# *Literature Review On The Video Captioning Problem*

A PREPRINT

**Supervisor : Prof. Mohamed El Shenawy** *
Department of Computer and information Engineering
University Of science Technology at Zewail City
melshenawy@zewailcity.edu.eg

**May Hammad , Menah hammad**
Department of Aerospace Engineering
University Of science Technology at Zewail City
s-may.mahmod@zewailcity.edu.eg
s-menah.hammad@zewailcity.edu.eg

March 17, 2019

**ABSTRACT**

With recent success of object recognition , activity recognition on images and videos as well as the cutting edge improvements in speech classification and natural language generation ,an increasing interest in combining those techniques and developing new techniques and models to conquer the video captioning problem . where a caption or a short sentence can be produced to describe the prevailing content in a short video .This produced sentence or caption should capture the visual semantics of the video frames . In this paper we present a detailed study of all techniques and methods that was developed up to date to solve the described problem along with future possible improvements in literature .We also present used benchmark data-sets and evaluation techniques that was used to evaluate the performances of presented models and architectures .

## 1 Introduction

Describing short and long videos in natural language is one of the challenging tasks for machines. Generating automatic video descriptions involves understanding of many background artifacts and detection of their occurrences in a video employing computer vision techniques.All the information extracted from the video should be joined using a comprehensible , correct text using Natural Language Processing techniques.Automatic video description has many applications in human-robot interaction, automatic video subtitling ,it can also be used to help those who are visually impaired by generating descriptions for their surroundings and transforming it to speech.Generating sentences to describe the video content has mainly two components, understanding the visual content of the video and describing it in grammatically correct and relevant sentences.Although video description was mainly inspired by the success in image captioning,generating video descriptions is a more challenging task.It requires also to capture the tangled relations between events, actions, and objects and there flow sequence .Generating visual descriptions could be divided mainly to two related areas "Video Captioning" , "Video Description".Video Captioning is about using the visual information of a video as all morphed together and output a single sentence to describe the main event in the video .While Video Description is making use of all available information including audio to output multiple sentences that could narrate even longer video. Those descriptions are more detailed and may used together to form a coherent paragraph .

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

## 2 Benchmark datasets

Starting with the introduced Benchmark datasets that have served a Lot of classification ,object recognition and activity recognition in images and then moving on to describ video based captioning benchmarks.

### 2.1 MSCOCO

Is a large-scale object detection, segmentation, and captioning dataset.It is composed of 80k pairs of images and captions where images are sorted using semantic similarity using fc7 features .Visual Genome [1] consisting of 108,077 Images, 4 Million Region Descriptions, 1.7 Million Visual Question Answers, 3.8 Million Object Instances, 2.8 Million Attributes, 2.3 Million Relationships, Everything Mapped to Wordnet Synsets.

### 2.2 FLICKR 30K

[2]Publically available dataset: http://nlp.cs.illinois.edu/HockenmaierGroup/8k-pictures.html The Flickr 8K dataset includes images obtained from the Flickr website. University of Illinois at Urbana, Champaign has the sole link of this dataset. The images do not contain any famous person or place so that the entire image can be learnt based on all the different objects in the image. Multiple captions for each image are taken because there is a great amount of variance that is possible in the captions that can be written to describe a single image. This also helps satisfy the dynamic nature of images. There are multiple objects in the image but in a caption usually the main subject and either one or two of the secondary subjects are included in the caption.

### 2.3 TACoS Dataset

[3] Is 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.

### 2.4 Microsoft Video Description Corpus (MSVD)

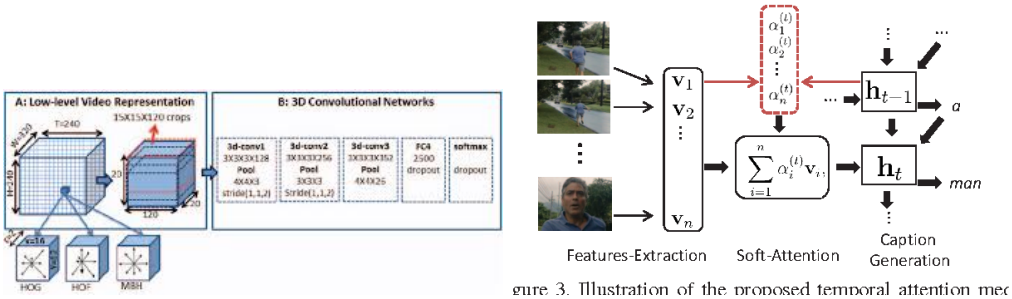
[4] Is also referred as Youtube Dataset in early works, is one of the earliest open world dataset. It is a collection of Youtube clips collected on Mechanical Turk by requesting workers to pick short clips depicting a single activity. As a result, each clip lasts between 10 seconds to 25 seconds, with 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.

