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

Angela Rossi

link to the paper

FIRST, what’s video retrieval?

Video retrieval refers to the task of finding and retrieving relevant videos from a database or collection of videos, based on a given text query. In the context of this paper, the authors are specifically focused on the "text-to-video retrieval" problem, where the user provides a text query and the system must retrieve the most relevant videos from the database.

FIRST, what’s video retrieval?

KEY POINTS ABOUT VIDEO RETRIEVAL

- It is a cross-modal task, meaning it involves bridging the gap between textual and visual data. The system needs to understand the semantics and content of both the text query and the video content in order to match them effectively.

- The goal is to find videos that are relevant and responsive to the user's text query. For example, if the query is "a man cooking in a kitchen", the system should retrieve videos that depict a man cooking in a kitchen setting.

FIRST, what's video retrieval?

KEY POINTS ABOUT VIDEO RETRIEVAL

- This is an important task with many practical applications, such as enabling users to search large video databases (e.g. YouTube, surveillance footage) using natural language queries.

- It poses several challenges, such as modeling the complex correspondences between language and visual content, handling varied video lengths and content, and doing this efficiently to support real-time retrieval.

Table of Contents

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

- What's CRET?

Motivation and Challenges

- Embedding-based (EDB) methods: These methods first extract features for text and video separately, then compute the similarity between them in a joint embedding space. EDB methods are generally efficient, but they often struggle to capture the fine-grained cross-modal correspondences between linguistic and visual elements.

- Model-based (MDB) methods: These methods use deep neural networks to directly predict the similarity between text-video pairs. MDB methods have demonstrated superior accuracy compared to EDB methods, especially when equipped with large-scale pre-training. However, MDB methods suffer from computational inefficiency during the inference stage, as they require exhaustively comparing the query text with every video in the database.

Introduction

- What's CRET?

Key Innovations of CRET

- Cross-modal Correspondence Modeling (CCM) Module: the CCM module uses transformer decoders and shared center queries to align the local features of video and text modalities. This allows the CRET framework to effectively model the intricate relationships between linguistic and visual elements, without suffering from the computational inefficiency of pairwise model-based approaches.

Introduction

- What's CRET?

Key Innovations of CRET

- Gaussian Estimation of Embedding Space (GEES) Loss: The CRET framework makes a key assumption that the frame-level features of a video follow a multivariate Gaussian distribution. Based on this assumption, the authors derive the GEES loss, which serves as a surrogate for the vanilla Noise Contrastive Estimation (NCE) loss. The GEES loss enables the model to efficiently leverage information from all video frames, even with a sparse sampling strategy, thereby mitigating the trade-off between information loss and computational cost.

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.

Introduction

- What's the goal of this paper?

  Plus this research aims to:

1) Exploit the local details and correspondences between video and text as previous embedding-based (EDB) methods often lack the ability to model the intricate relationships between local visual and linguistic features.

2) Achieve efficient retrieval without compromising accuracy, as model based methods (MDB) have shown superior accuracy but suffer from inefficient inference due to the need for exhaustive pairwise comparisons.

Introduction

- EDB Methods:
- These methods first extract features from the text and video modalities using separate encoders.

- They then map the text and video features into a common embedding space, where the similarity between a text query and a video can be computed using distance metrics like cosine similarity.

- EDB methods are generally easier to implement and more efficient during inference, as the text and video embeddings can be pre-computed.

- However, EDB methods often struggle to effectively model the local, fine-grained correspondences between the linguistic and visual elements.

Introduction

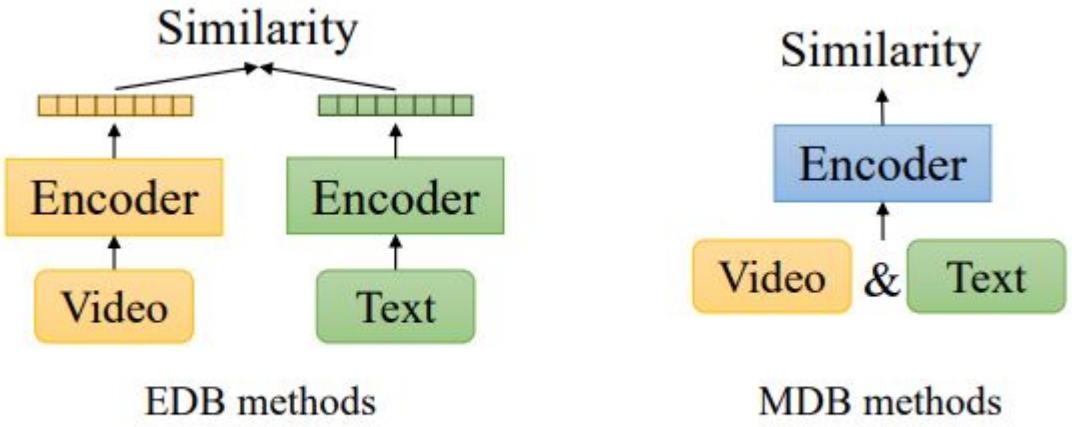
- MDB Methods:
- MDB methods take a text-video pair as input and use deep neural networks to directly predict the similarity or relevance between the two modalities.

