HiT: Hierarchical Transformer with Momentum Contrast for Video-Text Retrieval

Song Liu[1], Haoqi Fan[2], Shengsheng Qian[3], Yiru Chen[4], Wenkui Ding[4], Zhongyuan Wang[4]
[1]Peking University, [2]FAIR
[3]Institute of Automation, Chinese Academy of Sciences
[4]Kuaishou
shawnliu@pku.edu.cn, haoqifan@fb.com, shengsheng.qian@nlpr.ia.ac.cn
{chenyiru, dingwenkui, wangzhongyuan}@kuaishou.com

Abstract

Video-Text Retrieval has been a hot research topic with the explosion of multimedia data on the Internet. Transformer for video-text learning has attracted increasing attention due to the promising performance. However, existing cross-modal transformer approaches typically suffer from two major limitations: 1) Limited exploitation of the transformer architecture where different layers have different feature characteristics. 2) End-to-end training mechanism limits negative interactions among samples in a mini-batch. In this paper, we propose a novel approach named Hierarchical Transformer (HiT) for video-text retrieval. HiT performs hierarchical cross-modal contrastive matching in feature-level and semantic-level to achieve multi-view and comprehensive retrieval results. Moreover, inspired by MoCo, we propose Momentum Cross-modal Contrast for cross-modal learning to enable large-scale negative interactions on-the-fly, which contributes to the generation of more precise and discriminative representations. Experimental results on three major Video-Text Retrieval benchmark datasets demonstrate the advantages of our methods.

1. Introduction

Cross-modal Retrieval [57, 11, 14, 66, 8, 3, 36, 10, 12, 45, 24, 9, 58, 59, 27, 56] has attracted increasing attention with the aim to search the semantic similar instances from different modalities. The explosive growth of video contents on the Internet has brought great challenges to accurate video-text retrieval. This paper studies the learning of video-text retrieval.

Recent works [48, 69, 14, 40, 13] have shown that Transformer can learn high level video representations and capture semantically meaningful and temporally long-range structure for videos. It is feasible and efficient to adopt Transformer architectures to encode video and text features in video-text learning. This paper proposes Hierarchical Transformer (HiT) to achieve video-text retrieval by Hierarchical Cross-modal Contrastive Matching, which consists of semantic-level contrastive matching and feature-level contrastive matching. In addition, to enable large-scale negative interactions on-the-fly, Momentum Cross-modal Contrast is introduced.

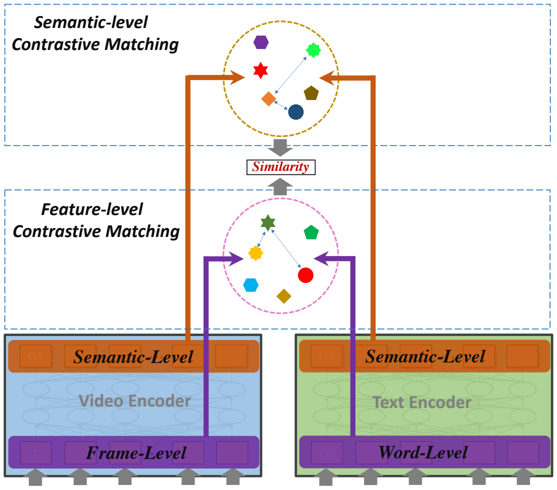


Figure 1. Hierarchical Cross-modal Contrastive Matching consists of Feature- and Semantic-level Contrastive Matching. *Notably, Momentum Cross-modal Contrast is not shown in this figure.*

Hierarchical Cross-modal Contrastive Matching. According to the attention allocated characteristics of different layers in Transformer architectures, the features in different layers emphasize different views for samples [16, 42, 50, 30]. For example, the features in lower layers tend to encode more local contents with basic syntactic representations. Higher layers capture more complex semantics and usually produce higher-level semantic representations, as most works [13, 40] performed. Based on these specialities, we propose Hierarchical Cross-modal Contrastive Matching to achieve comprehensive and accurate matching for cross-modal samples hierarchically. As the first contribution, a novel and simple cross-modal Transformer architecture named Hierarchical Transformer (HiT) is proposed, which is designed as Figure 1.

Momentum Cross-modal Contrast. Recently, a class of self-supervised methods for unsupervised visual representation learning [61, 5, 17, 4] emphasize the necessity of large-scale negative samples. Inspired by these works, we argue that large-scale negative interactions in training process have been neglected in cross-modal contrastive learning. In this paper, we introduce MoCo [17, 5] into HiT to enable large-scale negative interactions on-the-fly. We refer it as Momentum Cross-modal Contrast (MCC). In MCC, we build several negative memory banks to enable broader negative interactions during training. In addition, since video and text encoders with gradient descent (query encoders) are updated dramatically, they would largely affect the performance due to the result of representation inconsistency in banks. Thus, key encoders for two modalities with momentum update are required to maintain representation consistency.

In brief, the contributions are listed as follows:

- We propose Hierarchical Transformer (HiT), which jointly performs Hierarchical Cross-modal Contrastive Matching and Momentum Cross-modal Contrast.
- Hierarchical Cross-modal Contrastive Matching is presented to achieve multi-view and comprehensive video-text retrieval learning.
- Momentum Cross-modal Contrast (MCC) introduces MoCo into the cross-modal learning field. It enables large-scale negative interactions to obtain precise and discriminative representations.
- Extensive experiments demonstrate the advantages of the proposed methods on three benchmarks, including MSR-VTT, ActivityNet and LSMDC.

