IJMIR manuscript No.
(will be inserted by the editor)

Joint Embeddings with Multimodal Cues for Video-Text Retrieval

Niluthpol C. Mithun · Juncheng Li · Florian Metze · Amit K. Roy-Chowdhury

Abstract For multimedia applications, constructing a joint representation that could carry information for multiple modalities could be very conducive for downstream use cases. In this paper, we study how to effectively utilize available multimodal cues from videos in learning joint representations for the cross-modal video-text retrieval task. Existing hand labeled video-text datasets are often very limited by their size considering the enormous amount of diversity the visual world contains. This makes it extremely difficult to develop a robust video-text retrieval system based on deep neural network models. In this regard, we propose a framework that simultaneously utilizes multi-modal visual cues by a "mixture of experts" approach for retrieval. We conduct extensive experiments to verify that our system is able to boost the performance of the retrieval task compared to the state-of-the-art. In addition, we propose a modified pairwise ranking loss function in training the embedding and study the effect of various loss functions. Experiments on two benchmark datasets show that our approach yields significant gain compared to the state-of-the-art.

Keywords Video-Text Retrieval · Joint Embedding · Multimodal Cues

1 Introduction

The goal of this work is to retrieve the correlated text description given a random video, and vice versa, to retrieve the matching videos provided with text descriptions (See Fig. 1). While several computer vision tasks (e.g., image classification [20,37,23], object detection [47,46,36]) are now reaching maturity, cross-modal retrieval between visual data and natural language description remains a very challenging problem [64,35] due to the gap and ambiguity between different modalities and availability of limited training data. Some recent works [38,30,59,25,17] attempt to utilize cross-modal joint embeddings to address the gap. By projecting data from multiple modalities into the same joint space, the similarity of the resulting points would reflect the semantic closeness between their corresponding original inputs. In this work, we focus on learning joint video-text embedding models and combining video cues for different purposes effectively for developing robust video-text retrieval system.

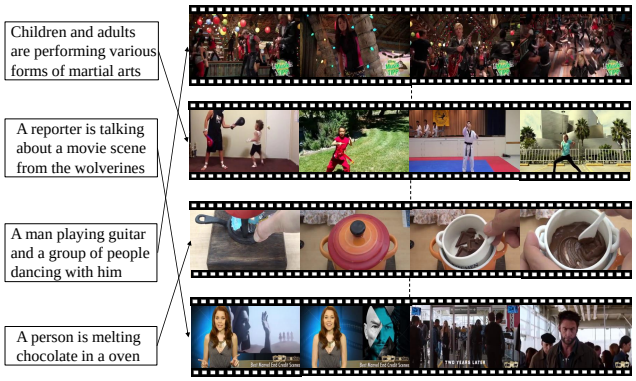

Fig. 1: Illustration of Video-Text retrieval task: given a text query, retrieve and rank videos from the database based on how well they depict the text, and vice versa.

The video-text retrieval task is one step further than the image-text retrieval task, which is a comparatively well-studied field. Most existing approaches for video-text retrieval are very similar to the image-text retrieval methods by design and focus mainly on the modification of loss functions [12,61,50,40,41]. We observe that simple adaptation of a state-of-the-art image-text embedding method [13] by mean-pooling features from video frames generates a better result than existing video-text retrieval approaches [12,40]. However, such methods ignore lots of contextual information in video sequences such as temporal activities or specific

Niluthpol C. Mithun
University of California, Riverside, CA
Email: nmith001@ucr.edu

Juncheng Li
Carnegie Mellon University, PA
Email: junchenl@cs.cmu.edu

Florian Metze
Carnegie Mellon University, PA
Email: fmetze@cs.cmu.edu

Amit K. Roy-Chowdhury
University of California, Riverside, CA
Email: amitrc@ece.ucr.edu

“A dog is barking in field”

“Gunshot broke out at the concert”

Fig. 2: Example frame from two videos and associated caption to illustrate the significance of utilizing supplementary cues from videos to improve the chance of correct retrieval.

scene entities, and thus they often can only retrieve some generic responses related to the appearance of static frame. They may fail to retrieve the most relevant information in many cases to understand important questions for efficient retrieval such as ‘What happened in the video’, or ‘Where did the video take place’. This greatly undermines the robustness of the systems; for instance, it is very difficult to distinguish a video with the caption “a dog is barking" apart from another “a dog is playing" based only on visual appearance (See Fig. 2). Associating video motion content and the environmental scene can give supplementary cues in this scenario and improve the chance of correct prediction. Similarly, to understand a video described by “gunshot broke out at the concert" may require analysis of different visual (e.g., appearance, motion, environment) and audio cues simultaneously. On the other hand, a lot of videos may contain redundant or identical contents, and hence, an efficient video-text retrieval should utilize the most distinct cues in the content to resolve ambiguities in retrieval.

While developing a system without considering most available cues in the video content is unlikely to be comprehensive, an inappropriate fusion of complementary features could adversely increase ambiguity and degrade performance. Additionally, existing hand labeled video-text datasets are very small and very restrictive considering the amount of rich descriptions that a human can compose and the enormous amount of diversity in the visual world. This makes it extremely difficult to train deep models to understand videos in general to develop a successful video-text retrieval system. To ameliorate such cases, we analyze how to judiciously utilize different cues from videos. We propose a mixture of experts system, which is tailored towards achieving high performance in the task of cross-modal video-text retrieval. We believe focusing on three major facets (i.e., concepts for Who, What, and Where) from videos is crucial for efficient retrieval performance. In this regard, our framework utilizes three salient features (i.e., object, action, place) from videos (extracted using pre-trained deep neural networks) for learning joint video-text embeddings and uses an ensemble approach to fuse them. Furthermore, we propose a modified pairwise ranking loss for the task that emphasizes on hard negatives and relative ranking of positive labels. Our approach shows significant performance improvement compared to previous approaches and baselines.

Contributions: The main contributions of this work can be summarized as follows.

• The success of video-text retrieval depends on more robust video understanding. This paper studies how to achieve the goal by utilizing multimodal features from a video (different visual features and audio inputs). Our proposed framework uses action, object, place, text and audio features by a fusion strategy for efficient retrieval.

• We present a modified pairwise loss function to better learn the joint embedding which emphasizes on hard negatives and applies a weight-based penalty on the loss based on the relative ranking of the correct match in the retrieval.

• We conduct extensive experiments and demonstrate a clear improvement over the state-of-the-art methods in the video to text retrieval tasks on the MSR-VTT dataset [60] and MSVD dataset [9].

This paper is an extended version of our work [35] with significantly more insights and detailed discussions about the proposed framework. The main extension in our pipeline is adding scene cues from videos, along with object and activity cues for learning joint embeddings to develop a more comprehensive video-text retrieval system. The previous version utilized object-text and activity-text embeddings which focused mainly on resolving ambiguities arising related to concepts for Who and What. We add a place-text embedding network in our framework to make it more robust which will help us resolve ambiguities arising from concepts for Where. Experiments show that this change results in a significant improvement over the previous works in two benchmark datasets.

2 Related Work

Image-Text Retrieval. Recently, there has been significant interest in learning robust visual-semantic embeddings for image-text retrieval [38,26,21,57]. Based on a triplet of object, action and, scene, a method for projecting text and image to a joint space was proposed in early work [14]. Canonical Correlation Analysis (CCA) and several extensions of it have been used in many previous works for learning joint embeddings for the cross-modal retrieval task [49,22,18,62, 44,19] which focuses on maximizing the correlation between the projections of the modalities. In [18], authors extended classic two-view CCA approach with a third view coming from high-level semantics and proposed an unsupervised way to derive the third view from clustering the tags. In [44], authors proposed a method named MACC (Multimedia Aggregated Correlated Components) aiming to reduce the gap between cross-modal data in the joint space by embedding

visual and textual features into a local context that reflects the data distribution in the joint space. Extension of CCA with deep neural networks named deep CCA (DCCA) has also been utilized to learn joint embeddings [62,1], which focus on learning two deep neural networks simultaneously to project two views that are maximally correlated. While CCA-based methods are popular, these methods have been reported to be unstable and incur a high memory cost due to the covariance matrix calculation with large-amount of data [58,32]. Recently, there are also several works leveraging adversarial learning to train joint image-text embeddings for cross-modal retrieval [57,10].

