sentences provided by human annotators. Therefore, this measure highly correlates with human judgments. This measure is different from others in the sense that it captures saliency and importance, accuracy, and grammatical correctness.

METEOR [9] is computed based on the alignment between a given hypothesis sentence and a set of candidate reference. METEOR compares exact token matches, stemmed tokens, paraphrase matches, as well as semantically similar matches using WordNet synonyms. This semantic aspect of METEOR distinguishes it from others. It is shown in the literature that METEOR is always better than BLEU and ROUGE and outperforms CIDEr when the number of references is small.

4. Methods

A good video captioning requires both local and global understanding, recognizing activities and reasoning dependencies between local activities and context. Each subsection below focuses on one methodology of approaching video captioning problem, and discusses both the backbone and various variants of it as well as its advantages and limitations, from classical ones to state-of-the-art ones.

4.1. Template-based Captioning

Following the success of image recognition and activity recognition, one naive approach is to synthesize the detected outputs into a sentence using a template to ensure grammatical correctness. Template-based language methods first split sentences into fragments (e.g. subject, verb and object) following specific rules of language grammar, and each fragment is associated with detected words (e.g. objects, actions and attributes) from visual content. Then generated fragments are composed to a sentence with predefined language template. As a result, the captioning quality highly depends on the templates of sentence and sentences are always generated with syntactical structure.

[4] is one of the earliest works that builds a concept hierarchy of actions for natural language description of human activities. [13] constructs a CRF semantic role representation for each video and uses a template model to generate a sentence. [16] proposes a two-step approach of Highest Vision Confidence (HVC) model and Factor Graph Model (FGM). It first obtains confidences on subject, verb, object and scene elements; then a factor graph model is used to infer the most likely SVO (subject, verb, object) triplet in the video; finally it generates sentence based on a template. Additionally, in the deep joint embedding model of [21], template is used for its language model to generate sentence based on SVO triplet.

Although template-based language can generate complete sentences, generated descriptions are very rigid. Meanwhile, the evaluation is usually limited to narrow domain with a small vocabulary, such as TACoS dataset. For any sufficiently rich domain, the required complexity of rules and templates makes manual design of templates unfeasible or too expensive.

4.2. Joint Embedding

Video captioning problem arises as a side product of video retrieval problem where a video is to be retrieved according to given text description. Since multi-model embedding is a common practice to solve video retrieval problem, some early works apply joint embedding approach to video and language for video captioning as well.

The framework of joint embedding consists of three components: (1) a visual model to map video to representation vector, (2) a language model to map text caption to representation vector, (3) a projection of visual representation vector and language representation vector to the shared space, by minimizing distance between the two projected vectors. The idea is that the joint embedding space is semantically continuous and ensures semantically similar items, regardless of being video or description, are close to each other. During inference time, an input video is mapped to a point in the shared space corresponding to a semantically close sentence description which is further converted to text in the inverse process of the language model.

There are many possible choices of visual model and language model as practiced in the literature. We will discuss them respectively in the following.

The simplistic form of language model could be taking bag of words or one-hot encoding as semantic representation. Based on the assumption that essential semantic meaning of a video can be captured by SVO (Subject, Verb, Object) triplets, [21] proposes a compositional language model with recursive neural network, which first extracts SVO triplets from language tree of the text description parsed using Stanford Parser and recursively applies composition function to word2Vec features of S-V pair and then SV-O pair to generate a single feature vector for the sentence. This approach helps to focus on the key words of sentence description in the training data, but is limited by rigid template. Therefore, in other works, richer language models have been applied, especially RNN-based models. For example, [6] uses a recurrent neural network to encode the words in a sentence which enables to capture all the details of the sentence. [8] additionally incorporates coherence loss, which is the perplexity of the generated sentence, to guarantee the contextual relationship among the words in the sentence so that it is coherent and smooth.

The visual model follows the progress of deep models in image domain. As usual, 4096-d fc7 layer of VGG-16 pretrained on ImageNet is extracted for each frame of a video. Aside from works that apply convolutional neural networks to a fixed temporal window of frames, [21] uses a temporal pyramid scheme to summarize the feature sequence and capture motion information.

