Domain-Specific Semantics Guided Approach to Video Captioning

M.Hemalatha and C.Chandra Sekhar
Indian Institute of Technology Madras
hemalatham.ch@gmail.com, chandra@cse.iitm.ac.in

Abstract

In video captioning, the description of a video usually relies on the domain to which the video belongs. Typically, the videos belong to wide range domains such as sports, music, news, cooking, etc. In many cases, a video can be associated with more than one domain. In this paper, we propose an approach to video captioning that uses domain-specific decoders. We build a domain classifier to obtain the estimates of probabilities of a video belonging to different domains. For each video, we identify the $top-k$ domains based on the estimated probabilities. Each video in the training data set is shared in training the domain-specific decoders of $top-k$ labels obtained from the domain classifier. The domain-specific decoders use the domain-specific semantic tags for generating captions. The proposed approach uses the Temporal VLAD for preprocessing the features extracted from 2D-CNN and 3D-CNN features. The preprocessed features provide better feature representation of the videos. The effectiveness of the proposed approach is demonstrated through the results of experimental studies on Microsoft Video Description (MSVD) corpus and MSR-VTT dataset.

1. Introduction

Video captioning is a process of generating a meaningful natural language sentence for a given video. Applications such as video understanding, human computer interaction, automatic video subtitling, and navigation by the visually challenged persons depend on video captioning. The task of video captioning is similar to image captioning. Image captioning is widely studied, and many approaches have been developed to generate captions for images [11], [12], [18],[20],[22], [35], [38], [39], [44], [45]. However, captioning a video is a more challenging task due to the temporal information present in the video in addition to the spatial information. Many approaches have been developed for video captioning [2], [3], [16], [17], [13], [21], [25], [40], [46], [47]. In many video captioning techniques, the input video is converted into an intermediate representation using an encoder. This encoder can be a handcrafted feature extractor, a CNN (Convolutional Neural Network), an RNN (Recurrent Neural Network) or a combination of them. Most of the existing video captioning techniques use a CNN developed for image classification in ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [31], [29] as an encoder to process every frame in a video. The encoder of a video captioning system usually focuses only on a single type of features like spatial features or spatio-temporal features or semantic features. A single type of features may not capture the diverse types of information present in a video. The video captioning model proposed in this work uses a fusion of different type of features such as spatial features from a 2-dimensional CNN (2D-CNN), spatio-temporal features from a 3-dimensional CNN (3D-CNN) and semantic features.

Captioning a video primarily depends on the knowledge of the domain to which the video belongs. Most of the existing approaches to video captioning assume that every video belongs only to a single domain. Very few approaches exploit the domain information of the videos. The approach in [30] uses a separate language model for every domain. In this approach, the language model for a domain is built using the captions of videos pertaining only to that domain. The approach in [5], [6] learns the topics/domains from the videos using captions and visual information during training. As the captions will not be available during testing, only the visual information is used to predict the topic of a video. It may result in an inaccurate prediction of topics for the videos.

Generating captions using domain-specific knowledge may give a better performance. While learning about a new domain, a person uses the prior knowledge related to the other domains. For example, the knowledge about the actions such as running, kicking, jumping, jogging, etc. is used to learn about sports domain. There may be an overlap in the semantic tags of different domains. Figure 1 shows frames of the videos taken from the *actions* domain and the *cooking* domain. The *cooking* domain video shows a person cooking and cutting vegetables. The *actions* domain video shows a person performing actions like cutting paper,

