Learning Joint Embedding with Multimodal Cues for Cross-Modal Video-Text Retrieval

Niluthpol Chowdhury Mithun
University of California, Riverside, CA
nmithun@ece.ucr.edu

Juncheng Li
Carnegie Mellon University, PA
Bosch Research and Technology Center, PA
junchenl@cs.cmu.edu

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

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

ABSTRACT

Constructing a joint representation invariant across different modalities (e.g., video, language) is of significant importance in many multimedia applications. While there are a number of recent successes in developing effective image-text retrieval methods by learning joint representations, the video-text retrieval task, however, has not been explored to its fullest extent. In this paper, we study how to effectively utilize available multimodal cues from videos for the cross-modal video-text retrieval task. Based on our analysis, we propose a novel framework that simultaneously utilizes multi-modal features (different visual characteristics, audio inputs, and text) by a fusion strategy for efficient retrieval. Furthermore, we explore several loss functions in training the embedding and propose a modified pairwise ranking loss for the task. Experiments on MSVD and MSR-VTT datasets demonstrate that our method achieves significant performance gain compared to the state-of-the-art approaches.

CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**; *Learning to rank*;

KEYWORDS

Video-Text Retrieval, Joint Embedding, Multimodal Cues.

ACM Reference Format:
Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K. Roy-Chowdhury. 2018. Learning Joint Embedding with Multimodal Cues for Cross-Modal Video-Text Retrieval. In *ICMR '18: 2018 International Conference on Multimedia Retrieval, June 11–14, 2018, Yokohama, Japan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3206025.3206064

1 INTRODUCTION

Cross-modal retrieval between visual data and natural language description remains a long-standing challenge in multimedia [12, 41]. Joint visual-semantic embeddings [9, 15, 20, 26, 35] underpin the building of most cross-modal retrieval methods as they can bridge the gap between different modalities. In this joint space, the similarity of different points reflects the semantic closeness between their corresponding original inputs. In this work, we focus on learning effective joint embedding models for the cross-modal video-text retrieval task (See Fig. 1).

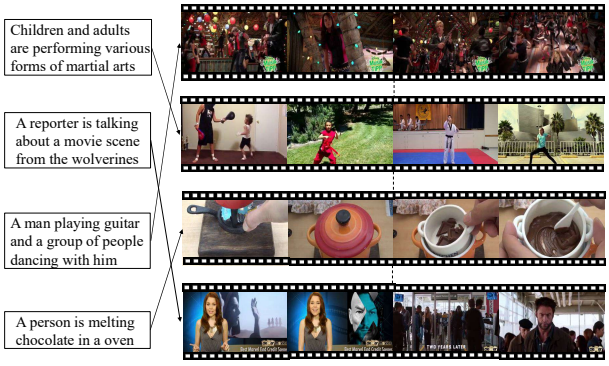

Figure 1: Illustration of Cross-Modal Video-Text retrieval task: given a text query, retrieve and rank videos from the database based on how well they depict the text or vice versa.

Recently, a few methods have been proposed for learning visual-semantic embeddings for the video-language matching task [6, 24, 25, 30, 37]. Most of these existing approaches are very similar to the image-text retrieval methods by design and focus mainly on the loss functions. We observe that simple adaptation of a state-of-the-art image-text embedding method [7] by mean-pooling features from video frames provides a better result than most existing video-text retrieval approaches [6, 24]. Image-text retrieval is a relatively mature field of study, and one might think that directly applying such techniques to video-text retrieval may give optimal performance. However, such methods fail to take advantage of the supplementary information such as the temporal dynamics and sounds, which are already included in the videos. This limits the robustness of the systems; for instance, it may be 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. Associating video motion content and sound 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 both visual and audio modalities simultaneously.

ICMR '18, June 11–14, 2018, Yokohama, Japan

ACM ISBN 978-1-4503-5046-4/18/06…$15.00
https://doi.org/10.1145/3206025.3206064

Developing a practical system for video-text retrieval without considering most available cues in the video content is unlikely to be comprehensive. However, an inappropriate fusion of complementary features may increase ambiguity and degrade performance. In this regard, we study how to judiciously utilize different cues from videos to develop a successful video-text retrieval system. We propose a novel framework, which is tailored towards achieving high performance in the task of cross-modal video-text retrieval. Compared to the existing methods, our framework fuses four types of feature (object, action, text and, audio) for efficient retrieval. Furthermore, we propose a modified pairwise ranking loss for the retrieval task that emphasizes on hard negatives and relative ranking of positive labels. Our approach shows significant performance improvement compared to previous approaches and baselines.

