Improving Transformer with Dynamic Convolution and Shortcut for Video-Text Retrieval

Zhi Liu[1*], Jincen Cai[1], and Mengmeng Zhang[1,2*]
[1] North China University of Technology, Beijing, 100144, China
[2] Beijing Polytechnic College, Beijing, 10853, China
[e-mail: lzliu@ncut.edu.cn, caijincen712@163.com, muchmeng@126.com]
[*]Corresponding author: Zhi Liu, Mengmeng Zhang

Received January 27, 2022; revised April 14, 2022; accepted June 25, 2022; published July 31, 2022

Abstract

Recently, Transformer has made great progress in video retrieval tasks due to its high representation capability. For the structure of a Transformer, the cascaded self-attention modules are capable of capturing long-distance feature dependencies. However, the local feature details are likely to have deteriorated. In addition, increasing the depth of the structure is likely to produce learning bias in the learned features. In this paper, an improved Transformer structure named TransDCS (Transformer with Dynamic Convolution and Shortcut) is proposed. A Multi-head Conv-Self-Attention module is introduced to model the local dependencies and improve the efficiency of local features extraction. Meanwhile, the augmented shortcuts module based on a dual identity matrix is applied to enhance the conduction of input features, and mitigate the learning bias. The proposed model is tested on MSRVTT, LSMDC and Activity-Net benchmarks, and it surpasses all previous solutions for the video-text retrieval task. For example, on the LSMDC benchmark, a gain of about 2.3% MdR and 6.1% MnR is obtained over recently proposed multimodal-based methods.

Keywords: Video representation, cross-modal retrieval, Multi-modal, Local Descriptors, Transformer.

This work is supported by MOE Planned Project of Humanities and Social Sciences (No.20YJA870014).

1. Introduction

Video is one of the most popular forms of media. There are billions of video clips on the internet and it is still keeping on growing rapidly every day. However, these massive amounts of videos will not be available to end-user if there is no effective way to access them. Therefore, efficient video retrieval is an important technology. Traditional video retrieval methods are mainly based on human-annotated keywords, which is a textual database retrieval system substantially and is hardly to utilize the semantic information of the videos. In the past years, deep learning has been applied to video retrieval tasks to improve retrieval efficiency.

The basic framework of the video retrieval model mainly consists of three parts, including text encoder, video encoder and similarity computation. The text encoder is to extract query information from text descriptions, and the video encoder is to learn video representation from video clips. The problem of learning text representation has been studied to a certain extent, and various approaches [1-3] have been proposed for encoding queries. In contrast, research on representation learning for video is still unsatisfactory. Since video is formed with multiple modalities, including appearance, motion, OCR, audio, scene, etc., how to build an effective video representation based on multi-modal information becomes a critical issue. Traditional studies [4-7] have separated video into several modal streams, and cannot fully utilize the information contained in videos. Multimodal representation methods have been studied to address this problem. Feichtenhofer C et al. [8] used multilayer convolution to process the input data and aggregate spatial as well as temporal features. Liu, Y et al. [9] used the Collaborative Gating mechanism to aggregate different modalities to form a video representation. Andres Mafla et al. [10] took visual regions and scene text instances as input to a Graph Convolutional Network (GCN) and used dot product to aggregate features across modalities.

Transformer [1] is a newly proposed architecture and has made great progress in natural language processing. It has been adopted in the field of video retrieval [11-14]. Based on the Transformer architecture, Valentin Gabeur et al. [14] proposed MMT, which uses a self-attention mechanism to collect cross-modal and temporal cues about the events that occur in the video and aggregate them in a compact representation. Linchao Zhu et al. [12] proposed a Tangled Transformer to further express and aggregate three inputs: action features, regional object features, and language features, and to enhance the interaction between language and visual features by adding an advanced interaction design (Multi-head Attention). As is shown in **Table 1**, the Multi-modal Transformer (MMT) [14] structure is considered one of the most promising methods for multimodal video representation, since it adopts the most number of modalities and uses an advanced Transformer model to efficiently process information, which in turn promises a higher level of semantic information.

At present, there still exists large room for improvement on Transformer. On one hand, the Transformer relies heavily on global self-attention blocks due to the lack of inductive bias, which leads to inefficiency in learning local dependencies; on the other hand, with the increasing of the depth, the Transformer is likely to produce learning bias in the learned features, which is mainly due to the fact that the deeper the layers, the more likely the gradient diverges and the larger the error.

To address the mentioned two problems, in this paper, by employing a span-based dynamic convolution and an augmented shortcuts module, an enhanced Transformer structure named Transformer with Dynamic Convolution and Shortcut (TransDCS) is proposed to improve the performance of multimodal video retrieval. The main contributions of this paper are as follows:

(1) By introducing span-based dynamic convolution to the original Transformer, an improved Transformer structure that integrates global, local, and contextual visual/textual information is proposed, which can produce features with enhanced discriminative power.

(2) A shortcut module named augmented shortcuts is introduced to Transformer either to effectively stabilize the capability of the model and suppress the learning bias that may arise from the deep network.

(3) The proposed Transformer is tested on the mainstream datasets for multimodal video retrieval, and the experiment results show that the proposed model outperforms the recent multimodal video representation models.

Table 1 shows the comparison of the proposed model with the existing works (including the most promising model MMT [14]).

Table 1. Comparison of the advantages/disadvantages between the existing works and the proposed one

Model	Number of Modes	Processing & Aggregation method	Advantages	Disadvantages
Liu, Y et al. [9]	7	— & projection and Hadamard product	More modalities, more types of information	Simple method to process and aggregate
Andres Mafla et al. [10]	2	GCN & dot product and add	Utilizes the graph technology	Fewer modalities & simple aggregation methods
Yale Song et al. [11]	3	Local Feature Transformer & residual learning	Using Transformer technology	Transformer has limited effect on local features
Linchao Zhu et al. [12]	3	Tangled Transformer & concatenate	Multi-head attention mechanism is used for information interaction	Fewer modalities & simple aggregation methods
Valentin Gabeur et al. [14]	7	Transformer & weighted sum	More modalities & using Transformer	Original Transformer lacks inductive bias
Ours	7	Transformer with dynamic convolution and shortcut & weighted sum	More modalities & added inductive bias structure to Transformer	—

The remainder of this paper is organized as follows. In section 2, some related work is addressed. In section 3, the adopted base model and its problems are explained. In section 4, the proposed TransDCS model is presented. In section 5, experimental results are illustrated and section 6 concludes this work.