Additionally, inspired by the rich online image data, [6] exploits web image search to help relating semantics of text with visual signals. In this configuration, the sentence embedding model consists of two branches that merge the output of visual model for web images and the output of language model for the input sentence. Specifically during training, top-K results of web image search with the input sentence as a query are collected, and embedding is computed on this set of images using the same architecture as for the video embedding. The idea is that the web images serve as priors to establish connections between visual concepts and sentences.

In general, the approach of joint embedding is effective in the scenario of videos within narrow domain since the embedding space can generalizes such finite domain well, and richer model structures boost up performance. However, it can easily fail when encountering videos with situations that haven't been seen before. Also since the embedding is of fixed length, it limits the amount of information that can be carried by video and text description.

4.3. Encoder-Decoder

Inspired by the progress in machine translation and image captioning, some other early works formulate video captioning problem partially as machine translation problem where a semantic representation is generated for a video and then is translated to natural language sentence. More formally speaking, the framework those works propose is an Encoder-Decoder structure that encodes video into semantic representation and then decodes into natural language. The benefit of translation is that now we can have an open-world vocabulary if we feed machine translation model with large text corpus, which is not hard to obtain.

[19] first proposes to use mean pooling features across all frames in the video as a simple yet reasonable semantic representation for short video clips. It translates to natural language via a two-layer LSTM using the global video semantic representation as input at every time step. LSTM decoder for text generation becomes a pretty common practice as it demonstrates its capability in natural language processing. As is later pointed out in the literature [22] [7], mean pooling collapses the temporal structure of a video, i.e. the dependencies and the ordering of activities. Therefore, many follow-up works explore improving model capability of encoding both local motion features and global temporal structures.

Attention Mechanism. The attention mechanisms in deep neural networks are inspired by humans attention that sequentially focuses on the most relevant parts of the information over time to make predictions. [22] adapts the recently proposed soft attention mechanism to balance exploitation of local temporal structure, which captures details of activities, and global temporal structure, which reflects long-term dependencies and ordering of activities. The framework first uses 3D-CNN to generate temporal features vectors which capture local temporal structure (motion features). The decoder is an LSTM with soft attention mechanism, which takes in the dynamic weighted sum of the temporal feature vectors according to attention weights generated at each time step. Specifically, attention weights are generated for all the frames based on hidden state of previous time step (which presumably summarizes all the previously generated words) and the corresponding frame's temporal feature vector. Soft attention mechanism enables the decoder to look at different temporal locations and relate activities occurring cross time span for global reasoning. It has become a common practice in future works.

As a complement to temporal attention mechanism, [24] proposes Gaze Encoding Attention Network (GEAN), which predicts spatial attention map per frame supervised by human gaze data and then generates caption from the pool of masked visual features using soft temporal attention as usual. It has been observed that gaze tracking is rather stable across subjects when watching a video, meaning that spatial attention can be connected with video captioning and human gaze data can be used as a strong training signal. Its Recurrent Gaze Prediction (RGP) model fuses the history of visual features of frames through time via GRU and predicts a gaze map per frame in temporal order. The gaze map is then applied to corresponding C3D motion feature map and fovea feature map, because the authors argue that human perceives focused regions in a high visual acuity with more neurons, while peripheral scene fields in a low resolution with fewer neurons. Independent from the rest of GEAN, RGP is trained with Hollywood2 EM dataset, which is a large-scale activity recognition dataset with human gaze data, as well as a self-constructed toy dataset VAS.

The experiments show that introducing spatial attention mechanism helps with performance for not only GEAN but also multiple other temporal attention-based video captioning methods. It demonstrate that "where to look" signals in both temporal dimension and spatial dimension are very important to solving video captioning. However, it should be noted that in contrast to soft temporal attention, spatial attention practiced here considers only visual context but not previously generated words. In addition, since RGP is trained separately from the rest of video captioning framework due to lack of large-scale video captioning dataset with human gaze data, RGP is unaware of the specific task of video captioning.

Hierarchical Neural Encoder. Another line of works focuses on refining neural encoder. Even though LSTM can deal with long video clips in principal, it has been reported that the favorable length of video clips to LSTM falls in the range of 30 to 80 frames [18]. Therefore, it's usually hard for a plain LSTM to capture the large number of long-range dependencies in video. Aiming at learning the visual features with multiple temporal granularities, [7] exploits Hierarchical Recurrent Neural Encoder (HRNE), which consists of a LSTM filter on sub-sequences of an input sequence to explore local temporal features within sub-sequences and then another layer of LSTM on top to summarize and learn temporal dependencies among sub-sequences. Such a hierarchical structure significantly reduces the length of input information flow but is still capable of exploiting temporal information over longer time. It has been noted that more LSTM layers could be added to HRNE to build multiple time-scale abstraction of the visual information. The method achieves state-of-the-art performance on video captioning benchmarks at that time. However, it requires fixed manual setting of the sub-sequence length, and thus it doesn't adapt to varying types of videos.

