Fine-grained Video-Text Retrieval with Hierarchical Graph Reasoning

Shizhe Chen[1]*, Yida Zhao[1], Qin Jin[1]†, Qi Wu[2]
[1]School of Information, Renmin University of China
[2]Australian Centre for Robotic Vision, University of Adelaide
{cszhe1, zyiday, qjin}@ruc.edu.cn, qi.wu01@adelaide.edu.au

Abstract

Cross-modal retrieval between videos and texts has attracted growing attentions due to the rapid emergence of videos on the web. The current dominant approach is to learn a joint embedding space to measure cross-modal similarities. However, simple embeddings are insufficient to represent complicated visual and textual details, such as scenes, objects, actions and their compositions. To improve fine-grained video-text retrieval, we propose a Hierarchical Graph Reasoning (HGR) model, which decomposes video-text matching into global-to-local levels. The model disentangles text into a hierarchical semantic graph including three levels of events, actions, entities, and generates hierarchical textual embeddings via attention-based graph reasoning. Different levels of texts can guide the learning of diverse and hierarchical video representations for cross-modal matching to capture both global and local details. Experimental results on three video-text datasets demonstrate the advantages of our model. Such hierarchical decomposition also enables better generalization across datasets and improves the ability to distinguish fine-grained semantic differences. Code will be released at https://github.com/cshizhe/hgr_v2t.

1. Introduction

The rapid emergence of videos on the Internet such as on YouTube and TikTok has brought great challenges to accurate retrieval of video contents. Traditional retrieval methods [2, 4, 12] are mainly based on keyword search. However, since keywords are limited and unstructured, it is difficult to retrieve various fine-grained contents, such as a compositional event "a white dog is chasing a cat". To address the limitation of keyword-based approach, more and more researchers are paying attention to video retrieval using natural language texts that contain richer and more structured details than keywords, a.k.a, cross-modal video-text retrieval [6, 28, 43].

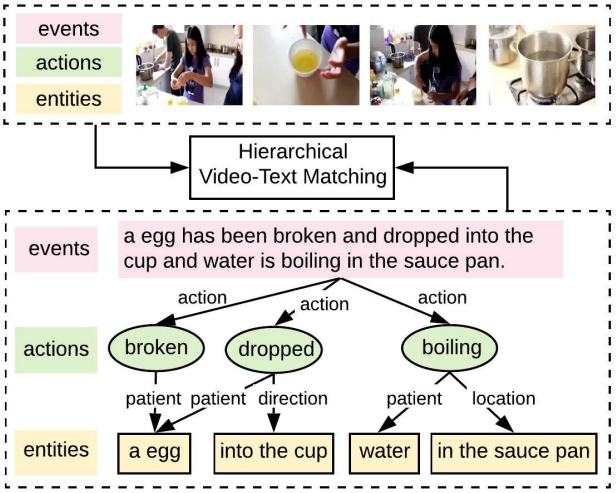

Figure 1. We factorize video-text matching into hierarchical levels including events, actions, and entities to form a global to local structure. On one hand, this enhances global matching with the help of detailed semantic components, on the other hand, it improves local matching with the help of global event structure.

The current dominant approach for cross-modal retrieval is to encode different modalities into a joint embedding space [9] to measure cross-modal similarities, which can be broadly classified into two categories. The first type of works [6, 27, 28] embeds videos and texts into global vectors. Despite of high efficiency, such global representation is hard to capture fine-grained semantic details. For example, understanding the video and text in Figure 1 involves complicated reasoning about different actions (break, drop, boil), entities (egg, into the cup *etc*.) as well as how all components compose to the event ('egg' is the patient of action 'break' and 'into the cup' is the direction). To avoid losing those details, another type of methods [32, 43] employs a sequence of frames and words to represent videos and texts respectively and aligns local components to compute overall similarities. Although these approaches have achieved improved performance for image-text retrieval [19, 22], learn-

*This work was partially performed while Shizhe Chen was visiting University of Adelaide.
†Qin Jin is the corresponding author.

ing semantic alignments between videos and texts is more challenging since video-text pairs are more weakly supervised than image-text pairs. Moreover, such sequential representations neglect topological structures in videos and texts, making it hard to understand relations between local components within an event.

In this work, we propose a Hierarchical Graph Reasoning (HGR) model which takes the advantage of above global and local approaches and makes up their deficiencies. As shown in Figure 1, we decompose video-text matching into three hierarchical semantic levels, which are responsible to capture global events, local actions and entities respectively. On the text side, the global event is represented by the whole sentence, actions are denoted by verbs and entities refer to noun phrases. We build a semantic role graph across levels to capture how local components composite an event and propose an attention-based graph reasoning method to generate hierarchical textual embeddings. Different levels of text are used to guide video encoding into corresponding hierarchical embeddings to distinguish different aspects in videos. We align cross-modal components at each semantic level via attention mechanisms to facilitate matching in weakly-supervised condition. Matching scores from all three levels are aggregated together in order to enhance fine-grained semantic coverage.

We carry out extensive experiments on three video-text datasets. Consistent improvements over previous approaches demonstrate the effectiveness of our proposed model. The hierarchical decomposition also enables better generalization ability in cross-dataset evaluation. To further evaluate fine-grained retrieval ability, we propose a new binary selection task [15, 16] which requires systems to select correct matching sentence for a given video from two similar sentences with subtle difference. Our model achieves better performance to recognize fine-grained semantic changes and prefers more comprehensive video descriptions due to the fusion of hierarchical matchings.