Most recent works relating to text and image modality are trained with ranking loss [28,17,58,13,39,52]. In [17], authors proposed a method for projecting words and visual content to a joint space utilizing ranking loss that applies a penalty when a non-matching word is ranked higher than the matching one. A cross-modal image-text retrieval method has been presented in [28] that utilizes triplet ranking loss to project image feature and RNN based sentence description to a common latent space. Several image-text retrieval methods have adopted a similar approach with slight modifications in input feature representations [39], similarity score calculation [58], or loss function [13]. VSEPP model [13] modified the pair-wise ranking loss based on violations caused by the hard-negatives (i.e., non-matching query closest to each training query) and has been shown to be effective in the retrieval task. For image-sentence matching, a LSTM based network is presented in [24] that recurrently selects pairwise instances from image and sentence descriptions, and aggregates local similarity. In [39], authors proposed a multimodal attention mechanism to attend to sentence fragments and image regions selectively for similarity calculation. Our method complements these works that learn joint image-text embedding using a ranking loss (e.g., [28,52,13]). The proposed retrieval framework can be applied to most of these approaches for improved video-text retrieval performance.

Video Hyperlinking. Video hyperlinking is also closely relevant to our work. Given an anchor video segment, the task is to focus on retrieving and ranking a list of target videos based on the likelihood of being relevant to the content of the anchor [2,5]. Multimodal representations have been utilized widely in video hyperlinking approaches in recent years [6, 56,2]. Most of these approaches rely heavily on multimodal autoencoders for jointly embedding multimodal data [55,15, 8]. Bidirectional deep neural network (BiDNN) based representations have also been shown to be very effective in video hyperlinking benchmarks [56,54]. BiDNN is also a variation of multimodal autoencoder, which performs multimodal fusion using a cross-modal translation with two interlocked deep neural networks [55,54]. Considering the input data, video-text retrieval is dealing with the same multimodal input as video hyperlinking in many cases. However, video-text retrieval task is more challenging than hyperlinking since it requires to distinctively retrieve matching data from a different modality, which requires effective utilization of the correlations in between cross-modal cues.

Video-Text Retrieval. Most relevant to our work are the methods that relate video and language modalities. Two major tasks in computer vision related to connecting these two modalities are video-text retrieval and video captioning. In this work, we only focus on the retrieval task. Similar to image-text retrieval approaches, most video-text retrieval methods employ a shared subspace. In [61], authors vectorize each subject-verb-object triplet extracted from a given sentence by word2vec model [34] and then aggregate the Subject, Verb, Object (SVO) vector into a sentence level vector using RNN. The video feature vector is obtained by mean pooling over frame-level features. Then a joint embedding is trained using a least squares loss to project the sentence representation and the video representation into a joint space. Web image search results of input text have been exploited by [40], which focused on word disambiguation. In [53], a stacked GRU is utilized to associate sequence of video frames to a sequence of words. In [41], authors propose an LSTM with visual-semantic embedding method that jointly minimizes a contextual loss to estimate relationships among the words in the sentence and a relevance loss to reflect the distance between video and sentence vectors in the shared space. A method named Word2VisualVec is proposed in [12] for the video to sentence matching task that projects vectorized sentence into visual feature space using mean squared loss. A shared space across image, text and sound modality is proposed in [4] utilizing ranking loss, which can also be applied to video-text retrieval task.

Utilizing multiple characteristics of video (e.g., activities, audio, locations, time) is evidently crucial for efficient retrieval [63]. In the closely related task of video captioning, dynamic information from video along with static appearance features has been shown to be very effective [65,45]. However, most of the existing video-text retrieval approaches depend on one visual cue for retrieval. In contrast to the existing works, our approach focuses on effectively utilizing different visual cues and audio (if available) concurrently for more efficient retrieval.

Ensemble Approaches. Our retrieval system is based on an ensemble framework [42,16]. A strong psychological context of the ensemble approach can be found from its intrinsic connection in decision making in many daily life situations [42]. Seeking the opinions of several experts, weighing them and combining to make an important decision is an innate behavior of human. The ensemble methods hinge on the same idea and utilize multiple models for making an optimized decision, as in our case diverse cues are available from videos and we would like to utilize multiple expert models which

focus on different cues independently to obtain a stronger prediction model. Moreover, ensemble-based systems have been reported to be very useful when dealing with a lack of adequate training data [42]. As diversity of the models is crucial for the success of ensemble frameworks [43], it is important for our case to choose a diverse set of video-text embeddings that are significantly different from one another.

3 Approach

In this section, we first provide an overview of our proposed framework (Section 3.1). Then, we describe the input feature representation for video and text (Section 3.2). Next, we describe the basic framework for learning visual-semantic embedding using pair-wise ranking loss (Section 3.3). After that, we present our modification on the loss function which improves the basic framework to achieve better recall (Section 3.4). Finally, we present the proposed fusion step for video-text matching (Section 3.5).

3.1 Overview of the Proposed Approach

In a typical cross-modal video-text retrieval system, an embedding network is learned to project video features and text features into the same joint space, and then retrieval is performed by searching the nearest neighbor in the latent space. Since in this work we are looking at videos in general, detecting most relevant information such as object, activities, and places could be very conducive for higher performance. Therefore, along with developing algorithms to train better joint visual-semantic embedding models, it is also very important to develop strategies to effectively utilize different available cues from videos for a more comprehensive retrieval system.

In this work, we propose to leverage the capability of neural networks to learn a deep representation first and fuse the video features in the latent spaces so that we can develop expert networks focusing on specific subtasks (e.g. detecting activities, detecting objects). For analyzing videos, we use a model trained to detect objects, a second model trained to detect activities, and a third model focusing on understanding the place. These heterogeneous features may not be used together directly by simple concatenation to train a successful video-text model as intra-modal characteristics are likely to be suppressed in such an approach. However, an ensemble of video-text models can be used, where a video-text embedding is trained on each of the video features independently. The final retrieval is performed by combining the individual decisions of several experts [42]. An overview of our proposed retrieval framework is shown in Fig. 3. We believe that such an ensemble approach will significantly reduce the chance of poor/wrong prediction.

We follow network architecture proposed in [28] that learns the embedding model using a two-branch network using image-text pairs. One of the branches in this network takes text feature as input and the other branch takes in a video feature. We propose a modified bi-directional pairwise ranking loss to train the embedding. Inspired by the success of ranking loss proposed in [13] in image-text retrieval task, we emphasize on hard negatives. We also apply a weight-based penalty on the loss according to the relative ranking of the correct match in the retrieved result.

3.2 Input Feature Representation

Text Feature. For encoding sentences, we use Gated Recurrent Units (GRU) [11]. We set the dimensionality of the joint embedding space, D, to 1024. The dimension of the word embeddings that are input to the GRU is 300. Note that the word embedding model and the GRU are trained end-to-end in this work.

Object Feature. For encoding image appearance, we adopt deep pre-trained convolutional neural network (CNN) model trained on ImageNet dataset as the encoder. Specifically, we utilize state-of-the-art 152 layer ResNet model ResNet152 [20]. We extract image features directly from the penultimate fully connected layer. We first rescale the image to 224x224 and feed into CNN as inputs. The dimension of the image embedding is 2048.

Activity Feature. The ResNet CNN can efficiently capture visual concepts in static frames. However, an effective approach to learning temporal dynamics in videos was proposed by inflating a 2-D CNN to a deep 3-D CNN named I3D in [7]. We use I3D model to encode activities in videos. In this work, we utilize the pre-trained RGB-I3D model and extract 1024 dimensional feature utilizing continuous 16 frames of video as the input.

Place Feature. For encoding video feature focusing on scene/place, we utilize deep pre-trained CNN model trained on Places-365 dataset as the encoder [66]. Specifically, we utilize 50 layer model ResNet50 [20]. We extract image features directly from the penultimate fully connected layer. We re-scale the image to 224x224 and feed into CNN as inputs. The dimension of the image embedding is 2048.

Audio Feature. We believe that by associating audio, we can get important cues to the real-life events, which would help us remove ambiguity in many cases. We extract audio feature using state-of-the-art SoundNet CNN [3], which provides 1024 dimensional feature from input raw audio waveform. Note that, we only utilize the audio which is readily available with the videos.

3.3 Learning Joint Embedding

In this section, we describe the basic framework for learning joint embedding based on bi-directional ranking loss.

