mathematics

MDPI

Article

Multi-Level Cross-Modal Semantic Alignment Network for Video–Text Retrieval

Fudong Nian [1,2,*], Ling Ding [1], Yuxia Hu [2] and Yanhong Gu [1]

[1] School of Advanced Manufacturing Engineering, Hefei University, Hefei 230601, China
[2] Anhui International Joint Research Center for Ancient Architecture Intellisencing and Multi-Dimensional Modeling, Anhui Jianzhu University, Hefei 230601, China
* Correspondence: nianfd@hfuu.edu.cn

Abstract: This paper strives to improve the performance of video–text retrieval. To date, many algorithms have been proposed to facilitate the similarity measure of video–text retrieval from the single global semantic to multi-level semantics. However, these methods may suffer from the following limitations: (1) largely ignore the relationship semantic which results in semantic levels are insufficient; (2) it is incomplete to constrain the real-valued features of different modalities to be in the same space only through the feature distance measurement; (3) fail to handle the problem that the distributions of attribute labels in different semantic levels are heavily imbalanced. To overcome the above limitations, this paper proposes a novel multi-level cross-modal semantic alignment network (MCSAN) for video–text retrieval by jointly modeling video–text similarity on global, entity, action and relationship semantic levels in a unified deep model. Specifically, both video and text are first decomposed into global, entity, action and relationship semantic levels by carefully designing spatial–temporal semantic learning structures. Then, we utilize KLDivLoss and a cross-modal parameter-share attribute projection layer as statistical constraints to ensure that representations from different modalities in different semantic levels are projected into a common semantic space. In addition, a novel focal binary cross-entropy (FBCE) loss function is presented, which is the first effort to model the unbalanced attribute distribution problem for video–text retrieval. MCSAN is practically effective to take the advantage of the complementary information among four semantic levels. Extensive experiments on two challenging video–text retrieval datasets, namely, MSR-VTT and VATEX, show the viability of our method.

Keywords: video–text retrieval; multi-level space learning; cross-modal similarity calculation

MSC: 68T09

Citation: Nian, F.; Ding, L.; Hu, Y.; Gu, Y. Multi-Level Cross-Modal Semantic Alignment Network for Video–Text Retrieval. *Mathematics* **2022**, *10*, 3346. https://doi.org/10.3390/math10183346

Academic Editor: Jakub Nalepa

Received: 22 July 2022
Accepted: 10 September 2022
Published: 15 September 2022

1. Introduction

With the popularity of online short video sharing platforms (e.g., TikTok), video–text retrieval (using a video clip as query to find similar sentences and vice versa) becomes a fundamental and crucial information retrieval problem for both users and platforms. As a specific cross-modal retrieval [1] task, many research papers about video–text retrieval have been presented and achieved remarkable progress from 2018 to the present [2–6].

Measuring the semantic similarity between any pair of video–text is very important for video–text retrieval. Basically, existing studies on video–text retrieval can be summarized into two categories according to the roughness of video and text semantic representations: (1) The first one is single-level encoding methods [2,6–8], which leverage convolutional neural networks (CNN), recurrent neural networks (RNN) or Transformer to learn global representation (i.e., single level) of video/text. However, compacting video/text into a single global level representation is insufficient due to neglecting the local and fine-grained static/dynamic information that widely existed in video/text. (2) The second one is multi-level encoding methods [3,4,9–11], which convert video/text into hybrid attribute spaces

(e.g., global, entity and action). The total similarity of the video–text pair is the fusion of the cross-modal similarities in different semantic levels.

Recently, large-scale visual-and-language pretraining (VLP) models (e.g., DALL-E [12], CLIP [13] and MVP [14]), which can learn representations grounded in both visual textual contexts via a self-supervised learning manner, have been successfully applied in various downstream multi-modal tasks such as visual question answering [15], video captioning [16] and visual reasoning [17]. Some video–text retrieval methods utilize VLP models to obtain more representative video and text features [18–21]. Although these approaches show promising performance on video–text retrieval tasks, training VLP means leveraging lots of extra multi-modal training data and a video–text retrieval framework with the VLP model also means adding a number of parameters.

Despite the significance and value of the methods in the above algorithms, they still suffer from three critical shortcomings: (1) *Semantic levels are insufficient*. Though action and entity semantic levels are considered in [3,9,11] and get better performance, the semantic similarity in relationship level is ignored. However, several works [22–24] demonstrate that modeling relationships can improve the performance of many multi-modal understanding tasks significantly. Similarly, as shown in Figure 1, relationships are also common clues in video–text retrieval scenarios. (2) *Incomplete joint space representation*. To compare the cross-modal representations in different semantic levels, most existing works [25,26] utilize ranking or classification losses. Although it has achieved certain results, there is no statistical guarantee that real-valued vectors encoded from different modalities are in a common space. (3) *Unbalanced data distribution problem is neglected*. Figure 2 shows the number of five entities in MSR-VTT [27] dataset, from it we can see that attributes distribution has a serious long tail effect. Unfortunately, to the best of our knowledge, no video–text retrieval frameworks are handling this problem explicitly.

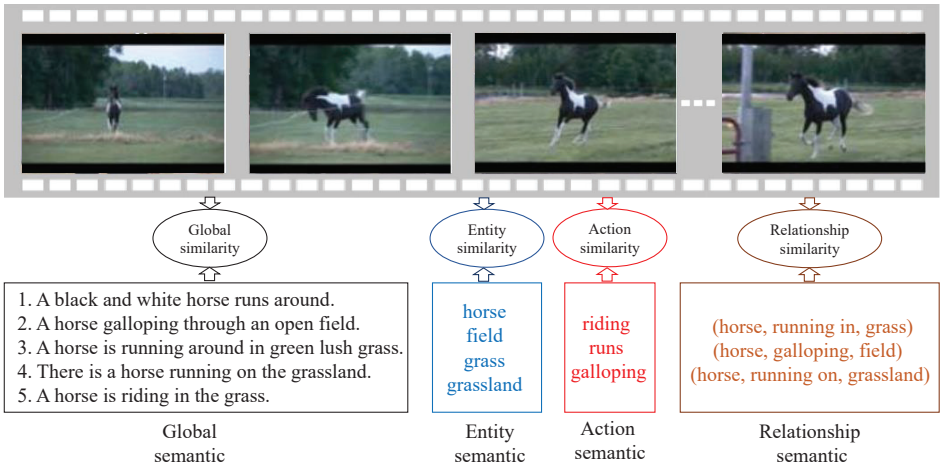


Figure 1. An illustration of both video and text are decomposed into global, entity, action and relationship semantic levels.

To tackle these limitations, in this paper, we present a novel **M**ulti-Level **C**ross-Modal **S**emantic **A**lignment **N**etwork (MCSAN) for video–text retrieval by jointly modeling the video/text context information and unbalanced attributes distribution challenges in a unified deep model. To achieve a robust video–text retrieval, as shown in Figure 1, besides the global representation, we first construct multi-level semantic dictionaries (entity, action and relationship) from the corpus. Then, we design several practical semantic encoding modules including CNN, RNN, GNN [28] and Multi-head self-attention [29], which are utilized to encode both video and text into multi-level semantic spaces (i.e., global, entity, action and relationship). Meanwhile, not only using feature distance, we also utilize KLDivLoss and a cross-modal parameter-share attribute projection layer as statistical constraints to ensure that representations from different modalities in different semantic

levels are projected into a common semantic space. In addition, we design a novel multi-label regression loss to handle the unbalanced attributes distribution problem. Extensive experimental results on two widely-used video–text retrieval benchmark datasets, i.e., MSR-VTT [27] and VATEX [30], to validate the effectiveness and superiority of our MCSAN.