2. Related Work

2.1. Video-Text Retrieval

Video-Text Retrieval has received wide attention with the exploitation of the huge multimedia data and rich application scenarios. Several excellent works [66, 8, 45, 57, 11, 14, 8, 3, 36, 10, 12] are introduced to address this task. JSFusion [66] proposes a joint sequence fusion model for sequential interaction of videos and texts. Dual Encoding [8] consists of mean pooling, biGRU and CNN models to encode sequential videos and texts in multiple levels. PVSE [45] presents a polysemous instance embedding network to learn multiple and diverse representations of videos and texts for the polysemous problem. A graph-based framework is proposed in [64] for matching between movie segments and synopsis paragraphs, which takes into account both the flow of events and the interactions among characters. HGR [3] is a Hierarchical Graph Reasoning model, which decomposes video-text matching into global-to-local levels and disentangles texts into a hierarchical semantic graph including three levels of events, actions and entities.

2.2. Transformer for Video-Text Learning

BERT [54] is a Transformer-based representation model for natural language process tasks. It falls into a line of works that learn a universal language encoder by pretraining with language modeling objectives. Recently, several attempts [30, 33, 46, 49, 13, 48, 69, 29, 20, 28] have been made which utilize BERTs and Transformers as the backbone for cross-modal tasks. In video-text learning tasks, VideoBERT [48] transforms a video into spoken words paired with a series of images and applies a Transformer to learn joint representations. ActBERT [69] learns a joint video-text representation that uncovers global and local visual clues from paired video sequences and text descriptions. Both the global and the local visual signals interact with the semantic stream mutually. MMT [13] proposes the multi-modal transformer which performs the task of processing features extracted from different modalities at different moments in videos, such as video, audio and speech. COOT [14] proposes a hierarchical model that exploits long-range temporal context in videos and texts, and uses intra-level and inter-level cooperation for producing the final video/text embeddings based on interactions between local and global contexts. Support-set [40] incorporates a auxiliary generative task, *i.e.,* cross-captioning task, to alleviate mismatching problems existed in recent works. Very recently, T2VLAD [62] uses a paradigm of global-local alignment to perform video retrieval, which is similar to our Hierarchical Cross-modal Contrastive Matching to some extent. However, we argue that it performs their global-local alignment in a very different way. We exhaustively delve deep into the potential of hierarchical matching in ablation study. Moreover, we further propose Momentum Cross-modal Contrast that enables more precise and comprehensive cross-modal contrastive learning.

2.3. Contrastive Learning

Contrastive Learning [4, 17, 5, 52, 22, 53, 37, 60, 6, 15] has made remarkable progress in unsupervised visual representation learning. We introduce several representative negative-with contrastive learning mechanisms which differ in contrastive samples construction and representation encoder updating method. In the *end-to-end* mechanism, the encoders used to learn query and key representations are updated end-to-end by back-propagation. In the *memory*

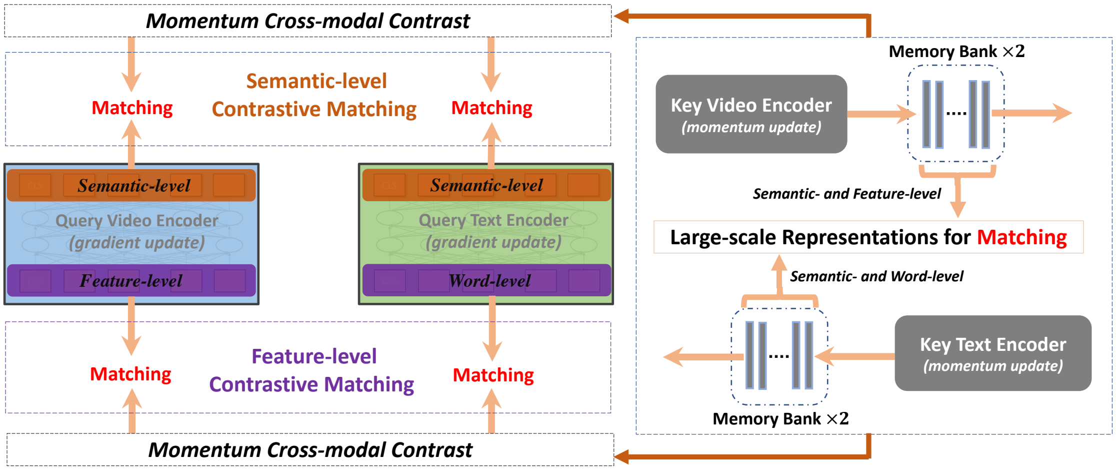


Figure 2. The pipeline of HiT. All encoders adopt transformer based architectures. **Video Encoding**: Query Video Encoder and Key Video Encoder. **Text Encoding**: Query Text Encoder and Key Text Encoder. **Momentum Cross-modal Contrast**: Four memory banks are built to save the key semantic-level and key feature-level representations. Two query encoders are gradient updated and two key encoders are momentum updated. **Hierarchical Cross-modal Contrastive Matching**: Semantic-level Contrastive Matching is performed between query video (text) semantic-level representations and key text (video) semantic-level representations in memory banks. Feature-level Contrastive Matching is performed between query video (text) feature-level representations and key text (video) feature-level representations in memory banks.

bank [61] mechanism, the key representations generated by the dynamically updated query encoder are sampled from a memory bank without the good representation consistency. MoCo [5, 17] mechanism obtains the new key representation on-the-fly by a momentum-updated key encoder, and maintains a dictionary as a queue to allow the training process to reuse the encoded key representations from the immediate preceding mini-batches. SimCLR [4] shows that contrastive learning in unsupervised visual representation learning benefits from large batch size negatives, stronger data augmentation and introducing a learnable nonlinear transformation, *i.e.,* using projection heads. Though recent works [6, 15] argue that contrastive learning can achieve decent performance even without negatives by using a momentum encoder or stop gradient to prevent collapse solutions, our HiT in video-text learning and [17, 4, 5, 61, 21] in visual representation learning indeed benefit from the large-scale negative interactions. The effects of cross-modal learning without negatives are not involved in this paper.