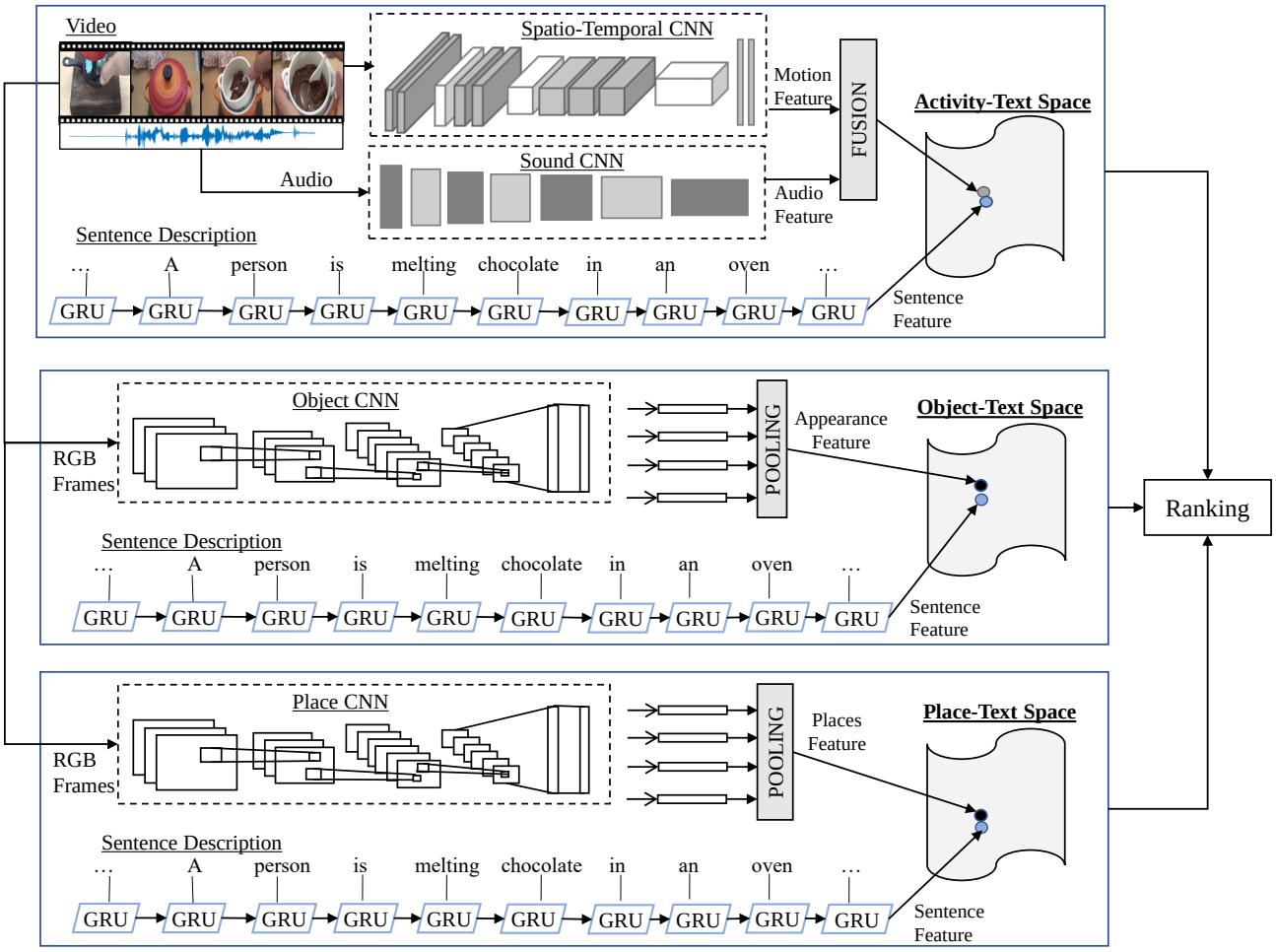

Fig. 3: An overview of the proposed retrieval process. We propose to learn three joint video-text embedding networks as shown in Fig. 3. One model learns a joint space (Object-Text Space) between text features and visual object features. Another model learns a joint space (Activity-Text Space) between text feature and activity features. Similarly, there is a third model which learns a joint space (Place-Text Space) between scene features and text features. Here, Object-Text space is the expert in solving ambiguity related to who is in the video, whereas Activity-Text space is the expert in retrieving what activity is happening and place-Text space is the expert in solving ambiguity regarding locations in the video. Given a query sentence, we calculate the sentence's similarity scores with each one of the videos in the entire dataset in all of the three embedding spaces and use a fusion of scores for the final retrieval result. Please see Section 3.1 for an overview and Section 3 for details.

Given a video feature representation (i.e., appearance feature, or activity feature, or scene feature) $\overline{v}$ ($\overline{v} \in \mathbb{R}^V$), the projection for a video feature on the joint space can be derived as $v = W^{(v)}\overline{v}$ ($v \in \mathbb{R}^D$). In the same way, the projection of input text embedding $\overline{t}(\overline{t} \in \mathbb{R}^T)$ to joint space is $t = W^{(t)}\overline{t}(t \in \mathbb{R}^D)$. Here, $W^{(v)} \in \mathbb{R}^{D \times V}$ is the transformation matrix that projects the video content into the joint embedding space, and D denotes the dimension of the joint space. Similarly, $W^{(t)} \in \mathbb{R}^{D \times T}$ maps input sentence/caption embedding to the joint space. Given feature representation for words in a sentence, the sentence embedding $\overline{t}$ is found from the hidden state of the GRU. Here, given the feature representation of both videos and corresponding text, the goal is to learn a joint embedding characterized by θ (i.e., $W^{(v)}$, $W^{(t)}$ and GRU weights) such that the video content and semantic content are projected into the joint embedding space. We keep image encoder (e.g., pre-trained CNN) fixed in this work, as the video-text datasets are small in size.

In the embedding space, it is expected that the similarity between a video and text pair to be more reflective of semantic closeness between videos and their corresponding texts. Many prior approaches have utilized pairwise ranking loss for learning joint embedding between visual input and textual input. They minimize a hinge based triplet ranking loss combining bi-directional ranking terms, in order to maximize the similarity between a video embedding and the corresponding text embedding, and while at the same time, minimize the similarity to all other non-matching ones. The optimization

problem can be written as,

$$\min_{\theta} \sum_{v} \sum_{t^-} [\alpha - S(v,t) + S(v,t^-)]_+ + \sum_{t} \sum_{v^-} [\alpha - S(t,v) + S(t,v^-)]_+ \quad (1)$$

where, $[f]_+ = max(0, f)$. t^- is a non-matching text embedding, and t is the matching text embedding for video embedding v. This is similar for text embedding t. α is the margin value for the pairwise ranking loss. The scoring function $S(v,t)$ is defined as the similarity function to measure the similarity between the videos and text in the joint embedded space. We use cosine similarity in this work, as it is easy to compute and shown to be very effective in learning joint embeddings. [28,13].

In Eq. (1), in the first term, for each pair (v,t), the sum is taken over all non-matching text embedding t^-. It attempts to ensure that for each visual feature, corresponding/matching text features should be closer than non-matching ones in the joint space. Similarly, the second term attempts to ensure that text embedding that corresponds to the video embedding should be closer in the joint space to each other than non-matching video embeddings.

3.4 Proposed Ranking Loss

Recently, focusing on hard-negatives has been shown to be effective in many embedding tasks [13,48,33]. Inspired by this, we focus on hard negatives (i.e., the negative video and text sample closest to a positive/matching (v,t) pair) instead of summing over all negatives in our formulation. For a positive/matching pair (v,t), the hardest negative sample can be identified using $\hat{v} = \arg\max_{v^-} S(t,v^-)$ and $\hat{t} = \arg\max_{t^-} S(v,t^-)$. The optimization problem can be rewritten as following to focus on hard-negatives,

$$\min_{\theta} \sum_{v} [\alpha - S(v,t) + S(v,\hat{t})]_+ + \sum_{t} [\alpha - S(t,v) + S(t,\hat{v})]_+ \quad (2)$$

The loss in Eq. 2 is similar to the loss in Eq. 1 but it is specified in terms of the hardest negatives [13]. We start with the loss function in Eq. 2 and further modify the loss function following the idea of weighted ranking [51] to weigh the loss based on the relative ranking of positive labels.

$$\min_{\theta} \sum_{v} L(r_v)[\alpha - S(v,t) + S(v,\hat{t})]_+ + \sum_{t} L(r_t)[\alpha - S(t,v) + S(t,\hat{v})]_+ \quad (3)$$

where $L(.)$ is a weighting function for different ranks. For a video embedding v, r_v is the rank of matching sentence t among all compared sentences. Similarly, for a text embedding t, r_t is the rank of matching video embedding v among all compared videos in the batch. We define the weighting function as $L(r) = (1 + \beta/(N - r + 1))$, where N is the number of compared videos and β is the weighting factor. Fig. 4 shows an example showing the significance of the proposed ranking loss.

