electronics

MDPI

Article

Cross-Modal Learning Based on Semantic Correlation and Multi-Task Learning for Text-Video Retrieval

Xiaoyu Wu [1,*], Tiantian Wang [1] and Shengjin Wang [2]

[1] School of Information and Communication Engineering, Communication University of China, Beijing 100024, China; tiantian_wong@cuc.edu.cn

[2] Department of Electronic Engineering, Tsinghua University, Beijing 100084, China; wgsgj@tsinghua.edu.cn

* Correspondence: wuxiaoyu@cuc.edu.cn

Received: 19 November 2020; Accepted: 9 December 2020; Published: 11 December 2020

Abstract: Text-video retrieval tasks face a great challenge in the semantic gap between cross modal information. Some existing methods transform the text or video into the same subspace to measure their similarity. However, this kind of method does not consider adding a semantic consistency constraint when associating the two modalities of semantic encoding, and the associated result is poor. In this paper, we propose a multi-modal retrieval algorithm based on semantic association and multi-task learning. Firstly, the multi-level features of video or text are extracted based on multiple deep learning networks, so that the information of the two modalities can be fully encoded. Then, in the public feature space where the two modalities information are mapped together, we propose a semantic similarity measurement and semantic consistency classification based on text-video features for a multi-task learning framework. With the semantic consistency classification task, the learning of semantic association task is restrained. So multi-task learning guides the better feature mapping of two modalities and optimizes the construction of unified feature subspace. Finally, the experimental results of our proposed algorithm on the Microsoft Video Description dataset (MSVD) and MSR-Video to Text (MSR-VTT) are better than the existing research, which prove that our algorithm can improve the performance of cross-modal retrieval.

Keywords: cross-model learning; text-video retrieval; semantic correlation; multi-task learning

1. Introduction

In today's era of the increasing scale of information and more diversified information forms, video media websites such as YouTube are developing rapidly, while TikTok and other short video applications are also popular with people. These platforms have a large number of videos or short videos. However, retrieving the content according to users' requirements from tens of millions of video data involves cross-modal video retrieval task. Specifically, a user can type a descriptive sentence in the search bar of a video website, such as "a baby is crying in the room", and then the results which meet the description of the sentence are retrieved. This is a kind of text-video retrieval task that has broad application prospects in intelligent editing, intelligent review and recommendation of media content. Figure 1 illustrates the text-video task more comprehensively and intuitively. As shown, a text query is used as an input to retrieve the corresponding video in a large dataset, and then, according to the ranking, related videos are returned to get the output results. The returned ranking results of task are sorted according to similarity score, that is, the first item in the result most closely match the query, the second item in the result partially match the query description, but not completely consistent, and so on. According to the example in Figure 2, the query is "a person is driving a car down the road", and its results with similarity rank can be seen. The top-1 result completely meets the description and realizes the best match. The top-2 shows a car that is driving on the lawn rather than on the road, so its

similarity is weaker than the former. However, the top-3 item is actually a news report about parking problem and the latter of this video is an indoor broadcast. Without doubt, its similarity score is worse.

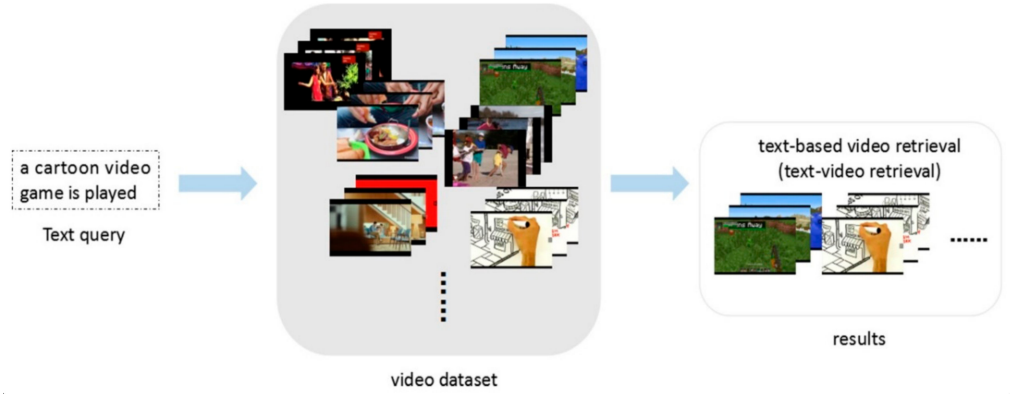


Figure 1. Illustration of text-based video retrieval task. On the left, the text is taken as an input query to retrieve the corresponding video in order.

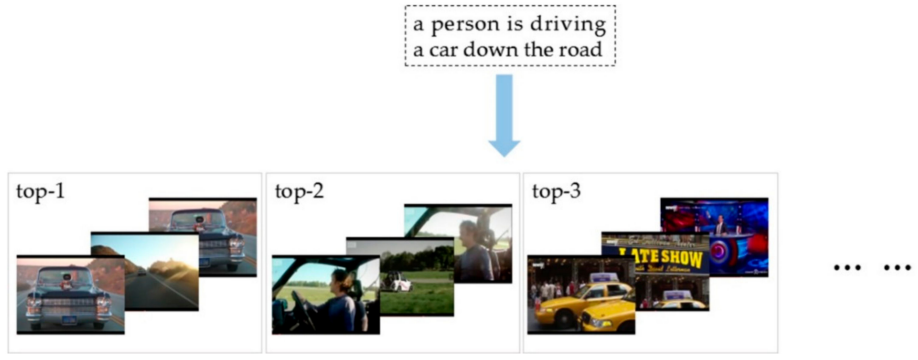


Figure 2. The video clips are ranked by similarity score for a query.

Here, as it is known, text and video are heterogeneous information; they cannot directly find their nearest neighbor as easily as single modality task. In order to solve this problem, traditional concept-based methods [1–5] transform the cross-model retrieval task into a series of detection tasks. As current researches have made some achievements in object detection domain, many mature deep neural network (DNN) models can be used. All of these kinds of methods select information from the input text query as target entities. Then, the pre-trained classifiers like Faster R-CNN [6] and Mask R-CNN [7] are used to detect these concepts in video content. By fusing all scores, the task of video retrieval is completed. However, concept-based methods meet several difficulties. As a natural language, the text query has semantic complexity, so when we specify some concepts from a whole text query, it is crucial to select relevant and detectable concepts more accurately [8]. Simply choosing words or phrases as targets may lose rich semantic information and cannot extract the implicit concept well, so that it leads to the decline of retrieval accuracy. For example, a query "a band is performing in a small club", the word "performing" in this sentence more likely means singing songs, playing the guitar or playing the drum while with small probability means dancing. So, if we choose "performing" as an action concept to detect in video dataset, the result cannot well match this query.

Since it is hard to exhaustively express the semantic information in the queries by concepts, the word embedding technique is utilized to integrate with the visual features and map them to a common space as a "bridge" for comparing the similarity between the text and visual data [9]. This kind of method is concept-free, and makes corresponding video retrieval by using the whole text query [10–12]. These methods use a variety of neural networks to fully encode video in spatial and temporal dimensions to represent its content. For the sentence, they use the method of natural language processing. The text query is vectorized by word embedding method and neural network [12,13].

