MSR-VTT: A Large Video Description Dataset for Bridging Video and Language

Jun Xu , Tao Mei , Ting Yao and Yong Rui
Microsoft Research, Beijing, China
{v-junfu, tmei, tiyao, yongrui}@microsoft.com

Abstract

While there has been increasing interest in the task of describing video with natural language, current computer vision algorithms are still severely limited in terms of the variability and complexity of the videos and their associated language that they can recognize. This is in part due to the simplicity of current benchmarks, which mostly focus on specific fine-grained domains with limited videos and simple descriptions. While researchers have provided several benchmark datasets for image captioning, we are not aware of any large-scale video description dataset with comprehensive categories yet diverse video content.

In this paper we present MSR-VTT (standing for "MSR-Video to Text") which is a new large-scale video benchmark for video understanding, especially the emerging task of translating video to text. This is achieved by collecting 257 popular queries from a commercial video search engine, with 118 videos for each query. In its current version, MSR-VTT provides 10K web video clips with 41.2 hours and 200K clip-sentence pairs in total, covering the most comprehensive categories and diverse visual content, and representing the largest dataset in terms of sentence and vocabulary. Each clip is annotated with about 20 natural sentences by 1,327 AMT workers. We present a detailed analysis of MSR-VTT in comparison to a complete set of existing datasets, together with a summarization of different state-of-the-art video-to-text approaches. We also provide an extensive evaluation of these approaches on this dataset, showing that the hybrid Recurrent Neural Network-based approach, which combines single-frame and motion representations with soft-attention pooling strategy, yields the best generalization capability on MSR-VTT.

1. Introduction

It has been a fundamental yet emerging challenge for computer vision to automatically describe visual content with natural language. Especially, thanks to the recent development of Recurrent Neural Networks (RNNs), there has been tremendous interest in the task of image captioning, where each image is described with a single natural sentence [7, 8, 11, 14, 22, 36]. Along with this trend, researchers have provided several benchmark datasets to boost research on image captioning (e.g., Microsoft COCO [21] and Flickr 30K [41]), where tens or hundreds of thousands of images are annotated with natural sentences.

While there has been increasing interest in the task of video to language, existing approaches only achieve severely limited success in terms of the variability and complexity of video contents and their associated language that they can recognize [2, 7, 15, 28, 35, 39, 34]. This is in part due to the simplicity of current benchmarks, which mostly focus on specific fine-grained domains with limited data scale and simple descriptions (e.g., cooking [5], YouTube [16], and movie [27, 32]). There are currently no large-scale video description benchmarks that match the scale and variety of existing image datasets because videos are significantly more difficult and expensive to collect, annotate and organize. Furthermore, compared with image captioning, the automatic generation of video descriptions carries additional challenges, such as modeling the spatiotemporal information in video data and pooling strategies.

Motivated by the above observations, we present in this paper the MSR-VTT dataset (standing for MSR-Video to Text), which is a new large-scale video benchmark for video understanding, especially the emerging task of translating video to text. This is achieved by collecting 257 popular queries from a commercial video search engine, with 118 videos for each query. In its current version, MSR-VTT provides 10K web video clips with 41.2 hours and 200K clip-sentence pairs in total, covering a comprehensive list of 20 categories and a wide variety of video content. Each clip was annotated with about 20 natural sentences.

From a practical standpoint, compared with existing datasets for video to text, such as MSVD [3], YouCook [5], M-VAD [32], TACoS [25, 28], and MPII-MD [27], our MSR-VTT benchmark is characterized by the following major unique properties. First, our dataset has the largest number of clip-sentence pairs, where each video clip is annotated with multiple sentences. This can lead to a better training of RNNs and in consequence the generation of

1. A black and white horse runs around.
2. A horse galloping through an open field.
3. A horse is running around in green lush grass.
4. There is a horse running on the grassland.
5. A horse is riding in the grass.

1. A woman giving speech on news channel.
2. Hillary Clinton gives a speech.
3. Hillary Clinton is making a speech at the conference of mayors.
4. A woman is giving a speech on stage.
5. A lady speak some news on TV.

1. A child is cooking in the kitchen.
2. A girl is putting her finger into a plastic cup containing an egg.
3. Children boil water and get egg whites ready.
4. People make food in a kitchen.
5. A group of people are making food in a kitchen.

1. A man and a woman performing a musical.
2. A teenage couple perform in an amateur musical
3. Dancers are playing a routine.
4. People are dancing in a musical.
5. Some people are acting and singing for performance.

1. A white car is drifting.
2. Cars racing on a road surrounded by lots of people.
3. Cars are racing down a narrow road.
4. A race car races along a track.
5. A car is drifting in a fast speed.

1. A player is putting the basketball into the post from distance.
2. The player makes a three-pointer.
3. People are playing basketball.
4. A 3 point shot by someone in a basketball race.
5. A basketball team is playing in front of speculators.