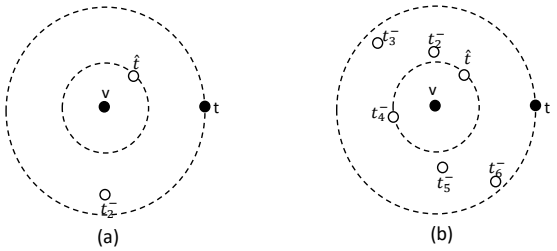

Fig. 4: An example showing the significance of the proposed ranking loss. The idea is that if a large number of non-matching instances are ranked higher than the matching one given the current state of the model, then the model must be updated by a larger amount (Case: (b). However, the model needs to be updated by a smaller amount if the matching instance is already ranked higher than most non-matching ones. (Case: (a))

It is very common, in practice, to only compare samples within a mini-batch at each iteration rather than comparing the entire training set for computational efficiency [33,48, 25]. This is known as semi-hard negative mining [33,48]. Moreover, selecting the hardest negatives in practice may often lead to a collapsed model and semi-hard negative mining helps to mitigate this issue [33,48]. We utilize a batch-size of 128 in our experiment.

It is evident from Eq. 3 that the loss applies a weight-based penalty based on the relative ranking of the correct match in retrieved result. If a positive match is ranked top in the list, then $L(.)$ will assign a small weight to the loss and will not cost the loss too much. However, if a positive match is not ranked top, $L(.)$ will assign a much larger weight to the loss, which will ultimately try to push the positive matching pair to the top of rank.

3.5 Matching and Ranking

The video-text retrieval task focuses on returning for each query video, a ranked list of the most likely text description from a dataset and vice versa. We believe, we need to understand three main aspects of each video: (1) Who: the salient objects of the video, (2) What: the action and events in the video and (3) Where: the place aspect of the video. To achieve this, we learn three expert joint video-text embedding spaces as shown in Fig. 3.

The Object-Text embedding space is the common space where both appearance features and text feature are mapped to. Hence, this space can link video and sentences focusing on the objects. On the other hand, the Activity-Text embedding space focuses on linking video and language description which emphasizes more on the events in the video. Action features and audio features both provide important cues for understanding different events in a video. We fuse action and audio features (if available) by concatenation and map the concatenated feature and text feature into a common space, namely, the Activity-Text space. If the audio feature is absent from videos, we only use the action feature as the video representation for learning the Activity-Text space. The Place-Text embedding space is the common space where visual features focusing on scene/place aspect and text feature are mapped to. Hence, this space can link video and sentences focusing on the entire scene. We utilize the same loss functions described in Sec. 3.4 for training these embedding models.

At the time of retrieval, given a query sentence, we compute the similarity score of the query sentence with each one of the videos in the dataset in three video-text embedding spaces and use a fusion of similarity scores for the final ranking. Conversely, given a query video, we calculate its similarity scores with all the sentences in the dataset in three embedding spaces and use a fusion of similarity scores for the final ranking.

$$S_{v-t}(v,t) = w_1 S_{o-t} + w_2 S_{a-t} + w_3 S_{p-t} \quad (4)$$

It may be desired to use a weighted sum when it is necessary in a task to put more emphasis on one of the facets of the video (objects or captions or scene). In this work, we empirically found putting comparatively higher importance to S_{o-t} (Object-Text) and S_{a-t} (Activity-Text), and slightly lower importance to S_{p-t} (Place-Text) works better in evaluated datasets than putting equal importance to all. We empirically choose $w_1 = 1$, $w_2 = 1$ and $w_3 = 0.5$ in our experiments based on our evaluation on the validation set.

4 Experiments

In this section, we first describe the datasets and evaluation metric (Section 4.1). Then, we describe the training details. Next, we provide quantitative results on MSR-VTT dataset (Section 4.3) and MSVD dataset (Section 4.4) to show the effectiveness of our proposed framework. Finally. we present some qualitative examples analyzing our success and failure cases (Section 4.5).

4.1 Datasets and Evaluation Metric

We present experiments on two standard benchmark datasets: Microsoft Research Video to Text (MSR-VTT) Dataset [60] and Microsoft Video Description dataset (MSVD) [9] to evaluate the performance of our proposed framework. We adopt rank-based metric for quantitative performance evaluation.

MSR-VTT. The MSR-VTT is a large-scale video description dataset. This dataset contains 10,000 video clips. The dataset is split into 6513 videos for training, 2990 videos for testing and 497 videos for the validation set. Each video has 20 sentence descriptions. This is one of the largest video captioning dataset in terms of the quantity of sentences and the size of the vocabulary.

MSVD. The MSVD dataset contains 1970 Youtube clips, and each video is annotated with about 40 sentences. We use only the English descriptions. For a fair comparison, we used the same splits utilized in prior works [53], with 1200 videos for training, 100 videos for validation, and 670 videos for testing. The MSVD dataset is also used in [40] for video-text retrieval task, where they randomly chose 5 ground-truth sentences per video. We use the same setting when we compare with that approach.

Evaluation Metric. We use the standard evaluation criteria used in most prior work on image-text retrieval and video-text retrieval task [40, 28, 12]. We measure rank-based performance by $R@K$, Median Rank($MedR$) and Mean Rank ($MeanR$). $R@K$ (Recall at K) calculates the percentage of test samples for which the correct result is found in the top-K retrieved points to the query sample. We report results for $R@1$, $R@5$ and $R@10$. Median Rank calculates the median of the ground-truth results in the ranking. Similarly, Mean Rank calculates the mean rank of all correct results.

4.2 Training Details

We used two Titan Xp GPUs for this work. We implemented the network using PyTorch following [13]. We start training with a learning rate of 0.002 and keep the learning rate fixed for 15 epochs. Then the learning rate is lowered by a factor of 10 and the training continued for another 15 epochs. We use a batch-size of 128 in all the experiments. The embedding networks are trained using ADAM optimizer [27]. When the L2 norm of the gradients for the entire layer exceeds 2, gradients are clipped. We tried different values for margin α in training and found $0.1 \leq \alpha \leq 0.2$ works reasonably well. We empirically choose α as 0.2. The embedding model was evaluated on the validation set after every epoch. The model with the best sum of recalls on the validation set is chosen as the final model.

4.3 Results on MSR-VTT Dataset

We report the result on MSR-VTT dataset [60] in Table 1. We implement several baselines to analyze different components of the proposed approach. To understand the effect of

Table 1: Video-to-Text and Text-to-Video Retrieval Results on MSR-VTT Dataset.

#	Method	Video-to-Text Retrieval					Text-to-Video Retrieval				
		$R@1$	$R@5$	$R@10$	$MedR$	$MeanR$	$R@1$	$R@5$	$R@10$	$MedR$	$MeanR$
1.1	VSE (Object-Text)	7.7	20.3	31.2	28.0	185.8	5.0	16.4	24.6	47.0	215.1
	VSEPP (Object-Text)	10.2	25.4	35.1	25.0	228.1	5.7	17.1	24.8	65.0	300.8
1.2	Ours (Object-Text)	10.5	26.7	35.9	25.0	266.6	5.8	17.6	25.2	61.0	296.6
	Ours (Audio-Text)	0.4	1.1	1.9	1051	2634.9	0.2	0.9	1.5	1292	1300
	Ours (Activity-Text)	8.4	22.2	32.3	30.3	229.9	4.6	15.3	22.7	71.0	303.7
	Ours (Place-Text)	7.1	19.8	28.7	38.0	275.1	4.3	14.0	21.1	77.0	309.6
1.3	CON(Object, Activity)-Text	9.1	24.6	36.0	23.0	181.4	5.5	17.6	25.9	51.0	243.4
	CON(Object, Activity, Audio)-Text	9.3	27.8	38.0	22.0	162.3	5.7	18.4	26.8	48.0	242.5
1.4	Joint Image-Text-Audio Embedding	8.7	22.4	32.1	31.0	225.8	4.8	15.3	22.9	73.0	313.6
1.5	Fusion [Object-Text, Activity (I3D)-Text]	12.3	31.3	42.9	16.0	145.4	6.8	20.7	29.5	39.0	224.7
	Fusion [Object-Text, Activity(I3d-Audio)-Text]	12.5	32.1	42.4	16.0	134.0	7.0	20.9	29.7	38.0	213.8
	Fusion [Object-Text, Place-Text]	11.8	30.1	40.8	18.0	172.1	6.5	19.9	28.5	43.0	234.1
	Fusion [Activity-Text, Place-Text]	11.0	28.4	39.3	20.0	152.1	5.9	18.6	27.4	44.0	224.7
1.6	**Fusion [Object-Text, Activity-Text, Place-Text]**	**13.8**	**33.5**	**44.3**	**14.0**	**119.2**	**7.3**	**21.7**	**30.9**	**34.0**	**196.1**
1.7	Rank Fusion [Object-Text, Activity-Text, Place-Text]	12.2	31.6	42.7	16.0	127.6	6.8	20.5	29.4	38.0	204.3

different loss functions, features, effect of feature concatenation and proposed fusion method, we divide the table into 7 rows (1.1-1.7). In row-1.1, we report the results on applying two different variants of pair-wise ranking loss. VSE[28] is based on the basic triplet ranking loss similar to Eq. 1 and VSEPP[13] is based on the loss function that emphasizes on hard-negatives as shown in Eq. 2. Note that, all other reported results in Table 1 are based on the modified pairwise ranking loss proposed in Eq. 3. In row-1.2, we provide the performance of different features in learning the embedding using the proposed loss. In row-1.3, we present results for the learned embedding utilizing a feature vector that is a direct concatenation of different video features. In row-1.4, we provide the result when a shared representation between image, text and audio modality is learned using proposed loss following the idea in [4] and used for video-text retrieval task. In row-1.5, we provide the result based on the proposed approach that employs two video-text joint embeddings for retrieval. In row-1.6, we provide the result based on the proposed ensemble approach that employs all three video-text joint embeddings for retrieval. Additionally, in row-1.7, we also provide the result for the case where the rank fusion has been considered in place of the proposed score fusion.