After getting the cross-modality encoding, it is necessary to establish connections between different modalities. So, features are projected into a common learning space by transformation to measure the similarity of text feature and video feature [12]. Recently, with big advances of deep learning in natural language processing and computer vision research, these kinds of methods based on similarity measures have an increasing use for video retrieval. However, due to the complexity of video and text, using only a single method cannot encode their features very well. Therefore, some researches combine different encoding strategies to form multi-scale structure and get powerful dense representations of text and video [10,14,15]. However, studies on this aspect are still limited. These existing algorithms are not accurate enough to extract features from text or video. In addition, all the methods based on similarity lack cross modal information constraint and the absolute distance constraint between the sample pairs.

Considering the problems above, this paper proposes a multi-level and multi-task learning based on semantic association to deal with text-video retrieval. Our framework uses the dual multi-level encoding structure [10]. It extracts the multi-level semantic features considering context relevance from the text and video input respectively. Then, it constructs the shared feature subspace with semantic maintenance. By feature mapping, we can transform the two model representations into the same dimension, which makes it easier to compare and measure the cross-modal similarity. In addition to similarity measurement, the multi-task learning model is formed to align the features of different modalities by combining semantic consistency hard classification task. Figure 3 illustrates the framework of our method. In this paper, experiments are carried out on the existing public datasets Microsoft Video Description dataset (MSVD) and MSR-Video to Text (MSR-VTT), and all the results show the effectiveness of the proposed algorithm.

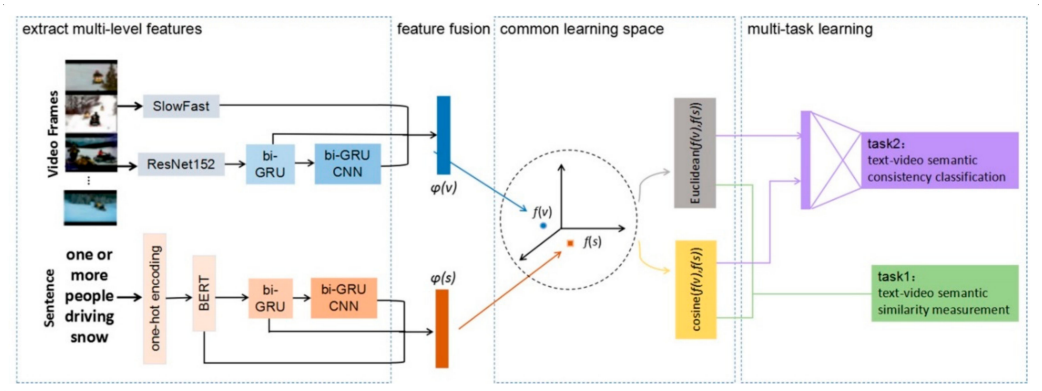


Figure 3. The framework of our method in this paper. By using dual multi-level encoder, we fully learn the multi-modal features, and then map them into the common learning subspace, and the cross-modal similarity is calculated to complete the retrieval task. At the same time, the text-video sample pairs are classified through the full connection layer to form a multi-task learning model.

The contributions of our paper can be summarized as follows:

- We propose a novel multi-task learning model that combines the original text-video retrieval task and a classification task. By exploiting the classification auxiliary task with semantic alignment consistency and constructing the constraint between text and video modalities, we achieve the semantic association between different modality and improve the retrieval performance through the jointly learning;
- A new loss function that restricts relative distance and absolute distance simultaneously is presented. The new triplet loss based on the hardest negative sample [16] and the absolute distance between different modality of any sample, that is, while correctly distinguishing the positive sample pair from the negative sample pair with the minimum cross-modal distance, we adjust the distance to fully consider the between-class distance of different modes in the

common subspace. The experimental results demonstrate our methods achieve competitive performances on two widely adopted datasets;

- In addition, dual multi-level feature representations for text and video are improved in contrast to the reference [10]. Based on our task, the slightly modified SlowFast [17] model is utilized to extract accurate video features in the spatial domain, and the BERT [18] model is used to embed the high-level text semantic embedding in sentences rather in words.

2. Related Work

For the text-video retrieval, international competition Text Retrieval Conference Video Retrieval Evaluation (TRECVid), led by the National Institute of Standards and Technology (NIST), proposed a new challenge named Ad-hoc Video Search (AVS) in 2016, which mainly focused on text-to-video retrieval [19]. Therefore, in this section, we briefly review the existing multi-model retrieval task methods and related achievements of the AVS competition.

For the task of text to visual retrieval, most of the traditional methods are concept-based approaches, and it is easy to observe that most of the top performers on AVS are concept-based. The methods that get well performance in AVS, such as [9,20–23], design relatively complex language rules according to the query text, and extract relevant concepts from a given query. For example, Nguyen et al. [21] set concept into three categories which includes: object (man, woman, cat, dog, etc.), scene (indoor/outdoor, park, kitchen etc.), action (lying on, near, etc.), so as to establish a large concept corpus. For each class, the appropriate model and dataset are used to train the corresponding classifiers. For instance, Ueki et al. [22] establish a concept bank consists of more than 50 k concepts, and, in addition to the pre-trained CNN model, they train a support vector machine (SVM) classifier to annotate video content automatically. Tao et al. [9] propose the just-in-time concept learning method to crawl the related images in an image search engine, and train a real-time SVM model for concepts that are not involved in the concept library in the query. Generally speaking, due to the diversity and contextual relevance of text description, it is nearly impossible to accurately parse a sentence into a series of concepts. Therefore, the concept-based model has a serious shortcoming in how to select concepts.

Compared with the concept-based approaches, the concept-free approaches directly encode the whole sentence to capture information from text modality. In these kind of methods, multi-modal inputs from different fields are embedded into the same feature space, and a common feature subspace is applied to realize semantic association. Finally, the similarity between the two is calculated there. These methods avoid the problem of concept selection and use feature mapping to unify the information representation of different modalities, rather than translate into detection. Therefore, we put the emphasis on this kind of methods. There are two significant problems for these methods: one is the feature encoding of text or videos, and the other is to measure the similarity between two modalities.

For text encoding, there are many effective strategies in natural language processing tasks, such as bag-of-words, word2vec and LSTM, while there are also many advanced models for video processing based on the deep neural network (CNN). For instance, Habibian et al. [24] use bag-of-words to encode sentence, and references [15,25] exploit different variants of recurrent neural networks. The multi-scale encoding strategy uses word2vec and the recurrent neural network (RNN) jointly in [12,13,15] and bag-of-words, word2vec and GRU in [14,26,27]. However, these methods still extract single-level visual features using pre-trained CNN model and then mean pooling [12–15,24,25]. For example, Yu et al. [15] exploit the LSTM to obtain the temporal information on the video after the features are extracted by ResNet-152.

For multi-level video encoding to obtain a powerful representation, Dong et al. [10] first propose a dual multi-level encoding method, which not only adopts various text encoding strategies, but also performs similarly in visual information. Specifically, it processes video and text separately and uses mean pooling, bi-GRU and CNN for three-level encoding on each side, so as to explicitly and progressively exploit global, local and temporal patterns features in both videos and sentences. Finally, the full representation of single-modality features formed. The triple loss function defined by

VSE++ [16] is used to focus on the hardest negative samples in its common learning space. On this basis, Wu et al. [28] propose a hybrid sequence encoder method jointly using GRU, VLAD [29] and graph modeling, and map video and language embedding into the semantic space of learning. For powerful features, dual encoding [10] is used as baseline to construct our dual multi-level encoding framework in this paper. However, all the methods above just consider semantic association through calculating the text-video similarity but neglect text-video semantic consistency cues that can be easily obtained from data without extra labelling. Moreover, multi-level video encoding should appropriately fuse the advanced algorithm like SlowFast [17] and BERT [18].