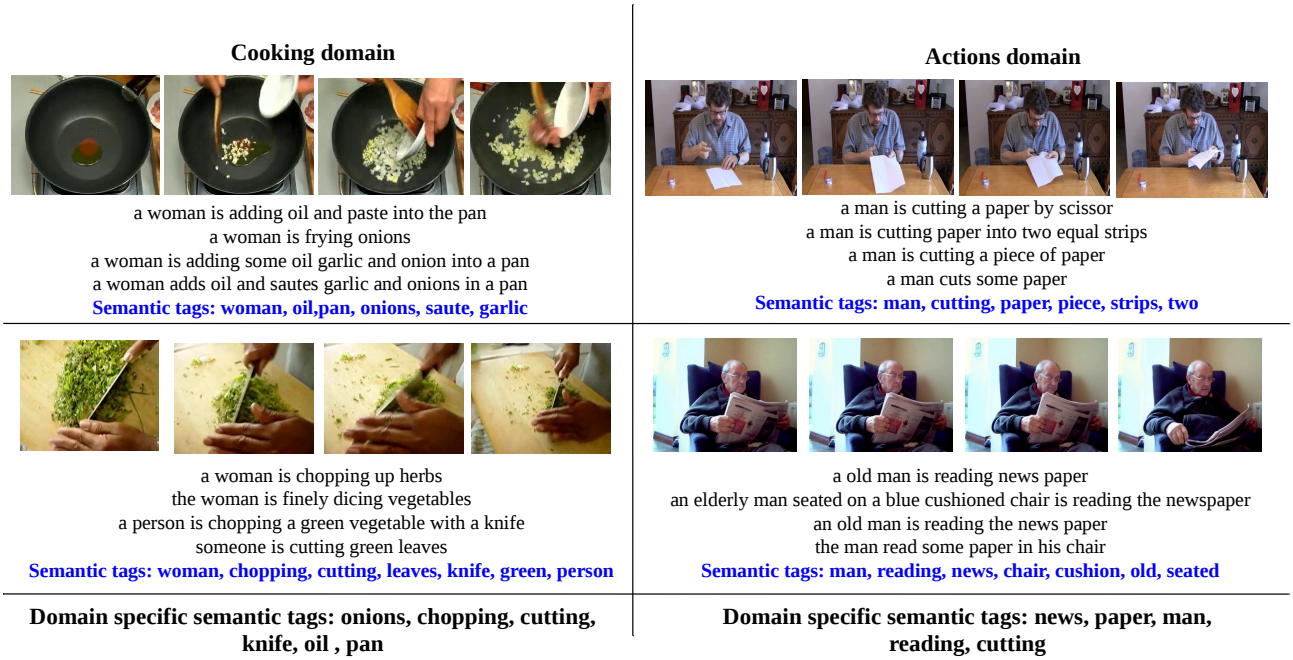


Figure 1: Sample video frames from *cooking* and *actions* domains along with the captions and semantic tags of the videos. The common semantic tags and the domain-specific semantic tags are also shown.

reading a newspaper, etc.

In Figure 1, the semantic tags *onions, chopping,* and *pan* are present only in the captions of videos belonging to the *cooking* domain. The semantic tags *reading, news paper,* and *chair* are present only in the captions of videos belonging to the *actions* domain. The semantic tags such as *man, woman, cutting* and *person* are common to both the domains. It is noted that the sentence styles for the videos of both the domains are similar. Training a language model for a domain with a small number of videos may lead to overfitting. We propose to use a video in training the language models for more than one domain to improve performance of video captioning. The main contributions of this work are as follows:

1. We propose a domain-specific approach to video captioning using the domain-specific semantic tags and language model for decoding.
2. We propose a method that uses both the visual and semantic features obtained from visual features for classifying the videos into different domains.
3. We propose a method to share the videos in building the decoders for different domains.
4. We propose to use the temporal VLAD method for aggregating the descriptors extracted frames of a video using the 2D-CNN and 3D-CNN models.

The remaining part of the paper is organized as follows. Section 2 presents a review of the approaches to video captioning. Section 3 presents the proposed approach. Section 4 presents the details of the experimental studies. Section 5 presents the results of the experimental studies.

2. Approaches to video captioning

The template-based approaches to video captioning deal with the subject, verb and object parts of the description. The visual content of the video is used to identify the *subject, verb,* and *object* parts of a caption. These *subject, verb* and *object* are then fit into the predefined sentence templates [41], [15], [28]. The template-based approaches are good at generating captions that are grammatically correct. These approaches are limited to the *subject, verb,* and *object* triplets. Many end-to-end approaches have been proposed for video captioning[36], [24], [43]. The sequence to sequence video to text (S2VT) approach [36] is the first end-to-end approach proposed for video captioning. It uses a stack consisting of two Long Short Term Memory (LSTM) models for both encoding and decoding.

Several attention-based approaches have been proposed for image captioning. The *show, attend and tell* approach [39] provides attention to different parts of an image. The approach in [16] proposes an attention based LSTM (aLSTMs) to build a decoder for video captioning. The approach in [46] proposes a Hierarchical Recurrent Neural Network (H-RNN) for video captioning. It provides spatial attention by computing features for multiple image patches located at different positions in the frames of a video. The approach in [17] uses an attention-based fusion of multimodal features extracted from a video like 2D-CNN features, 3D-CNN features, and audio features. These ap-

proaches do not use the domain-specific information that plays an important role in captioning of a video. The approach in [23] uses a multi-faceted attention on 2D-CNN, 3D-CNN and semantic features to generate captions for images.

In the Semantic Compositional Network (SCN) based approach to video captioning [13], the semantic attributes are first obtained from a video using a multi-label classification model. These semantic attributes are used to determine the weights of LSTM in the decoder. In the dual stream approach to video captioning [40], one stream is the visual features and another stream is the semantic attributes obtained from the video. A semantic guided LSTM [47] uses the semantic tags predicted from a video to guide the LSTM based decoder in the language model. In this approach, the language model uses the semantic tags obtained from the entire dataset. In our proposed approach, we use only the domain-specific semantic tags in training the domain-specific decoder.

In [9] an approach that uses a joint encoding of features in two streams is proposed. The two streams are 3D-CNN features and Motion History Images (MHI) features. This approach does not use the semantic features. Sequential Vector of Locally Assigned Descriptors(VLAD) approach [42] integrates the VLAD into the Recurrent Convolutional Neural Network for captioning. The object aggregation based approach using VLAD is proposed in [48]. It uses Bi-directional temporal graph and object aggregation for video captioning. These approaches does not use the domain-specific information of the videos. The PickNet approach [8] uses the reinforcement learning to pick the informative frames from an input video for caption generation. Here, there is a trade-off between the number of frames picked and the information gained. The approach in [32] uses a multi-modal stochastic LSTM to learn visual and textual features. Then a backward LSTM is used to perform the uncertainty propagation.

