ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10094821

C2KD: CROSS-LINGUAL CROSS-MODAL KNOWLEDGE DISTILLATION FOR MULTILINGUAL TEXT-VIDEO RETRIEVAL

Andrew Rouditchenko[1], *Yung-Sung Chuang*[1], *Nina Shvetsova*[2], *Samuel Thomas*[3,4], *Rogerio Feris*[3,4],
Brian Kingsbury[3,4], *Leonid Karlinsky*[3,4], *David Harwath*[5], *Hilde Kuehne*[2,4], *James Glass*[1]

MIT[1] Goethe University Frankfurt[2] IBM Research AI[3] MIT-IBM Watson AI Lab[4] UT Austin[5]

ABSTRACT

Multilingual text-video retrieval methods have improved significantly in recent years, but the performance for languages other than English still lags. We propose a Cross-Lingual Cross-Modal Knowledge Distillation method to improve multilingual text-video retrieval. Inspired by the fact that English text-video retrieval outperforms other languages, we train a student model using input text in different languages to match the cross-modal predictions from teacher models using input text in English. We propose a cross entropy based objective which forces the distribution over the student's text-video similarity scores to be similar to those of the teacher models. We introduce a new multilingual video dataset, Multi-YouCook2, by translating the English captions in the YouCook2 video dataset to 8 other languages. Our method improves multilingual text-video retrieval performance on Multi-YouCook2 and several other datasets such as Multi-MSRVTT and VATEX. We also conducted an analysis on the effectiveness of different multilingual text models as teachers.

Index Terms— Cross-Lingual, Cross-Modal, Knowledge Distillation, Multilingual, Retrieval

1. INTRODUCTION

Text-video retrieval, or the task of searching for videos with text queries, is becoming increasingly important as more videos are uploaded to the internet. Currently, most methods developed for this task are trained and evaluated with English text. Our focus is to improve the performance of text-video retrieval on more languages.

Learning a multilingual multimodal embedding space [2, 3] has been useful for multilingual text-video retrieval. Text in different languages and video are processed by separate encoders and projected into a shared embedding space where text and video that are semantically related should be close together regardless of the language. During inference, text queries and candidate videos are projected into the embedding space, and videos are ranked according to the similarity scores between the text and video embeddings. These methods are trained with a cross-modal contrastive objective on video datasets with parallel text translations in multiple languages, which are often derived from the original captions in English using machine translation. They leverage recently available multilingual models pre-trained on many languages [4, 5] to process text in different languages with only a single encoder.

While these methods have improved multilingual text-video retrieval, the performance for English is usually higher than for other languages. To address the gap in performance between English and multilingual text-video retrieval, we propose C2KD: Cross-Lingual Cross-Modal Knowledge Distillation. Our method trains a *student* model to learn better multilingual text-video similarity scores by learning from the English text-video scores of multiple trained and frozen *teachers*. The student learns to pull together video and multilingual text embeddings by optimizing their text-video scores through the contrastive loss. We introduce a framework where several trained and frozen teachers simultaneously process the English translations of the student's inputs and predict English text-video scores. Further, we propose a cross entropy based objective between the student's multilingual text-video scores and the teachers' English text-video scores. This teaches the student to learn multilingual text-video scores which are more aligned with the English scores, thus improving the multilingual text-video retrieval performance.

We applied our method to two existing multilingual text-video datasets: Multi-MSRVTT [2] and VATEX [6]. Since these datasets are mainly focused on open-domain videos, we collected the Multi-YouCook2 dataset as an extension of the YouCook2 [7] cooking video dataset to test the model in a domain which requires more fine-grained reasoning, such as understanding specific ingredients in recipes. Our results show that C2KD can improve multilingual text-video retrieval performance on all datasets, despite the variety in languages, domains, and sizes. We plan to release the code, data, and pre-trained models.

2. RELATED WORK

Recent work introduced methods to improve multilingual text-video retrieval. Huang et al. [2] demonstrated text-video retrieval in 9 languages. Their model is trained with a cross-modal contrastive objective to pull together the embeddings of parallel text translations and video inputs. Akula et al. [3] augment a text-video triplet loss with hard negatives which improved performance on low-resource languages. We observed that English text-video retrieval outperformed other languages, which motivated our approach.