Generally, the existing methods lack the hidden semantic consistency information among the data. The reasonable multiple tasks are not exploited as regular items, to get the better multi-level feature encoding and more accurately similarity for multiple modalities. Thus, we propose multiple distances to measure the similarity of text and video, and design multi-task constraint to establish public space in this paper. To the best of our knowledge, this paper is the first work for multi-task learning to text-video retrieval task.

3. Cross-Modal Learning Based on Semantic Correlation and Multi-Task Learning

In this section, we describe our approach in more detail. As shown in Figure 3, we use the dual multi-level encoding framework, and propose a cross-modal retrieval algorithm based on multi-task learning. Firstly, with the advanced deep neural networks to represent the video feature, a multi-level feature description of video clips is established (in Section 3.1). Secondly, using pre-trained language model and recurrent neural network, we extract the feature of a given query text (in Section 3.2). Next, the features of the video and text are mapped into the common learning space, respectively. We focus on this common space and come up with a novel idea, which is to judge the text-video similarity with absolute distance and relative distance at the same time. The two types of distance are also used to train a multi-task to classify positive and negative text-video sample pairs to establish a semantic alignment relationship between different modalities (in Section 3.3). In the end, we get an integrated cross-modal retrieval network model.

3.1. Multi-Level Video Semantic Feature Encoding

For a given video, we uniformly extract a sequence of n frames from every 0.5 s as $frame(t)$ where $t = 1, \ldots, n$ indicates the specific time step. By three-level encoding strategy, we get video feature of Level 1, Level 2 and Level 3. The final multi-scale video semantic feature is formed by concatenating them together:

$$\varphi(v) = [\varphi_1(v),\ \varphi_2(v),\ \varphi_3(v)], \tag{1}$$

where $\varphi(v)$ is the multi-scale video encoding, $\varphi_1(v)$ is the global level encoding, $\varphi_2(v)$ is the temporal-aware level encoding, $\varphi_3(v)$ is the temporal-domain multi-scale level encoding.

3.1.1. Global Encoding

The model utilizes SlowFast to extract the Level 1 encoding feature of video. We use the SlowFast model pre-trained on Kinetics-400 to extract the deep feature of each frame. The structure of SlowFast network is mainly composed of two channels, including slow pathway, with a low frame rate but a high channel number, and fast pathway, with a high frame rate but a low channel number. The data of the fast pathway is sent to the slow pathway through a lateral connection [17]. However, the fast pathway mainly increases training time through dense sampling, so as to extract the features of rapidly changing scenes or actions more accurately. The video feature $\{v_1, v_2, \ldots, v_n\}$ of $2048 \times n$ dimension can be obtained by only using the output of the slow pathway. After average pooling the global level feature, the result of 2048×1 dimension is $\varphi_1(v)$:

$$\varphi_1(v) = \frac{1}{n} \sum_{t=1}^{n} v_t, \tag{2}$$

where v_t indicates the SlowFast slow pathway feature vector of the t-th frame, $t \in [1, n]$.

3.1.2. Temporal-Aware Encoding

Due to the temporal correlation of video sequence, we use ResNet152 pre-trained on ImageNet to extract $2048 \times n$ dimension feature as input like [12], and send to the recurrent neural network (RNN) [30]. Taking into account the amount of calculation parameters and the bidirectional of video association between front and back frames, we use bi-GRU [31] with hidden vectors size of 512 to obtain the forward hidden state $\overrightarrow{h_t}$ and backward hidden state $\overleftarrow{h_t}$. Finally, the output of bi-GRU and the mean pooling representation of Level 2 are calculated as follows:

$$h_t = \left[\overrightarrow{h_t},\ \overleftarrow{h_t}\right], \tag{3}$$

$$\varphi_2(v) = \frac{1}{n}\sum_{t=1}^{n} h_t, \tag{4}$$

where h_t is in $1024 \times n$ dimensional $\{h_1, h_2, \ldots, h_n\}$ that indicates the t-th frame output of bi-GRU, $t \in [1, n]$ and the size of $\varphi_2(v)$ is 1024×1.

3.1.3. Temporal-Domain Multi-Scale Encoding

In order to enhance the feature with different time ranges in the whole video that help discriminate between videos of subtle difference, we build one-dimensional convolutions [32] of the output $H = \{h_1, h_2, \ldots, h_n\}$ of bi-GRU with multi-scale time domain, and then concatenate the convolution results with different kernel size as follows:

$$c_k = \text{max pooling}\Big(relu\Big(Conv1d_{k,r}(H)\Big)\Big), \tag{5}$$

$$\varphi_3(v) = [c_2,\ c_3,\ c_4,\ c_5], \tag{6}$$

where $Conv1d_{k,r}$ is a 1-d convolution block and the number of the filter is $r = 512$, their kernel size $k = 2, 3, 4, 5$. After *relu* activation and max pooling, each k will get a result c_k with the size of 512×1, so $\varphi_3(v)$ is 2048×1 by concatenation operation.

3.2. Multi-Level Text Semantic Feature Encoding

Given a sentence with the input length of m, on the text side, we get $\varphi_1(s)$, $\varphi_2(s)$ and $\varphi_3(s)$ through three-level encoding, similarly to that in Section 3.1. Finally, there is the text feature encoding $\varphi(s)$ with size of 3324×1:

$$\varphi(s) = [\varphi_1(s),\ \varphi_2(s),\ \varphi_3(s)], \tag{7}$$

Specifically, we first use a one-hot vector to represent each word of the input sentence. Then—unlike the common word2vec [33] or Glove [34] embedding methods, which generate a fixed vector for each word—we use the BERT model [18] to obtain the word embedding vector. The model calculates the vector of the word in the current context according to the context position of the word to get the sentence level word vector representation as $W = \{w_1,\ w_2,\ \cdots,\ w_m\}$. W is a $768 \times m$ dimensional vector, and the average pooling result shows the 768×1 dimensional global level encoding $\varphi_1(s)$. On the basis of Level 1, W is taken as input to extract feature with bi-GRU model as same as that used on the video side. The size of hidden vectors is also set to 512, and the output $1024 \times m$ dimension encoding is sent to the average pooling layer to obtain 1024×1 temporal-aware encoding $\varphi_2(s)$. Level 3 feature extraction takes $k = 2, 3, 4$ as the kernel size of one-dimensional convolution filter, which is similar to the video encoding in Formula (5). The three scales get 512×1 dimensional features, respectively, and then they are concatenated to form a temporal-domain multi-scale encoding $\varphi_3(s)$ in 1536×1.