3. Domain-specific semantics guided approach to video captioning

This section presents the proposed domain-specific semantics guided approach to video captioning. Figure 2 shows the architecture of the proposed approach. We use the spatial (2D-CNN), spatio-temporal (3D-CNN) and semantic features for domain classification and captioning of a video. Given an input video, the domain classifier generates probabilities of the video belonging to different domains.The top-k probabilities from the classifier are chosen and the video is assigned to the domains of the top-k probabilities. Using this approach we identify a new set of training videos for each domain. The domain-specific semantic tags are extracted from the captions of the training videos of a particular domain. The domain-specific decoder uses the domain-specific semantic tags during caption generation. During testing, captions for a given video are generated using decoders of all the domains. Then we choose the sentence with maximum probability as the final caption for the video.

3.1. Domain classification of videos

The domain classification of videos plays a significant role in the proposed domain-specific approach to video captioning. As shown 'in figure 2, the 2D-CNN and 3D-CNN features are first extracted from a video. The 2D-CNN and 3D-CNN features are pre-processed using the temporal VLAD method to obtain a representation of the video.

The semantic features extracted using the approach in [13] are also given as input to the domain classifier. Let N be the number of videos available. Let M be the number of semantic tags identified from the captions. The target semantic tag label vector for n^{th} video is $\mathbf{x}_n = [x_{n1}, x_{n2}, \cdots x_{nM}]^T$, where x_{ni} is 1 if the label i is present in the descriptions of the n^{th} video. A Multi-layer Perceptron (MLP) is trained as a multi-label classifier with $\mathbf{x}_n$ as the target vector for n^{th} video in the training dataset. The output of the MLP $\mathbf{s}_n = [s_{n1}, s_{n2}, \cdots s_{nM}]^T$ is the semantic feature vector of the n^{th} video. Here s_{nm} is the probability of the semantic tag m being associated with video n. The semantic features play a vital role in captioning as they explicitly indicate the presence of specific objects or actions in the video.

A set of VLAD processed 2D-CNN and 3D-CNN features and semantic features is given as input to the domain classifier. As there are no ground truth labels available for domain classification in MSVD, we use K-means clustering to give pseudo domain labels to the training videos. The K-means clustering is used to cluster the videos using the representation as obtained above. The clustering method clusters the videos with similar visual features, semantic tags and descriptions. The cluster index is assigned as the pseudo domain label of all the videos in a cluster. In MSR-VTT dataset, each video is assigned a category label like music, people, gaming, etc. Many of the videos are labeled erroneously. The tagging do not precisely determine the categories of the video. It is also found that the number of videos in some of the categories is large, whereas in few categories the number of videos is very small. Hence we use the same approach to obtain the domain labels for MSVD and MSR-VTT dataset.

The most frequent words in the captions of videos of a domain constitute the domain-specific semantic tags of that domain. Figure 3 shows the word clouds of four domains in the MSVD dataset. The font size of a word in a word cloud indicates the frequency of occurrence of the word in that particular domain. Words with large font size represent more frequent words, whereas the words with small font

