HANet: Hierarchical Alignment Networks for Video-Text Retrieval

Peng Wu[*]
xdwupeng@gmail.com
Guangzhou Institute of Technology,
Xidian University
Guangzhou, China

Xiangteng He[†]
xiangteng.hxt@alibaba-inc.com
DAMO Academy, Alibaba Group
Hangzhou, China

Mingqian Tang
mingqian.tmq@alibaba-inc.com
DAMO Academy, Alibaba Group
Hangzhou, China

Yiliang Lv
yiliang.lyl@alibaba-inc.com
DAMO Academy, Alibaba Group
Hangzhou, China

Jing Liu[†]
neouma@163.com
Guangzhou Institute of Technology,
Xidian University
Guangzhou, China

ABSTRACT

Video-text retrieval is an important yet challenging task in vision-language understanding, which aims to learn a joint embedding space where related video and text instances are close to each other. Most current works simply measure the video-text similarity based on video-level and text-level embeddings. However, the neglect of more fine-grained or local information causes the problem of insufficient representation. Some works exploit the local details by disentangling sentences, but overlook the corresponding videos, causing the asymmetry of video-text representation. To address the above limitations, we propose a Hierarchical Alignment Network (HANet) to align different level representations for video-text matching. Specifically, we first decompose video and text into three semantic levels, namely event (video and text), action (motion and verb), and entity (appearance and noun). Based on these, we naturally construct hierarchical representations in the individual-local-global manner, where the individual level focuses on the alignment between frame and word, local level focuses on the alignment between video clip and textual context, and global level focuses on the alignment between the whole video and text. Different level alignments capture fine-to-coarse correlations between video and text, as well as take the advantage of the complementary information among three semantic levels. Besides, our HANet is also richly interpretable by explicitly learning key semantic concepts. Extensive experiments on two public datasets, namely MSR-VTT and VATEX, show the proposed HANet outperforms other state-of-the-art methods, which demonstrates the effectiveness of hierarchical representation and alignment. Our code is publicly available. [1]

1 INTRODUCTION

"Hey, Siri, I want to watch a video of fighting in the desert." Recently, cross-modal retrieval has attracted increasing attention due to the explosive growth of online videos and advances in artificial intelligence technology. In addition to speech recognition, video-text retrieval is a key technique in the above scenario, which aims to search related videos given a natural-language sentence as the query. This task is challenging since the video and text are two different modalities, how to encode and match them in the joint space is the key.

Many efforts to make a reliable and accurate video-text retrieval system have been done. Recently, a typical practice is to encode videos and texts into compact representations and measure their similarities in a joint latent common space using metric learning. In this case, most existing works [8, 11, 22, 29] focus on how to learn global representations of videos and texts, and achieve this goal by leverage various embedding networks, e.g., convolutional neural networks (CNN), gated recurrent units (GRU), Transformer, Bidirectional Encoder Representations from Transformers (BERT).

[*]Work done during an internship at Alibaba DAMO Academy.
[†]Corresponding authors

[1]https://github.com/Roc-Ng/HANet

Figure 1: Illustration of the hierarchical alignment. The individual level focuses on the alignment between frame and word, local level focuses on the alignment between video clip and textual context, and global level focuses on the alignment between the whole video and text. Moreover, concept vocabulary includes entity and action concepts, which are used for video parsing and hierarchical alignment.

However, such compact global representations neglect the more fine-grained or local information that existed in videos and texts, which may cause the problem of insufficient representation.

To mitigate this problem, some other works utilize local semantic information for fine-grained video-text retrieval. For example, Wray et al. [41] break the sentence into nouns and verbs using part-of-speech (pos) parsing and operate multiple cross-modal matching. Chen et al. [4] make further efforts for fine-grained retrieval by exploiting semantic alignments for both global event and local action and entity through hierarchical graph reasoning. Nevertheless, video-text retrieval is a cross-modal task, these methods only focus on text parsing, yet overlook video parsing, causing the asymmetry of video-text representation. Therefore, simultaneously parsing text and video is a more general solution.

To address the above issues, we propose the hierarchical alignment network (HANet), which aims to simultaneously parse text and video into different semantic levels, and then generates individual-, local- and global-level representations, finally hierarchically aligns different level representations in separate joint spaces. We illustrate the hierarchical alignment in Figure 1. The contributions of this paper are as follows:

- **Multi-semantic-level representation** is proposed to simultaneously parse videos and texts into different semantic levels, i.e. event-level on the whole video and text, action- and entity-level on the parsing parts of video and text, where

we subtly parse videos by means of concept-based classification under weak supervision without other complicated steps. Here, the concept is from predefined concept vocabularies which drive the model to learn the detailed components in video and text, so that the cross-modal associations between video frames and keywords could be established via concepts, which makes our model more interpretable.

- **Hierarchical Alignment** is proposed for cross-modal matching on top of multi-semantic-level representation. The individual level focuses on the alignment between frame and word on the concept-specific prediction features, local level focuses on the alignment between video clip and textual context, and global level focuses on the alignment between the whole video and text. Different level alignments not only capture fine-to-coarse correlations between videos and texts, but also take the advantage of the complementary information among three semantic levels.

We show the superiority of our HANet on two popular video-text retrieval datasets, i.e., MSR-VTT, VATEX. Without additional features and pre-training, HANet achieves clear performance improvements over state-of-the-art methods.