3.3. Cross Modal Multi-Task Learning

According to Formulas (1) and (7), we obtain the 5120×1-dimensional video encoding $\varphi(v)$ and 3324×1-dimensional text encoding $\varphi(s)$. To unite different modalities feature, we establish semantic association in a sharing subspace, that is mean, video feature $\varphi(v)$ and text feature $\varphi(s)$ are mapped into the same public space by affine transformation [16]. Then, in order to calculate the similarity distance on the same scale, it needs to normalize them. We get $f(v)$ and $f(s)$ as follows:

$$f(v) = \text{Norm}_{l_2}(W_v\varphi(v) + b_v) \tag{8}$$

$$f(s) = \text{Norm}_{l_2}(W_s\varphi(s) + b_s) \tag{9}$$

where W_v, W_s are parameters of affine transformation on video and text side, respectively, b_v and b_s are the bias terms, $\text{Norm}_{l_2}(\cdot)$ means l_2 normalization.

In the stage of learning parameters of the feature fusion layer, we propose a multi-task learning method based on semantic association, including the text-video semantic consistency classification task and the similarity measurement task. The similarity task is to determine the similarity between two modalities by calculating the text-video distance in common learning space, while, the classification task is to add constraints to the former through binary classification. Here, semantic consistency means that the video and caption match with each other, in other words, it is the text-video positive sample pair during the training process. On the contrary, the semantic inconsistency indicates the text-video negative sample pair.

For the multi-task learning method above, we establish a loss function hoping that the model will not only classify positive and negative test-video pairs as accurately as possible, but also reduce the internal distance of sample pairs and expand the internal distance of negative sample pairs. The specific loss function is defined as:

$$loss(v, s; \theta) = w_{sim} loss_{sim}(v, s; \theta) + w_{clf} loss_{clf}(v, s; \theta), \tag{10}$$

where the overall loss function of the model is denoted as $loss(v, s; \theta)$. $loss_{sim}(v, s; \theta)$ is the text-video similarity task loss, $loss_{clf}(v, s; \theta)$ is the loss of the text-video semantic consistency classification task, and w_{si} and w_{clf} are the weight between multi-task respectively. In our method, the similarity task is the main task, supplemented by the classification consistency task for regular optimization, so w_{sim} and w_{clf} are set to 1, 0.5. The loss of each task is shown in the following Formulas (11)–(14) in details, where θ represents all the trainable parameters.

3.3.1. Text-Video Similarity Task Loss

The text-video similarity task loss function can be obtained by using the cosine distance and Euclidean distance which calculated by $f(v)$ and $f(s)$ in common subspace. Specifically, the task consists of two parts: one is cosine distance to calculate the marginal ranking loss [16]; the second is the Euclidean distance to measure the absolute distance between its own text and video. $loss_{sim}$ is shown as follows:

$$loss_{sim}(v, s; \theta) = w_{mrl} loss_{mrl}(v, s; \theta) + w_{ad} loss_{ad}(v, s; \theta), \tag{11}$$

$$\begin{aligned} loss_{mrl}(v, s; \theta) = &\max(0,\ \alpha_1 + S_\theta(v, s^-)_{max} - S_\theta(v, s)) \\ &+\max(0,\ \alpha_1 + S_\theta(v^-, s)_{max} - S_\theta(v, s)) \end{aligned} \tag{12}$$

$$loss_{ad}(v, s; \theta) = D^2_\theta(v, s) + \max^2(0, \alpha_2 - D_\theta(v, s^-)) + \max^2(0, \alpha_2 - D_\theta(v^-, s)) \tag{13}$$

where $loss_{mrl}$ represents the margin ranking loss, and $loss_{ad}$ represents the loss defined by the absolute distance, the weights w_{mrl} and w_{ad} are set to 1, 1.5 in our experiments. In Formulas (13) and (14), (v, s) represents the text-video positive sample pair, and s^- and v^-, respectively, represent the negative text sample matching v and the negative video sample of matching s, so as to form (v, s^-) and (v^-, s),

two types of negative sample pair. The two kinds of $(v, s^-)_{\max}$ or $(v^-, s)_{\max}$ negatives are not randomly sampled, but selected in the current batch, which is the most similar negative sample pair named hardest negative sample. The video text similarity represented by $S_\theta(\cdot)$ and $D_\theta(\cdot)$ is computed by using cosine similarity and Euclidean distance between $f(v)$ and $f(s)$ in the common learning space, where the margin constant is set $\alpha_1 = 0.2$, $\alpha_2 = 0.4$ in our experiments.

3.3.2. Text-Video Semantic Consistency Classification Task Loss

We add the text-video semantic consistency classification, and use binary cross-entropy loss as our loss function. This task is to increase the consistency constraints in the process of feature mapped into the shared feature subspace, maintain the semantic information of feature data between text-video mode and each single mode better, and improve the expression ability in common learning space. At the same time, the addition of semantic consistency classification loss function is equal to regularization from another point of view, which optimizes the solution space of similarity task. It should be noted that we only add semantic consistency classification tasks in the network model training stage. Similar to the above, cosine distance and Euclidean distance of the positive sample pairs and the hardest negative sample pairs are chosen here. The classification loss function is:

$$loss_{clf}(v, s; \theta) = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})), \tag{14}$$

where y is the actual label of the sample pairs, we record the text-video positive sample pair label as 1, otherwise negative as 0, and $\hat{y}$ is the probability that the model predicts the sample is a positive example.

4. Experiments

We have carried out experiments on the public datasets MSVD [35] (https://www.cs.utexas.edu/users/ml/clamp/videoDescription/) and MSR-VTT [36] (http://ms-multimedia-challenge.com/2016/dataset), which are suitable for the text-video retrieval task. Through ablation experiments, it proves the effectiveness of the proposed method. After the performance comparison and analysis with other methods, the superiority of our method is finally reflected.

4.1. Dataset

MSVD. The Microsoft Video Description dataset (MSVD) is a public video description dataset that connects videos and languages proposed by Microsoft, but with a relatively small scale. The dataset contains 2089 video clips from YouTube and their corresponding natural language descriptions. Unfortunately, due to the volatility of YouTube, some of the videos were removed before we could archive them. A total of 1970 out of 2089 video clips are downloaded in our MSVD. We divided 1200 clips for training, 100 clips for validation and 670 clips for testing in experiments.

MSR-VTT. The MSR-Video to Text dataset (MSR-VTT) is another similar but larger text video data set established by Microsoft after MSVD. This is achieved by collecting 257 popular queries from a commercial video search engine, with 118 videos for each query. Videos are edited into clips. The current version provides a total of 10 k video clips, each with about 20 English sentences. The total length of the video is 41.2 h and the number of segment sentence pair is 200 k, covering the most comprehensive category and diverse visual content, with huge sentence and vocabulary components. We use the official data partition that 6513 clips for training, 497 clips for validation and 2990 clips for testing.

4.2. Measurements

The experimental measurements use mAP, and return the results of *R@K* and Med r as a reference for model performance at the same time. The higher *R@K* and mAP, the better the performance of the model, while the lower Med r, the better.

mAP. AP is average precision and mAP is mean average precision. They are computed as follows:

$$AP = \frac{1}{m}\sum_{i=1}^{m}\frac{i}{rank_i}, \tag{15}$$

$$mAP = \frac{1}{M}\sum AP, \tag{16}$$

where for a query input, the number of all returned retrieval results is n. The results are ranked in order. Among them, there are m results correctly relevant to this query. For the i-th in m, the corresponding rank in n is $rank_i$, so $i \in [1, m]$. M represents the total number of all queries, and mAP is the mean of AP in all queries.