3. Problem Definition

For Video-Text Retrieval task, we are given M videos $V = \{V_i\}_{i=0}^{M-1}$ and N captions $T = \{T_i\}_{i=0}^{N-1}$. Each video has several kinds of expert embeddings to represent videos in multiple views, e.g., motion, appearance and audio. Each caption is represented by the natural language in English. Formally, the target of our methods for video-text retrieval is to obtain two *query encoders* $f\colon V \to \mathbf{Z} = \{Z_i\}_{i=1}^{l}$ and $g\colon T \to \mathbf{Z} = \{Z_i\}_{i=1}^{l}$ jointly, where V and T are video and text domains respectively, and $\mathbf{Z}$ consists of l common embedding spaces. In the common embedding spaces, cross-modal instances are represented by a series of compact embeddings. Meanwhile, the distance among similar cross-modal instances are smaller than that of among dissimilar cross-modal instances in the common embedding spaces. The constraint can be formulated as follows:

$$d(f(V_i), g(T_i)) < d(f(V_i), g(T_j)) \ s.t. \ i \neq j \qquad (1)$$

where $d(\cdot,\cdot)$ is a distance measurement. The overall similarity between two cross-modal instances is decided by hierarchical contrastive matching results.

4. Hierarchical Transformer

Figure 2 illustrates the structure of the Hierarchical Transformer (HiT) for video-text retrieval. For video encoding, there are *Query Video Encoder* and *Key Video Encoder*. Both two video encoders utilize the same architecture. For text encoding, there are *Query Text Encoder* and *Key Text Encoder* that adopt the same architecture. Notably, Siamese encoders, *a.k.a.,* key encoders, are shown for the utilization of Momentum Cross-modal Contrast (MCC), which will be discussed later. There are only two query encoders left if we remove MCC, as shown in Figure 1.

4.1. Video Encoders

The video encoders, including query and key video encoders, are designed as Transformer based architectures and

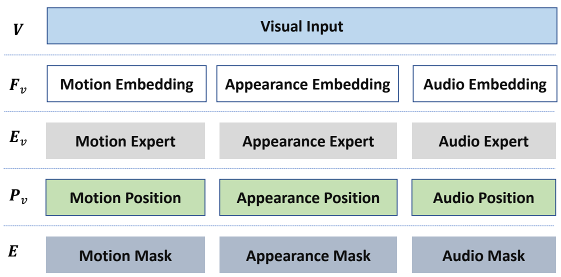


Figure 3. The visual input of video encoders.

we transform the raw visual features into a discrete sequence of tokens as inputs. To this end, we generate a sequence of pre-trained video-related features, including motion, appearance and audio features, to obtain *Visual Embeddings* $\mathbf{F}_v$ as the inputs. *Visual Segment Masks* $\mathbf{M}_v$ and *Visual Position Embeddings* $\mathbf{P}_v$ are needed to indicate the real numbers and positions of input features respectively. We append *Expert Embeddings* $\mathbf{E}$ to identity the attending expert. The final visual input $\mathbf{V}$ can be formulated as follows, also shown in Figure 3:

$$\mathbf{V} = \mathbf{F}_v + \mathbf{M}_v + \mathbf{P}_v + \mathbf{E} \tag{2}$$

• **Video Feature-level Feature.** As studied in [51, 41, 55], in the Transformer-based architectures, the features in lower layers capture low-level patterns that describe basic syntactic information. We obtain these visual token features in the first layer of the query video encoder and the key video encoder. Then we do *Average Pooling* and *Nonlinear Projection* for them and obtain $v_f^q \in \mathbb{R}^{D_v}$ and $v_f^k \in \mathbb{R}^{D_v}$, respectively. MLPs are adopted as the nonlinear projection heads to do nonlinear transformations. [4] has proved that a nonlinear projection head can improve the representation quality of the layer before it.

• **Video Semantic-level Feature.** Higher layer features in Transformer-based architectures capture higher-level representations with more complex semantic meanings. We do mean pooling for the contextual tokens in the last layer to represent the semantic-level features. Then two projection heads are used to do nonlinear transformations for obtaining $v_s^q \in \mathbb{R}^{D_v}$ and $v_s^k \in \mathbb{R}^{D_v}$ generated by the query video encoder and the key video encoder respectively.

4.2. Text Encoders

We leverage `BERT-base-uncased` [7] as the text encoders and fine-tune it. It's worth noting that the video features are generated by pre-trained deep neural networks and already have higher level semantic representation ability. While the text modality has different inherent complexity from the video modality and needs more Transformer blocks to model semantic relations among words. Thus, text encoders are deeper than video encoders.

Each word in a caption will be embedded into a D_t-dimension word embedding vector and we obtain *Token Embeddings* $\mathbf{F}_t$. `[CLS]` and `[END]` are embedded into the first and last positions. *Text Segment Mask* is needed to indicate the real length of the input sequence. *Text Position Embedding* is used to represent the word indexes of the input sequence in text encoders. The final input for text encoders is defined as:

$$\mathbf{T} = \mathbf{F}_t + \mathbf{M}_t + \mathbf{P}_t \tag{3}$$

• **Text Word-level Feature.** We obtain text word-level features from the first layer of query text encoder and key text encoder. Similar to the acquisition of video feature-level features, we utilize two projection heads to do nonlinear transformations and obtain $t_w^q \in \mathbb{R}^{D_t}$, $t_w^k \in \mathbb{R}^{D_t}$.