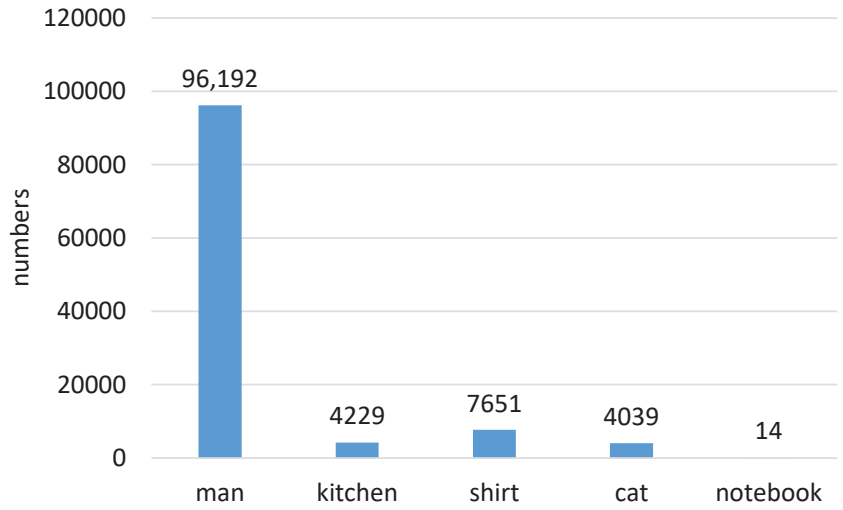


Figure 2. The number of 5 entities in MSR-VTT corpus, we can see that attributes distribution is very uneven.

In summary, this paper makes the following contributions.

- A novel Multi-Level Cross-Modal Semantic Alignment Network (MCSAN) is developed for video–text retrieval tasks. The semantic similarity between video and text is fully considered at four semantic levels (global, action, entity and relationship) by carefully designing spatial–temporal semantic learning structures and cross-modal representation distribution constraints in an end-to-end framework, which offers an alternative orientation for video–text retrieval.
- Compared with existing works, we are the first to measure the similarity of video–text pairs using relationship semantic level. We demonstrate that relationship, which is also a complementary and crucial clue for video–text retrieval.
- We introduce a novel focal binary cross-entropy (FBCE) loss function for multi-label attributes regression. Though several papers adopt attributes to improve the performance of video–text retrieval, to the best of our knowledge, this is the first effort to consider and handle the unbalanced attributes distribution problem in video–text retrieval scenarios.
- By carefully incorporating spatial–temporal semantic learning structures, cross-modal representation distribution constraints, novel multi-label classification loss function, adaptive graph learning, and cross-modal contextual attention module, our proposed method can jointly enjoy the merits of alleviate semantic gap and long-tailed multi-level multi-modal attributes learning for video–text retrieval. Moreover, we conduct extensive experiments on two popular benchmarks, i.e., MSR-VTT [27] and VATEX [30], demonstrating the rationality and effectiveness of the proposed multi-level cross-modal semantic alignment network.

The remainder of this paper is organized as follows: Section 2 reviews the related work. We present the detailed architecture of the multi-level cross-modal semantic alignment network in Section 3, while Section 4 describes the datasets, settings, implementation details, experimental results, and limitations. Finally, we conclude the paper with future work in Section 5.

2. Related Work

The general process of the current video–text retrieval framework can be roughly divided into two blocks, namely, modality encoding (embedding video and text into vectors), and similarity calculation (predicting the similarity score of two modalities). Therefore, in this section, we briefly review typical video–text retrieval algorithms in the above two views.

2.1. Modality Encoding

Existing video and text encoding methods can be roughly divided into two categories. The first one is single-level encoding methods [2,6–8,31,32], which leverage convolutional neural networks (CNN), recurrent neural networks (RNN) or Transformer to learn global representation (i.e., single level) of video/text. However, compacting video/text into a single global level representation is insufficient due to neglecting the local and fine-grained static/dynamic information that widely existed in video/text. The second one is multi-level encoding methods [3,4,9–11], which convert video/text into hybrid attribute spaces (e.g., global, entity and action). The total similarity of the video–text pair is the fusion of the cross-modal similarities in different semantic levels. The popular pipeline for video/text encoding is that: (1) For video encoding, using the pre-trained image classification or video clip classification model to extract the frame-wise feature. For text encoding, using Word2Vec, Glove, or Bert to extract the word-wise feature. (2) Utilizing CNN, RNN, or Transformer to handle temporal information of both video and text.

To enhance the representation ability, Yang et al. [33] constructed a tree structure to model text. Dong et al. [34] proposed a reading-strategy-inspired visual representation learning framework to represent videos. Some works [35–38] made full use of multi-modal cues, e.g., appearance, audio, motion, OCR and face, for video encoding. Although these methods have made some progress, the training pipeline of these methods are very complex, and even a large amount of additional annotation information is required.

Recently, thanks to the rapid development of VLP models [12–14], there are some works [18–21] tend to utilize the VLP model encoding the video and text directly. Although these approaches show promising performance on video–text retrieval tasks since they have better feature representation ability, training VLP means leveraging lots of extra multi-modal training data and the video–text retrieval framework with the VLP model also means adding a mount of parameters. Moreover, due to the training codes of these VLP based video–text retrieval algorithms are not publicly available, we have not compared these methods in this paper.

2.2. Similarity Calculation

The most common method of calculating cross-modal similarity in video–text retrieval task is that first project video and text into a common semantic space and then compute the feature distance in the common space [34,39–42]. Additionally, Lei et al. [39] concatenated video and text features, and fed them into a transformer for multi-modal fusion. Directly predicting the cross-modal similarity by fusing multi-modal representations without learning the common semantic space explicitly in other cross-modal understanding tasks [43,44] is not uncommon. Yu et al. [43] proposed a guided attention module to fuse text and image representations for visual question answering. Yu et al. [44] presented a multi-modal Transformer to cascade the image and text features in depth. for image captioning. Though the recent image-text retrieval literature [45] has demonstrated that multi-modal fusion can make the model more stable, extending it for video retrieval is unwise because it is at the expense of efficiency. It should be noted that all existing methods (whether it is common space base framework) calculate the similarity score either in global or entity and action semantic levels, which neglect the relationship information widely existing in multimedia data. Instead, we are the first to measure the similarity of video–text pairs using relationship semantic level, which is also a complementary and crucial clue for video–text retrieval.

3. Methodology

3.1. Problem Statement

Video–text retrieval includes two aspects: (1) Providing a sentence as a query and return the corresponding videos. (2) Providing a video as a query and return the corresponding sentences. The major challenge to achieve these goals is modeling the similarity between any video–text pairs. Formally, given a video V and a sentence T, the video–text retrieval model will output $S(V,T) = (0,1)$ to indicate the similarity of this video–text pair, where the larger the value of $S(V,T)$, the more similar the semantics of video V and text T.

3.2. Overall Framework

We introduce a novel multi-level cross-modal semantic alignment network (MCSAN) to improve the performance of video–text retrieval tasks. In order to precisely handle both the semantics of video and sentence, we employ several carefully designed spatial–temporal representation structures to encode both video and sentence into four semantic levels (global, action, entity and relationship). To calculate the similarity of video and sentence, we utilize metric learning, statistical distribution constraints and cross-modal parameter-share attribute projection layers to ensure multi-level representations extracted from different modalities are in multi-level common spaces. As the semantic distribution is seriously uneven, we propose a novel multi-label regression loss that can automatical focus on samples that are hard to be identified. The overall architecture of MCSAN is illustrated in Figure 3; in that which follows, we depict it in detail.

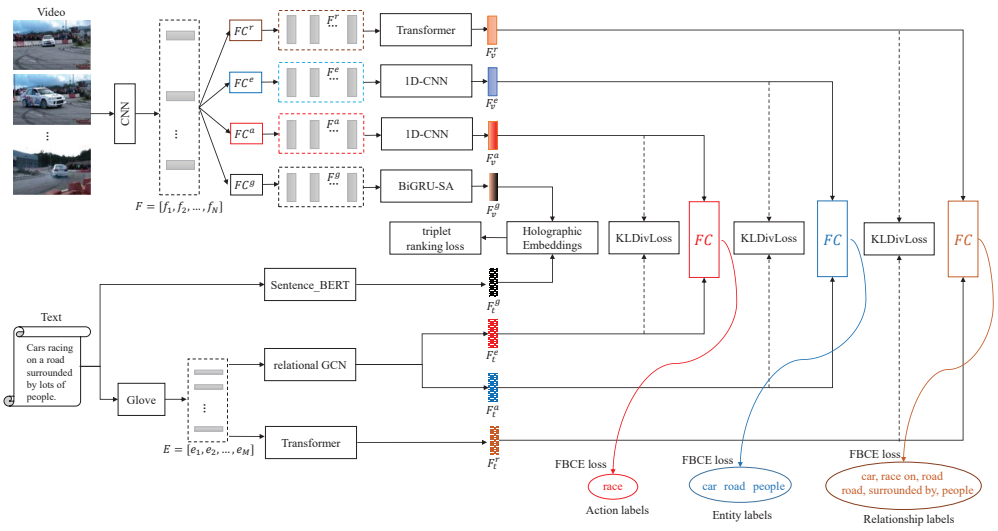


Figure 3. Framework of the proposed multi-level cross-modal semantic alignment network (MCSAN).