2. Related Work

For video retrieval applications, the key problem is to find an efficient method to describe the rich information contained in multi-modals of videos. This is a task called representation learning. There are many approaches on how to conduct representation learning on video.

Some traditional deep learning methods have been applied in video retrieval. Yu et al. [15] employed CNNs to encode both visual and audial information in videos. Song et al. [11] used Polysemous Instance Embedding Network (PIE-Net) to learn video representations, which takes as input a global context vector and multiple local feature vectors. Dong et al. [16] adopted three branches, i.e., mean pool, biGRU, and CNN to encode sequential videos. Mithun et al. [17] used multimodal cues with CNN (1D), CNN (2D), and CNN (3D) to process audio, image and action information for video representation. Shizhe Chen et al. [18] decomposed

the video into a hierarchical semantic graph including events, actions and entities.

By designing an attention mechanism and stacking Encoder-Decoder structures, Google proposed Transformer [1] in 2017, which achieves new high BLEU values in machine translation tasks. Transformer is widely adopted in several fields [19-22], and the related researches [23-27] on Transformer show that there is still exists much room for improvement. In video representation learning, the proposed algorithms based on Transformer can be classified into unimodal-based and multimodal-based methods.

The unimodal-based method considered the video as one modality. Han Fang et al. [28] proposed a Transformer-based image encoder named Spatial Transformer (ViT) for video representation. Luo et al. [29] took a Linear Projection named Flattened Patches module to video and used the ViT to get Frame representation. The multimodal-based method divides the video into several modalities, such as appearance, OCR, audio, events, etc. Max Bain et al. [30] proposed a model that manipulates temporal and spatial information separately at the self-attention level. Xing Cheng et al. [31] used Transformer Encode and Mean to aggregate video frames with three modal features, including entity expert, fusion expert and action expert, expecting to achieve an integrated video representation with multiple views. Valentin Gabeur et al. [14] collected seven modal features as input and used a Multi-modal Transformer to achieve information interaction among different modalities, and can make full use of the cross-modal cues present in the video. In this paper, by addressing the disadvantages of the Transformer, a network named TransDCS is proposed for multimodal-based video retrieval.

3. Preliminaries and Motivation

Multi-modal Transformer (MMT) [14] is a Transformer-based architecture that uses a self-attention mechanism to collect valuable cues across modalities and aggregate them in a compact representation. The video is divided into seven modalities: motion, RGB, scene, face, OCR, speech, and audio, and each modality is pre-processed by a pre-trained network in its domain. For example, motion features are pre-trained on the Kinetics action recognition dataset using the S3D network[32]; scene features are pre-trained on the Places365 dataset by the DenseNet-161 network[33]; face features are pre-trained on the VGGFace2 dataset using SSD face detection and ResNet50 network, and audio features are pre-trained on the YT8M dataset with the VGGish network[34], etc. Once the embedding features have been extracted from the input data, they are augmented by adding a position encoding (indicating time) and a label encoding (labeling which modality is used). The enhanced embeddings are then passed through the MMT backbone network, which is a standard Transformer encoder architecture. Each modality is pooled equally as an input, and each input modality produces one embedding so that there are a total of seven output embeddings from the MMT. The MMT schematic is shown in **Fig. 1**.

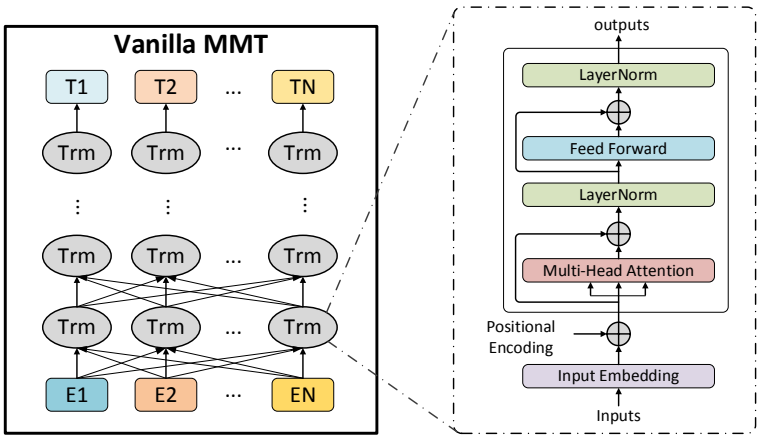


Fig. 1. Vanilla MMT structure.

Problem statement. The vanilla MMT structure consists of multiple Transformer (Trm) encoders. The optimization of a Transformer has three main considerations, including model efficiency, model generalization and model adaptation. Solutions to three considerations are performed mostly at the model and architecture level. In this paper, we focus on how to improve the multimodal Transformer at the model level for video representation.

Self-attention is the basic building block of Transformer to efficiently model the global dependencies in token input sequences. The self-attention module of Vaswani et al. [1] applies three projections to the input $X \in R^{n \times d}$ to obtain key (K), query (Q), and value (V) representations, where n is the number of time steps, d the input/output dimension, as shown in **Fig. 2** (a). The output of the self-attention module is as follows:

$$SA(Q,K,V) = softmax\left(\frac{Q^T K}{\sqrt{d_k}}\right)V. \tag{1}$$

Multi-modal Transformer (MMT) [14] successfully employs self-attention to achieve high performance. Thanks to the Self-Attention mechanism (SA) and the Feed Forward Network (FFN), MMT is able to reflect long-range feature dependencies and thus obtain a global feature representation. Due to the inherent nature of video clips, besides the global features, plenty of local dependencies contained in video clips are also important for video retrieval. However, since the self-attention module used in MMT lacks inductive bias, the local feature details have been ignored. In contrast, the success of Convolutional Neural Networks (CNN) relies on its two inherent inductive biases, i.e., translational invariance and local relevance, which are good at capturing local dependencies. Therefore, CNN-based local features can enhance representation learning when combined with Transformer-based global representation.

In addition, the Convolution structure always requires much deeper neural networks to increase the capacity and performance of the model. However, as the depth of the Transformer structure increases, it is prone to learning bias, which is a degradation problem in deep models: the deeper the layers, the more likely the gradient will diverge and the larger the error will be, making it difficult to train. Theoretically, the deeper the model layer, the smaller the error should be. However, in some cases, the error of the deep model may be larger than that of the shallow one. The reason is that it is difficult to achieve a constant transformation (x=y) in the deep network. As the depth of the network increases, more and more activation functions are introduced and the data are mapped to a more discrete space, which makes it difficult to return

the data to the origin (constant transformation). Introducing linear transformation branches in the Transformer structure, i.e., augmented shortcuts module, can prevent the network parameters from falling into a pseudo-local optimum and make the network parameters converge to the global optimum.