The contributions of this work are as follows:

- We propose a Hierarchical Graph Reasoning (HGR) model for fine-grained video-text retrieval, which decomposes video-text matching into global-to-local levels. It improves global matching with the help of detailed semantics and local matching with the help of global event structures.
- The three disentangled levels in texts (event, actions and entities) interact with each other via attention-based graph reasoning and align with corresponding levels of videos for cross-modal matching.
- The HGR model achieves improved performance on different video-text datasets and better generalization ability on unseen dataset. A new binary selection task further demonstrates the ability of our model to distinguish fine-grained semantic differences.

2. Related Works

Image-Text Matching. Most of previous works [7, 9, 10, 18, 21] for image-text matching encode images and sentences as fix-dimensional vectors in a common latent space for similarity measure. Frome *et al*. [9] firstly propose the joint embedding framework for images and words, and train the model with contrastive ranking loss. Kiros *et al*. [21] extend the framework to match images and sentences with CNN to encode images and RNN for sentences. Faghri *et al*. [7] improve training strategy with hard negative mining. Huang *et al*. [18] and Gu *et al*. [10] explore reconstructions in multi-task framework to enrich global representations. However, it is hard to cover complicated semantics only using fixed-dimensional vectors. Therefore, Karpathy *et al*. [19] decompose image and sentences as multiple regions and words, and propose using maximum alignment to compute global matching similarity. Lee *et al*. [22] improve the alignment with stacked cross-attention. Wu *et al*. [40] factorize image descriptions into objects, attributes, relations and sentences, however, they do not consider interactions across levels and the decomposition might not be optimal for video descriptions that focus on actions and events.

Video-Text Matching. Though sharing similarities with image-text matching, the video-text matching task is more challenging because videos contain multi-modalities and spatial-temporal evolution [3, 26, 28]. Mithun *et al*. [28] and Liu *et al*. [27] employ multimodal cues such as image, motion and audio for video encoding. To encode sequential videos and texts, Dong *et al*. [6] utilize three branches, *i.e.* mean pooling, biGRU and CNN to encode them. Yu *et al*. [43] propose a joint sequence fusion model for sequential interaction of videos and texts. Song *et al*. [32] employ multiple diverse representations for videos and texts for the polysemous problem. Chen *et al*. [3] tackles the weakly-supervised spatial-temporal grounding in videos. The most similar work to ours is Wray *et al*. [39] and Zhang *et al*. [44]. Wray *et al*. [39] disentangles action phrases into verbs and nouns for fine-grained action retrieval, which however is hard to apply on sentences with more complicated compositions. Zhang *et al*. [44] propose hierarchical modeling of videos and paragraphs, but are not applicable to decompose single sentences. Therefore, in this work we propose to decompose a sentence as a hierarchical semantic graph and integrate video-text matching at different levels.

Graph-based Reasoning. Graph convolutional network (GCN) [20] is firstly proposed for graph recognition, which employs convolution on neighbourhoods of nodes. Graph attention networks [34] are further introduced to dynamically attend over neighborhoods' features. In order to model graphs with different edge types, relational GCN [30] is proposed to learn specific contextual transformation for each relation type. The graph-based reasoning has great ap-

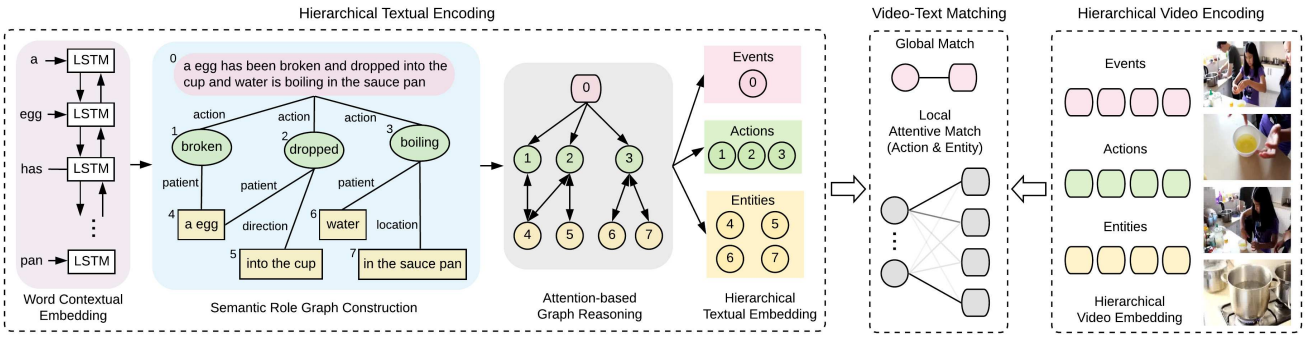

Figure 2. Overview of the proposed Hierarchical Graph Reasoning (HGR) model for cross-modal video-text retrieval.

plications in computer vision tasks such as action recognition [33, 37], scene graph generation [42], referring expression grounding [23, 36], visual question answering [17, 24] etc. Most of them [17, 23, 24, 36, 42] apply graph reasoning on image regions to learn visual relationships. In this work, we focus on reasoning over hierarchical graph structures on video descriptions for fine-grained video-text matching.