2 RELATED WORK

2.1 Video-Text Retrieval

Video-text retrieval is a non-trivial branch of cross-modal retrieval [3, 13, 15, 21, 27, 31, 36, 40, 46]. The typical methods can be divided into three components, namely, text encoding, video encoding, and joint space learning. Recently, many works focus on designing powerful text and video encoding. For example, Dong et al. [8] proposed multi-level encodings of video and text in advance to learning shared representations. Li et al. [22] concatenated the bag-of-words vector, word2vec embedding, and Recurrent neural networks (RNN) vector as the final text representations. Similarly, Li et al. [23] incorporated several sentence encoders and measured similarities in multiple encoder-specific common spaces rather than a single common space. Some works [12, 25, 28] made full use of multimodal cues, e.g., motion, appearance, face, OCR, for video encoding. Other works attempted to decompose texts into some semantic phrases. Yang et al. [48] constructed a latent semantic tree to describe the text and used a temporal attentive encoder to obtain the temporal-attentive video representation. Xu et al. [47] proposed a compositional semantics language model to parse the sentence into Subject-Verb-Object structure. Wray et al. [41] disentangled sentences into verbs and nouns for fine-grained video retrieval, and Chen et al. [4] disentangled texts into events, actions and entities. The above two methods are similar to ours, but they overlook disentangling videos and developing interpretability. Another two works [24, 50] are also hierarchical models, however, the method in [50] is not applicable to decompose single sentences, and the hierarchical transformer in [24] only focus on global features.

As for joint space learning, Chen et al. [2] designed a new ranking loss that assigns weights to the relative similarities between positive and negative pairs. An analogous work [39] introduced a new polynomial loss with the universal weighting framework. With the help of these new losses, traditional methods achieve clear performance improvements.

2.2 Cross-modal Concept Learning

In the last few years, cross-modal concept learning is usually utilized for a new challenge in TRECVID, i.e., Ad-hoc Video Search (AVS). The majority of the top-ranked solutions [19, 26, 30, 37] for AVS focused on computing the similarity between a textual query and a specific video via concepts. In terms of video, constructing visual concept classifier to detect concepts; In terms of text, designing complex linguistic rules to extract relevant concepts. The merit of these concept-based AVS methods is the good interpretability, but the weakness is that using predefined concepts to describe videos and texts is insufficient. Recently, two hybrid works [9, 42] employed both concept-based and concept-free strategies and achieved better performance. Similar to them, Our HANet is also considered as a hybrid model. Another interesting work [49] proposed a high-level concept word detector to generates a series of concept words as useful semantic priors for cross-modal tasks.

3 HANET

We propose the hierarchical alignment network (HANet) for video-text retrieval, whose goal is to hierarchically align different level of video-text features and measure the similarity in different common spaces. The overview of HANet is illustrated in Figure 2, which consists of four components: 1) Video-text parsing (Section 3.1), i.e., parsing video and text with video-specific and text-specific parse manners respectively. 2) Video representations (Section 3.2), i.e., constructing frame, clip and video levels of features on the basis of video parsing. 3) Text representations (Section 3.3), i.e., constructing word, phrase, and sentence levels of features on the basis of text parsing. 4) Hierarchical alignment (Section 3.4), i.e., aligning different semantic levels of video and text representations to compute their similarity. Finally, training and inference are introduced in Section 3.5.

3.1 Video-Text Parsing

Given a video V of length N and its corresponding caption S of length M, we expect that not only global representations of the video and text are close in their common space, but also more fine-grained representations should be close. To achieve this goal, we first parse the video and text.

3.1.1 Video Parsing. Unlike text parsing, video parsing is challenging since videos are more complicated but lacking the distinct semantic structure than texts. A possible way to video parsing is to introduce some existing vision operations, such as temporal segmentation, object detection, tracking, which are time-consuming and complex for practice application. To address this issue, we propose to project video frames into the concept space with the predefined concept vocabulary, where each frame is explicitly associated with the specific concepts. That is, we establish cross-modal associations between video frames (or clips) with key words (such as nouns and verbs) via concepts.

Formally, given the predefined action concept vocabulary of size K_a, we project the corresponding action feature map $v_a^{Ind} \in \mathbb{R}^{N \times D^v}$ of video into the K_a dimensional action concept space using CNN. Likewise, we adopt another CNN to project entity feature map $v_e^{Ind} \in \mathbb{R}^{N \times D^v}$ into the K_e dimensional entity concept space. That

is,

$$l_a^v = \sigma \left(BN \left(Conv1d_{k=5} \left(v_a^{Ind} \right) \right) \right) \tag{1}$$

$$l_e^v = \sigma \left(BN \left(Conv1d \left(v_e^{Ind} \right) \right) \right) \tag{2}$$

where $l_a^v \in \mathbb{R}^{N \times K_a}$ and $l_e^v \in \mathbb{R}^{N \times K_e}$ are denoted as the confidence for action and entity concepts, respectively. σ is the sigmoid activation, BN is the batch normalization. The detailed descriptions of v_a^{Ind} and v_e^{Ind} can be referred to Section 3.2. It is worth mentioning that we employ the convolution layer with kernel size of $k = 5$ to obtain the probabilistic output of action concept, which is to capture the intrinsic motion information in continuous video frames.

3.1.2 Text Parsing. Following the prior work [4], we employ the off-the-shelf semantic role parsing toolkit [35] to obtain verbs, nouns as well as the semantic role of each noun to the corresponding verb. We refer the reader to [4, 35] for detailed descriptions. Here, verbs are considered as actions, likewise, nouns are entities. Then, we project sentence words into the action and entity concepts, and their confidences are obtained as follows,

$$l_a^s = \sigma \left(BN \left(Conv1d \left(s_a^{Ind} \right) \right) \right) \tag{3}$$

$$l_e^s = \sigma \left(BN \left(Conv1d \left(s_e^{Ind} \right) \right) \right) \tag{4}$$

where $l_a^s \in \mathbb{R}^{M \times K_a}$ and $l_e^s \in \mathbb{R}^{M \times K_e}$ are denoted as the confidence for action and entity concepts, respectively. The detailed descriptions of s_a^{Ind} and s_e^{Ind} can be referred to Section 3.3.

