Learning Linguistic Association Towards Efficient Text-Video Retrieval

Sheng Fang[1,2], Shuhui Wang[1,3(✉)], Junbao Zhuo[1], Xinzhe Han[2,1], and Qingming Huang[2,1,3]

[1] Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] Peng Cheng Laboratory, Shenzhen, China
{sheng.fang, junbao.zhuo, xinzhe.han}@vipl.ict.ac.cn, wangshuhui@ict.ac.cn, qmhuang@ucas.ac.cn

Abstract. Text-video retrieval attracts growing attention recently. A dominant approach is to learn a common space for aligning two modalities. However, video deliver richer content than text in general situations and captions usually miss certain events or details in the video. The information imbalance between two modalities makes it difficult to align their representations. In this paper, we propose a general framework, **LIN**guistic **AS**sociation (LINAS), which utilizes the complementarity between captions corresponding to the same video. Concretely, we first train a teacher model taking extra relevant captions as inputs, which can aggregate language semantics for obtaining more comprehensive text representations. Since the additional captions are inaccessible during inference, Knowledge Distillation is employed to train a student model with a single caption as input. We further propose Adaptive Distillation strategy, which allows the student model to adaptively learn the knowledge from the teacher model. This strategy also suppresses the spurious relations introduced during the linguistic association. Extensive experiments demonstrate the effectiveness and efficiency of LINAS with various baseline architectures on benchmark datasets. Our code is available at https://github.com/silenceFS/LINAS.

Keywords: Text-Video Retrieval; Knowledge Distillation

1 Introduction

Due to the popularity of online video sharing websites such as TikTok and YouTube, video has become one of the most informative data sources that contain rich visual content. That's the reason why video tasks attract lots of attention recently [1,3,6,9,33,34,35,42,15]. Especially, the explosive increase of the video data makes effective and efficient video retrieval technologies in urgent need. In this paper, we focus on the text-video retrieval task, which aims to find the video in the candidate pool that best matches the semantics of the given natural language description, and vice versa. A general approach for text-video retrieval is to learn a similarity function on video and text that best describes

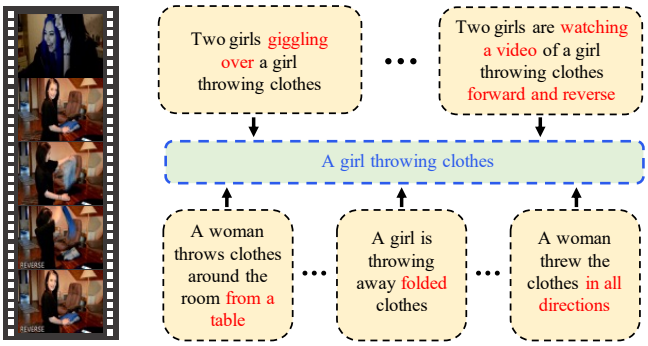

Fig. 1. Illustration of the information imbalance. The sentence in the green box is the query caption and all the other captions are relevant descriptions to the same video. The words marked in red represent the missing information.

their semantic relevance, so that the documents can be ranked according to the similarities. Typically, the videos and captions are encoded separately and then projected into a common embedding space, where the similarities can be calculated by dot product of their corresponding representations. The major challenge of the common latent space approach is how to align the data pairs from the two modalities, and great endeavors have been devoted in this direction [44,38,5,13,10,36].

In general situations, video delivers richer content than text. The former is a consecutive photometric record of events in a physical world, while the latter is the abstract description of the events that a person sees or experiences. So there naturally exists prominent information imbalance between video and text modalities. On the language side, it is natural for a human to describe an event with missing details of actions, attributes and objects. Moreover, different individuals may describe an event with different focuses and language habits. This further leads to the lack of information during the description process.

Figure 1 is an example for demonstrating the imbalance between video and text. In the video, two women laugh together and a video tape about a girl throwing clothes is played forward and reverse. The query caption 'A girl throwing clothes' only briefly describes the action of throwing clothes and ignores other information in the video, like 'laughing together', 'watching a video', 'forward and reverse', 'folded clothes', 'from a table', 'in all directions', *etc.* We further observe the diversity and complementarity among different language descriptions in all the associated captions of the same video in Figure 1. This diversity is of great value for us to enrich the text representations and alleviate the imbalance between two modalities.

In this paper, we propose **LIN**guistic **AS**sociation (LINAS) framework towards efficient text-video retrieval. It includes a teacher model for aggregating the diversified captions for training and a student model for text-video retrieval

without the support of extra captions for inference. First, for learning the teacher model, a support set is constructed for each caption which consists of complementary descriptions from crowdsourcing annotation or the captions of similar videos. With language cross-attention, different captions in the support set are given different weights, according to their degrees of complementarity. In this way, the teacher model learns to combine complementary semantics in different captions. Afterward, the learned enriched text representations will encourage better alignment between two modalities.

However, the teacher model requires additional captions that are inaccessible during inference. To facilitate efficient retrieval in real situations, we introduce Knowledge Distillation to train a student model with a single caption as input. The text and video embeddings of the student model are expected to be as close as possible to those of the teacher model, so the ability of linguistic association can be transfered from the teacher model to the student model without taking extra complementary descriptions at inference stage.

Moreover, the teacher model introduces some spurious correlation when aggregating additional captions, which will inevitablly confuse the student model without careful treatment. Therefore, we further propose Adaptive Distillation strategy which allows the student model to adaptively learn the relational knowledge from the teacher model. By learning the weight (or mask) on each pairwise similarity, our model gradually pays more attention to the diagonal elements of the similarity matrix, which strengthens the transfer of positive relational knowledge. The Adaptive Distillation strategy not only maintains the richness of text representations, but also suppresses the spurious relations introduced during the linguistic association. In together, LINAS achieves better performance.