R@K. *R@K* represents the proportion of queries that have at least one relevant result among the top k retrieval results in all queries. The calculation formula is as follows:

$$R@K = \frac{1}{M}\sum_{q=1}^{M}\frac{1}{S_q} \times 100\%, \tag{17}$$

where M is the number of all queries. S_q denotes the recall score, $q \in [1, M]$. For a query input, the number of all returned retrieval results in order is n, we record the rank of the first relevant item in n results as $rank_1$, if $rank_1 \leq K$, then $S_q = 1$, otherwise $S_q = 0$. Here $K = 1,\ 5,\ 10$.

Med r. Med r is named median rank, which means the median rank of the first relevant item in the retrieval results.

4.3. Implementation Details

Our experiments are carried out on a GPU of NVIDIA GeForce RTX 2080 with 10 GB memory. The process of mapping the feature into the common space uses the size of 2048. For training, we set the mini-batch size to 128 and use SGD with Adam. Each round of learning rate changes according to the exponential decay strategy. Our initial learning rate is set to 0.0001, and the decay rate coefficient is 0.90. The maximum number of epoch is 50, however, if the training accuracy rate is not improved within 10 epochs, the training will be stopped in advance. All the experiments are carried out with Pytorch.

4.4. Experiment Results

To prove the usefulness of each strategy in our method, we conduct ablation experiments on the MSVD (Table 1) and compare the mAP results with several existing algorithm models (Table 2). All experiments are repeated on MSR-VTT again (Tables 3 and 4).

Table 1. Ablation study on the Microsoft Video Description (MSVD) dataset.

Index	Baseline	Our Works				Results				
	Dual Encoding	Semantic Similarity			Multi-Task [4]	*R@1*(%)	*R@5*(%)	*R@10*(%)	Med r	mAP
		fea_s [1]	fea_v [2]	*loss* [3]						
1	√					12.7	34.5	46.4	13.0	0.234
2	√	√				13.7	34.8	47.0	12.0	0.242
3	√	√	√			14.7	39.3	52.4	9.0	0.268
4	√	√	√	√		15.3	39.5	52.7	9.0	0.272
5	√	√	√	√	√	15.6	39.9	53.3	9.0	0.276

[1] $feat_s$ represents using BERT to extract text features; [2] $feat_v$ represents using the slow pathway in SlowFast to encode video global features; [3] *loss* represents adding the absolute distance of the text-video in common learning space compared with the original loss function in baseline according to Formula (11). [4] multi-task means the improvement method of utilizing semantic consistency classification task loss.

Table 2. State-of-the-art on MSVD dataset.

Methods	*R@1*(%)	*R@5*(%)	*R@10*(%)	Med r	mAP
W2VV [14]	/	/	/	/	0.100
VSE [37]	12.3	30.1	42.3	14.0	/
VSE++ [16]	15.4	39.6	53.0	9.0	0.218
Dual Encoding [10]	12.7	34.5	46.4	13.0	0.234
Ours [1]	15.6	39.9	53.3	9.0	0.276

[1] Ours means the final method in Table 1 index 5.

Table 3. Ablation study on the MSR-Video to Text (MSR-VTT) dataset.

Index	Baseline	Our Works				Results				
	Dual Encoding	Semantic Similarity			Multi-Task [4]	*R@1*(%)	*R@5*(%)	*R@10*(%)	Med r	mAP
		fea_s [1]	fea_v [2]	*loss* [3]						
1	√					8.0	22.9	32.6	32.0	0.159
2	√	√				8.6	24.3	34.2	27.0	0.169
3	√	√	√			8.7	24.8	34.9	26.0	0.172
4	√	√	√	√		8.9	24.9	34.9	27.0	0.173
5	√	√	√	√	√	9.0	25.1	35.1	26.0	0.174

[1] $feat_s$ represents using BERT to extract text features; [2] $feat_v$ represents using the slow pathway in SlowFast to encode video global features; [3] *loss* represents adding the absolute distance of the text-video in common learning space compared with the original loss function in baseline according to Formula (11). [4] multi-task means the improvement method of utilizing semantic consistency classification task loss.

Table 4. State-of-the-art on the MSR-VTT dataset.

Methods	*R@1*(%)	*R@5*(%)	*R@10*(%)	Med r	mAP
W2VV [14]	1.8	7.0	10.9	193.0	0.052
VSE [37]	5.0	16.4	24.6	47	/
VSE++ [16]	5.7	17.1	24.8	65	/
$W2VV_{imrl}$ [10]	6.1	18.7	27.5	45.0	0.131
Mithun et al. [12]	6.8	20.7	29.5	39.0	/
Dual Encoding [10]	7.7	22.0	31.8	32.0	0.155
HYBRID [28]	7.8	22.5	32.0	31.0	0.158
Ours [1]	9.0	25.1	35.3	26.0	0.175

[1] Ours means the final method in Table 3 index 5.

The baseline method dual encoding in Tables 1 and 3 represents the implementation of its model in our experimental environment. We divide the proposed method into four sub steps for ablation experiments. These four steps can be regarded as two aspects: the innovation of text-video cross-model retrieval task and the structure of multi-task with semantic consistency classification task. It can be concluded from Table 1 that our four strategies are all effective.

We utilize BERT [18] as the level-1 encoding model to get high-quality text feature and as an input to downstream models. So that there will be more accurate text information to retrieve video. As a widely used word embedding, word2vec has a fixed representation for each word regardless of the context within where the word appears. However, BERT produces word representations that are dynamically informed by the words around them [38]. That is to say, if a word appears in two sentences at the same time, the word embedding they get through word2vec are the same, but different through BERT. By jointly using BERT as the global encoding, we get 0.8% mAP improvement.

A better feature is also needed in visual, so we utilize advanced SlowFast model for level-1 video encoding. This network is based on 3D ResNet, which captures the spatiotemporal feature in both the slow and fast pathways. The slow pathway focuses on spatial domain and semantic information in visual and has a lateral connection with the fast pathway [17]. When we integrate SlowFast as a level-1 encoder, we do not explicitly get the fast pathway output feature, but it is still fused into output of

the slow pathway by the lateral connection. In this way, we obtain the spatiotemporal visual features which contain rapidly changing information without increasing dimension. The improvement of 2.6% proves the effectiveness of this visual feature encoding in MSVD.

On the basis of high-quality feature, modifying the original loss function $loss_{sim}$ also brings partial advancement to our model. As a common loss function in cross-modal retrieval, marginal ranking loss only considers the intra-distance between videos. Its purpose is to make a distance between the relevant video and the nearest negative video as much as possible. However, there is no way to discuss the distance between text-video. Our loss function (in Section 3.3.1) with the above two kinds of distance is helpful. The measure *R@5* or mAP is improved 0.2% at least.

Then by adding the classification task of text-video semantic consistency, we make the multi-task joint loss function which shown in Formula (10). The mAP of the model before and after the multi-task experiment is 27.2% and 27.6%, respectively. That is because the classification task is equal to a regularization term, so it can constrain and optimize the learning of the semantic association task. In fact, because of the fusion of various encoding strategies in multi-level, the complexity of our model is relatively high. In order to get better generalization ability, we make multi-task as a constraint. Classifying text-video pairs can restrict the establishment of common learning space.