Multilingual text-video retrieval methods rely on pre-trained multilingual text encoders to handle many languages with a single model. MBERT [4] and XLM-R [5] learn multilingual representations through masked language modeling. LaBSE [8] is instead trained to maximize the similarity of translation pairs in a shared embedding space. Another approach for training a multilingual text model is to distill the knowledge [9] from a monolingual model. Distill Sentence BERT [10] is initialized from XLM-R and trained to output similar multilingual embeddings to Sentence BERT [11] using English translations as input. Our approach has a similar idea, but it incorporates visual context. We use English text as input to several cross-modal teachers, and train a student to output similar text-video scores using text in other languages.

Of most relevance to our work, TeachText [12] introduced cross-modal Knowledge Distillation for English text-video retrieval. They

This research was supported by the MIT-IBM Watson AI Lab and the MIT SuperCloud [1]. We thank Alex H. Liu and Yuan Gong for helpful comments.

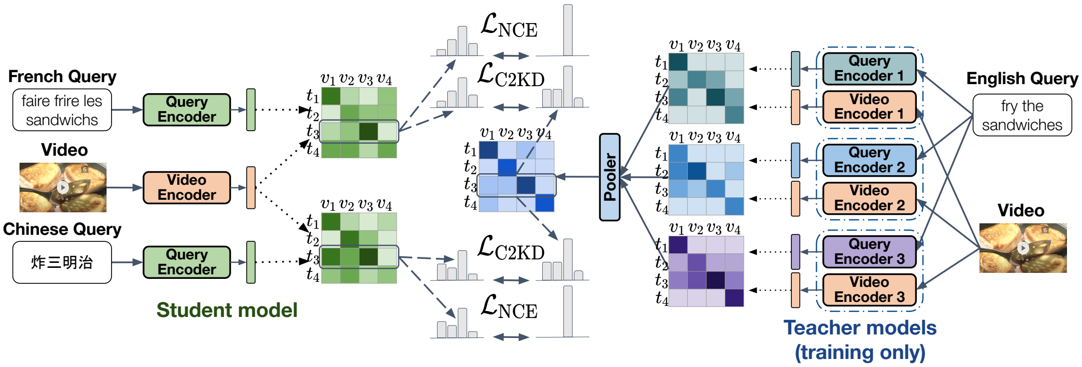

Fig. 1. **Overview of C2KD.** A multilingual student model computes text-video scores for a batch of video and text inputs, while teacher models process the same video and English translations. The student is trained with two objectives. $\mathcal{L}_{NCE}$ (described in Section 3.1) trains the model to have high text-video scores for text and video pairs using the cross entropy loss. $\mathcal{L}_{C2KD}$ (described in Section 3.3) distills the knowledge from the teacher English text-video scores using a cross entropy loss.

use teacher retrieval models with various English text embeddings and train a student to output similar text-video scores with a regression loss. Our approach has several major differences. First, our text and models are multilingual. Second, our teachers use English input instead of using the same multilingual input as the students. Third, we use a cross entropy objective between the student and teacher text-video scores instead of using a regression loss, which is more effective since it considers the context of all of the text-video pairs in the batch. We compare our objective to theirs in Section 4.3.

3. METHOD

3.1. Text-Video Contrastive Loss

We handle the problem of learning multilingual text-video representations. For simplicity, we first describe the approach for learning with English text and then explain how to extend it to more languages. We consider a dataset $D_{en} = \{(t_i, v_i)\}_{i=1}^{N}$ of paired videos and English captions. The goal of text-video retrieval is to learn text and vision models, $f(\cdot)$ and $g(\cdot)$ respectively, which output embeddings that are similar to each other when the input text caption t_i and video v_i are semantically related (ie. describing similar concepts), and have low similarity when they are unrelated. In this work, we use cosine similarity by L2-normalizing the outputs of $f(\cdot)$ and $g(\cdot)$ and taking the dot-product.