Our main contributions can be summarized as follows:

- We propose a general framework LINAS for text-video retrieval that utilizes the complementarity between relevant captions. It encourages the model to learn the ability of linguistic association for better aligning two modalities.
- We introduce Knowledge Distillation to train a student model without extra input. We further propose Adaptive Distillation strategy for suppressing the spurious correlation in the teacher model.
- Consistent improvements brought by LINAS with various baseline architectures on benchmark datasets demonstrate its effectiveness. Moreover, experimental results validate the efficiency and generalization ability of the proposed LINAS.

2 Related Work

Text-video Retrieval. Compared to text-image retrieval, text-video retrieval is more challenging and in line with the current trend of shot video. Text-video retrieval attracts growing attention recently. Early work for text-video retrieval is mostly concept-based [16,12,39,30,25,37]. The recent dominant methods are latent-space-based, which aim to project video and text into a joint embedding

space for measuring similarities [9,10,29,28,45,19,38,5,44,23,43]. Dong *et al.* [9] compose parallel multi-level encodings, *i.e.*, mean pooling, biGRU and biGRU-CNN, for comprehensive representations. Projections into a common space are learned afterward. They further propose a hybrid common space which consists of a latent subspace and a concept subspace [10]. Wray *et al.* [44] decompose text into different parts like nouns and verbs for fined-grained action retrieval. Similarly, Chen *et al.* [5] model the text-video matching at levels of events, actions and entities. Considering the multi-modality, Mithun *et al.* [29] employ image, motion and audio modalities to obtain video representations. Liu *et al.* [23] further exploit more multi-modal cues like speech content, OCR, *etc.* Gabeur *et al.* [13] use transformer [40] to aggregate multi-modal features. Wang *et al.* [43] design a global-local alignment method with VLAD encoding. More recent work utilize BERTs or fransformers as the backbone and finetune large-scale pretrained model for cross-modal retrieval [17,32,22,2,24]. Lei *et al.* [17] propose ClipBERT by employing sparse sampling. Liu *et al.* [22] model the feature-level and semantic-level cross-modal matching through Hierarchical Transformer. Luo *et al.* [24] transfer the knowledge of the CLIP model to video-language retrieval in an end-to-end manner. In this work, we propose a general framework for learning linguistic association, which utilizes the complementarity between relevant captions to the same video. Though pretrained models have the ability of language association to a certain extent which is consistent with our method, they require large-scale data and we can capture the association with limited data.

Knowledge Distillation. Knowledge Distillation refers to the methods that train a smaller student network under the supervision of a larger teacher network. Buciluǎ *et al.* [4] first propose model compression for classification and regression tasks. Hinton *et al.* [14] expand this idea and transfer knowledge from the teacher model to the student model by minimizing the difference between classification logits produced by two models. Afterward, Knowledge Distillation is formalized as a pattern for downsizing a network regardless of the structural differences [41]. Park *et al.* [31] propose distance-wise and angle-wise distillation losses for transferring the mutual relations of samples from teacher model to student model instead of simply closing the outputs of them. Knowledge Distillation is widely used in various computer vision tasks such as image classification [20], object detection [18], cross-modal retrieval [26], *etc.* TeachText [7] is the most similar work to ours, which employs Knowledge Distillation for leveraging complementary cues from multiple text encoders. In our work, Knowledge Distillation is utilized to teach the student model the ability of language association. The proposed Adaptive Distillation strategy allows the student model to adaptively learn the relational knowledge from the teacher model.

3 Method

3.1 Problem Description

Due to the richness of video content, there are usually multiple captions corresponding to the same video in the crowdsourcing annotation process. Let

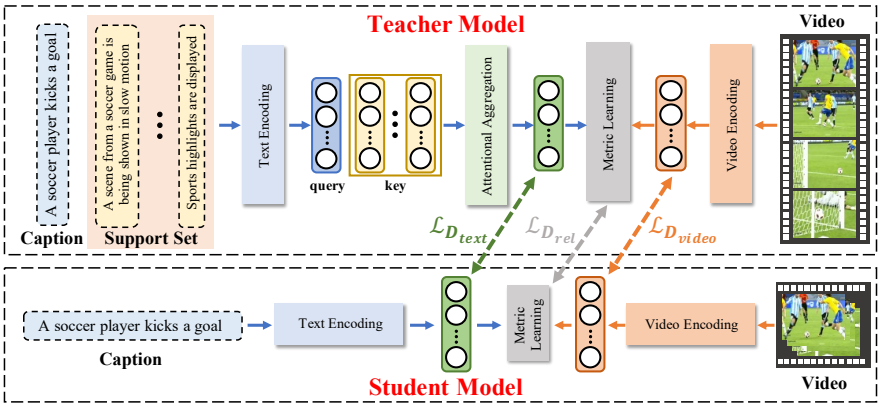

Fig. 2. Overview of LINAS.

$\mathcal{D} = \{(c_i, v_i)\}$ be a dataset where (c_i, v_i) represents a postive caption-video pair. For a video v_i, $\mathcal{C}_i$ is the set consists of all captions corresponding to v_i.

A general approach for text-video retrieval is to encode videos and captions separately. Then the similarities of their representations are measured for ranking. We summarize the framework of mainstream methods into three modules—text encoding, video encoding, and metric learning, noted as T_E, V_E, and M respectively. Metric learning here represents the module for learning similarities between two modalities. Any model that has above three modules can be used as our baseline in LINAS.

Figure 2 is an overview of our proposed LINAS framework. We first train a teacher model which takes query and support set captions as inputs. It achieves more comprehensive text embeddings through attentional aggregation. Afterward, Knowledge Distillation is used to obtain a more efficient student model whose input is a single query caption. In this approach, the student model can learn the ability of linguistic association.