Table 2 shows the comparison with the state-of-the-art. All the methods in this table are concept-free not concept-based. Dong et al. [14], Krios et al. [37] and Faghri et al. do not jointly use multiple encoding strategies in both modalities; while Dong et al. [10] first propose dual multi-level encoding, and these four approaches have different loss functions. Finally, our method has achieved the best performance compared with them, which further illustrates the effectiveness of our algorithm.

Due to the scale of the MSVD dataset, and in order to further verify the universality of the model, we then do experiments on the MSR-VTT dataset. There are more advanced methods carrying out experiments on this dataset, which is helpful for comparing the performance of different models. Tables 3 and 4 are the experimental results of this algorithm on MSR-VTT. The ablation experiments prove the four strategies that feature encoding of video and text, the optimization of the loss function, and the improvement of multi-task learning by adding classification consistency tasks are all useful. In Table 4, the result mAP of Wu et al. [28], which combines several algorithms, is 0.158, which has the best performance on the AVS task in TRECVid 2019, but our method still gets a 1.7% improvement. The results show that our model is outstanding compared with the existing method mAP, and the retrieval result is better, which confirms the effectiveness of the algorithm proposed in this paper.

In addition to mAP evaluation, *R@K* is also a public metric reflecting the performance of the model. Figures 4 and 5 show the *R@K* curves of different methods on two datasets. By analyzing the trend of all single curves, we can see that *R@K* is higher with the larger range of *K*. It shows that the recall rate in top 1 of text video retrieval task is not high, but it will be improved when we calculate in the top 10 or top five results. Furthermore, comparing the increase from *R@1* to *R@5* with the increase from *R@5* to *R@10*, it is easy to find that the improvement between *R@1* and *R@5* is significantly greater than between *R@5* and *R@10*. This shows that more correctly retrieval appear in the top five results. By analyzing *R@K* in different methods, it is known that the higher mAP, the higher *R@K*. Our approach gets the best *R@K* in both datasets. The results are consistent with the conclusion by mAP, which fully proves the effectiveness of our method again.

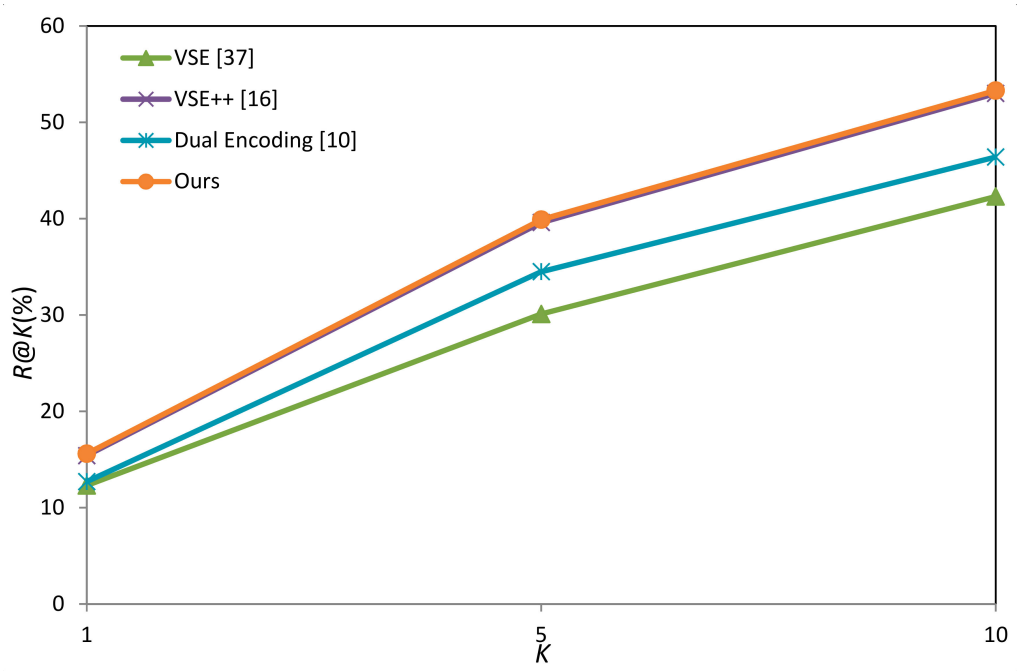


Figure 4. Trend curves of *R@K* in the MSVD dataset.

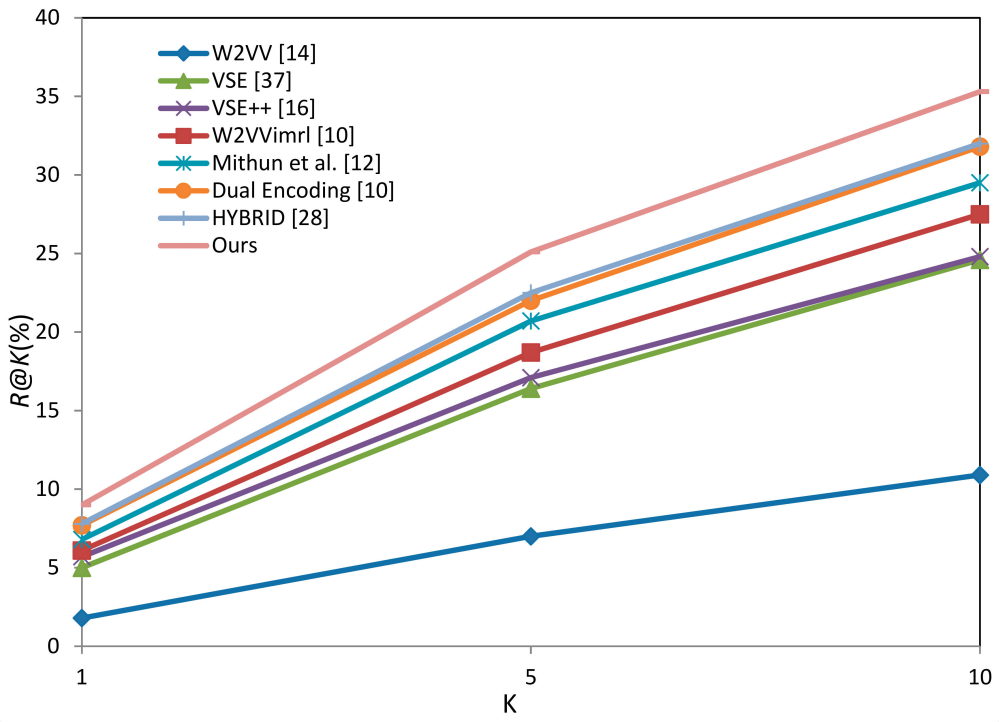
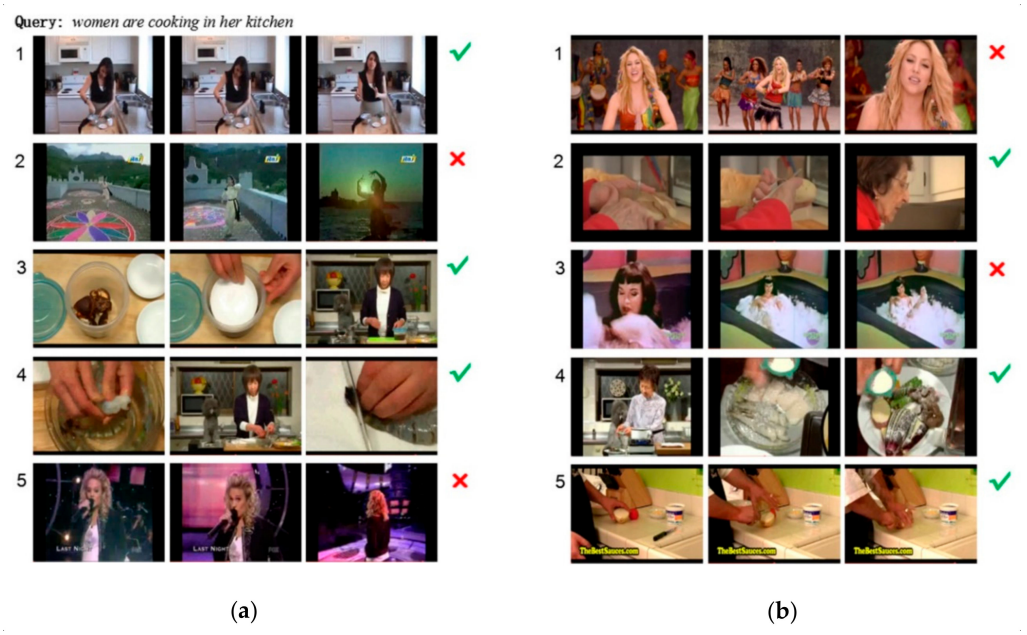