## 3 Methods

A good video captioning needs to incorporate both local and global features as activity features and reasoning dependencies between local activities and context. Each subsection below focuses on one methodology of approaching video captioning problem, and discusses both the state of the art and various variants of proposed architectures .

The first method proposed in [5] is based on joint embedding and used 19-layer VGG, the hidden layer size dh of embedding$phi_v$ and $phi_s$ was set to 1,000 and the dimension of the embedding space de was set to 300. For model using GoogLeNet, we used $dh = 600$ and $de = 300$. They implemented our model using Chainer . They used Adam for optimization with a learning rate of $10^-4$ . The parameters of the CNNs and skip-thought were fixed. They applied dropout with a ratio of 0.5 to the input of the first and second layers of $phi_v$ and $phi_s$. These models were trained for 15 epochs, and their parameters were saved at every 100 updates. They took the model parameters whose performance was the best on the validation set.

The third method in [6] Have proposed to model the local temporal structure of videos at the level of the temporal features $V = v1, ..., vn$ that are extracted by the encoder. They proposed to use a spatio-temporal convolutional neural network (3-D CNN) which has shown a good capacity to capture well the temporal dynamics in video. They used a 3-D CNN to construct the higher-level representations that retain and summarize the local motion features of short frame sequences. This is done by first dividing the input video clip into a 3-D spatio-temporal grid of $16 * 12 * 2$ ,$(width * height * timesteps)$ cuboids. Each cuboid is represented by concatenating the histograms of oriented gradients, oriented flow and motion boundary (HoG, HoF, MbH) with 33 bins. This transformation is done to make sure that local temporal structure (motion features) are well extracted and to reduce the computation of the sub-sequence 3-D CNN.The used 3-D CNN architecture is composed of three 3-D convolutional layer, each followed by ReLU activation and local max-pooling. From the activation of the last 3-D convolution+ReLU+pooling layer and used as pretrained , which preserves the temporal sequential structure of the input video and abstracts the local motion features, They also obtained a set of temporal feature vectors by max-pooling along the spatial dimensions to get feature vectors that each summarize the content over short frame sequences within the video. Finally, these feature vectors are concatenated , with the image features extracted from single frames taken at similar positions across the video. They have also applied adapting the recently proposed soft attention mechanism from which allows the decoder to weight each temporal feature vector . This approach has been used successfully by for exploiting spatial structure underlying an image.




(a) . Illustration of the spatio-temporal convolutional neural network (3-D CNN).

gure 3. Illustration of the proposed temporal attention mec

(b) Illustration of the proposed temporal attention mechanism in the LSTM decoder

Figure 1

The fourth method presented in [7] is based on the detection of semantic concepts, i.e., tags, in the image under test. In order to detect such from an image, they select a set of tags from the caption text in the training set. They also use the K most common words in the training captions to determine the vocabulary of tags, which includes the most frequent nouns, verbs, or adjectives. They treat this problem as a multi-label classification task to give the model the capacity to predict semantic concepts given a test image Suppose there are N training examples, and

$$yi = [yi1, ..., yiK] \in \{0, 1\}^k$$

is the label vector of the i-th image, where $yik = 1$ if the image is annotated with tag $k$, and $yik = 0$ otherwise. Let $vi$ and $si$ represent the image feature vector and the semantic feature vector for the i-th image, the cost function to be minimized is

$$(\frac{1}{N}) * \sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} * log(s_{ik}) + (1 - y_{ik}) * log(1 - s_{ik})$$

where $si = \sigma$ ,$f(vi)$ is a K-dimensional vector with $si = [si1, ..., siK]$, $\sigma(\Delta)$ is the logistic sigmoid function and f$(\cdot)$ is implemented as a multilayer perceptron (MLP). In testing, for each input image, we compute a semantic concept vector $s$, formed by the probabilities of all tags,computed by the semantic-concept detection model. The SCN extends each weight matrix of the conventional RNN to be an ensemble of a set of tag-dependent weight matrices, subjective to the probabilities that the tags are present in the image. Specifically, the SCN-RNN computes the hidden states as follows:

$$ht = \sigma(W(s)xt - 1 + U(s)ht - 1 + z)$$

(4) where $z = 1(t = 1) \cdot Cv$, and $W(s)$ and $U(s)$ are ensembles of tag-dependent weight matrices, subjective to the probabilities that the tags are present in the image, according to the semantic-concept vector