We will introduce the teacher model for aggregating textual semantics in Section 3.2. Then the student model follows in Section 3.3. Section 3.4 is about the distillation process for learning linguistic association.

3.2 Teacher Model

Support set. Before training the teacher model, we first construct a support set for each caption which can provide complementary semantics. The support set consists of descriptions belonging to the same video with the query caption. For caption c_i, N captions are selected from $\mathcal{C}_i \setminus c_i$ to compose the support set, noted as $\{s_i^n\}_{n=1}^N$. Figure 2 provides an instance of support set.

The above scheme has a premise that the videos in the training set must have multiple captions. To make our framework adaptive to datasets where each video

has a unique caption, we propose an alternative approach to construct support sets. The more generalized version of LINAS will be introduced in Section 4.4.

Training. The three modules in our teacher model are denoted as T_E^t, V_E^t, and M^t respectively. Embeddings of query caption and support set captions are obtained by $q_i = T_E^t(c_i)$, $k_i^n = T_E^t(s_i^n)$. Since it is expected that the teacher model can aggregate the complementary semantics to the query caption, we design an attentional aggregation module for combining the representations of query caption and support set captions.

$$x_i^t = q_i + \sum_{n=1}^{N} \frac{\exp(Q(q_i)^T K(k_i^n))}{\sum_{l=1}^{N} \exp(Q(q_i)^T K(k_i^l))} k_i^n, \tag{1}$$

where Q and K are learnable linear projections. We treat the original caption as query and the support set captions as keys for cross-attention learning. Afterward, video embeddings are obtained through $y_i^t = V_E^t(v_i)$. Then the similarity matrix S^t can be achieved by $S^t(i,j) = M^t(x_i^t, y_j^t)$. The objective function for training the teacher model is the same with the original baseline, noted as $\mathcal{L}_O$.

3.3 Student Model

Training. The student model also has text encoding, video encoding, and metric learning three modules noted as T_E^s, V_E^s, and M^s, which need to train from scratch. The text embeddings, video embeddings and similarity matrix of student model can be obtained by $x_i^s = T_E^s(c_i)$, $y_i^s = V_E^s(v_i)$, $S^s(i,j) = M^s(x_i^s, y_j^s)$. Different from training the teacher model, the objective function here is

$$\mathcal{L}_S = \mathcal{L}_O + \mathcal{L}_D \tag{2}$$

where $\mathcal{L}_D$ represents the distillation loss which will be introduced in Section 3.4.

Inference. Only the student model is employed for inference because the support set is inaccessible while testing. Since the student model has the same structure as the chosen baseline, LINAS brings no extra computation cost but significant performance enhancement.

Efficiency. Actually, the student model does not have to be consistent with the teacher model. If we utilize a stronger teacher model for distillation and a lightweight student model for inference, LINAS will achieve efficient retrieval with the performance approaching more complex models.

3.4 Learning to Associate

Start from our motivation of learning linguistic association, we hope that the student model can imitate the teacher model. In order to make the student model to mimic the enriched text representations of the teacher model, we first propose a text distillation loss, denoted as $\mathcal{L}_{D_{text}}$. Since the main goal of text-video retrieval is to better align the two modalities, we raise a similar loss $\mathcal{L}_{D_{video}}$

for supervising the video embeddings of the student model.

$$\mathcal{L}_{D_{text}} = \sum_{i=1}^{B}(||x_i^t - x_i^s||_2^2), \quad \mathcal{L}_{D_{video}} = \sum_{i=1}^{B}(||y_i^t - y_i^s||_2^2), \tag{3}$$

where B represents the batch size.

On the other hand, retrieval is a bidirectional task which is different from classic Knowledge Distillation applications [14]. The correlation between samples is particularly important, because the essence of retrieval is ranking. Motivated by the Relational Knowledge Distillation proposed by Park *et al.* [31], we carry out a relational distillation loss $\mathcal{L'}_{D_{rel}}$. It aims to minimize the distance between the similarity matrixs of the student model and the teacher model, which is calculated by

$$\mathcal{L'}_{D_{rel}} = \sum_{i=1}^{B}\sum_{j=1}^{B} L_\delta(S^t(i,j), S^s(i,j)), \tag{4}$$

where L_δ represents the Huber loss, defined as

$$L_\delta(a,b) = \begin{cases} \frac{1}{2}(a-b)^2, & for\ |a-b| \le \delta \\ \delta|a-b| - \frac{1}{2}\delta^2, & otherwise \end{cases} \tag{5}$$

However, there are some spurious relations in our teacher model introduced during the process of enriching textual representations. Besides, the teacher model contains extra information that the student model cannot perceive. Therefore, using all the correlation from the teacher model for supervision will confuse the student model. The experimental results are the evidence of this problem. When we use the whole similarity matrix to supervise the training of the student model with $\mathcal{L'}_{D_{rel}}$, the performance will decline (as shown in Section 4.3). We further propose Adaptive Distillation strategy which allows the student model to selectively learn the relational knowledge from the teacher model. In this way, the transfer of positive knowledge can be strengthened and the spurious correlation can be suppressed in the distillation process.

Adaptive Distillation. In order to enable the model to adaptively learn the required relational knowledge, we assign a weight to each element in the similarity matrix, that is

$$\mathcal{L}_{D_{rel}} = \sum_{i=1}^{B}\sum_{j=1}^{B} m(i,j) L_\delta(S^t(i,j), S^s(i,j)), \tag{6}$$

where m is a mask matrix whose elements are between 0 and 1.