[1] extends the idea of hierarchical LSTM, but instead of fixed length, it uses a trainable boundary detector cell to dynamically identify discontinuity points between frames and decide when to re-initialize memory of bottom-level LSTM. The top-level LSTM still keeps track of the summarized representations and contextual information even though the bottom-level LSTM has cleared its history. The idea is that visual content as well as semantic content of a video may change abruptly across frames. Such architecture ensures that the input data following a time boundary are not misled by those seen before the boundary, and generates a hierarchical representation of the video in which each chunk is composed by homogeneous frames. The method boosts performance of HRNE, and achieves state-of-the-art performance at that time on movie description dataset like M-VAD and MPII-MD, and competitive performance on MSVD. The reason is that movie videos have more underlying hierarchies, while MSVD mainly contains short video clips with a single action, and is therefore less appropriate than M-VAD and MPII-MD to evaluate the effectiveness of the method in identifying the video temporal structure. Meanwhile, the architecture enables greater interpretability, as now the layered structure of video also gets revealed alongside the encoded semantic representation and the correspondence of words to video frames can be inferred.

The authors specifically investigate the learned boundaries, and find that the proposed boundary-aware LSTM cell can identify camera changes and appearance variations, but also detects more soft boundaries which do not correspond to shots. Replacing learned boundary detector with shot detector reduces captioning scores. Therefore, even though shots give a reasonable decomposition of the video, there are more implicit video structures that can be utilized for better captioning performance.

Aside from those progresses, [18] demonstrates that a sequence-to-sequence model (S2VT) which uses a single two-layer stacked LSTM integrating both encoding stage and decoding stage works for video to text, with the benefit of weight sharing. Such architecture has been practiced in machine translation, but not in video captioning before. The first LSTM layer in the architecture is used to model the visual frame sequence, and the next layer is used to model the output word sequence. Specifically, the model has two stages: the first LSTM layer receives a sequence of frames and encodes them while the second LSTM layer receives

Method	MSVD	MSR-VTT	M-VAD	MPII-MD
Mean-pooling [19]	26.9	23.7	6.1	6.7
Temporal Attention [22]	29.6	-	5.7	-
HRNE [7]	33.9	-	5.8	-
Boundary-aware NE [1]	32.4	-	7.3	7.0
S2VT [18]	29.8	-	6.7	7.1
GEAN [24]	-	-	7.2	7.2

Figure 2. Reported METEOR scores of Encoder-Decoder methods on common benchmark datasets. METEOR is used as it's shown in the literature to be more reliable as measurement than the others. There are many variants of the models and different runs of experiments, and we show the most commonly referred numbers in the literature for comparison. The purpose of the figure is to take a glance at how video captioning performance evolves over iterations.

the hidden representation and null padded input words; after all the frames in the video clip are exhausted, the second LSTM layer is fed with the beginning-of-sentence tag, which prompts it to start decoding its current hidden representation into a sequence of words. The experiments show that it achieves state-of-the-art performance on MSVD, M-VAD and MPII-MD, but it requires incorporating heavy-weight optical flows as input along with RGB frames.

Figure 2 shows reported METEOR scores of the discussed Encode-Decoder methods, evaluated on the benchmark datasets appearing in the related literature. It's notable that methods performing relatively well on MSVD can do poorly on M-VAD and MPII-MD. This shows the discrepancy of characteristics between MSVD and movie description datasets. In general, the performance improvement on movie description datasets is marginal compared to that on MSVD. Meanwhile, it can be observed that the major gains of performance come from introducing hierarchical structure for encoding stage and attention mechanisms.

The Encoder-Decoder framework opens up possibilities for open-domain video captioning, so that the generated description can have open-world vocabulary. The early model has the limitation of degenerating video captioning problem to image captioning problem, where all the temporal ordering and dependencies of activities of video get ignored. While follow-up works of soft attention mechanism and neural encoder have alleviated the limitation, it should be recognized that there is still a lot of temporal information of video not fully utilized, since this approach still doesn't handle well long videos or videos with complex sequence of activities. This asks for not only more advanced model but also re-posing the problem as we will discuss below.