Video-text parsing introduces a new concept-based match space between videos and texts on action and entity concept-level. In this way, good interpretability is also introduced.

3.2 Video Representations

We design three different granularities of representations, namely individual, local and global, corresponding to the video frame, video clip, and whole video respectively, which capture fine-to-coarse information and are complementary to each other.

3.2.1 Individual-level Representation. Formally, given the video V, we first use the pre-trained CNN to extract frame-level features $F^v = \{f_i^v\}^N$. Following [4], we then employ different fully connected (FC) layers to encode the video into three semantic-level embeddings as follows:

$$v_x^{Ind} = FC \left(F^v \right), \quad x \in \{a, e, g\} \tag{5}$$

where $v_x^{Ind} \in \mathbb{R}^{N \times D^v}$. Since v_a^{Ind} and v_e^{Ind} only contain frame-level information, here, they are considered as individual-level representations. v_g^{Ind} is used for the global-level representation in the later section.

3.2.2 Local-level Representation. To further explore the contextual information between frames, we propose a Selecting and Merging (SeMe) module to generate the local-level representation. The SeMe module takes individual-level representation as input, and outputs concept confidence. To extract more high-level representation, we employ a convolution layer to project v_a^{Ind} and v_e^{Ind} into another spaces. That is,

$$v_a^{SE} = SE \left(v_a^{Ind} \right) \tag{6}$$

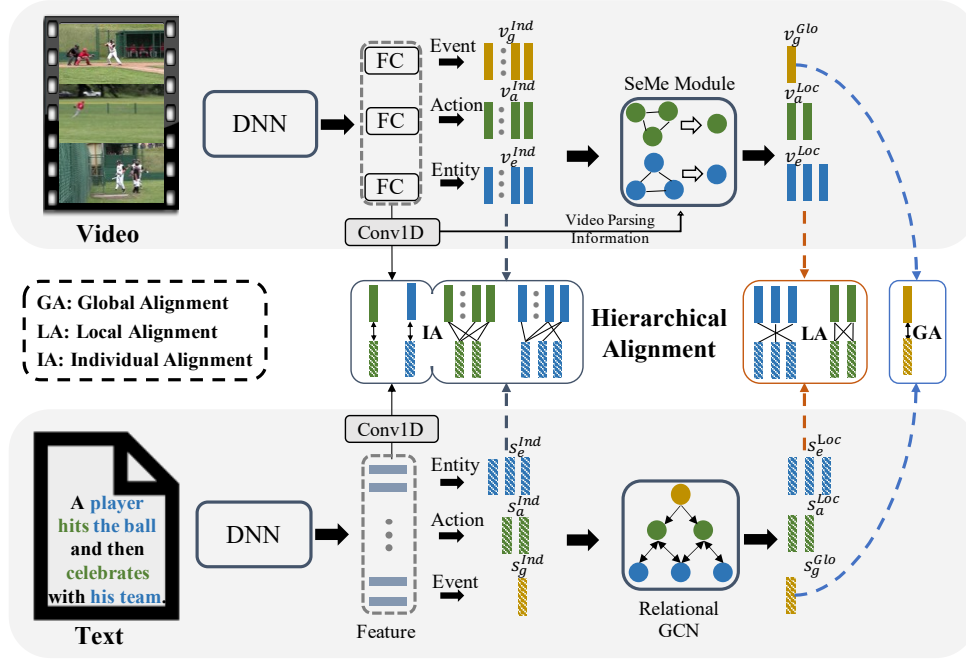$$v_e^{SE} = SE \left(v_e^{Ind} \right) \tag{7}$$

Figure 2: Framework of the proposed Hierarchical Alignment Network (HANet).

where SE is the simple yet effective Squeeze-and-Excitation block [16] to explicitly model inter-dependencies between channels. After that, we obtain the local-level representation with the help of video parsing. Specifically, each frame has K_a dimensional action concept confidence and K_e dimensional entity concept confidence, here we expect to know which action and entity concepts are associated with each video rather than each frame. To this end, we introduce a multiple instance learning (MIL) based mechanism inspired by [32, 43, 44], which can be presented as follows,

$$p_{a,i}^v = \frac{1}{\tau} \sum_{\forall z \in Z_i} z \qquad (8)$$

where $\tau = \lfloor \frac{N}{8} \rfloor$, and Z_i is the set of τ-max frame-level confidence scores for the ith action concept, which is selected from frame-level action concept confidence l_a^v, and the size of Z_i is τ. p_a^v is the video-level action concept confidence given the video V. For the ith action concept, we obtain the ith video-level action concept confidence $p_{a,i}^v$ by averaging Z_i. In the similar vein, we obtain p_e^v. Given p_a^v and p_e^v, N_a and N_e action and entity concepts with the highest confidence are selected as the reliable concepts for the video V. For each selected action concept, we select the video clip of size $k = 5$ in v_a^{SE} which corresponds to the highest confidence, and employ the average pooling to obtain a feature vector. Through this operation, we finally obtain the local-level action representation $v_a^{Loc} \in \mathbb{R}^{N^a \times D^v}$. The main difference between obtaining v_a^{Loc} and obtaining local-level entity representation v_e^{Loc} is that we select 3 frames in v_e^{SE}, which may not be continuous, corresponding to the top 3 highest confidences and obtain a feature vector by the average pooling. By merging adjacent and semantically similar frames, local-level representations contain more rich information

and capture local range dependencies, they can explicitly align with nouns and verbs in texts.