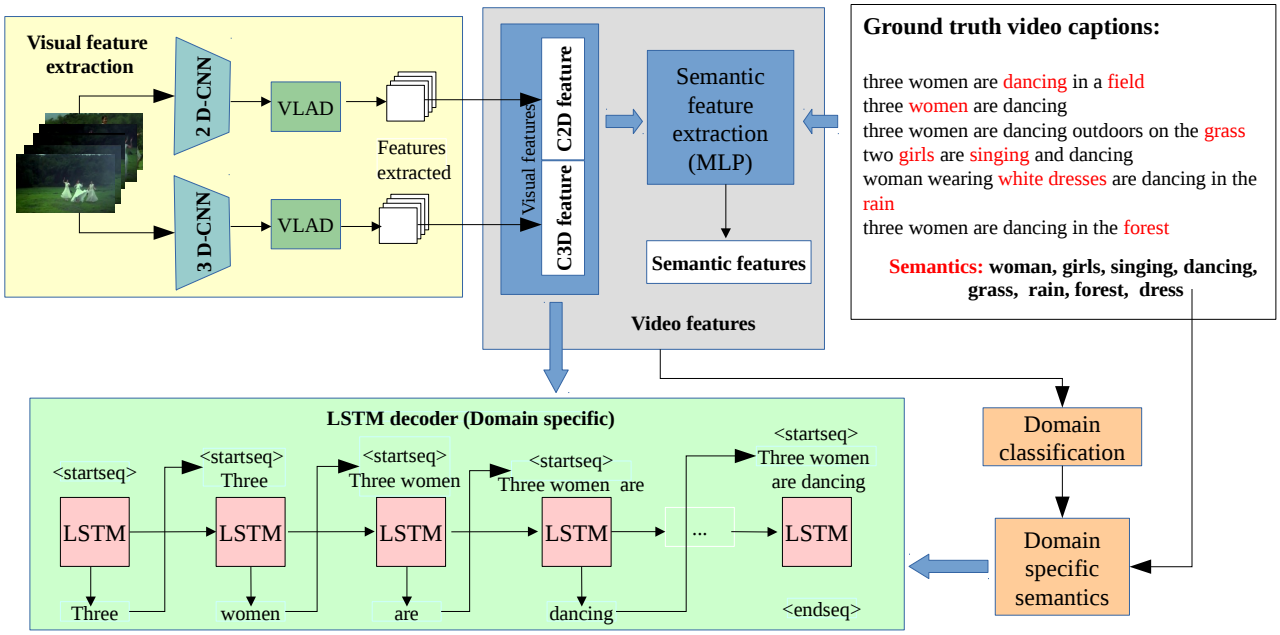


Figure 2: Framework of the proposed approach video captioning. The visual features and semantic features are used jointly to identify the video domains. Then the domain-specific decoder is used for generating captions.

size represent less frequent words. From the word cloud, it can be seen that the categories of domains are *cooking, action, music, and riding*. After clustering we inspected the videos and the descriptions of videos in different clusters and are identified with the domain titles like *cooking, music, sports, animals, actions, etc.*

A video can be associated with more than one domain using the visual and semantic information available in the video. These videos are called as multi-domain videos. An example of such an association is shown in Figure 4. Here, a video in the *animals* domain has semantic tags *cat, playing, trampoline, man and boy*, and a video in *actions* domain has semantic tags *man, playing, ball, kicking, around, boy, and soccer*. The semantic tags *man, playing, and boy* are present in both the videos. The cross-domain semantic tags are common to both *animals* and *sports* domain, so both the videos are associated with *animals* and *actions* domains.

The pseudo domain labels obtained using the clustering method are considered as desired output for the domain classification model. Hard assignment of only one domain label for a video confines the video to one domain only. In this work, a video is associated with more than one domain by soft assignment. Here, the domain classification is considered as a multi-class classification task. Let $\mathbf{y}_n = [y_{n1}, y_{n2}, \cdots y_{nL}]^T$ be the domain label vector, where L is the number of domains. The probability $p(d_{nl}|\mathbf{r}_n)$ of the video n^{th} to be assigned the domain l is calculated by the multi-class classifier. Here $\mathbf{r}_n$ is the feature representation of n^{th} video given by $\mathbf{r}_n = [\mathbf{v}_n, \mathbf{t}_n, \mathbf{s}_n]$, where $\mathbf{v}_n$ are VLAD aggregated 2D-CNN features, $\mathbf{t}_n$ are the VLAD aggregated 3D-CNN features, and the $\mathbf{s}_n$ are the semantic features of the n^{th} video.

The final output of the classifier is represented as $\mathbf{d}_n = [d_{n1}, d_{n2}, \cdots d_{nL}]$, which is the vector of probabilities of assigning the n^{th} video to the l^{th} domain by the multi-class classifier. From the final output of the classifier, we obtain the top-k labels and assign the n^{th} video to the k domains. In our work, the semantic features obtained from the entire dataset are used for domain classification. So, the semantic tags available in the captions indirectly contribute to domain classification through the semantic features.

3.2. Domain-specific decoder

The encoder of the video captioning model is common for all the domains. A separate domain-specific decoder is trained for each domain. In this work, a novel method is presented to train a domain-specific language model. We identify a new set of training videos for each domain using top-k probabilities obtained using the domain classifier. These videos are used in training the domain-specific decoder. The domain-specific decoder uses domain-specific vocabulary derived from the captions of training videos as the possible words to be predicted by the language model.

A video is assigned to the training set of a domain only if the probability for that domain is greater than a threshold. If the probability is very small, then the contribution of

(a) Cooking domain (b) Riding vehicles domain (c) Music domain (d) Actions domain

Figure 3: Word clouds for different domains

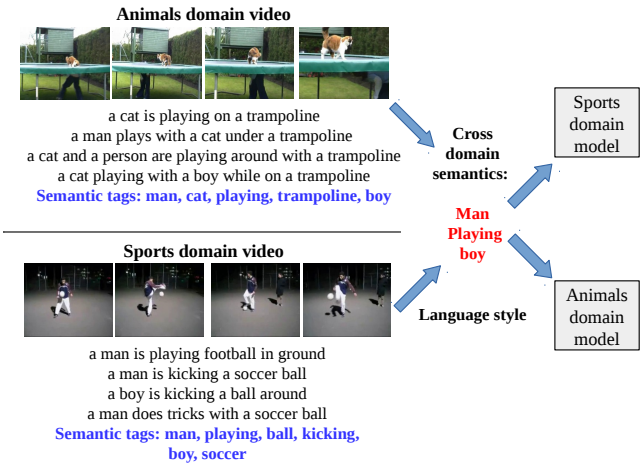


Figure 4: An illustration of videos to be shared among domains for the semantic tags information

the video to that particular domain is not significant. In this work, the threshold is empirically chosen as 0.3. The decoder is a caption generation model using LSTM as shown in Figure 2. The decoder predicts one word at a time. For the n^{th} video, the m^{th} word of the description is given by

$$c_{nm} = \underset{\mathbf{w}_t}{\operatorname{argmax}}\, p(\mathbf{w}_t|\mathbf{h}_t, \mathbf{c}_{<m}\mathbf{v}_n, \mathbf{t}_n, \mathbf{s}_n)) \tag{1}$$

where V is the vocabulary, $\mathbf{c}_{<m}$ is set of words generated upto m^{th} time step, $\mathbf{h}_t$ is the internal state of LSTM at time t. $\mathbf{w}_t$ is the vector of final scores obtained from the domain-specific decoder at t^{th} time step. The descriptions of the training videos of a particular domain determine domain-specific semantic features $\mathbf{s}_n$ used in the decoder.