Loss Function. For evaluating the performance of different ranking loss functions in the task, we can compare results reported in row-1.1 and row-1.2. We can choose only results based on Object-Text spaces from these two rows for a fair comparison. We see that VSEPP loss function and proposed loss function performs significantly better than the traditional VSE loss function in $R@1$, $R@5$, $R@10$. However, VSE loss function has better performance in terms of the mean rank. This phenomenon is expected based on the characteristics of the loss functions. As higher $R@1$, $R@5$ and $R@10$ are more desirable for a efficient video-text retrieval system than the mean rank, we see that our proposed loss function performs better than other loss functions in this task. We observe similar performance improvement using our loss function in other video-text spaces too.

Video Features. We can compare the performance of different video features in learning the embedding using the proposed loss from row-1.2. We observe that object feature and activity feature from video performs reasonably well in learning a joint video-text space. The performance is very low when only audio feature is used for learning the embedding. It can be expected that the natural sound associated in a video alone does not contain as much information as videos in most cases. However, utilizing audio along with i3d feature as activity features provides a slight boost in performance as shown in row-1.3 and row-1.4.

Feature Concatenation for Representing Video. Rather than training multiple video-semantic spaces, one can argue that we can simply concatenate all the available video features and learn a single video-text space using this concatenated video feature [12, 60]. However, we observe from row-1.3 that integrating complementary features by static concatenation based fusion strategy fails to utilize the full potential of different video features for the task. Comparing row-1.2 and row-1.3, we observe that a concatenation of object feature, activity feature and Audio feature performs even worse than utilizing only object feature in $R@1$. Although we see some improvement in other evaluation metrics, overall the improvement is very limited. We believe that both appearance feature and action feature gets suppressed in such concatenation as they focus on representing different entities of a video.

Table 2: Video-to-Text Retrieval Results on MSVD Dataset. We highlight the proposed **method**. The methods which has 'Ours' keyword in name are trained with the proposed loss.

Method	*R@1*	*R@5*	*R@10*	*MedR*	*MeanR*
Results Using Partition used by JMET and JMDV					
CCA					245.3
JMET					208.5
JMDV					224.1
W2VV-ResNet152	16.3		44.8	14.0	110.2
VSE (Object-Text)	15.8	30.2	41.4	12.0	84.8
VSEPP(Object-Text)	21.2	43.4	52.2	9.0	79.2
Ours(Object-Text)	23.4	45.4	53.0	8.0	75.9
Ours(Activity-Text)	21.3	43.7	53.3	9.0	72.2
Ours(Place-Text)	11.2	25.1	34.3	27.0	147.7
Ours-Fusion(O-T, P-T)	25.7	45.4	54.0	7.0	65.4
Ours-Fusion(A-T, P-T)	26.0	46.1	55.8	7.0	53.5
Ours-Fusion(O-T, A-T)	31.5	51.0	61.5	5.0	41.7
Ours-Fusion(O-T, A-T, P-T)	**33.3**	**52.5**	**62.5**	**5.0**	**40.2**
Rank-Fusion(O-T, A-T, P-T)	30.0	51.3	61.8	5.0	42.3
Results Using Partition used by LJRV					
ST	2.99	10.9	17.5	77.0	241.0
LJRV	9.85	27.1	38.4	19.0	75.2
W2VV(Object-Text)	17.9	-	49.4	11.0	57.6
Ours(Object-Text)	20.9	43.7	54.9	7.0	56.1
Ours(Activity-Text)	17.5	39.6	51.3	10.0	54.8
Ours(Place-Text)	8.5	23.3	32.7	26.0	99.3
Ours-Fusion(O-T, A-T)	25.5	51.3	61.9	5.0	32.5
Ours-Fusion(O-T, A-T, P-T)	**26.4**	**51.9**	**64.5**	**5.0**	**31.1**
Rank-Fusion(O-T, A-T, P-T)	24.3	49.3	62.4	6.0	34.6

Learning a Shared Space across Image, Text and Audio. Learning a shared space across image, text and sound modality is proposed for cross-modal retrieval task in [4]. Following the idea, we trained a shared space across video-text-sound modality using the pairwise ranking loss by utilizing video-text and video-sound pairs. The result is reported in row-1.4. We observe that performance in video-text retrieval task degrades after training such an aligned representation across 3 modalities. Training such a shared representation gives the flexibility to transfer across multiple modalities. Nevertheless, we believe it is not tailored towards achieving high performance in a specific task. Moreover, aligning across 3 modalities is a more computationally difficult task and requires many more examples to train.

Proposed Fusion. The best result in Table. 1 is achieved by our proposed fusion approach as shown in row-1.6. We see that the proposed method achieves 31.43% improvement in $R@1$ for text retrieval and 25.86% improvement for video retrieval in $R@1$ compared to best performing Ours(Object-text) as shown in row-1.2, which is the best among the other methods which use a single embedding space for the retrieval task. In row-1.5, Fusion[Object-text & Activity(I3D-Audio)-text] differs from Fusion[Object-text & Activity(I3D)-text] in the feature used in learning the activity-text space. We see that utilizing audio in learning the embedding improves the result slightly. However, as the retrieval performance of individual audio feature is very low (shown in row-1.2), we did not utilize audio-text space separately in fusion as we found it degraded the performance significantly.

Table 3: Text-to-Video Retrieval Results on MSVD Dataset. We highlight the proposed **method**.

Method	*R@1*	*R@5*	*R@10*	*MedR*	*MeanR*
Results Using Partition used by JMET and JMDV					
CCA					251.3
JMDV					236.3
VSE(Object-Text)	12.3	30.1	42.3	14.0	57.7
VSEPP(Object-Text)	15.4	39.6	53.0	9.0	43.8
Ours(Object-Text)	16.1	41.1	53.5	9.0	42.7
Ours(Activity-Text)	15.4	39.2	51.4	10.0	43.2
Ours(Place-Text)	7.9	24.5	36.0	21.0	64.6
Ours-Fusion(O-T, P-T)	17.0	42.2	56.0	8.0	36.5
Ours-Fusion(A-T, P-T)	17.2	42.6	55.6	8.0	34.1
Ours-Fusion(O-T, A-T)	20.3	47.8	61.1	6.0	28.3
Ours-Fusion(O-T, A-T, P-T)	**21.3**	**48.5**	**61.6**	**6.0**	**26.3**
Rank-Fusion(O-T, A-T, P-T)	19.4	45.8	59.4	7.0	29.2
Results Using Partition used by LJRV					
ST	2.6	11.6	19.3	51.0	106.0
LJRV	7.7	23.4	35.0	21.0	49.1
Ours(Object-Text)	15.0	40.2	51.9	9.0	45.3
Ours(Activity-Text)	14.6	38.9	51.0	10.0	45.1
Ours(Place-Text)	7.9	24.5	36.0	21.0	64.6
Ours-Fusion(O-T, A-T)	20.2	47.5	60.7	6.0	29.0
Ours-Fusion(O-T, A-T, P-T)	**20.7**	**47.8**	**61.9**	**6.0**	**26.8**
Rank-Fusion(O-T, A-T, P-T)	18.5	44.9	58.8	7.0	30.2

Comparing row-1.6, row-1.5 and row-1.2, we find that the ensemble approach with score fusion results in significant improvement in performance, although there is no guarantee that the combination of multiple models will perform better than the individual models in the ensemble in every single case. However, the ensemble average consistently improves performance significantly.