Figure 1. Examples of the clips and labeled sentences in our MSR-VTT dataset. We give six samples, with each containing four frames to represent the video clip and five human-labeled sentences.

more natural and diverse sentences. Second, our dataset contains the most comprehensive yet representative video content, collected by 257 popular video queries in 20 representative categories (including cooking and movie) from a real video search engine. This will benefit the validation of the generalization capability of any approach for video to language. Third, the video content in our dataset is more complex than any existing dataset as those videos are collected from the Web. This plays as a ground challenge for this particular research area. Last, in addition to video content, we keep audio channel for each clip, which leaves a door opened for related areas. Fig. 1 shows some examples of the videos and their annotated sentences. We will make this dataset publically available to the research community to support future work in this area.

From a methodology perspective, we are interested in answering the following questions regarding the best performing RNN-based approaches for video to text. What is the best video representation for this specific task, either single-frame-based representation learned from the Deep Convolutional Neural Networks (DCNN) or temporal features learned from 3D CNN? What is the best pooling strategy over frames? What is the best network structure for learning spatial representation? What is the best combination of different components if considering performance and computational costs? We examine these questions empirically by evaluating multiple RNN architectures that each takes a different approach to combining learned representation across the spatiotemporal domain.

In summary, we make the following contributions in this work: 1) we build to-date the largest dataset called MSR-VTT for the task of translating video to text, which contains diverse video content corresponding to various categories and diverse textual descriptions, and 2) we summarize existing approaches for translating video to text into a single framework and comprehensively investigate several state-of-the-art approaches on different datasets.

The remaining of this paper are organized as follows. Section 2 reviews related work on vision to text. Section 3 describes the details of MSR-VTT dataset. Section 4 introduces the approaches to video to text. Section 5 presents evaluations, followed by the conclusions in Section 6.

2. Related Work

The research on vision to language has proceeded along two dimensions. One is based on language model which first detects words from visual content by object recognition and then generates a sentence with language constraints, while the other is leveraging sequence learning models (e.g., RNNs) to directly learn an embedding

between visual content and sentence. We first review the state-of-the-art research along these two dimensions. We then briefly introduce a collection of datasets for videos.

Image captioning has been taken as an emerging ground challenge for computer vision. In the language model-based approaches, objects are first detected and recognized from the images, and then the sentences can be generated with syntactic and semantic constraints [9, 18, 20, 30, 38, 42]. For example, Farhadi *et al.* perform object detection to infer a triplet of S-V-O and convert it into a sentence by predefined language templates [9]. Li *et al.* move one step further to consider the relationships among the detected objects for composing phrases [20]. Recently, researchers have explored to leverage sequence learning to generate sentences [4, 8, 11, 13, 19, 22, 36, 37]. For example, Kiros *et al.* propose to use a log-bilinear model with bias features derived from the image to model the embedding between text and image [13]. Hao *et al.* propose a three-step approach to image captioning including word detection by multiple instance learning, sentence generation by language models, and sentence reranking by deep embedding [8]. Similar works have started to adopt RNNs for generating image descriptions by conditioning the output from RNN on the image representation learned from the Convolutional Neural Network (CNN) [11, 22, 36]. In [19], Lebret *et al.* propose to use a phrase-based model rather than single word to generate sentences, while Xu *et al.* leverage visual attention mechanism to mimic human ability to compress salient visual information into descriptive language [37].

In the video domain, similar approaches have been proposed for video description generation. The first research dimension applies video representation to template-based or statistical machine translations [2, 15, 28]. These approaches generate sentences by mapping semantic sentence representation, modeled with a Conditional Random Field (CRF), to high-level concepts such as the actors, actions and objects. On the other hand, sequence learning can be applied to video description as video is naturally a sequence of objects and actions [7, 34, 35, 39, 40]. Donahue *et al.* leverage CNN to learn the single frame representation as the input to the long-term recurrent convolutional networks to output sentences [7]. In [35], Venugopalan *et al.* design an encoder-decoder neural network to generate descriptions. By mean pooling, the features over all frames can be represented by one single vector, which is the input of the RNN. Compared to mean-pooling, Li *et al.* propose to utilize the temporal attention mechanism to exploit temporal structure as well as a spatiotemporal convolutional neural network [10] to obtain local action features [39]. Besides, Venugopalan *et al.*[34] propose a end-to-end sequence-to-sequence model to generate captions for videos.

There are several existing datasets for video to text. The YouTube cooking video dataset, named YouCook [5], contains the videos about the scenes where people are cooking various recipes. Each video has a number of human-annotated descriptions about actions and objects. Similarly, TACos [25, 28] and TACos Multi-Level [26] include a set of video descriptions and temporal alignment. MSVD is a collection of 1,970 videos from Youtube with multiple categories [3]. The sentences in this dataset are annotated by AMT workers. On the other hand, M-VAD [32] and MPII-MD [27] provide movie clips with aligned Audio Description (AD) from the Descriptive Video Service (DVS) and scripts rather than human-labeled sentences. We will provide comparison and statistics for all these datasets later. In this work, we build to-date the largest video description dataset with comprehensive video context and well-defined categories. Besides, we summarize the existing approaches for translating video to text into one single framework and conduct a comprehensive evaluation of different methods.