4. Proposed TransDCS Model

Based on the analysis presented in the previous section, in this section, a model named TransDCS is proposed.

For the problem of lost local dependency in Transformer, a convolution-based model is introduced in this work. Span-based dynamic convolution [25] not only has the flexibility of dynamic convolution but also allows kernels to be generated in the local range of current tokens. It can efficiently exploit local dependencies and are capable of distinguishing different meanings of the same token. Span-based dynamic convolution first uses deeply separable convolution to collect span-based token information, as shown in **Fig. 2** (b), and then generates convolutional kernels dynamically. By generating the local relation of the input token conditioned on its local context instead of a single token, the kernel can capture local dependency effectively. Specifically, with query and key pair *Q*, *Ks* as input, the kernel is generated by (2):

$$f(Q, K_s) = softmax\big(W_f(Q \odot K_s)\big). \tag{2}$$

where $\odot$ denotes point-wise multiplication. As illustrated in **Fig. 2** (b), this operator is called span-based dynamic convolution. The output can be written as:

$$SDC\big(Q, K_s, V; W_f, i\big) = LConv\left(V, softmax\left(W_f(Q \odot K_s)\right), i\right). \tag{3}$$

A linear layer is applied afterward for further process. If not specifically stated, always keep the same kernel size for depth-separable convolution and span-based dynamic convolution.

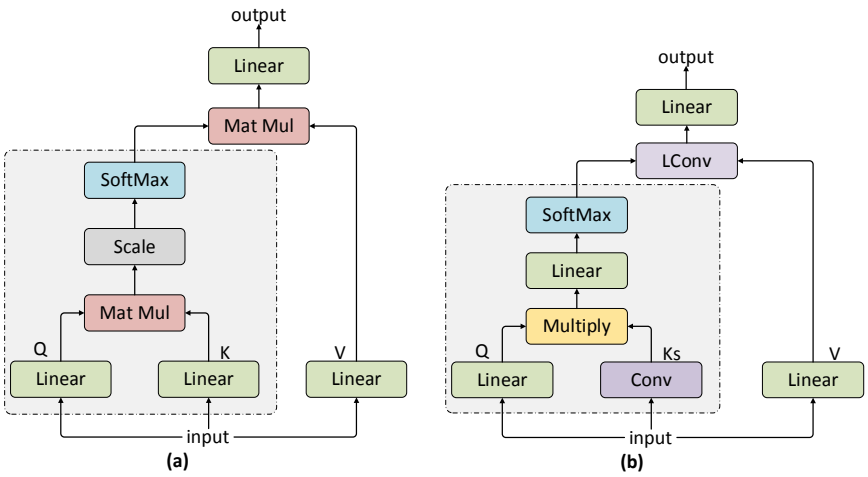


Fig. 2. (a) Self-Attention (SA) (b) Span-based Dynamic Convolution (SDC).

Multi-head Conv-Self-Attention (MCSA). The MCSA attention block integrates the span-based dynamic convolution and self-attention to model both global and local dependencies, as shown in **Fig. 3**. The self-attention and span-based dynamic convolution share the same query but use different keys as a reference to generate attention maps and convolution kernels. In addition, the essence of the multi-head mechanism is the computation of multiple solitary attentions that act as an integration. In this paper, the embedding dimension is partitioned into *h* copies (*h* stands for head, *h*=8), and the output of the Conv-Self-Attention module is 8 subspaces, which means that there may be 8 meanings of information, and the concatenation operation makes the input containing multiple semantic information.

In the MMT model, *h*=4 is used in self-attention (SA), and self-attention has 4 subspaces. In this paper, since the span-based dynamic convolution (SDC) is introduced in the attention module, it consists of two parts, SA and SDC. If the heads are set to 4 (*h*=4), either, to make the overall dimensionality remains unchanged, only two subspaces are generated for SA and SDC respectively, which limits the amount of semantics. Therefore, the model proposed in this paper adopts an 8-head structure, assigning 4 subspaces to SA and SDC respectively, to ensure that there are enough subspaces to generate multiple semantic information.

Denote *Cat*( ) as the concatenate operation, and MCSA attention is expressed as:

$$MCSA(K, Q, K_s, V; W_f) = Cat\left(SA(Q, K, V), SDC(Q, K_s, V; W_f)\right). \tag{4}$$

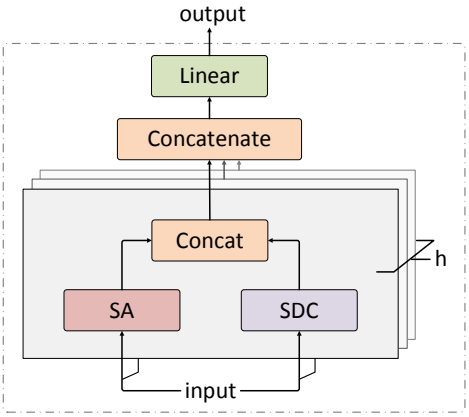


Fig. 3. Multi-head Conv-Self-Attention (MCSA).

The introduction of the span-based dynamic convolution module solves the problem of the lack of inductive bias in self-attention, and it can help to collect contextual information efficiently. The final outputs of this module are fed to the feed-forward layer for further process.

Meanwhile, the inclusion of a parallel shortcut connection in the Multi-head Self-Attention (MSA) module empirically guarantees the transferability of features, especially in deep networks. The MSA module with shortcuts can be formulated as follow:

$$ShortcuMSA(Z_l) = MSA(Z_l) + Z_l\,. \tag{5}$$

where the identity projection (Z_l) is parallel to the MSA module. Intuitively, the shortcut connection bypasses the MSA module and provides another alternative path, where features can be directly delivered to the next layer without the interference of other tokens. The success

of shortcuts shows that bypassing the attention layers with extra paths is an effective way to enhance feature representation and improve the performance of Transformer-like structures.

In the following paragraphs, we aim to refine the existing shortcut connections in the Multi-modal Transformer and explore effective augmented shortcuts to reduce the learning bias introduced by the deep network.