1.1 Overview of the Proposed Approach

In the cross-modal video-text retrieval task, 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. Utilizing multiple characteristics of video (e.g., objects, actions, place, time) is evidently crucial for efficient retrieval [39] and has become a common practice. In the closely related task of video captioning, dynamic information from video along with static appearance features has been shown to be effective [27, 42]. Methods have been developed to extract features from videos that focus on different characteristics. For example, ResNet feature focuses on identifying objects in the frames, whereas I3D feature focuses on identifying the activities. Since in this work we are looking at videos in general, detecting both objects and activities from the video is very important for higher performance. Therefore, we need to develop a strategy that fuses information from different video features efficiently for the target task.

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). We propose to learn two joint video-text embedding networks as shown in Fig. 2. One model learns a joint space (Object-Text Space in Fig. 2) between text features and visual appearance features. Another model learns a joint space (Activity-Text Space in Fig. 2) between text feature and a combination of activity and audio features. Here, Object-Text space is the expert in solving ambiguity between objects in a video, whereas Activity-Text space is the expert in solving ambiguity between actions/events 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 both Object-Text and Activity-Text embedding spaces and use the sum of similarity scores for final ranking.

We follow network architecture proposed in [18] that learns the embedding model using a two-branch network using image-text pair. 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 [7] 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.

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, text and audio features by a fusion strategy for efficient retrieval. We also present a modified pairwise loss to better learn the joint embedding.
- We demonstrate a clear improvement over the state-of-the-art methods in the video to text retrieval tasks with the MSR-VTT dataset [36] and MSVD dataset [4].

2 RELATED WORK

Image-Text Retrieval. Recently, there has been significant interest in learning robust visual-semantic embeddings for image-text retrieval [12, 16]. 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 [8]. Canonical Correlation Analysis (CCA) was used in several previous works for learning joint embeddings for the retrieval task [10, 13, 29, 38]. However, CCA based methods have been reported to be unstable and incur a high memory cost due to the covariance matrix calculation [34].

Most recent works relating to text and image modality are trained with ranking loss [7, 9, 18, 23, 32, 34]. In [9], 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 [18] 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 [23], similarity score calculation [34], or loss function [7]. VSEPP model [7] 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 [14] that recurrently selects pairwise instances from image and sentence descriptions, and aggregates local similarity. In [23], authors proposed a multimodal attention mechanism to attend to sentence fragments and image regions selectively for similarity calculation. Our method complements these works that learns joint image-text embedding using a ranking loss (e.g., [7, 18, 32]). The framework can be applied to most of these approaches for improved video-text retrieval.

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 [37], authors vectorize each subject-verb-object triplet extracted from a given sentence by word2vec model [22] 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

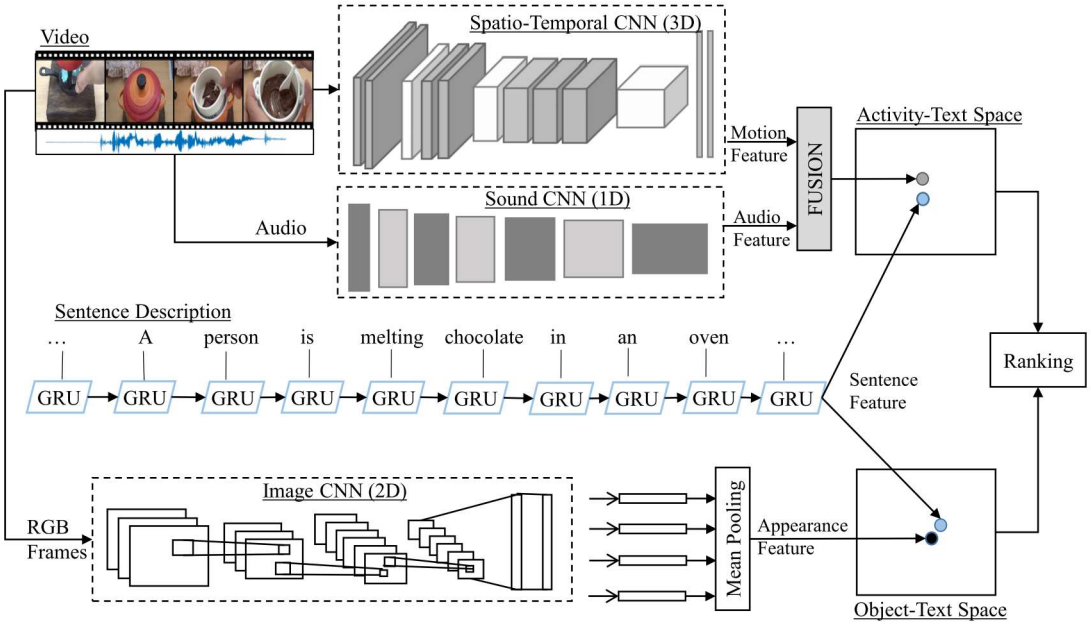

Figure 2: An overview of the proposed retrieval process. Please see Section 1.1 for an overview and Section 3 for details.

