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

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

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.

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 sblation studies. The process and results of experiments are documented in the paper, which support the effectiveness of CRET.

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

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 as a result of 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 of 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.