While testing, all the domain-specific decoders are used for generating the captions of a single video. The caption for a testing video is obtained as

$$\begin{aligned} l^* &= \underset{l}{\operatorname{argmax}}\, p(sent_{nl}) * d_{nl}, \qquad l = 1, 2, \cdots L \\ sent_n^* &= sent_{nl^*} \end{aligned} \tag{2}$$

where $sent_n^*$ is the predicted sentence, and $p(sent_{nl})$ is the probability of the sentence generated by the domain l for the n^{th} video.

3.3. Temporal VLAD preprocessing of features

The vector of locally aggregated descriptors (VLAD) approach creates a vector representation that aggregates descriptors based on a locality criterion in a feature space. In [19], VLAD was proposed to create a compact representation of the features obtained from an image. Action-VLAD in [14] is proposed to create VLAD representation of a video. In this work, VLAD is used in temporal aggregation of the feature vectors extracted from 2D-CNN or 3D-CNN. The feature vectors extracted from the frames of a video are considered as separate descriptors. Different videos have different number of frames. So the number of descriptors for different videos are different.

Let $\mathbf{x}_{ni}$ be the feature vector extracted from i^{th} frame of the n^{th} video, where $i = 1, 2, \cdots T_n$. Here T_n is the number of frames in the n^{th} video. The feature vectors extracted from the frames of all the videos are clustered using K-means clustering. The feature vectors of the frames from a single video may be assigned to more than one cluster. The centers of the clusters $C = \{\mathbf{c}_1, \mathbf{c}_2 \cdots \mathbf{c}_K\}$ are used for VLAD aggregation, where $\mathbf{c}_k$ is the k^{th} cluster center and K is the number of clusters. The VLAD descriptor accumulates the differences $\mathbf{x}_{ni} - \mathbf{c}_k$. The descriptor $\mathbf{a}_{nk}$ is obtained as

$$\mathbf{a}_{nk} = \sum_{i=1}^{T_n} d_k(\mathbf{x}_{ni})(\mathbf{x}_{ni} - \mathbf{c}_k) \tag{3}$$

where $d_k(\mathbf{x}_{ni}) = 1$ if $\mathbf{c}_k$ is the nearest cluster center for $\mathbf{x}_{ni}$, otherwise $d_k(\mathbf{x}_{ni}) = 0$. The VLAD descriptor $\mathbf{b}_n$ for the n^{th} video is obtained by summing up the descriptors for all K clusters as follows

$$\mathbf{b}_n = \sum_{k=1}^{K} \mathbf{a}_{nk} \tag{4}$$

In this work, the features extracted from 2D-CNN and 3D-CNN are processed using this method of VLAD aggregation.

4. Experiments

4.1. Dataset

The Microsoft Research Video Description Corpus (MSVD) [4] is a collection of about 1970 You-tube video clips. The duration of each clip is about 10 to 25 seconds. These clips include various activities such as "cooking", "people playing instruments" and "activities by the animals". Each clip has about 40 annotations. The total number of annotations is 80,839. The data is split as follows: 1200 clips for training, 100 clips for validation and 670 clips for testing.

The MSR-VTT (MSR Video to Text) is a large scale video captioning dataset. The MSR-VTT dataset provides 10,000 videos, with 20 human annotated descriptions for each video. Each video is labeled into one of the 20 category like music, people, gaming, actions, etc. In this work we use the standard split of MM2016 challenge[1], i.e., 6513 videos for training, 2990 videos for testing and 497 videos for validation.

4.2. Evaluation metrics

Several standard metrics such as BLEU (precision-based) (BL-4) [26], METEOR (harmonic mean of precision and recall) (MET) [10], CIDEr (consensus-based) (CIDr) and ROUGE-L (recall-based) (RG-L) are used for evaluating the video captioning systems. The performance of the proposed approach is evaluated using all the four metrics and the results are given as percentage(%) scores. All the four measures are evaluated in [34] for the image captioning problem and it was shown that the METEOR score is better than BLEU and ROUGE-L. Hence we use METEOR and CIDEr as the comparison metrics. We have used the Microsoft COCO evaluation server [7] implementation of the metrics to evaluated the performance video captioning.