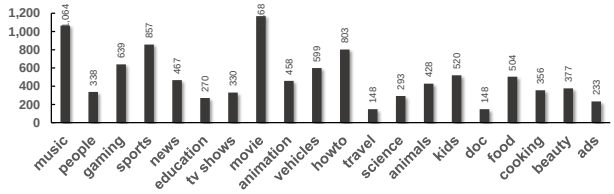


Figure 2. The distribution of video categories in our MSR-VTT dataset. This distribution well aligns with the real data statistics in a commercial video site.

3. The MSR-VTT Datataset

The MSR-VTT dataset is characterized by the unique properties including the large scale clip-sentence pairs, comprehensive video categories, diverse video content and descriptions, as well as multimodal audio and video streams. We next describe how we collect representative videos, select appropriate clips, annotate sentences, and split the dataset. Finally, we would like to summarize the existing video description datasets together and make a comparison.

3.1. Collection of Representative Videos

Current datasets for video to text mostly focus on specific fine-grained domains. For example, YouCook [5], TACoS [25, 28] and TACoS Multi-level [26] are mainly designed for cooking behavior. MSR-VTT focuses on general videos in our life, while MPII-MD [27] and M-VAD [32] on movie domain. Although MSVD [3] contains general web videos which may cover different categories, the very limited size (1,970) is far from representativeness. To collect representative videos, we obtain the top 257 representative queries from a commercial video search engine, corresponding to 20 categories [1]. We then crawl the top 150

[1] These 20 categories include music, people, gaming, sports (actions), news (events/politics), education, TV shows, movie, animation, vehicles,

Dataset	Context	Sentence Source	#Video	#Clip	#Sentence	#Word	Vocabulary	Duration (hrs)
YouCook [5]	cooking	labeled	88	–	2,668	42,457	2,711	2.3
TACos [25, 28]	cooking	AMT workers	123	7,206	18,227	–	–	–
TACos M-L [26]	cooking	AMT workers	185	14,105	52,593	–	–	–
M-VAD [32]	movie	DVS	92	48,986	55,905	519,933	18,269	84.6
MPII-MD [27]	movie	DVS+Script	94	68,337	68,375	653,467	24,549	73.6
MSVD [3]	multi-category	AMT workers	–	1,970	70,028	607,339	13,010	5.3
MSR-VTT-10K	20 categories	AMT workers	7,180	10,000	200,000	1,856,523	29,316	41.2

Table 1. Comparison of video description datasets. Please note that TACos M-L means TACos Multi-Level dataset. Although MSVD dataset has multiple video categories, the category information is not provided. In our MSR-VTT-10K dataset, we provide the category information for each clip. Among all the above datasets, MPII-MD, M-VAD and MSR-VTT contain audio information.

video search results for each query. We remove duplicate and short videos, as well as the videos with bad visual quality, to maintain the data quality. As a result, we have 30,404 representative videos. All the videos were downloaded with high quality and audio channel.

3.2. Clip Selection and Sentence Annotation

Since our goal is to collect short video clips that each can be described with one single sentence in our current version of MSR-VTT, we adopt color histogram-based approach to segmenting each video (Section 3.1) into shots [23]. As a result, there are 3,590,688 shots detected. As one video clip could have multiple consecutive shots, we asked 15 subjects to watch the videos and select appropriate consecutive shots to form video clips. For each video, at most three clips are selected to ensure the diversity of the dataset. In total there are 30K clips selected, among which we randomly selected 10K clips (originated from 7,180 videos) in the current version of MSR-VTT and left the remaining clips in our second version. The median number of shots for single video clip is 2. The duration of each clip is between 10 and 30 seconds, while the total duration is 41.2 hours.

Although one can leverage Audio Descriptions (AD) to annotate movies [27, 32], it is difficult to obtain quality sentence annotation for web videos. Therefore, we rely on Amazon Mechanical Turk (AMT) workers(1317) to annotate these clips. Each video clip is annotated by multiple workers after being watched. In the post processing, duplicated sentences and too short sentences are removed. As a result, each clip is annotated with 20 sentences by different workers. There are 200K clip-sentence pairs (corresponding to 1.8M words and 29,316 unique words) which represents the dataset with the largest number of sentences and vocabulary. Fig. 2 shows the category distribution of these 10K clips.