4.4. Paragraph/Dense Captioning

The semantics complexity of a single sentence usually doesn't match that of a video in the wild that spans longer than magnitude of seconds. Therefore, the most recent works have shifted focus on to generating paragraph or dense captioning for a video. We'd like all the activities and interactions of them in a video being described in natural language, and more preferably a paragraph that has internal coherence across sentences. With introduction of large scale dataset with dense annotations of sentence descriptions, such problems become feasible to solve.

Paragraph Description. This line of works focuses on generating a long story-like caption. Some works first temporally segment the video with action localization [15] or different levels of details [11], and then generate multiple captions for those segments and connect them with natural language processing techniques. However, these methods are usually limited in inter-sentence dependencies, even though they explicitly enforce some kind of consistency criteria.

The key framework proposed by [23] is hierarchical RNN (h-RNN) for describing a long video with a paragraph consisting of multiple sentences. This framework consists of two generators: (1) a sentence generator which produces single short sentences that describe specific time intervals and video regions, and (2) a paragraph generator which takes the sentential embedding as input and uses another recurrent layer to output the paragraph state; such state is then used to initialize the sentence generator. In addition, both sentence and paragraph generators adopt recurrent layers for language modeling. It uses C3D features to model video motion and activities, and applies soft temporal attention to the feature pool before feeding into Hierarchical RNN. The model is evaluated on TACoS-Multi Dataset which provides paragraph description to video clips and MSVD which provides parallel sentences to video clip and is used as a special case where the number of sentence in the paragraph is 1. Interestingly, the experiments show that the special case h-RNN outperforms state-of-the-art single-sentence captioning methods on MSVD dataset at that time, which means the hierarchy helps not only inter-sentence dependencies but also intra-sentence dependencies. Meanwhile, h-RNN definitely outperforms baseline methods that have no hierarchy, i.e. with only the sentence generator, but not the paragraph generator.

The evaluation of paragraph generation has only been conducted on closed-domain dataset, and thus the conclusion is not necessarily applicable to general open domain dataset. This calls for large-scale open domain video dataset with paragraph description annotations. Additionally, one possible improvement identified by the authors is to leverage bi-directional RNN. In h-RNN, the sentential information flows uni-directionally through the paragraph recurrent layer, and thus misleading information will be potentially passed down when the first several sentences in a paragraph are generated incorrectly. Using bidirectional RNN for sentence generation would possibly alleviate the issue and make the model more robust to drifting.

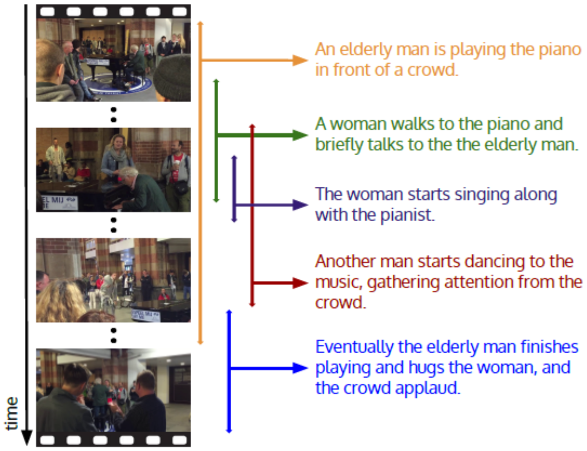
An elderly man is playing the piano in front of a crowd.
A woman walks to the piano and briefly talks to the the elderly man.
The woman starts singing along with the pianist.
Another man starts dancing to the music, gathering attention from the crowd.
Eventually the elderly man finishes playing and hugs the woman, and the crowd applaud.
time

Figure 3. Example of dense captioning of events for a video clip. Here each event has its own temporal localization, indicated by start time and end time, and caption. Multiple events can overlap in time, and can be causally related to each other.

