Cross-Lingual Cross-Modal Consolidation for Effective Multilingual Video Corpus Moment Retrieval

Jiaheng Liu[1,2], **Tan Yu**[1], **Hanyu Peng**[1], **Mingming Sun**[1], **Ping Li**[1]
[1]Cognitive Computing Lab, Baidu Research
[2]State Key Lab of Software Development Environment, Beihang University
liujiaheng@buaa.edu.cn
{tanyu01,penghanyu,sunmingming01,liping11}@baidu.com

Abstract

Existing multilingual video corpus moment retrieval (mVCMR) methods are mainly based on a two-stream structure. The visual stream utilizes the visual content in the video to estimate the query-visual similarity, and the subtitle stream exploits the query-subtitle similarity. The final query-video similarity ensembles similarities from two streams. In our work, we propose a simple and effective strategy termed as Cross-lingual Cross-modal Consolidation (C^3) to improve mVCMR accuracy. We adopt the ensemble similarity as the teacher to guide the training of each stream, leading to a more powerful ensemble similarity. Meanwhile, we use the teacher for a specific language to guide the student for another language to exploit the complementary knowledge across languages. Extensive experiments on mTVR dataset demonstrate the effectiveness of our C^3 method.

1 Introduction

Video Corpus Moment Retrieval (VCMR) task has been proposed in (Escorcia et al., 2019; Lei et al., 2020), which aims to retrieve a short moment from a large video corpus given a natural language query. Recently, (Lei et al., 2021a) introduces a multilingual Video Corpus Moment Retrieval (mVCMR) task. Compared with VCMR, mVCMR supports queries in multiple languages. It is more useful in practice, especially in international applications.

To facilitate the research in mVCMR, (Lei et al., 2021a) builds a large-scale mTVR dataset, where queries are in two languages. Apart from the video's visual content, the video's textual subtitles are provided as auxiliary information to help the query-to-video retrieval. (Lei et al., 2021a) proposes an mXML model to generate the query-video similarity for video retrieval and the query-clip similarity for moment localization with a two-stream structure. The subtitle stream captures the similarity between the text query and the video's textual subtitles. In parallel, the visual stream describes the similarity between the text query and the video's visual content. The final text-video similarity is a summation of the similarities from two streams.

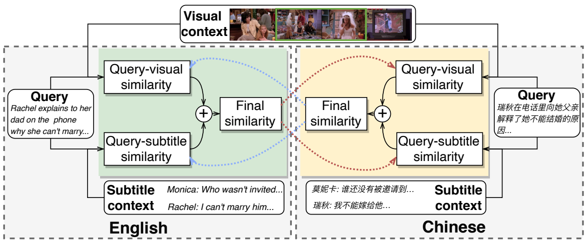
Visual context
Query
Rachel explains to her dad on the phone why she can't marry...
Query-visual similarity
Query-subtitle similarity
Final similarity
Subtitle context
Monica: Who wasn't invited...
Rachel: I can't marry him...
English
Query-visual similarity
Query-subtitle similarity
Final similarity
Query
瑞秋在电话里向她父亲解释了她不能结婚的原因...
莫妮卡: 谁还没有被邀请到...
瑞秋: 我不能嫁给他...
Subtitle context
Chinese

Figure 1: Illustration of our Cross-lingual Cross-modal Consolidation (C^3). The left green and right yellow boxes contain similarities using English and Chinese queries, respectively. Query-visual similarity measures the relevance between the query and visual context, and query-subtitle similarity denotes that between the query and subtitle context. The final similarity is obtained by summing up the query-visual and query-subtitle similarities. The final similarity in a specific language serves as the teacher to guide the learning of query-subtitle/query-visual similarity in another language.

As the final similarity is obtained from summing up similarities of two streams, straightforwardly, it is reasonable to hypothesize that improving the effectiveness of the similarity from each stream is beneficial to enhancing the performance of the final similarity. Meanwhile, since the final similarity fuses the information from two modalities, it is also reasonable to hypothesize that the final similarity is more reliable than the similarity from each stream. Based on the above two hypotheses, we propose a simple approach to improve the performance of two-stream architecture for the mVCMR task. To be specific, we use the final similarity fusing two modalities as the teacher and the similarity from each stream as the student. We train the student through the guidance of the teacher's knowledge. Meanwhile, to exploit the natural compensation across languages, we devise a student in one language and the teacher in another language. We term our method as cross-lingual cross-modal con-

Findings of the Association for Computational Linguistics: NAACL 2022, pages 1854 - 1862
July 10-15, 2022 ©2022 Association for Computational Linguistics

solidation (C^3), as visualized in Figure 1. Comprehensive experimental results on the mTVR dataset demonstrate the effectiveness of our C^3 method.

2 Related Work

Text-video retrieval. Traditionally, text-video retrieval (Rohrbach et al., 2015; Xu et al., 2016) is normally tackled through two mainstream methods: joint-embedding methods (Xu et al., 2015; Torabi et al., 2016; Pan et al., 2016; Plummer et al., 2017; Miech et al., 2019) and attention-based methods (Yu et al., 2017, 2018; Hori et al., 2017; Krishna et al., 2017). Recently, inspired by the great success of pre-training achieved by Transformer (Vaswani et al., 2017) and BERT (Devlin et al., 2019) in natural language processing, Transformer/BERT-based models emerge for solving text-video retrieval (Sun et al., 2019b,a; Li et al., 2020; Luo et al., 2020; Lei et al., 2021b).