3.3. Dataset Split

To split the dataset to training, validation and testing sets, we separate the video clips according to the corresponding searched queries. The clips from the same video or the same queries will not appear solely in the training or testing set to avoid overfitting. We split the data according to 65%:30%:5%, corresponding to 6,513, 2,990 and 497 clips in the training, testing and validation sets, respectively.

how-to, travel, science (technology), animal, kids (family), documentary, food, cooking, beauty (fashion), advertisement.

3.4. Data Statistics

Table 1 lists the statistics and comparison among different datasets. We will release more data in the future. In this work, we denote our dataset MSR-VTT-10K as it contains 10, 000 video clips. Our MSR-VTT is the largest dataset in terms of clip-sentence pairs (200K) and word vocabulary (29,316). A major limitation for existing datasets is limited domain and annotated sentences [5, 25, 26, 28]. Although MPII-MD and M-VAD contain a number of clips, both of them are originated from one single domain (i.e., movie). The MSR-VTT is derived from a wide variety of video categories (7,180 videos from 20 general domains/categories), this can benefit the generalization capability of model learning. In addition, compared with the scripts and DVS sentences in the MPII-MD and M-VAD, since MSR-VTT has the largest vocabulary with each clip annotated with 20 different sentences, it can lead to a better training of RNNs and in consequence the generation of more natural and diverse sentences. MSVD [3], which is the most similar dataset to ours, has a small number of clips and sentences. In summary, the MSR-VTT represents the most comprehensive, diverse, and complex dataset for video to language.

4. Approaches to Video Descriptions

We briefly describe different video-to-text approaches that we benchmark on our proposed MSR-VTT dataset. Most of state-of-the-art methods for video to text are based on the Long-Short Term Memory (LSTM), which is a variant of RNN and can capture long-term temporal information by mapping sequences to sequences. As this dimension of research achieves better performance than language model-based approaches, we summarize all the RNN-based approaches in one single framework, as shown in Fig. 3. Specifically, given an input video, 2-D CNN is utilized to

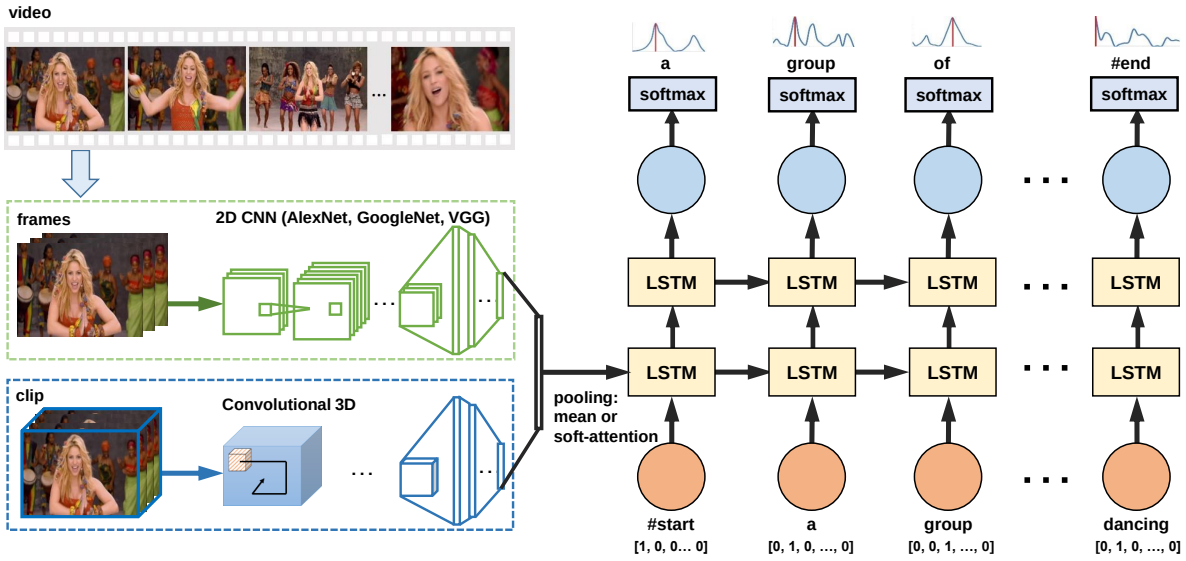


Figure 3. Summarization of RNN-based approaches for video to text. The left side reflects how to learn the representation of whole video as an input. The right side in the figure tells about the procedure of how LSTM works and the how the sentence is generated.

extract the visual representation for single frames. Moreover, optical flow or 3-D CNN is exploited to represent the motion information in the video. Then the video representation, which is comprised of 2-D CNN and/or 3-D CNN outputs, is feed into RNN. There are multiple ways that can be used to combine all the frame-level or clip-level representation to generate the video-level visual representation. The first one is mean pooling strategy which can easily create a fixed length vector as video representation [35]. The second one uses soft attention mechanism to selectively focus on only a small subset of frames instead of the simple mean pooling over the frame-level representations [39]. RNN model is trained to predict each word of the sentence after it has seen the entire video as well as all the preceding words. Please note that the video-level representation can be only input once or at each time step.

