From Within to Between: Knowledge Distillation for Cross Modality Retrieval

Vinh Tran, Niranjan Balasubramanian, and Minh Hoai

Stony Brook University, Stony Brook, NY 11790, USA
{tquangvinh, niranjan, minhhoai}@cs.stonybrook.edu

Abstract. We propose a novel loss function for training text-to-video and video-to-text retrieval networks based on knowledge distillation. This loss function addresses an important drawback of the max-margin loss function often used in existing cross-modality retrieval methods, in which a fixed margin is used in training to separate matching video-and-caption pairs from non-matching pairs, treating all non-matching pairs the same and failing to account for the different degrees of non-matching. We address this drawback by introducing a novel loss for the non-matching pairs; this loss leverages the knowledge within one domain to train a better network for matching between two domains. This proposed loss does not require extra annotation. It is complementary to the existing max-margin loss, and it can be integrated into the training pipeline of any cross-modality retrieval method. Experimental results on four cross-modal retrieval datasets namely MSRVTT, ActivityNet, DiDeMo, and MSVD show the effectiveness of the proposed method.
Code is available at: https://github.com/tqvinhcs/CrossKD

Keywords: Text-video retrieval · Knowledge distillation.

1 Introduction

Given that videos have become a big part of our lives with hundred of thousands of video hours produced, uploaded, and consumed every day, a problem of growing importance in computer vision is to index and search for videos based on their content. In this paper, we tackle two important cross modal retrieval problems: text-to-video and video-to-text. In the former problem, the input is a text query, and the system has to retrieve a list of videos with relevant content [15, 35]. In the latter, the input is a video, and the system has to rank a list of captions based on how likely it is for the captions to be used to describe the content of the video [15, 35]. This video-to-text retrieval task is useful for automatically captioning a video based on its content. Hereafter, for brevity, we will refer to these two problems as Text-and-Video Retrieval, or TVR for short.

TVR can be tackled by projecting text and video into a joint embedding space and learning a similarity scoring function for text and video embedding vectors. At test time, the retrieved candidates, either text or video, are ranked based on their similarity with respect to the input query. Usually, an encoder

such as an LSTM/RNN [30,57,58,60] or a language model such as BERT [11,15] is used for encoding the caption. Whereas, the video representation is often a composition of multiple types of features such as faces, scene, motion, and sound. A feature aggregation method, such as NetVLAD [40] or a Transformer [15], is then used for encoding the video features into a single representation.

The key question in this setting is how the joint embedding space and the similarity scoring function are learned. Usually, one can assume that there is labeled training data containing *matching* video-caption pairs i.e., the video and the captions associated with it. Non-matching video-caption pairs can be constructed by pairing a random video with a random caption of another video. A common approach in cross-modal retrieval is to treat matching and non-matching pairs as positive and negative samples respectively, and a binary classifier is trained to separate the two sets using the well-known max-margin loss. One major problem of this approach is that it treats all non-matching pairs the same, demanding a fixed separating margin for all of them. However, not all non-matching pairs are alike. A caption C_A of video A might be similar to a caption C_B of video B, so the pair (C_A, B) would be a noisy negative sample. This severely affects the performance of the learned embedding space and scoring function.

In this work, we propose a simple but effective technique to address this problem. We use the similarity of the captions (or videos) as a better indicator for the degree of matching and non-matching. In other words, we use the knowledge of one domain (either text or video) to guide the training of neural networks that connect between two domains. This is a type of knowledge distillation, where knowledge *within* one domain is distilled to the *between* of two domains. The main rationale behind this approach is as follows: we can more reliably learn a within domain similarity function than a cross-domain similarity function, since there is often larger amounts of within-domain training data. This suggests that the former can be used as a teacher similarity function to train the latter. Fig. 1 illustrates this idea.

The proposed within-to-between knowledge distillation can be implemented as a loss function and added to the existing training loss of any TVR framework to learn the joint embedding of text and video. Experimental results on four TVR benchmarks, MSRVTT [57], Activity Net [30], DiDeMo [20], and MSVD [5] show that the proposed knowledge distillation loss improves the performance of two cross modal retrieval frameworks.

2 Related Work

Text-Video Retrieval (TVR) is an emerging research area [15,17,30,35,39,41, 58,60]. Early works use language model such as LSTM [23] to represent text for capturing the sequential properties in a sentence [6, 30, 58, 60]. Before feeding to the language model to extract sentence features, Word2Vec [42] is often used for representing word embedding. Recent methods use more advanced architectures such as OpenAI-GPT [35], BERT [15, 61] or GPT2-XL and GPT2-XL-F [9] for text encoding. The video side is more complex where multiple modalities are used