Inspired by Neural Architecture Search [21], the mask m can be treated as architecture parameters for optimization. θ represents the student model parameters. Since the search space in our case is continuous, we adopt EM (Expectation-Maximization) Algorithm to iteratively optimize m and θ.

Algorithm 1 shows the overall procedure of our Adaptive Distillation strategy. For training the model parameters θ, $\mathcal{L}_{train}(\theta, m)$ is exactly the same as $\mathcal{L}_{D_{rel}}$.

For optimizing the mask m, we found that the diagonal elements in S^t are relatively large. In order to reduce the numerical influence in the learning process, we reweight m in $\mathcal{L}_{val}$, that is

$$\mathcal{L}_{val}(\theta, m) = \sum_{i=1}^{B}\sum_{j=1}^{B} \frac{1}{S^t(i,j)} m(i,j) L_\delta(S^t(i,j), S^s(i,j)). \tag{7}$$

Considering that the samples of each training batch are randomly selected, we can use two values to represent the diagonal and non-diagonal elements of m respectively due to the exchange symmetry. Figure 3 is the visualization of the values in the mask during training. Note that m is normalized and uniformly initialized. When the model converges, an adaptively learned mask tends to transfer the relational knowledge from the diagonal elements from the teacher model. Finally, based on learned mask m, we retrain the student model from scratch. The whole distillation loss for learning linguistic association in our LINAS is

$$\mathcal{L}_D = \alpha * (\mathcal{L}_{D_{text}} + \mathcal{L}_{D_{video}}) + \beta * \mathcal{L}_{D_{rel}}, \tag{8}$$

where α and β are hyperparameters for balancing losses. Detailed analysis about the Adaptive Distillation strategy is available in *Supplementary Material*.

Algorithm 1: Adaptive Distillation

Create a mask m which is uniformly initialized. θ represents the model parameters.
while *not converged* **do**
  $\theta \leftarrow \theta - \eta_\theta \nabla_\theta \mathcal{L}_{train}(\theta, m)$;
  $m \leftarrow m - \eta_m \nabla_m \mathcal{L}_{val}(\theta, m)$;
end
Based on the learned mask m, retrain the model parameters θ from scratch.

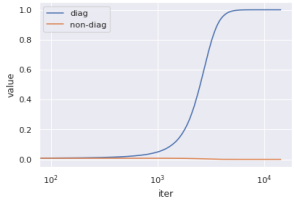

Fig. 3. Visualization of the mask learning Process.

4 Experiments

4.1 Experimental Settings

We conduct experiments on MSR-VTT, MSVD and VATEX datasets in this work. To measure the performance of retrieval models, we employ commonly used metrics including Recall at K (R@K), Median Rank (MedR) and mean Average Precision (mAP). For fair comparison, we adopt the same video features and training strategy with the chosen baseline. The hyperparameters for balancing distillation losses α and β are set to be 0.2 and 1. The δ in Huber loss is set to be 1. Each caption has a support set consists of 8 corresponding descriptions in our experiments. More details are available in *Supplementary Material*.

Table 1. Comparison on MSR-VTT.

Method	Text2Video					Video2Text					SumR
	R@1	R@5	R@10	MedR	mAP	R@1	R@5	R@10	MedR	mAP	
VSE++[11]	5.7	17.1	24.8	65	-	10.2	25.4	35.1	25	-	118.3
Mithun *et al.*[29]	7.0	20.9	29.7	38	-	12.5	32.1	42.4	16	-	144.6
W2VV[8]	6.1	18.7	27.5	45	-	11.8	28.9	39.1	21	-	132.1
CE[23]	10.0	29.0	41.2	16	-	15.6	40.9	55.2	8.3	-	191.9
HGR[5]	9.2	26.2	36.5	24	-	15.0	36.7	48.8	11	-	172.4
Dual Encoding[9]	11.0	29.3	39.9	19	20.3	19.7	43.6	55.6	8	9.3	199.0
LINAS - Dual Encoding	11.9	31.0	42.1	17	21.6	22.0	46.9	59.2	6	10.4	213.1
Hybrid Space[10]	11.6	30.3	41.3	17	21.2	22.5	47.1	58.9	7	10.5	211.7
LINAS - Hybrid Space	12.3	31.6	42.8	16	22.1	22.3	47.8	60.4	6	10.6	217.2
CE+[7]	14.4	37.4	50.2	**10**	-	22.7	52.6	66.3	5	-	243.6
TeachText - CE+[7]	14.9	38.3	51.5	**10**	-	**24.9**	54.1	67.6	5	-	251.3
LINAS - CE+	**15.2**	**38.9**	**52.0**	**10**	-	24.7	**55.2**	**68.0**	**4**	-	**254.0**

Table 2. Comparison on MSVD.

Method	Text2Video		Video2Text	
	R@5	MedR	R@5	MedR
VSE++[11]	39.6	9	-	-
M-Cues[29]	47.8	6	-	-
MoEE[27]	52.0	5	-	-
CE[23]	52.3	5	-	-
Support Set[32]	52.8	5	50.7	5
Dual Encoding[9]	32.6	13	29.4	21
LINAS - Dual Encoding	33.9	12	33.7	17
CE+[7]	56.5	**4**	54.3	5
TeachText - CE+[7]	56.9	**4**	55.0	**4**
LINAS - CE+	**57.6**	**4**	**55.7**	**4**

Table 3. Comparison on VATEX.