3. Hierarchical Graph Reasoning Model

Figure 2 illustrates the overview of the HGR model which consists of three blocks: 1) hierarchical textual encoding (Section 3.1) that constructs semantic role graphs from texts and applies graph reasoning to obtain hierarchical text representations; 2) hierarchical video encoding (Section 3.2) that maps videos into corresponding multi-level representations; and 3) video-text matching (Section 3.3) which aggregates global and local matchings at different levels to compute overall cross-modal similarities.

3.1. Hierarchical Textual Encoding

Video descriptions naturally contain hierarchical structures. The overall sentence describes the global event in the video which is composed of multiple actions in temporal dimensions, and each action is composed of different entities as its arguments such as agent and patient of the action. Such global-to-local structure is beneficial for accurate and comprehensive understanding of the semantic meanings of video descriptions. Therefore, in this section, we introduce how to obtain hierarchical textual representations from a video description in a global-to-local topology.

Semantic Role Graph Structure. Given a video description C that consists of N words $\{c_1, \cdots, c_N\}$, we consider C as a global event node in the hierarchical graph. Then we employ an off-the-shelf semantic role parsing toolkit [31] to obtain verbs, noun phrases in C as well as the semantic role of each noun phrase to the corresponding verb (details of semantic roles are given in the supplementary). The verbs are considered as action nodes and connected to event node with direct edges, so that temporal relations of different actions can be implicitly learned from event node in following graph reasoning. The noun phrases are entity nodes that are connected with different action nodes. The edge type r_{ij} from entity node i to action node j is decided by the semantic role of the entity in reference to the action, while the edge type r_{ji} from action node j to any entity node i is unified as an action type for simplicity. If an entity node serves multiple semantic roles to different action nodes, we duplicate the entity node for each semantic role. Such semantic role relations are important to understand the event structure, for example, "a dog chasing a cat" is apparently different from "a cat chasing a dog" which only switches semantic roles of the two entities. In the left side of Figure 2, we present an example of the constructed graph.

Initial Graph Node Representation. We embed semantic meaning of each node into a dense vector as initialization. For the global event node, we aim to summarize the salient event described in the sentence. Therefore, we first utilize an bidirectional LSTM (Bi-LSTM) [14] to generate a sequence of contextual-aware word embeddings $\{w_1, \cdots, w_N\}$ as follows:

$$\overrightarrow{w}_i = \overrightarrow{\text{LSTM}}(W_c c_i, \overrightarrow{w}_{i-1}; \overrightarrow{\theta}) \quad (1)$$

$$\overleftarrow{w}_i = \overleftarrow{\text{LSTM}}(W_c c_i, \overleftarrow{w}_{i+1}; \overleftarrow{\theta}) \quad (2)$$

$$w_i = (\overrightarrow{w}_i + \overleftarrow{w}_i)/2 \quad (3)$$

where W_c is word embedding matrix, $\overrightarrow{\theta}$ and $\overleftarrow{\theta}$ are parameters in the two LSTMs. Then we average the word embeddings via an attention mechanism that focuses on important words in the sentence as the global event embedding g_e:

$$g_e = \sum_{i=1}^{N} \alpha_{e,i} w_i \quad (4)$$

$$\alpha_{e,i} = \frac{\exp(W_e w_i)}{\sum_{j=1}^{N} \exp(W_e w_j)} \quad (5)$$

where W_e is the parameter to be learned. For action and entity nodes, though different LSTMs can be employed to only encode their constitutive words independently, since

semantic role parsing might separate words with mistakes, contextual word representations can be beneficial to resolve such negative influences. Therefore, we reuse the above Bi-LSTM word embeddings w_i and apply max pooling over words in each node as action node representations $g_a = \{g_{a,1}, \cdots, g_{a,N_a}\}$ and entity node representations $g_o = \{g_{o,1}, \cdots, g_{o,N_o}\}$, where N_a and N_o are numbers of action and entity nodes respectively.

Attention-based Graph Reasoning. The connections across different levels in the constructed graph not only explain how local nodes compose the global event, but also are able to reduce ambiguity for each node. For example, the entity "egg" in Figure 2 can have diverse appearances without context, but the context from action "break" constrains its semantics, so that it should have high similarity with visual appearance of a "broken egg" rather than a "round egg". Therefore, we propose to reason over interactions in the graph to obtain hierarchical textual representations.

Since edges in our graph are of different semantic roles, a straightforward approach to model interactions in graph is to utilize relational GCN [30], which requires to learn separate transformation weight matrix for each semantic role. However, it can lead to rapid growth of parameters, which makes it inefficient to learn from limited amount of video-text data and prone to over-fitting on rare semantic roles.