Augmented shortcuts. We propose augmented shortcuts to mitigate learning bias by parallelizing the Multi-head Conv-Self-Attention (MCSA) module with more identity projection. The MCSA module equipped with *N* augmented shortcuts can be formulated as:

$$AugMCSA(Z_l) = MCSA(Z_l) + Z_l + \sum_{i=1}^{N} F_{l_i}(Z_l; W_{l_i}), l \in [1,2,\dots,L]. \quad (6)$$

where $F_{l_i}(\cdot)$ is the *i-th* augmented shortcuts connection of layer *l* and W_{l_i} denotes its parameter. In addition to the original shortcut, the augmented shortcuts provide more alternative paths to bypass the attention mechanism. Unlike the identity projection that directly copies the input tokens to the corresponding output, the parameterized projection $F_{l_i}(\cdot)$ can transform the input features in to another feature space.

A simple formulation of $F_{l_i}(\cdot)$ is a sequence of linear projections and activation functions, i.e.,

$$F_{l_i}(Z_l; W_{l_i}) = \sigma(Z_l; W_{l_i}), l \in [1,2,\dots,L], i \in [1,2,\dots,N]. \quad (7)$$

where$W_{l_i} \in R^{d \times d}$ is the weight matrix and σ is a nonlinear activation function (e.g., GELU) or another module. In Equation (7), $F_{l_i}(\cdot)$ processes each token independently and preserves their specificity, which is complementary to the MCSA module for aggregating different tokens. Note that the identity mapping is a special case of (7), i.e., $\sigma(x) = x$ and W_{l_i} is the identity matrix. The ablation experiments in Section 5.5 show that augmented shortcuts of the form σ(x)=x outperform others.

The augmented shortcuts provide additional alternative paths to bypass the attention module, Meanwhile, the augmented shortcuts with the same parameters can reinforce the original features, and suppress gradient divergence, and prevent the tokens from producing learning bias gradually as the network deepens.

Considering that shortcut connections exist in both MCSA and Feed Forward network (FFN), the proposed augmented shortcuts can also be embedded into FFN similarly, i.e.,

$$AugFFN(Z'_l) = FFN(Z'_l) + Z'_l + \sum_{i=1}^{N} F_{l_i}(Z'_l; W'_{l_i}), l \in [1,2,\dots,L]. \quad (8)$$

where Z'_l is the input feature of the FFN module in the *l-th* layer. Paralleling the FFN module with the augmented shortcuts can further alleviate the learning bias.

By stacking the new attention mechanism and the grouped feedforward module together in an iterative manner, and by constructing the augmented shortcuts on the whole Transformer framework in the manner of **Fig. 4**, i.e., $\sigma(x) = x$, we build the Transformer with Dynamic Convolution and Shortcut (TransDCS) model. With the introduction of the span-based dynamic convolution to the attention module, it not only has the flexibility of dynamic convolution but also makes the kernel be generated in the local range of current tokens, which

can exploit local dependencies and be capable distinguish different meanings of the same token. The necessity of the shortcut module will be demonstrated in section 5.5, and dual identical shortcut paths are used in this paper, which allows the fundamental properties of the input features to be maintained to some extent as the deep network is constructed. In conclusion, the proposed TransDCS model is effective in capturing the global and local information as well as in learning representations and is stable and robust in training, which will be demonstrated in Section 5.

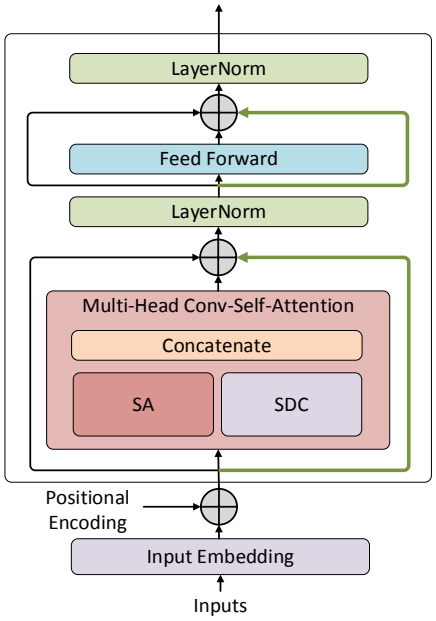


Fig. 4. The diagram of MCSA module is equipped with augmented shortcuts (green lines).

5. Experimental Evaluation

5.1 Architecture

In this paper, the experiments are performed based on the video retrieval framework from [14]. Compared with the basic framework, the improvements of this paper are listed as follows. Firstly, on the video encoding side, the Transformer with Dynamic Convolution and Shortcut (TransDCS) model is designed to improve the video representation efficiency, as described in the previous section. Secondly, we use a more aggressive dropout for the text-based BERT [2] and the video-based Transformer encoder, which corresponds to 0.2, compared to the value of 0.1 in the original framework. Thirdly, since it is observed in this paper that the deeper and wider Transformer encoder in the improved video network will provide improved results, we use 6 layers and 8 heads (compared to 4 layers and 4 heads in the original implementation) in the network. Finally, we use the Mish activation function as a nonlinear function for the output layer of the video representation.

5.2 Datasets

We conduct the experiments on three video benchmarks: LSMDC, Activity-Net and MSR-VTT, covering a challenging set of domains which include videos from YouTube, personal collections and movies.

(1) Large Scale Movie Description (LSMDC) [35]: We employ the LSMDC as the primary dataset, which will be used for all subsequent ablation experiments in the video-text retrieval task. The dataset contains 118081 videos and equivalent captions extracted from 202 movies. They are divided into 109673, 7408 and 1000, as training, validation and test sets. Each video was selected from movies ranging from 2 to 30 seconds.

(2) A Large-Scale Video Benchmark for Human Activity Understanding (Activity-Net) [36]: The dataset consists of 20K YouTube videos annotated with sentence descriptions in time. We concatenated all descriptions of a video to form a paragraph. The training set has 10009 videos. It is used for ab initio training and fine-tuning. We evaluate our video-paragraph retrieval on a "val1" split (4917 videos).

(3) Microsoft Research Video to Text (MSR-VTT) [37]: This dataset consists of 10000 videos, each with 10s to 32s in length and 20 items in cation. To be precise, not all the cations are paired with the whole content of the corresponding video, and some may describe only a short clip, which increases the difficulty of the task. We report the result on different splits of the dataset, where '1k-B' follows the data splits from [38] and '1k-A' follows the data splits from [15].

5.3 Evaluation

The performance of the proposed model is evaluated with standard retrieval metrics: recall at rank N (R@N, higher is better), median rank (MdR, lower is better) and mean rank (MnR, lower is better). In the paper, we report R@1, R@5 and R@10, median and mean ranking. For each metric, the mean of the value is obtained for experiments conducted with 3 random seeds.

