CRET: Cross-Modal Retrieval Transformer for Efficient Text-Video Retrieval

Angela Rossi

Table of Contents

- Introduction (text-to-video retrieval, EDB and MDB methods, goal of the paper)
- CRET Framework overview (architecture and components)
- Gaussian Estimation of Embedding Space (GEES) Loss
- Cross-modal Correspondence Modeling (CCM) Module
- Experimental Evaluation
- Conclusion and Future Work

Introduction

- What's CRET?

  Short for "Cross-Modal Retrieval Transformer", is a novel framework proposed in the research paper "CRET: Cross-Modal Retrieval Transformer for Efficient Text-Video Retrieval". The primary goal of the CRET framework is to effectively model the local correspondences between video and text modalities, while maintaining efficient retrieval performance.

Introduction

Bibliography

Ji K., Liu J., Hong W., Zhong L., Wang J., Chen J., Chu W,
CRET: Cross-Modal Retrieval Transformer for Efficient Text-Video Retrieval

link to the paper

Introduction

- What's the goal of this paper?

  The main goal of this paper was to propose a new embedding-based framework called CRET (Cross-Modal Retrieval Transformer) that can effectively model the local correspondences between video and text modalities, while also maintaining efficient retrieval performance, using an Embedding Based Method.

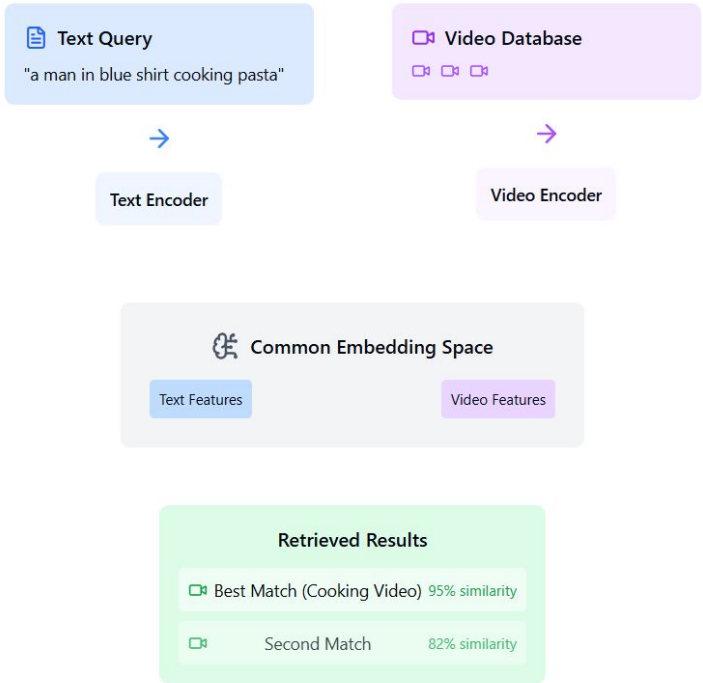
Introduction

- What's text-to-video retrieval?

  - Task of finding and retrieving relevant videos from a database based on a text query
  - Enables searching video content using natural language descriptions
  - Cross-modal task: bridges between text and visual data

Introduction

- What's text-to-video retrieval?



Introduction

- For example:

(b) The words or phrases in the caption are connected to the corresponding local visual parts in the video frames.

Introduction

KEY POINTS ABOUT TEXT-TO-VIDEO RETRIEVAL

- Cross-modal task = it involves bridging the gap between textual and visual data.
- It has to find videos that are relevant.
- Enables users to search large videos databases (such as YouTube).
- Has to match abstract concepts with visual features.
- It has to support real-time retrieval.