To address this problem, we propose to factorize multi-relational weights in GCN into two parts: a common transformation matrix $W_t \in \mathbb{R}^{D\times D}$ that is shared for all relationship types and a role embedding matrix $W_r \in \mathbb{R}^{D\times K}$ that is specific for different semantic roles, where D is the dimension of node representation and K is the number of semantic roles. For inputs to the first GCN layer, we multiply initialized node embeddings $g_i \in \{g_e, g_a, g_o\}$ with their corresponding semantic roles as:

$$g_i^0 = g_i \odot W_r r_{ij} \tag{6}$$

where r_{ij} is an one-hot vector denoting the edge type from node i to j. Suppose g_i^l is the output representation of node i at l-th GCN layer, we employ a graph attention network to select relevant contexts from neighbor nodes to enhance the representation for each node:

$$\tilde{\beta}_{ij} = (W_a^q g_i^l)^T (W_a^k g_j^l)/\sqrt{D} \tag{7}$$

$$\beta_{ij} = \frac{\exp(\tilde{\beta}_{ij})}{\sum_{j\in\mathcal{N}_i}\exp(\tilde{\beta}_{ij})} \tag{8}$$

where $\mathcal{N}_i$ is neighborhood nodes of node i, W_a^k and W_a^q are parameters to compute graph attention. Then the shared W_t is utilized to transform contexts from attended nodes to node i with residual connection:

$$g_i^{l+1} = g_i^l + W_t^{l+1}\sum_{j\in\mathcal{N}_i}(\beta_{ij} g_j^l) \tag{9}$$

Putting together Eq (6) and Eq (9), we can see that the transformation from nodes in lower layer is specific for different semantic role edges. Take the first GCN layer as an example, the computation is as follows:

$$g_i^1 = g_i^0 + \sum_{j\in\mathcal{N}_i}(\beta_{ij}(W_t^1 \odot W_r r_{ij}) g_j) \tag{10}$$

where $\odot$ is element-wise multiplication with broadcasting, $W_t^1 \odot W_r r_{ij}$ is the edge specific transformation at layer 1. In this way, we significantly reduce the size of parameters from $L \times K \times D \times D$ to $L \times D \times D + K \times D$ where L is the number of layers of GCN, but still maintain role-awareness when reasoning over graph. The outputs from the L-th GCN layer are our final hierarchical textual representations, which are denoted as c_e for global event node, c_a for action nodes and c_o for entity nodes.

3.2. Hierarchical Video Encoding

Videos also contain multiple aspects such as objects, actions and events. However, it is challenging to directly parse video into hierarchical structures as in texts which requires temporal segmentation, object detection, tracking and so on. We thus build three independent video embeddings instead to focus on different level of aspects in the video.

Given video V as a sequence of frame-wise features $\{f_1, \cdots, f_M\}$, we utilize different weights W_e^v, W_a^v and W_o^v to encode videos into three level of embeddings:

$$v_{x,i} = W_x^v f_i, \quad x \in \{e, a, o\} \tag{11}$$

For the global event level, we employ the attention mechanism similar to Eq (4) to obtain one global vector to represent the salient event in the video as v_e. And for the action and entity level, the video representations are a sequence of frame-wise features $v_a = \{v_{a,1}, \cdots, v_{a,M}\}$ and $v_o = \{v_{o,1}, \cdots, v_{o,M}\}$ respectively. These features will be sent to the following matching module to match with their corresponding textual features at different levels, which guarantees different transformation weights can be learned to focus on different level video information with the guidance of corresponding textual representation.

3.3. Video-Text Matching

In order to cover both local and global semantics to match videos and texts, we aggregate results from the three hierarchical levels for the overall cross-modal similarity.

Global Matching. At the global event level, the video and text are encoded into global vectors that capture salient event semantics with attention mechanism. Therefore, we simply utilize cosine similarity $cos(v, c) \equiv \frac{v^T c}{||v||||c||}$ to measure the cross-modal similarity for global video and text contents. The global matching score is $s_e = cos(v_e, c_e)$.

Table 1. Cross-modal retrieval comparison with state-of-the-art methods on MSR-VTT testing set.

Model	Text-to-Video Retrieval					Video-to-Text Retrieval					rsum
	R@1	R@5	R@10	MedR	MnR	R@1	R@5	R@10	MedR	MnR	
VSE [21]	5.0	16.4	24.6	47	215.1	7.7	20.3	31.2	28	185.8	105.2
VSE++ [7]	5.7	17.1	24.8	65	300.8	10.2	25.4	35.1	25	228.1	118.3
Mithun *et al*. [28]	5.8	17.6	25.2	61	296.6	10.5	26.7	35.9	25	266.6	121.7
W2VV [5]	6.1	18.7	27.5	45	-	11.8	28.9	39.1	21	-	132.1
Dual Encoding [6]	7.7	22.0	31.8	32	-	13.0	30.8	43.3	15	-	148.6
Our HGR	**9.2**	**26.2**	**36.5**	**24**	**164.0**	**15.0**	**36.7**	**48.8**	**11**	**90.4**	**172.4**

Local Attentive Matching. At the action and entity level, there are multiple local components in the video and text. Therefore, an alignment between cross-modal local components is supposed to be learned to compute overall matching score. For each $c_{x,i} \in c_x$ where $x \in \{a, o\}$, we first compute local similarities between each pair of cross-modal local components $s^x_{ij} = cos(v_{x,j}, c_{x,i})$. Such local similarities implicitly reflect the alignment between local texts and videos such as how strong a text node is relevant to a video frame, but they lack proper normalization. Therefore, we normalize s^x_{ij} inspired by stacked attention [22] as follows:

$$\varphi^x_{ij} = \text{softmax}(\lambda([s^x_{ij}]_+ / \sqrt{\sum_j [s^x_{ij}]^2_+})) \tag{12}$$

where $[\cdot]_+ \equiv \max(\cdot, 0)$. The φ^x_{ij} is then utilized as attention weights over video frames for each local textual node i, which dynamically aligns $c_{x,i}$ to video frames. We then compute the similarity between $c_{x,i}$ and v_x as weighted average of local similarities $s_{x,i} = \sum_j \varphi^x_{ij} s^x_{ij}$. The final matching similarity summarizes all local component similarities of text $s_x = \sum_i s_{x,i}$. The local attentive matching does not require any local text-video groundings, and can be learned from the weakly supervised global video-text pairs.