Among all the RNN-based state-of-the-art methods for translating video to text, we mainly investigate and evaluate in terms of two directions on our MSR-VTT dataset: the mean pooling model proposed in [35] and the soft attention method proposed in [39]. For mean pooling method, we design 7 runs, i.e., MP-LSTM (AlexNet), MP-LSTM (GoogleNet), MP-LSTM (VGG-16), MP-LSTM (VGG-19), MP-LSTM (C3D), MP-LSTM (C3D + VGG-16) and MP-LSTM (C3D + VGG-19). The first 5 runs employ mean pooling over the frame/clip-level features from AlexNet [17], GoogleNet [31], VGG-16 [29], VGG-19 [29] and C3D [33] networks, respectively. The last 2 runs feed the concatenations of C3D and VGG-16, C3D and VGG-19 into the LSTM model. Similarly, for soft attention strategy, we also compare 7 runs with different input frame/clip-level features, named SA-LSTM (Alex), SA-LSTM (Google), SA-LSTM (VGG-16), SA-LSTM (VGG-19), SA-LSTM (C3D), SA-LSTM (C3D + VGG-16) and SA-LSTM (C3D + VGG-19), respectively.

5. Evaluations

We conducted all the experiments on our newly created MSR-VTT-10K dataset [2] and empirically verify the RNN-based video sentence generation from three aspects: 1) when different video representation is used, 2) when different pooling strategy is exploited, and 3) how the performance is affected when using different size of hidden layer of LSTM.

5.1. Experiment Settings

To describe the visual appearances of frames in video, we adopt the output of 4096-dimensional fc6 layer from AlexNet, VGG-16 and VGG-19 and $pool5/7x7_{s1}$ layer of GoogleNet which are all pre-trained on ImageNet dataset [6]. C3D architecture, which is pre-trained on Sports-1M video dataset [12] and has been proved to be powerful in action recognition tasks, is utilized to model the temporal information across frames in video. Specifically, each continuous 16 frames are treated as one short video clip and taken as the inputs of C3D. The 4,096-dimensional outputs of fc6 layer in C3D are regarded as the representations of each video clip. Each sentence is represented as a vector of words, and each word is encoded by one-hot vector. In our experiments, we use about 20,000 most frequent words

[2] In addition to MSR-VTT-10K, we will release more data in the future.

Feature	BLEU@4	METEOR
AlexNet	6.3	14.1
GoogleNet	8.1	15.2
VGG-16	8.7	15.5
VGG-19	7.3	14.5
C3D	7.5	14.5

Table 2. The performance of KNN baselines with different video representations and mean-pooling strategy.

as the word vocabulary. In our experiments, with an initial learning rate 0.01 and mini-batch size set $1,024$, the objective value can decrease to 20% of the initial loss and reach a reasonable result after $5,000$ iterations (about 100 epochs). All videos are resized to resolution 320×240 and 30 fps.

To evaluate the generated sentences, we use the BLEU@N [24] and METEOR [1] metrics against all ground truth sentences. Both metrics are widely used in machine translation literature and already shown to be well correlated with human judgement. Specifically, BLEU@N measures the fraction of N-gram (up to 4-gram) that are in common between a hypothesis and a reference or set of references, while METEOR computes unigram precision and recall, extending exact word matches to include similar words based on WordNet synonyms and stemmed tokens.

5.2. Performance Comparison between Different Video Representations

Table 2 show the results using different video feature with mean pooling method. Since its weak performance, we will show RNN based method in the below parts.

The first experiment was conducted to examine how different video representations work on sentence generation. Table 3 shows the performances of five runs averaged over all the test videos in our dataset. It is worth noting that the performances in Table 3 are all with mean pooling. The performance trend is similar with that using soft attention.

Overall, the results across BLEU@4 and METEOR consistently indicate that video representations learnt from a temporal clip using C3D leads to a performance boost against frame-based representations. There is a slightly performance difference between VGG-19 and VGG-16. Though both runs utilize VGG network, VGG-19 is deeper than VGG-16 and thus learns a more powerful frame representations. Similar observations are also found when comparing to AlexNet and GoogleNet. The results indicate that improvement can be generally expected when learning frame-based representations by a deeper CNN.

Figure 4 (a) details the performance comparison by using VGG-19 and C3D across different categories in terms of METEOR. The two kinds of video representations show different characteristics in different types of categories. For instance, the videos in category "sports/actions" are diverse in appearance, resulting in poor performance by VGG-19.

Feature	BLEU@4	METEOR
AlexNet	35.4	26.3
GoogleNet	36.7	27.5
VGG-16	37.2	28.6
VGG-19	37.3	28.7
C3D	39.9	29.3

Table 3. BLEU@4 and METEOR for comparing the quality of sentence generation on different video representations. The experiments are all based on mean-pooling strategy, and the size of hidden layer in LSTM is set to 512. All values are reported as percentage (%).