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 [24], which focused on word disambiguation. In [33], a stacked GRU is utilized to associate sequence of video frames to a sequence of words. In [25], 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 [6] 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 [2] utilizing ranking loss, which can also be applied to video-text retrieval task.

Most of these video-text retrieval approaches are very similar to the image-text retrieval methods by design and fail to utilize video dynamics for retrieval. In contrast to the existing works, our approach is capable of utilizing different visual cues and audio (if available) concurrently for more efficient retrieval.

3 APPROACH

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

3.1 Input Feature Representation

Text Feature. For encoding sentences, we use Gated Recurrent Units (GRU) [5]. We set the dimensionality of the joint embedding space, D, to 1024. The dimensionality 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 [11]. 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 [3]. 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.

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 [1], which provides 1024 dimensional feature from audio. Note that, we only utilize the audio sound which is readily available with the videos.

3.2 Learning Joint Embedding

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

Given a video feature representation (e.g., appearance feature, or a combination of action and audio features) $\overline{v}$ ($\overline{v} \in \mathbb{R}^V$), 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 and D is the dimensionality 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 $\bar{t}$ 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.

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 learn to maximize the similarity between a video embedding and corresponding text embedding and minimize similarity to all other non-matching ones. The optimization problem can be written as,

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

where, $[f]_+ = max(0, f)$. t^- is a non-matching text embedding, and t is the matching text embedding. This is the same for video embedding v. 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 as it is easy to compute and shown to be very effective in learning joint embedding [7, 18].

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.

3.3 Proposed Ranking Loss

Recently, focusing on hard-negatives has been shown to be effective in many embedding tasks [7, 21, 28]. Inspired by this, we focus on hard negatives (i.e., the negative video/text sample closest to a positive/matching (v, t) pair) instead of summing over all negatives in our formulation. For a positive 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 written as following to focus on hard-negatives,

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

The loss in Eq. 2 is similar to the loss in Eq. 1 but it is specified in terms of the hardest negatives [7]. We start with the loss function in Eq. 2 and modify the loss function following the idea of weighted ranking [31] to weight the loss based on the relative ranking of positive labels.

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

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 $L(r) = (1+1/(N-r+1))$, where N is the number of compared videos.

It is very common, in practice, to only compare samples within a mini-batch at each iteration rather than comparing the whole training set for computational efficiency [15, 21, 28]. This is known as semi-hard negative mining [21, 28]. Moreover, selecting the hardest negatives in practice may often lead to a collapsed model and semi-hard negative mining helps to mitigate this issue [21, 28]. 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 ultimately try to push the positive matching pair to the top of rank.

3.4 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 two main aspects of each video: (1) the salient objects of the video and (2) the action and events in the video. To achieve this, we learn two joint video-text embedding spaces as shown in Fig. 2.

The Object-Text embedding space is the common space where both appearance features and text 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 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 a video, we only use the action feature as the video representation for learning the Activity-Text embeddings. We utilize the same loss functions described in Sec. 3.3 for training both the Object-Text and Activity-Text 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 both Object-Text and Activity-Text embedding spaces and use a sum 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 both embedding spaces and use a sum of similarity scores for final ranking. 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). In this work, we put equal importance to both facets in ranking.

4 EXPERIMENTS

4.1 Datasets and Evaluation Metric

We present experiments on two benchmark datasets: Microsoft Research Video to Text (MSR-VTT) Dataset [36] and Microsoft Video Description dataset (MSVD) [4] to evaluate the performance of our