• **Text Semantic-level Feature.** The mean pooling of token features from the last layer are referred as text semantic-level features. These contextual tokens represent the higher-level meaning of the given caption. Two projection heads are used to do nonlinear transformations for obtaining $t_s^q \in \mathbb{R}^{D_s}$ and $t_s^k \in \mathbb{R}^{D_s}$.

4.3. Momentum Cross-modal Contrast

The end-to-end training as most methods do largely limits the negative interactions. To enable large-scale negative interactions for generating more precise and discriminative representations, Momentum Cross-modal Contrast (MCC) is proposed. Four memory banks are firstly built as queues for saving negative representations dynamically.

• **Text Memory Banks.** Text memory banks B_T^w for storing key text word-level features and B_T^s for storing key text semantic-level features are built as two queues. In per training iteration, the current mini-batch key text representations t_w^k (or t_s^k) encoded by the key text encoder will be enqueued into B_T^w (or B_T^s) and the oldest mini-batch will be dequeued. The key text representations in B_T^w and B_T^s will be interacted with the current mini-batch video representation v_f^q and v_s^q encoded by the query video encoder.

•**Video Memory Banks.** Similarly, video memory banks B_V^f for saving key video feature-level features v_f^k, and B_V^s for saving key video semantic-level features v_s^k are built as two queues.

Moreover, to maintain the representation consistency in the memory banks, two key encoders, which perform momentum update [17, 5], are required. We denote θ_q^v and θ_k^v as the parameters of the query and key video encoders. While θ_q^t and θ_k^t are the parameters of the query and key text encoders. We formulate the momentum update for θ_k^v and θ_k^t as:

$$\begin{aligned} \theta_k^v &\leftarrow m\theta_k^v + (1-m)\theta_q^v \\ \theta_k^t &\leftarrow m\theta_k^t + (1-m)\theta_q^t \end{aligned} \tag{4}$$

where $m \in [0, 1)$ is a momentum coefficient, which is a relatively large value. We set $m = 0.999$ in this paper. The parameters θ_q^v and θ_q^t are updated by back-propagation. The momentum update makes θ_k^v and θ_k^t evolve more smoothly than θ_q^v and θ_q^t. As a result, though the key representations in the memory banks are encoded by different encoders (in different mini-batches), the difference among these encoders can be made small.

4.4. Hierarchical Cross-modal Contrastive Matching

We propose hierarchical cross-modal contrastive matching for video-text retrieval learning. Specifically, we utilize video feature-level features and text word-level features for feature-level contrastive matching. The video and text semantic-level features are used for semantic-level contrastive matching.

Feature-level Contrastive Matching. For the view of retrieval texts with videos, we achieve *positive similarity* s^{vt+} by calculating cosine similarity between v_f^q and t_w^k. Then, we obtain *negative similarity* $S_{vt-} = \{s_i^{vt-}\}_{i=1}^{K_t}$ by calculating cosine similarity among v_f^q and all key text representations in B_T^w. Thus, we achieve $S_{vt} = [s^{vt+}, S_{vt-}] \in \mathbb{R}^{1+K_t}$, where K_t is the queue size of B_T^f. Similarly, for the view of retrieval videos with texts, we get $S_{tv} = [s^{tv+}, S_{tv-}] \in \mathbb{R}^{1+K_v}$, where K_v is the queue size of B_V^f.

The InfoNCE [39], a form of contrastive loss functions, is adopted as our objective function for feature-level contrastive matching:

$$L_1 = -log\frac{\exp(s_{vt+}/\gamma)}{\sum_{i=0}^{K_t}\exp(S_{vt}/\gamma)} - log\frac{\exp(s_{tv+}/\gamma)}{\sum_{i=0}^{K_v}\exp(S_{tv}/\gamma)} \tag{5}$$

where γ is a temperature hyper-parameter, which is set to 0.07 in this paper.

Semantic-level Contrastive Matching. We achieve positive and negative similarity $C_{vt} = [c^{tv+}, C_{tv-}] \in \mathbb{R}^{1+K_t}$ and $C_{tv} = [c^{tv+}, C_{tv-}] \in \mathbb{R}^{1+K_v}$. The objective function L_2 of semantic-level contrastive matching is defined as:

$$L_2 = -log\frac{\exp(c_{vt+}/\gamma)}{\sum_{i=0}^{1+K_t}\exp(C_{vt}/\gamma)} - log\frac{\exp(c_{tv+}/\gamma)}{\sum_{i=0}^{1+K_v}\exp(C_{tv}/\gamma)} \tag{6}$$

Thus, the overall objective function is L.

$$L = \alpha L_1 + \beta L_2 \tag{7}$$

where α and β are two hyper-parameters to balance two objectives. We set both α, β to 1 in our experiments.

5. Experiments

5.1. Datasets and Evaluation Metrics

We adopt video-text retrieval experiments on three datasets. Pre-training experiments are conducted on HowTo100M [36].

- **MSR-VTT** [65] contains 10,000 videos, where each video is annotated with 20 captions in English. We follow the training protocol defined in [13, 31, 36] to evaluate on text-to-video and video-to-text retrieval tasks on the 1k-A testing split with 1,000 video or text candidates defined by [66].
- **ActivityNet Captions** [26] consists of 20K YouTube videos temporally annotated with sentence descriptions. We follow the approach of [47, 13], where all the descriptions of a video are concatenated to form a paragraph. The training set has 10,009 videos. We evaluate our video-paragraph retrieval on the "val1" split (4917 videos).
- **LSMDC** [43] contains 118,081 short video clips (∼45s) extracted from 202 movies. Each clip is annotated with a caption, extracted from either the movie script or the audio description. The testing set is composed of 1,000 videos, from movies not present in the training set.
- **Metric.** We measure the retrieval performance with common metrics in information retrieval, including Recall at K (R@K and K=1, 5, 10), and Median Rank (MedR). R@K is the percentage of test queries that at least one relevant item is found among the top-K retrieved results. The MedR measures the median rank of correct items in the retrieved ranking list, where lower score indicates a better model. We also take the sum of all R@K as rsum to reflect the overall retrieval performance.