Given $s \in R^K$, we define two weight tensors $W_T \in R^{n*x*K}$ ,$W_T[k]$ and $U_T[k]$ denote the k-th 2D "slice" of $W_T$ and $U_T$ , respectively. The probability of the k-th semantic concept, $s_k$, is associated with a pair of RNN weight matrices $WT[k]$ and $UT[k]$, implicitly specifying K RNNs in total. Consequently, training such a model as defined in (4) and (5) can be interpreted as jointly training an ensemble of K RNNs. Model learning :Given the image I and associated caption X, the objective function is the sum of the log-likelihood of the caption conditioned on the image representation:

$$logp(X|I) = \sum_{t=1}^{N} p(xt|x0, ..., xt-1, v, s)$$

This objective corresponds to a single image caption pair.
In training, we average over all training pairs. Training Procedure :For image representation, we take the output of the 2048- way pool5 layer from ResNet-152 , pretrained on the ImageNet dataset . For video representation, in addition to using the 2D ResNet-152 to extract features on each video frame, we also utilize a 3D CNN (C3D) to extract features on each video. The C3D is pretrained on Sports-1M video dataset , and we take the output of the 4096-way f c7 layer from C3D as the video representation. We consider the RGB frames of videos as input, with 2 frames per second. Each video frame is resized as $112 * 112$ and $224 * 224$ for the C3D and ResNet-152 feature extractor, respectively. The C3D feature extractor is applied on video clips of length 16 frames with an overlap of 8 frames. We use the procedure described in Section 3.2 for semantic concept detection. The semantic-concept vocabulary size is determined to reflect the complexity of the dataset, which is set to 1000, 200 and 300 for COCO, Flickr30k and Youtube2Text, respectively. Since Youtube2Text is a relatively small dataset, we found that it is very difficult to train a reliable semantic-concept detector using the Youtube2Text dataset alone, due to its limited amount of data. In experiments, we utilize additional training data from COCO. For model training, all the parameters in the SCN-LSTM are initialized from a uniform distribution in [-0.01,0.01]. All bias terms are initialized to zero. Word embedding vectors are initialized with the publicly available word2vec vectors . The embedding vectors of words not present in the

8-In [8]This model is basicaly a CNN-RNN encoder decoder .In order to effectively represent the visual content of a video, it first uses a 2-D and/or 3-D CNN, which produces a rich representation of each sampled frame/clip from the video. Then, it performs "mean pooling" process over all the frames/clips to generate a single Dv-dimensional vector v for each video V.It then uses a Long Shot-Term Memory with the video representation from the convolutional encoder to output the desired output sentence by using LSTM-type RNN model.In particular, the training of model is performed by simultaneously minimizing the relevance loss and coherence loss.Lstm -E METEOR score is 29.9

$$E(V, S) = (1 - (\lambda)) \times ||T_v v - T_s s||^2 - \lambda \times \sum_{t=1}^{N_s} [logPr_t(w_t)]) \tag{1}$$

Let N denote the number of video-sentence pairs in the training set, we have the following optimization problem

$$min(\frac{1}{N} \times (\sum_{t=1}^{+N} [E(V, S)] + ||T_s||^2 + ||T_v||^2 + ||T_h||^2 + \theta^2)) \tag{2}$$