Dense Captioning. The pioneer work in dense captioning for video is [5], which proposes the benchmark dense captioning dataset ActivityNet Captions and the task of dense captioning for events. The task is in analogy to dense captioning for image, where captions as well as their temporal localization should be provided for a video. Each caption describes an event involving a single major activity and can be overlapping with each other in temporal axis. Figure 3 shows an example of dense captioning for video. The network architecture is inspired by Deep Activity Proposal network [3], which consists of an event proposal module and a captioning module. The proposal module takes in C3D video features and generates proposals of temporal segments in terms of start time, end time and hidden vector representation at varying temporal scales. The neighboring proposals to a single proposal can be categorized into the past bucket and the future bucket based on start time. For each proposal, the captioning module uses attention mechanism to generalize past vector representation from the past bucket and future vector representation from the future bucket in order to provide temporal context, and generates caption using LSTM by feeding in the concatenation of past representation, proposal representation and future representation. This architecture allows to detect all events in a single pass of the video without need for heavy temporal sliding window.

The experiments show that including temporal context improves performance in general. This is consistent with our intuition that most events in a video are highly correlated and can even cause one another. Also the deep captioning model using ground truth proposal significantly outperforms all the other state-of-the-art methods, demonstrating the capability of its captioning module and that jointly learning localization and captioning helps video captioning quality. The reported scores on ActivityNet Captions are shown in Figure 4. This serves as a baseline benchmark for future dense captioning methods.

While temporal localization greatly helps learning of captioning, such information is not always available, and thus this is where weak supervision plays a role. [14] is the first work for dense video captioning by weakly supervised learning with only video-level sentence annotations. The architecture of the proposed approach consists of three major components: visual model, region-sequence model and language model. The visual model is a lexical-FCN trained with weakly supervised multi-instance multi-label learning, which builds the weak mapping between sentence lexical words and grid regions. The second component solves the region-sequence generation problem, which uses submodular maximization scheme to automatically generate informative and diverse region-sequences based on Lexical-FCN outputs. A winner-takes-all scheme is proposed to weakly associate sentences to region-sequences in the training phase. The third component generates sentence output for each region-sequence with a sequence-to-sequence language model.

Although the approach is trained with weakly supervised signal, the experiments show that the best single caption by the proposed approach outperforms the state-of-the-art results on the MSR-VTT challenge by a large margin. One limitation is that the framework doesn't leverage the temporal context among the dense captions, and thus it doesn't necessarily produce a consistent narration for the video clips. This is one possible future extension as is pointed out by authors. Meanwhile, the experiments are conducted on MSR-VTT, with the assumption that the 20 parallel sentences for each clip in the dataset contain very diversified annotations and can be used in the task of dense captioning. However, ActivityNet Captions may be more appropriate for the purpose.

In general, generating multiple sentences per video clip produce the most state-of-the-art results. Such problem definition allows each sentence to cover a reasonable and adaptable time scope. It provides more freedom in captioning and more detailed instructions to learning. making it possible to scale up to much longer videos. Dense captioning for video is especially an on-going topic, with many new works quickly arising.

5. Future Directions

Video captioning problem is not yet solved, as the best performance so far is still far from human-level captioning. Here, we identify several possible future directions, according to discussions in the literature and progress in related fields.

	with GT proposals						with learned proposals					
Method	B@1	B@2	B@3	B@4	M	C	B@1	B@2	B@3	B@4	M	C
Mean-pooling [19]	18.22	7.43	3.24	1.24	6.56	14.86	-	-	-	-	-	-
S2VT [18]	20.35	8.99	4.60	2.62	7.85	20.97	-	-	-	-	-	-
H-RNN [23]	19.46	8.78	4.34	2.53	8.02	20.18	-	-	-	-	-	-
Deep Captioning [5]	**26.45**	**13.48**	**7.12**	**3.98**	**9.46**	**24.56**	**17.95**	**7.69**	**3.86**	**2.20**	**4.82**	**17.29**

Figure 4. Reported Bleu (B), METEOR (M) and CIDEr (C) captioning scores for the task of dense-captioning events in [5]. The left is performances of just captioning module with ground truth proposals. The right is the combined performances of event proposal module and captioning module. The prior work has focused only on describing entire videos and not also detecting a series of events, and thus only performance using ground truth proposals get reported.

- Dense captioning is one promising direction to go. It's shown in the literature that there is high agreement in the temporal event segments among human subjects, which is in line with research suggesting that brain activity is naturally structured into semantically meaningful events. Therefore, dense captioning is a more well-defined task than general video captioning. ActivityNet Challenge starts in 2017 to include dense captioning for video as part of its competition, and therefore we are expecting more exciting works will appear in this field. With ActivityNet Captions dataset ready at hand, we are able to employ deeper models.