Method	Text2Video		Video2Text	
	R@5	MedR	R@5	MedR
CE[23]	68.7	-	71.0	-
W2VV++[19]	68.2	-	75.1	-
HGR[5]	73.5	2	-	-
TeachText - CE+[7]	**87.4**	**1**	-	-
VSE++[11]	65.8	3	75.1	**2**
LINAS - VSE++	70.1	3	78.9	**2**
Dual Encoding[9]	71.3	2	75.7	**2**
LINAS - Dual Encoding	72.7	2	79.1	**2**
Hybrid Space[10]	73.6	-	75.7	-
LINAS - Hybrid Space	74.4	2	**79.4**	**2**

4.2 Comparison with Existing Methods

In this section, we compare the results of our methods applied to various backbones with existing text-video retrieval methods on different datasets. We first make comparisons with methods that do not utilize pretrained large-scale model.

Results on MSR-VTT can be seen in Table 1. Note that CE+ is an improved version of CE proposed by Croitoru *et al.*, which utilizes more high-quality multimodal features and more powerful text embedding. All the methods in Table 1 are trained using only the samples from the target datasets for fair comparison. We can see that LINAS is model-agnostic and the application of LINAS on Dual Encoding, Hybrid Space and CE+ three different baseline models has significantly improved the performance. Moreover, the student model in our method

Table 4. Comparison with pretraining methods.

Method	Dataset	Text2Video				Video2Text				SumR
		R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR	
ClipBERT[17]	MSR-VTT 1k-A	22.0	46.8	59.9	6	-	-	-	-	-
MMT[13]		25.8	57.3	69.3	4	26.1	57.8	68.5	4	304.8
LINAS - MMT		27.1	59.8	71.7	4	28.3	60.3	72.0	3	319.2
Frozen in time[2]	MSVD	33.7	64.7	76.3	3	-	-	-	-	-
CLIP4Clip[24]		46.2	76.1	84.6	2	-	-	-	-	-
LINAS - CLIP4Clip		46.7	76.8	85.6	2	47.3	75.0	83.2	2	414.6

is exactly the same as the baseline, which means LINAS can improve the performance without bringing additional cost at inference. TeachText is a similar work to ours which utilizes Knolwedge Distillation for levaraging complementary cues from various text encoders, while LINAS attempts to learn the language association through relevant captions. We can draw a conclusion that, they are both effective and our LINAS can boost the performance more comprehensively. When applied on CE+, our method outperforms all the competitors in Table 1 which do not use cross-modal pretraining models. The results show the effctiveness of proposed LINAS.

The results on MSVD and VATEX are shown in Table 2 and Table 3. Note that, only a part of the results are reported for saving space. Since these two datasets are in a smaller scale compared with MSR-VTT, the performances on them are much higher than those on MSR-VTT. Nevertheless, the utilization of LINAS still improves the overall performance significantly, which illustrates the robustness of LINAS. On MSVD dataset, 'LINAS - CE+' outperforms all competitors including Support Set, which utilizes more data for training. On VATEX dataset, we apply LINAS on VSE++, which is a quite basic architecture. The consistent promising improvements show the validity of language association. [4]

Recently, pretraining methods are dominant in performance, *e.g.* MMT [13], ClipBERT [17], Frozen in time [2], *etc.* In order to better prove the generalization ability of our method, we further choose some stronger models as baselines to apply proposed LINAS. MMT utilizes transformer architecture to aggregate features from different modalities, *e.g.* OCR, Face, Speech, *etc.* Moreover, it uses a pretrained model on HowTo100M [28]. As shown in Table 4, on MSR-VTT 1k-A (another split edition of MSR-VTT dataset), LINAS further improves the performance of MMT. CLIP4Clip [24] utilizes large-scale pretrained model CLIP for text-video retrieval. Our LINAS still achieves constant gains when applied to CLIP4Clip on MSVD, which shows the effectiveness of LINAS. Actually, the pretrained model has the ability of linguistic association to some extent after scanning large-scale data, which is consistent to our motivation. However,

[4] 'TeachText-CE+' achieves the best perfomrance on VATEX. However, the authors have not provided corresponding multi-modal features of VATEX dataset.

LINAS can further improve the performance on pretraining methods and it can capture the association with limited data.

4.3 Ablation Study

To be clear in advance, all the experiments in this section are under the same experimental settings. 'Dual Encoding' is chosen as the baseline and the experiments are conducted on MSR-VTT dataset.

Distillation strategy. The distillation loss for training the student model in LINAS is $\mathcal{L}_D = \alpha * (\mathcal{L}_{D_{text}} + \mathcal{L}_{D_{video}}) + \beta * \mathcal{L}_{D_{rel}}$. Ablation studies on the distillation strategy are reported in Table 5. The last row shows the performance of the teacher model. We observe that the teacher model is good at text-to-video retrieval but obtains unfavorable performance in the other direction. The amazing performance at T2V is because the teacher model makes use of ground truth information to construct the support set in both training and testing stages. Meanwhile, the aggregation of captions will reduce the discrimination between texts and introduce spurious relations, which results in the reduction of performance at V2T.

We can see that $\mathcal{L}_{D_{text}}$, $\mathcal{L}_{D_{video}}$, and $\mathcal{L}_{D_{rel}}$ are all beneficial to the retrieval performance of the student model from Table 5. It shows that our poposed distillation losses are effective in the process of learning lingusitic association.

Moreover, the 6th row is trained with $\mathcal{L}'_{D_{rel}}$. The performance drop compared with the 4th row shows that using all the similarities from the teacher model for supervision is harmful. It is caused by introduced spurious correlation in the teacher model. The 7th row replaces $\mathcal{L}'_{D_{rel}}$ with $\mathcal{L}_{D_{rel}}$ which further improve the performance. It proves the validity of our Adaptive Distillation strategy. Through the mask learning procedure, we draw the conclusion that taking diagonal elements for supervision is helpful for strengthening the transfer of positive relational knowledge and suppressing the spurious correlation in the teacher model. On the other hand, the R@1 metrics at both directions of the teacher model is considerable which proves the reliability of the similarities of

Table 5. Ablation studies on disillation loss.