Instead, temporal representations by C3D is found to be more helpful for this category. In the case of category "documentary," where temporal information is relatively few, frame-based representations by VGG-19 show better performance. Moreover, the complementarity between frame-based visual representations and clip-based temporal representations is generally expected.

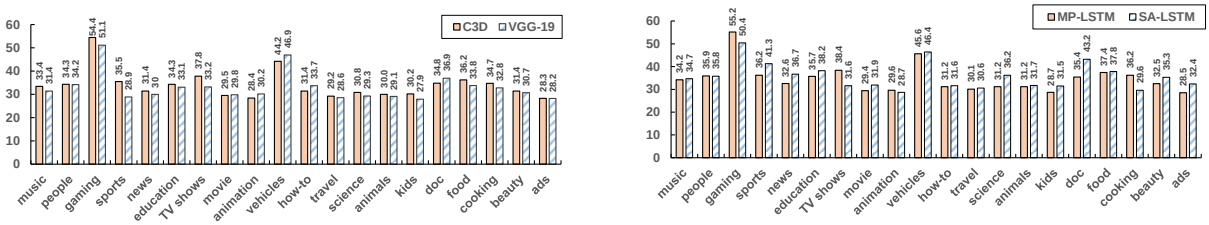
5.3. Performance Comparison between Different Pooling Strategies

We second investigated how the performance is affected with different pooling strategies. Two pooling approaches, i.e., mean pooling and soft attention model, are compared.

Table 4 lists the performances of seven video representations with mean pooling and soft attention method, respectively. Soft attention model consistently outperforms mean pooling across different video representations. In particular, the METEOR of soft attention model (SA-LSTM (C3D + VGG-19)) can achieve 29.9%, making the improvement over mean pooling (MP-LSTM (C3D + VGG-19)) by 1.4%, which is generally considered as a good progress on sentence generation task. Similar to the observations in Section 5.2, clip-based temporal representation by C3D exhibits better performance than frame-based visual representations and VGG-19 achieves the best performance across all the frame-based representations. The performances could be further boosted when concatenating the video representations learnt by C3D and VGG-19 with both mean pooling and soft attention models.

Figure 4 (b) further details the METEOR performances of mean pooling and soft attention model for all the 20 categories. Note that all the performances are given on video representations of C3D+VGG-19. Basically, different categories respond differently to the two pooling strategies. For instance, videos in the category "news" are better presented with soft attention model as there are always multiple scenes in the videos of this category. On the other hand, videos in the category "cooking" are often in a single scene and thus mean pooling shows much better results.

We also present a few sentence examples generated by different methods and ground truth in Figure 5. From



(a) Comparison between VGG-19 and C3D with mean pooling.

(b) Comparison between mean pooling and soft attention.

Figure 4. Per-category METEOR scores across all the 20 categories.

Model	BLEU@1	BLEU@2	BLEU@3	BLEU@4	METEOR
MP-LSTM (AlexNet) [35]	75.9	60.6	46.5	35.4	26.3
MP-LSTM (GoogleNet)	76.8	61.3	47.2	36.7	27.5
MP-LSTM (VGG-16)	78.0	62.0	48.7	37.2	28.6
MP-LSTM (VGG-19)	78.2	62.2	48.9	37.3	28.7
MP-LSTM (C3D)	79.8	64.7	51.7	39.9	29.3
MP-LSTM (C3D+VGG-16)	79.8	64.7	52.0	40.1	29.4
MP-LSTM (C3D+VGG-19)	79.9	64.9	52.1	40.1	29.5
SA-LSTM (AlexNet)	76.9	61.1	46.8	35.8	27.0
SA-LSTM (GoogleNet) [39]	77.8	62.2	48.1	37.1	28.4
SA-LSTM (VGG-16)	78.8	63.2	49.0	37.5	28.8
SA-LSTM (VGG-19)	79.1	63.3	49.3	37.6	28.9
SA-LSTM (C3D)	80.2	64.6	51.9	40.1	29.4
SA-LSTM (C3D+VGG-16)	81.2	65.1	52.3	40.3	29.7
SA-LSTM (C3D+VGG-19)	81.5	65.0	52.5	40.5	29.9

Table 4. Performance comparison on our MSR-VTT dataset of seven video representations with mean pooling and soft attention method, respectively. The number of the hidden layer of LSTM is set to 512 in all the experiments.

Feature	BLEU@4	METEOR
Single frame	32.4	22.6
Mean pooling	37.3	28.7
Soft-Attention	37.6	28.9

Table 5. Performance comparison among different pooling methods (with VGG-19 feature and 512 hidden layers in LSTM).