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*	*Recall*	*R@1*	*R@5*	*R@10*	*MedR*	*MeanR*	*Recall*
3.1	VSE-ResNet	7.7	20.3	31.2	28.0	185.8	19.7	5.0	16.4	24.6	47.0	215.1	15.3
	VSEPP ResNet	10.2	25.4	35.1	25.0	228.1	23.5	5.7	17.1	24.8	65.0	300.8	15.8
3.2	ResNet	10.5	26.7	35.9	25.0	266.6	24.4	5.8	17.6	25.2	61.0	296.6	16.2
	Audio	0.4	1.1	1.9	1051	2634.9	1.1	0.2	0.9	1.5	1292	1300	0.8
	I3D	8.4	22.2	32.3	30.3	229.9	21.0	4.6	15.3	22.7	71.0	303.7	14.2
3.3	CON(ResNet-I3D)	9.1	24.6	36.0	23.0	181.4	23.2	5.5	17.6	25.9	51.0	243.4	16.3
	CON(ResNet-I3D-Audio)	9.3	27.8	38.0	22.0	162.3	25.0	5.7	18.4	26.8	48.0	242.5	16.9
3.4	Joint Image-Text-Audio Embedding	8.7	22.4	32.1	31.0	225.8	21.0	4.8	15.3	22.9	73.0	313.6	14.3
3.5	Fusion [ResNet & I3D]	12.3	31.3	42.9	16.0	145.4	28.9	6.8	20.7	29.5	39.0	224.7	19.0
	Fusion [ResNet & CON(I3d-Audio)]	**12.5**	**32.1**	**42.4**	**16.0**	**134.0**	**29.0**	**7.0**	**20.9**	**29.7**	**38.0**	**213.8**	**19.2**

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 around 40 sentences. We use only the English descriptions. For a fair comparison, we used the same splits utilized in prior works [33], with 1200 videos for training, 100 videos for validation, and 670 videos for testing. The MSVD dataset is also used in [24] 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 [6, 18, 24]. 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. We also report the average of *R*@1, *R*@5 and *R*@10 as Recall in the tables.

4.2 Training Details

We used two GTX 1080 Ti GPUs for this work. We implemented the network using PyTorch following [7]. 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 [17]. 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 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 finally.

4.3 Results on MSR-VTT Dataset

We report the result on MSR-VTT dataset [36] in Table 1. We implement several baselines to analyze different components of the proposed approach. To understand the effect of different loss functions, features, effect of feature concatenation and proposed fusion method, we divide the table into 5 rows (1.1-1.5). In row-1.1, we report the results on applying two different variants of pair-wise ranking loss. VSE[18] is based on the basic triplet ranking loss similar to Eq. 1 and VSEPP[7] 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 is 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 embedding learned utilizing video feature 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 [2] and used for video-text retrieval task. Finally, in row-1.5, we provide the result based on the proposed approach that employs two video-text space for retrieval.

Loss Function. For evaluating the performance of different ranking loss in the task, we can compare results reported in row-1.1 and row-1.2. We can choose only ResNet feature based method 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, and recall. However, VSE loss function has better performance in terms of 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.

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 ResNet feature and I3D feature performs reasonably well. 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. However, utilizing audio along with other feature provides a boost in performance as shown in row-1.3 and row-1.4.

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*	*Recall*
Results Using Partition used by JMET and JMDV						
CCA					245.3	
JMET					208.5	
JMDV					224.1	
W2VV-ResNet	16.3		44.8	14.0	110.2	
VSE-ResNet	15.8	30.2	41.4	12.0	84.8	34.5
VSEPP-ResNet	21.2	43.4	52.2	9.0	79.2	39.0
Ours-ResNet	23.4	45.4	53.0	8.0	75.9	40.6
Ours-I3D	21.3	43.7	53.3	9.0	72.2	39.5
Ours-(ResNet-I3d Fusion)	**31.5**	**51.0**	**61.5**	**5.0**	**41.7**	**48.0**
Results Using Partition used by LJRV						
ST	2.99	10.9	17.5	77.0	241.0	10.5
LJRV	9.85	27.1	38.4	19.0	75.2	25.1
W2VV-ResNet	17.9		49.4	11.0	57.6	
Ours-ResNet	20.9	43.7	54.9	7.0	56.1	39.9
Ours-I3D	17.5	39.6	51.3	10.0	54.8	36.1
Ours-(ResNet-I3d Fusion)	**25.5**	**51.3**	**61.9**	**5.0**	**32.5**	**46.3**

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 [6, 36]. 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 ResNet feature, I3D feature and Audio feature performs even worse than utilizing only ResNet 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.

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 [2]. Following the idea, we trained a shared space across image-text-sound modality using the pairwise ranking loss by utilizing image-text and image-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.