The proposed MCSAN is reasonable: (1) Obviously, as shown in Figure 3, a query sentence (e.g., "*Cars racing on a road surrounded by lots of people.*") corresponding to a specific video clip has multiple semantic levels including global-level (the whole sentence), verb-level (actions), and noun-level (entities). The multiple semantic levels involves complicated interactions between them, which are actually organized as a relationship structure, indicating that a textual query can be effectively grounded onto the video by properly aligning different semantic levels with the corresponding video parts. Although some previous methods attempt to utilize the multi-level semantic structure, they fail to capture the fine-grained and explicit relationship information. (2) Intuitively, in video retrieval or video captioning tasks, humans usually takes the relation-aware context of the visual structure into consideration and select the correct temporal moment.

3.3. Multi-Level Attributes Vocabulary Construction

As mentioned above, besides global representation, we also represent both video and text into three attribute spaces, i.e., action, entity and relationship. To achieve this goal, attribute vocabulary is necessary for training the encoding modules. In this work, we employ the Stanford CoreNLP Natural Language Processing Toolkit [46] to build action, entity and relationship vocabularies. Specifically, for a video–text retrieval dataset, we collect all sentences as original corpus, and extract nouns, verbs and subject–verb–object (SVO) triplets, which are deemed as entity, action and relationship, respectively. Then, we lemmatize all words to eliminate duplication of attributes. Finally, based on the frequency statistics, we select the top K_e nouns, K_a verbs and K_r SVO triplets as the final entity, action and relationship vocabularies, respectively.

3.4. Multi-Level Text and Video Semantics Encoding

In this section, we detail how to encode both video/text into global, entity, action and relationship semantic spaces.

3.4.1. Multi-Level Text Semantics Encoding

(1) Text parsing and initial embedding.

For a given sentence T, we first utilize the pre-trained word embedding model as the textual encoder to extract word initial embeddings $E = [e_1, e_2, ..., e_M]$, where M denotes the number of words in sentence T.

(2) Text global semantic encoding.

The text global semantic encoding F_t^g is calculated by pre-trained *Sentence_BERT* [47] since it sets a new state-of-the-art performance on semantic textual similarity task:

$$F_t^g = Sentence\_BERT(T), \tag{1}$$

(3) Text entity, action semantic encoding.

Since prior works [10,11] demonstrate that relational GCN [48] performs well on local-level text representations, we adopt relational GCN to encode the semantics of text entity and action, which are denoted as F_t^e and F_t^a, respectively. The structure of relational GCN is the same as [11] except not handle global feature, please refer to [11] for details.

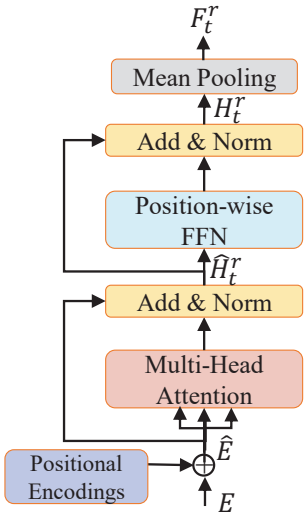


Figure 4. Text relationship semantic encoding architecture.

(4) Text relationship semantic encoding.

Compared with entity or action, the relationship in a sentence is complex, e.g., "*A dog is chasing a frisbee on the grass*" includes two relationships ("*dog chasing frisbee*" and "*dog on grass*") and words in these relationships are nonadjacent. Addressing it, we utilize a Transformer with multi-head self-attention block (as shown in Figure 4) to encode the text relationship semantic.

Specifically, the self-attention layer produces the attention output as follows:

$$S_t^r = softmax(\frac{Q_t K_t^\top}{\sqrt{d}})V_t, \tag{2}$$

where d indicates the feature dimension. Q_t, K_t and V_t are Query, Key and Value matrixes of the self-attention layer, which are based on the original word initial embedding with positing encoding $\widehat{E}$:

$$Q_t = W_t^q \widehat{E}, K = W_t^k \widehat{E}, V = W_t^v \widehat{E}, \tag{3}$$

where W_t^q, W_t^k, and W_t^v are the embedding weights of the self-attention layer.

Based on several affinity matrixes, then compute the context representation H_t^r:

$$\widehat{H}_t^r = LayerNorm(S_t^r + \widehat{E}), \tag{4}$$

$$H_t^r = LayerNorm(\widehat{H}_t^r + Linear(ReLU(Linear(\widehat{H}_t^r)))), \tag{5}$$

where *LayerNorm* is the layer normalization operation and *Linear(ReLU())* is a fully connected position-wise FFN layer.

Finally, the text relationship semantic encoding F_t^r is obtained by applying mean pooling on H_t^r, that is:

$$F_t^r = \frac{1}{N}\sum_{i=1}^{N} H_{t_i}^r, \tag{6}$$

where $H_{t_i}^r$ is the context representation of i-th word.

3.4.2. Multi-Level Video Semantics Encoding

Parallel to multi-level text semantics encoding, we also encode video into global, action, entity and relationship spaces.

(1) Video parsing and initial embedding.

Given a video V, we first uniformly sample N frames, then use the pre-trained feature extraction model to extract general frame-wise features, i.e., $F = [f_1, f_2, ..., f_N]$, here, f_i represents the i-th frame general feature of video V. We then leverage four residual sub-networks to convert the general frame-level features to level-specific frame initial embeddings as follows:

$$F^g = ReLU(F + FC^g(F)), \tag{7}$$

$$F^e = ReLU(F + FC^e(F)), \tag{8}$$

$$F^a = ReLU(F + FC^a(F)), \tag{9}$$

$$F^r = ReLU(F + FC^r(F)), \tag{10}$$

where $F^g = [f_1^g, f_2^g, ..., f_N^g] \in R^{N\times D_v}$, $F^e = [f_1^e, f_2^e, ..., f_N^e] \in R^{N\times D_v}$, $F^a = [f_1^a, f_2^a, ..., f_N^a] \in R^{N\times D_v}$ and $F^r = [f_1^r, f_2^r, ..., f_N^r] \in R^{N\times D_v}$ are frame-wise global, entity, action and relationship initial embeddings, respectively. Here, D_v is the dimension of the extracted visual feature. FC^g, FC^e, FC^a and FC^r are four different fully connected layers.

(2) Video global semantic encoding.

Bi-directional recurrent neural network [49] is an effective structure to handle both past and future contextual information of a given video sequence. We adopt a bidirectional GRU with soft attention (BiGRU-SA) structure to learn video global semantic encoding from frame-wise global initial embedding F^g.

As shown in Figure 5, to extract more salient global semantic and suppress unimportant frames automatically, we first utilize the attention mechanism to refine the F^g, i.e.,

$$\widehat{F}^g = \alpha F^g, \tag{11}$$

where $\widehat{F}^g$ is frame-wise global initial embedding after soft attention, and α is the attention score which is computed as follows:

$$\alpha = softmax(F^g W), \tag{12}$$

where W is the learnable weight matrix.