Hidden layer size	BLEU@4	METEOR	Parameters
128	32.5	26.6	3.7M
256	38.0	29.0	7.6M
512	39.9	29.3	16.3M

Table 6. Performance comparison of different size of hidden layer in LSTM. The video representation here is the clip-based temporal representations by C3D and the pooling strategy is mean pooling.

these exemplar results, it is easy to see that SA-LSTM (C3D+VGG-19) can generate more accurate sentences. For instance, compared to the sentence "Kids are playing toys" by MP-LSTM (AlexNet) and "People are playing in the room" by SA-LSTM (GoogleNet), the generated sentence "Children are painting in room" by SA-LSTM (C3D + VGG-19) encapsulates the first video more clearly. For the last video, the sentences generated by all the methods are not accurate as the video contains multiple diverse scenes. Therefore, there is still much space for researchers in this area to design new algorithms to boost performance on this dataset, especially dealing with complex visual content.

Besides, in Table 5, we have compared the performance of single frame (middle frame) with mean-pooling and soft-attention on the VGG-19 feature. It can prove approaches by pooling method can achieve a better performance than single frame.

5.4. The Size of hidden Layer of LSTM

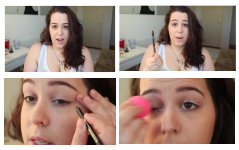
In order to show the relationship between the performance and hidden layer size of LSTM, we compare the results of the hidden layer size in the range of 128, 256, and 512. The results shown in Table 6 indicate increasing the hidden layer size can lead to the improvement of the performance with respect to both BLEU@4 and METEOR. Therefore, in our experiments, the hidden layer size is empirically set to 512 which achieves the best performance.

1. Kids are playing toys.
2. People are playing in the room.
3. **Children are painting in room.**
GT: Kids are in a classroom finger painting.

1. A man is talking.
2. A man is singing on stage.
3. **People are singing and dancing.**
GT: Two teens dancing in musical scene.

1. A woman is talking.
2. A person is folding a paper.
3. **A person is showing how to fold up a piece of paper.**
GT: A person is showing how to make an easy paper airplane.

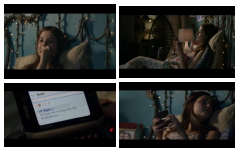
1. People playing football.
2. A group of people are playing football
3. **A group of people are playing a football game.**
GT: Two teams are playing a football game.



1. A woman Is talking.
2. A woman is talking about her.
3. **A woman is showing how to apply make up.**
GT: The woman is giving a make up tutorial.

1. A man is talking.
2. A man is talking about a computer.
3. **An advertisement for the phone.**
GT: Someone giving demo about phone.

1. Someone is making food.
2. A person is cooking.
3. **Some one is showing how to prepare a dish.**
GT: A woman shows how to prepare potatoes.



1. A woman is talking about her hair.
2. Someone is talking in bed.
3. **A woman is laying in bed.**
GT: A girl sitting in bed knocks on the wall and then begins texting someone on her phone.

Figure 5. Examples of sentence generation results from different approaches and ground truth. Each video clip is represented by four frames. (1) refers to the sentence generated by MP-LSTM (AlexNet) [35]. (2) refers to the sentence generated by SA-LSTM (GoogleNet) [39]. (3) refers to the sentence generated by SA-LSTM (C3D+VGG-19). GT is a random human generated sentence in the ground truth. Sentences in bold highlight the most accurate sentences.

Feature	Correctness	Grammar	Relevance
AlexNet	7.8	7.0	7.9
GoogleNet	6.2	6.8	6.4
VGG-16	5.3	6.9	5.4
VGG-19	5.4	6.7	5.2
C3D	5.1	6.4	5.3
C3D+VGG-16	5.1	6.1	5.0
C3D+VGG-19	4.9	6.1	5.1

Table 7. Human evaluation of different methods on MSR-VTT. Each method is evaluated by 5 persons (scale 1-10, lower is better).

5.5. Human Evaluations

We have extracted the SVO parts from all the sentences and calculated the overlapped percentages of SVO on the 20 annotated sentences to show the human consistency for our dataset. The mean overlapping percentage is 62.7%, which proves the good human consistency for all annotated sentences. Besides, we have conducted human evaluations on different approaches in our dataset in Table 7 in terms of correctness, grammar, and relevance, which shows similar results compared with the above metrics.

6. Conclusions

We introduced a new dataset for describing video with natural language. Utilizing over 3,400 worker hours, a vast collection of video-sentence pairs was collected, annotated and organized to drive the advancement of the algorithms for video to text. This dataset contains the most representative videos covering a wide variety of categories and to-date the largest amount of sentences. We comprehensively evaluated RNN-based approaches with variant components on related and our dataset. We found that the temporal representation learned from convolutional 3D networks plays strong complement to the spatial representation, and the soft-attention pooling strategy shows powerful capability to model complex and long video data.