	Distillation Loss				Text2Video					Video2Text					SumR
	$\mathcal{L}_{D_{text}}$	$\mathcal{L}_{D_{video}}$	$\mathcal{L}_{D_{rel}}$	$\mathcal{L}'_{D_{rel}}$	R@1	R@5	R@10	MedR	mAP	R@1	R@5	R@10	MedR	mAP	
1					10.9	29.3	39.8	20	20.2	19.5	42.8	55.8	8	9.3	199.0
2	✓				11.3	30.0	40.8	18	20.8	21.1	44.6	56.7	7	10.1	204.4
3			✓		11.3	30.1	41.1	18	20.9	20.8	44.5	58.2	7	9.8	205.9
4	✓	✓			11.7	30.6	41.6	**17**	21.3	21.9	45.2	58.3	7	10.2	209.3
5	✓		✓		11.5	30.2	41.1	18	21.0	20.4	45.8	57.7	7	10.2	206.7
6	✓	✓		✓	11.5	30.2	41.1	18	21.0	21.8	45.5	58.2	7	10.0	208.3
7	✓	✓	✓		**11.9**	**31.0**	**42.1**	**17**	**21.6**	**22.0**	**46.9**	**59.2**	**6**	**10.4**	**213.1**
Teacher Model					19.0	44.6	57.7	7	31.3	23.7	39.0	46.9	13	18.9	231.0

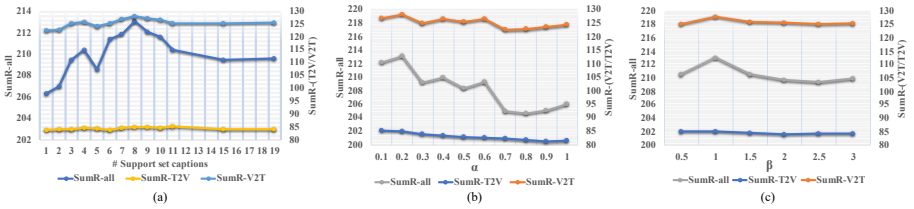

Fig. 4. The influence of support set size and hyperparameters α, β.

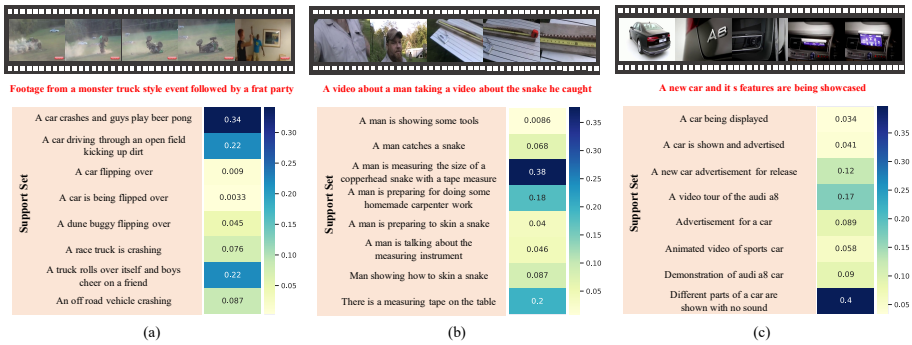

Fig. 5. Weights of support set captions in attentional aggregation.

positive video-caption pairs. It experimentally supports the conclusion of our Adaptive Distillation strategy.

Support set size. Each video in MSR-VTT has 20 relevant captions. Apart from the query caption itself, there are up to 19 captions to compose the support set. Extensive experiments are conducted to explore the impact of support set size by random sampling. The results of different support set sizes are shown in Figure 4 (a). With the increase of the number of support set captions, the performance roughly shows a trend of increasing first and then decreasing. The overall performance reaches the peak when the support set has 8 captions. When there are not enough support set captions, the model can not capture the correlation information from insufficient data. When there are too many captions, too many noise and distractive information is introduced which makes it hard to learn the linguistic association.

Loss weight. Extensive experiments are conducted to evaluate the effects of hyperparameters α and β. The results are shown in Figure 4 (b) and (c). We can draw a conclusion that, the proper values of α and β are 0.2 and 1.

Attentional aggregation. To validate the effectiveness of the attentional aggregation mechanism, a comparison experiment employing mean pooling for aggregation is designed. The results are shown in Table 6, which show the advantage of the attentional aggregation. Moreover, to demonstrate that the teacher

model does concentrate on the complementary information through the aggregation process, we visualize some text-video pairs and the attention weights of support set captions in Figure 5. Taking Figure 5 (b) as an example, the original query text only expresses the message that a man catches a snake. However, the video contains more events. In the video, the man continues to measure the length of the snake. In the figure, the support set captions containing elements like 'tape', 'carpenter', 'table' are given higher attention weights as we expected. Since the aggregation module is designed for concentrating on the complementary information which is relevant to the video but not involved in the caption.

Table 6. Experiments of different aggretation mechanism.

Method	Text2Video					Video2Text					SumR
	R@1	R@5	R@10	MedR	mAP	R@1	R@5	R@10	MedR	mAP	
Mean Pooling	11.5	30.4	41.4	17	21.3	21.1	45.3	57.4	7	10.1	207.1
Attentional	11.9	31.0	42.1	17	21.6	22.0	46.9	59.2	6	10.4	213.1

4.4 General Applicability

Efficiency. We carry out a lightweight model named 'Base' in this experiment, which simply adopts mean pooling for video representations and biGRU for text representations. The 1st row in Table 7 shows the results of the 'Base' model without distillation. Then LINAS is utilized on the 'Base' model and brings significant improvement. Afterward, we utilize a more complex teacher model 'Dual Encoding' for distillation while remain the structure of the student model unchanged. We observe that the model is further improved, whose performance is comparable or even slightly surpasses that of the stronger baseline model 'Dual Encoding'. Note that the number of parameters in 'Base' (14.9M) is less than 20% of that in 'Dual Encoding' (81.4M), which results in 3× speed-up during inference. The experimental results demonstrate that LINAS can achieve efficient text-video retrieval.