- They have the advantage of being able to capture the intricate local relationships between the text and video features.

- MDB methods have demonstrated superior accuracy compared to EDB methods, especially when equipped with large-scale pre-training on extensive video-text datasets.

- However, MDB methods suffer from inefficient inference, as they require exhaustively comparing the query text with every video in the database to compute the similarities.

Introduction



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

Introduction

- What's text to video retrieval?

  Text-to-video retrieval is a challenging cross-modal information retrieval task that aims to find the most relevant videos in a database given a textual query. This differs from traditional unimodal retrieval tasks, such as image retrieval or document retrieval, where the query and the target content are from the same modality.

Introduction

- For example:

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

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.

- 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) is a groundbreaking language model developed by researchers at Google. It has become one of the most influential and widely used models in the field of natural language processing (NLP).

It is a transformer-based model that is trained on a large corpus of text data in an unsupervised manner. The key innovation of BERT is its use of bidirectional training, which allows the model to capture contextual information from both the left and right sides of a given word in a sentence.

CRET Framework Overview

Key Components

TEXT ENCODER

WHAT'S BERT?

BERT is trained using two main objectives:

- Masked Language Model (MLM): During training, BERT randomly masks a certain percentage of the tokens in the input sequence, and then the model is trained to predict these masked tokens based on the surrounding context.

- Next Sentence Prediction (NSP): BERT is also trained to predict whether two given sentences are consecutive in the original text or not. This helps the model learn relationships between sentences.

CRET Framework Overview

Key Components

Cross-modal Correspondence Modeling (CCM) Module

- The CCM module aligns the local features of the video and text modalities using transformer decoders and shared center queries.

- It maps the local text and video features into a common embedding space, emphasizing the correspondences between linguistic and visual elements.

CRET Framework Overview

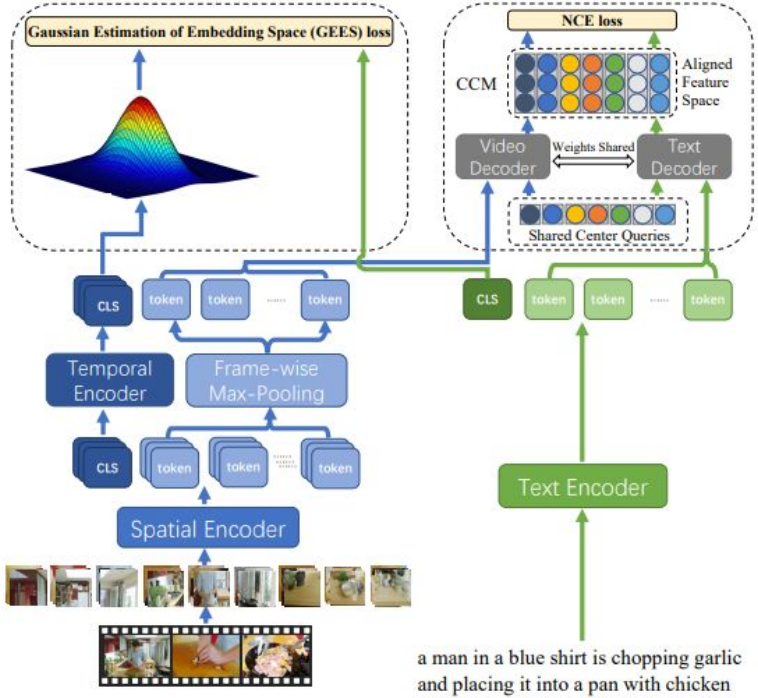
Key Components

Gaussian Estimation of Embedding Space (GEES) Loss

- The GEES loss is a novel loss function that addresses the trade-off between information loss and computational cost in video frame sampling.

- It builds upon the Noise Contrastive Estimation (NCE) loss, but makes a Gaussian assumption about the distribution of video frame features to enable efficient optimization.