Figure 5. Trend curves of *R@K* in the MSR-VTT dataset.

We randomly select a query text on the corresponding dataset and display the top five video shots from the returned retrieval results, as shown in Figures 6 and 7. The left video shots correspond to our multi-task cross-modal retrieval method, and the right is dual encoding [10]. In the sample query, the top-1 result of dual encoding in either dataset is wrong, while our method is also in error in MSR-VTT, but retrieves correctly in MSVD. Observing the top five results returned, there are more videos matching the query text. Combined with *R@K* evaluation analysis above, it can be seen that the value of *R@1* for any model is low, but the rate of *R@5* is two to three times that of R@1, as shown in Figure 4. As a result, the top-1 shot cannot be correctly matched with the query text, and the retrieval

error rate is higher. However, once the requirement is adjusted to the top five, there are more relevant items. According to Figure 6, for the query text "women are cooking in her kitchen" on the MSVD dataset, our method, with the top-1, top-3 and top-4 correct results, performs better than the results of dual encoding, of which the top-2, top-4 and top-5 results are right. In the same way, the sample query text is "a band is performing in a small club" in Figure 7. Our method has an error only in the top-1 video, while the dual encoding has in both the top-1 and top-3. The examples of visualization more intuitively show the performance of our method.



(**a**) (**b**)

Figure 6. Top five video shots returned on MSVD dataset with a query. (**a**) is our method and (**b**) is dual encoding. The video clip with $\surd$ in green means correct result, with × in red means error.

(**a**) (**b**)

Figure 7. Top five video shots returned on MSR-VTT dataset with a query. (**a**) is our method and (**b**) is dual encoding. The video clip with $\surd$ in green means correct result, with × in red means error.

5. Conclusions

For text-video cross-modal retrieval, this paper proposes a cross-modal video retrieval algorithm based on semantic association and multi-task learning. Firstly, a multi-level video feature encoding at the global video level, temporal-aware level and time-domain multi-scale level is proposed. Then, based on the BERT model, we further get the text multi-level feature encoding. Finally, in the process of cross-modal semantic association, a multi-task learning method considering semantic similarity measurement and semantic consistency classification is proposed to construct common learning space with semantic preservation. In experiments, by testing on two public datasets MSVD and MSR-VTT, it is proved that the method we proposed has achieved better results than the existing methods. The results indicate the effectiveness of the algorithm in this paper.

At present, the research in this paper is still based on labeled training data, but the feature and knowledge learned by this method is limited by the size and distribution of data. In the next step, we will consider integrating external knowledge, making the external prior information of knowledge graph embed into the network structure [39,40] to explore the effective fusion of external knowledge and labeled data information, and try to further improve cross-modal retrieval.

Author Contributions: Conceptualization, X.W. and T.W.; methodology, X.W. and T.W.; software, T.W.; validation, T.W. and X.W.; formal analysis, X.W. and T.W.; investigation, X.W. and T.W.; resources, X.W. and S.W.; data curation, T.W.; writing—original draft preparation, T.W.; writing—review and editing, X.W. and S.W.; visualization, T.W.; supervision, X.W.; project administration, X.W.; funding acquisition, X.W. and S.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by National Natural Science Foundation of China (No. 61801441, No. 61701277, No. 61771288), National Key R & D plan of the 13th Five-Year plan (No. 2017YFC0821601), cross media intelligence special fund of Beijing National Research Center for Information Science and Technology (No. BNR2019TD01022), discipline construction project of "Beijing top-notch" discipline (Internet information of Communication University of China) and in part by the State Key Laboratory of Media Convergence and Communication, Communication University of China.

Acknowledgments: The authors are thankful for organizing Committee of international competition Text Retrieval Conference Video Retrieval Evaluation (TRECVid) and the teachers and students who communicate and coordinate during the outbreak of COVID-19.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chang, X.; Yang, Y.; Hauptmann, A.; Xing, E.P.; Yu, Y.L. Semantic concept discovery for large-scale zero-shot event detection. In Proceedings of the Twenty-fourth International Joint Conference on Artificial Intelligence (IJCAI'15), Buenos Aires, Argentina, 25–31 July 2015; pp. 2234–2240.
2. Dalton, J.; Allan, J.; Mirajkar, P. Zero-shot video retrieval using content and concepts. In Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM'13), San Francisco, CA, USA, 27 October–1 November 2013; pp. 1857–1860.
3. Habibian, A.; Mensink, T.; Snoek, C.G.M. Composite concept discovery for zero-shot video event detection. In Proceedings of the 4th ACM International Conference on Multimedia Retrieval (ICMR'14), Glasgow, UK, 1–4 April 2014; pp. 17–24.
4. Markatopoulou, F.; Moumtzidou, A.; Galanopoulos, D.; Mironidis, T.; Kaltsa, V.; Ioannidou, A.; Symeonidis, S.; Avgerinakis, K.; Andreadis, S.; Gialampoukidis, I.; et al. ITICERTH Participation in TRECVID 2016. Available online: https://www-nlpir.nist.gov/projects/tvpubs/tv16.slides/tv16.avs.iti-certh.slides.pdf (accessed on 18 October 2020).
5. Jiang, L.; Meng, D.; Mitamura, T.; Hauptmann, A.G. Easy Samples First: Self-paced Reranking for Zero-Example Multimedia Search. In Proceedings of the 22nd ACM International Conference on Multimedia (MM'14), Orlando, FL, USA, 4 November 2014; pp. 547–556.
6. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
7. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV'17), Venice, Italy, 22–29 October 2017; pp. 2980–2988.