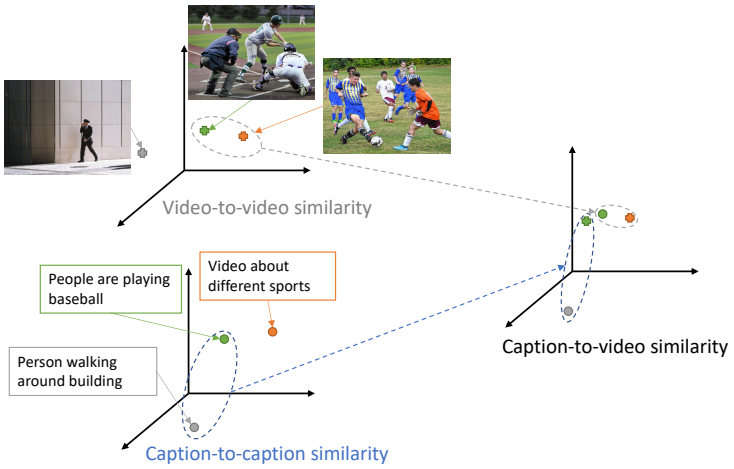

Fig. 1: **Knowledge distillation for learning similarity scoring function between a caption and a video.** The distillation, illustrated as dash lines, can be drawn from either the text domain or video domain. Our idea is based on the observation that the similarity scoring function in the same domain is more reliable than the similarity scoring function between two domains. Hence, the former can be used as the teacher to train the latter.

for extracting video features. Then a multi-expert model such as NetVLAD [40], Collaborative Experts [35], or Transformer [15] is used for representing video features. However, previous works mainly focus on the features aggregation. Instead, we propose a new learning objective for better learning the similarity of text and video across modalities.

Knowledge distillation [22] has widely been used to transfer knowledge learned between different deep learning models. This method was originally introduced for decision tree simplification [2] and model compression [3]. This method has been proven to be useful in many visual recognition problems [7, 27, 29, 31, 32, 38, 44, 48]. Conventionally, the knowledge is transferred from the teacher to the student model by considering each data sample individually, but Park *et al.* [45] propose a method that distills the mutual relation by penalizing structural differences of data samples. To some extent, we also utilize the relational knowledge from captions to captions and from videos to videos, but we perform knowledge distillation across the domains of text and video. In recent works, knowledge distillation has also been employed in the problem of cross modal learning [7]. However, this work focuses on aligning audio, image, and video representations, which is different from the TVR task. Moreover, their proposed method is semi-supervised and relies on class label prediction of human action for distillation. Whereas our method is unsupervised and is trained without using any class label. Recently, Croitoru *et al.* [9] proposed TEACHTEXT, a generalized distillation

method utilizing multiple teachers to improve the retrieval performance. Our knowledge distillation method differs from this by not using external teachers during training. Alternatively, our distillation comes directly from the available structure of the video and caption domains, which is very efficient to compute and does not require an external model for distillation. Along with our work, [8, 55] shows that related captions can also help improve the result embedding.

Contrastive learning for cross modal retrieval. To learn a joint embedding space for cross modal retrieval, a common approach is to use contrastive learning [26, 33, 38–41, 51, 61]. This learning method aims to maximize the similarity of text and video representations extracted from the same instance while maximizing the difference between text and video representations from different instances. To this end, bidirectional max-margin loss is used for training [6, 7, 9, 15, 35, 41]. By pushing dissimilar text-video pair away, the model implicitly learns the ranking of video and text model based on their correspondent similarity. Beside max-margin loss, recent works also use normalized softmax loss [59] for contrastive learning in a similar fashion [1]. Instead of learning ranking function, some previous works also directly learn the similarity scores between the two models using regression loss [54].

3 Knowledge Distillation for Cross Modal Retrieval

3.1 Framework for cross modal retrieval

Our method is applicable to different retrieval frameworks, including Multimodal Transformer (MMT) [15] and TEACHTEXT [9]. For brevity, we will describe our method with the MMT framework in this section, but we will demonstrate its benefits for both MMT, CLIP4Clip and TEACHTEXT in the experiment section.

Video encoder. MMT [15] uses a Transformer architecture that combines multiple video embedding experts, with each expert corresponding to one type of video feature among motion [56], audio [21], scene [25], OCR [10,36], face [19,34], speech [42], and appearance [24]. The output of this video encoder is a representation that consists of N different embeddings, denoted by $\Psi(v) = \{\psi_i\}_{i=1}^{N}$. Please refer to [15] for the implementation details of the multi-modal experts.

Text encoder. MMT uses a pre-trained BERT [11] model for encoding the caption text. BERT is a transformer-based model that has been shown to produce effective text representations for a wide variety of tasks [11]. Each caption is represented by BERT's output vector for the "[CLS]" token. Subsequently, gated embedding modules [40] are used to generate N different embeddings of this caption representation corresponding to N video experts. The caption embedding is denoted by $\Phi(c) = \{\phi_i\}_{i=1}^{N}$, where ϕ_i is the embedding vector of the i^{th} expert.

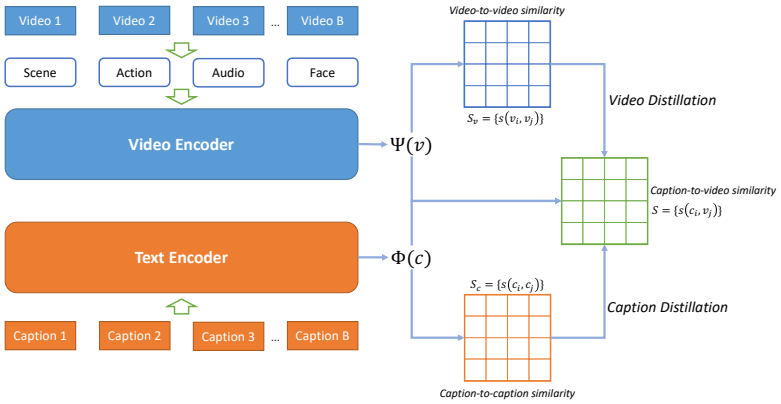

Fig. 2: **The proposed distillation methods for cross modal text-video retrieval**. We use a BERT model for caption representation and Multi-modal transformer for video representation. The video representation is the aggregation of multiple experts. Using these output representations, the caption-to-caption and video-to-video similarity can be computed efficiently within a training batch. The compatibility function $s(c, v)$ between caption and video can be trained with the additional training signal.