To summarize, local-level representation is the aggregation of several frame-level features corresponding to reliable concepts. First, we obtain the video-level concept confidence based on frame-level concept confidence via Equation (8). Then, we select highly confident action and entity concepts as reliable concepts based on video-level concept confidence. For each reliable concept, we choose some highly confident frame features based on the frame-level concept confidence, and aggregate them to generate the final local-level representation.

3.2.3 Global-level Representation. For the global event level, we adopt an attention mechanism to average the frame-level features as a single global vector v_g^{Glo} that represents the salient event in the video, which is presented as follows,

$$v_g^{Glo} = \alpha^\top v_g^{Ind} \qquad (9)$$

$$\alpha = softmax\left(v_g^{Ind} W\right) \qquad (10)$$

where α is the attention score, and W is the learnable weight.

3.3 Text Representations

Parallel to video representations, text representations consist of three levels of granularity, i.e., the individual level corresponds to the word, the local level corresponds to the context, and the global level corresponds to the sentence. Three levels of text representations are used to align with corresponding video representations in the hierarchical alignment. Formally, given the sentence S, we employ the pre-trained model to extract the word embeddings

$F^s = \{f_i^s\}^M$, then generate three different levels of representations on top of text parsing.

3.3.1 Individual-level Representation.
We utilize an bidirectional GRU (Bi-GRU) [5] to generate a sequence of contextual-aware word embeddings as follows,

$$\overrightarrow{s^{Ind}} = \overrightarrow{GRU}\left(F^s\right) \tag{11}$$

$$\overleftarrow{s^{Ind}} = \overleftarrow{GRU}\left(F^s\right) \tag{12}$$

$$s^{Ind} = \left(\overrightarrow{s^{Ind}} + \overleftarrow{s^{Ind}}\right)/2 \tag{13}$$

Based on the text parsing, we select features corresponding to verbs and nouns as the individual-level representations $s_a^{Ind} \in \mathbb{R}^{M^a \times D^s}$ and $s_e^{Ind} \in \mathbb{R}^{M^e \times D^s}$ that are subsets of s^{Ind}, which correspond to v_a^{Ind} and v_e^{Ind} in videos. Meanwhile, we employ the attention mechanism similar to Equations (9)-(10) to obtain the global event embedding s_g^{Ind} for the following local and global-level representations.

3.3.2 Local and Global-level Representation.
We follow the pioneer work [4] and employ modified relational GCN [34] to obtain local and global-level representations. To be specific, we gather three features, i.e., s_g^{Ind}, s_a^{Ind} and s_e^{Ind}, and use $g \in \mathbb{R}^{(1+M^a+M^e)\times D^s}$ to denote this feature set, which is initialized node embeddings of graph. Different semantic roles by text parsing are the edges of graph.

Here, we only use one GCN layer, which is presented as follows,

$$g_i^1 = g_i^0 + \sum_{j \in N_i} \left(\beta_{ij}\left(W_t \odot W_r r_{ij}\right) g_j\right) \tag{14}$$

where $W_t \in \mathbb{R}^{D^s \times D^s}$ is the transformation matrix, $W_r \in \mathbb{R}^{D^s \times K^r}$ is role embedding matrix, K^r is the number of semantic roles, N_i is neighborhood nodes of node i, r_{ij} an one-hot vector of length K^r denoting the edge type from node i to j, and β_{ij} is the similarity between node i to j, which is computed as follows,

$$\beta_{ij} = softmax\left(\frac{\phi\left(g_i\right)\varphi\left(g_j\right)^\top}{\sqrt{D^s}}\right) \tag{15}$$

here $\phi\left(g_i\right) = W_\phi g_i$ and $\varphi\left(g_j\right) = W_\varphi g_j$ are two embeddings. We refer the reader to [4] for detailed descriptions of the modified relational GCN.

The outputs from the GCN layer are the final different levels of representations, which are denoted as s_a^{Loc} for local-level action representation, s_e^{Loc} for local-level entity representation and s_g^{Glo} for global-level representation. Notably, we do not use the relational GCN in videos since the semantic role of each entity to the corresponding action is unknown.

3.4 Hierarchical Alignment
After the aforementioned text encoding and video encoding, we obtain three levels of representations, namely, individual, local, and global levels. In this section, we introduce how to hierarchically align representations at three different level.

3.4.1 Individual Alignment.
Since there are multiple components in the video and text at the individual level, following [4, 6, 20], we use the stack attention mechanism to align multiple components and compute the overall similarity score. For the sake of clarity, we use v^{Ind} to denote v_a^{Ind} and v_e^{Ind}, in the similar vein, we define s^{Ind}. We use cosine similarity to compute similarities between each pair of cross-modal components $c_{ij}^{Ind} = cos\left(v_i^{Ind}, s_j^{Ind}\right)$. Then we compute the attention weight that dynamically aligns sentence words and video frames as follows,

$$\gamma_{ij} = softmax\left(\lambda[c_{ij}^{Ind}]_+ \Big/ \sqrt{\sum_j [c_{ij}^{Ind}]_+^2}\right) \tag{16}$$

where λ is the temperature parameter, $[\cdot]_+ \equiv max(\cdot, 0)$. The final similarity summarizes all individual component similarities and is shown as follows,

$$c^{Ind} = \sum_i \left(\sum_j \gamma_{ij} c_{ij}^{Ind}\right) \tag{17}$$

More importantly, we introduce the concept-based similarity based on the concept confidence, where the concept confidence develops from individual-level representations. For simplicity, we use p^v to denote p_a^v and p_e^v. Following [9], we employ generalized Jaccard similarity to compute the concept-based similarity,

$$c^p = \frac{\sum_i min\left(p_i^v, p_i^s\right)}{\sum_i max\left(p_i^v, p_i^s\right)} \tag{18}$$

where p^s is denoted p_a^s and p_e^s, which is obtained as in Equation (8).