5.2. Implementation Details

- **Features.** We follow MMT [13] to conduct pre-trained feature extraction. Motion features are extracted from S3D [63] trained on the Kinetics action recognition dataset. Audio features are extracted using VGGish model [18] trained on YT8M. Appearance features are extracted from the final global average pooling layer of SENet-154 [19] trained for classification on ImageNet.

For MSRVTT and LSMDC, we use all motion, appearance and audio experts. We employ 30 features for each type of visual features as the visual inputs, and first the 25 wordpieces from captions as the text inputs. For HowTo100M and ActivityNet, we only use motion and audio experts, and limit our visual input to 100 features per expert and our text input to the first 100 wordpieces.

- **Backbone.** Text encoders `BERT-base-uncased` [7] have 12-layers and we finetune it. Video encoders have 4 transformer layers with 4 attention heads. The hidden size and the intermediate size are set to 512. We set the hidden size of projection heads to 8,192. D_v and D_t are both set to

Table 1. The experimental results on MSR-VTT. Larger R@1,R@5,R@10 and smaller MedR indicate better retrieval performance.

Methods	Video-to-Text Retrieval				Text-to-Video Retrieval				rsum
	R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR	
AM [2]	6.8	18.1	26.5	42	7.0	18.1	27.0	40	103.5
LJE [38]	9.2	27.6	39.1	22	6.9	22.5	29.8	32	134.9
ActBERT [69]	-	-	-	-	8.6	23.4	33.1	36	-
JSFusion [66]	9.5	28.6	40.2	18	9.6	29.8	42.1	20	159.8
HowTo100M [36]	12.2	33.5	47.5	13	12.6	36.2	48.1	13	190.2
CE [31]	20.9	48.8	62.4	6	20.6	50.3	64.0	5.3	267.0
MMT [13]	24.4	56.0	67.8	4	24.6	54.0	67.1	4	293.9
SUPPORT-SET [40]	26.6	55.1	67.5	3	27.4	56.3	67.7	3	300.6
HiT	**28.8**	**60.3**	**72.3**	**3**	**27.7**	**59.2**	**72.0**	**2.9**	**320.3**
HowTo100M [36]	16.8	41.7	55.10	8	14.9	40.2	52.8	9	221.5
NoiseEstimation [1]	-	-	-	-	17.4	41.6	53.6	8	-
UniVL [34]	-	-	-	-	21.2	49.6	63.1	6	-
AVLnet [44]	28.5	54.6	65.2	4	27.1	55.6	66.6	4	297.5
MMT [13]	27.0	57.5	69.7	3.7	26.6	57.1	69.6	4	307.5
SUPPORT-SET [40]	28.5	58.6	71.6	3	30.1	58.5	69.3	3	316.6
HiT Pre-trained on HT100M	**32.1**	**62.7**	**74.1**	**3**	**30.7**	**60.9**	**73.2**	**2.6**	**333.7**

Table 2. Text-to-video retrieval results on ActivityNet.

Methods	R@1	R@5	R@50	MedR
FSE [68]	18.2	44.8	89.1	7.0
CE [31]	18.2	47.7	91.4	6.0
HSE [68]	20.5	49.3	-	-
MMT [13]	22.7	54.2	93.2	5.0
SUPPORT-SET [40]	26.8	58.1	93.5	**3.0**
HiT	**27.7**	**58.6**	**94.7**	4.0
HiT Pre-trained	**29.6**	**60.7**	**95.6**	**3.0**

Table 3. Text-to-video retrieval results on LSMDC.

Methods	R@1	R@5	R@10	MedR
CT-SAN [67]	5.1	16.3	25.2	46
JSFusion [66]	9.1	21.2	34.1	36
CCA [25]	7.5	21.7	31.0	33
MEE [35]	9.3	25.1	33.4	27
MEE-COCO [35]	10.1	25.6	34.6	27
CE [31]	11.2	26.9	34.8	25.3
MMT [13]	13.2	29.2	38.8	21.0
HiT	**14.0**	**31.2**	**41.6**	**18.5**

2,048. The ReLU is used as the activation function and BN layers are used in hidden layers.

- **Optimization.** The initial learning rate is set to 2e-5 and the network is optimized by AdamW [32] optimizer. The batch size K is 256 and train 40 epochs. Pre-training requires 8 GPUs and others require 1 GPU. All experiments are conducted on NVIDIA 3090Ti GPUs.
- K_v **and** K_t **in MCC .** For MSR-VTT, we report retrieval results when we set K_v and K_t to 4,096. K_v and K_t in ActivityNet are set to 512. In LSMDC, K_v and K_t are 1024.

5.3. Compare to state-of-the-arts

The Table 1- 3 present the retrieval results of HiT on MSR-VTT, ActivityNet Captions and LSMDC under the data-specific traning and pre-training protocols. We also compare HiT with other state-of-the-art methods.

As shown in the results, HiT outperforms all comparison methods by a clear margin. For MSR-VTT, we report video-to-text retrieval and text-to-video retrieval results. In particular, our retrieval performance at rsum is 320.3, exceeding recent state-of-the-art methods [40] by a margin of 20.3%. It well reflects the overall retrieval quality of HiT. With pre-training on HowTo100M, HiT further boosts the retrieval performance. For ActivityNet Captions and LSMDC, we report the retrieval performance in terms of text-to-video retrieval. HiT still outperforms comparison method. We find that the growth of retrieval performance benefits from the proposed components, including Hierarchical Cross-modal Contrastive Matching and Momentum Cross-modal Contrast. To demonstrate the effectiveness and robustness of two components, we exhaustively and comprehensively ablate HiT in the following sections.