The caption-video similarity is taken as the weighted sum of the experts' caption-video similarity values:

$$s(c, v) = \sum_{i=1}^{N} w_i \langle \phi_i, \psi_i \rangle, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product, and w_i is the weight for the i^{th} expert. Given a query item in one domain, these similarity scores are then used to rank items in the other domain. Fig. 2 depicts the processing pipeline of our cross modal retrieval framework.

3.2 Bidirectional max-margin ranking loss

To encode the relationships and to measure similarity between a caption and a video, existing methods use bidirectional max-margin loss to separate matching caption-video pairs (called positive pairs) from the non-matching ones (called negative pairs). Positive and negative pairs of caption-video are created within each input training batch. A positive pair is the one where the caption and the video come from the same training data instance. Whereas, a negative caption-video pair consists of a caption and a video from different training instances. An instance is a pair of a caption and a video in each training batch. For each training instance, the similarity $s_{ij} = s(c_i, v_j)$ between a caption i and a video

j is computed. Then, a bidirectional max-margin loss is used to train the compatibility function as:

$$\mathcal{L}_{\text{margin}} = \frac{1}{B}\sum_{i=1}^{B}\sum_{j\neq i}[\max(0, s_{ij} - s_{ii} + m) + \max(0, s_{ji} - s_{ii} + m)], \quad (2)$$

where B is the number of instances in each training batch and m the margin. The bidirectional max-margin loss function requires that the compatibility of a caption and a video from each positive pair to be higher than those from a negative pair by at least a margin m. In other frameworks [1, 37], normalized softmax loss function is used for contrastive learning instead of bidirectional max-margin ranking loss.

3.3 Knowledge distillation loss from caption and video

Since videos and captions come from two different domains, learning a compatibility function based solely on the max-margin loss may be sub-optimal. We propose a complementary loss function that exploits the relationship between data instances in the same domain and distill this knowledge between domains. This is to use the similarity between captions (or between videos) from different training samples as a guidance for training the compatibility function between captions and videos. To this end, we compute the caption similarity of a given caption to the other captions in an input training batch. For a caption c that is represented by embedding vectors $\{\phi_i\}_{i=1}^{N}$ and a caption c' represented by embedding vectors $\{\phi'_i\}_{i=1}^{N}$, the caption-to-caption similarity is computed as:

$$s(c, c') = \sum_{i=1}^{N} w_i \langle \phi_i, \phi'_i \rangle. \quad (3)$$

Given a training batch B consisting of multiple caption and video pairs $\{(c_j, v_j)\}$, and a particular caption c_i, we can compute the similarity between c_i to other c_j's, which can then be normalized to get a proper probability distribution:

$$\mathcal{P}_{ij}^{cc} = \frac{\exp(s(c_i, c_j)/\tau)}{\sum_{l=1}^{B}\exp(s(c_i, c_l)/\tau)}, \forall j \in B, \quad (4)$$

where $\tau > 0$ is a temperature parameter controlling the smoothness of the distribution. Similarly, we can also compute the cross-domain caption-to-video similarity probability distribution as:

$$\mathcal{Q}_{ij}^{cv} = \frac{\exp(s(c_i, v_j)/\tau)}{\sum_{l=1}^{B}\exp(s(c_i, v_l)/\tau)}, \forall j \in B, \quad (5)$$

Given the above two distributions, we use the knowledge distillation loss to measure the dissimilarity between the caption-to-caption distribution $\mathcal{P}$ and the

caption-to-video distribution $\mathcal{Q}$ using Kullback–Leibler divergence as:

$$\mathcal{L}_{\text{cap\_distill}} = \frac{1}{B}\sum_{i=1}^{B} KL(\mathcal{P}^{cc}_{i,:}||\mathcal{Q}^{cv}_{i,:}). \tag{6}$$

Analogously, the similarity between videos features can be used as another training signal. The video-to-video and video-to-caption similarity can also be written in form of two distributions as:

$$\mathcal{P}^{vv}_{ij} = \frac{\exp(s(v_i, v_j)/\tau)}{\sum_{l=1}^{B}\exp(s(v_i, v_l)/\tau)}, \forall j \in B, \tag{7}$$

and

$$\mathcal{Q}^{vc}_{ij} = \frac{\exp(s(c_j, v_i)/\tau)}{\sum_{l=1}^{B}\exp(s(c_l, v_i)/\tau)}, \forall j \in B. \tag{8}$$

Again, the video distillation loss is computed as the Kullback–Leibler divergence between the two distributions:

$$\mathcal{L}_{\text{vid\_distill}} = \frac{1}{B}\sum_{i=1}^{B} KL(\mathcal{P}^{vv}_{i,:}||\mathcal{Q}^{vc}_{i,:}). \tag{9}$$

3.4 Composition loss for training retrieval model

We add the two new distillation loss functions in Eq. (6) and Eq. (9) to the standard bidirectional max-margin ranking loss for training the compatibility function between a caption and a video. The distillation loss for the captions is given by:

$$\mathcal{L}_{\text{cap\_compose}} = \mathcal{L}_{\text{margin}} + \mathcal{L}_{\text{cap\_distill}}. \tag{10}$$

In the same way, the knowledge distillation for videos is given by:

$$\mathcal{L}_{\text{vid\_compose}} = \mathcal{L}_{\text{margin}} + \mathcal{L}_{\text{vid\_distill}}. \tag{11}$$

In our experiments, we train with both losses and also show an ablation study for each loss function. During development, we explored several ways for combining the two losses and find that treating both of them equally yields the best results.