Training and Inference. We take the average of cross-modal similarities at all levels as final video-text similarity:

$$s(v, c) = (s_e + s_a + s_o)/3 \tag{13}$$

The contrastive ranking loss is employed as training objective. For each positive pair (v^+, c^+), we find its hardest negatives in a mini-batch (v^+, c^-) and (v^-, c^+), and push their distances from the positive pair (v^+, c^+) further away than a pre-defined margin Δ as follows:

$$\begin{aligned} L(v^+, c^+) = [\Delta + s(v^+, c^-) - s(v^+, c^+)]_+ \\ + [\Delta + s(v^-, c^+) - s(v^+, c^+)]_+ \end{aligned} \tag{14}$$

4. Experiments

To demonstrate the effectiveness of our HGR model, we compare it with state-of-the-art (SOTA) methods on three video-text datasets for text-to-video retrieval and video-to-text retrieval. Extensive ablation studies are conducted to investigate each component of our model. We also propose a binary selection task to evaluate fine-grained discrimination ability of different models for cross-modal retrieval.

4.1. Experimental Settings

Datasets. We carry out experiments on MSR-VTT [41], TGIF [25] and recent VATEX [38] video-text datasets. The MSR-VTT dataset contains 10,000 videos with 20 text descriptions for each video. We follow the standard split with 6,573 videos for training, 497 for validation and 2,990 for testing. The TGIF dataset contains gif format videos, where there are 79,451 videos for training, 10,651 for validation and 11,310 for testing in the official split [25]. Each video is annotated with 1 to 3 text descriptions. The VATEX dataset includes 25,991 videos for training, 3,000 for validation and 6,000 for testing. Since the annotations on testing set are private, we randomly split the validation set into two equal parts with 1,500 videos as validation set and other 1,500 videos as our testing set. There are 10 sentences in English and Chinese languages to describe each video. In this work, we only utilize the English annotations.

Evaluation Metrics. We measure the retrieval performance with common metrics in information retrieval, including Recall at K (R@K), Median Rank (MedR) and Mean Rank (MnR). R@K is the fraction of queries that correctly retrieve desired items in the top K of ranking list. We utilize K = 1, 5, 10 following the tradition. The MedR and MnR measures the median and average rank of correct items in the retrieved ranking list respectively, where lower score indicates a better model. We also take the sum of all R@K as rsum to reflect the overall retrieval performance.

Implementation Details. For the video encoding, we use Resnet152 pretrained on Imagenet [13] to extract frame-wise features for MSR-VTT and TGIF. We utilize the officially provided I3D [1] video feature for VATEX dataset. For the text encoding, we set the word embedding size as 300 and initialize with pretrained Glove embeddings [29]. We use two layers of attentional graph convolutions. The dimension of joint embedding space for each level is 1024.

Table 2. Generalization on unseen Youtute2Text testing set using different pre-trained models on MSR-VTT dataset.

Model	Text-to-Video Retrieval					Video-to-Text Retrieval					rsum
	R@1	R@5	R@10	MedR	MnR	R@1	R@5	R@10	MedR	MnR	
VSE [21]	11.0	28.6	39.9	18	48.7	15.4	31.0	42.4	19	128.0	168.3
VSE++ [7]	13.8	34.6	46.1	13	48.4	20.8	37.6	47.8	12	108.3	200.6
Dual Encoding [6]	12.7	32.0	43.8	15	52.7	18.7	37.2	45.7	15	142.6	190.0
Our HGR	**16.4**	**38.3**	**49.8**	**11**	**49.2**	**23.0**	**42.2**	**53.4**	**8**	**77.8**	**223.2**

Table 3. Text-to-video retrieval comparison with state-of-the-art methods on TGIF and VATEX testing set.

Dataset	Model	R@1	R@5	R@10	MedR
TGIF	DeViSE [9]	0.8	3.5	6.0	379
	VSE++ [7]	0.4	1.6	3.6	692
	Order [35]	0.5	2.1	3.8	500
	Corr-AE [8]	0.9	3.4	5.6	365
	PVSE [32]	2.3	7.5	11.9	162
	HGR	**4.5**	**12.4**	**17.8**	**160**
VATEX	VSE [21]	28.0	64.2	76.9	3
	VSE++ [7]	33.7	70.1	81.0	**2**
	Dual Encoding [6]	31.1	67.4	78.9	3
	HGR	**35.1**	**73.5**	**83.5**	**2**

We set $\lambda = 4$ in local attentive matching. For training, we set the margin $\Delta = 0.2$, and train the model for 50 epochs with mini-batch size of 128. The epoch with the best rsum on validation set is selected for inference.