6. Ablation Study

Hierarchical Matching. As mentioned above, we use token features from the first layers to perform Feature-level Contrastive Matching while token features from the last layers are adopted for Semantic-level Contrastive Matching. In this section, we design several variants to verify the impacts of Hierarchical Cross-modal Contrastive Matching. *Note that we do not perform MCC here for efficiency.*

- **HiT**-*sl*. We only implement semantic-level matching

Table 4. Ablation study on MSR-VTT to investigate contributions of Momentum Cross-modal Contrast.

Methods	**Memory Bank**			**Video-to-Text Retrieval**			**Text-to-Video Retrieval**			**rsum**
	Use	Qk	Qv	R@1	R@5	R@10	R@1	R@5	R@10	
HiT *w/o MCC*	✗	-	-	27.1	55.3	68.3	27.0	58.0	70.8	306.5
HiT *w MCC*	✓	256	256	26.9	56.1	69.0	27.0	58.6	71.0	308.6
HiT *w MCC*	✓	512	512	27.6	58.3	70.0	27.4	58.7	70.8	312.8
HiT *w MCC*	✓	1,024	1,024	27.7	57.9	70.3	27.3	**59.7**	71.8	314.7
HiT *w MCC*	✓	2,048	2,048	28.0	59.6	71.9	27.4	59.0	71.5	317.4
HiT *w MCC*	✓	4,096	4,096	**28.8**	**60.3**	72.3	**27.7**	59.2	**72.0**	**320.3**
HiT *w MCC*	✓	8,192	8,192	28.1	58.9	**72.5**	27.0	58.7	71.0	316.2

Table 5. The investigation of Hierarchical Cross-modal Contrastive Matching in Text-to-Video Retrieval.

Methods	R@1	R@5	R@10	MedR
HiT-*sl*	23.5	56.2	68.8	4
HiT-*fl*	25.1	53.6	67.2	4.3
HiT-*4-level*	27.1	**59.2**	71.0	2.7
HiT-*3-level-a*	**28.5**	58.4	71.0	**2.3**
HiT-*3-level-b*	26.7	58.5	**71.4**	2.7
HiT	27.0	58.0	70.8	3

while feature-level matching is removed.

- **HiT**-*fl*. Only feature-level matching is implemented.
- **HiT**-*4-level*. *To investigate the potential of hierarchical matching for transformer architectures, contrastive matching with respect to more levels is conducted.* Since a text encoder has 12 transformer blocks and a video encoder has 4 blocks, except feature-level (between *layer-1 in text encoder* and *layer-1 in video encoder*) and semantic-level (between *layer-12 in text encoder* and *layer-4 in video encoder*), we append contrastive matching with more levels between *layer-5 in text encoder* and *layer-2 in video encoder*, *layer-9 in text encoder* and *layer-3 in video encoder*.
- **HiT**-*3-level-a*. We append contrastive matching between *layer-9 in text encoder* and *layer-3 in video encoder*.
- **HiT**-*3-level-b*. Contrastive matching between *layer-5 in text encoder* and *layer-2 in video encoder* is appended.
- **HiT**. Original HiT in Table 1.

Table 5 presents the ablation results on MSR-VTT in text-to-video retrieval. We find that using more levels to conduct contrastive matching is able to obtain clear improvements. However, n-level matching requires n times retrieval during inference. In addition, significant improvements are not shown in 3-level and 4-level matching results. For the sake of retrieval efficiency and efficient training with Momentum Cross-modal Contrast, we select 2-level matching in this paper to report the main results.

Momentum Cross-modal Contrast. To explore the impacts of the size of the memory banks, sufficient experiments are carried out. The experimental results are shown in Table 4. We vary the queue size of K_v and K_t from 0 to 8,192 and evaluate R@K and rsum. As shown in the results, it deserves attention that the introduction of large-scale negatives for similarity learning indeed achieves considerable performance improvements, in which we attribute it to broader negative interactions for obtaining precise and discriminative representations. In addition, with the growth of queue size K_v and K_t, retrieval performance is slightly degraded after the growth which is probably due to some positive samples are misclassified as negative samples.

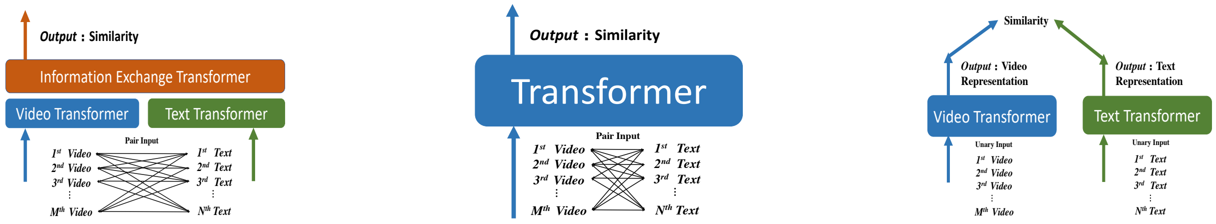
Momentum Encoders. For maintaining representation consistency in memory banks, we introduce two key encoders for two modalities where momentum update is performed. In this section, we abate two momentum encoders to explore their effectiveness in terms of maintaining representation consistency by evaluating the retrieval performance. We achieve the ablation by using query encoders to produce representations for memory banks. Table 6 presents the ablation results. We can find that it shows the degraded performance when we do not use momentum encoders. Particularly, it degrades performance at R5 to 48.4%, which clearly demonstrates the necessity of momentum encoders.