CRET Framework Overview



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

- The authors of the CRET paper made a key assumption: the frame-level features of a video follow a multivariate Gaussian distribution.

- Let's denote the frame-level features as **υ**, which follow a Gaussian distribution with mean **μ** and covariance **σ: υ ~ N(μ, σ)**

Gaussian Estimation of Embedding Space (GEES) Loss

GAUSSIAN ASSUMPTION

GEES LOSS DERIVATION

- Using the properties of Gaussian distributions, the authors were able to derive an upper bound for the expectation of the vanilla NCE loss, which becomes the GEES loss:

  $L_{GEES} \leq -\log(\exp(\langle T^g, \mu \rangle + 1/2 \langle T^g, \sigma T^g \rangle) / \Sigma \exp(\langle T^g_j, \mu \rangle + 1/2 \langle T^g_j, \sigma T^g_j \rangle))$

Gaussian Estimation of Embedding Space (GEES) Loss

GAUSSIAN ASSUMPTION

GEES LOSS DERIVATION

About the formula:

- **T^g** represents the global text features, which is a 1D vector capturing the overall semantics of the text.
- **μ** is the mean vector of the video frame-level features, which follows the assumed Gaussian distribution.
- **⟨T^g, μ⟩** is the dot product between the global text features and the mean video features, capturing the similarity between the text and the average video representation.

Gaussian Estimation of Embedding Space (GEES) Loss

GAUSSIAN ASSUMPTION

GEES LOSS DERIVATION

About the formula:

- **⟨T^g, σ T^g⟩** is the dot product between the global text features and the matrix-vector product of the covariance matrix σ and the text features T^g. This term captures the impact of the variance of the video features on the text-video similarity.
- The numerator **exp(⟨T^g, μ⟩ + 1/2 ⟨T^g, σ T^g⟩)** represents the similarity between the global text features and the Gaussian-distributed video features.

Gaussian Estimation of Embedding Space (GEES) Loss

GAUSSIAN ASSUMPTION

GEES LOSS DERIVATION

About the formula:

- The denominator **Σ exp(⟨T^g_j, μ⟩ + 1/2 ⟨T^g_j, σ T^g_j⟩)** sums up the similarities between the global text features and the Gaussian-distributed video features for all text-video pairs in the batch.
- Taking the negative logarithm of this ratio gives us the GEES loss, which is an upper bound of the expected vanilla NCE loss.

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

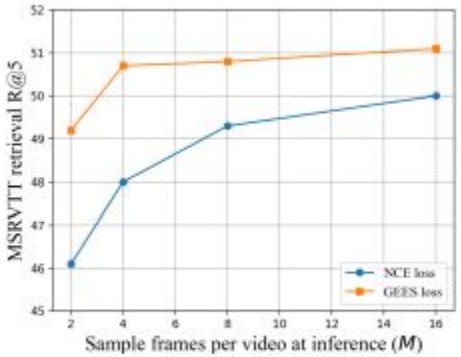
GAUSSIAN ASSUMPTION

VALIDATION

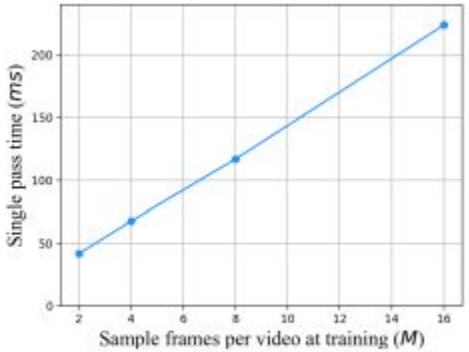
Three-step validation process to support the Gaussian assumption:

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

Gaussian Estimation of Embedding Space (GEES) Loss



(a) The effectiveness of GEES loss.



(b) Time cost in the training stage.

Validation of Gaussian Assumption

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.