Fusion. The best result in Table. 1 is achieved by our proposed fusion approach as shown in row-1.5. We see that the proposed method achieves 19.5% improvement in *R*@1 for text retrieval and 20.68% improvement for video retrieval in *R*@1 compared to ResNet(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[Resnet & CON (I3D-Audio)] differs from Fusion[ResNet & I3d] 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 degrade 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*	*Recall*
Results Using Partition used by JMET and JMDV						
CCA					251.3	
JMDV					236.3	
VSE-ResNet	12.3	30.1	42.3	14.0	57.7	30.2
VSEPP-ResNet	15.4	39.6	53.0	9.0	43.8	36.0
Ours-ResNet	16.1	41.1	53.5	9.0	42.7	36.9
Ours-I3D	15.4	39.2	51.4	10.0	43.2	35.3
Ours-(ResNet-I3d Fusion)	**20.3**	**47.8**	**61.1**	**6.0**	**28.3**	**43.1**
Results Using Partition used by LJRV						
ST	2.6	11.6	19.3	51.0	106.0	11.2
LJRV	7.7	23.4	35.0	21.0	49.1	22.0
Ours-ResNet	15.0	40.2	51.9	9.0	45.3	35.7
Ours-I3D	14.6	38.9	51.0	10.0	45.1	34.8
Ours-(ResNet-I3d Fusion)	**20.2**	**47.5**	**60.7**	**6.0**	**29.0**	**42.8**

4.4 Results on MSVD Dataset

We report the results of Video-to-Text retrieval task on MSVD dataset [4] 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[29], ST [19], JMDV [37], LJRV [24], JMET [25], and W2VV [6]. For these approaches, we directly cite scores from respective papers when available. We report score for JMET from [6]. The score of CCA is reported from [37] and the score of ST is reported from [24]. If the score for multiple models is reported, we select the score of the best performing method from the paper. We could not compare with method [40] which focuses on movie retrieval task and results are not available on our experimented datasets. We did not re-implement this method in our setting as their method is based on an ensemble of several models and it is very difficult to exactly emulate the implementation details.

We also implement and compare results with state-of-the-art image-embedding approach VSE[18] and VSEPP[7] using ResNet152 feature as the video feature following publicly available code [7]. Our proposed fusion method is named as Ours-(ResNet-I3D Fusion) in the Table. 2 and Table. 3. Our method utilizes the proposed loss and employs two embedding spaces for calculating similarity between video and text. We use ResNet-152 feature as the appearance feature in training Object-Text space. As the audio is muted in this dataset, we train the Activity-Text space utilizing only I3D feature from videos. To show the impact of only using the proposed loss in retrieval, we also report results based on each of these two embedding space (Ours-ResNet and Ours-I3D) in the tables.

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 both the 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.