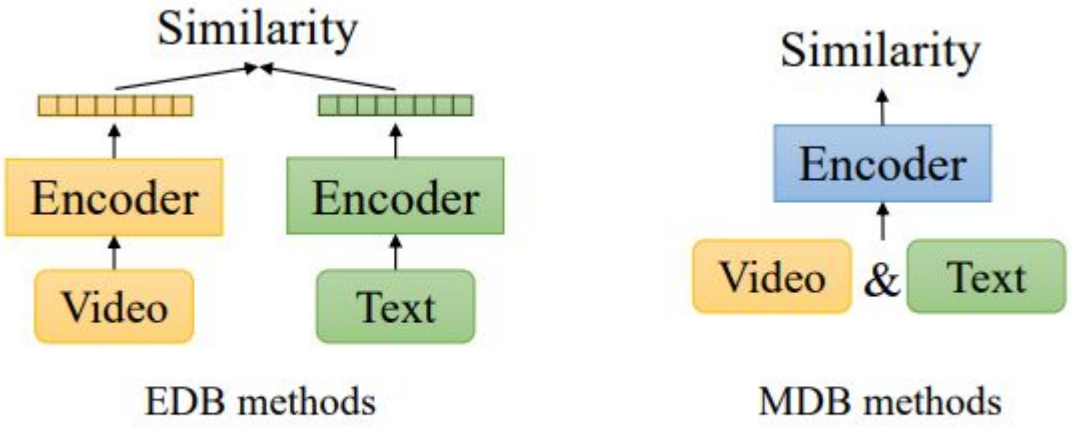
Introduction

- What's CRET?
- There are two methods: Embedding Based (EDB) and Model Based (MDB), the second one is more efficient. The authors of the paper chose EDB.

- Cross-modal Correspondence Modeling (CCM) Module: uses transformer decoders and shared center queries to align the local features of video and text modalities.

- Gaussian Estimation of Embedding Space (GEES) Loss: A technique that treats video frames as following a Gaussian distribution, enabling dense frame sampling benefits while only processing a few frames.

Introduction

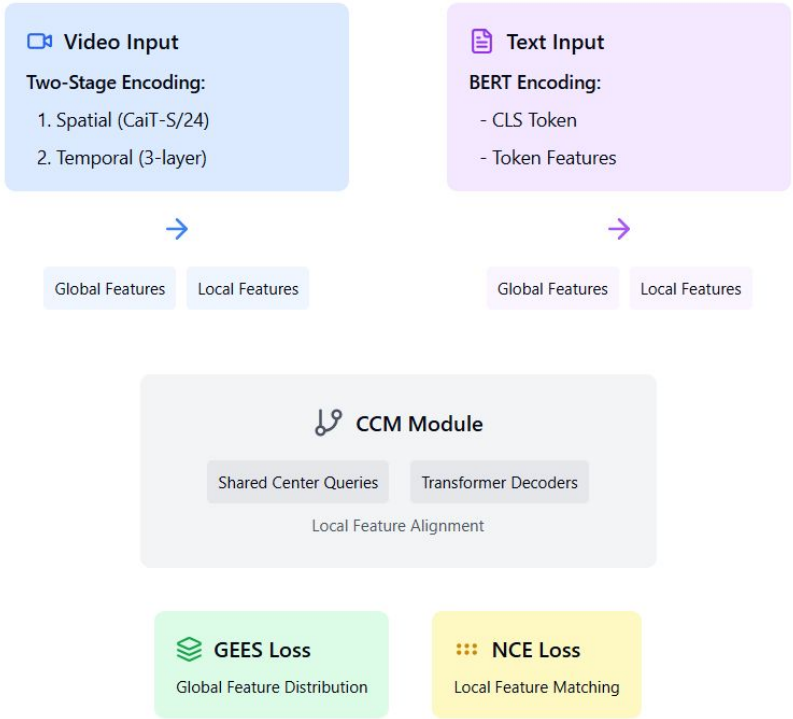


(a) EDB and MDB methods mainly differ at whether explicit embeddings of text/video are generated.

Table of Contents

- Introduction (text-to-video retrieval, EDB and MDB methods, goal of the paper)
- CRET Framework overview (architecture and components)
- Gaussian Estimation of Embedding Space (GEES) Loss
- Cross-modal Correspondence Modeling (CCM) Module
- Experimental Evaluation
- Conclusion and Future Work

CRET Framework Overview



CRET Framework Overview

Key Components

VIDEO ENCODER

It is composed of two stages:

- Spatial Encoder: A CaiT-S/24 transformer model encodes individual video frames, extracting both global and local features.

- Temporal Encoder: A 3-layer transformer model takes the frame-level features and fuses them to produce the final video embeddings.

CRET Framework Overview

Key Components

TEXT ENCODER

- The text encoder uses a pre-trained BERT-base-uncased model to encode the input text captions.

- About the captions: the videos have them because of the datasets used (MSRVTT has about 20 captions per video).

- Similar to the video encoder, the text encoder outputs both global and local text features.

CRET Framework Overview

Key Components

TEXT ENCODER

WHAT'S BERT?

BERT (Bidirectional Encoder Representations from Transformers) language model developed by researchers at Google. It is a transformer-based model that is trained on a large corpus of text data in an unsupervised manner.