Video corpus moment retrieval. Video corpus moment retrieval (Escorcia et al., 2019; Lei et al., 2020) aims to retrieve the ground-truth video from the whole corpus and predict the moment with high Intersection-of-Union (IoU) with the ground-truth moment using the natural language query. In practice, videos are often associated with other modalities (e.g., subtitles), so the multi-modal moment retrieval task with both visual and text contexts has been proposed. The recent work mTVR (Lei et al., 2021a) extends the monolingual moment retrieval task to the multilingual setting, and introduces a large-scale multilingual moment retrieval (i.e., mTVR) dataset, where the language queries and subtitles are in two languages (i.e., English and Chinese). Meanwhile, it proposes the mXML model jointly trained on the English and Chinese data for multilingual video moment retrieval.

Knowledge distillation. (Hinton et al., 2015) proposes knowledge distillation (KD), where the student network is trained by the soft output of the teacher network. Recently, knowledge distillation has demonstrated the effectiveness for many vision and language tasks (Tan et al., 2018; Hu et al., 2020; Fu et al., 2021; Wang et al., 2020; Fei et al., 2021; Sun et al., 2019c; Hou et al., 2020; Sanh et al., 2019; Peng et al., 2019; Jin et al., 2019; Liu et al., 2022, 2020, 2021). For example, (Sun et al., 2019c) proposes a Patient Knowledge Distillation method to compress an original large model into an equally-effective lightweight shallow network for pre-trained language models (Devlin et al., 2019).

3 Method

3.1 Preliminary

Problem Definition. The current multilingual moment retrieval model mXML (Lei et al., 2021a) performs video retrieval at its shallow layers and moment localization at its deep layers. We denote a query by q_g, where g denotes the language type, *e.g.*, English. A video v consists of L consecutive moments $\{c_l\}_{l=1}^{L}$. Each moment c_l is paired with subtitle $s_{g,l}$. mXML generates the query-video score $S(q_g, v)$ for video retrieval and the index of the start/end frame $t_{\text{st}}/t_{\text{ed}}$ for moment localization:

$$[S(q_g, v), t_{\text{st}}, t_{\text{ed}}] = \text{mXML}(q_g, \{(c_l, s_{g,l})\}_{l=1}^{L}).$$

mXML supports English (en) and Chinese (zh), *i.e.*, $g \in \{\text{en}, \text{zh}\}$. Below, we describe the video retrieval and moment localization in mXML briefly. You can refer to the Appendix A for more details.

Input Feature. ResNet-152 (He et al., 2016) and I3D (Carreira and Zisserman, 2017) extract the visual features of each video moment. The language features are extracted by RoBERTa-base (Liu et al., 2019) for English (Liu et al., 2019) and Chinese (Cui et al., 2020), respectively. Self-Encoder (SE) based on Transformer (Vaswani et al., 2017) and modular attention (Lei et al., 2020) are used to further encode the visual and text features.

Video Retrieval. The subtitle-based score for video retrieval $S_s(q_g, v)$ and the visual-based score $S_v(q_g, v)$ are obtained by two streams, respectively. The details of obtaining $S_s(q_g, v)$ and $S_v(q_g, v)$ are shown in Appendix A.3. The final video retrieval score $S(q_g, v)$ using both contexts is devised as

$$S(q_g, v) = S_s(q_g, v) + S_v(q_g, v). \quad (1)$$

Moment Localization. The subtitle-based query-clip score $S_s(q_g, c_l)$ and the visual-based query-clip score $S_v(q_g, c_l)$ are computed by two streams, respectively. The details of computing $S_s(q_g, c_l)$ and $S_v(q_g, c_l)$ are shown in Appendix A.4. The final query-clip score is devised as

$$S(q_g, c_l) = S_s(q_g, c_l) + S_v(q_g, c_l). \quad (2)$$

Then, to produce moment localization predictions from $S(q_g, c_l)$, mXML predicts the start and end probabilities $\mathbf{p}_g^{\text{st}}, \mathbf{p}_g^{\text{ed}} \in \mathbb{R}^L$ for each query.

For mXML, the video retrieval loss $\mathcal{L}^{\text{vr}}$, moment localization loss $\mathcal{L}^{\text{loc}}$ and language neighborhood constraint loss $\mathcal{L}^{\text{nc}}$ are illustrated in Appendix A.5. The final loss function of mXML is devised as

$$\mathcal{L}_{\text{mXML}} = \mathcal{L}^{\text{vr}} + \lambda_1 \mathcal{L}^{\text{loc}} + \lambda_2 \mathcal{L}^{\text{nc}}, \tag{3}$$

where λ_1 and λ_2 are the loss weights.

3.2 C^3 in Video Retrieval

In Eq. (1), the final video-level similarity $S(q_g, v)$ is a summation of scores from two modalities, $S_v(q_g, v)$ and $S_s(q_g, v)$. We use $S(q_g, v)$ with knowledge of two modalities as the teacher and distill (Hinton et al., 2015) its knowledge to the score with information of only a single modality.