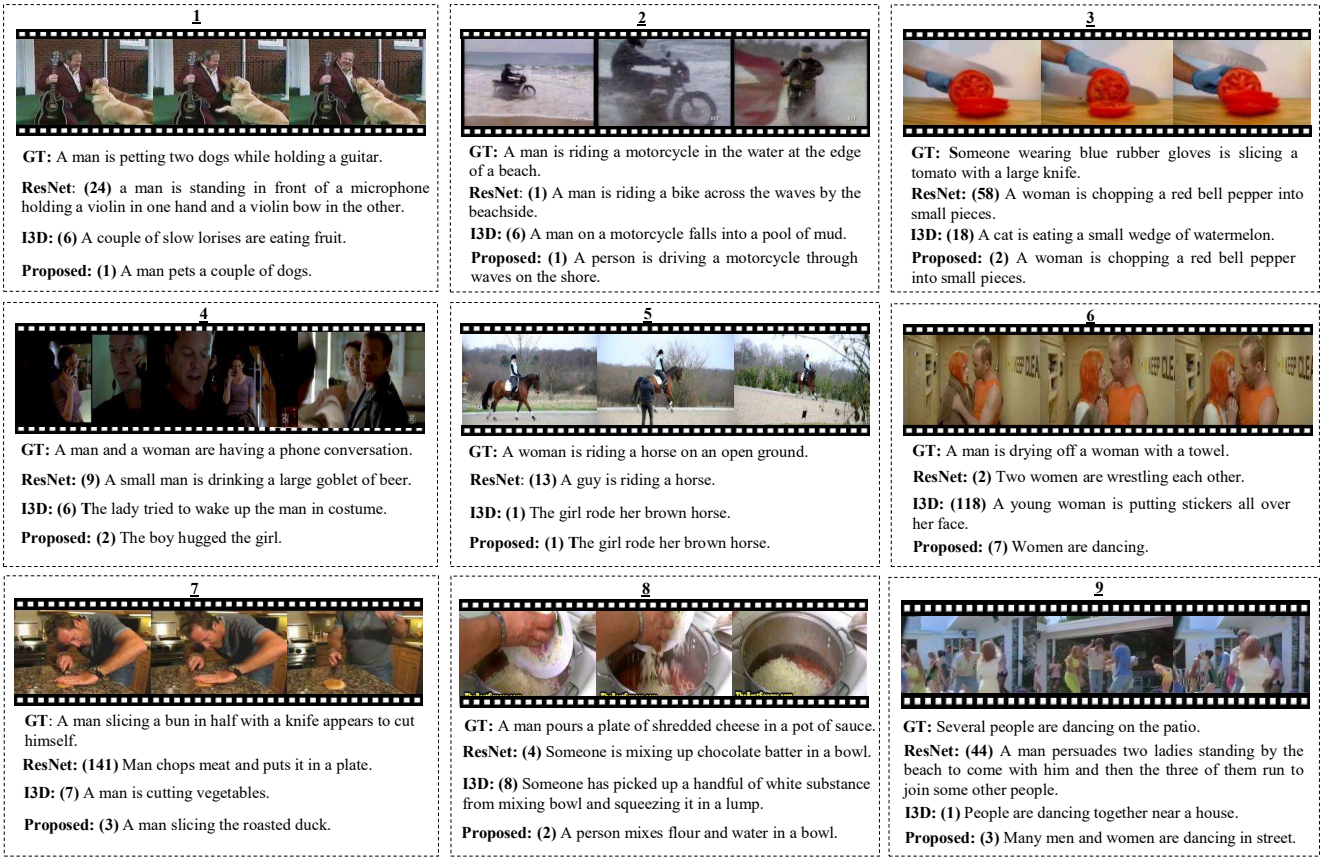

Figure 3: Examples of 9 test videos from MSVD dataset and the top 1 retrieved captions for Ours-ResNet, Ours-I3D, and the proposed method as shown in Table. 2. The value in brackets is the rank of the highest ranked ground-truth caption. Ground Truth (GT) is a sample from the ground-truth captions. Among the approaches, ResNet and I3D are methods where single video-text space is used for retrieval (ResNet is trained using ResNet feature as video feature and I3D is trained using the I3D feature). We also report result for the proposed fusion approaches where both video-text spaces are used for retrieval.

4.5 Qualitative Results

MSVD Dataset. In Fig. 3, we show examples of 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 both video-text spaces helps to match the correct caption. It can be easily identified from the top-1 retrieved caption that the projection learned between video and text by utilizing appearance feature (ResNet) is significantly different from that learned using Activity feature (I3D). The variation between the rank of the highest matching caption further strengthens this observation. We also cannot claim that one video feature is better than others for this task. ResNet feature performs better than the I3D feature in retrieval for some videos. For other videos, the I3D feature achieves higher performance. However, combining knowledge from two video-text spaces, we have consistently better performance than utilizing one of the features. We see from Fig. 3 that, among 9 videos, the retrieval performance is improved or higher recall is retained for 7 videos. Video-6 and video-9 show two of our failure cases, where utilizing multiple video-text spaces degrades the performance slightly than ResNet in Video-6 and I3D in Video-9.

MSR-VTT Dataset. Similar to Fig. 3, we also show qualitative results for a few test videos from MSR-VTT dataset in Fig. 4. Video 1-6 in Fig. 4 shows a few examples where utilizing cue from both 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. 4. 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. 4, 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 improving the result. Since we did not exploit the structure of the audio and analyze the structural