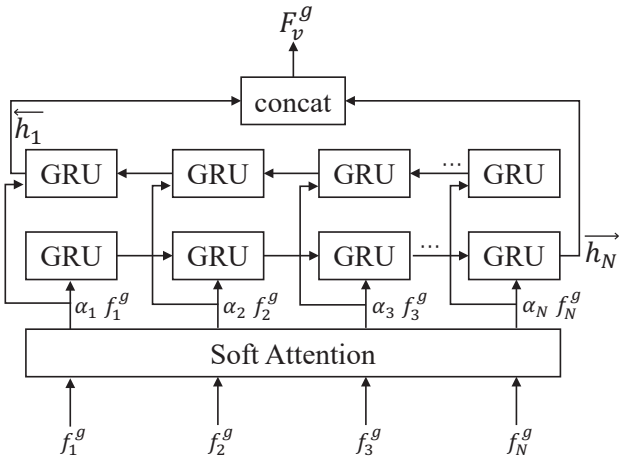


Figure 5. Video global semantic encoding architecture.

Let $\overrightarrow{GRU}$ and $\overleftarrow{GRU}$ indicate the forward and backward GRUs, $\overrightarrow{h}$ and $\overleftarrow{h}$ are corresponding hidden states generated as follows:

$$\overrightarrow{h} = \overrightarrow{GRU}(\widehat{F}^g), \tag{13}$$

$$\overleftarrow{h} = \overleftarrow{GRU}(\widehat{F}^g), \tag{14}$$

Finally, we concatenate the last hidden states of $\overrightarrow{h}$ and $\overleftarrow{h}$ to obtain the final video global semantic encoding:

$$F_v^g = concat(\overrightarrow{h_N}, \overleftarrow{h_1}), \tag{15}$$

where *concat* denotes the concatenate operation, $\overrightarrow{h_N}$ represents the hidden vector at the N-th time step of the forward GRU layer, and $\overleftarrow{h_1}$ represents the hidden vector at the 1-th time step of the backward GRU layer.

(3) Video entity semantic encoding.

Intuitively, the video entity depends on each frame. To be specific, on the one hand, if any frame contains a certain entity means that the entire video also has this entity concept. On the other hand, if the video does not contain a certain entity concept means that all frames do not contain this entity concept. We adapt 1D CNN with kernel size 1 to model this characteristic, as illustrated in Figure 6a, the final video entity semantic encoding F_v^e can be represented as follows:

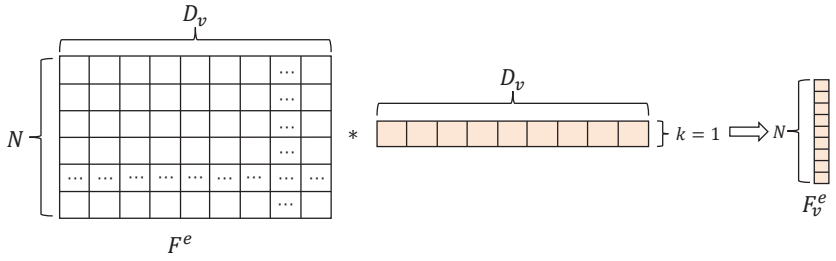
$$F_v^e = ReLU(Conv1d_{k=1}(F^e)), \tag{16}$$

(4) Video action semantic encoding.

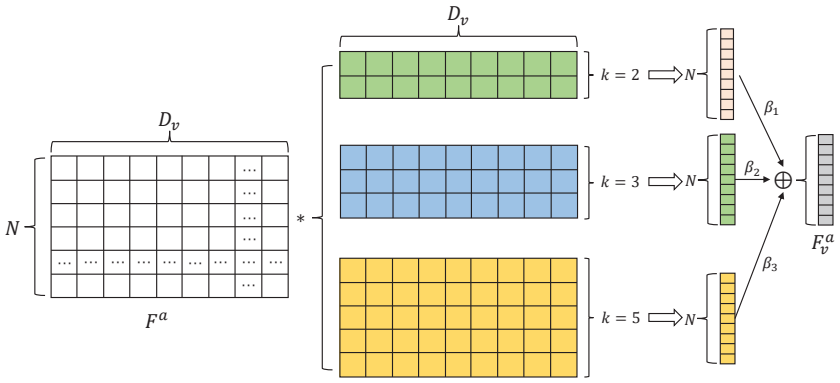
We also adapt 1D CNN to encode video action semantic. However, different from entity, action is motion information, which depends on several continuous video frames. Therefore, we integrate three 1D CNN with different kernel sizes to handle this challenge, the final video action semantic encoding F_v^a can be represented as follows:

$$F_v^a = \beta_1 ReLU(Conv1d_{k=2}(F^e)) + \beta_2 ReLU(Conv1d_{k=3}(F^e)) + \beta_3 ReLU(Conv1d_{k=5}(F^e)), \tag{17}$$

where β_1, β_2, and β_3 serve as the balance factor.



(**a**) Video entity semantic encoding architecture.



(**b**) Video action semantic encoding architecture.

Figure 6. The structures of video entity and action semantic encoding using different 1D CNN.

(5) Video relationship semantic encoding.

Different from entity and action, the relationship semantics associated two objects may appear in discontinuous frames. To address this challenge, consistent with text relation semantic extraction, we also present a self-attention based transformer encoder (as shown in Figure 7) to compute the final video relationship semantic encoding F_v^r from the frame-wise relationship initial embedding F^r.

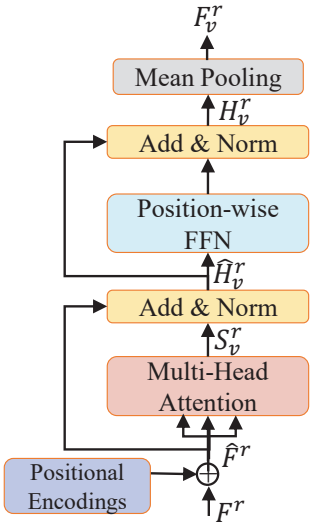


Figure 7. Video relationship semantic encoding architecture.

Specifically, the self-attention layer produces the attention output S_v^r as follows:

$$S_v^r = softmax(\frac{Q_v K_v^\top}{\sqrt{d}})V_v, \tag{18}$$

where d indicates the feature dimension. Q_v, K_v and V_v are query, key and value matrixes of the self-attention layer, which are based on the original frame-wise relationship initial embedding with positing encoding $\widehat{F}^r$:

$$Q_v = W_v^q \widehat{F}^r, K = W_v^k \widehat{F}^r, V = W_v^v \widehat{F}^r, \tag{19}$$

where W_v^q, W_v^k, and W_v^v are the embedding weights of the self-attention layer.

Based on several affinity matrixes, then compute the context representation H_v^r:

$$\widehat{H}_v^r = LayerNorm(S_v^r + \widehat{F}^r), \tag{20}$$

$$H_v^r = LayerNorm(\widehat{H}_v^r + Linear(ReLU(Linear(\widehat{H}_v^r)))), \tag{21}$$

where *LayerNorm* is the layer normalization operation and *Linear(ReLU())* is a fully connected position-wise FFN layer.

Finally, the video relationship semantic encoding F_v^r is obtained by applying mean pooling on H_v^r, that is:

$$F_v^r = \frac{1}{N}\sum_{i=1}^{N} H_{v_i}^r, \tag{22}$$

where $H_{v_i}^r$ is the context representation of ith frame.

3.5. Multi-Level Text and Video Semantics Alignment

In this section, we detail how to project the real-valued semantic vectors extracted from different modalities into a common space and calculate their similarity at different semantic levels.

3.5.1. Global Alignment

In order to make different dimensions in F_v^g and F_t^g fully interactive, we first use Holographic Embeddings [50] to fusion F_v^g and F_t^g, then leverage a fully connected layer (FC) to calculate the similarity of F_v^g and F_t^g, formally:

$$sim(F_v^g, F_t^g) = \sigma(FC(F_v^g \star F_t^g)), \tag{23}$$

where $sim(F_v^g, F_t^g)$ denotes the global semantic similarity, σ is Sigmoid non-linear function which projects the similarity score into $(0, 1)$, $\star$ represents circular correlation operation, i.e., $[a \star b]_k = \sum_{i=0}^{d_f - 1} a_i b_{k+i\%d_f}$, here d_f is feature dimension.

3.5.2. Entity, Action and Relationship Alignment

(1) Statistical distribution constraints.