In the multilingual scenario, to exploit more complementary knowledge, a more effective approach is to distill the scores from a language $g \in \{\text{en}, \text{zh}\}$ to the scores from another language $h \in \{\text{en}, \text{zh}\}$:

$$\begin{aligned} S(q_h, v) &\xrightarrow{\text{distill}} S_v(q_g, v), \\ S(q_h, v) &\xrightarrow{\text{distill}} S_s(q_g, v). \end{aligned} \tag{4}$$

Given a mini-batch of query-video pairs $\{(q_g^i, v^i)\}_{i=1}^n$, where n is the batchsize, $S(q_h^i, v^k)$ is the similarity score between q_h^i and v^k of the teacher model based on two modalities (i.e., visual and subtitle contexts), where $k \in [1, n]$. $S_v(q_g^i, v^k)$ and $S_s(q_g^i, v^k)$ are the corresponding similarity scores of the student model from visual and subtitle contexts, respectively. Then, for each query q_h^i, we can generate the teacher scores $\{S(q_h^i, v^k)\}_{k=1}^n$, and perform the softmax function with temperature τ_{vr} on the scores to obtain the normalized score:

$$\hat{S}(q_h^i, v^i) = \frac{e^{S(q_h^i, v^i)/\tau_{\text{vr}}}}{\sum_{k=1}^n e^{S(q_h^i, v^k)/\tau_{\text{vr}}}}. \tag{5}$$

In the same manner, we obtain the normalized student scores $\hat{S}_v(q_h^i, v^i)$ and $\hat{S}_s(q_h^i, v^i)$. Finally, the C^3 loss for video retrieval is devised as

$$\begin{aligned} \mathcal{L}^{\text{vr}}_{\text{C}^3} = \sum_{i=1}^{n} \sum_{g\in\{\text{en,zh}\}} \sum_{h\neq g} \frac{-1}{n} [&\hat{S}(q_h^i, v^i)\log(\hat{S}_v(q_h^i, v^i)) \\ &+ \hat{S}(q_h^i, v^i)\log(\hat{S}_s(q_h^i, v^i))]. \end{aligned} \tag{6}$$

The loss for video retrieval is devised as

$$\mathcal{L}^{\text{vr}}_{+} = \mathcal{L}^{\text{vr}} + \alpha \mathcal{L}^{\text{vr}}_{\text{C}^3}, \tag{7}$$

where α is a pre-defined positive constant.

3.3 C^3 in Moment Localization

Similarly, C^3 can also be used on the moment localization. In Eq. (2), we generate the query-clip similarity score using two contexts, and then produce the start and end probabilities. In the same way, based on $S_s(q_g, c_l)$ with single subtitle context, we can generate the start and end probabilities $\mathbf{p}^{\text{st}}_{g,s}, \mathbf{p}^{\text{ed}}_{g,s} \in \mathbb{R}^L$, and based on $S_v(q_g, c_l)$ with single visual context, we can generate $\mathbf{p}^{\text{st}}_{g,v}, \mathbf{p}^{\text{ed}}_{g,v} \in \mathbb{R}^L$. Note that we use the softmax function with temperature τ_{loc} to generate the start and end probabilities in the similar way of Eq. (5). We define the start and end probabilities of the teacher model from language h as $\mathbf{p}^{\text{st}}_h$ and $\mathbf{p}^{\text{ed}}_h$, and the start and end probabilities of the student model from language g as $\mathbf{p}^{\text{st}}_{g,v}, \mathbf{p}^{\text{ed}}_{g,v}$ using visual context and $\mathbf{p}^{\text{st}}_{g,s}, \mathbf{p}^{\text{ed}}_{g,s}$ using subtitle context. Thus, the C^3 loss for moment localization is defined as follows:

$$\begin{aligned} \mathcal{L}^{\text{loc}}_{\text{C}^3} = \sum_{g\in\{\text{en,zh}\}} \sum_{h\neq g} [&\text{CE}(\mathbf{p}^{\text{st}}_h, \mathbf{p}^{\text{st}}_{g,v}) + \text{CE}(\mathbf{p}^{\text{st}}_h, \mathbf{p}^{\text{st}}_{g,s}) \\ &+ \text{CE}(\mathbf{p}^{\text{ed}}_h, \mathbf{p}^{\text{ed}}_{g,v}) + \text{CE}(\mathbf{p}^{\text{ed}}_h, \mathbf{p}^{\text{ed}}_{g,s})], \end{aligned} \tag{8}$$

where CE() is cross-entropy function defined as

$$\text{CE}(\mathbf{x}, \mathbf{y}) = -\sum_{l=1}^{L} \mathbf{x}[l]\log(\mathbf{y}[l]). \tag{9}$$

The loss for moment localization is as follows:

$$\mathcal{L}^{\text{loc}}_{+} = \mathcal{L}^{\text{loc}} + \beta \mathcal{L}^{\text{loc}}_{\text{C}^3}, \tag{10}$$

where β is a pre-defined positive constant.

3.4 Training and Inference

Final loss function. Considering our proposed C^3, the final loss function is constructed as follows:

$$\mathcal{L}_{\text{mXML+}} = \mathcal{L}^{\text{vr}}_{+} + \lambda_1 \mathcal{L}^{\text{loc}}_{+} + \lambda_2 \mathcal{L}^{\text{nc}}, \tag{11}$$

which is similar to the formulation of $\mathcal{L}_{\text{mXML}}$ defined in Eq. (3) but replaces $\mathcal{L}^{\text{vr}}$ by its counterpart $\mathcal{L}^{\text{vr}}_{+}$ and $\mathcal{L}^{\text{loc}}$ by $\mathcal{L}^{\text{loc}}_{+}$ based on our C^3.