CRET Framework Overview

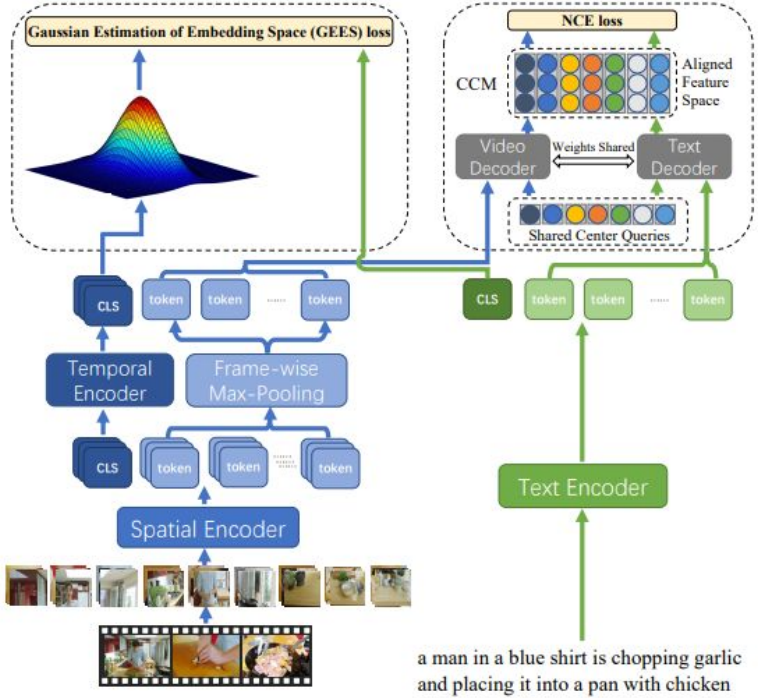


Table of Contents

- Introduction (text-to-video retrieval, EDB and MDB methods, goal of the paper)
- CRET Framework overview (architecture and components)
- Gaussian Estimation of Embedding Space (GEES) Loss
- Cross-modal Correspondence Modeling (CCM) Module
- Experimental Evaluation
- Conclusion and Future Work

Gaussian Estimation of Embedding Space (GEES) Loss

- The key challenge in text-to-video retrieval is balancing the information loss and computational cost when sampling frames from a video.

- The vanilla Noise Contrastive Estimation (NCE) loss, commonly used in these tasks, is highly dependent on the sampling frequency.

Gaussian Estimation of Embedding Space (GEES) Loss

- Why is it called “vanilla” Noise Contrastive Estimation (NCE) loss?

The term "vanilla" is used to refer to the original or basic version of the Noise Contrastive Estimation (NCE) loss, in contrast to the more specialized GEES loss proposed in the CRET paper. The vanilla NCE loss is a widely used loss function in various machine learning tasks, including text-to-video retrieval. It aims to discriminate between positive (matched) text-video pairs and artificially generated negative pairs.

Gaussian Estimation of Embedding Space (GEES) Loss

GAUSSIAN ASSUMPTION

ADVANTAGES OF GEES LOSS

- The GEES loss allows the model to effectively leverage information from all the video frames, even with a sparse sampling strategy, without incurring a significant computational cost.

- This is in contrast to the vanilla NCE loss, which would require dense sampling to achieve comparable performance.

Gaussian Estimation of Embedding Space (GEES) Loss

GAUSSIAN ASSUMPTION

VALIDATION

Three-step validation process to support the Gaussian assumption:

- Theoretical Grounding
- Statistical Testing (Henze-Zirkler test)
- Binomial Distribution Analysis

Gaussian Estimation of Embedding Space (GEES) Loss

Theoretical Grounding:

- Showed that Gaussian random vectors remain Gaussian after linear transformations, providing mathematical justification for the assumption.

- IN THE PICTURE: The four subfigures (a)-(d) show the estimated probability density of the video frame features, illustrating the authors' assumption that the video frame features follow a multivariate Gaussian distribution.