4.2. Comparison with State of The Arts

Table 1 compares the proposed HGR model with SOTA methods on the MSR-VTT testing set. For fair comparison, all the models utilize the same video features. Our model achieves the best performance across different evaluation metrics on the MSR-VTT dataset. It outperforms the state-of-the-art Dual Encoding [6] method even with half less parameters and computations, which obtains 19.5% and 15.4% relative gains on R@1 metric for text-to-video and video-to-text retrieval respectively. The overall retrieval quality reflected by the rsum metric is also boosted by a large margin (+23.8). We believe the major gain comes from our global-to-local matching and attention-based graph reasoning to learn hierarchical textual representations. Though Dual Encoding enhances global video and sentence features via ensembling different networks such as mean pooling, RNNs and CNNs, it may still focus on the global event level and thus not as efficient as ours to capture fine-grained semantic details in text for cross-modal video-text retrieval.

To demonstrate the robustness of our approach on different datasets and features, we further provide quantitative results on TGIF and VATEX datasets in Table 3. The models employ Resnet152 image features on the TGIF dataset and I3D video features on the VATEX dataset. We can see that our HGR model achieves consistent improvements across different datasets and features compared to SOTA models, which demonstrates that it is beneficial to improve the cross-modal retrieval accuracy by decomposing videos and texts into global-to-local hierarchical graph structures.

4.3. Generalization on Unseen Dataset

Current video-text retrieval methods are mainly evaluated on the same dataset. However, it is important for the model to generalize to out-of-domain data. Therefore, we further conduct generalization evaluations: we first pretrain a model on one dataset and then measure its performance on another dataset that is unseen in the training. Specifically, we utilize the MSR-VTT dataset for training and test models on the Youtube2Text testing split [11], which contains 670 videos and 41.5 descriptions per video on average.

Table 2 presents retrieval results on the Youtube2Text dataset. The hard negative training strategy proposed in VSE++ [7] enables the model to learn visual-semantic matching more effectively, which also improves model's generalization ability on unseen data. The Dual Encoding model though achieves better retrieval performance on the MSR-VTT dataset as show in Table 1, it does not generalize well on a new dataset compared with VSE++ with overall 10.6 points decrease on rsum metric. Our HGR model instead not only outperforms previous approaches on in-domain evaluation, but also achieves significantly better retrieval performance on out-of-domain dataset. This property proves that improvements of our model does not result from using more complicated networks that might overfit datasets. Since we decompose texts into structures of events, actions and entities from global to local and match them with hierarchical video embeddings, our model is capable of learning better alignments of local components as well as global event structures, which improves the generalization ability on new compositions.

4.4. Ablation Studies

In order to investigate contributions of different components in our proposed model, we carry out ablation studies on the MSR-VTT dataset in Table 4. The Row 1 in Table 4 replaces graph attention mechanism in graph rea-

Table 4. Ablation studies on MSR-VTT dataset to investigate contributions of different components of our HGR model.

	Model	Text-to-Video Retrieval					Video-to-Text Retrieval					rsum
		R@1	R@5	R@10	MedR	MnR	R@1	R@5	R@10	MedR	MnR	
1	w/o graph attention	8.9	25.3	35.6	25	173.5	14.5	35.7	47.1	12	96.5	167.1
2	w/o role awareness	9.1	25.7	36.3	24	171.3	14.2	34.7	46.8	12	98.0	166.8
3	w/o hierarchical video	8.8	25.5	36.2	24	170.2	**15.2**	35.1	47.2	12	108.9	167.9
4	full HGR model	**9.2**	**26.2**	**36.5**	**24**	**164.0**	15.0	**36.7**	**48.8**	**11**	**90.4**	**172.4**

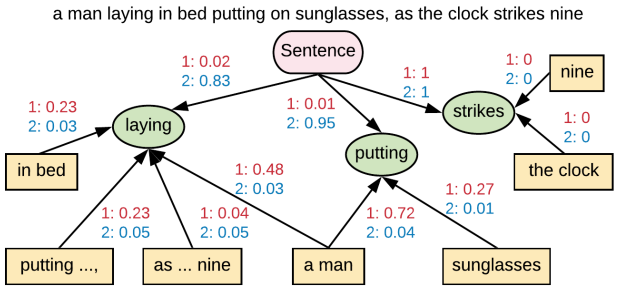

Figure 3. The attention distributions of action nodes at different graph reasoning layers to gather contexts from other nodes. The number in red after 1 denotes attention score in the first attention layer, while the number in blue after 2 denotes attention score in the second attention layer.

soning and simply utilizes average pooling over neighbor nodes, which reduces the retrieval performance with 0.9 and 1.7 on R@10 metric than the full model in Row 4 for text-to-video and video-to-text retrieval respectively. The role awareness in Eq (6) is also beneficial in graph reasoning comparing Row 2 and Row 4, which enables the model to understand how different components relate with each other within an event. In Figure 3, we present a learned pattern on how action nodes interacting with neighbor nodes in graph reasoning at different layers, which is strongly relevant to semantic roles. At the first attention layer, the action node such as "laying", "putting" focuses more on its main arguments such as agent "man". Then at the second layer, action nodes begin to reason over their temporal relations and thus pay more attention to temporal arguments as well as implicit contexts from global event node.