In order to ensure entity, action and relationship features extracted from different modalities that are projected into a common space, we leverage cross-modal parameter-share attribute projection layer (i.e., FC_*^{share}) in different semantic levels. That is,

$$p_v^e = ReLU(FC_e^{share}(F_v^e)), p_t^e = ReLU(FC_e^{share}(F_t^e)), \tag{24}$$

$$p_v^a = ReLU(FC_a^{share}(F_v^a)), p_t^a = ReLU(FC_a^{share}(F_t^a)), \tag{25}$$

$$p_v^r = ReLU(FC_r^{share}(F_v^r)), p_t^r = ReLU(FC_r^{share}(F_t^r)), \tag{26}$$

where p_v^e, p_v^a and p_v^r are the predicted entity, action and relationship of the video. p_t^e, p_t^a and p_t^r are the predicted entity, action and relationship of the text.

In addition, we also consider minimizing the cross-modal feature distribution in different semantic levels as follows:

$$min(distribution(\mathbf{F}_v^e), distribution(\mathbf{F}_t^e)), \tag{27}$$

$$min(distribution(\mathbf{F}_v^a), distribution(\mathbf{F}_t^a)), \tag{28}$$

$$min(distribution(\mathbf{F}_v^r), distribution(\mathbf{F}_t^r)), \tag{29}$$

where $\mathbf{F}$ denotes semantic representation matrix of a mini-batch.

(2) Similarity calculation.

We use the cosine similarity to measure cross-modal similarity in entity, action and relationship levels, that is:

$$sim(F_v^e, F_t^e) = cos(F_v^e, F_t^e), \tag{30}$$

$$sim(F_v^a, F_t^a) = cos(F_v^a, F_t^a), \tag{31}$$

$$sim(F_v^r, F_t^r) = cos(F_v^r, F_t^r), \tag{32}$$

where $sim(F_v^e, F_t^e)$, $sim(F_v^a, F_t^a)$ and $sim(F_v^r, F_t^r)$ denote the entity, action and relationship semantic similarities of the video V and the sentence T, respectively.

3.6. Training and Inference

3.6.1. Training

Once all similarity scores in four semantic levels are computed, the similarity between the video V and the sentence T is calculated by:

$$sim(V, T) = (sim(F_v^g, F_t^g) + sim(F_v^e, F_t^e) + sim(F_v^a, F_t^a) + sim(F_v^r, F_t^r))/4, \tag{33}$$

where $sim(V, T)$ denotes the overall similarity of the the video V and the sentence T.

We use the triplet ranking loss with hard negative sampling strategy (i.e., L_{rank}) to optimize MCSAN, that is:

$$L_{rank} = max(0, m + sim(V, T^-) - sim(V, T)) + max(0, m + sim(V^-, T) - sim(V, T)), \quad (34)$$

where (V, T) are the positive pair, T^- and V^- are the hardest negatives of text and video in a mini-batch, and m is a positive constant, which we term the margin.

In addition, we also use a novel multi-label regression loss for entity, action and relationship semantic learning. The original video–text retrieval dataset has no attribute annotations. Therefore, annotating entity, action and relationship for each video/text is crucial for learning multi-level representations. To ensure video and text share a common attribute space, we annotate video/text with the same vocabulary (Section 3.3). To be specific, for each sentence, we also use Standford NLP ToolKit to parse its entity, action and relationship as annotations. For each video, we parse all corresponding sentences and take their union as the annotations. Finally, all annotations are labeled in one-hot format based on the vocabulary (that is, entity, action and relationship can be predicted by a multi-label regression manner).

As illustrated in Section 1, the distribution of attributes (i.e., entity, action and relationship) is very imbalanced. Thus, the widely-used binary cross-entropy (BCE) loss may affect the generalization performance of the video–text retrieval model. To address it, inspired by the Focal loss [51], which performed successfully on unbalanced object detection tasks, we propose a novel focal binary cross-entropy (FBCE) loss for entity, action and relationship learning as:

$$\begin{aligned} L_e = &-\frac{1}{K}\sum_{i=1}^{K}(y^e_{v,i}\alpha_e(1-p^e_{v,i})^{\gamma_e}log(p^e_{v,i}) + (1-y^e_{v,i})\alpha_e(1-p^e_{v,i})^{\gamma_e}log(1-p^e_{v,i})) \\ &-\frac{1}{K}\sum_{i=1}^{K}(y^e_{t,i}\alpha_e(1-p^e_{t,i})^{\gamma_e}log(p^e_{t,i}) + (1-y^e_{t,i})\alpha_e(1-p^e_{t,i})^{\gamma_e}log(1-p^e_{t,i})), \end{aligned} \quad (35)$$

$$\begin{aligned} L_a = &-\frac{1}{K}\sum_{i=1}^{K}(y^a_{v,i}\alpha_a(1-p^a_{v,i})^{\gamma_a}log(p^a_{v,i}) + (1-y^a_{v,i})\alpha_a(1-p^a_{v,i})^{\gamma_a}log(1-p^a_{v,i})) \\ &-\frac{1}{K}\sum_{i=1}^{K}(y^a_{t,i}\alpha_a(1-p^a_{t,i})^{\gamma_a}log(p^a_{t,i}) + (1-y^a_{t,i})\alpha_a(1-p^a_{t,i})^{\gamma_a}log(1-p^a_{t,i})), \end{aligned} \quad (36)$$

$$\begin{aligned} L_r = &-\frac{1}{K}\sum_{i=1}^{K}(y^r_{v,i}\alpha_r(1-p^r_{v,i})^{\gamma_r}log(p^r_{v,i}) + (1-y^r_{v,i})\alpha_r(1-p^r_{v,i})^{\gamma_r}log(1-p^r_{v,i})) \\ &-\frac{1}{K}\sum_{i=1}^{K}(y^r_{t,i}\alpha_r(1-p^r_{t,i})^{\gamma_r}log(p^r_{t,i}) + (1-y^r_{t,i})\alpha_r(1-p^r_{t,i})^{\gamma_r}log(1-p^r_{t,i})), \end{aligned} \quad (37)$$

where K is the batch size. $y^e_{v,i}$, $y^a_{v,i}$and $y^r_{v,i}$ are the ground-truth entity, action and relationship of the i-th video. $y^e_{t,i}$, $y^a_{t,i}$and $y^r_{t,i}$ are the ground-truth entity, action and relationship of the i-th text. Please refer to [51] for the details of factors α and γ.

In addition, we leverage Kullback–Leibler divergence loss (KLDivLoss) to constrain that features from different modalities have the same distribution. That is:

$$L_{e\_kl} = KLDivLoss(\mathbf{F}^e_v, \mathbf{F}^e_t), \quad (38)$$

$$L_{a\_kl} = KLDivLoss(\mathbf{F}^a_v, \mathbf{F}^a_t), \quad (39)$$

$$L_{r\_kl} = KLDivLoss(\mathbf{F}^r_v, \mathbf{F}^r_t), \quad (40)$$

Finally, we combine the above losses to obtain the total objective function L_{total} as follows:

$$L_{total} = L_{rank} + \lambda_1(L_e + L_a + L_r) + \lambda_2(L_{e\_kl} + L_{a\_kl} + L_{r\_kl}), \quad (41)$$

where λ_1, λ_2 are trade-off hyper-parameters.

3.6.2. Inference

We simply sort the similarity scores of all video–text pairs for video–text retrieval.

4. Experiments and Results

To justify the effectiveness of our MCSAN model, we carried out extensive experiments under the bidirectional retrieval scenario including (1) text-to-video retrieval, i.e., retrieving videos that are semantically consistent with a given sentence query; and (2) video-to-text retrieval, i.e., retrieving sentences that are semantically consistent with a given video query. In this section, we first give the details about experimental settings, including datasets, evaluation metrics and implementation details. Then, we compare our MCSAN with several state-of-the-art baselines. Next, we explore the effectiveness of each component in our model via ablation studies. Finally, several qualitative and visualization experimental results are reported.

4.1. Experimental Settings

4.1.1. Datasets

We conduct extensive experiments on two most widely-used video–text retrieval datasets, namely, MSR-VTT [27] and VATEX [30], to evaluate our proposed method and several state-of-the-art baselines.

MSR-VTT consists of 10,000 video clips, where each video is described by 20 different sentences. Following the official setting, this dataset is split into 6573 training videos, 497 validation videos, and 2990 testing videos.