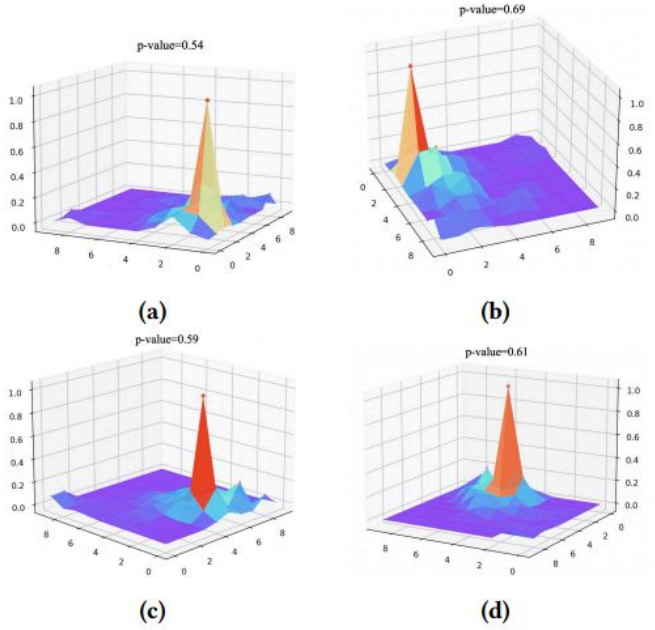
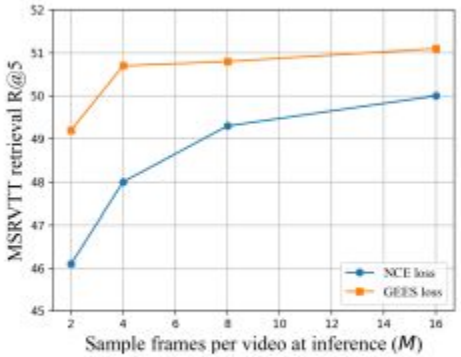
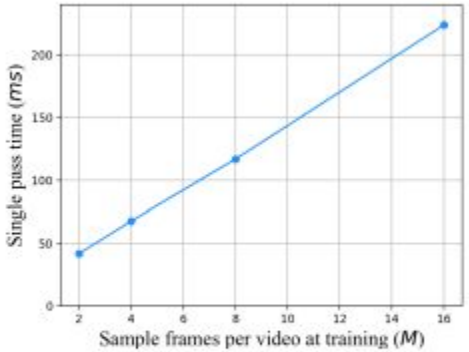


Figure 4: Probability density of video frames features. We utilize the histogram to estimate the probability density. Z-axis represents the probability density.

Gaussian Estimation of Embedding Space (GEES) Loss



(a) The effectiveness of GEES loss.



(b) Time cost in the training stage.

Table of Contents

- Introduction (text-to-video retrieval, EDB and MDB methods, goal of the paper)
- CRET Framework overview (architecture and components)
- Gaussian Estimation of Embedding Space (GEES) Loss
- Cross-modal Correspondence Modeling (CCM) Module
- Experimental Evaluation
- Conclusion and Future Work

Cross-modal Correspondence Modeling (CCM) Module

The CCM (Cross-modal Correspondence Modeling) module is a core component of CRET that aligns local features between video and text. It uses transformer decoders and shared center queries to match specific video elements (like objects or actions) with corresponding words in the text.

Note: A key difference from other methods is that CCM can process video/text independently and still capture local correspondences, making it both effective and efficient.

Table of Contents

- Introduction (text-to-video retrieval, EDB and MDB methods, goal of the paper)
- CRET Framework overview (architecture and components)
- Gaussian Estimation of Embedding Space (GEES) Loss
- Cross-modal Correspondence Modeling (CCM) Module
- Experimental Evaluation
- Conclusion and Future Work

Experimental Evaluation

Datasets:

- **MSRVTT**: A general video dataset with 10,000 videos and 200,000 captions.
- **LSMDC**: A video dataset extracted from 202 movies, with 118,081 videos.
- **MSVD**: A dataset with 1,970 videos, each with 1-62 seconds in length
- **DiDeMo**: A dataset of 10,000 Flickr videos annotated with 40,000 sentences

Experimental Evaluation

Evaluation Metrics:

R@K (Recall at rank K):

- Percentage of test samples where the correct result is in the top-K retrieved items.
- Measures percentage of test samples where the correct video is in top-K retrieved results
- Higher is better
- Paper reports R@1, R@5, and R@10
- Example: R@5 = 50% means for 50% of queries, the correct video was among top 5 results

Experimental Evaluation

Evaluation Metrics:

MdR (Median Rank):

- Median rank of the correct item in the retrieved ranking

- Lower is better

- Example: MdR = 6 means on average, the correct video appears at position 6 in results