4.3. Training procedure

In this work, 2D-CNN features are extracted using ResNet152. We remove the final dense layer from the pre-trained model and use it as the feature extractor. From an input video, every 10^{th} frame is retrieved and processed by the feature extraction model. The temporal features are extracted using a 3D-CNN of [33]. As the 3D-CNN processes 16 frames at a time step, the input video is split into 16 frame clips. Both the 2D-CNN and 3D-CNN features are processed by VLAD to obtain a representation of the video. A multi-layer perceptron is used to learn the semantic features from the training videos. We identify the top 300 most frequently used words from all the training captions as semantic tags of the videos. The final layer of the MLP uses the sigmoid activation function, and the model is trained using binary cross-entropy loss. After domain clustering we obtain 8 domains with around 150 training samples each for MSVD dataset. For MSR-VTT dataset we obtained 10 domains with around 650 training videos for each domain.

After pre-processing the descriptions, the number of words in the vocabulary is 9842, and the maximum length of description is 49. GloVe embedding [27] is used to embed the words of the vocabulary. The dimension of GloVe embedding is 100 for every word. The size of the final dense layer in the decoder is equivalent to the size of the vocabulary of the training captions available in the domain. The final dense layer uses the soft-max activation function. The decoder predicts the probabilities of all the words in the domain-specific vocabulary. At a particular time step, the word with the maximum probability is identified as the predicted word. The parameters for the domain-specific decoder model are initialized using the parameters of a baseline model. The optimizer for training the decoder model is Adam optimizer, and the loss function is the categorical cross-entropy loss function.

5. Results

5.1. Ablation Study

Table 1 presents the performance of video captioning models using different features. It also presents the results of proposed approach based on the domain-specific decoders (DSD) and the domain-specific semantics(DS-SEM) on MSVD and MSR-VTT dataset. The different methods listed in the table are as follows:

- **2D+3D+SEM:** The video captioning model trained using 2D-CNN, 3D-CNN and semantic features without VLAD aggregation . The model has a common decoder for all the domains.
- **Temporal VLAD on 2D+3D+SEM:** The video captioning model trained using VLAD aggregated 2D-CNN, 3D-CNN features and also the semantic features. It does not use domain-specific decoder for caption generation.
- **DSD Top-1:** The DSD model uses a separate domain-specific decoder for every domain. All the domain-specific decoder models use the VLAD aggregated 2D-CNN and 3D-CNN features. It chooses the domain label with maximum probability in the domain classifier. Hence, it does not allow sharing a video with more than one domain for training the domain-specific decoders.
- **DSD Top-3 and DSD Top-5:** The DSD Top-3 chooses the domain labels with top 3 probabilities and the DSD Top-5 chooses the domain labels with top 5 probabilities in the multi-class domain classification. A single video is shared for training among domains associated with the Top-3 and Top-5 domain probabilities.

Table 1: Performance of captioning using different features and the proposed domain-specific decoder based video captioning systems on MSVD and MSRVTT dataset

	MSVD				MSRVTT			
Method	BL-4	MET	CIDr	RG-L	BL-4	MET	CIDr	RG-L
2D+3D+SEM	41.5	31.5	56.2	69.1	33.4	24.5	28.7	56.8
Temporal VLAD on 2D+3D+SEM	50.0	33.5	70.8	71.4	38.7	27.5	41.6	60.7
DSD Top-1	46.7	32.6	67	70.3	39.6	28.0	44.3	61.0
DSD Top-3	49.5	34.1	74.8	71.6	43.6	29.1	47.2	61.7
DSD Top-5	48.2	33.8	68.8	71.7	40.6	28.3	44.8	61.2
DSD Top-5 DS-SEM	49.6	34.0	70.4	72.0	42.6	28.6	46.7	60.5
DSD Top-3 DS-SEM	**50.1**	**34.7**	**76.0**	**73.1**	**45.2**	**29.9**	**51.1**	**64.2**

- **DSD Top-3 DS-SEM and DSD Top-5 DS-SEM:** The models are similar to DSD models in choosing the domain-specific training videos. The domain-specific semantic features are obtained from domain-specific semantics, to guide the decoder in predicting the words of a description.

From Table 1 it is seen that the VLAD aggregation on 2D-CNN and 3D-CNN features increases the BLEU and METEOR scores compared to the model using features without VLAD aggregation. Hence, the VLAD aggregation provides a better representation for the 2D-CNN and 3D-CNN features. The domain-specific decoder improves the captioning ability of the model significantly compared to the models without the domain-specific decoder. DSD Top-1 has lower performance compared to DSD Top-3 and DSD Top-5. Since each domain has limited number of training videos, if we set k as 1 the language model is trained using a limited number of videos which leads to overfitting.

Increasing the number of domains to which a video is assigned improves the performance but, if the video is assigned to a large number of domains (here 5), it reduces the performance. It is inferred that if the video is shared by a large number of domains, it results in poor performance as the decoder loses its specificity. Table 1 also shows that the use of domain-specific semantics provides a better performance. Use of domain-specific semantics in the language model reduces the impact of the semantics in other domains, thus improving the language model's performance.

5.2. Comparison with state-of-the-art methods

Table 2 and Table 3 show the performance of the proposed approach and the other state-of-the-art methods using BLEU-4, METEOR, CIDEr and ROUGE-L metrics on MSVD dataset and MSR-VTT dataset respectively.

The following inferences are made by analyzing the results in Table 2 and Table 3.