The Noise-Contrastive Estimation loss (NCE) [13, 14, 15] has been commonly used to learn text-video representations [16, 17]. Given a batch of B text-video pairs, let $\mathbf{S}$ be the text-video similarity matrix, with $\mathbf{S}_{ij} = f(t_i)^\top g(v_j)$. With temperature τ, the NCE loss is given as:

$$\mathcal{L}_{NCE} = -\sum_{i=1}^{B} \log \frac{\exp(\mathbf{S}_{ii}/\tau)}{\sum_{k=1}^{B} \exp(\mathbf{S}_{ik}/\tau)}. \tag{1}$$

This can be interpreted as the cross entropy loss between the distribution over normalized text-video scores in $\mathbf{S}$ and the one-hot distribution. Specifically, let $Q_{t_i}(v_j)$ be the probability that video v_j matches with text t_i :

$$Q_{t_i}(v_j) = \frac{\exp(\mathbf{S}_{ij}/\tau)}{\sum_{k=1}^{B} \exp(\mathbf{S}_{ik}/\tau)}. \tag{2}$$

The target distribution, $P_{t_i}(v_j)$, is one-hot (since the correct match for text t_i is video v_i):

$$P_{t_i}(v_j) = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Given the equation for cross entropy,

$$\mathcal{L}_{CE} = -\sum_{i=1}^{B} \sum_{j} P_{t_i}(v_j) \log Q_{t_i}(v_j), \tag{4}$$

we can see immediately that Eq. 1 is equivalent to Eq. 4. In Section 3.3, we introduce an additional cross entropy based objective between a new target distribution $P'_{t_i}(v_j)$ and $Q_{t_i}(v_j)$.

To extend this to a dataset of videos paired with captions in L languages, ie. $D_{multi} = \{(t_i^1, t_i^2, \ldots, t_i^L, v_i)\}_{i=1}^{N}$, we compute a text-video similarity matrix for each language, ie. $\mathbf{S}^l$, where $\mathbf{S}_{ij}^l = f(t_i^l)^\top g(v_j)$. Then we apply $\mathcal{L}_{NCE}$ to each matrix and take the sum of the losses. This pulls together the embeddings of videos and their paired captions in different languages.

During inference, f and g are used to encode text and video inputs. For a given text query, videos are ranked by their cosine similarity to the text.

3.2. C2KD Method

Although $\mathcal{L}_{NCE}$ can be used to learn multilingual text-video representations, the performance for English text-video retrieval is usually higher than for other languages. This implies that the English text-video scores are most accurate. Our key idea is to use the English text-video scores to improve the scores for other languages.

The method is illustrated in Figure 1. We first train M teacher models using D_{multi} and $\mathcal{L}_{NCE}$, and then freeze their parameters. The teacher models have the same architecture, except the text encoders are different so that complementary information from different models can be used. Next, we begin training a student model with D_{multi} and $\mathcal{L}_{NCE}$. For each batch of video and multilingual text, the teachers are simultaneously provided with the video and English translations as input. Each teacher produces an English text-video similarity matrix. We apply a pooler function $\Psi : \mathbb{R}^{M\times B\times B} \rightarrow \mathbb{R}^{B\times B}$ to the M teacher similarity matrices to get a single similarity matrix $\mathbf{S}'$, where $\mathbf{S}'_{ij}$ is the similarity score at row i and column j. In our experiments, we experimented with different pooler functions such as mean, max, and min. We train the student with $\mathcal{L}_{NCE}$ and a 2nd objective, $\mathcal{L}_{C2KD}$ (introduced in Section 3.3), which encourages the student's text-video scores from captions in different languages to be similar to the teacher English text-video scores in $\mathbf{S}'$. Note that only the student model is used during inference.

3.3. Knowledge Distillation Objective

We introduce a distillation objective that encourages the student's multilingual text-video scores to be similar to the teacher English text-video scores in $\mathbf{S}'$. The main idea is that instead of using the one-hot distribution $P_{t_i}(v_j)$ in $\mathcal{L}_{NCE}$, we use a new distribution $P'_{t_i}(v_j)$ obtained from the teacher English text-video scores in $\mathbf{S}'$. Specifically, let $P'_{t_i}(v_j)$ be the probability that video v_j matches with text t_i :

$$P'_{t_i}(v_j) = \frac{\exp(\mathbf{S}'_{ij}/\tau)}{\sum_{k=1}^{B} \exp(\mathbf{S}'_{ik}/\tau)} \tag{5}$$