5.4 Video-Text Retrieval Results

In this subsection, experimental results comparing the aforementioned methods on three benchmark datasets, LSMDC, Activity-Net and MSR-VTT are shown, and the experimental results of the different datasets are described. The reference methods are briefly listed as follows:

(1) **MMT** MMT is a model proposed by Valentin Gabeur et al. [14] for building information interactions between multimodalities and compressing them into dense representations for video retrieval tasks. The model proposed in this paper mainly optimizes its video encoding part by adding dynamic convolution and enhanced shortcuts to the original one to enhance video representation, thus improving the retrieval performance (more details in Section 5.1).

(2) **CE** The CE proposed by Liu, Y et al [9] is the main data source for multiple modalities of MMT. The CE model uses a simple aggregation mechanism for multiple modes, i.e., a Collaborative gate mechanism consisting of projection and Hadamard product.

(3) **MEE** MEE is a NetVLAD-based video retrieval model proposed by Antoine Miech et al. [38] in 2020. It uses appearance, motion, face and audio as input and applies max-pooling to aggregate the first three modalities and NetVLAD to aggregate the audio modality.

(4) **NoiseE** NoiseE proposed by Elad Amrani et al. [39] is an early model to study multimodality. It reduces the noise problem of multimodal data to a multimodal density estimation task that focuses on learning the connection between two modalities (text and visual).

In addition, there are many other models for video retrieval. AvLnet [40] is a model that uses audio and visual to represent the video, and FSE [41], UniVL [42], HSE [41], and ClipBERT [43] are those models that study the correspondence between video and text. Performance of these models has been considered in the experiments, either. However, since

the datasets and the settings used in some of these models are different from those used in this paper, their results are listed and compared only when these conditions are matched.

Table 2 shows the performance of the proposed algorithm on LSMDC dataset. LSMDC is different from the other datasets used in this paper as it includes the largest number of videos and only one subtitle per video pair. In **Table 2**, TD represents the training dataset, L represents the LSMDC benchmark dataset and H represents the large-scale HowTo100M pre-training dataset. MCSA represents the proposed Multi-head Conv-Self-Attention module and AugS represents the proposed augmented shortcuts module.

As is shown in **Table 2**, by dividing the original attentional mechanism into global and local attention, the proposed MCSA exceeds the reference methods in both metrics. With the addition of augmented shortcuts, the performance has been further improved. In the text-to-video task, compared with the pre-trained MMT model on Howto100M, the proposed TransDCS model improved 2.3% and 6.1% in MdR and MnR respectively and improved 1.8% on average in other metrics. While in the video-to-text task, compared with the un-pretrained MMT, it improved by 2.2%, 2.7%, and 2.7% in R@1, 5, and 10, respectively, and decreased by 4.0% and 8.6% in the median and mean.

Table 2. Experimental results of comparison with previous excellent methods on LSMDC dataset

		Text	→		**Video**		**Video**	→		**Text**	
Model	TD	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
NoiseE [39]	H+L	6.4	19.8	28.4	39.0	—	—	—	—	—	—
MEE [38]	C+L	10.1	25.6	34.6	27.0	—	—	—	—	—	—
CE [9]	L	11.2	26.9	34.8	25.3	96.8	—	—	—	—	—
MMT [14]	L	13.2	29.2	38.8	21.0	76.3	12.1	29.3	37.9	22.5	77.1
MMT-pretrained [14]	H+L	12.9	29.9	40.1	19.3	75.0	—	28.6	—	20.0	76.0
MCSA (ours)	L	13.8	31.1	41.7	**16.5**	69.1	13.9	31.8	40.3	18.5	68.8
AugS(ours)	L	12.8	**32.0**	41.7	17.0	**67.1**	14.1	31.3	40.5	18.0	**67.0**
AugS+MCSA (ours)	L	**14.2**	**32.0**	**42.0**	17.0	68.9	**14.3**	**32.0**	**40.6**	**18.0**	68.5

Table 3. Comparison of paragraph-video retrieval methods trained with video-level information on the Activity-Net dataset (val1 test-split)

	Text	→		**Video**		**Video**	→		**Text**	
Model	R@1	R@5	R@50	MdR	MnR	R@1	R@5	R@50	MdR	MnR
FSE [41]	18.2	44.8	89.1	7.0	—	16.7	43.1	88.4	7.0	—
CE [9]	18.2	47.7	91.4	6.0	23.1	17.7	46.6	90.9	6.0	24.4
HSE [41]	20.5	49.3	—	—	—	18.7	48.1	—	—	—
CLIPBERT [43]	21.3	49.0	—	6.0	—	—	—	—	—	—
MMT [14]	22.7	54.2	93.2	5.0	20.8	22.9	54.8	93.1	4.3	21.2
Ours	**23.7**	**55.7**	**93.8**	**4.0**	**18.5**	**24.3**	**57.2**	**93.8**	**4.0**	**19.0**

As shown in **Table 3**, on the test dataset of Activity-Net, the proposed TransDCS improved the text-to-video R@1 to 23.7, R@5 to 55.7 and reduced the mean rank of prediction by 2.3%. Meanwhile, the video-to-text R@1 improved to 24.3, R@5 improved to 57.2 and reduced the mean rank by 2.2%.

Referring to 1k-A in **Table 4**, the proposed TransDCS model increased by 0.2% and 0.4% on the R@5 and R@10 metrics, respectively, remained unchanged on the median, and decreased by 1.3% on the mean. On 1k-B, the TransDCS model obtained improvements in the metrics R@5, R@10, 2.2% and 2.6%, respectively. There are also some decreases in the mean and median. The advantage of TransDCS is mainly in the mitigation of overfitting. The main reason is that MCSA provides more complex but critical targets, while augmented shortcuts prevent the model from falling into a local optimum by learning futility words.