Rank vs Similarity Score in Fusion. We provide the retrieval result based on weighted rank aggregation of three video-text spaces in row-1.7. Comparing the effect of rank fusion in replacement of the score fusion from row-1.6 and row-1.7 in Table. 1, it is also evident that the proposed score fusion approach shows consistent performance improvement over rank fusion approach. It is possible that exploiting similarity score to combine multiple evidences may be less effective than using rank values in some cases, as score fusion approach independently weights scores and does not consider overall performance in weighting [31]. However, we empirically find that utilizing score fusion is more advantageous than rank fusion in our system in terms of retrieval effectiveness.

4.4 Results on MSVD Dataset

We report the results of video to text retrieval task on MSVD dataset [9] in Table 2 and the results for text to video retrieval in Table 3.

We compare our approach with existing video-text retrieval approaches, CCA[49], ST [29], JMDV [61], LJRV [40], JMET [41], and W2VV [12]. For these approaches, we directly cite scores from respective papers when available.

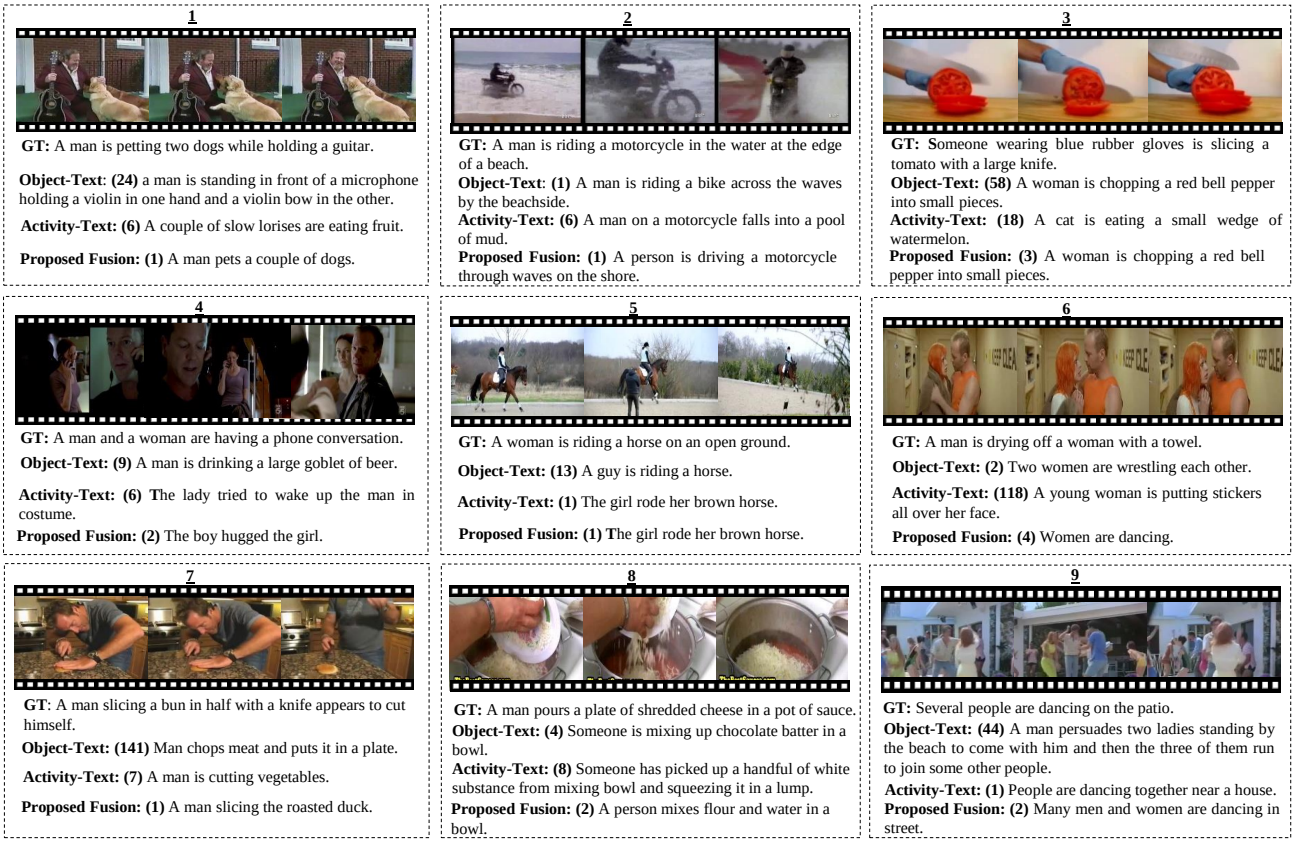

Fig. 5: Examples of 9 test videos from MSVD dataset and the top 1 retrieved captions by using a single video-text space and the fusion approach with our loss function. The value in brackets is the rank of the highest ranked ground-truth caption. Ground Truth (GT) is a sample from the ground-truth captions. Among all the approaches, object-text (ResNet152 as video feature) and activity-text (I3D as video feature) are systems where single video-text space is used for retrieval. We also report result for the fusion system where three video-text spaces (object-text, activity-text and place-text) are used for retrieval.

We report score for JMET from [12]. The score of CCA is reported from [61] and the score of ST is reported from [40]. If scores for multiple models are reported, we select the score of the best performing method from the paper.

We also implement and compare results with state-of-the-art image-embedding approach VSE[28] and VSEPP[13] in the Object-Text(O-T) embedding space. Additionally, to show the impact of only using the proposed loss in retrieval, we also report results based on the Activity-Text(A-T) space and Place-Text(P-T) space in the tables. Our proposed fusion is named as Ours-Fusion(O-T,A-T,P-T) in the Table. 2 and Table. 3. The proposed fusion system utilizes the proposed loss and employs three video-text embedding spaces for calculating the similarity between video and text. As the audio is muted in this dataset, we train the Activity-Text space utilizing only I3D feature from videos. We also report results for our fusion approach using any two of the three video-text spaces in the tables. Additionally, we report results of Rank-Fusion(O-T, A-T, P-T), which uses rank in place of similarity score in combining retrieval results of three video-text spaces in the fusion system.

From Table 2 and Table 3, it is evident that our proposed approach performs significantly better than existing ones. The result is improved significantly by utilizing the fusion proposed in this paper that utilizes multiple video-text spaces in calculating the final ranking. Moreover, utilizing the proposed loss improves the result over previous state-of-the-art methods. It can also be identified that our loss function is not only useful for learning embedding independently, but also it is useful for the proposed fusion. We observe that utilizing the proposed loss function improves the result over previous state-of-the-art methods consistently, with a minimum improvement of 10.38% from best existing method VSEPP(Object-Text) in Video-to-Text Retrieval and 4.55% in Text-to-Video Retrieval. The result is improved further by utilizing the proposed fusion framework in this paper that utilizes multiple video-text spaces in an ensemble fusion approach in calculating the final ranking, with an improvement of 57.07% from the best existing method in the video to text retrieval and 38.31% in the text to video retrieval. Among the video-text spaces, object-text and activity-text space show better performance in retrieval, compared to place-text space