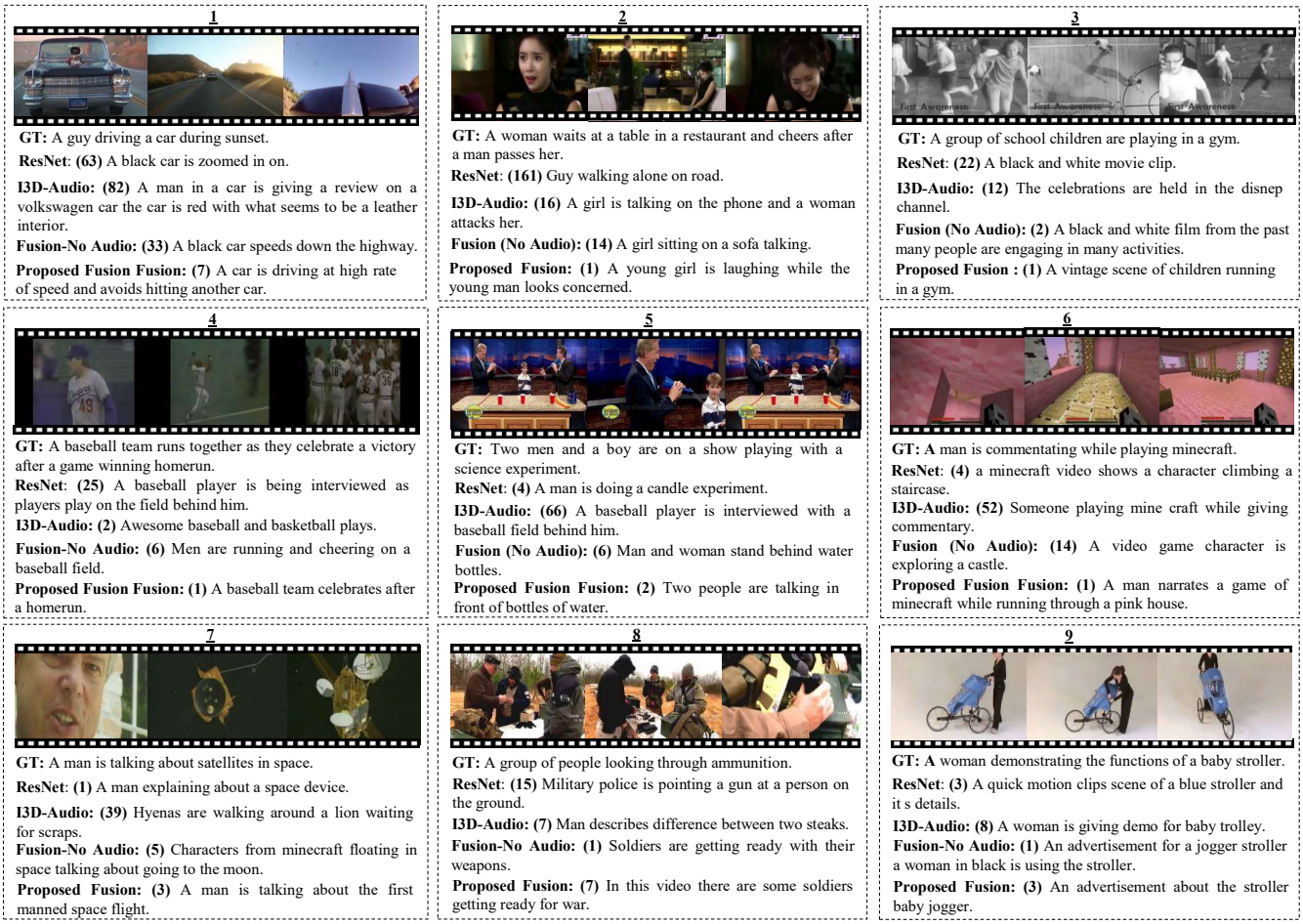

Figure 4: A snapshot of 9 test videos from MSR-VTT dataset with success and failure cases, the top 1 retrieved captions for four approaches based on the proposed loss function and the rank of the highest ranked ground-truth caption inside the bracket. Among the approaches, ResNet and I3D-Audio are methods where single video-text space is used for retrieval (ResNet is trained using ResNet feature as video feature and I3D-Audio is trained using the concatenated I3D feature and Audio feature as video feature). We also report results for two fusion approaches where two video-text spaces are used for retrieval. Both fusion approaches use a object-text space trained with ResNet feature, while in our proposed approach, the activity-text space is trained using audio along with I3D feature. Fusion (No Audio) uses activity-text space that is trained with only I3D feature.

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 the visual aspects. We observe 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). We see an overall improvement in the quantitative result (Table 1) which also supports our idea of using audio. These failure cases provide future directions of this work focusing on developing more sophisticated algorithms to combine similarity scores from multiple joint spaces and further analyze the role of associated audio for video-text retrieval.

5 CONCLUSIONS

We propose an approach for efficient cross-modal video-text retrieval using joint embeddings by effectively including features from different visual entities and features from audio. We also propose a new loss function that further exploits the multimodal correlation. Experiments on two benchmark datasets demonstrate that our proposed loss function learns better embeddings for the video-text retrieval task than existing ones. Moreover, our overall framework achieves significant performance improvement compared to several state-of-the-art approaches.

Acknowledgement. This work was partially supported by NSF grants IIS-1746031 and CNS-1544969. Juncheng Li was supported by the Bosch Graduate Fellowship to CMU LTI. We thank the anonymous reviewers for insightful suggestions. We would also like to thank Rameswar Panda and Sujoy Paul for helpful comments.