There are several promising directions for future study on our dataset. There remains space to boost the performance on certain categories (corresponding to complex video content) that the approaches introduced in this paper cannot work well. The audio information has not been exploited for video description generation. Using audio and its AD information may further improve existing performance. The dataset can also be utilized for video summarization if one can build the embedding between video frames and the words. Furthermore, emotion and action recognition could be integrated into existing framework to make the generated language more diverse and natural.

References

[1] S. Banerjee and A. Lavie. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of ACL Workshop*, pages 65–72, 2005. 6

[2] A. Barbu, A. Bridge, Z. Burchill, D. Coroian, S. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanaswamy, D. Salvi, et al. Video in sentences out. *Proceedings of UAI*, 2012. 1, 3

[3] D. L. Chen and W. B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of ACL*, pages 190–200, 2011. 1, 3, 4

[4] X. Chen and C. L. Zitnick. Mind's Eye: A recurrent visual representation for image caption generation. In *Proceedings of CVPR*, 2015. 3

[5] P. Das, C. Xu, R. F. Doell, and J. J. Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of CVPR*, pages 2634–2641, 2013. 1, 3, 4

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of CVPR*, pages 248–255, 2009. 5

[7] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of CVPR*, 2015. 1, 3

[8] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. Platt, et al. From captions to visual concepts and back. In *Proceedings of CVPR*, 2015. 1, 3

[9] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *Proceedings of ECCV*, pages 15–29, 2010. 3

[10] S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013. 3

[11] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of CVPR*, 2015. 1, 3

[12] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of CVPR*, pages 1725–1732, 2014. 5

[13] R. Kiros, R. Salakhutdinov, and R. Zemel. Multimodal neural language models. In *Proceedings of ICML*, pages 595–603, 2014. 3

[14] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *TACL*, 2015. 1

[15] 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. 1, 3

[16] N. Krishnamoorthy, K. S. Girish Malkarnenkar, Raymond J. Mooney, and S. Guadarrama. Generating natural-language video descriptions using text-mined knowledge. In *Proceedings of AAAI*, 2013. 1

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of NIPS*, pages 1097–1105, 2012. 5

[18] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, 2013. 3

[19] R. Lebret, P. O. Pinheiro, and R. Collobert. Phrase-based image captioning. *Proceedings of ICML*, 2015. 3

[20] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. Composing simple image descriptions using web-scale N-grams. In *Proceedings of International Conference on Computational Natural Language Learning*, pages 220–228, 2011. 3

[21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of ECCV*, pages 740–755, 2014. 1

[22] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). In *Proceedings of ICLR*, 2015. 1, 3

[23] T. Mei, Y. Rui, S. Li, and Q. Tian. Multimedia search reranking: A literature survey. *ACM Computing Surveys (CSUR)*, 46(3):38, 2014. 4

[24] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318, 2002. 6

[25] 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*, 1:25–36, 2013. 1, 3, 4

[26] A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, and B. Schiele. Coherent multi-sentence video description with variable level of detail. *Pattern Recognition*, pages 184–195, 2014. 3, 4

[27] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele. A dataset for movie description. *Proceedings of CVPR*, 2015. 1, 3, 4

[28] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele. Translating video content to natural language descriptions. In *Proceedings of ICCV*, pages 433–440, 2013. 1, 3, 4

[29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of ICLR*, 2015. 5

[30] C. Sun, C. Gan, and R. Nevatia. Automatic concept discovery from parallel text and visual corpora. In *ICCV*, pages 2596–2604, 2015. 3

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of CVPR*, 2015. 5

[32] A. Torabi, C. J. Pal, H. Larochelle, and A. C. Courville. Using descriptive video services to create a large data source for video annotation research. *arXiv:1503.01070*, 2015. 1, 3, 4

[33] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. C3D: generic features for video analysis. In *Proceedings of ICCV*, 2015. 5

[34] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence–video to text. In *Proceedings of ICCV*, 2015. 1, 3

[35] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. Translating videos to natural language using deep recurrent neural networks. In *Proceedings of ACL*, 2015. 1, 3, 5, 7, 8

[36] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and Tell: A neural image caption generator. In *Proceedings of CVPR*, 2015. 1, 3

[37] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of ICML*, 2015. 3

[38] Y. Yang, C. L. Teo, H. Daumé III, and Y. Aloimonos. Corpus-guided sentence generation of natural images. In *Proceedings of Intl Conference on Empirical Methods in Natural Language Processing*, pages 444–454, 2011. 3

[39] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In *Proceedings of ICCV*, 2015. 1, 3, 5, 7, 8

[40] T. Yao, T. Mei, C.-W. Ngo, and S. Li. Annotation for free: Video tagging by mining user search behavior. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 977–986. ACM, 2013. 3

[41] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 1

[42] Z.-J. Zha, T. Mei, Z. Wang, and X.-S. Hua. Building a comprehensive ontology to refine video concept detection. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 227–236. ACM, 2007. 3