We apply the cross entropy loss between $P'_{t_i}(v_j)$ (generated by the teacher English text-video scores) and $Q_{t_i}(v_j)$ (generated by the student multilingual text-video scores):

$$\mathcal{L}_{C2KD} = -\sum_{i=1}^{B} \sum_{j} P'_{t_i}(v_j) \log Q_{t_i}(v_j), \tag{6}$$

Note that the temperature τ in $\mathcal{L}_{C2KD}$ is controlled independently of the one in $\mathcal{L}_{NCE}$. We apply $\mathcal{L}_{C2KD}$ to each of the student text-video similarity matrices using text in different languages and take the sum of the losses. The final objective is given by:

$$\mathcal{L} = \alpha \mathcal{L}_{NCE} + (1 - \alpha) \mathcal{L}_{C2KD} \tag{7}$$

where α is a balance hyperparameter.

The difference between $\mathcal{L}_{NCE}$ and $\mathcal{L}_{C2KD}$ is the target distribution; the former uses a one-hot distribution while the latter uses soft-labels produced by the teachers. $\mathcal{L}_{NCE}$ makes rigid assumptions about which captions are similar to which video clips (only paired examples should match), whereas $\mathcal{L}_{C2KD}$ enables the model to have leeway in assigning higher scores to pairs which are not ground-truth pairs, but still have some semantic similarity. Also, $\mathcal{L}_{C2KD}$ shares the same cross entropy objective as the original KD [9], but it is more technically advanced since it distills teacher cross-modal matrices instead of just the logits from uni-modal encoders. Further, our cross-modal distillation is consistent with the retrieval task.

4. EXPERIMENTS

4.1. Datasets

Multi-MSRVTT [2] is a multilingual version of the MSRVTT [18] video dataset. The video categories are general, such as "sports" and "vehicles." The original dataset contains 10k videos from YouTube, each annotated with 20 captions in English. The captions were translated to 8 other languages with machine translation. We followed the setup in prior work [2] and used a training set of 6.5k videos, validation set of 497 videos, and test set of 1k videos.

Multi-YouCook2 is our multilingual extension of the YouCook2 [7] video dataset. The original dataset contains 2k cooking videos from YouTube. The video categories are about recipes, such as "spaghetti and meatballs." Each video was segmented into smaller clips containing recipe steps and annotated with text captions of the recipe steps in English. Inspired by the procedure to collect Multi-MSRVTT [2], we translated the captions to 8 other languages using machine translation. Following the setup in prior English text-video work [19], we used 9,586 training clips and 3,350 evaluation clips.

VATEX [6] contains videos each with 10 English and 10 Chinese captions. The videos were selected from an action classification dataset [20]. Following prior work [2], we use the official training set of 26k videos and split the validation set equally into 1.5k validation and 1.5k test videos. Note that we made our own split since theirs was not released, and we will release our split.

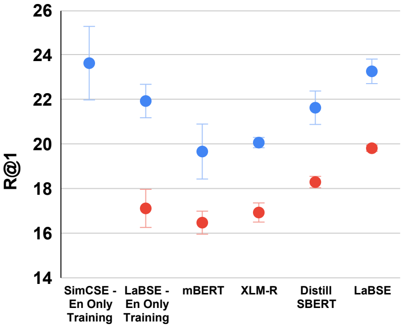

Fig. 2. Performance on MSRVTT with different students. Blue dots: English-only performance, Red dots: multilingual performance.

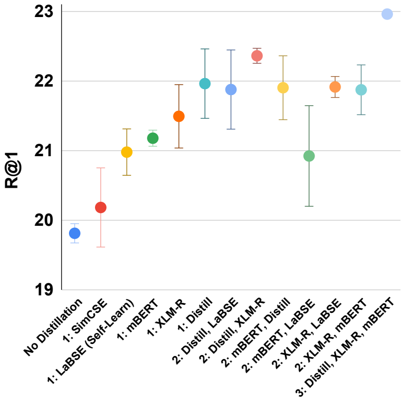

Fig. 3. Ablation on the number of teachers.