4 Experimental Results

4.1 Text-video retrieval frameworks

Retrieval frameworks. One advantage of the proposed method is its ability to be used with different TVR frameworks. To demonstrate this, we experiment with three recent TVR frameworks: Multi-modal Transformer (MMT) [15],

Table 1: **Benefits of knowledge distillation for the MMT framework** on the MSRVTT and ActivityNet datasets. Both type of distillations effectively improve the performances of text-to-video and video-to-text retrieval on all datasets. Caption Distillation is slightly better than Video Distillation

Datasets & Methods	*Text* → *Video*			*Video* → *Text*		
MSRVTT 1k-A	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
MMT [15]	26.6	57.1	69.6	**27.0**	57.5	69.7
MMT + Caption Distillation	**27.8**	58.4	70.4	**27.0**	**58.8**	70.2
MMT + Video Distillation	26.7	**59.0**	**71.8**	26.4	58.1	**71.7**
MSRVTT 1k-B	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
MMT [15]	20.3	49.1	63.9	21.1	49.4	63.2
MMT + Caption Distillation	20.8	**52.4**	**66.5**	**22.7**	**52.5**	**67.4**
MMT + Video Distillation	**21.3**	51.4	66.2	21.3	**52.5**	66.3
ActivityNet	R@1↑	R@5↑	R@50↑	R@1↑	R@5↑	R@50↑
MMT [15]	22.7	54.2	93.2	22.9	54.8	93.1
MMT + Caption Distillation	24.3	57.1	**94.1**	**26.5**	**58.7**	**94.1**
MMT + Video Distillation	**24.4**	**58.0**	94.0	25.7	58.0	93.3

CLIP4Clip [37], and TEACHTEXT [9]. Both MMT and TEACHTEXT are based on Collaborative Experts (CE) architecture [35] with multiple experts for video encoding. Meanwhile, CLIP4Clip [37] is the most recent framework for TVR with a single video encoder based on vision transformer.

Video encoders. MMT [15] uses seven pretrained experts for video encoding, namely: **Motion** from S3D [56], **Audio** from VGGish [21], **Scene** from DenseNet [25], **OCR** from the output of text detector embedded with Word2Vec [42], **Face** from ResNet50 [18] trained on VGGFace2, **Speech** using Google Cloud Speech, and **Appearance** from SENet-154 [24].

TEACHTEXT [9] also uses seven experts for video encoding. Most of these video experts are similar to that of MMT, except for the **Motion** expert which comprises of two action experts *Action(KN)* and *Action(IG)*. The first action expert, *Action(KN)*, is an I3D model trained on Kinetics [4], and the second expert, *Action(IG)*, is a 34-layer R(2+1)D model [52] pretrained on IG-65m dataset [16]. TEACHTEXT establishes a stronger baseline than CE (denoted as CE+) by using the more powerful text embeddings from GPT2-XL [50]. The full TEACHTEXT framework (denoted as TT-CE+) is trained with additional knowledge given by multiple text encoders as teachers including Word2Vec [42], GPT2-XL [50], and GPT2-XL-F.

CLIP4Clip [37] uses the pretrained CLIP [49] model as a single video encoder. The encoder is based on Vision Transformer (ViT) [12]. In our experiment, we use the pretrained CLIP (ViT-B/16) [49] as our backbone to encode video. The visual encoder has 12 layers and patch size of 16. Unfortunately, due to our

Table 2: **Text-to-video retrieval performance for CLIP4CLIP** with and without knowledge distillation on the MSRVTT, MSVD and DiDeMo datasets. Both type of distillations effectively improve the performances on all datasets. All the results are obtained using ViT-B/16 as video encoder. Caption Distillation is slightly better than Video Distillation

Datasets and Methods	R@1↑	R@5↑	R@10↑	MnR↓
MSRVTT 1k-A				
CLIP4Clip [37]	43.4	72.3	80.5	15.4
CLIP4Clip + Caption Distillation	**46.2**	**73.1**	**81.2**	**13.2**
CLIP4Clip + Video Distillation	44.7	72.8	81.1	13.8
MSVD				
CLIP4Clip [37]	48.6	78.6	87.2	9.0
CLIP4Clip + Caption Distillation	48.8	**79.2**	**87.5**	**8.7**
CLIP4Clip + Video Distillation	**48.9**	79.1	87.4	9.0
DiDeMo				
CLIP4Clip [37]	42.0	69.1	78.1	18.8
CLIP4Clip + Caption Distillation	**43.2**	**69.7**	79.2	**17.5**
CLIP4Clip + Video Distillation	**43.2**	69.2	**79.3**	17.9

limited computational resources, we have to keep the first 6 layers of the ViT-B/16 frozen during training. Furthermore, we can only train our model with relatively smaller batch sizes in comparison to [37]. The batch sizes are set to be 24 for MSRVTT and MSVD datasets. For DiDeMo, we can only use the batch size of 6, since the videos in this dataset are relatively long and the video encoder requires a long observation window of 64 frames. As a result, our reproduced results for CLIP4CLIP on DiDeMo are not as good as the previously reported results [37].