Table 6. The impacts of Momentum Encoders for generating key representations.

Encoders	R@1	R@5	R@10	MedR
Query Encoders	21.1	48.4	60.9	6
Key Encoders	**27.7**	**59.2**	**72.0**	**2.6**

Contrastive Loss. In Equation 5 and 6, InfoNCE is adopted as the Contrastive losses to perform common space learning. In this section, we use another commonly used loss function, *i.e.,* Triplet Ranking Loss, as the objectives and present the retrieval performance for MSR-VTT in Table 7. Though existing the difficulty in tuning the right combination of temperature and batch-size as analysed in [40], we find that InfoNCE achieves better performance than Triplet Ranking Loss in HiT, likely due to the better adaptiveness of InfoNCE in MCC.

The temperature γ in InfoNCE is a sensitive parameter. To show how γ affects retrieval performance, the impacts



(a) Two-stream (b) Single-stream (c) Dual Streams

Figure 4. Three types of transformer based approaches for cross-modal learning. Assume that there are M videos and N texts, (a) Two-stream and (b) Single-stream require pairwise inputs with $\mathcal{O}(MN)$ time complexity. (c) Dual Streams require unary inputs with $\mathcal{O}(M+N)$ time complexity.

Table 7. The selection of Contrastive losses.

Encoders	R@1	R@5	R@10	MedR
Triplet Ranking Loss	25.6	56.7	69.1	5
InfoNCE	**27.7**	**59.2**	**72.0**	**2.6**

of γ with regard to rsum are presented in Table 8. We can observe that the best performance can be achieved when we set γ to 0.07.

Table 8. Parameter analysis for temperature γ.

γ	0.0007	0.007	0.07	0.7	7
rsum	285.1	311.2	**320.3**	155.4	112.2

Expert Utilization. In MSR-VTT, we use three types of expert embeddings as the visual inputs, including motion features, appearance features and audio features. This section explores the comparison of the different experts. The comparison results are in Table 9.

Table 9. Ablation study on different experts.

Experts	R@1	R@5	R@10	MedR
Motion only	25.1	51.6	65.0	4
Appearance only	18.2	41.9	55.5	5
Audio only	10.9	22.1	31.1	14
Motion + Appearance	24.2	52.5	65.1	4.3
Motion + Audio	**28.1**	57.8	71.5	3.0
Appearance + Audio	20.1	46.9	58.7	5
All	27.7	**59.2**	**72.0**	**2.6**

From the results, we find that the motion expert provides the best results when we only use one of three experts. Using audio features solely shows the worst performance. When using two experts, the combination of Motion and Audio features achieves best results. We note that audio features contribute most when used in conjunction with the others, most likely due to the complementary cues it provides, compared to the other experts.

Feature Aggregation. As illustrated in section 4.1 and 4.2, we leverage *Mean Pooling* to produce aggregated features before projection heads, in the sense of capturing important features from tokens. In this section, we evaluate three more aggregation methods, including *Max Pooling*, *1D-CNN* [23] (kernel sizes: [2,3,4,5]) and using a `[CLS]` aggregated token. To obtain aggregated visual features from `[CLS]` token, similar to the text inputs, here we need embed a `[CLS]` and a `[END]` token into the first and last positions of the visual inputs. We initialize them with random vectors. Table 10 presents comparison results in terms of text-video retrieval. It notes that the decent results are not presented in `[CLS]`. We suppose the reason is that the features are not well aggregated in the `[CLS]` of feature-level.

Table 10. Feature aggregation method comparison.

Aggregation	R@1	R@5	R@10	MedR
Mean Pooling	27.7	59.2	72.0	2.6
Max Pooling	26.8	60.1	71.2	3.0
1D-CNN	24.4	55.6	68.2	4.0
`[CLS]`	24.2	53.1	65.0	5.5

7. Retrieval Complexity Analysis

In this section, we analyze the retrieval complexities of recent transformer based state-of-the-art methods. Existing transformer architectures for cross-modal learning can be roughly categorized as Two-stream, Single-stream and Dual Streams. Two-stream architectures, showing in Figure 4-(a), utilize the visual encoder and the text encoder to learn visual and textual representations independently, and then jointly elaborate information exchange transformers (e.g., Co-attention Transformer in [33], Tangled Transformer in [69] and Cross-Modality Encoder in [49]) to achieve the fusion of cross-modal representations and information exchange. Singe-stream architectures [30, 29, 46, 20], as shown in Figure 4-(b), fuse visual and textual representations at the initial stage of the model. We argue that these methods are not suitable for cross-modal retrieval tasks, due to the requirement of pairwise input and intra-model information exchange with $\mathcal{O}(MN)$ time complexity. Approaches with Dual Streams [14, 40, 13, 62] and HiT, as shown in Figure 4-(c) have become a recent trend for cross-modal retrieval with excellent efficiency. Dual Streams require a time complexity of $\mathcal{O}(M+N)$.

8. Conclusion

This paper proposes Hierarchical Transformer (HiT) for video-text retrieval. In HiT, Hierarchical Cross-modal Contrastive Matching is proposed that uses feature-level and semantic-level features to perform contrastive matching. Moreover, to achieve large-scale negative sample interactions in video-text representation learning, we propose Momentum Cross-modal Contrast. Sufficient experiments demonstrate the advantages of our methods.