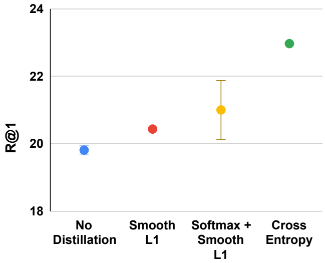

Fig. 4. Ablation on the KD objective.

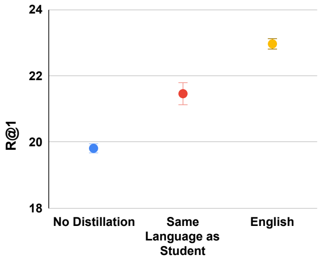

Fig. 5. Ablation on the language used with the teachers.

4.2. Implementation Details and Experimental Setup

For the student text model f, we use LaBSE [8]. We discuss the teacher text models in Section 4.3. For the video model g, we first extract features from CLIP ViT-B/32 [21] at 1 FPS and process them with a 2-layer Transformer [22]. Due to GPU memory limitations, we do not update the weights of the CLIP model. We set τ in $\mathcal{L}_{NCE}$ to 0.05 and τ in $\mathcal{L}_{C2KD}$ to 0.1. We found the best pooler function Ψ and balance α to be different for each dataset (shown in Table 1). We trained the models for 20 epochs for MSR-VTT, 10 epochs for Multi-YouCook2 and 30 epochs for VATEX. The batch size was 64 videos. The initial learning rate was 10^{-4} with an exponential decay of 0.9. We use the standard R@K metrics (recall at rank K, higher is better). All of our reported results are the average of three runs. In the zero-shot setting, models are trained on English text-video pairs only and evaluated using captions in all languages. In the translate-train setting, the models are trained on text-video pairs in all languages. Note that C2KD is only applicable to the translate-train setting since it requires multilingual text during training.

Parameter	MSRVTT	YouCook2	VATEX
α (Balance)	0.5	0.1	0.1
Ψ (Pooler)	Min	Min	Max

Table 1. Balance and pooler hyperparameters.

Model	Set	*en*	*de*	*fr*	*cs*	*zh*	*ru*	*vi*	*sw*	*es*	Avg↑
Rand.	-	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
[3]†	Z	17.7	15.1	14.8	13.0	11.6	12.6	7.1	4.9	15.6	12.5
[2]	Z	**23.8**	**19.4**	**20.7**	**19.3**	**18.2**	**19.1**	8.2	8.4	20.4	**17.5**
NCE	Z	21.9	18.9	18.7	18.2	16.3	17.5	**9.1**	**12.8**	**20.5**	17.1
[3]†	T	17.0	17.0	17.2	16.1	14.6	16.0	8.6	11.5	16.8	15.0
[2]	T	23.1	21.1	21.8	20.7	20.0	**20.5**	10.9	14.4	**21.9**	19.4
NCE	T	**23.3**	**21.1**	**22.3**	**20.9**	**20.3**	19.6	**12.1**	**17.2**	21.5	**19.8**
C2KD	T	**26.4**	**24.7**	**25.4**	**24.0**	**23.4**	**23.1**	**13.6**	**20.3**	**25.5**	**23.0**

Table 2. **Text-video retrieval on Multi-MSRVTT (R@1).** †: our implementation, Set=Setting, Z=Zero-Shot, T=Translate-Train.

Model	Set	*en*	*de*	*fr*	*cs*	*zh*	*ru*	*vi*	*ja*	*es*	Avg↑
Rand.	-	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
[3]†	Z	10.1	2.5	2.7	2.1	1.4	1.6	2.2	1.2	2.3	2.9
[2]†	Z	12.7	3.7	3.3	2.7	2.0	2.5	2.3	1.8	2.4	3.7
NCE	Z	**14.4**	**7.0**	**6.4**	**5.1**	**3.5**	**4.7**	**5.0**	**2.7**	**6.3**	**6.1**
[3]†	T	10.0	9.1	9.1	8.6	6.7	9.0	6.3	7.5	9.1	8.4
[2]†	T	11.3	10.4	10.6	10.1	8.3	9.3	8.4	9.1	10.4	9.8
NCE	T	**14.9**	**13.1**	**13.0**	**12.1**	**9.6**	**12.1**	**10.9**	**10.0**	**13.2**	**12.1**
C2KD	T	**15.5**	**14.0**	**13.9**	**12.8**	**10.4**	**13.1**	**11.4**	**11.3**	**14.1**	**12.9**

Table 3. **Text-video retrieval on Multi-YouCook2 (R@1).** †: our implementation, Set=Setting, Z=Zero-Shot, T=Translate-Train.