Training details. Our training procedure is based on the implementation of MMT[1], TEACHTEXT[2], and CLIP4Clip[3]. We train all the models on PyTorch [46] with Adam optimizer [28]. The bidirectional max-margin ranking loss is used for MMT and TEACHTEXT (CE+ and TT-CE+), while the normalized softmax loss is used in CLIP4Clip for contrastive learning. If not otherwise specified, all training parameters are the same as reported in MMT [15], TEACHTEXT [9], and CLIP4Clip [37].

4.2 Datasets

We perform experiments on four challenging TVR benchmarks: MSRVTT [57], ActivityNet [30], DiDeMo [20], and MSVD [5]. We report performances on both

[1] https://github.com/gabeur/mmt
[2] https://github.com/albanie/collaborative-experts
[3] https://github.com/ArrowLuo/CLIP4Clip

text-to-video and video-to-text retrieval tasks. For experiments using TEACHTEXT and CLIP4Clip, we follow [9,37] to report only text-to-video performance.

MSRVTT [57] is a large-scale dataset for video understanding, especially for TVR. The dataset contains 10,000 video clips crawled from web. Each video clip is associated with 20 natural sentences annotated by AMT workers. In the MMT framework, we follow [15, 40, 58] and perform experiments on Split 1k-A and Split 1k-B. Split 1k-A [58] uses 9000 videos for training and 1000 for testing. Meanwhile, Split 1k-B [40] uses 6656 videos for training and 1000 for testing. In addition, for direct comparison with prior work, we also provide the retrieval performance on the full split of this dataset with TEACHTEXT framework (6513 videos for training, 497 for validation, and 2990 for testing).

ActivityNet [30] contains 20,000 Youtube videos amounting to 849 video hours with temporally annotated sentence descriptions. Following [60] and [15], we concatenate all the sentence descriptions for each video to form a paragraph. There are 10009 instances in the training set. Following the same paragraph-video retrieval setup in [15, 35, 60], we perform evaluation on the "val1" split (4917 videos) of this dataset.

DiDeMo [20] stands for Distinct Describable Moments. This dataset contains unedited 10,464 personal videos in multiple content such as sports, concerts, and pets. Each video in the dataset has three to five captions. We follow [9, 35, 60] and use 8392 videos for training, 1065 for validation, and 1004 for testing.

MSVD [5] dataset has 1970 video clips associated with 80K English captions. The setup is similar to [9, 35], where 1200 clips are used for training, 100 for validation, and 670 for testing. Since MSVD videos do not contain sound, audio-based features are not used for this dataset.

Evaluation metrics. The performance of all models are evaluated with recall at rank N (R@N), a standard retrieval metric. A better model should achieve higher recall. We also report median rank (MdR) and mean rank (MnR) of the correct results. For median rank and mean rank, a lower number indicates better performance. Similar to [9], we also report the geometric mean of R@1, R@5, and R@10 for conciseness. The geometric mean summarizes the overall retrieval performance at multiple recall ranking steps. Since each TVR framework is performed on a different subset of datasets, we will perform our experiments following the setups in previous works [9, 15, 37].

4.3 Knowledge Distillation for the MMT retrieval framework

We first evaluate the benefits of knowledge distillation for MMT on the MSRVTT and ActivityNet datasets. As can be seen from Table 1, training retrieval models with knowledge distillation improves the retrieval performance. Our models yield significant improvement over its direct baseline MMT on both text-to-video and video-to-text retrieval tasks. For Split 1k-A, Caption Distillation improves the text-to-video retrieval performance from 26.6% to 27.8% at R@1. For R@5 and R@10, the performance gain brought by Video Distillation are even higher, from

Table 3: **Text-to-video retrieval performance of TEACHTEXT** [9] without and with Caption Distillation on four datasets

Methods	R@1↑	R@5↑	R@10↑	MdR↓	Methods	R@1↑	R@5↑	R@10↑	MdR↓
MSRVTT									
CE+ [9]	13.8	36.5	49.4	11.0	TT-CE+ [9]	14.6	37.9	50.9	**10.0**
+Our Distillation	**14.7**	**37.8**	**50.6**	**10.0**	+Our Distillation	**14.7**	**38.1**	**51.1**	**10.0**
ActivityNet									
CE+ [9]	19.4	49.3	65.4	6.0	TT-CE+ [9]	23.5	57.2	**73.6**	**4.0**
+Our Distillation	**20.6**	**50.6**	**66.9**	**5.0**	+Our Distillation	**23.9**	**57.3**	73.5	**4.0**
MSVD									
CE+ [9]	25.1	56.5	70.9	**4.0**	TT-CE+ [9]	25.1	56.8	71.2	**4.0**
+Our Distillation	**26.0**	**58.3**	**72.9**	**4.0**	+Our Distillation	**25.5**	**57.1**	**71.7**	**4.0**
DiDeMo									
CE+ [9]	18.2	43.9	57.1	7.9	TT-CE+ [9]	21.6	48.6	**62.9**	6.0
+Our Distillation	**20.2**	**45.2**	**58.8**	**7.0**	+Our Distillation	**21.7**	**49.2**	62.4	**5.7**