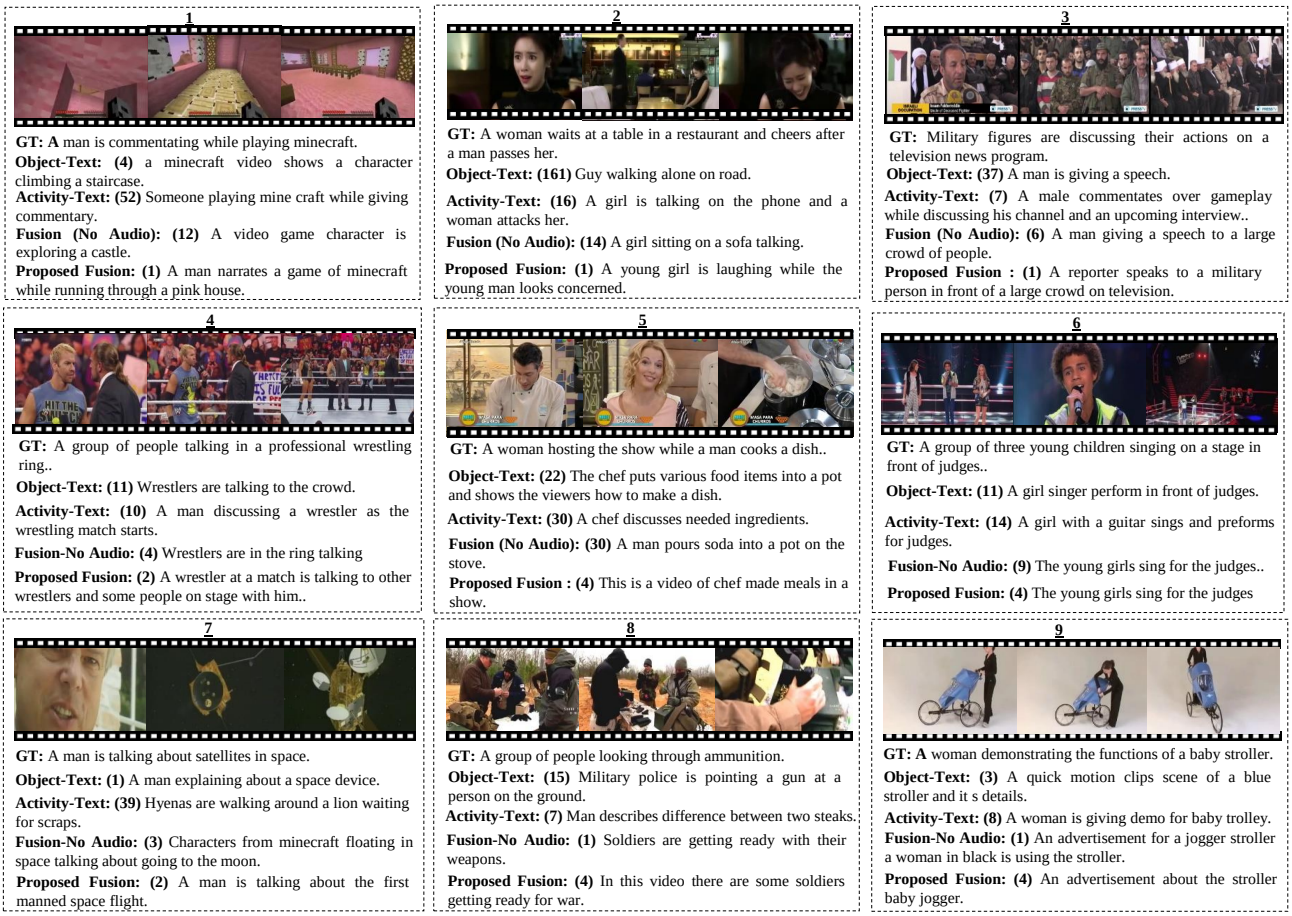

Fig. 6: A snapshot of 9 test videos from MSR-VTT dataset with success and failure cases, the top 1 retrieved captions for four approaches based on the proposed loss function and the rank of the highest ranked ground-truth caption inside the bracket. Among the approaches, Object-Text is trained using ResNet feature as video feature and Activity-Text is trained using the concatenated I3D feature and Audio feature as the video feature. We also report results for fusion approaches where three video-text spaces are used for retrieval. The fusion approaches use an object-text space trained with ResNet feature and place-text space trained with ResNet50(Place) feature, while in the proposed fusion, the activity-text space is trained using concatenated I3D and Audio feature. Fusion (No Audio) utilizes activity-text space trained with only the I3D feature.

which indicates that the annotators focused more on object and activity aspects in annotating the videos. Similar to the results of MSR-VTT dataset, we observe that the proposed score fusion approach consistently shows superior performance than rank fusion approach in both video to text and text to video retrieval.

4.5 Qualitative Results

We report the qualitative results on MSVD dataset in Fig. 5 and the results on MSR-VTT dataset in Fig. 6.

MSVD Dataset. In Fig. 5, we show examples of a few test videos from MSVD dataset and the top 1 retrieved captions for the proposed approach. We also show the retrieval result when only one of the embeddings is used for retrieval. Additionally, we report the rank of the highest ranked ground-truth caption in the figure. We can observe from the figure that in most of the cases, utilizing cue from multiple video-text spaces helps to match the correct caption. We see from Fig. 5 that, among 9 videos, the retrieval performance is improved or higher recall is retained for 7 videos. Video-6 and video-9 show two failure cases, where utilizing multiple video-text spaces degrades the performance slightly than using object-text in Video-6 and activity-text space in Video-9.

MSR-VTT Dataset. Similar to Fig. 5, we also show qualitative results for a few test videos from MSR-VTT dataset in Fig. 6. Video 1-6 in Fig. 6 shows a few examples where utilizing cue from multiple video-text spaces helps to match the correct caption compared to using only one of the video-text space. Moreover, we also see the result was improved after utilizing audio in learning the second video-text space (Activity-text space). We observe this improvement for most of the videos, as we also observe from Table. 1. Video

7-9 shows some failure cases for our fusion approach in Fig. 6. Video 7 shows a case, where utilizing multiple video-text spaces for retrieval degrades the performance slightly compared to utilizing only one of the video-text space. For Video-8 and video-9 in Fig. 6, we observe that the performance improves after fusion overall, but the performance is better when the audio is not used in learning video-text space. On the other hand, video 1-6 includes cases where utilizing audio helped to improve the result.

4.6 Discussion

The experimental results are aligned with our rationale that utilizing multiple characteristics of a video is crucial for developing an efficient video-text retrieval system. Experiments also demonstrate that our proposed ranking loss function is effective in learning video-text embeddings better than existing ones. However, we observe that major improvement in experimental performance comes from our mixture of experts system which utilizes evidence from three complementary video-text spaces for retrieval. Our mixture of expert video-text model may not outperform the performance of a single video-text model in the ensemble in every single case, but it is evident from experiments that our system significantly reduces the overall risk of making a particularly poor decision.

From qualitative results, we observe it cannot be claimed in general that one video feature is consistently better than others for the task of video-text retrieval. It can be easily identified from the top-1 retrieved captions in Fig. 5 and Fig. 6 that the video-text embedding (Object-Text) learned utilizing object appearance feature (ResNet) as video feature is significantly different from the joint embedding (Activity-Text) learned using Activity feature (I3D) as video feature. The variation between the rank of the highest matching caption further strengthens this observation. Object-text space performs better than the activity-text space in retrieval for some videos. For other videos, the activity-text space achieves higher performance. However, it can be claimed that combining knowledge from multiple video-text embedding spaces consistently shows better performance than utilizing only one of them.

We observe from Fig. 6 that using audio is crucial in many cases where there is deep semantic relation between visual content and audio (e.g., the audio is from the third person narration of the video, the audio is music or song) and it gives important cues in reducing description ambiguity (e.g., video-2, video-5 and video-6 in Fig. 6). We observe that the performance degrades in some cases when audio is utilized in the system (e.g., video-8 in Fig. 6). We see an overall improvement in the quantitative result (Table 1) which also supports our idea of using audio. Since we did not exploit the structure of the audio and analyze the structural alignment between audio and video, it is difficult to determine whether audio is always helpful. For instance, audio can come from different things (persons, animals or objects) in a video, and it might shift our attention away from the main subject. Moreover, the captions are provided mostly based on visual aspects, which makes audio information very sparse. Hence, the overall improvement using audio was limited.

5 Conclusions

In this paper, we study how to leverage diverse video features effectively for developing a robust cross-modal video-text retrieval system. Our proposed framework learns three expert video-text embedding models focusing on three salient video cues (i.e., object, activity, place) and uses a combination of these models for high-quality prediction. A modified pairwise ranking loss function is also proposed for better learning the joint embeddings, which focuses on hard negatives and applies a weight-based penalty based on the relative ranking of the correct match. Extensive quantitative and qualitative evaluations of MSVD and MSR-VTT datasets demonstrate that our framework performs significantly better than baselines and state-of-the-art systems. Moving forward, we would like to improve our system by developing more sophisticated algorithms to combine evidence from multiple joint spaces and further analyze the role of associated audio for video-text retrieval.