Training. The training consists of two stages. In the first stage, we train the standard mXML model using the loss function $\mathcal{L}_{\text{mXML}}$ in Eq. (3) to obtain the teacher model, which produces the query-video score $S(q_g, v)$ and the query-clip score $S(q_g, c_l)$ with knowledge of two modalities. Then, in the second stage, we use the teacher mXML model to distill the training process of the randomly initialized student mXML model using the loss function $\mathcal{L}_{\text{mXML+}}$ in Eq. (11). After training, the student

Algorithm 1: The training process.

```
Train the teacher $\mathrm{mXML_T}$, a standard mXML model, using $\mathcal{L}_{\mathrm{mXML}}$ in Eq. (3).
for $i \in [1, Q]$ do
    Initialize a student model $\mathrm{mXML_S}$.
    $\mathrm{mXML_T} \xrightarrow{\text{distill}} \mathrm{mXML_S}$ using $\mathcal{L}_{\mathrm{mXML+}}$ in Eq. (11).
    $\mathrm{mXML_T} = \mathrm{mXML_S}$.
end
return $\mathrm{mXML_S}$
```

Table 1: R@1 on the test-public split of mTVR.

Methods	English		Chinese	
	IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7
MCN	0.02	0.00	0.13	0.02
CAL	0.09	0.04	0.11	0.04
MEE+MCN	0.92	0.42	1.43	0.64
MEE+CAL	0.97	0.39	1.51	0.62
MEE+ExCL	0.92	0.33	1.43	0.72
XML	7.25	3.25	5.91	2.57
mXML	8.30	3.82	6.76	3.20
C^3	**9.11**	**4.72**	**7.05**	**4.08**

Table 2: R@10 on the test-public split of mTVR.

Methods	English		Chinese	
	IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7
XML	22.79	13.96	18.93	11.13
mXML	23.27	13.98	18.99	11.52
C^3	**24.72**	**16.00**	**20.62**	**13.30**

mXML model will perform better than the teacher model. Thus, we utilize the trained student mXML model as the new teacher mXML model to guide the training of a new randomly initialized student mXML model from the beginning. We repeat the distillation process in the second stage for Q iterations until the performance saturates. For better clarification, we summarize the training process of our proposed C^3 method as shown in Algorithm 1.

Inference. Since our method is orthogonal to the model, the inference process is same as the mXML.

4 Experiments

Dataset. mTVR (Lei et al., 2021a) is a large-scale multilingual video moment retrieval dataset, which contains 218 thousand English and Chinese queries from 21.8 thousand TV show video clips. This dataset extends the TVR dataset (in English) (Lei et al., 2020) and adds Chinese queries and subtitles. (Lei et al., 2021a) proposed to split the mTVR dataset into 80% train, 10% val, 5% test-public and 5% test-private datasets. We use this dataset to validate the effectiveness of our method for research purpose. All experiments in our work are conducted on one NVIDIA Tesla V100 GPU.

Experimental setting. We report the average recall at K (i.e., R@K) for multilingual Video Corpus Moment Retrieval (mVCMR) task on the mTVR (Lei et al., 2021a) dataset, where the predicted moment is right when it has high Intersection-over-Union (IoU) with the ground-truth. We use the same training strategy and network architecture of mXML (Lei et al., 2021a). λ_1 and λ_2 of Eq. (3) and Eq. (11) are set as 0.01, 1, respectively. The loss weights α and β of $\mathcal{L}_{\mathrm{C3}}^{\mathrm{vr}}$ and $\mathcal{L}_{\mathrm{C3}}^{\mathrm{loc}}$ for video retrieval and moment localization are set as 1.0, 100, respectively. The τ_{vr} is set as 0.02, and τ_{loc} is set as 1. The number of distillation iterations, Q, is set as 2. We deploy our proposed C^3 in mXML and abbreviate the mXML with our C^3 to C^3.

4.1 Main Results

In Table 1, on the test-public split[1] of the mTVR dataset, we compare C^3 with existing methods, including proposal-based approaches (MCN (Anne Hendricks et al., 2017) and CAL (Escorcia et al., 2019)), reranking-based methods (MEE (Miech et al., 2018)+MCN, MEE+CAL, MEE+ExCL (Ghosh et al., 2019) and XML (Lei et al., 2020)) and the state-of-the-art mXML (Lei et al., 2021a). We observe that our C^3 achieves consistently higher R@1 in both English and Chinese than the baseline methods. In Table 2, C^3 also considerably improves the R@10 of the XML and mXML on the test-public split of the mTVR dataset. Note that the R@10 results of XML is based on our re-implementation. In Table 3, we compare the performance on the val split of the mTVR dataset, and C^3 also outperforms the mXML a lot, which further shows the advantage of our method.