Table 4: **Text-to-video retrieval performance for TEACHTEXT methods with and without knowledge distillation**. The performance measure is the geometric mean of R@1, R@5 and R@10. The left columns on each dataset are the base models of CE+ and TT-CE+ from TEACHTEXT. The right columns are the results obtained by adding our distillation loss. Our proposed method improves the performance on all base models in all datasets

Method	MSRVTT		ActivityNet		DiDeMo		MSVD	
	Base	Ours	Base	Ours	Base	Ours	Base	Ours
CE+ [9]	29.2	**30.4**	39.7	**41.1**	35.8	**37.7**	46.5	**47.9**
TT-CE+ [9]	30.4	**30.6**	46.3	**46.5**	40.4	**40.5**	46.6	**47.1**

57.1% to 59.0% and 69.6% to 71.8%, respectively. On the ActivityNet dataset, the performance gaps are even wider, clearly demonstrating the benefits of our proposed distillation loss. Our models perform better than the direct baseline MMT at every recall step. Both types of knowledge distillation provide benefits for training the retrieval system. Between the two, Caption Distillation performs slightly better than Video Distillation. There are two possible reasons that can explain this. One reason is that caption similarity starts with a pretrained BERT model, whereas video similarity is trained from scratch. The other is that the text similarity computation is arguably simpler in that it is done over a single representation, while the video similarity is computed over a composition of features from multiple experts.

4.4 Knowledge Distillation for the CLIP4Clip retrieval framework

We also consider a recent TVR framework CLIP4Clip [37]. For this framework, the bidirectional max-margin loss is replaced by the normalized softmax loss for contrastive learning. Hence, the composition losses $\mathcal{L}_{\text{norm\_softmax}} + \mathcal{L}_{\text{cap\_distill}}$

Table 5: **Comparison to other methods on MSRVTT 1k-A dataset.** Results are obtained by applying caption distillation on CLIP4Clip [37] framework.

Method	R@1↑	R@5↑	R@10↑	MnR↓
ActBERT [61]	8.6	23.4	33.1	-
MIL-NCE [39]	9.9	24.0	32.4	-
JSFusion [58]	10.2	31.2	43.2	-
HT [41]	12.1	35.0	48.0	-
HT-pretrained [41]	14.9	40.2	52.8	-
CE [35]	20.9	48.8	62.4	28.2
CLIP [49]	22.5	43.3	53.7	61.7
MMT [15]	24.6	54.0	67.1	-
MMT-pretrained [15]	26.6	57.1	69.6	24.0
TT-CE+ [9]	29.6	61.6	74.2	-
SSB [47]	30.1	58.5	69.3	-
Frozen [1]	31.0	59.5	70.5	-
MDMMT [13]	38.9	69.0	79.7	16.5
CLIP4Clip [37]	44.5	71.4	**81.6**	15.3
Ours	**46.2**	**73.1**	81.2	**13.2**

and $\mathcal{L}_{\text{norm\_softmax}} + \mathcal{L}_{\text{vid\_distill}}$ are used instead. As can be seen from Table 2, both proposed distillation losses, especially the Caption Distillation loss, improve the retrieval performance on the three datasets.

4.5 Knowledge Distillation for the TEACHTEXT Framework

We also evaluate the benefits of knowledge distillation for the TEACHTEXT [9] framework. As before, the distillation losses are added to the training loss of this method. Since caption distillation is slightly better than video distillation, we perform further experiments with caption distillation only. The experiments are conducted for two training settings: *(1) without external teachers (CE+)*, and *(2) with external teachers (TT-CE+)*. Specifically, we use the loss $\mathcal{L}_{\text{margin}} + \mathcal{L}_{\text{cap\_distill}}$ when applying our method on the CE+ setting, and the loss $\mathcal{L}_{\text{margin}} + \mathcal{L}_{\text{d}} + \mathcal{L}_{\text{cap\_distill}}$ when applying our method on the TT-CE+ setting. Here, $\mathcal{L}_{\text{d}}$ is the distillation loss from external teachers [9].

Following [9], we perform experiments on the four TVR benchmarks. For a fair comparison with most previous methods, we do not use the denoising trick [9] when training on the crowd-sourced datasets MSRVTT and MSVD.

Performance without using external teachers (CE+). Tables 3 and 4 show the results of this experiment. On all four datasets, the CE+ method with the proposed caption distillation outperforms the baseline CE+ method without any knowledge distillation. On average, caption distillation brings a 1.4% gain over the direct baseline CE+. We also show some qualitative results of our proposed method in Fig. 3. Notably, the proposed method is a form of

Table 6: **Comparison with the other methods on the MSVD dataset.**

Method	R@1↑	R@5↑	R@10↑	MnR↓
VSE++ [14]	15.4	39.6	53.0	-
M-Cues [43]	20.3	47.8	61.1	-
MEE [40]	21.1	52.0	66.7	-
CE [9]	21.5	52.3	67.5	-
TT-CE [9]	22.1	52.2	67.2	-
CE+ [9]	25.1	56.5	70.9	-
TT-CE+ [9]	25.1	56.8	71.2	-
SSB [47]	28.4	60.0	72.9	-
Frozen [1]	33.7	64.7	76.3	-
CLIP4Clip [37]	46.2	76.1	84.6	10.0
Ours	**48.8**	**79.2**	**87.5**	**8.7**