Generalization. We propose an alternative approach to construct support sets for the situation where each video only has one corresponding caption. Given a trained text-video retrieval model, top-N relevant videos to the query caption can be obtained. Afterward, the support set of the query caption is composed of the corresponding captions of these N videos. For validating generalized LINAS, we construct a demo dataset by randomly choosing one caption for each video in MSR-VTT. The generalized LINAS is employed on this dataset and N is set to be 8. As shown in Table 8, the promising improvement brought by the generalized LINAS on the chosen baseline model illustrates its effectiveness. In this approach, LINAS can be extended to applications on more datasets. Visualizations of the support sets in our demo dataset can be found in *Supplementary Material*.

Table 7. Experiments of utilizing a lightweight student model.

Teacher Model	Student Model	Text2Video				Video2Text				SumR
		R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR	
-	Base	9.7	26.8	37.0	23	17.7	40.0	51.7	10	182.9
Base	Base	10.3	27.9	38.9	**19**	19.3	42.9	54.1	**8**	193.4
Dual Encoding[9]	Base	10.7	28.9	**39.9**	**19**	**20.0**	**43.9**	**56.4**	**8**	**199.8**
-	Dual Encoding[9]	**10.9**	**29.3**	39.8	20	19.5	42.8	55.8	**8**	199.0

Table 8. Experiments of generalized LINAS.

Method	Text2Video				Video2Text				SumR
	R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR	
Dual Encoding[9]	4.8	17.0	25.0	55	4.7	16.4	25.3	59	91.2
generalized LINAS - Dual Encoding	5.6	17.4	25.6	45	5.3	17.3	26.4	48	97.5

5 Conclusion

In this paper, we propose LINAS towards efficient text-video retrieval. A teacher model which takes extra relevant captions as inputs is trained first. It can aggregate complementary semantics of the diversified captions for text enrichment. Afterward, Knowledge Distillation is introduced to teach the student model the ability of linguistic association, which has only one query caption as input. We further design Adaptive Distillation strategy which allows the student model to adaptively learn the relational knowledge from the teacher model. It aims to strengthen the transfer of positive knowledge and suppress the spurious correlation introduced by linguistic association. LINAS can be applied to most mainstream methods and bring no extra computation cost during inference. Moreover, LINAS can achieve efficient text-video retrieval by adopting a lightweight student model. Additionally, we propose Generalized LINAS for applications on datasets where each video only has one caption. It employs an alternative scheme for constructing support sets. Extensive experimental results demonstrate the effectiveness, efficiency and generalization ability of LINAS.

Acknowledgements. This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0102000, in part by National Natural Science Foundation of China: 62022083, U21B2038 and 61931008, and in part by the Fundamental Research Funds for the Central Universities.

References

1. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lucic, M., Schmid, C.: Vivit: A video vision transformer. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 6816–6826. IEEE (2021). https://doi.org/10.1109/ICCV48922.2021.00676, https://doi.org/10.1109/ICCV48922.2021.00676
2. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1728–1738 (2021)
3. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 813–824. PMLR (2021), http://proceedings.mlr.press/v139/bertasius21a.html
4. Buciluǎ, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 535–541 (2006)
5. Chen, S., Zhao, Y., Jin, Q., Wu, Q.: Fine-grained text-video retrieval with hierarchical graph reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10638–10647 (2020)
6. Chen, W., Li, G., Zhang, X., Yu, H., Wang, S., Huang, Q.: Cascade cross-modal attention network for video actor and action segmentation from a sentence. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 4053–4062 (2021)
7. Croitoru, I., Bogolin, S.V., Leordeanu, M., Jin, H., Zisserman, A., Albanie, S., Liu, Y.: Teachtext: Crossmodal generalized distillation for text-video retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11583–11593 (2021)
8. Dong, J., Li, X., Snoek, C.G.: Predicting visual features from text for image and video caption retrieval. IEEE Transactions on Multimedia **20**(12), 3377–3388 (2018)
9. Dong, J., Li, X., Xu, C., Ji, S., He, Y., Yang, G., Wang, X.: Dual encoding for zero-example video retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9346–9355 (2019)
10. Dong, J., Li, X., Xu, C., Yang, X., Yang, G., Wang, X., Wang, M.: Dual encoding for video retrieval by text. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
11. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: VSE++: improving visual-semantic embeddings with hard negatives. In: British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018. p. 12. BMVA Press (2018), http://bmvc2018.org/contents/papers/0344.pdf
12. Foteini, M., Anastasia, M., Damianos, G., Theodoros, M., Vagia, K., Anastasia, I., Symeonidis, S.: Iti-certh participation in trecvid 2016. In: TRECVID 2016 Workshop (2016)
13. Gabeur, V., Sun, C., Alahari, K., Schmid, C.: Multi-modal transformer for video retrieval. In: European Conference on Computer Vision (ECCV). vol. 12349, pp. 214–229. Springer (2020)
14. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)