3.4.2 Local Alignment.
For the local level, we employ the stack attention mechanism similar to Equations (16)-(17) to obtain c_a^{Loc} and c_e^{Loc}.

3.4.3 Global Alignment.
At the global event level, the video and text are encoded into global vectors. We use the cosine similarity to measure the cross-modal similarity between global video and global text $c_g^{Glo} = cos\left(v_g^{Glo}, s_g^{Glo}\right)$.

3.5 Training and Inference.
3.5.1 Training.
Once all similarity scores are computed, we obtain the two similarity between the video V and the sentence S, that is,

$$c_l(V, S) = \left(c_a^{Ind} + c_e^{Ind} + c_a^{Loc} + c_e^{Loc} + c_g^{Glo}\right)/5 \tag{19}$$

$$c_p(V, S) = \left(c_a^p + c_e^p\right)/2 \tag{20}$$

The widely used ranking loss with hard negative sampling strategy is used to optimize HANet, here $\mathcal{L}_l$ and $\mathcal{L}_p$ are presented as follows,

$$\mathcal{L}_l = [\Delta + c_l(V, S^-) - c_l(V, S)]_+ + [\Delta + c_l(V^-, S) - c_l(V, S)]_+ \tag{21}$$

$$\mathcal{L}_p = [\Delta + c_p(V, S^-) - c_p(V, S)]_+ + [\Delta + c_p(V^-, S) - c_p(V, S)]_+ \tag{22}$$

where (V, S) are the positive pair, and V^- and S^- are the hardest negatives in a mini-batch.

Besides, we use the binary cross-entropy (BCE) loss for concept learning,

$$
\mathcal{L}_a = -\frac{1}{K^a} \sum_i \left(y_{a,i} log \left(p_a^v \right) + (1 - y_{a,i}) log \left(1 - p_a^v \right) \right)
$$
$$
-\frac{1}{K^a} \sum_i \left(y_{a,i} log \left(p_a^s \right) + (1 - y_{a,i}) log \left(1 - p_a^s \right) \right) \tag{23}
$$

$$
\mathcal{L}_e = -\frac{1}{K^e} \sum_i \left(y_{e,i} log \left(p_e^v \right) + (1 - y_{e,i}) log \left(1 - p_e^v \right) \right)
$$
$$
-\frac{1}{K^e} \sum_i \left(y_{e,i} log \left(p_e^s \right) + (1 - y_{e,i}) log \left(1 - p_e^s \right) \right) \tag{24}
$$

where y_a and y_e are the ground-truth.

The overall loss is the combination of aforementioned losses, which is shown as follows,

$$
\mathcal{L}_{total} = \mathcal{L}_l + \eta \mathcal{L}_p + \mu \left(\mathcal{L}_a + \mathcal{L}_e \right) \tag{25}
$$

where η and μ is the trade-off hyper-parameters.

3.5.2 Inference. we simply take the average of $c_p (V, S)$ and $c_l (V, S)$ as the final video-text similarity between V and S for video-text retrieval.

4 EXPERIMENTS

In this section, we first introduce two datasets (MSR-VTT [45] and VATEX [38]) performed in our experiments. Then we compare our HANet with recent state-of-the-art methods and analyze its effectiveness. We also investigate each component in our HANet by ablation studies.

4.1 Datasets and Evaluation Metrics

4.1.1 Datasets. We conduct extensive experiments on two popular datasets, namely MSR-VTT and VATEX, for the video-text retrieval task.

- **MSR-VTT** dataset [45] is composed of 10000 video clips with 20 text descriptions per clip. We follow the official data split, where 6573, 497 and 2990 videos are used for training, validation and testing, respectively.
- **VATEX** dataset [38] is a large-scale bilingual video description dataset, each clip is accompanied by 10 English text descriptions and 10 Chinese text descriptions. Here, we only use the English text descriptions. Following the partition provided by [4, 9], we use 25991 video clips for training, 1500 clips for validation and 1500 clips for testing, where validation and test sets are obtained from the official validation set since the annotations on test set are private.

4.1.2 Evaluation Metrics. Following prior works, we report the results using the rank-based performance metric, i.e., Recall at K (R@K, K=1, 5, 10, higher is better), Median Rank (MdR, lower is better), and Sum of all Recalls (SumR, higher is better) to measure the overall performance.

4.2 Implementation Details

4.2.1 Video-Text Features. For fair comparison, we apply the same feature in our HANet to all the compared methods. For MSR-VTT,

we utilize the visual feature provided by [4] with dimension of 2048, which is extracted with ResNet152 pre-trained on ImageNet [14]. For VATEX, we use the officially provided I3D [1] video feature. For the text features on both MSR-VTT and VATEX, we set the word embedding size as 300 and initialize with pre-trained Glove embeddings [33].