References

[1] Elad Amrani, Rami Ben-Ari, Daniel Rotman, and Alex Bronstein. Noise estimation using density estimation for self-supervised multimodal learning. *CoRR*, abs/2003.03186, 2020. 6

[2] Jingjing Chen, Chong-Wah Ngo, Fuli Feng, and Tat-Seng Chua. Deep understanding of cooking procedure for cross-modal recipe retrieval. In Susanne Boll, Kyoung Mu Lee, Jiebo Luo, Wenwu Zhu, Hyeran Byun, Chang Wen Chen, Rainer Lienhart, and Tao Mei, editors, *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, pages 1020–1028. ACM, 2018. 6

[3] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10638–10647, 2020. 1, 2

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 2, 3, 4

[5] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2, 3, 4

[6] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *CoRR*, abs/2011.10566, 2020. 2, 3

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4, 5

[8] 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

[9] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. 2018. 1

[10] 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, 2014. 1, 2

[11] Zerun Feng, Zhimin Zeng, Caili Guo, and Zheng Li. Exploiting visual semantic reasoning for video-text retrieval. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 1005–1011. ijcai.org, 2020. 1, 2

[12] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013. 1, 2

[13] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 5, 6, 8

[14] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. COOT: cooperative hierarchical transformer for video-text representation learning. In *Advances in neural information processing systems*, 2020. 1, 2, 8

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2, 3

[16] Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Visualizing and understanding the effectiveness of bert. *arXiv preprint arXiv:1908.05620*, 2019. 1

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 2, 3, 4

[18] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin W. Wilson. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 131–135. IEEE, 2017. 5

[19] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(8):2011–2023, 2020. 5

[20] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 2, 8

[21] Yannis Kalantidis, Mert Bülent Sariyildiz, Noé Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural*

Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020. 3

[22] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *CoRR*, abs/2004.11362, 2020. 2

[23] Yoon Kim. Convolutional neural networks for sentence classification. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751. ACL, 2014. 8

[24] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 1

[25] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4437–4446. IEEE Computer Society, 2015. 6

[26] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 706–715. IEEE Computer Society, 2017. 5

[27] 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 (ECCV)*, pages 201–216, 2018. 1

[28] Jie Lei, Linjie Li, Luowei Zhou, Mohit Bansal Zhe Gan, Tamara L. Berg, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *2021 IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[29] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, pages 11336–11344, 2020. 2, 8

[30] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 1, 2, 8

[31] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 279. BMVA Press, 2019. 5, 6

[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 6

[33] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019. 2, 8

[34] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Xilin Chen, and Ming Zhou. Univilm: A unified video and language pre-training model for multimodal understanding and generation. *CoRR*, abs/2002.06353, 2020. 6

[35] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *CoRR*, abs/1804.02516, 2018. 6

[36] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE international conference on computer vision*, pages 2630–2640, 2019. 1, 2, 5, 6

[37] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 6706–6716. IEEE, 2020. 2

[38] 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 Kiyoharu Aizawa, Michael S. Lew, and Shin'ichi Satoh, editors, *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, ICMR 2018, Yokohama, Japan, June 11-14, 2018*, pages 19–27. ACM, 2018. 6

[39] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[40] Mandela Patrick, Po-Yao Huang, Yuki Markus Asano, Florian Metze, Alexander G. Hauptmann, João F. Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *ICLR*, 2021. 1, 2, 6, 7, 8

[41] Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. Dissecting contextual word embeddings: Architecture and representation. In *EMNLP*, 2018. 4

[42] Yifan Qiao, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. Understanding the behaviors of bert in ranking. *arXiv preprint arXiv:1904.07531*, 2019. 1

[43] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212, 2015. 5

[44] Andrew Rouditchenko, Angie W. Boggust, David Harwath, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Rogério Feris, Brian Kingsbury, Michael Picheny, Antonio Torralba, and James R. Glass. Avlnet: Learning audio-visual language representations from instructional videos. *CoRR*, abs/2006.09199, 2020. 6

[45] 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

[46] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 2, 8

[47] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019. 5

[48] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7464–7473, 2019. 1, 2

[49] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2, 8

[50] Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*, 2019. 1

[51] Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *ACL*, 2019. 4

[52] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *CoRR*, abs/1906.05849, 2019. 2

[53] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. *CoRR*, abs/2005.10243, 2020. 2

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2

[55] Jesse Vig. A multiscale visualization of attention in the transformer model. In *ACL*, 2019. 4

[56] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 154–162, 2017. 1

[57] Wei Wang, Junyu Gao, Xiaoshan Yang, and Changsheng Xu. Learning coarse-to-fine graph neural networks for video-text retrieval. *IEEE Transactions on Multimedia*, PP:1–1, 07 2020. 1, 2

[58] Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. Position focused attention network for image-text matching. *arXiv preprint arXiv:1907.09748*, 2019. 1

[59] Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang. Learning fragment self-attention embeddings for image-text matching. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2088–2096, 2019. 1

[60] Zhirong Wu, Yuanjun Xiong, X Yu Stella, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[61] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 2, 3

[62] Yi Yang Xiaohan Wang, Linchao Zhu. T2vlad:global-local sequence alignment for text-video retrieval. In *2021 IEEE Conference on Computer Vision*, 2021. 2, 8

[63] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV*, volume 11219 of *Lecture Notes in Computer Science*, pages 318–335. Springer, 2018. 5

[64] Yu Xiong, Qingqiu Huang, Lingfeng Guo, Hang Zhou, Bolei Zhou, and Dahua Lin. A graph-based framework to bridge movies and synopses. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4592–4601, 2019. 2

[65] 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

[66] 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 (ECCV)*, pages 471–487, 2018. 1, 2, 5, 6

[67] Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. End-to-end concept word detection for video captioning, retrieval, and question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3261–3269. IEEE Computer Society, 2017. 6

[68] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII*, volume 11217 of *Lecture Notes in Computer Science*, pages 385–401. Springer, 2018. 6

[69] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8746–8755, 2020. 1, 2, 6, 8