4.2 Ablation Study and Analysis

Effect of different components. By default, we exploit the C^3 in both video retrieval ($\mathcal{L}_{\mathrm{C3}}^{\mathrm{vr}}$ in Eq. (7)) and moment localization ($\mathcal{L}_{\mathrm{C3}}^{\mathrm{loc}}$ in Eq. (10)). We conduct the experiments to evaluate the importance of $\mathcal{L}_{\mathrm{C3}}^{\mathrm{loc}}$ and $\mathcal{L}_{\mathrm{C3}}^{\mathrm{vr}}$ by removing one of them, respectively. From Table 4, we observe that, by removing $\mathcal{L}_{\mathrm{C3}}^{\mathrm{loc}}$ or $\mathcal{L}_{\mathrm{C3}}^{\mathrm{vr}}$, the R@1 of video corpus moment retrieval considerably deteriorates. It validates the

[1] https://competitions.codalab.org/competitions/33493

Table 3: R@1 on the val split of mTVR.

Methods	English		Chinese	
	IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7
mXML	6.22	2.96	5.17	2.41
C^3	**7.44**	**3.85**	**5.70**	**2.86**

Table 4: R@1 of the proposed C^3 and its alternative variants by removing one of the components in video corpus moment retrieval on the val split of mTVR.

Methods	English		Chinese	
	IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7
w/o $\mathcal{L}_{C^3}^{\mathrm{loc}}$	6.55	3.13	5.55	2.54
w/o $\mathcal{L}_{C^3}^{\mathrm{vr}}$	6.71	3.54	5.53	2.72
C^3	**7.44**	**3.85**	**5.70**	**2.86**

Table 5: R@1 and R@5 in the video retrieval (VR) of different methods on the val split of mTVR.

Method	English		Chinese	
	R@1	R@5	R@1	R@5
mXML	19.35	42.32	17.75	39.34
C^3	**21.51**	**45.30**	**19.42**	**41.68**

Table 6: R@1 in the single video moment retrieval (SVMR) on the val split of mTVR.

Method	English		Chinese	
	IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7
mXML	29.05	13.20	26.31	11.46
C^3	**30.13**	**13.99**	**27.89**	**12.92**

necessity of using both $\mathcal{L}_{C^3}^{\mathrm{loc}}$ and $\mathcal{L}_{C^3}^{\mathrm{vr}}$ in our C^3.

Analysis on video retrieval. It is worth noting that we consider both the video retrieval and the moment localization in a single video for mVCMR task. To more comprehensively demonstrate the advantage of our C^3, we also investigate its influence in the video retrieval task. From Table 5, we can observe that our C^3 outperforms the baseline model, mXML, by a large margin.

Analysis on single video moment retrieval. To further demonstrate the effectiveness of the proposed C^3 in moment localization, we also report the single video moment retrieval (i.e., SVMR) results. From Table 6, we observe that our C^3 achieves significant performance improvement compared with mXML. It indicates that our C^3 predicts more accurate moment localization results, which shows the effectiveness of our C^3 in moment localization.

Extension on monolingual video corpus moment retrieval. Despite that the proposed C^3 method is devised to mVCMR, it is also naturally applicable to monolingual video corpus moment retrieval. Here, we evaluate the effectiveness of our C^3 in the monolingual setting. TVR dataset (Lei et al., 2020) contains 109 thousand queries collected on 21.8 thousand videos from 6 TV shows of diverse genres, where each video is associated with subtitles and each query is associated with a tight temporal window. Specifically, we use the XML (Lei et al., 2020) model as the baseline method, and propose an alternative variant by applying the our C^3 loss functions between the outputs of the multi-modality and to the outputs of the single modality for both the video retrieval and moment localization tasks following the similar strategy of the Eq. (6) and Eq. (8). In other words, this alternative variant is called as the intra-lingual cross-modal consolidation (IC^2) method. In Table 7, we compare our alternative variant IC^2 with the baseline model XML on both val and test-public splits of the TVR dataset for monolingual video corpus moment retrieval. As shown in Table 7, our IC^2 also achieves significant performance improvements on the TVR dataset, which further demonstrates the versatility of our proposed consolidation strategy.

Table 7: R@1 of different methods on both val and test-public splits of the TVR dataset.

Method	val		test-public	
	IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7
XML	5.28	2.62	7.25	3.25
IC^2	**6.27**	**2.93**	**8.45**	**4.00**

5 Limitations and Potential risks

Although our C^3 has achieved substantial improvement based on mXML on the mTVR dataset, we find that there exists some hyper-parameters (e.g., the τ_{vr}, τ_{loc}) to tune in C^3, which may be time-consuming. Besides, we develop the C^3 strategy to improve the performance of mVCMR task, and we have not seen the potential ricks in our paper.

6 Conclusion

In our work, for the multilingual video corpus moment retrieval (mVCMR), we introduce a simple and effective Cross-Lingual and Cross-Modal Consolidation (i.e., C^3) strategy. It enhances the reliability of the similarity score from a single modality through the knowledge distillation from the similarity score with access to the multi-modal information. Meanwhile, it exploits the complementary information across languages by cross-lingual knowledge distillation for both video retrieval and moment localization. Extensive experimental results demonstrate the effectiveness of our method.

References

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812.

Andrea Burns, Donghyun Kim, Derry Wijaya, Kate Saenko, and Bryan A Plummer. 2020. Learning to scale multilingual representations for vision-language tasks. In *European Conference on Computer Vision*, pages 197–213. Springer.

