Assignment: Critical Review
CRET: Cross-Modal Retrieval Transformer for Efficient Text-Video Retrieval

Siwen Tu (s3631400) [1] **Shupei Li (s3430863)** [1]

1. Motivation

Information can be stored and delivered through various carriers. For example, visual information is usually shown by images and acoustic information is usually saved in audios. In machine learning area, a modality refers to the data collected via the same information channel. Because of the diversity of information carriers, searching for information involves interactions between different modalities in general. During the class, we mainly focus on classical models whose input and output are both text. They have achieved satisfactory performance on document retrieval, however, can't be easily generalized to multimodality information retrieval tasks.

To process information from different modalities, we need to design a model that can represent and fuse data efficiently. Here we only consider two modalities — text and videos, due to the wide application of the text-to-video retrieval operation. The task is formulated as follows. Given queries with the textual content, we are supposed to find relevent videos in databases. There are three challenges to complete this task: data representation, modality infusion, and feature alignment. Ji et al. propose a novel method called CRET in (Ji et al., 2022), which successfully tackles challenges and achieves the state-of-the-art performance on text-to-video retrieval task. Next, we will introduce the CRET method in detail.

2. Summary

The embedding-based (EDB) methods and the model-based (MDB) methods are two popular methods in tackling the text-to-video retrieval tasks. The EDB method utilizes the embeddings extracted from the text and video separately. While the EDB method suffers from the loss of accuracy which results from the loss of correspondence details, it shows the significant advantage in terms of efficiency. In contrast, the MDB method superiors the EDB in retrieval accuracy but suffers from low efficiency since it iterates all the text-video pairs to evaluate the distance without extracting explicit embeddings.

Authors of this paper propose a novel EDB method named cross-modal retrieval transformer (CRET), which solves the conflict between efficiency and accuracy. The main contribution of this novel method consists of two parts. One is Cross-modal Correspondence Modeling (CCM) module and the other is Gaussian Estimation of Embedding Space (GEES) loss.

The proposed CRET model encodes the video and text separately. The text encoder applies the BERT as the base model to encode features into global and local features. Meanwhile, the video encoder consists of the spatial and temporal encoders which extract the spatial and temporal features separately. The spatial encoder (CaiT-S/24) encodes all features into global and local features. Then the model feeds the global features into a temporal encoder to get the temporal embeddings. On the other hand, the local feature embeddings are projected to the same dimension as the text embeddings.

The CCM module plays an important role in smoothing the gap between the domains of video and text modalities. The principle behind CCM is multi-head self-attention mechanism. Authors align features using the transformer in which the text and video encoders share the same weights. In detail, they first calculate the distance between the token features and the query center, then regard the distance as the weights of this feature. Afterwards, they concatenate and project the aligned features from the multi-head module, and calculate the similarity score of these aligned features from text and video modalities.

The GEES samples the video frame embeddings densely to reduce the computing costs. The traditional NCE loss function requires a trade-off between the accuracy and the computational burden. Authors improve this loss function by introducing the multivariate Gaussian assumption about the frame distribution and approximating the loss function with the upper bound, which simplifies the calculation process. The improvement enhances the optimization efficiency of the SGD algorithm during the training.

[1]LIACS, Leiden University, Leiden, Netherlands. Correspondence to: Siwen Tu <s3631400@vuw.leidenuniv.nl>, Shupei Li <s3430863@vuw.leidenuniv.nl>.

Information Retrieval, Master CS, Spring 2023, Leiden, the Netherlands, IR '23.

The model considers the GEES loss and CCM loss together with a balancing parameter during the training stage. In the inference stage, the model adopts a linear regression model with an adjustable hyperparameter to infer the similarity score, whose inputs are global and local features.

Authors conduct experiments on four benchmark data sets to verify the effectiveness of CRET method. They apply muiltiple metrics: R@1, R@5, R@10, and MdR. Experimental results show that the CERT model outperforms other baselines on all data sets, even if some of these models are pre-trained on other large video data sets. Authors also conduct ablation studies to validate the effects of CCM and GEES loss.

At the end of the paper, authors test the hypothesis in which the GEES loss is based on and analyze the quality of model according to experiments.

3. Assessment

This paper clearly illustrates the principle of CRET method. And its structure is consistent with requirements of the scientific paper. First, authors review the previous works and summarize their contributions. Then, they make necessary assumptions to simplify the problem and describe CRET method in detail. It is worth mentioning that the validity of assumptions is proved by empirical experiments. Authors compare CRET and benchmarks on four data sets. They also conduct ablation studies. The process and results of experiments are documented in the paper, which support the effectiveness of CRET.

We examine the reproducibility of the paper from three aspects: source code, the availability of data sets, and experimental settings. It is unfortunate that authors don't make their code publicly available. Besides, authors affiliate with a financial services company, which increases the difficulty of accessing the source code. In the experimental setup section, authors introduce four datasets (MSRVTT, LSMDC, MSVD, DiDeMo) and implementation details. We check the accessibility of data sets one by one. LSMDC and DiDeMo can be downloaded from links listed in references. Official websites of MSRVTT and MSVD are no longer available. However, they can be downloaded from the open source community alternatively. As for experimental settings, authors have reported main parameter settings in the paper. But it is hard to use exactly the same experimental settings as the paper due to the inaccessibility of the source code.

The main contribution of this paper is the design of CRET model that achieves the state-of-the-art performance on text-to-video retrieval. CRET successfully deals with three common challenges in text-to-video retrieval operations. Specifically, CRET utilizes CaiT-S/24 to encode the video data and BERT to encode the text data. And the CCM module in CRET solves the problem of modality fusion as well as feature alignment. The paper points out that CRET can be extended to more methods and scenarios (Ji et al., 2022). From the perspective of application, we think the CRET method has broad prospect in video-streaming websites and search engines.

Finally, we will discuss strong and weak points of this paper. One major advantage of this paper is fully considering the algorithm efficiency, while maintaining the satisfactory model performance. In the model design stage, authors introduce the multivariate Gaussian distribution assumption to approximate the loss function, which largely reduce the computing costs. In the experiment stage, authors accelerate the model training process by adopting pre-trained models, setting hyperparameters carefully, and distributing the training on several servers. Compared to baseline models, the CRET method doesn't need the extra pre-training on other data sets, which enhances its competitiveness and practicability. Another strong point of the paper is that authors conduct extensive ablation studies and hypothesis tests, which proves the validity of each part of the model. However, this paper also has some shortcomings that can be improved. First, the inaccessibility of the code has a great impact on reproducibility. Second, authors don't report the standard variance of experimental results and don't clearly state the number of times for running each model. Therefore, it is hard to assess the stability of CRET based on tables provided in the paper.

References

Ji, K., Liu, J., Hong, W., Zhong, L., Wang, J., Chen, J., and Chu, W. Cret: Cross-modal retrieval transformer for efficient text-video retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pp. 949–959, New York, NY, USA, 2022. Association for Computing Machinery.