		English (*en*)			Chinese (*zh*)		
Model	Set	R@1	R@5	R10	R@1	R@5	R@10
Rand.	-	0.07	0.33	0.67	0.07	0.33	0.67
[2]	Z	**44.4**	80.5	88.7	29.7	63.2	75.5
[2]	T	44.3	**80.7**	**88.9**	**40.5**	**76.4**	**85.9**
[3]†	Z	37.7	77.0	87.7	25.7	57.3	72.5
[2]†	Z	39.9	79.1	89.3	26.9	60.4	75.3
NCE	Z	**42.0**	**81.0**	**90.6**	**28.0**	**63.4**	**75.6**
[3]†	T	37.5	77.1	88.2	33.2	70.9	83.9
[2]†	T	41.3	78.9	88.8	34.1	74.3	85.2
NCE	T	**42.6**	**81.0**	**90.6**	**38.0**	**75.4**	**88.0**
C2KD	T	**43.1**	**82.1**	**91.5**	**39.6**	**77.0**	**88.6**

Table 4. **Multilingual text-video retrieval on VATEX.** Upper and lower halves separated due to different test splits. †: our implementation, Set=Setting, Z=Zero-Shot, T=Translate-Train.

4.3. Ablation Studies and Analysis

We conduct an analysis on Multi-MSRVTT to justify our design choices. The bars in the figures are the standard deviation of three runs.

Text encoders. We compare text encoders in Figure 2 when trained for text-video retrieval ($\mathcal{L}_{NCE}$ only, α=1). LaBSE and Distill SBERT outperformed mBERT and XLM-R, which are not trained with sentence level objectives. When trained with multilingual captions, LaBSE's performance on English is comparable to SimCSE's, a recent English-only sentence embedding model [23]. Finally, LaBSE's performance across all languages, including English, improved when trained on multilingual captions. Since LaBSE is the strongest multilingual model, we use it as our student text encoder.

Teacher models. In Figure 3, we show the performance of our method with different teachers. With one teacher, we found that SimCSE was the worst teacher, which was surprising considering its strong English-only performance. Using LaBSE as its own teacher is feasible, but it is better to use a different model as the teacher. Distill SBERT was the best teacher, which is reasonable considering it is the most similar to LaBSE in using a sentence-level objective. With two teachers, any combination of Distill SBERT, mBERT, and XLM-R improved performance over any individual teacher. Given these results, we used Distill SBERT, mBERT, and XLM-R as the final set of teachers, which obtained the best results.

Knowledge Distillation objective. We compare distillation objectives in Figure 4. For English text-video retrieval, TeachText [12] proposed to regress the teacher text-video scores using a Smooth L1 Loss. We found it only gave a minor improvement over the baseline without distillation. While the TeachText approach considers each text-video score independently, our proposed $\mathcal{L}_{C2KD}$ loss instead considers the context of all the text-video scores by normalizing them with softmax and applying the cross entropy loss. This significantly outperforms the regression based method. We also tried an intermediate approach of combining softmax normalization and Smooth L1 Loss, which performed only slightly better than Smooth L1 loss. This shows that it is essential to use the distribution over the text-video scores instead of treating them independently.

Teacher language. In Figure 5, we compare the results when different languages are used by the teachers. Using the same multilingual text as input to the student and teachers improves the results over no distillation, likely due to the complementary information provided by different text encoders. However, our proposed method of using English with the teachers performs better. This result matches our intuition that English should be the best language to use with the teachers since English text-video retrieval is highest.

4.4. Main Results

We tested C2KD on three datasets with the best student (LaBSE) and teacher models (Distill SBERT, mBERT, and XLM-R). We also implemented the baselines [2, 3] since their code was not released. The "NCE" method corresponds to our baseline without distillation ($\mathcal{L}_{NCE}$ only, α=1). We applied C2KD to this method.