Joao Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Victor Escorcia, Mattia Soldan, Josef Sivic, Bernard Ghanem, and Bryan Russell. 2019. Temporal localization of moments in video collections with natural language. *arXiv preprint arXiv:1907.12763*.

Hongliang Fei, Tan Yu, and Ping Li. 2021. Cross-lingual cross-modal pretraining for multimodal retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3644–3650.

Hao Fu, Shaojun Zhou, Qihong Yang, Junjie Tang, Guiquan Liu, Kaikui Liu, and Xiaolong Li. 2021. Lrc-bert: Latent-representation contrastive knowledge distillation for natural language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12830–12838.

Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander G Hauptmann. 2019. Excl: Extractive clip localization using natural language descriptions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1984–1990.

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

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. 2017. Attention-based multimodal fusion for video description. In *Proceedings of the IEEE international conference on computer vision*, pages 4193–4202.

Lu Hou, Zhiqi Huang, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. 2020. Dynabert: Dynamic bert with adaptive width and depth. *Advances in Neural Information Processing Systems*, 33.

Hengtong Hu, Lingxi Xie, Richang Hong, and Qi Tian. 2020. Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3123–3132.

Xiao Jin, Baoyun Peng, Yichao Wu, Yu Liu, Jiaheng Liu, Ding Liang, Junjie Yan, and Xiaolin Hu. 2019. Knowledge distillation via route constrained optimization. In *ICCV*.

Donghyun Kim, Kuniaki Saito, Kate Saenko, Stan Sclaroff, and Bryan Plummer. 2020. Mule: Multimodal universal language embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11254–11261.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.

Jie Lei, Tamara L Berg, and Mohit Bansal. 2021a. mtvr: Multilingual moment retrieval in videos. *ACL*.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021b. 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*, pages 7331–7341.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 447–463. Springer.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *EMNLP*.

Haoliang Liu, Tan Yu, and Ping Li. 2021. Inflate and shrink: Enriching and reducing interactions for fast text-image retrieval. In *Proceedings of the 2021*

Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9796–9809, Virtual Event / Punta Cana, Dominican Republic.

Jiaheng Liu, Haoyu Qin, Yichao Wu, Jinyang Guo, Ding Liang, and Ke Xu. 2022. Coupleface: Relation matters for face recognition distillation. *arXiv preprint arXiv:2204.05502*.

Jiaheng Liu, Shunfeng Zhou, Yichao Wu, Ken Chen, Wanli Ouyang, and Dong Xu. 2020. Block proposal neural architecture search. *IEEE TIP*, 30:15–25.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. 2020. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*.

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

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640.

Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. 2016. Jointly modeling embedding and translation to bridge video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4594–4602.

Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. 2019. Correlation congruence for knowledge distillation. In *ICCV*.

Bryan A Plummer, Matthew Brown, and Svetlana Lazebnik. 2017. Enhancing video summarization via vision-language embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5781–5789.

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

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

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

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

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019c. Patient knowledge distillation for bert model compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4323–4332.

Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2018. Multilingual neural machine translation with knowledge distillation. In *International Conference on Learning Representations*.

Atousa Torabi, Niket Tandon, and Leonid Sigal. 2016. Learning language-visual embedding for movie understanding with natural-language. *arXiv preprint arXiv:1609.08124*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Fei Huang, and Kewei Tu. 2020. Structure-level knowledge distillation for multilingual sequence labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3317–3330.

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

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

Youngjae Yu, Jongseok Kim, and Gunhee Kim. 2018. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

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

A More detailed preliminary on mXML

The current multilingual moment retrieval model mXML (Lei et al., 2021a) is built upon the Cross-modal Moment Localization (XML) model (Lei et al., 2020), which performs efficient video-level retrieval at its shallow layers and accurate moment-level localization at its deep layers. To adjust to the multilingual settings in the video corpus moment retrieval (VCMR) task and improve the efficiency and effectiveness, mXML employs two strategies (i.e., encoder parameter sharing and language neighborhood constraint loss) to better utilize the multilingual data while maintaining smaller model size. In this section, we briefly review the mXML model, which is also illustrated in Figure 2.

A.1 Input Feature

ResNet-152 (He et al., 2016) and I3D (Carreira and Zisserman, 2017) extract the visual features of each video moment. The generated video moment visual features are denoted by $\mathbf{E}^v = [\mathbf{e}_1^v, \cdots, \mathbf{e}_L^v] \in \mathbb{R}^{d \times L}$, where d is the feature dimension. The language features are extracted by RoBERTa-base (Liu et al., 2019) for English (Liu et al., 2019) and Chinese (Cui et al., 2020), respectively. For queries, token-level features are used. The query features are denoted by $\mathbf{E}_g^q = [\mathbf{e}_{g,1}^q, \cdots, \mathbf{e}_{g,L_q}^q] \in \mathbb{R}^{d \times L_q}$, where L_q is the number of tokens and $g \in \{\mathrm{en}, \mathrm{zh}\}$. For subtitles, token-level features in a video moment are max-pooled into a single vector. The subtitle features are denoted by $\mathbf{E}_g^s = [\mathbf{e}_{g,1}^s, \cdots, \mathbf{e}_{g,L}^s] \in \mathbb{R}^{d \times L}$.