4.2.2 Concept Vocabulary. The concept vocabulary is constructed from all training sentences. Specifically, we first remove all English stop-words and punctuations, and use NLTK toolkit [2] to find the part-of-speech tags. After that, only nouns and verbs are retained, which correspond to the entities and actions, respectively. To avoid duplication of concepts, we also lemmatize these nouns and verbs by NLTK. Finally, the top K_a =512 frequent verbs and K_e =1024 nouns are selected as the final action and entity concept vocabularies, respectively. Based on data statistics of training samples, we found that the top K_a =512 frequent verbs and K_e =1024 nouns cover the vast majority of high-frequency words.

4.2.3 Training. We implement HANet using PyTorch [3] on the NVIDIA V100 GPU. We use Adam [17] to optimize HANet, with learning rate of 1e-4 and batch size of 64. The maximal number of epochs is set to 50, and early stop occurs if the validation performance (SumR) does not improve in ten consecutive epochs. As for hyper-parameters, N_a and N_e are set to 10 and 20 respectively based on data statistics of training samples. The weight η and μ in the combined losses is empirically set to 0.1 and 0.01, respectively. Following [4], the temperature parameter λ is set to 4, and the margin Δ is set to 0.2.

4.3 Comparison with State-of-the-Art Methods

We compare our HANet with state-of-the-art methods on the MSR-VTT and VATEX datasets. Notably, all comparison methods take the same visual feature as input. The comparison results on the MSR-VTT are shown in Table 1. We observe that HANet achieves significant performance improvements over comparison methods on text-to-video and video-to-text retrieval tasks. All baselines except HGR [4] only use the global features to compute the similarity between the video and text. HGR is similar to our HANet among these baselines, which also decomposes the text and aligins the video and text in a global-to-local fashion. Compared with HGR, our HANet is equipped with concept-guiding video parsing, and aligns videos and texts in a more fine-grained and more precise fashion, i.e., individual level (frame and word), local level (clip and phrase), and global level (video and sentence). As a consequence, HANet significantly outperforms HGR in all evaluation metrics, especially, boosts the overall retrieval quality by a margin of 9.4 in SumR.

To demonstrate the reliability of HANet, we also carry out experiments on another dataset, i.e., VATEX, using different visual features I3D. From results in Table 2, it is also easy to notice that HANet outperforms other existing methods by a large margin. Here HSL [9] stresses the importance of video-text representation and also uses different semantic levels of features. However, HSL simply

---

[2] http://www.nltk.org
[3] https://pytorch.org

Table 1: Comparisons with the state-of-the-art methods on the MSR-VTT dataset.

Method	Text-to-Video				Video-to-Text				SumR
	R@1	R@5	R@10	MdR	R@1	R@5	R@10	MdR	
VSE [18]	5.0	16.4	24.6	47	7.7	20.3	31.2	28	105.2
VSE++ [10]	5.7	17.1	24.8	65	10.2	25.4	35.1	25	118.3
Mithum et al. [29]	5.8	17.6	25.2	61	10.5	26.7	35.9	25	121.7
W2VV [7]	6.1	18.7	27.5	45	11.8	28.9	39.1	21	132.1
Dual Encoding [8]	7.7	22.0	31.8	32	13.0	30.8	43.3	15	148.6
TCE [48]	7.7	22.5	32.1	30	-	-	-	-	-
Zhao et al.[51]	8.8	25.5	36.5	22	14.0	33.1	44.9	14	162.8
HGR [4]	9.2	26.2	36.5	24	15.0	36.7	48.8	11	172.4
HANet	**9.3**	**27.0**	**38.1**	**20**	**16.1**	**39.2**	**52.1**	**9**	**181.8**

Table 2: Comparisons with the state-of-the-art methods on the VATEX dataset.

Method	Text-to-Video				Video-to-Text				SumR
	R@1	R@5	R@10	MdR	R@1	R@5	R@10	MdR	
W2VV [7]	14.6	36.3	46.1	-	39.6	69.5	79.4	-	285.5
VSE++ [10]	31.3	65.8	76.4	-	42.9	73.9	83.6	-	373.9
CE [25]	31.1	68.7	80.2	-	41.3	71.0	82.3	-	374.6
W2VV++ [22]	32.0	68.2	78.8	-	41.8	75.1	84.3	-	380.2
Dual Encoding [8]	31.1	67.4	78.9	-	-	-	-	-	-
HGR [4]	35.1	73.5	83.5	-	-	-	-	-	-
HSL [9]	**36.8**	73.6	83.7	-	46.8	75.7	85.1	-	401.7
HANet	36.4	**74.1**	**84.1**	**2**	**49.1**	**79.5**	**86.2**	**2**	**409.4**

Table 3: Ablation studies on the MSR-VTT dataset to investigate the effectiveness of hierarchical alignment.

Individual Alignment	Local Alignment	Global Alignment	Text-to-Video				Video-to-Text				SumR
			R@1	R@5	R@10	MdR	R@1	R@5	R@10	MdR	
√			8.6	25.7	36.4	23	14.1	36.4	48.6	11	169.8
	√		8.4	24.6	34.9	25	13.2	35.2	47.0	12	163.3
		√	8.4	24.9	35.6	24	14.3	33.9	46.1	13	163.2
√	√		9.2	26.7	37.6	21	15.6	38.8	51.4	10	179.3
√		√	9.1	26.6	37.8	21	15.5	38.0	50.6	10	177.6
	√	√	8.6	25.3	36.2	23	13.6	35.6	48.0	11	167.3
√	√	√	**9.3**	**27.0**	**38.1**	**20**	**16.1**	**39.2**	**52.1**	**9**	**181.8**

Table 4: Ablation studies on the MSR-VTT dataset to investigate the effectiveness of SeMe and relational GCN.