We also show that representing videos as hierarchical embeddings is important to capture different aspects in the video, which improves overall rsum performance from 167.9 in Row 3 to 172.4 in row 4. Since our video-text similarities are aggregated from different levels, in Table 5 we break down the performance at each level for video-text retrieval. We can see that the global event level performs the best alone on rsum metric since local levels might not contain overall event structures on itself. But different levels are complementary with each other and their combination significantly improves the retrieval performance.

Table 5. Break down of retrieval performance at different levels on MSR-VTT testing set.

	Text-to-Video			Video-to-Text		
	rsum	MedR	MnR	rsum	MedR	MnR
event	57.6	43	267.8	77.8	20.5	258.0
action	50.4	77	441.6	80.7	22	241.4
entity	44.7	62	251.3	58.4	37	230.0
fusion	**71.9**	**24**	**164.0**	**100.6**	**11**	**90.4**

4.5. Fine-grained Binary Selection

To prove the ability of our model for fine-grained retrieval, we further propose a binary selection task that requires the model to select a sentence that better matches with a given video from two very similar but semantically different sentences. We utilize testing videos from the Youtube2Text dataset and randomly select one ground-truth video description for each video as positive sentence. The negative sentence is generated by perturbing the ground-truth sentence in one of the following ways:

1. switch roles: switching agent and patient of an action;
2. replace actions: replacing action with random action;
3. replace persons: replacing agent or patient entities with random agents or patients;
4. replace scenes: randomly replacing scene entities;
5. incomplete events: only keeping part of all actions, entities in the sentence;

We then ask human workers to ensure the automatic generated sentences are syntactically correct but indeed semantically inconsistent with the video content. Examples can be found in the supplementary material.

Table 6 presents results in different binary selection tasks. For the switching roles task, our model outperforms VSE++ model with absolute 4.87%, but is slightly inferior to Dual Encoding model. We suspect the reason is that video descriptions in Youtube2Text are relatively short (7 words on average per sentence), which makes sequential models with local contexts such as LSTM, CNN in Dual Encoding model sufficient to capture the event structure. For the replacing tasks, the HGR model achieves the best performance to distinguish entity replacement especially for scenes. The largest improvement of our HGR model lies in

Table 6. Performance of different models on fine-grained binary selection task.

Model	switch roles	replace actions	replace persons	replace scenes	incomplete events	average
# of triplets	616	646	670	539	646	623.4
VSE++ [7]	64.61	**74.46**	85.67	83.30	78.79	77.37
Dual Encoding [6]	**71.92**	71.52	86.12	82.00	70.59	76.43
Our HGR	69.48	71.21	**86.27**	**84.05**	**82.04**	**78.61**

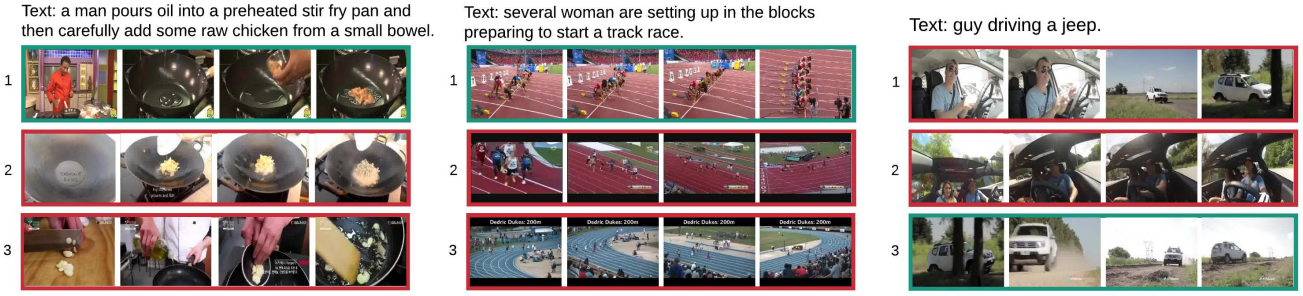

Figure 4. Text-to-video retrieval examples on MSR-VTT testing set. We visualize top 3 retrieved videos (green: correct; red: incorrect).

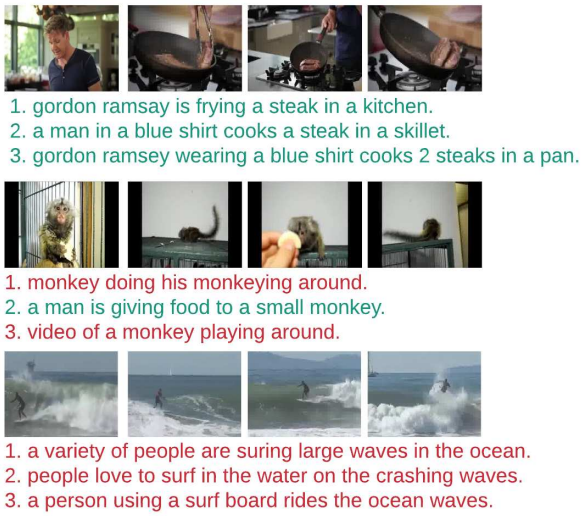

Figure 5. Video-to-text retrieval examples on MSR-VTT testing set with top 3 retrieved texts (green: correct; red: incorrect).

the incomplete events task, where both the two sentences are relevant to video contents but one captures more details. Due to the fusion of hierarchical levels from global to local, our model can select the more comprehensive sentence.