A.2 Encoding

mXML uses Self-Encoder (SE) implemented by Transformer (Vaswani et al., 2017) to further encode the query's token features:

$$\mathbf{H}_g^q = \mathrm{SE}(\mathbf{E}_g^q) = [\mathbf{h}_{g,1}^q, \cdots, \mathbf{h}_{g,L_q}^q].$$

A modular attention (Lei et al., 2020) is conducted on the query token features $\mathbf{H}_g^q$, generating two modularized query vectors $\mathbf{q}_g^v, \mathbf{q}_g^s \in \mathbb{R}^d$. In parallel, mXML encodes the moment subtitle features $\mathbf{E}_g^s$ and moment visual features through a stack of two Self-Encoders:

$$\mathbf{H}_{g,0}^s = \mathrm{SE}(\mathbf{E}_g^s),\ \mathbf{H}_{g,1}^s = \mathrm{SE}(\mathbf{H}_{g,0}^s),$$
$$\mathbf{H}_0^v = \mathrm{SE}(\mathbf{E}^v),\ \mathbf{H}_1^v = \mathrm{SE}(\mathbf{H}_0^v).$$

Among them, the output of the first Self-Encoder, $\mathbf{H}_{g,0}^s$ and $\mathbf{H}_0^v$, are used for video retrieval. The output of the second Self-Encoder, $\mathbf{H}_{g,1}^s$ and $\mathbf{H}_1^v$ are used for moment localization.

A.3 Video Retrieval

Given the modularized queries $\mathbf{q}_g^v, \mathbf{q}_g^s$ and the encoded contexts $\mathbf{H}_0^v$, $\mathbf{H}_{g,0}^s$, the video-level retrieval (VR) scores $S_s(q_g, v)$ and $S_v(q_g, v)$ using the subtitle context and the visual context are computed as follows, respectively:

$$\begin{aligned} S_s(q_g, v) &= \max_{l \in [1,L]} \cos(\mathbf{q}_g^s, \mathbf{H}_{g,0}^s[:, l]), \\ S_v(q_g, v) &= \max_{l \in [1,L]} \cos(\mathbf{q}_g^v, \mathbf{H}_0^v[:, l]), \end{aligned} \tag{12}$$

where $\mathbf{H}_{g,0}^s[:, l]$ denotes the l-th column vector in $\mathbf{H}_{g,0}^s$ and $\cos(\cdot, \cdot)$ measures the cosine similarity between two vectors. The score essentially computes the cosine similarity between each clip and query and picks the maximum. Then, the final video-level retrieval (VR) score $S(q_g, v)$ using both subtitle and visual contexts is defined as follows:

$$S(q_g, v) = S_s(q_g, v) + S_v(q_g, v). \tag{13}$$

A.4 Moment Localization

Given the modularized queries $\mathbf{q}_g^v, \mathbf{q}_g^s$ and the encoded contexts $\mathbf{H}_1^v$, $\mathbf{H}_{g,1}^s$, mXML computes the query-clip similarity scores $S_s(q_g, c_l)$ and $S_v(q_g, c_l)$ using the subtitle and visual contexts as follows, respectively:

$$\begin{aligned} S_s(q_g, c_l) &= \langle \mathbf{H}_1^v[:, l], \mathbf{q}_g^v \rangle, \\ S_v(q_g, c_l) &= \langle \mathbf{H}_{g,1}^s[:, l], \mathbf{q}_g^s \rangle, \end{aligned} \tag{14}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product between two vectors. Similarly, the final query-clip similarity score using both contexts is defined as follows:

$$S(q_g, c_l) = S_s(q_g, c_l) + S_v(q_g, c_l), \tag{15}$$

which is also the summation of these query-clip similarity scores. Then, to produce moment localization predictions from the final query-clip score $S(q_g, c_l)$, mXML adopts the Convolutional Start-End detector (ConvSE) with two 1D convolution filters for learning to detect start (up) and end (down) edges in the score curves and generate the start (st) $\mathbf{s}_g^{\mathrm{st}}$ and end (ed) scores $\mathbf{s}_g^{\mathrm{ed}}$, which are also shown as follows, respectively:

$$\begin{aligned} \mathbf{s}_g^{\mathrm{st}} &= \mathrm{ConvSE}_{\mathrm{st}}(S(q_g, c_l)), \\ \mathbf{s}_g^{\mathrm{ed}} &= \mathrm{ConvSE}_{\mathrm{ed}}(S(q_g, c_l)). \end{aligned} \tag{16}$$

Then, these scores are normalized with the softmax function to output the start and end probabilities $\mathbf{p}_g^{\mathrm{st}}, \mathbf{p}_g^{\mathrm{ed}} \in \mathbb{R}^L$ for each query.