VATEX is a large-scale dataset including more than 30,000 video clips, where each video is associated with 10 English and 10 Chinese sentences. Here, only English sentences are used in our experiments. Following the settings in previous work [3,9,11], where 25,991, 1500, 1500 video clips are utilized for training, validation and testing, respectively.

4.1.2. Evaluation Metrics

Following prior cross-modal retrieval works [11,52], rank-based metric, i.e., Recall at K (R@K for short), Sum of all Recalls (SumR for short), are adopted to measure the overall video–text retrieval performance. The higher R@K and SumR indicate that the model has better retrieval performance.

4.1.3. Implementation Details

For a fair comparison with all the compared methods, we apply the same feature as HANet [11]. To be specific, for visual embedding, we use the pre-trained ResNet152 [53] and I3D [54] to extract frame features of MSR-VTT and VATEX, respectively. For text embedding, we use the pre-trained Glove [55] to extract word features of both MSR-VTT and VATEX. Here, the dimension of text embedding is 300 (extracted from Glove), and the dimension of visual embedding are 2048 (extracted from ResNet152) and 1024 (extracted from I3D) for MSR-VTT and VATEX, respectively. We optimized our MCSAN using PyTorch on 1 GeForce RTX 2080Ti GPU. Following HANet [11], the Adam [56] optimizer is employed with learning rate of 1×10^{-4} and the model is trained for 50 epochs with the batch size of 64. The trade-off hyper-parameters λ_1 and λ_2 are empirically set to 0.1 and 0.001, respectively. We estimate the hyperparameter margin m by running a grid search on the corresponding dataset, and the optimal value of m is empirically set to 0.25. In our FBCE loss, following the original Focal loss [51], α_e, α_a and α_r are empirically set to 0.25, γ_e, γ_a and γ_r are empirically set to 2. Note that, the model with the highest SumR value on validation set is deemed as the best model to report the performance on the test set.

4.2. Comparison with State-of-the-Art Methods

We compare our model with two types of baselines: compute the similarity between video and text only using the global features, i.e., VSE [57], VSE++ [58], Mithum et al. [2], W2VV [41], W2VV++ [8], Dual Encoding [3], CE [36], TCE [33], Zhao et al. [59], and compute the similarity between video and text using several level features, i.e., HGR [9], HANet [11], HSL [60]. The experimental results of our proposed MCSAN and all baseline approaches are reported in Tables 1 and 2. Note that all models are trained using the same initial video/text features on the same dataset. From Tables 1 and 2, we can have the following observations:

(1) In the two datasets, methods utilizing multi-level features perform better than these algorithms only use the global feature significantly, indicating that representations extracted from different semantic levels are complementary for video–text retrieval task.

(2) For all video–text retrieval methods, video-to-text has better performance than text-to-video. We think the reason is that on existing video–text retrieval dataset, one video corresponds to multiple sentences; however, one text can only correspond to one video. The results indicate that collecting a new video–text retrieval dataset with proportional numbers of video and text may be a feasible research direction.

(3) The proposed MCSAN outperforms all the baselines on MSR-VTT and VATEX datasets. The results demonstrate that the proposed model can jointly model multi-modal context information and multi-level semantics of both video and text in a unified deep model, which can better capture the underlying complementary representation of video/text, so as to improve the performance of video–text retrieval.

Table 1. Comparisons with the state-of-the-art video–text retrieval methods on the MSR-VTT dataset. The best performance are marked in bold. Our proposed model performs the best.

Method	Text-to-Video			Video-to-Text			SumR
	R@1	R@5	R@10	R@1	R@5	R@10	
VSE [57]	5.0	16.4	24.6	7.7	20.3	31.2	105.2
VSE++ [58]	5.7	17.1	24.8	10.2	25.4	35.1	118.3
Mithum et al. [2]	5.8	17.6	25.2	10.5	26.7	35.9	121.7
W2VV [41]	6.1	18.7	27.5	11.8	28.9	39.1	132.1
Dual Encoding [3]	7.7	22.0	31.8	13.0	30.8	43.3	148.6
TCE [33]	7.7	22.5	32.1	-	-	-	-
Zhao et al. [59]	8.8	25.5	36.5	14.0	33.1	44.9	162.8
HGR [9]	9.2	26.2	36.5	15.0	36.7	48.8	172.4
HANet [11]	9.3	27.0	38.1	16.1	39.2	52.1	181.8
HSL [60]	10.8	29.2	38.5	20.3	45.1	55.9	201.7
MCSAN	**10.9**	**29.6**	**40.1**	**21.1**	**46.8**	**57.4**	**205.3**

Table 2. Comparisons with the state-of-the-art video–text retrieval methods on the VATEX dataset. The best performance are marked in bold. The overall performance are indicated by Sum of Recalls (SumR). MCSAN is the best.

Method	Text-to-Video			Video-to-Text			SumR
	R@1	R@5	R@10	R@1	R@5	R@10	
W2VV [41]	14.6	36.3	46.1	39.6	69.5	79.4	285.5
VSE++ [58]	31.3	65.8	76.4	42.9	73.9	83.6	373.9
CE [36]	31.1	68.7	80.2	41.3	71.0	82.3	374.6
W2VV++ [8]	32.0	68.2	78.8	41.8	75.1	84.3	380.2
Dual Encoding [3]	31.1	67.4	78.9	-	-	-	-
HGR [9]	35.1	73.5	83.5	-	-	-	-
HSL [60]	**36.8**	73.6	83.7	46.8	75.7	85.1	401.7
HANet [11]	36.4	74.1	84.1	49.1	79.5	86.2	409.4
MCSAN	36.6	**74.2**	**84.1**	**51.4**	**81.3**	**86.8**	**411.8**

4.3. Ablation Studies

Because the proposed MCSAN contains multiple key components, we additional compare variants of MCSAN with respect to the following perspectives to verify the effectiveness of each component in MCSAN: (1) the effect of the relationship information, (2) the impact of the statistical distribution constraints, and (3) the effect of the proposed FBCE loss for handling unbalanced attributes. The following MCSAN variants are designed for comparison. The following MCSAN variants are designed for comparison.

(1) MCSAN$^{\neg}$R: A variant of MCSAN with the relationship encoding level being removed.

(2) MCSAN$^{\neg}$S: A variant of MCSAN with the statistical distribution constraints being removed.

(3) MCSAN$^{\neg}$L: A variant of MCSAN with the FBCE loss being removed, and only using the conventional BCE loss for entity, action and relationship learning.

The ablation study results are shown in Table 3. We can observe that the combination of all three components has the best performance and the performance will degrade if remove any components. The results demonstrate that all components in our MCSAN are effective. Moreover, we can also observe that if the relationship encoding is removed, the performance will degrade the most, which indicates that explicitly relationship understanding is crucial for video–text retrieval.

Table 3. Results of comparison among different variants in MCSAN on MSR-VTT dataset. The best performance are marked in bold. MCSAN exploiting all the four semantic levels (global, entity, action and realtionship) is the best.

Method	Text-to-Video			Video-to-Text			SumR
	R@1	R@5	R@10	R@1	R@5	R@10	
MCSAN$^{\neg}$R	9.4	27.1	37.9	17.2	42.3	54.7	195.8
MCSAN$^{\neg}$S	10.1	27.8	39.8	19.9	45.9	57.0	202.6
MCSAN$^{\neg}$L	10.6	28.8	39.2	20.8	46.4	56.6	203.7
MCSAN	**10.9**	**29.6**	**40.1**	**21.1**	**46.8**	**57.4**	**205.3**

4.4. Qualitative Analyses

We visualize three text-to-video retrieval examples on the MSR-VTT test set in Figure 8. For each query sentence, Figure 8 shows the top-3 scored videos evaluated by our MCSAN. We find that the proposed model is able to recognize the relationship "*man at top speed*" since comprehensive relationship between concepts is explicitly modeled in our framework. The results demonstrate that the proposed method can handle different types of video/text semantics well.