1. The FGM and S2VT approaches use only the spatial features. The LSTM-E and joint stream models use the spatial and temporal features. Models that use the spatial, temporal and semantic features like SCN provide a better performance compared to the models that use only spatial and temporal features. Thus the use of

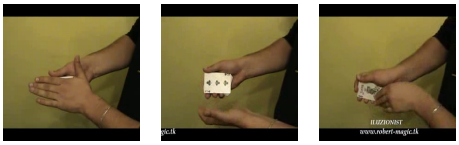
Table 2: Comparison of the performance for different approaches to video captioning on MSVD dataset

Approach	BL-4	MET	CIDr	RG-L
FGM [41]	13.7	23.9	-	-
LSTM-YT [37]	33.3	29.1	-	-
S2VT [36]	-	29.8	-	-
LSTM-E [24]	45.3	31.0	-	-
Joint stream [9]	-	31.1	-	-
MM-Att [17]	**53.9**	32.2	67.4	-
h-RNN [46]]	49.9	32.6	65.8	-
aLSTMs [16]	50.8	33.3	74.8	-
SCN-LSTM [13]]	51.1	33.5	**77.7**	-
Topic-Guid [6]	49.2	34.2	77.6	71.0
Less-is-more [8]]	52.3	33.3	76.5	69.6
DSD-3 DS-SEM	50.1	**34.7**	76.0	**73.1**

Table 3: Comparison of the performance for different approaches to video captioning on MSR-VTT dataset

Approach	BL-4	MET	CIDr	RG-L
MM-Att [17]	39.7	25.5	40.0	-
aLSTMs [16]	38.0	26.1	43.2	-
Stochastic -RNN [32]	39.8	26.1	40.9	59.3
Less-is-more [8]	41.3	27.7	44.1	59.8
DS-RNN [40]	42.3	29.4	46.1	62.3
Topic-Guid[6]	44.9	29.6	51.8	62.8
DSD-3 DS-SEM	**45.2**	**29.9**	51.1	**64.2**

MSVD dataset



Ground truth: a person is playing with cards
Domain: Actions
DSD Top-1 man playing a picture
DSD Top-3 DS-SEM a man is playing with cards
DSD Top-5 DS-SEM a man is rotating hands

Ground truth: the space shuttle is flying
Domain: Technology
DSD Top-1 a bird is flying a runway
DSD Top-3 DS-SEM a shuttle is flying
DSD Top-5 DS-SEM a space is flies

MSRVTT dataset

Ground truth: a person is showing a stroller
Domain: Actions
DSD Top-1 a woman is talking a a car
DSD Top-3 DS-SEM a woman is showing how to use a stroller
DSD Top-5 DS-SEM a man is demonstrating a stroller

Ground truth: a man is lecturing about underprivileged kids
Domain: Technology
DSD Top-1 a man is talking about a car
DSD Top-3 DS-SEM a man is giving a speech
DSD Top-5 DS-SEM a woman is talking a speech

Figure 5: Captions generated for videos by assigning videos to multiple domains in MSVD and MSR-VTT dataset

spatial, temporal and semantic features in the proposed approach improves the performance.

2. The SCN [13] uses semantic features in the LSTM based decoder. It assumes that all the videos belong to a single domain. The use of domain-specific semantic features improves the performance significantly in comparison to the SCN.

3. A previous approach to domain-specific captioning [30] uses only the visual features for domain classification. The topic based approach in [5], [6] uses captions and visual features for topic prediction. However, the captions will not be available during testing. In the proposed approach, the semantic features are used for domain classification. Hence, it is seen that the proposed approach provide better scores than the previous topic based approaches to video captioning.

4. Using only the domain-specific semantic features in guiding the LSTM of decoder increases the METEOR score. The decrease in CIDEr score may be because of using only the domain-specific semantic features. The ability to predict cross-domain semantic words is reduced.

5.3. Sample results for the videos

Figure 5 shows frames of a few videos and their captions generated by different models. It is seen that the captions generated by the model trained using DSD-Top-3 DS-SEM are better compared to other models. For videos from *technology* domain the descriptions are incorrect, as it has very few training videos in both the datasets. Though the domain-specific decoder allows overlap, it falls short in performance for domains which have very few videos.

6. Conclusion

In this paper a domain-specific semantics guided approach is proposed for video captioning. Here videos are classified into domains using a multi-class classifier. Each video is included in the training set of top-k domains predicted by domain classifier. The domain-specific features like vocabulary and semantic tags are obtained from the new training set of videos and their captions. The inter-domain videos are shared among more than one domain for training the language models of the domain-specific decoder. This allows the models to learn domain-specific features from the inter-domain videos, which improves the performance of the domain-specific decoders. The results of studies on the MSVD and MSRVTT datasets shows that the proposed approach generates better captions compared to the state-of-the-art approaches.

References

[1] Microsoft research, ACM multimedia MSR video to language challenge, 2016.

[2] S. Aakur, F. D. M. d. Souza, and S. Sarkar. Towards a knowledge-based approach for generating video descriptions. In *Proceedings of the 2017 14th Conference on Computer and Robot Vision (CRV)*, pages 24–31, May 2017.