Table 4. Text-to-video retrieval results on the MSR-VTT test set. Larger R@1, R@5, R@10 and smaller MdR, MnR indicate better retrieval performance

Model	Split	Vis Enc. Init.	Visual-Text PT	R@1	R@5	R@10	MdR	MnR
NoiseE [39]	1k-A	ImageNet, Kinetics	HowTo100M	17.4	41.6	53.6	8.0	—
CE [9]	1k-A	Numerous experts	—	20.9	48.8	62.4	6.0	28.2
UniVL [42]	1k-A	—	HowTo100M	21.2	49.6	63.1	6.0	—
CLIPBERT [43]	1k-A	—	COCO, VisGenome	22.0	46.8	59.9	6.0	—
AVLnet [40]	1k-A	ImageNet, Kinetics	HowTo100M	**27.1**	55.6	66.6	**4.0**	—
MMT [14]	1k-A	Numerous experts	—	24.1	56.4	69.6	**4.0**	25.8
Ours	1k-A	Numerous experts	—	25.2	**56.6**	**70.0**	**4.0**	**24.5**
MEE [38]	1k-B	Numerous experts	COCO	14.2	39.2	53.8	9.0	—
CE [9]	1k-B	Numerous experts	—	18.2	46.0	60.7	7.0	35.3
MMT [14]	1k-B	Numerous experts	—	20.3	49.1	63.9	6.0	29.5
Ours	1k-B	Numerous experts	—	**20.5**	**51.6**	**66.5**	**5.0**	**27.5**

5.5 Ablation Results

The efficiency of the span-based dynamic convolution module, augmented shortcuts module, and the activation function used in the proposed model is validated in this subsection. In addition, we adjust the number of layers and heads of the attention mechanism to show the influence of different parameters.

The detailed configurations and results are shown in **Table 5**. It can be found that the changing of hyper-parameters such as the number of layers and heads on the original model or using the Mish activation function cannot efficiently improve the retrieval performance. Meanwhile, adding only Multi-head Conv-Self-Attention without building a deep network cannot efficiently improve the performance. In contrast, by constructing a deep network with 6 layers and 8 heads, with the same Mish activation function, a 0.7% improvement in R@1 and a 2.5% reduction in MnR are achieved. That is to say, span-based dynamic convolution

must be optimized with properly selected activation functions and hyper-parameters to get a performance improvement. Meanwhile, further addition of the augmented shortcuts module results in an average increment of 0.4% on R@1.

Table 5. Detailed configuration of different sized models. MCSA is the proposed Multi-head Conv-Self-Attention based on span-based dynamic convolution. AugS represents augmented shortcuts. Larger indicates increasing the number of layers and heads of the model

Model	Modification	(L, H, D, E)	Act.	R@1	MnR
Baseline	—	(4,4,0.1,50)	Tanh	13.2	76.3
	—	(4,4,0.1,50)	Mish	13.6	73.9
	—	(6,8,0.2,65)	Tanh	12.4	**65.4**
	—	(6,8,0.2,65)	Mish	12.4	68.3
TransDCS	+MCSA	(4,4,0.1,50)	Tanh	13.6	71.7
	+MCSA	(4,4,0.1,50)	Mish	13.1	71.6
	+MCSA	(6,8,0.2,65)	Tanh	12.8	70.6
	+MCSA	(6,8,0.2,65)	Mish	13.8	69.1
	+AugS	(6,8,0.2,65)	Mish	12.8	67.1
	+MCSA, +AugS	(6,8,0.2,65)	Mish	**14.2**	68.9

(1) Different convolution modules. We tested integrating different convolution modules into the self-attention mechanism. Besides span-based dynamic convolution, Local Token Interaction (LTI) is another efficient convolution module. LTI is derived from the Local Patch Interaction (LPI) block [24], which is designed to make inter-patch communication explicit. Local Token Interaction consists of two one-dimensional convolutional layers of depth 3×3 with batch normalization and GELU nonlinearity. As shown in **Table 6**, adding the local dependency of span-based dynamic convolution to the self-attention module outperforms LTI.

Table 6. Comparison of TCS with different convolutions

Model	Convolution	R@1	R@10	MdR	MnR
Baseline	—	13.2	38.8	21.0	76.3
TransDCS	Span-based Dynamic	**13.8**	**41.7**	16.5	**69.1**
	Local Token Interaction	13.3	**41.7**	**16.0**	72.2

(2) Number of the augmented shortcuts. **Table 7** shows the results of the retrieval performance with the number of augmented shortcuts based on the inclusion of span-based dynamic convolution. When there are no shortcuts, the overall retrieval model works with a mere 0.5% recall in the R@5 metric, which indicates that shortcuts are an indispensable part of the overall model. In addition to the original identity shortcut, doubling the augmented shortcuts improves model performance (e.g., a 0.9% increase in metric R@5 accuracy compared to the baseline). Further increasing the number of augmented shortcuts, the experimental results drop significantly (e.g., the video-to-text metric R@5 drops by 1.1%). Based on the experimental results, it is found that doubling the augmented shortcuts (i.e., one each for MCSA, FFN modules, and two in total) is a good choice.

Table 7. Number of the augmented shortcuts

Num.	Text → Video R@5	MdR	MnR	Video → Text R@5	MdR	MnR
w/o shortcut	0.5	500.2	500.5	0.5	500.5	500.5
0	31.1	**16.5**	69.1	31.8	18.5	68.8
2	**32.0**	17.0	**68.9**	**32.0**	**18.0**	**68.5**
4	31.8	17.0	71.2	30.9	**18.0**	70.0

(3) Formulation of the augmented shortcuts. The augmented shortcuts are implemented sequentially with SDC, activation functions (e.g., GeLU), and identity matrices. **Table 8** shows the impact of each component. Here the span-based dynamic convolution is included in the Baseline. Z_l stands for using the identity matrix directly, without any processing, and it can be found that this shortcut performs the best among the others.

Table 8. Formulation of the augmented shortcuts

Form	R@5	MdR	MnR
Baseline	31.1	**16.5**	69.1
Act($\mathbf{Z_l}$)	30.5	17.0	70.1
SDC($\mathbf{Z_l}$)	30.7	19.0	71.2
$\mathbf{Z_l}$	**32.0**	17.0	**68.9**

(4) Placement of the augmented shortcuts. The location of the augmented shortcuts is chosen to be implemented sequentially in the form of pre, post and realformer. **Table 9** shows the impact of each method based on the addition of span-based dynamic convolution. It can be found that adding the augmented shortcuts along the post form performs the best. The accuracy on metrics R@5 is 0.9% higher than the baseline model, indicating the effectiveness of the post structure. Realformer, which iterates by feeding the output of the previous layer and intermediate variables (attention scores) together to the next layer, results in a decrease in metrics compared to the post structure.

Table 9. Placement of the augmented shortcuts

Place	R@5	R@10	MnR
Baseline	31.1	41.7	69.1
pre	31.2	41.4	69.9
post	**32.0**	**42.0**	**68.9**
realformer	30.2	41.5	70.6

Fig. 5 shows some examples of frames of the top three videos results for each query. In the figure, if the ground truth video is presented, it is indicated by green boxes around the similarity scores. In the first row of **Fig. 5**, the proposed model retrieves reasonable videos even when the ranking is not perfect. In the second row of **Fig. 5**, the proposed model assigns high similarity to the correct video.