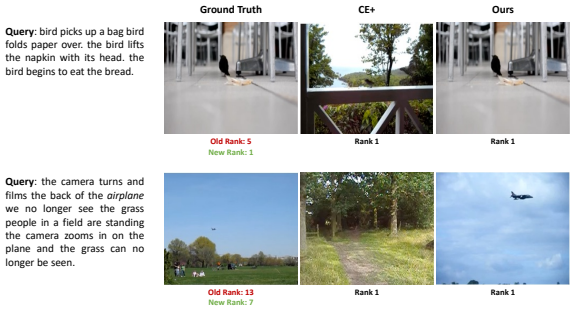

Fig. 3: **Qualitative retrieval performance on DiDeMo dataset with CE+ baseline.** On the left is the query. The first column is the correct video clip. The next two columns show the top 1 retrieved clips by CE+ and our proposed method respectively.

self-distillation and does not use any external information for training, while the full TT-CE+ employs the external teachers for distillation.

Performance using external teachers (TT-CE+). As can be also seen in Tables 3 and 4, caption distillation is still beneficial when combining with the external teachers of TEACHTEXT. However, the additional benefit is not as large as when training without external teachers (i.e., with the direct CE+ baseline). This is likely due to the saturation of information, since we already have the extra distillation from the three external teachers. A similar saturation phenomenon was also reported in [9]. The performance of TEACHTEXT plateaued after three external teachers had been used for distillation, and no significant further improvement was observed.

Table 7: **Comparison with other methods on the DiDeMo dataset.**

Method	R@1↑	R@5↑	R@10↑	MnR↓
S2VT [53]	11.9	33.6	-	-
FSE [60]	13.9	36.0	-	-
MEE [40]	16.1	41.2	55.2	43.7
CE [9]	17.1	41.9	56.0	-
TT-CE [9]	21.0	47.5	61.9	-
CE+ [9]	18.2	43.9	57.1	-
ClipBERT [33]	20.4	48.0	60.8	-
TT-CE+ [9]	21.6	48.6	62.9	-
Frozen [1]	34.6	65.0	74.7	-
MDMMT [13]	38.9	69.0	79.7	-
CLIP4Clip [37] (reported in [37])	**43.4**	**70.2**	**80.6**	**17.5**
CLIP4Clip-rerun (frozen layers + smaller batches)	42.0	69.1	78.1	18.8
CLIP4Clip-rerun + Caption Distillation (Ours)	43.2	69.7	79.2	**17.5**

4.6 Comparison to other methods

Using the proposed caption distillation loss with CLIP4CLIP, we achieve better text-to-video retrieval performance than the previous state-of-the-art results, as can be seen in Table 5 for MSRVTT 1k-A and Table 6 for MSVD datasets. The results on the DiDeMo dataset is shown in Table 7, and our method is not better than the current state-of-the-art due to the lack of computational resources to follow the recommended experiment setting. As explained in Sec. 4.1, for DiDeMo, we has to freeze some network layers and use much smaller batch size. This method is denoted as CLIP4CLIP-rerun in Table 7, and our method outperforms this direct baseline.

5 Conclusions

In this paper, we proposed a novel knowledge distillation loss for cross modal text-to-video and video-to-text retrieval. The new loss exploits the information from features of the same domain as knowledge to guide the similarity learning for the cross domain matching. This information does not require any external data or additional annotation and can be drawn directly from either the text or video features for distillation. This loss function can be combined with the original loss function of a retrieval framework. More importantly, our proposed knowledge distillation loss is framework agnostic, and is applicable to any retrieval framework. Extensive experiments on three retrieval frameworks and four large-scale datasets for cross modal retrieval show the benefits of our method.

Acknowledgements This material is based on research that is supported in part by the Air Force Research Laboratory (AFRL), DARPA, for the KAIROS program under agreement number FA8750-19-2-1003.

References

1. 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 International Conference on Computer Vision (2021)
2. Breiman, L., Shang, N.: Born again trees. University of California, Berkeley, Berkeley, CA, Technical Report **1**(2) (1996)
3. 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)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
5. Chen, D., Dolan, W.B.: Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (2011)
6. Chen, S., Zhao, Y., Jin, Q., Wu, Q.: Fine-grained video-text retrieval with hierarchical graph reasoning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2020)
7. Chen, Y., Xian, Y., Koepke, A., Shan, Y., Akata, Z.: Distilling audio-visual knowledge by compositional contrastive learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2021)
8. Chun, S., Oh, S.J., De Rezende, R.S., Kalantidis, Y., Larlus, D.: Probabilistic embeddings for cross-modal retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8415–8424 (2021)
9. 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 International Conference on Computer Vision (2021)
10. Deng, D., Liu, H., Li, X., Cai, D.: Pixellink: Detecting scene text via instance segmentation. In: Proceedings of AAAI Conference on Artificial Intelligence (2018)
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Association for Computational Linguistics (2018)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of International Conference on Learning and Representation (2021)
13. Dzabraev, M., Kalashnikov, M., Komkov, S., Petiushko, A.: Mdmmt: Multidomain multimodal transformer for video retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2021)
14. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: Improving visual-semantic embeddings with hard negatives. In: Proceedings of the British Machine Vision Conference (2018)
15. Gabeur, V., Sun, C., Alahari, K., Schmid, C.: Multi-modal Transformer for Video Retrieval. In: Proceedings of the European Conference on Computer Vision (2020)
16. Ghadiyaram, D., Tran, D., Mahajan, D.: Large-scale weakly-supervised pre-training for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
17. Ging, S., Zolfaghari, M., Pirsiavash, H., Brox, T.: Coot: Cooperative hierarchical transformer for video-text representation learning. In: Advances in Neural Information Processing Systems. vol. 33, pp. 22605–22618 (2020)