REFERENCES

[1] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*. 892–900.

[2] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2017. See, Hear, and Read: Deep Aligned Representations. *arXiv preprint arXiv:1706.00932* (2017).

[3] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 4724–4733.

[4] David L Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. ACL, 190–200.

[5] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[6] Jianfeng Dong, Xirong Li, and Cees GM Snoek. 2016. Word2VisualVec: Image and Video to Sentence Matching by Visual Feature Prediction. *arXiv preprint arXiv:1604.06838* (2016).

[7] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. VSE++: Improved Visual-Semantic Embeddings. *arXiv preprint arXiv:1707.05612* (2017).

[8] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *European Conference on Computer Vision*. Springer, 15–29.

[9] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. 2013. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*. 2121–2129.

[10] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. 2014. A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision* 106, 2 (2014), 210–233.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 770–778.

[12] Christian Andreas Henning and Ralph Ewerth. 2017. Estimating the information gap between textual and visual representations. In *International Conference on Multimedia Retrieval*. ACM, 14–22.

[13] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research* 47 (2013), 853–899.

[14] Yan Huang, Wei Wang, and Liang Wang. 2017. Instance-aware image and sentence matching with selective multimodal lstm. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2310–2318.

[15] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3128–3137.

[16] Andrej Karpathy, Armand Joulin, and Fei Fei F Li. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in Neural Information Processing Systems*. 1889–1897.

[17] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[18] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539* (2014).

[19] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*. 3294–3302.

[20] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. 2015. Associating neural word embeddings with deep image representations using fisher vectors. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 4437–4446.

[21] R Manmatha, Chao-Yuan Wu, Alexander J Smola, and Philipp Krahenbuhl. 2017. Sampling Matters in Deep Embedding Learning. In *2017 IEEE International Conference on Computer Vision*. IEEE, 2859–2867.

[22] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[23] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual Attention Networks for Multimodal Reasoning and Matching. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 299–307.

[24] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. 2016. Learning joint representations of videos and sentences with web image search. In *European Conference on Computer Vision*. Springer, 651–667.

[25] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. 2016. Jointly modeling embedding and translation to bridge video and language. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 4594–4602.

[26] Bryan Plummer, Matthew Brown, and Svetlana Lazebnik. 2017. Enhancing Video Summarization via Vision-Language Embedding. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

[27] Vasili Ramanishka, Abir Das, Dong Huk Park, Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, and Kate Saenko. 2016. Multimodal video description. In *ACM Multimedia Conference*. ACM, 1092–1096.

[28] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 815–823.

[29] Richard Socher and Li Fei-Fei. 2010. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 966–973.

[30] Atousa Torabi, Niket Tandon, and Leonid Sigal. 2016. Learning language-visual embedding for movie understanding with natural-language. *arXiv preprint arXiv:1609.08124* (2016).

[31] Nicolas Usunier, David Buffoni, and Patrick Gallinari. 2009. Ranking with ordered weighted pairwise classification. In *International Conference on Machine Learning*. ACM, 1057–1064.

[32] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2016. Order-embeddings of images and language. In *International Conference on Learning Representations*.

[33] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *IEEE International Conference on Computer Vision*. IEEE, 4534–4542.

[34] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. 2018. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018).

[35] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 5005–5013.

[36] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5288–5296.

[37] Ran Xu, Caiming Xiong, Wei Chen, and Jason J Corso. 2015. Jointly Modeling Deep Video and Compositional Text to Bridge Vision and Language in a Unified Framework.. In *AAAI*, Vol. 5. 6.

[38] Fei Yan and Krystian Mikolajczyk. 2015. Deep correlation for matching images and text. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3441–3450.

[39] Rong Yan, Jun Yang, and Alexander G Hauptmann. 2004. Learning query-class dependent weights in automatic video retrieval. In *ACM Multimedia Conference*. ACM, 548–555.

[40] Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. 2017. End-To-End Concept Word Detection for Video Captioning, Retrieval, and Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3165–3173.

[41] Liang Zhang, Bingpeng Ma, Guorong Li, Qingming Huang, and Qi Tian. 2017. Multi-Networks Joint Learning for Large-Scale Cross-Modal Retrieval. In *ACM Multimedia Conference*. ACM, 907–915.

[42] Xishan Zhang, Ke Gao, Yongdong Zhang, Dongming Zhang, Jintao Li, and Qi Tian. 2017. Task-Driven Dynamic Fusion: Reducing Ambiguity in Video Description. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3713–3721.