15. Junbao, Z., Yan, Z., Shuhao, C., Shuhui, W., Bin, M., Qingming, H., Xiaoming, W., Xiaolin, W.: Zero-shot video classification with appropriate web and task knowledge transfer. In: Proceedings of the 30th ACM International Conference on Multimedia (2022)
16. Le, D.D., Phan, S., Nguyen, V.T., Renoust, B., Nguyen, T.A., Hoang, V.N., Ngo, T.D., Tran, M.T., Watanabe, Y., Klinkigt, M., et al.: Nii-hitachi-uit at trecvid 2016. In: TRECVID. vol. 25 (2016)
17. Lei, J., Li, L., Zhou, L., Gan, Z., Berg, T.L., Bansal, M., Liu, J.: Less is more: Clipbert for video-and-language learning via sparse sampling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7331–7341 (2021)
18. Li, Q., Jin, S., Yan, J.: Mimicking very efficient network for object detection. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 6356–6364 (2017)
19. Li, X., Xu, C., Yang, G., Chen, Z., Dong, J.: W2vv++ fully deep learning for ad-hoc video search. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 1786–1794 (2019)
20. Li, Z., Hoiem, D.: Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence **40**(12), 2935–2947 (2017)
21. Liu, H., Simonyan, K., Yang, Y.: DARTS: differentiable architecture search. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net (2019), https://openreview.net/forum?id=S1eYHoC5FX
22. Liu, S., Fan, H., Qian, S., Chen, Y., Ding, W., Wang, Z.: Hit: Hierarchical transformer with momentum contrast for video-text retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11915–11925 (2021)
23. Liu, Y., Albanie, S., Nagrani, A., Zisserman, A.: 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. p. 279. BMVA Press (2019), https://bmvc2019.org/wp-content/uploads/papers/0363-paper.pdf
24. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of CLIP for end to end video clip retrieval. CoRR **abs/2104.08860** (2021), https://arxiv.org/abs/2104.08860
25. Markatopoulou, F., Galanopoulos, D., Mezaris, V., Patras, I.: Query and keyframe representations for ad-hoc video search. In: Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval. pp. 407–411 (2017)
26. Miech, A., Alayrac, J.B., Laptev, I., Sivic, J., Zisserman, A.: Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9826–9836 (2021)
27. Miech, A., Laptev, I., Sivic, J.: Learning a text-video embedding from incomplete and heterogeneous data. arXiv preprint arXiv:1804.02516 (2018)
28. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2630–2640 (2019)
29. Mithun, N.C., Li, J., Metze, F., Roy-Chowdhury, A.K.: Learning joint embedding with multimodal cues for cross-modal text-video retrieval. In: Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval. pp. 19–27 (2018)

30. Nguyen, P.A., Li, Q., Cheng, Z.Q., Lu, Y.J., Zhang, H., Wu, X., Ngo, C.W.: Vireo@ trecvid 2017: Video-to-text, ad-hoc video search, and video hyperlinking. In: TRECVID (2017)
31. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3967–3976 (2019)
32. Patrick, M., Huang, P., Asano, Y.M., Metze, F., Hauptmann, A.G., Henriques, J.F., Vedaldi, A.: Support-set bottlenecks for video-text representation learning. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), https://openreview.net/forum?id=EqoXe2zmhrh
33. Qi, Z., Wang, S., Su, C., Su, L., Huang, Q., Tian, Q.: Towards more explainability: concept knowledge mining network for event recognition. In: Proceedings of the ACM International Conference on Multimedia (ACM MM). pp. 3857–3865 (2020)
34. Qi, Z., Wang, S., Su, C., Su, L., Huang, Q., Tian, Q.: Self-regulated learning for egocentric video activity anticipation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021). https://doi.org/10.1109/TPAMI.2021.3059923
35. Qi, Z., Wang, S., Su, C., Su, L., Zhang, W., Huang, Q.: Modeling temporal concept receptive field dynamically for untrimmed video analysis. In: Proceedings of the ACM International Conference on Multimedia (ACM MM). pp. 3798–3806 (2020)
36. Sheng, F., Shuhui, W., Junbao, Z., Qingming, H., Bin, M., Xiaoming, W., Xiaolin, W.: Concept propagation via attentional knowledge graph reasoning for video-text retrieval. In: Proceedings of the 30th ACM International Conference on Multimedia (2022)
37. Snoek, C.G., Li, X., Xu, C., Koelma, D.C.: University of amsterdam and renmin university at trecvid 2017: Searching video, detecting events and describing video. In: TRECVID (2017)
38. Song, Y., Soleymani, M.: Polysemous visual-semantic embedding for cross-modal retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1979–1988 (2019)
39. Ueki, K., Hirakawa, K., Kikuchi, K., Ogawa, T., Kobayashi, T.: Waseda_meisei at trecvid 2017: Ad-hoc video search. In: TRECVID (2017)
40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
41. Wang, L., Yoon, K.J.: Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
42. Wang, T., Zhang, R., Lu, Z., Zheng, F., Cheng, R., Luo, P.: End-to-end dense video captioning with parallel decoding. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 6827–6837. IEEE (2021). https://doi.org/10.1109/ICCV48922.2021.00677, https://doi.org/10.1109/ICCV48922.2021.00677
43. Wang, X., Zhu, L., Yang, Y.: T2vlad: global-local sequence alignment for text-video retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5079–5088 (2021)
44. Wray, M., Larlus, D., Csurka, G., Damen, D.: Fine-grained action retrieval through multiple parts-of-speech embeddings. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 450–459 (2019)

45. Yang, X., Dong, J., Cao, Y., Wang, X., Wang, M., Chua, T.: Tree-augmented cross-modal encoding for complex-query video retrieval. In: Huang, J., Chang, Y., Cheng, X., Kamps, J., Murdock, V., Wen, J., Liu, Y. (eds.) Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020. pp. 1339–1348. ACM (2020). https://doi.org/10.1145/3397271.3401151, https://doi.org/10.1145/3397271.3401151