18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
19. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Proceedings of the European Conference on Computer Vision (2016)
20. Hendricks, L.A., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with temporal language. In: Proceedings of the International Conference on Computer Vision (2017)
21. Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: Cnn architectures for large-scale audio classification. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (2017)
22. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
23. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)
24. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. pp. 7132–7141 (2018)
25. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
26. Huang, Z., Niu, G., Liu, X., Ding, W., Xiao, X., Wu, H., Peng, X.: Learning with noisy correspondence for cross-modal matching. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 29406–29419 (2021)
27. Jha, A., Kumar, A., Banerjee, B., Namboodiri, V.: Sd-mtcnn: Self-distilled multitask cnn. In: Proceedings of the British Machine Vision Conference (2020)
28. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
29. Koepke, A., Wiles, O., Zisserman, A.: Visual pitch estimation. In: Proceedings of the Sound and Music Computing Conference (2019)
30. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Niebles, J.C.: Dense-captioning events in videos. In: Proceedings of the International Conference on Computer Vision (2017)
31. Le, Q.V., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G.S., Dean, J., Ng, A.Y.: Building high-level features using large scale unsupervised learning. In: Proceedings of the International Conference on Machine Learning (2012)
32. Lee, S., Song, B.C.: Graph-based knowledge distillation by multi-head attention network. In: Proceedings of the British Machine Vision Conference (2019)
33. 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 Conference on Computer Vision and Pattern Recognition (2021)
34. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: Proceedings of the European Conference on Computer Vision (2016)
35. Liu, Y., Albanie, S., Nagrani, A., Zisserman, A.: Use what you have: Video retrieval using representations from collaborative experts. In: Proceedings of the British Machine Vision Conference (2019)
36. Liu, Y., Wang, Z., Jin, H., Wassell, I.: Synthetically supervised feature learning for scene text recognition. In: Proceedings of the European Conference on Computer Vision (2018)

37. 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. arXiv preprint arXiv:2104.08860 (2021)
38. 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 Conference on Computer Vision and Pattern Recognition. pp. 9826–9836 (2021)
39. Miech, A., Alayrac, J.B., Smaira, L., Laptev, I., Sivic, J., Zisserman, A.: End-to-end learning of visual representations from uncurated instructional videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9879–9889 (2020)
40. Miech, A., Laptev, I., Sivic, J.: Learning a text-video embedding from incomplete and heterogeneous data. arXiv preprint arXiv:1804.02516 (2018)
41. 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 International Conference on Computer Vision (2019)
42. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: ICLR Workshop (2013)
43. Mithun, N.C., Li, J., Metze, F., Roy-Chowdhury, A.K.: Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In: ACM International Conference on Multimedia Retrieval (2018)
44. Nguyen, N., Nguyen, T., Tran, V., Tran, T., Ngo, T., Nguyen, T., Hoai, M.: Dictionary-guided scene text recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2021)
45. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
46. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems. vol. 32, pp. 8026–8037 (2019)
47. Patrick, M., Huang, P.Y., Asano, Y., Metze, F., Hauptmann, A., Henriques, J., Vedaldi, A.: Support-set bottlenecks for video-text representation learning. In: Proceedings of International Conference on Learning and Representation (2021)
48. Phuong, M., Lampert, C.H.: Distillation-based training for multi-exit architectures. In: Proceedings of the International Conference on Computer Vision (2019)
49. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)
50. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
51. Tang, Z., Lei, J., Bansal, M.: Decembert: Learning from noisy instructional videos via dense captions and entropy minimization. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 2415–2426 (2021)
52. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)

53. Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., Saenko, K.: Translating videos to natural language using deep recurrent neural networks. arXiv preprint arXiv:1412.4729 (2014)
54. Wang, L., Li, Y., Huang, J., Lazebnik, S.: Learning two-branch neural networks for image-text matching tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(2), 394–407 (2018)
55. Wray, M., Doughty, H., Damen, D.: On semantic similarity in video retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3650–3660 (2021)
56. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: Proceedings of the European Conference on Computer Vision (2018)
57. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
58. Yu, Y., Kim, J., Kim, G.: A joint sequence fusion model for video question answering and retrieval. In: Proceedings of the European Conference on Computer Vision (2018)
59. Zhai, A., Wu, H.Y.: Classification is a strong baseline for deep metric learning. In: Proceedings of the British Machine Vision Conference (2018)
60. Zhang, B., Hu, H., Sha, F.: Cross-modal and hierarchical modeling of video and text. In: Proceedings of the European Conference on Computer Vision (2018)
61. Zhu, L., Yang, Y.: Actbert: Learning global-local video-text representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8746–8755 (2020)