Table 2 shows the multilingual text-video retrieval results on Multi-MSRVTT. C2KD improves performance across languages, with average R@1 improving from 19.8 to 23.0 (+16.2% relative). The largest improvement is on Spanish (*es*), from 21.5 to 25.5 (+18.6% relative).

Table 3 shows the results on Multi-YouCook2. Applying our C2KD method to the baseline, we see improvements for all languages. The average R@1 improves from 12.1 to 12.9 (+6.6% relative). The largest improvement is on Japanese, from 10.1 to 11.3 (+11.9% relative).

Table 4 shows the results on VATEX. The retrieval performance is generally higher than on the other datasets, which could be attributed to the large training set. Nonetheless, C2KD can improve the performance for both English and Chinese in all metrics. Chinese R@1 is improved from 38 to 39.6 (+4.2% relative).

Overall, C2KD consistently improved performance across languages and domains, with significant improvements on some languages. Also, our results accurately represent the performance since we ran each experiment three times and report the average.

5. CONCLUSION

We introduce Cross-Lingual Cross-Modal Knowledge Distillation (C2KD) to improve multilingual text-video retrieval performance. Our method trains a student using input multilingual text to output similar text-video similarity scores compared with teachers using input English text. We obtained an improvement in multilingual text-video retrieval across languages and domains. Finally, we introduce the Multi-YouCook2 dataset with captions in 9 languages and will make the data public to spur more research in this direction. Ideas for future work include applying multilingual text augmentation and paraphrasing strategies to generate more data.

6. REFERENCES

[1] Albert Reuther, Jeremy Kepner, Chansup Byun, Siddharth Samsi, William Arcand, David Bestor, Bill Bergeron, Vijay Gadepally, Michael Houle, Matthew Hubbell, Michael Jones, Anna Klein, Lauren Milechin, Julia Mullen, Andrew Prout, Antonio Rosa, Charles Yee, and Peter Michaleas, "Interactive supercomputing on 40,000 cores for machine learning and data analysis," in *2018 IEEE High Performance extreme Computing Conference (HPEC)*. IEEE, 2018, pp. 1–6.

[2] Po-Yao Huang, Mandela Patrick, Junjie Hu, Graham Neubig, Florian Metze, and Alexander G Hauptmann, "Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models," in *2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 2443–2459.

[3] Jayaprakash Akula, Rishabh Dabral, Preethi Jyothi, and Ganesh Ramakrishnan, "Cross lingual video and text retrieval: A new benchmark dataset and algorithm," in *2021 International Conference on Multimodal Interaction*, 2021, pp. 595–603.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, June 2019, pp. 4171–4186, Association for Computational Linguistics.

[5] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020, pp. 8440–8451, Association for Computational Linguistics.

[6] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang, "Vatex: A large-scale, high-quality multilingual dataset for video-and-language research," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4581–4591.

[7] Luowei Zhou, Chenliang Xu, and Jason J Corso, "Towards automatic learning of procedures from web instructional videos," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[8] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang, "Language-agnostic BERT sentence embedding," in *60th Annual Meeting of the Association for Computational Linguistics*, Dublin, Ireland, May 2022, pp. 878–891, Association for Computational Linguistics.

[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *NIPS Deep Learning Workshop*, vol. 2, no. 7, 2014.

[10] Nils Reimers and Iryna Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, Nov. 2020, pp. 4512–4525, Association for Computational Linguistics.

[11] Nils Reimers and Iryna Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, Nov. 2019, pp. 3982–3992, Association for Computational Linguistics.

[12] Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu, "Teachtext: Crossmodal generalized distillation for text-video retrieval," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11583–11593.

[13] Michael Gutmann and Aapo Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 297–304.

[14] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu, "Exploring the limits of language modeling," *arXiv preprint arXiv:1602.02410*, 2016.

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

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

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

[18] Jun Xu, Tao Mei, Ting Yao, and Yong Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *IEEE conference on computer vision and pattern recognition*, 2016, pp. 5288–5296.

[19] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2630–2640.

[20] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al., "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[23] Tianyu Gao, Xingcheng Yao, and Danqi Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic, Nov. 2021, pp. 6894–6910, Association for Computational Linguistics.