Acknowledgements This work was partially supported by NSF grants 33384, IIS-1746031, CNS-1544969, ACI-1548562 and ACI-1445606. J. Li was supported by the Bosch Graduate Fellowship to CMU LTI. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

1. Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: International Conference on Machine Learning, pp. 1247–1255 (2013)
2. Awad, G., Butt, A., Fiscus, J., Joy, D., Delgado, A., Michel, M., Smeaton, A.F., Graham, Y., Kraaij, W., Quénot, G., et al.: Trecvid 2017: Evaluating ad-hoc and instance video search, events detection, video captioning and hyperlinking. In: Proceedings of TRECVID (2017)
3. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: Advances in Neural Information Processing Systems, pp. 892–900 (2016)
4. Aytar, Y., Vondrick, C., Torralba, A.: See, hear, and read: Deep aligned representations. arXiv preprint arXiv:1706.00932 (2017)
5. Bois, R., Vukotić, V., Simon, A.R., Sicre, R., Raymond, C., Sébillot, P., Gravier, G.: Exploiting multimodality in video hyperlinking to improve target diversity. In: International Conference on Multimedia Modeling, pp. 185–197. Springer (2017)
6. Budnik, M., Demirdelen, M., Gravier, G.: A study on multimodal video hyperlinking with visual aggregation. In: 2018 IEEE International Conference on Multimedia and Expo, pp. 1–6. IEEE (2018)

7. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4724–4733. IEEE (2017)
8. Cha, M., Gwon, Y., Kung, H.: Multimodal sparse representation learning and applications. arXiv preprint arXiv:1511.06238 (2015)
9. Chen, D.L., Dolan, W.B.: Collecting highly parallel data for paraphrase evaluation. In: Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pp. 190–200. ACL (2011)
10. Chi, J., Peng, Y.: Dual adversarial networks for zero-shot cross-media retrieval. In: International Joint Conferences on Artificial Intelligence, pp. 663–669 (2018)
11. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
12. Dong, J., Li, X., Snoek, C.G.: Word2visualvec: Image and video to sentence matching by visual feature prediction. arXiv preprint arXiv:1604.06838 (2016)
13. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: Improved visual-semantic embeddings. British Machine Vision Conference (BMVC) (2018)
14. Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.: Every picture tells a story: Generating sentences from images. In: European Conference on Computer Vision, pp. 15–29. Springer (2010)
15. Feng, F., Wang, X., Li, R.: Cross-modal retrieval with correspondence autoencoder. In: ACM Multimedia Conference, pp. 7–16. ACM (2014)
16. Fraz, M.M., Remagnino, P., Hoppe, A., Uyyanonvara, B., Rudnicka, A.R., Owen, C.G., Barman, S.A.: An ensemble classification-based approach applied to retinal blood vessel segmentation. IEEE Transactions on Biomedical Engineering **59**(9), 2538–2548 (2012)
17. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al.: Devise: A deep visual-semantic embedding model. In: Advances in Neural Information Processing Systems, pp. 2121–2129 (2013)
18. Gong, Y., Ke, Q., Isard, M., Lazebnik, S.: A multi-view embedding space for modeling internet images, tags, and their semantics. International Journal of Computer Vision **106**(2), 210–233 (2014)
19. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: An overview with application to learning methods. Neural Computation **16**(12), 2639–2664 (2004)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. IEEE (2016)
21. Henning, C.A., Ewerth, R.: Estimating the information gap between textual and visual representations. In: International Conference on Multimedia Retrieval, pp. 14–22. ACM (2017)
22. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research **47**, 853–899 (2013)
23. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2261–2269. IEEE (2017)
24. Huang, Y., Wang, W., Wang, L.: Instance-aware image and sentence matching with selective multimodal lstm. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2310–2318. IEEE (2017)
25. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128–3137. IEEE (2015)
26. Karpathy, A., Joulin, A., Li, F.F.F.: Deep fragment embeddings for bidirectional image sentence mapping. In: Advances in Neural Information Processing Systems, pp. 1889–1897 (2014)
27. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
28. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539 (2014)
29. Kiros, R., Zhu, Y., Salakhutdinov, R.R., Zemel, R., Urtasun, R., Torralba, A., Fidler, S.: Skip-thought vectors. In: Advances in Neural Information Processing Systems, pp. 3294–3302 (2015)
30. Klein, B., Lev, G., Sadeh, G., Wolf, L.: Associating neural word embeddings with deep image representations using fisher vectors. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4437–4446. IEEE (2015)
31. Lee, J.H.: Analyses of multiple evidence combination. In: ACM SIGIR Forum, vol. 31, pp. 267–276. ACM (1997)
32. Ma, Z., Lu, Y., Foster, D.: Finding linear structure in large datasets with scalable canonical correlation analysis. In: International Conference on Machine Learning, pp. 169–178 (2015)
33. Manmatha, R., Wu, C.Y., Smola, A.J., Krahenbuhl, P.: Sampling matters in deep embedding learning. In: IEEE International Conference on Computer Vision, pp. 2859–2867. IEEE (2017)
34. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp. 3111–3119 (2013)
35. Mithun, N.C., Li, J., Metze, F., Roy-Chowdhury, A.K.: Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In: ACM International Conference on Multimedia Retrieval (2018)
36. Mithun, N.C., Munir, S., Guo, K., Shelton, C.: Odds: real-time object detection using depth sensors on embedded gpus. In: ACM/IEEE International Conference on Information Processing in Sensor Networks, pp. 230–241. IEEE Press (2018)
37. Mithun, N.C., Panda, R., Roy-Chowdhury, A.K.: Generating diverse image datasets with limited labeling. In: ACM Multimedia Conference, pp. 566–570. ACM (2016)
38. Mithun, N.C., Rameswar, P., Papalexakis, E., Roy-Chowdhury, A.: Webly supervised joint embedding for cross-modal image-text retrieval. In: ACM International Conference on Multimedia (2018)
39. Nam, H., Ha, J.W., Kim, J.: Dual attention networks for multimodal reasoning and matching. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 299–307. IEEE (2017)
40. Otani, M., Nakashima, Y., Rahtu, E., Heikkilä, J., Yokoya, N.: Learning joint representations of videos and sentences with web image search. In: European Conference on Computer Vision, pp. 651–667. Springer (2016)
41. Pan, Y., Mei, T., Yao, T., Li, H., Rui, Y.: Jointly modeling embedding and translation to bridge video and language. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4594–4602. IEEE (2016)
42. Polikar, R.: Ensemble based systems in decision making. IEEE Circuits and systems magazine **6**(3), 21–45 (2006)
43. Polikar, R.: Bootstrap inspired techniques in computational intelligence: ensemble of classifiers, incremental learning, data fusion and missing features. IEEE Signal Processing Magazine **24**(4), 59–72 (2007)
44. Quynh Nhi Tran, T., Le Borgne, H., Crucianu, M.: Aggregating image and text quantized correlated components. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2046–2054 (2016)
45. Ramanishka, V., Das, A., Park, D.H., Venugopalan, S., Hendricks, L.A., Rohrbach, M., Saenko, K.: Multimodal video description. In: ACM Multimedia Conference, pp. 1092–1096. ACM (2016)
46. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788. IEEE (2016)
47. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp. 91–99 (2015)

48. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823. IEEE (2015)
49. Socher, R., Fei-Fei, L.: Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 966–973. IEEE (2010)
50. Torabi, A., Tandon, N., Sigal, L.: Learning language-visual embedding for movie understanding with natural-language. arXiv preprint arXiv:1609.08124 (2016)
51. Usunier, N., Buffoni, D., Gallinari, P.: Ranking with ordered weighted pairwise classification. In: International Conference on Machine Learning, pp. 1057–1064. ACM (2009)
52. Vendrov, I., Kiros, R., Fidler, S., Urtasun, R.: Order-embeddings of images and language. In: International Conference on Learning Representations (2016)
53. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Sequence to sequence-video to text. In: IEEE International Conference on Computer Vision, pp. 4534–4542. IEEE (2015)
54. Vukotić, V., Raymond, C., Gravier, G.: Bidirectional joint representation learning with symmetrical deep neural networks for multimodal and crossmodal applications. In: ACM International Conference on Multimedia Retrieval, pp. 343–346. ACM (2016)
55. Vukotić, V., Raymond, C., Gravier, G.: Generative adversarial networks for multimodal representation learning in video hyperlinking. In: ACM International Conference on Multimedia Retrieval, pp. 416–419. ACM (2017)
56. Vukotić, V., Raymond, C., Gravier, G.: A crossmodal approach to multimodal fusion in video hyperlinking. IEEE MultiMedia **25**(2), 11–23 (2018)
57. Wang, B., Yang, Y., Xu, X., Hanjalic, A., Shen, H.T.: Adversarial cross-modal retrieval. In: ACM Multimedia Conference, pp. 154–162. ACM (2017)
58. Wang, L., Li, Y., Huang, J., Lazebnik, S.: Learning two-branch neural networks for image-text matching tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence (2018)
59. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5005–5013. IEEE (2016)
60. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5288–5296 (2016)
61. Xu, R., Xiong, C., Chen, W., Corso, J.J.: Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In: AAAI, vol. 5, p. 6 (2015)
62. Yan, F., Mikolajczyk, K.: Deep correlation for matching images and text. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3441–3450. IEEE (2015)
63. Yan, R., Yang, J., Hauptmann, A.G.: Learning query-class dependent weights in automatic video retrieval. In: ACM Multimedia Conference, pp. 548–555. ACM (2004)
64. Zhang, L., Ma, B., Li, G., Huang, Q., Tian, Q.: Multi-networks joint learning for large-scale cross-modal retrieval. In: ACM Multimedia Conference, pp. 907–915. ACM (2017)
65. Zhang, X., Gao, K., Zhang, Y., Zhang, D., Li, J., Tian, Q.: Task-driven dynamic fusion: Reducing ambiguity in video description. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3713–3721. IEEE (2017)
66. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)