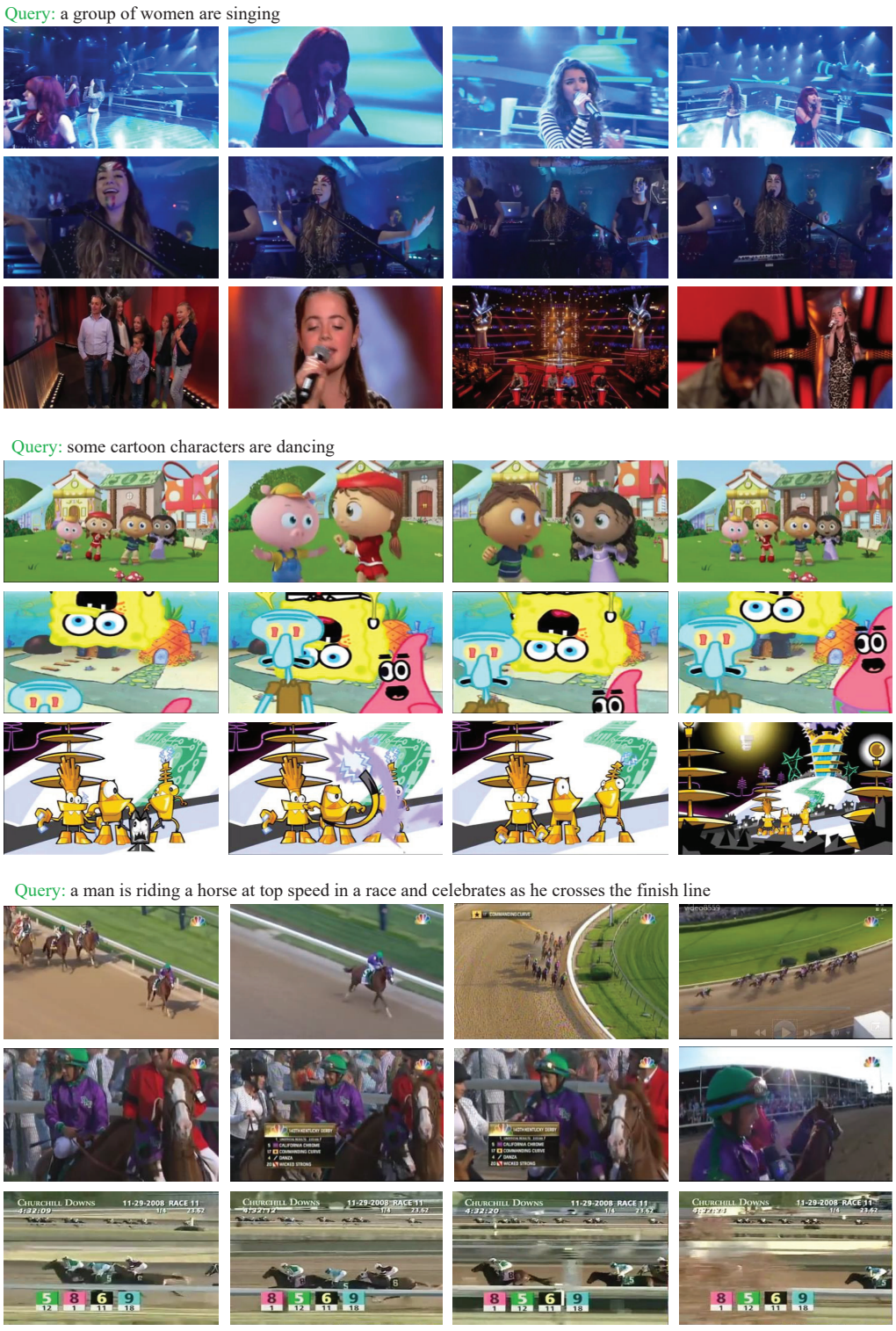


Figure 8. Top 3 text-to-video retrieval examples on the MSR-VTT test set.

4.5. Limitations

Although we achieve good results in the previous experiments, there are still some inevitable problems. Note that Rao et al. [45] demonstrated that modality embedding is crucial for the cross modal retrieval performance improvement. However, only simple CNN/Transformer structures are adopted in this paper. Thus, we will explore a more effective way to extract text/video semantic features in the future. Moreover, we will try to leverage additional knowledge information (e.g., knowledge embedding learnt from

the external sentence corpus) since it can provide useful complementary information for video–text retrieval.

5. Conclusions

In this paper, we propose a novel Multi-Level Cross-Modal Semantic Alignment Network (MCSAN) for video–text retrieval tasks. The semantic similarity between video and text is fully considered at four semantic levels (global, action, entity and relationship) by carefully designed spatial–temporal semantic learning structures and cross-modal representation distribution constraints in an end-to-end framework. We make the first attempt to handle the cross-modal relationship semantic consistency without relying on any extern corpus in video–text retrieval task. Moreover, we introduce a novel focal binary cross-entropy (FBCE) loss function, which also is the first effort to model the unbalanced attribute distribution problem for video–text retrieval. We conduct extensive experiments on two video–text retrieval benchmarks, i.e., MSR-VTT and VATEX. The quantitative and qualitative experimental results demonstrate the effectiveness of the proposed framework. We achieve state-of-the-art performance against existing non-VLP-based video–text retrieval frameworks. Due to that most recent VLP based video–text retrieval algorithms are not publicly available, we hope that future work in the field of video–text retrieval should pay more attention to the reproducibilities and capabilities of the model, rather than use all the tricks to report a hard-to-reproduce better performance.

Author Contributions: Conceptualization, methodology, and writing—original draft preparation, F.N.; writing—review and editing, Y.G.; visualization, L.D.; funding acquisition, Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation (NSF) of China (No. 61902104), the Anhui Provincial Key Research and Development Program (No. 2022a05020042, No. 2022a05020043), the Anhui Provincial Natural Science Foundation (No. 2008085QF295, No. 2008085QF316), the University Natural Sciences Research Project of Anhui Province (No. KJ2020A0651), New Energy Vehicle and Intelligent Networked Vehicle Innovation Project (No. wfgcyh2020477), and Anhui International Joint Research Center for Ancient Architecture Intellisencing and Multi-Dimensional Modeling (No. GJZZX2021KF01).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available in a publicly accessible repository. The data presented in this study are openly available at https://github.com/Roc-Ng/HANet, accessed on 26 July 2021.

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

References

1. Kaur, P.; Pannu, H.S.; Malhi, A.K. Comparative analysis on cross-modal information retrieval: A review. *Comput. Sci. Rev.* **2021**, *39*, 100336. [CrossRef]
2. 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, Yokohama, Japan, 11–14 June 2018; pp. 19–27.
3. 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/CVF Conference On Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 9346–9355.
4. Wang, W.; Gao, J.; Yang, X.; Xu, C. Learning coarse-to-fine graph neural networks for video–text retrieval. *IEEE Trans. Multimed.* **2020**, *23*, 2386–2397. [CrossRef]
5. Jin, W.; Zhao, Z.; Zhang, P.; Zhu, J.; He, X.; Zhuang, Y. Hierarchical cross-modal graph consistency learning for video–text retrieval. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, 11–15 July 2021; pp. 1114–1124.
6. Gorti, S.K.; Vouitsis, N.; Ma, J.; Golestan, K.; Volkovs, M.; Garg, A.; Yu, G. X-Pool: Cross-Modal Language-Video Attention for Text-Video Retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 21–24 June 2022; pp. 5006–5015.
7. Feng, Z.; Zeng, Z.; Guo, C.; Li, Z. Exploiting visual semantic reasoning for video–text retrieval. *arXiv* **2020**, arXiv:2006.08889.