8. Lu, Y.J.; Zhang, H.; de Boer, M.H.T.; Ngo, C.W. Event detection with zero example: Select the right and suppress the wrong concepts. In Proceedings of the 6th ACM International Conference on Multimedia Retrieval (ICMR'16), New York, NY, USA, 6–9 June 2016.
9. Tao, Y.; Wang, T.; Machado, D.; Garcia, R.; Tu, Y.; Reyes, M.P.; Chen, Y.; Tian, H.; Shyu, M.L.; Chen, S.C. Florida International University—University of Miami Participation in TRECVID 2019. Available online: https://www-nlpir.nist.gov/projects/tvpubs/tv19.papers/fiu_um.pdf (accessed on 7 March 2020).
10. Dong, J.; Li, X.; Xu, C.; Ji, S.; He, Y.; Yang, G.; Wang, X. Dual encoding for zero-example video retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9346–9355.
11. Liu, Y.; Albanie, S.; Nagrani, A.; Zisserman, A. Use what you have: Video retrieval using representations from collaborative experts. In Proceedings of the 30th British Machine Vision Conference (BMVC'19), Cardiff, Wales, UK, 9–12 September 2019.
12. Mithun, N.C.; Li, J.; Metze, F.; Roy-Chowdhury, A.K. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval (ICMR'18), Yokohama, Japan, 11–14 June 2018; pp. 19–27.
13. 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 Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI'15), Austin, TX, USA, 25–29 January 2015.
14. Dong, J.; Li, X.; Snoek, C.G.M. Predicting Visual Features from Text for Image and Video Caption Retrieval. *IEEE Trans. Multimed.* **2018**, *20*, 3377–3388. [CrossRef]
15. Yu, Y.; Ko, H.; Choi, J.; Kim, G. End-to-End Concept Word Detection for Video Captioning, Retrieval, and Question Answering. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3261–3269.
16. Faghri, F.; Fleet, D.J.; Kiros, J.R.; Fidler, S. VSE++: Improved Visual-Semantic Embeddings. In Proceedings of the 29th British Machine Vision Conference (BMVC'18), Newcastle upon Tyne, UK, 3–6 September 2018.
17. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. SlowFast Networks for Video Recognition. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 6201–6210.
18. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
19. TRECVid AVS Task. Available online: https://www-nlpir.nist.gov/projects/tv2016/tv2016.html (accessed on 6 March 2020).
20. Le, D.D.; Phan, S.; Nguyen, V.T.; Renoust, B.; Nguyen, T.A.; Hoang, V.N.; Ngo, T.D.; Tran, M.T.; Watanabe, Y.; Klinkigt, M.; et al. NII-HITACHI-UIT at TRECVID 2016. Available online: https://www-nlpir.nist.gov/projects/tvpubs/tv16.papers/nii-hitachi-uit.pdf (accessed on 13 November 2020).
21. Nguyen, P.A.; Li, Q.; Cheng, Z.Q.; Lu, Y.J.; Zhang, H.; Wu, X.; Ngo, C.W. VIREO @ TRECVID 2017: Video-to-Text, Ad-hoc Video Search and Video Hyperlinking. Available online: https://www-nlpir.nist.gov/projects/tvpubs/tv17.papers/vireo.pdf (accessed on 5 December 2020).
22. Ueki, K.; Hirakawa, K.; Kikuchi, K.; Ogawa, T.; Kobayashi, T. Waseda_Meisei at TRECVID 2017: Ad-hoc video search. Available online: https://www-nlpir.nist.gov/projects/tvpubs/tv17.papers/waseda_meisei.pdf (accessed on 13 November 2020).
23. Nguyen, P.A.; Wu, J.; Ngo, C.W.; Danny, F.; Huet, B. VIREO-EURECOM @ TRECVID 2019: Ad-hoc Video Search (AVS). Available online: https://www-nlpir.nist.gov/projects/tvpubs/tv19.papers/eurecom.pdf (accessed on 7 March 2020).
24. Habibian, A.; Mensink, T.; Snoek, G.C.M. Video2vec Embeddings Recognize Events When Examples Are Scarce. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2089–2103. [CrossRef] [PubMed]
25. Yu, Y.; Kim, J.; Kim, G. A joint sequence fusion model for video question answering and retrieval. In Proceedings of the 15th European Conference on Computer Vision (ECCV'18), Munich, Germany, 8–14 September 2018.

26. Li, X.; Dong, J.; Xu, C.; Wang, X.; Yang, G. Renmin University of China and Zhejiang Gongshang University at TRECVID 2018: Deep Cross-Modal Embeddings for Video-Text Retrieval. Available online: https://www-nlpir.nist.gov/projects/tvpubs/tv18.papers/rucmm.pdf (accessed on 13 November 2020).
27. Hernandez, R.; Perez-Martin, J.; Bravo, N.; Barrios, J.M.; Bustos, B. IMFD_IMPRESEE at TRECVID 2019: Ad-Hoc Video Search and Video to Text. Available online: https://www-nlpir.nist.gov/projects/tvpubs/tv19.papers/imfd_impresee.pdf (accessed on 7 March 2020).
28. Wu, X.; Chen, D.; He, Y.; Xue, H.; Song, M.; Mao, F. Hybrid Sequence Encoder for Text Based Video Retrieval. Available online: https://www-nlpir.nist.gov/projects/tvpubs/tv19.papers/atl.pdf (accessed on 7 March 2020).
29. Gong, Y.; Wang, L.; Guo, R.; Lazebnik, S. Multi-scale Orderless Pooling of Deep Convolutional Activation Features. In Proceedings of the 13th European Conference on Computer Vision (ECCV'14), Zurich, Switzerland, 6–12 September 2014; pp. 392–407.
30. Schuster, M.; Paliwal, K.K. Bidirectional Recurrent Neural Networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [CrossRef]
31. Cho, K.; Van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In Proceedings of the 19th Conference on Empirical Methods in Natural Language (EMNLP), Doha, Qatar, 25–29 October 2014.
32. Kim, Y. Convolutional Neural Networks for Sentence Classification. In Proceedings of the 19th Conference on Empirical Methods in Natural Language (EMNLP), Doha, Qatar, 25–29 October 2014.
33. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. In Proceedings of the 2013 International Conference on Learning Representations (ICLR), Scottsdale, AZ, USA, 2–4 May 2013.
34. Pennington, J.; Socher, R.; Manning, C. Glove: Global Vectors for Word Representation. In Proceedings of the Conference on Empirical Methods in Natural Language Proceeding (EMNLP'14), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
35. Chen, D.L.; Dolan, W.B.; Yao, T. Collecting highly parallel data for paraphrase evaluation. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL'11), Portland, OR, USA, 19 June 2011; pp. 190–200.
36. Xu, J.; Mei, T.; Yao, T.; Rui, Y. MSR-VTT: A large video description dataset for bridging video and language. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 5288–5296.
37. Kiros, R.; Salakhutdinov, R.; Zemel, R.S. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *arXiv* **2014**, arXiv:1411.2539.
38. BERT Word Embeddings Tutorial. Available online: https://mccormickml.com/2019/05/14/BERT-word-embeddings-tutorial/ (accessed on 4 December 2020).
39. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv* **2016**, arXiv:1609.02907.
40. Chen, S.; Zhao, Y.; Jin, Q.; Wu, Q. Fine-Grained Video-Text Retrieval with Hierarchical Graph Reasoning. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 10635–10644.