- Attention mechanism plays a very important role in video understanding. While works have been shown on either soft temporal attention or static spatial attention, one possible extension is soft spatio-temporal attention which dynamically looks at spatio-temporal regions, especially in the scope of dense captioning where multiple events can refer similar temporal segments but different spatial regions.

- Audio that accompany visual frames have been utilized in activity detection problem to achieve state-of-the-art performance. Audio is even more related to video captioning, as it to some extent reveals the story line and offers semantic cues. Therefore, one possible attempt is to incorporate audio data into learning.

- We have seen some works on learning with web image search to help model connect semantic concepts with visual cues, or human gaze behaviors to imitate human visual focus when describing a video. Those practices improve the performance of the underlying architectures. Therefore, one possible direction is to involve generally speaking common sense knowledge into video captioning model. This has been practiced in visual question answering, which brings improvement to performance.

- Most existing works focus on discovering objects, actions and their interactions. While they are important to semantics of captioning, high-level abstract concepts have rarely been touched. Human-level captioning is not only able to narrate all details but also able to summarize with high-level reasoning based on needs. Especially when forming a paragraph description, major events should be kept while minor events could be omit.

- The temporal structure of video is intrinsically layered. Whether being able to capture temporal structure and context greatly influences the captioning quality of the methods as is demonstrated in previous discussions. It involves local temporal structure, which requires semantically richer motion feature representations, and global temporal structure, which requires richer model of hierarchy to model diverse temporal granularities. This is the fundamental challenge we'd like to tackle.

References

[1] L. Baraldi, C. Grana, and R. Cucchiara. Hierarchical boundary-aware neural encoder for video captioning. *arXiv preprint arXiv:1611.09312*, 2016.

[2] D. L. Chen and W. B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 190–200. Association for Computational Linguistics, 2011.

[3] V. Escorcia, F. C. Heilbron, J. C. Niebles, and B. Ghanem. Daps: Deep action proposals for action understanding. In *European Conference on Computer Vision*, pages 768–784. Springer, 2016.

[4] A. Kojima, T. Tamura, and K. Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision*, 50(2):171–184, 2002.

[5] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 1, page 6, 2017.

[6] M. Otani, Y. Nakashima, E. Rahtu, J. Heikkilä, and N. Yokoya. Learning joint representations of videos and sentences with web image search. In *European Conference on Computer Vision*, pages 651–667. Springer, 2016.

[7] P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang. Hierarchical recurrent neural encoder for video representation with application to captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1029–1038, 2016.

[8] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui. Jointly modeling embedding and translation to bridge video and language. 2016.

[9] J. Park, C. Song, and J.-h. Han. A study of evaluation metrics and datasets for video captioning. In *Intelligent Informatics and Biomedical Sciences (ICIIBMS), 2017 International Conference on*, pages 172–175. IEEE, 2017.

[10] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics (TACL)*, 1:25–36, 2013.

[11] 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*, pages 184–195. Springer, 2014.

[12] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212, 2015.

[13] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele. Translating video content to natural language descriptions. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 433–440. IEEE, 2013.

[14] Z. Shen, J. Li, Z. Su, M. Li, Y. Chen, Y.-G. Jiang, and X. Xue. Weakly supervised dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, page 10, 2017.

[15] A. Shin, K. Ohnishi, and T. Harada. Beyond caption to narrative: Video captioning with multiple sentences. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 3364–3368. IEEE, 2016.

[16] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R. Mooney. Integrating language and vision to generate natural language descriptions of videos in the wild. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1218–1227, 2014.

[17] A. Torabi, C. Pal, H. Larochelle, and A. Courville. Using descriptive video services to create a large data source for video annotation research. *arXiv preprint arXiv:1503.01070*, 2015.

[18] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542, 2015.

[19] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*, 2014.

[20] J. Xu, T. Mei, T. Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 5288–5296. IEEE, 2016.

[21] 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. 2015.

[22] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*, pages 4507–4515, 2015.

[23] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4584–4593, 2016.

[24] Y. Yu, J. Choi, Y. Kim, K. Yoo, S.-H. Lee, and G. Kim. Supervising neural attention models for video captioning by human gaze data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017). Honolulu, Hawaii*, pages 2680–8, 2017.