8. Li, X.; Xu, C.; Yang, G.; Chen, Z.; Dong, J. W2vv++ fully deep learning for ad-hoc video search. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 1786–1794.
9. Chen, S.; Zhao, Y.; Jin, Q.; Wu, Q. Fine-grained video–text retrieval with hierarchical graph reasoning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 10638–10647.
10. Wray, M.; Larlus, D.; Csurka, G.; Damen, D. Fine-grained action retrieval through multiple parts-of-speech embeddings. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 450–459.
11. Wu, P.; He, X.; Tang, M.; Lv, Y.; Liu, J. HANet: Hierarchical Alignment Networks for Video–text Retrieval. In Proceedings of the 29th ACM international conference on Multimedia, Virtual Event, 20–24 October 2021; pp. 3518–3527.
12. Reddy, M.D.M.; Basha, M.S.M.; Hari, M.M.C.; Penchalaiah, M.N. Dall-e: Creating images from text. *UGC Care Group I J.* **2021**, *8*, 71–75.
13. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning (PMLR), Virtual Event, 18–24 July 2021; pp. 8748–8763.
14. Wei, L.; Xie, L.; Zhou, W.; Li, H.; Tian, Q. MVP: Multimodality-guided Visual Pre-training. *arXiv* **2022**, arXiv:2203.05175.
15. Yang, Z.; Garcia, N.; Chu, C.; Otani, M.; Nakashima, Y.; Takemura, H. Bert representations for video question answering. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 1556–1565.
16. Tang, M.; Wang, Z.; Liu, Z.; Rao, F.; Li, D.; Li, X. Clip4caption: Clip for video caption. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, 20–24 October 2021; pp. 4858–4862.
17. Wang, Z.; Codella, N.; Chen, Y.C.; Zhou, L.; Yang, J.; Dai, X.; Xiao, B.; You, H.; Chang, S.F.; Yuan, L. CLIP-TD: CLIP Targeted Distillation for Vision-Language Tasks. *arXiv* **2022**, arXiv:2201.05729.
18. Luo, J.; Li, Y.; Pan, Y.; Yao, T.; Chao, H.; Mei, T. CoCo-BERT: Improving video-language pre-training with contrastive cross-modal matching and denoising. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, 20–24 October 2021; pp. 5600–5608.
19. Fang, H.; Xiong, P.; Xu, L.; Chen, Y. Clip2video: Mastering video–text retrieval via image clip. *arXiv* **2021**, arXiv:2106.11097.
20. Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; Li, T. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv* **2021**, arXiv:2104.08860.
21. Gao, Z.; Liu, J.; Chen, S.; Chang, D.; Zhang, H.; Yuan, J. Clip2tv: An empirical study on transformer-based methods for video–text retrieval. *arXiv* **2021**, arXiv:2111.05610.
22. Nian, F.; Bao, B.K.; Li, T.; Xu, C. Multi-modal knowledge representation learning via webly-supervised relationships mining. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 411–419.
23. Yao, T.; Pan, Y.; Li, Y.; Mei, T. Exploring visual relationship for image captioning. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 684–699.
24. Wang, S.; Wang, R.; Yao, Z.; Shan, S.; Chen, X. Cross-modal scene graph matching for relationship-aware image-text retrieval. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 1508–1517.
25. Shvetsova, N.; Chen, B.; Rouditchenko, A.; Thomas, S.; Kingsbury, B.; Feris, R.S.; Harwath, D.; Glass, J.; Kuehne, H. Everything at Once-Multi-Modal Fusion Transformer for Video Retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 21–24 June 2022; pp. 20020–20029.
26. Wray, M.; Doughty, H.; Damen, D. On semantic similarity in video retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 3650–3660.
27. Xu, J.; Mei, T.; Yao, T.; Rui, Y. Msr-vtt: A large video description dataset for bridging video and language. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5288–5296.
28. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Philip, S.Y. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4–24. [CrossRef]
29. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December2017; pp. 6000–6010.
30. Wang, X.; Wu, J.; Chen, J.; Li, L.; Wang, Y.F.; Wang, W.Y. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 4581–4591.
31. Li, T.; Ni, B.; Xu, M.; Wang, M.; Gao, Q.; Yan, S. Data-driven affective filtering for images and videos. *IEEE Trans. Cybern.* **2015**, *45*, 2336–2349. [CrossRef]
32. Nian, F.; Li, T.; Wu, X.; Gao, Q.; Li, F. Efficient near-duplicate image detection with a local-based binary representation. *Multimed. Tools Appl.* **2016**, *75*, 2435–2452. [CrossRef]
33. Yang, X.; Dong, J.; Cao, Y.; Wang, X.; Wang, M.; Chua, T.S. Tree-augmented cross-modal encoding for complex-query video retrieval. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Xi'an, China, 25–30 July 2020; pp. 1339–1348.

34. Dong, J.; Wang, Y.; Chen, X.; Qu, X.; Li, X.; He, Y.; Wang, X. Reading-strategy inspired visual representation learning for text-to-video retrieval. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 5680–5694. [CrossRef]
35. Gabeur, V.; Sun, C.; Alahari, K.; Schmid, C. Multi-modal transformer for video retrieval. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 214–229.
36. Liu, Y.; Albanie, S.; Nagrani, A.; Zisserman, A. Use what you have: Video retrieval using representations from collaborative experts. *arXiv* **2019**, arXiv:1907.13487.
37. Miech, A.; Laptev, I.; Sivic, J. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv* **2018**, arXiv:1804.02516.
38. Gabeur, V.; Nagrani, A.; Sun, C.; Alahari, K.; Schmid, C. Masking modalities for cross-modal video retrieval. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January2022; pp. 1766–1775.
39. Song, X.; Chen, J.; Wu, Z.; Jiang, Y.G. Spatial–temporal graphs for cross-modal text2video retrieval. *IEEE Trans. Multimed.* **2021**, *14*, 2914–2923. [CrossRef]
40. Miech, A.; Zhukov, D.; Alayrac, J.B.; Tapaswi, M.; Laptev, I.; Sivic, J. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 2630–2640.
41. Dong, J.; Li, X.; Snoek, C.G. Predicting visual features from text for image and video caption retrieval. *IEEE Trans. Multimed.* **2018**, *20*, 3377–3388. [CrossRef]
42. Dong, J.; Long, Z.; Mao, X.; Lin, C.; He, Y.; Ji, S. Multi-level alignment network for domain adaptive cross-modal retrieval. *Neurocomputing* **2021**, *440*, 207–219. [CrossRef]
43. Yu, Z.; Yu, J.; Cui, Y.; Tao, D.; Tian, Q. Deep modular co-attention networks for visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 6281–6290.
44. Yu, J.; Li, J.; Yu, Z.; Huang, Q. Multimodal transformer with multi-view visual representation for image captioning. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 4467–4480. [CrossRef]
45. Rao, J.; Wang, F.; Ding, L.; Qi, S.; Zhan, Y.; Liu, W.; Tao, D. Where Does the Performance Improvement Come From?—A Reproducibility Concern about Image-Text Retrieval. *arXiv* **2022**, arXiv:2203.03853.
46. Manning, C.D.; Surdeanu, M.; Bauer, J.; Finkel, J.R.; Bethard, S.; McClosky, D. The Stanford CoreNLP natural language processing toolkit. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, MD, USA, 23–24 June 2014; pp. 55–60.
47. Reimers, N.; Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv* **2019**, arXiv:1908.10084.
48. Schlichtkrull, M.; Kipf, T.N.; Bloem, P.; Berg, R.v.d.; Titov, I.; Welling, M. Modeling relational data with graph convolutional networks. In Proceedings of the European Semantic Web Conference, Heraklion, Greece, 3–7 June 2018; pp. 593–607.
49. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [CrossRef]
50. Nickel, M.; Rosasco, L.; Poggio, T. Holographic embeddings of knowledge graphs. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; pp. 1955–1961.
51. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
52. Qu, L.; Liu, M.; Wu, J.; Gao, Z.; Nie, L. Dynamic modality interaction modeling for image-text retrieval. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual, 11–15 July 2021; pp. 1104–1113.
53. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
54. Carreira, J.; Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
55. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
56. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
57. Kiros, R.; Salakhutdinov, R.; Zemel, R.S. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv* **2014**, arXiv:1411.2539.
58. Faghri, F.; Fleet, D.J.; Kiros, J.R.; Fidler, S. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv* **2017**, arXiv:1707.05612.
59. Zhao, R.; Zheng, K.; Zha, Z.J. Stacked convolutional deep encoding network for video–text retrieval. In Proceedings of the 2020 IEEE International Conference on Multimedia and Expo (ICME), Virtual, 6–10 July 2020; pp. 1–6.
60. Dong, J.; Li, X.; Xu, C.; Yang, X.; Yang, G.; Wang, X.; Wang, M. Dual encoding for video retrieval by text. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 4065–4080. [CrossRef]