Model	Text-to-Video				Video-to-Text				SumR
	R@1	R@5	R@10	MdR	R@1	R@5	R@10	MdR	
FC → SeMe	9.2	26.7	37.8	21	14.7	38.3	50.9	10	177.6
FC → relational GCN	8.9	26.6	37.6	21	14.9	38.1	50.7	10	176.8
HANet	**9.3**	**27.0**	**38.1**	**20**	**16.1**	**39.2**	**52.1**	**9**	**181.8**

concatenates all levels of features and overlooks the hierarchical alignment.

4.4 Ablation Studies

In this section, we conduct experiments on the MSR-VTT to verify the effectiveness of each component in HANet.

4.4.1 Effectiveness of Hierarchical Alignment. We first investigate the effectiveness of hierarchical alignment. The results of using different alignments of our HANet are shown in Table 3. As we can see that, only using a single alignment achieves worse performance. In particular, only using global alignment is similar to most previous works that simply use a single representation, which results in

Query: A girl playing guitar and singing a song.

Query: An adult brown horse stand in the barn and his father horse jumps a barrier and his mother horse lay his head toward ground and walking.

Query: two teams playing volleyball.

Figure 3: Examples of text-to-video retrieval on the MSR-VTT test set. Top 3 retrieved videos are visualized. (green box: correct; gray box: incorrect)

Figure 4: Examples of concept prediction in video parsing. Four different videos are taken from the MSR-VTT test set. Bigger font meaning larger predicted scores in each sample. (green: action concepts; blue: entity concepts)

Talk	0.68	1.00	0.51	0.85	0.99	0.00	0.04	0.06	0.06	0.10
Perform	0.03	0.12	0.00	0.13	0.24	0.35	1.00	0.96	0.90	0.45
Sing	0.32	0.51	0.00	0.75	0.49	0.86	1.00	1.00	0.97	0.84

Action

Entity

Woman	1.00	0.99	0.87	0.93	0.00	0.99	0.98	0.92	0.99	0.86
Girl	1.00	0.87	0.62	0.90	0.00	0.99	0.95	0.82	0.99	0.87
Man	0.04	0.02	0.01	0.02	1.00	0.01	0.01	0.03	0.00	0.04
Judge	0.37	0.71	1.00	0.00	0.60	0.11	0.06	0.40	0.43	0.06

Figure 5: Examples of concept prediction of video frames in video parsing. A video is taken from the MSR-VTT test set. We normalize the concept confidence to [0,1] for each concept.

an obvious drop of 18.6 in terms of SumR. It convincingly demonstrates the improvement contributed by the proposed hierarchical alignment which provides more fine-grained information. Besides, we notice that any combination of two different alignments can gain performance improvements, which demonstrates all three alignments are effective. Finally, the combination of all three alignments further improves performance, which demonstrates individual, local and global-level information are complementary.

4.4.2 Effectiveness of Local-level Representation. Here we use two FC layers to replace our designed SeMe module in terms of video and relational GCN in terms of text, respectively. We show results in Table 4. Compared with HANet, they both suffer performance degradation, which convincingly demonstrates that, 1) video contextual information is introduced by selecting and merging concept-based relevant frames in our SeMe module; 2) text contextual information is also captured by relational GCN.

4.5 Qualitative Analyses

4.5.1 Visualization of the Text-to-Video Retrieval. We visualize several examples on the MSR-VTT test set for text-to-video retrieval in Figure 3. In the left and middle samples, our HANet successfully retrieve the correct video given query. The right sample shows an ambiguous case, where all top 3 retrieved videos present a scene of "two teams playing volleyball". We argue that this was caused by the instance-based assumption [40] in current video-text retrieval, namely only a single video is relevant to a query. In fact, these three videos can be deemed equally relevant.

4.5.2 Visualization of Action- and Entity-level Concept. Since video parsing in HANet is based on the concept-based classification, in this part, we visualize some examples of concept prediction at the video and frame-level. In Figure 4, we observe that some relevant concepts are predicted with high confidence, for example, "play", "soccer" in the example of the top right corner. However, there are also some irrelevant predicted concepts, such as "tennis" and "throw". In general, concepts predicted by HANet are reasonable, and helpful for understanding the cross-modal retrieval.

The frame-level concept prediction is shown in Figure 5, we highlight the areas with higher concept confidence in red. For the action concept, since we take as input the 5-frame clip to obtain the confidence of action concept, actions, e.g., "talk", "sing", contained in consecutive frames are retrieved. Intriguingly, based on the frame-level concept confidence, not only "woman", "girl" and "man" are retrieved, but also the "judge" that is gender-neutral is retrieved. which demonstrates the reliability of MIL based mechanism in Equation (8) for weakly supervised concept classification as well as the practicability of SeMe module.

5 CONCLUSION

In this paper, we propose the hierarchical alignment network (HANet) to make full use of complementary information of different semantic levels of representations for video-text retrieval. To this end, we first parse the video and text by concept-based weakly supervised classification and existing text parsing toolkit, respectively. Then we introduce the hierarchical alignment to align representations at the individual, local and global levels for computing cross-modal similarity. The quantitative and qualitative results on two popular text-video retrieval benchmarks significantly demonstrate the effectiveness of HANet. In the future, the more precise and efficient hierarchical alignment is yet to be explored since pairwise matching is relatively expensive.

ACKNOWLEDGMENTS

REFERENCES

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

[2] Feiyu Chen, Jie Shao, Yonghui Zhang, Xing Xu, and Heng Tao Shen. 2020. Interclass-Relativity-Adaptive Metric Learning for Cross-Modal Matching and Beyond. *IEEE Transactions on Multimedia* (2020).