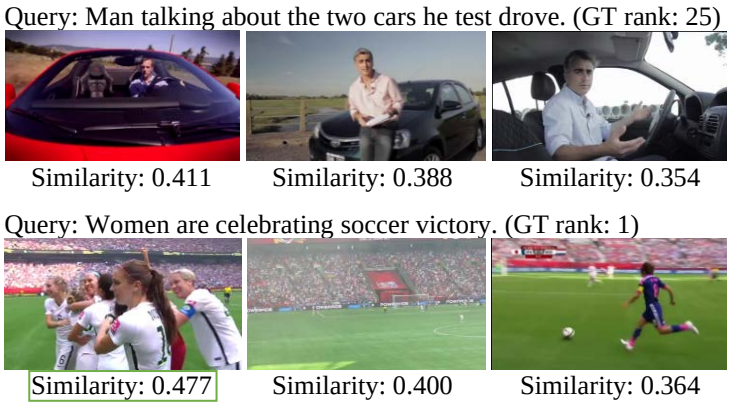


Fig. 5. Qualitative results of MSR-VTT.

6. Conclusion

In this paper, a feature-level multimodal retrieval method based on Transformer is proposed. To effectively utilize the local dependencies, a span-based dynamic convolution is introduced in the attention mechanism, in which the kernel is generated within the local range of the current token, and can efficiently exploit the local dependency. In addition, to overcome the problem of learning bias, the dual identity matrix is used to build augmented shortcuts. In the experiments, the proposed algorithm can obtain about 2.3% MdR and 6.1% MnR gains over the previous multimodal-based method on the LSMDC benchmark, and the R@k metrics are improved by 1.8% on average, either. In future work, we plan to further explore the design of interactions between global features and local dependencies to improve the performance of video retrieval.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," in *Proc. of 31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA., vol.30, pp. 6000-6010, 2017. Article (CrossRef Link)

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv:1810.04805v2 [cs.CL]*, 2019. Article (CrossRef Link)

[3] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," in *Proc. of 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, Canada., 2020. Article (CrossRef Link)

[4] Antoine Miech, Ivan Laptev, and Josef Sivic, "Learnable pooling with Context Gating for video classification," in *Proc. of 2017 IEEE Conference on Computer Vision and Pattern Recognition (Cvpr)*, 2017. Article (CrossRef Link)

[5] Christoph Feichtenhofer, Axel Pinz, and Richard P. Wildes, "Spatiotemporal Multiplier Networks for Video Action Recognition," in *Proc. of 30th IEEE Conference on Computer Vision and Pattern Recognition (Cvpr 2017)*, pp. 4768-4777, 2017. Article (CrossRef Link)

[6] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool, "Temporal Segment Networks: Towards Good Practices for Deep Action Recognition," *Computer Vision - Eccv 2016, Pt Viii*, vol.9912, pp. 20-36, 2016. Article (CrossRef Link)

[7] Karen Simonyan, and Andrew Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," *Advances in Neural Information Processing Systems 27 (Nips 2014)*, vol.1, pp. 568-576, 2014. Article (CrossRef Link)
[8] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition," in *Proc. of Cvpr Ieee*, pp. 1933-1941, 2016. Article (CrossRef Link)
[9] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman, "Use What You Have: Video retrieval using representations from collaborative experts," in *Proc. of BMVC2019*, 2019. Article (CrossRef Link)
[10] Andr´es Mafla, Rafael S. Rezende, Llu´ıs G´omez, Diane Larlus, and Dimosthenis Karatzas, "StacMR: Scene-Text Aware Cross-Modal Retrieval," in *Proc. of Ieee Wint Conf Appl*, pp. 2220-2230, 2021. Article (CrossRef Link)
[11] Yale Song, and Mohammad Soleymani, "Polysemous Visual-Semantic Embedding for Cross-Modal Retrieval," in *Proc. of 2019 IEEE/Cvf Conference on Computer Vision and Pattern Recognition (Cvpr 2019)*, pp. 1979-1988, 2019. Article (CrossRef Link)
[12] Linchao Zhu, and Yi Yang, "ActBERT: Learning Global-Local Video-Text Representations," in *Proc. of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8743-8752, 2020. Article (CrossRef Link)
[13] Song Liu, Haoqi Fan, Shengsheng Qian, Yiru Chen, Wenkui Ding, and Zhongyuan Wang, "HiT: Hierarchical Transformer with Momentum Contrast for Video-Text Retrieval," in *Proc. of 2021 IEEE/Cvf International Conference on Computer Vision (Iccv 2021)*, pp. 11915-11925, 2021. Article (CrossRef Link)
[14] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid, "Multi-modal Transformer for Video Retrieval," in *Proc. of ECCV 2020: Computer Vision – ECCV 2020*, vol.12349, pp. 214-229, 2020. Article (CrossRef Link)
[15] Youngjae Yu, Jongseok Kim, and Gunhee Kim, "A Joint Sequence Fusion Model for Video Question Answering and Retrieval," in *Proc. of Computer Vision – ECCV 2018*, vol.11211, pp. 487-503, 2018. Article (CrossRef Link)
[16] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang, "Dual Encoding for Zero-Example Video Retrieval," in *Proc. of 2019 IEEE/Cvf Conference on Computer Vision and Pattern Recognition (Cvpr 2019)*, pp. 9338-9347, 2019. Article (CrossRef Link)
[17] 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 *Proc. of Icmr '18: Proceedings of the 2018 Acm International Conference on Multimedia Retrieval*, pp. 19-27, 2018. Article (CrossRef Link)
[18] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu, "Fine-grained Video-Text Retrieval with Hierarchical Graph Reasoning," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10638-10647, 2020. Article (CrossRef Link)
[19] Qipeng Guo, Xipeng Qiu, Pengfei Liu, Xiangyang Xue, and Zheng Zhang, "Multi-Scale Self-Attention for Text Classification," in *Proc. of AAAI2020*, 2020. Article (CrossRef Link)
[20] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. of Interspeech*, pp. 5036–5040, 2020. Article (CrossRef Link)
[21] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran, "Image Transformer," in *Proc. of ICML*, pp. 4052–4061, 2018. Article (CrossRef Link)
[22] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A. Hunter, Costas Bekas, and Alpha A. Lee, "Molecular Transformer: A Model for Uncertainty Calibrated Chemical Reaction Prediction," *ACS Central Science*, vol. 5(9), pp. 1572–1583, 2019. Article (CrossRef Link)