Experimental Evaluation

	Weight Initialization		E2E	Type	R@1↑	R@5↑	R@10↑	MdR↓
	Visual	Textual						
MIL-NCE [36]	K+H	G+H	✓	EDB	9.9	24.0	32.4	29.5
JSFusion [54]	I	N.A.	✓	MDB	10.2	31.2	43.2	13.0
HT [38]	I	G	✓	EDB	12.4	36.0	52.0	10.0
HT [38]	I+H	G+H	✓	EDB	14.9	40.2	52.8	9.0
ActBERT [59]	I+H	B+H		MDB	16.3	42.8	56.9	10.0
HiT(appearance-only) [30]	I+H	B+H	✓	EDB	18.2	41.9	55.5	5.0
TACo(R-152) [53]	I+H	B+H	✓	MDB	18.9	46.2	58.8	7.0
UniVL(FT-Joint) [33]	K+H	B+H		EDB	20.6	49.1	62.9	6.0
UniVL(FT-Align) [33]	K+H	B+H		MDB	21.2	49.6	63.1	6.0
ClipBERT [28]	I+C+V	C+V	✓	MDB	22.0	46.8	59.9	6.0
Ours	I	B	✓	EDB	**23.9**	**50.8**	**63.4**	**5.0**

Table 1: Results of text-to-video retrieval on MSRVTT 1K test set, and we follow the split from ClipBERT. The two columns of "Weight Initialization" list the datasets that are exploited to pre-train visual/textual backbones before training on the MSRVTT dataset, where I, K, G, C, V, B, H stand for ImageNet [43], Kinetics [25], GoogleNews [39], COCO Captions [5], Visual Genome Captions [27], BERT-base [48] and HowTo100M [38], respectively. The column "E2E" indicates whether a method can be trained in an end-to-end manner. The column "Type" distinguishes model-based methods (MDB) and embedding-based methods (EDB).

Experimental Evaluation

	shared queries	shared weights	R@1↑	R@5↑	R@10↑	MdR↓
CCM			21.6	47.1	57.5	6.0
		✓	22.5	49.5	60.3	6.0
	✓		22.8	47.8	58.9	6.0
	✓	✓	**23.9**	**50.8**	**63.4**	**5.0**

Table 7: The ablation studies on the MSRVTT 1K test set to investigate the different designs of CCM module.

Experimental Evaluation

Method	ClipBERT [28]	Ours
Inference Time	12.5 minutes	20 seconds

Table 5: Total retrieval time for text queries on the MSRVTT 1K test set.

Experimental Evaluation

	E2E	Type	R@1↑	R@5↑	R@10↑	MdR↓
S2VT [49]	✓	EDB	11.9	33.6	-	13.0
FSE [57]	✓	EDB	13.9	36.0	-	11.0
ClipBERT‡ [28]	✓	MDB	20.4	48.0	60.8	6.0
Ours	✓	EDB	**21.2**	**50.3**	**63.5**	**6.0**

Table 4: Results of text-to-video retrieval on DiDeMo dataset. ‡ denotes pretraining on COCO Captions and Visual Genome Captions.

Table of Contents

- Introduction (text-to-video retrieval, EDB and MDB methods, goal of the paper)
- CRET Framework overview (architecture and components)
- Gaussian Estimation of Embedding Space (GEES) Loss
- Cross-modal Correspondence Modeling (CCM) Module
- Experimental Evaluation
- Conclusion and Future Work

Conclusion

Key Achievements:

- Successfully addressed limitations of existing text-to-video retrieval approaches
- Achieved state-of-the-art results without pre-training on external datasets

Main Innovations:

- CCM Module: Effectively models local video-text correspondences through shared queries
- GEES Loss: Balances computational cost and information preservation in frame sampling
- Efficient Architecture: Maintains high accuracy while enabling fast retrieval

Results:

- Outperformed existing methods on multiple datasets
- Demonstrated superior efficiency compared to model-based approaches
- Validated both mathematically and experimentally

Future Work

1. Model Extensions
- Apply CCM module and GEES loss to other model-based (MDB) methods
- Integrate multi-source expert features (motion, sound, OCR) into CRET

2. Pre-training & Generalization
- Explore large-scale video-text pre-training datasets like HowTo100M
- Improve model's ability to generalize across different domains

3. Broader Applications
- Apply GEES loss to other video-related tasks (e.g., video captioning)
- Test methodology on different types of video content and queries

I tried to create a functioning CRET Model as well

link to the Google Colab

I used the MSRVTT Database (not all the 7000 videos but 1000 of them), it is one of the databases that the writers of the paper used for their CRET Model.

any QUESTIONS?