[3] Hui Chen, Guiguang Ding, Zijia Lin, Sicheng Zhao, and Jungong Han. 2019. Cross-modal image-text retrieval with semantic consistency. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1749–1757.

[4] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. 2020. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10638–10647.

[5] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[6] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. 2021. Similarity Reasoning and Filtration for Image-Text Matching. *arXiv preprint arXiv:2101.01368* (2021).

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

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

[9] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. 2021. Dual Encoding for Video Retrieval by Text. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[10] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612* (2017).

[11] Zerun Feng, Zhimin Zeng, Caili Guo, and Zheng Li. 2020. Exploiting Visual Semantic Reasoning for Video-Text Retrieval. *arXiv preprint arXiv:2006.08889* (2020).

[12] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. Multimodal transformer for video retrieval. In *European Conference on Computer Vision (ECCV)*, Vol. 5. Springer.

[13] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. 2020. Coot: Cooperative hierarchical transformer for video-text representation learning. *arXiv preprint arXiv:2011.00597* (2020).

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[15] Xiangteng He, Yuxin Peng, and Liu Xie. 2019. A new benchmark and approach for fine-grained cross-media retrieval. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1740–1748.

[16] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.

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

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

[19] Duy-Dinh Le, Sang Phan, Vinh-Tiep Nguyen, Benjamin Renoust, Tuan A Nguyen, Van-Nam Hoang, Thanh Duc Ngo, Minh-Triet Tran, Yuki Watanabe, Martin Klinkigt, et al. 2016. NII-HITACHI-UIT at TRECVID 2016.. In *TRECVID*.

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

[21] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. *arXiv preprint arXiv:2102.06183* (2021).

[22] Xirong Li, Chaoxi Xu, Gang Yang, Zhineng Chen, and Jianfeng Dong. 2019. W2vv++ fully deep learning for ad-hoc video search. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1786–1794.

[23] Xirong Li, Fangming Zhou, Chaoxi Xu, Jiaqi Ji, and Gang Yang. 2020. SEA: Sentence Encoder Assembly for Video Retrieval by Textual Queries. *IEEE Transactions on Multimedia* (2020).

[24] Song Liu, Haoqi Fan, Shengsheng Qian, Yiru Chen, Wenkui Ding, and Zhongyuan Wang. 2021. HiT: Hierarchical Transformer with Momentum Contrast for Video-Text Retrieval. *arXiv preprint arXiv:2103.15049* (2021).

[25] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. 2019. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487* (2019).

[26] Foteini Markatopoulou, Damianos Galanopoulos, Vasileios Mezaris, and Ioannis Patras. 2017. Query and keyframe representations for ad-hoc video search. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*. 407–411.

[27] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9879–9889.

[28] Antoine Miech, Ivan Laptev, and Josef Sivic. 2018. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516* (2018).

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

[30] Phuong Anh Nguyen, Qing Li, Zhi-Qi Cheng, Yi-Jie Lu, Hao Zhang, Xiao Wu, and Chong-Wah Ngo. 2017. VIREO@ TRECVID 2017: Video-to-Text, Ad-hoc Video Search, and Video hyperlinking.. In *TRECVID*.

[31] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, João Henriques, and Andrea Vedaldi. 2020. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824* (2020).

[32] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. 2018. W-talc: Weakly-supervised temporal activity localization and classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 563–579.

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

[34] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*. Springer, 593–607.

[35] Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255* (2019).

[36] Yale Song and Mohammad Soleymani. 2019. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1979–1988.

[37] Kazuya Ueki, Koji Hirakawa, Kotaro Kikuchi, Tetsuji Ogawa, and Tetsunori Kobayashi. 2017. Waseda_Meisei at TRECVID 2017: Ad-hoc Video Search.. In *TRECVID*.

[38] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4581–4591.

[39] Jiwei Wei, Xing Xu, Yang Yang, Yanli Ji, Zheng Wang, and Heng Tao Shen. 2020. Universal weighting metric learning for cross-modal matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13005–13014.

[40] Michael Wray, Hazel Doughty, and Dima Damen. 2021. On Semantic Similarity in Video Retrieval. *arXiv preprint arXiv:2103.10095* (2021).

[41] Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. 2019. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 450–459.

[42] Jiaxin Wu and Chong-Wah Ngo. 2020. Interpretable embedding for ad-hoc video search. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3357–3366.

[43] Peng Wu and Jing Liu. 2021. Learning Causal Temporal Relation and Feature Discrimination for Anomaly Detection. *IEEE Transactions on Image Processing* 30 (2021), 3513–3527.

[44] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. 2020. Not only Look, but also Listen: Learning Multimodal Violence Detection under Weak Supervision. In *European Conference on Computer Vision*. Springer, 322–339.

[45] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5288–5296.

[46] Ruicong Xu, Li Niu, Jianfu Zhang, and Liqing Zhang. 2020. A Proposal-Based Approach for Activity Image-to-Video Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 12524–12531.

[47] Ran Xu, Caiming Xiong, Wei Chen, and Jason Corso. 2015. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29.

[48] Xun Yang, Jianfeng Dong, Yixin Cao, Xun Wang, Meng Wang, and Tat-Seng Chua. 2020. 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*. 1339–1348.

[49] Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. 2017. End-to-end concept word detection for video captioning, retrieval, and question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3165–3173.

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

[51] Rui Zhao, Kecheng Zheng, and Zheng-jun Zha. 2020. Stacked Convolutional Deep Encoding Network For Video-Text Retrieval. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.