[3] L. Baraldi, C. Grana, and R. Cucchiara. Hierarchical boundary-aware neural encoder for video captioning. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3185–3194, July 2017.

[4] 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*, HLT '11, pages 190–200. Association for Computational Linguistics, 2011.

[5] S. Chen, J. Chen, and Q. Jin. Generating video descriptions with topic guidance. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, ICMR 17, pages 5–13. ACM, 2017.

[6] S. Chen, Q. Jin, J. Chen, and A. G. Hauptmann. Generating video descriptions with latent topic guidance. *IEEE Transactions on Multimedia*, 21(9):2407–2418, Sep. 2019.

[7] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015.

[8] Y. Chen, S. Wang, W. Zhang, and Q. Huang. Less is more: Picking informative frames for video captioning. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Proceedings of the 2018 European Conference on Computer Vision(ECCV)*, pages 367–384, Cham, 2018. Springer International Publishing.

[9] Chenyang Zhang and Yingli Tian. Automatic video description generation via lstm with joint two-stream encoding. In *Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2924–2929, Dec 2016.

[10] M. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

[11] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):677–691, April 2017.

[12] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollr, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. From captions to visual concepts and back. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1473–1482, June 2015.

[13] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng. Semantic compositional networks for visual captioning. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1141–1150, July 2017.

[14] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3165–3174, July 2017.

[15] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV)*, pages 2712–2719, Dec 2013.

[16] Z. Guo, L. Gao, J. Song, X. Xu, J. Shao, and H. T. Shen. Attention-based lstm with semantic consistency for videos captioning. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, pages 357–361. ACM, 2016.

[17] C. Hori, T. Hori, T. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi. Attention-based multimodal fusion for video description. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4203–4212, Oct 2017.

[18] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4565–4574, June 2016.

[19] H. Jgou, M. Douze, C. Schmid, and P. Prez. Aggregating local descriptors into a compact image representation. In *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3304–3311, June 2010.

[20] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676, April 2017.

[21] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. Dense-captioning events in videos. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 706–715, Oct 2017.

[22] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1601–1608, June 2011.

[23] X. Long, C. Gan, and G. de Melo. Video captioning with multi-faceted attention. *Transactions of the Association for Computational Linguistics*, 6:173–184, 2018.

[24] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui. Jointly modeling embedding and translation to bridge video and language. *CoRR*, abs/1505.01861, 2015.

[25] Y. Pan, T. Yao, H. Li, and T. Mei. Video captioning with transferred semantic attributes. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 984–992, July 2017.

[26] K. Papineni, S. Roukos, T. Ward, and W. jing Zhu. Bleu: a method for automatic evaluation of machine translation. In

Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pages 311–318, 2002.

[27] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, Oct. 2014. Association for Computational Linguistics.

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

[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal for Computer Vision*, 115(3):211–252, Dec. 2015.

[30] Z. Shen, J. Li, Z. Su, M. Li, Y. Chen, Y. Jiang, and X. Xue. Weakly supervised dense video captioning. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5159–5167, July 2017.

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[32] J. Song, Y. Guo, L. Gao, X. Li, A. Hanjalic, and H. T. Shen. From deterministic to generative: Multimodal stochastic rnns for video captioning. *IEEE Transactions on Neural Networks and Learning Systems*, 30(10):3047–3058, Oct 2019.

[33] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, Dec 2015.

[34] R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, June 2015.

[35] S. Venugopalan, L. A. Hendricks, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko. Captioning images with diverse objects. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1170–1178, July 2017.

[36] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence – video to text. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 4534–4542, Washington, DC, USA, 2015. IEEE Computer Society.

[37] 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 the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1494–1504. Association for Computational Linguistics, 2015.

[38] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, June 2015.

[39] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 2048–2057, 2015.

[40] N. Xu, A. Liu, Y. Wong, Y. Zhang, W. Nie, Y. Su, and M. Kankanhalli. Dual-stream recurrent neural network for video captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(8):2482–2493, Aug 2019.

[41] R. Xu, C. Xiong, W. Chen, and J. J. Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2346–2352. AAAI Press, 2015.

[42] Y. Xu, Y. Han, R. Hong, and Q. Tian. Sequential video vlad: Training the aggregation locally and temporally. *IEEE Transactions on Image Processing*, 27(10):4933–4944, Oct 2018.

[43] Y. Yang, J. Zhou, J. Ai, Y. Bin, A. Hanjalic, H. T. Shen, and Y. Ji. Video captioning by adversarial lstm. *IEEE Transactions on Image Processing*, 27(11):5600–5611, Nov 2018.

[44] T. Yao, Y. Pan, Y. Li, and T. Mei. Incorporating copying mechanism in image captioning for learning novel objects. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5263–5271, July 2017.

[45] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4651–4659, June 2016.

[46] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4584–4593, June 2016.

[47] J. Yuan, C. Tian, X. Zhang, Y. Ding, and W. Wei. Video captioning with semantic guiding. In *Proceedings of the IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, pages 1–5, 09 2018.

[48] J. Zhang and Y. Peng. Object-aware aggregation with bidirectional temporal graph for video captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.