[23] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller, "Rethinking Attention with Performers," in *Proc. of ICLR 2021*, 2021. Article (CrossRef Link)

[24] Alaaeldin El-Nouby, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Gabriel Synnaeve, Mathilde Caron, Ivan Laptev, Natalia Neverova, Jakob Verbeek, and Hervé Jégou, "XCiT: Cross-Covariance Image Transformers," *arXiv:2106.09681v2 [cs.CV]*, 2021. Article (CrossRef Link)

[25] Zihang Jiang, Weihao Yu, Daquan Zhou, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan, "ConvBERT: Improving BERT with Span-based Dynamic Convolution," in *Proc. of 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada, 2020. Article (CrossRef Link)

[26] Nikita Kitaev, Łukasz Kaise, and Anselm Levskaya, "Reformer : the efficient transformer," in *Proc. of ICLR2020*, 2020. Article (CrossRef Link)

[27] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid, "VideoBERT: A Joint Model for Video and Language Representation Learning," in *Proc. of Ieee I Conf Comp Vis*, pp. 7463-7472, 2019. Article (CrossRef Link)

[28] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen, "CLIP2Video: Mastering Video-Text Retrieval via Image CLIP," *arXiv:2106.11097v1 [cs.CV]*, 2021. Article (CrossRef Link)

[29] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li, "CLIP4Clip: An Empirical Study of CLIP for End to End Video Clip Retrieval," *arXiv:2104.08860v2 [cs.CV]*, 2021. Article (CrossRef Link)

[30] Max Bain, Arsha Nagrani, G¨ul Varol, and Andrew Zisserman, "Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval," in *Proc. of 2021 IEEE/Cvf International Conference on Computer Vision (Iccv 2021)*, pp. 1728-1738, 2021. Article (CrossRef Link)

[31] Xing Cheng, HeZheng Lin, XiangYu Wu, Fan Yang, and Dong Shen, "Improving Video-Text Retrieval by Multi-Stream Corpus Alignment and Dual Softmax Loss," *arXiv:2109.04290v2 [cs.CV]*, 2021. Article (CrossRef Link)

[32] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy, "Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification," in *Proc. of Computer Vision - Eccv 2018*, vol.11219, pp. 318-335, 2018. Article (CrossRef Link)

[33] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger, "Densely Connected Convolutional Networks," in *Proc. of 30th IEEE Conference on Computer Vision and Pattern Recognition (Cvpr 2017)*, pp. 2261-2269, 2017. Article (CrossRef Link)

[34] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, RifA. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson, "Cnn Architectures for Large-Scale Audio Classification," in *Proc. of Int Conf Acoust Spee and Signal Processing*, pp. 131-135, 2017. Article (CrossRef Link)

[35] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele, "A Dataset for Movie Description," in *Proc. of 2015 IEEE Conference on Computer Vision and Pattern Recognition (Cvpr)*, pp. 3202-3212, 2015. Article (CrossRef Link)

[36] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles, "Dense-Captioning Events in Videos," in *Proc. of 2017 IEEE International Conference on Computer Vision (Iccv)*, pp. 706-715, 2017. Article (CrossRef Link)

[37] Jun Xu, Tao Mei, Ting Yao, and Yong Rui, "MSR-VTT: A Large Video Description Dataset for Bridging Video and Language," in *Proc. of Cvpr Ieee*, pp. 5288-5296, 2016. Article (CrossRef Link)

[38] Antoine Miech, Ivan Laptev, and Inria Josef Sivic, "Learning a Text-Video Embedding from Incomplete and Heterogeneous Data," *arXiv:1804.02516v2 [cs.CV]*, 2020. Article (CrossRef Link)

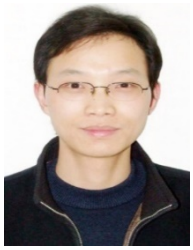
[39] Elad Amrani, Rami Ben-Ari, Daniel Rotman, and Alex Bronstein, "Noise Estimation Using Density Estimation for Self-Supervised Multimodal Learning," in *Proc. of Thirty-Fifth Aaai Conference on Artificial Intelligence, Thirty-Third Conference on Innovative Applications of Artificial Intelligence and the Eleventh Symposium on Educational Advances in Artificial Intelligence*, vol.35, pp. 6644-6652, 2021. Article (CrossRef Link)

[40] Andrew Rouditchenko, Angie Boggust, David Harwath, Brian Chen, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Hilde Kuehne, Rameswar Panda, Rogerio Feris, Brian Kingsbury, Michael Picheny, Antonio Torralba, and James Glass, "AVLnet: Learning Audio-Visual Language Representations from Instructional Videos," in *Proc. of Interspeech 2021*, pp. 1584-1588, 2021. Article (CrossRef Link)

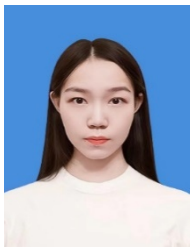
[41] Bowen Zhang, Hexiang Hu, and Fei Sha, "Cross-Modal and Hierarchical Modeling of Video and Text," in *Proc. of Computer Vision - Eccv 2018, Pt Xiii*, vol.11217, pp. 385-401, 2018. Article (CrossRef Link)

[42] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou, "UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation," *arXiv:2002.06353v3 [cs.CV]*, 2020. Article (CrossRef Link)

[43] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu, "Less is More: CLIPBERT for Video-and-Language Learning via Sparse Sampling," in *Proc. of 2021 IEEE Conference on Computer Vision and Pattern Recognition (Cvpr)*, pp. 7331-7341, 2021. Article (CrossRef Link)



Zhi Liu received the B.S. degree in electronic information technology and the Ph.D. degree in signal and information processing from Beijing Jiaotong University, China in 2001 and 2011 respectively. Currently, he is an associate professor in North China University of Technology.



Jincen Cai received the B.S. degrees in communication engineering from Shijiazhuang University, China in 2016. Now, she is a graduate student at the School of Information, North China University of Technology. Her research interests include computer vision, representation learning, video captioning and video retrieval.

Mengmeng Zhang received the B.S. degree in electronic information technology and M.S. degree in signal and information processing from Beijing Jiaotong University, China in 2000 and 2003 respectively, and the Ph.D. degree in communication and information systems from University of Science & Technology Beijing, China in 2012. Currently, he is a professor in North China University of Technology and Beijing Polytechnic College.