4.6. Qualitative Results

We visualize some examples on the MSR-VTT testing split for text-to-video retrieval in Figure 4. In the left example, our model successfully retrieves the correct video which contains all actions and entities described in the sentence, while the second video only lacks "pour oil" action and the third video does not contain "chicken" entity. In the middle example, the HGR model also distinguishes different relationship of actions such as "prepare to start a track race" and "run in a track race". The right example shows a fail case, where the top retrieved videos are largely relevant to the text query though are not ground-truth. In Figure 5, we provide qualitative results on video-to-text retrieval as well, which demonstrate the effectiveness of our HGR model for cross-modal retrieval on both directions.

5. Conclusion

Most successful cross-modal video-text retrieval systems are based on joint embedding approaches. However, simple embeddings are insufficient to capture fine-grained semantics in complicated videos and texts. Therefore, in this work, we propose a Hierarchical Graph Reasoning (HGR) model which decomposes videos and texts into hierarchical semantic levels including events, actions, and entities. It then generates hierarchical textual embeddings via attention-based graph reasoning and aligns texts with videos at different levels. Superior experimental results on three video-text datasets demonstrate the advantages of our model. The proposed HGR model also achieves better generalization performance on unseen dataset and is capable of distinguishing fine-grained semantic differences. In the future, we will improve video encoding with multi-modalities and spatial-temporal reasoning.

6. Acknowledgment

This work was supported by National Natural Science Foundation of China (No. 61772535), Beijing Natural Science Foundation (No. 4192028), and National Key Research and Development Plan (No. 2016YFB1001202). Qi Wu is funded by DE190100539 and NSFC 61877038.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 5

[2] Xiaojun Chang, Yi Yang, Alexander Hauptmann, Eric P Xing, and Yao-Liang Yu. Semantic concept discovery for large-scale zero-shot event detection. In *Proceedings of the Twenty-fourth International Joint Conference on Artificial Intelligence*, 2015. 1

[3] Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee Kenneth Wong. Weakly-supervised spatio-temporally grounding natural sentence in video. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1884–1894, 2019. 2

[4] Jeffrey Dalton, James Allan, and Pranav Mirajkar. Zero-shot video retrieval using content and concepts. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 1857–1860. ACM, 2013. 1

[5] Jianfeng Dong, Xirong Li, and Cees GM Snoek. Predicting visual features from text for image and video caption retrieval. *IEEE Transactions on Multimedia*, 20(12):3377–3388, 2018. 5

[6] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. Dual encoding for zero-example video retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9346–9355, 2019. 1, 2, 5, 6, 8

[7] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference*, 2018. 2, 5, 6, 8

[8] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 7–16. ACM, 2014. 6

[9] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 2121–2129, 2013. 1, 2, 6

[10] Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7181–7189, 2018. 2

[11] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2712–2719, 2013. 6

[12] Amirhossein Habibian, Thomas Mensink, and Cees GM Snoek. Composite concept discovery for zero-shot video event detection. In *Proceedings of International Conference on Multimedia Retrieval*, page 17. ACM, 2014. 1

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[15] Micah Hodosh and Julia Hockenmaier. Focused evaluation for image description with binary forced-choice tasks. In *Proceedings of the 5th Workshop on Vision and Language*, pages 19–28, 2016. 2

[16] Hexiang Hu, Ishan Misra, and Laurens van der Maaten. Evaluating text-to-image matching using binary image selection (bison). In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019. 2

[17] Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. Language-conditioned graph networks for relational reasoning. *arXiv preprint arXiv:1905.04405*, 2019. 3

[18] Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. Learning semantic concepts and order for image and sentence matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2018. 2

[19] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015. 1, 2

[20] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 2

[21] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 2, 5, 6

[22] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision*, pages 201–216, 2018. 1, 2, 5

[23] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4654–4662, 2019. 3

[24] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. *arXiv preprint arXiv:1903.12314*, 2019. 3

[25] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4641–4650, 2016. 5

[26] An-An Liu, Yu-Ting Su, Wei-Zhi Nie, and Mohan Kankanhalli. Hierarchical clustering multi-task learning for joint human action grouping and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):102–114, 2016. 2

[27] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *Proceedings of the British Machine Vision Conference*, 2019. 1, 2

[28] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 19–27. ACM, 2018. 1, 2, 5

[29] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing*, pages 1532–1543, 2014. 5

[30] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pages 593–607. Springer, 2018. 2, 4

[31] Peng Shi and Jimmy Lin. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*, 2019. 3

[32] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1979–1988, 2019. 1, 2, 6

[33] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid. Actor-centric relation network. In *Proceedings of the European Conference on Computer Vision*, pages 318–334, 2018. 3

[34] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017. 2

[35] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015. 6

[36] Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1960–1968, 2019. 3

[37] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European Conference on Computer Vision*, pages 399–417, 2018. 3

[38] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 5

[39] Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, October 2019. 2

[40] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6609–6618, 2019. 2

[41] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5288–5296, 2016. 5

[42] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European Conference on Computer Vision*, pages 670–685, 2018. 3

[43] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision*, pages 471–487, 2018. 1, 2

[44] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 374–390, 2018. 2