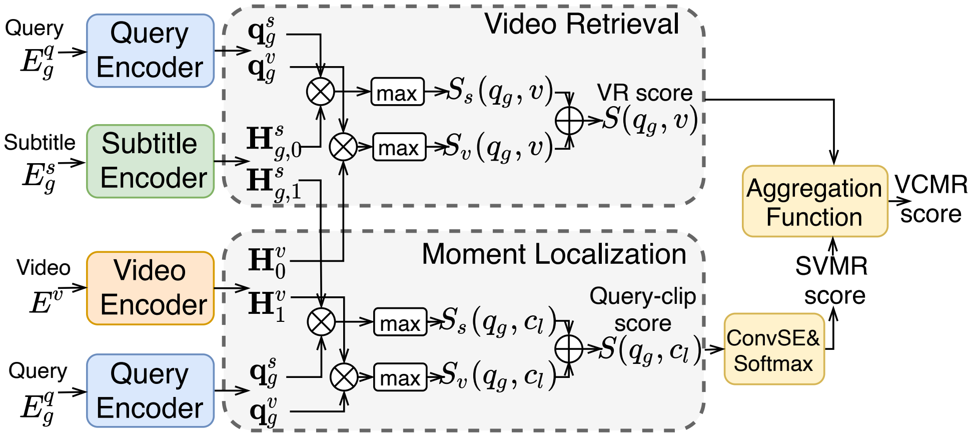

Figure 2: Illustration of inference process for multilingual video corpus moment retrieval (i.e., mVCMR) task. Here, we take the language $g \in \{\text{en}, \text{zh}\}$ as an example to show the processes of video retrieval and moment localization. For the mVCMR setting, all encoders are shared for different languages, and language neighborhood constraint is used on both query and subtitle embeddings. "ConvSE" and "Aggregation Function" operations are proposed in (Lei et al., 2020). "SVMR" denotes single video moment retrieval.

A.5 Training and Inference

In Figure 2, in the training process, mXML optimizes the video retrieval score and the moment localization probabilities based on the triplet loss and the cross-entropy loss, respectively. Besides, to facilitate more stronger multilingual learning, mXML also utilizes language neighborhood constraint loss for both query and subtitle embeddings based on the triplet loss. Below we introduce these loss functions in detail.

Video Retrieval Loss. For each positive pair (q_g^i, v^i), mXML samples two negative pairs (q_g^i, v^j) and (q_g^k, v^i) from the same mini-batch to calculate the combined hinge loss as follows:

$$\mathcal{L}^{\text{vr}} = \sum_{g \in \{\text{en},\text{zh}\}} \sum_{i=1}^{n} \frac{-1}{n} \{[S(q_g^i, v^i) - S(q_g^i, v^j)) + m_{\text{vr}}]_+ \\ + [S(q_g^i, v^i) - S(q_g^k, v^i)) + m_{\text{vr}}]_+\},$$

where $[x]_+ = \max(x, 0)$, m_{vr} is the margin and n is the number of samples for each mini-batch.

Moment Localization Loss. Given the start and end probabilities $\mathbf{p}_g^{\text{st}}, \mathbf{p}_g^{\text{ed}} \in \mathbb{R}^L$, the moment localization loss is defined as follows:

$$\mathcal{L}^{\text{loc}} = \sum_{g \in \{\text{en},\text{zh}\}} \sum_{i=1}^{n} \frac{-1}{n} [\log(\mathbf{p}_g^{\text{st}}(t_{\text{st}}^i)) + \log(\mathbf{p}_g^{\text{ed}}(t_{\text{ed}}^i))],$$

where t_{st}^i and t_{ed}^i are the ground-truth indices of the start and the end, respectively.

Language Neighborhood Constraint Loss. Following (Kim et al., 2020; Burns et al., 2020), mXML additionally adopts language neighborhood constraint loss for multilingual learning. It encourages sentences that express the same or similar meanings to be close to each other in the embedding space via a triplet loss. Given the i-th paired sentence embeddings $\mathbf{e}_{\text{en}}^i \in \mathbb{R}^d$ and $\mathbf{e}_{\text{zh}}^i \in \mathbb{R}^d$ from each mini-batch, mXML samples the j-th and the k-th negative sentence embeddings $\mathbf{e}_{\text{en}}^j$ and $\mathbf{e}_{\text{zh}}^k$ from this mini-batch, where $i \neq j$ and $i \neq k$. The language neighborhood constraint loss $\mathcal{L}^{\text{nc}}$ can be formulated as follows:

$$\mathcal{L}^{\text{nc}} = \sum_{i=1}^{n} \frac{-1}{n} \{[\cos(\mathbf{e}_{\text{en}}^i, \mathbf{e}_{\text{zh}}^k) - \cos(\mathbf{e}_{\text{en}}^i, \mathbf{e}_{\text{zh}}^i) + m_{\text{nc}}]_+ \\ + [\cos(\mathbf{e}_{\text{en}}^j, \mathbf{e}_{\text{zh}}^i) - \cos(\mathbf{e}_{\text{en}}^i, \mathbf{e}_{\text{zh}}^i) + m_{\text{nc}}]_+\},$$

where m_{nc} is the margin. The language neighborhood constraint loss is applied on both query and subtitle embeddings.

Overall, the final loss function of mXML $\mathcal{L}_{\text{mXML}}$ is defined as follows:

$$\mathcal{L}_{\text{mXML}} = \mathcal{L}^{\text{vr}} + \lambda_1 \mathcal{L}^{\text{loc}} + \lambda_2 \mathcal{L}^{\text{nc}}, \qquad (17)$$

where λ_1 and λ_2 are the loss weights of the moment localization loss and the language neighborhood constraint loss, respectively.

Inference. At inference, in Figure 2, for the video corpus moment retrieval task, the predicted start and end probabilities are employed to generate the single video moment retrieval (SVMR) score, where SVMR is to localize a video segment from a video under the language query. Then, the video retrieval score and the SVMR score are used to produce the final VCMR score using the aggregation function proposed in (Lei et al., 2020).