9-Similar to previous video-to-text approaches.This method [9] applies a convolutional neural network (CNN) to input images and provide the output of the top layer as input to the LSTM unit.It uses CNNs that are pretrained on the 1.2M image subset of the ImageNet dataset . Each input video frame is scaled to 256x256, and is cropped to a random 227x227region. It is then processed by the CNN. It removes the original last fully-connected classification layer of Alexnet and learn a new linear embedding of the features to a 500 dimensional space. The lower dimension features form the input frame to the first LSTM layer. The weights of the embedding are learned jointly with the LSTM layers during training.In the first several time steps,the top LSTM layer receives a sequence of frames and encodes them while the second LSTM layer receives the hidden representation and concatenates it with null padded input words (zeros), which it then encodes. There is no loss during this stage when the LSTMs are encoding. After all the frames in the video clip are exhausted, the second LSTM layer is fed the beginning-ofsentence (<BOS>) tag, which prompts it to start decoding its current hidden representation into a sequence of words. While training in the decoding stage, the model maximizes for the log-likelihood of the predicted output sentence given the hidden representation of the visual frame sequence, and the previous words it has seen.This log-likelihood is optimized over the entire training dataset using stochastic gradient descent. The loss is computed only when the LSTM is learning to decode.The METEOR score of S2VT is 29.8

10-In this model they sample video frames once in every ten frames, then these frames could represent given video and 28.5 frames for each video averagely. Then they extract frame-wise caffe fc7 layer features using VGG-16 layers model, then feed the sequential feature into the video caption system. Where they employ two LSTMs, forward pass and backward pass, to encode CNNs features of video frames,and then merge the output sequences at each time point with a learnt weight matrix. They set the size of hidden unit of all LSTMs to 512 as except for the first video encoder in unidirectional joint LSTM. During training phrase, they set 80 as maximum number of time steps of LSTM in the model and a mini-batch with 16 video-sentence pairs.The METEOR score of Joint-LSTM unidirectinal 29.5 ;

## 4 Conclusion

Among These proposed methods we chose to use the model described as Semantic Compositional Networks for Visual Captioning( SCN) because it has achieved notable performance on competing with other encoder-decoder architectures with high BELU and Cider scores as follows:

- Model :SCN-LSTM
- BLEU-1 :0.740 ,0.917
- BLEU-2:0.575 , 0.839
- BLEU-3 :0.436, 0.739
- BLEU-4: 0.331, 0.631
- METEOR:0.257, 0.348
- ROUGE-L: 0.543, 0.696
- CIDEr-D: 1.003 ,1.013

## References


[1] Ranjay Krishna , Yuke Zhu , Oliver Groth , Justin Johnson · Kenji Hata , Joshua Kravitz , Stephanie Chen , Yannis Kalantidis · Li-Jia Li , David A. Shamma Michael S. Bernstein , Li Fei-Fei Visual Genome:Connecting Language and Vision Using Crowd sourced Dense Image Annotations Applications *arXiv preprint arXiv:1602.07332*, 2016.

[2] Emiel van Miltenburg VStereotyping and Bias in the Flickr30K Dataset *arXiv preprint arXiv:1605.06083*, 2016.

[3] Mayu Otani1, Yuta Nakashima1 , Esa Rahtu2,Janne Heikkila2 , and Naokazu Yokoya1
Joint Representations of Videos and Sentences with Web Image Search
*arXiv preprint arXiv:1608.02367*,2016

[4] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, Aaron Courville Describing Videos by Exploiting Temporal Structure *arXiv preprint arXiv:1502.08029*,2016

[5] Zhe Gan , Chuang Gan , Xiaodong He , Yunchen Pu† Kenneth Tran , Jianfeng Gao , Lawrence Carin , Li Deng Duke University, Tsinghua University, Microsoft Research, Redmond Semantic Compositional Networks for Visual Captioning
*arXiv preprint arXiv:611.08002*,2017

[6] Guy Hadash, Einat Kermany, Boaz Carmeli, Ofer Lavi, George Kour, and Alon Jacovi. Estimate and replace: A novel approach to integrating deep neural networks with existing applications. *arXiv preprint arXiv:1804.09028*, 2018.

[7] Subhashini Venugopalan and Marcus Rohrbach and Jeff Donahue and Raymond J. Mooney and Trevor Darrell and Kate Saenko Sequence to Sequence - Video to Text *http://arxiv.org/abs/1505.00487*

[8] Pan, Yingwei and Mei, Tao and Yao, Ting and Li, Houqiang and Rui, Yong Jointly Modeling Embedding and Translation to Bridge Video and Language. *http://arxiv.org/abs/1505.01861*

[9] Yi Bin ,Yang Yang ,Zi Huang , Fumin Shen , Xing Xu , Heng Tao Shen Bidirectional Long-Short Term Memory for Video Description
*http://arxiv.org/abs/1505.00487*