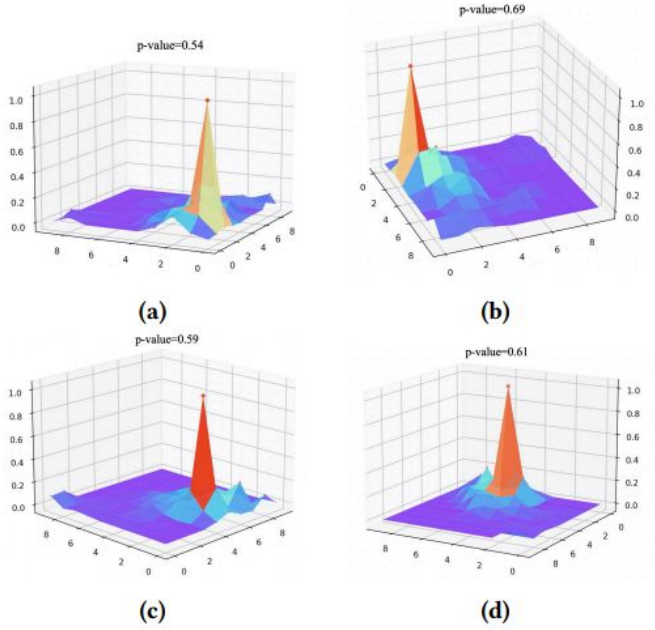


Figure 4: Probability density of video frames features. We utilize the histogram to estimate the probability density. Z-axis represents the probability density.

Validation of Gaussian Assumption

Statistical Testing:

- Performed the Henze-Zirkler test, a well-established statistical test for multivariate normality, on a sample of 300 videos from the dataset.

- All 300 videos passed the test at a 95% confidence level, supporting the Gaussian assumption.

Validation of Gaussian Assumption

Statistical Testing:

WHAT IS THE HENZE-ZIRKLER TEST?

The Henze-Zirkler test is a statistical test used to determine whether a set of random vectors follows a multivariate normal (Gaussian) distribution. It's a well-established technique for validating the multivariate normality assumption.

Validation of Gaussian Assumption

Statistical Testing:

HOW DOES THE HENZE-ZIRKLER TEST WORK?

- **Hypothesis testing**: The null hypothesis (H0) is that the random vectors follow a multivariate normal distribution, and the alternative hypothesis (H1) is that the random vectors do not follow a multivariate normal distribution.

Validation of Gaussian Assumption

Statistical Testing:

HOW DOES THE HENZE-ZIRKLER TEST WORK?

- **Test statistic**: The Henze-Zirkler test statistic is calculated based on the sample data, taking into account the sample size and the dimensionality of the random vectors, and they follow an approximate normal distribution under the null hypothesis.

Validation of Gaussian Assumption

Statistical Testing:

HOW DOES THE HENZE-ZIRKLER TEST WORK?

- **p-value calculation**: The test statistic is used to compute the p-value, which represents the probability of observing the given test statistic (or more extreme values) under the null hypothesis.

Validation of Gaussian Assumption

Statistical Testing:

HOW DOES THE HENZE-ZIRKLER TEST WORK?

- **Decision rule**: If the p-value is greater than the chosen significance level (e.g., 0.05 or 5%), the null hypothesis is not rejected, and the data is considered to be consistent with a multivariate normal distribution. And if it is less than the significance level, the null hypothesis is rejected, and the data is considered to deviate from a multivariate normal distribution.

Validation of Gaussian Assumption

Binomial Distribution Analysis:

- Treated the Gaussian/non-Gaussian classification of video frame features as a Bernoulli trial.

- Using the "rule of three" from Binomial distribution theory, the authors concluded that with 95% confidence, the frame features of every video in the test set follow a multivariate Gaussian distribution with a probability of at least 0.99.

Validation of Gaussian Assumption

Binomial Distribution Analysis:

BERNOULLI TRIAL

In a Bernoulli trial, an experiment has two possible outcomes - "success" or "failure". In the context of the CRET paper, the authors considered a "success" to be the case where the video frame features follow a Gaussian distribution, and a "failure" to be the case where they do not.

Validation of Gaussian Assumption

Binomial Distribution Analysis:

BERNOULLI TRIAL

Specifically, the authors randomly selected 300 videos from the MSRVTT dataset and ran the Henze-Zirkler test on each one. Since all 300 videos passed the test at a 95% confidence level, the authors could infer the following:

1. The probability (p) of the video frame features following a Gaussian distribution is at least 0.99 (99%).
2. This is because if the true probability p was less than 0.99, then the probability of observing 300 "successes" (all 300 videos passing the Henze-Zirkler test) would be less than 5% - violating the 95% confidence level.

Validation of Gaussian Assumption

Binomial Distribution Analysis:

RULE OF THREE

The rule of three states that if an event has not been observed in n independent trials, then with 95% confidence, the probability p of that event occurring is less than 3/n.

So:

The authors observed 300 "successes" (all 300 videos passing the Henze-Zirkler test)

Since there were no "failures" (videos not passing the test), they could apply the rule of three

This allowed them to conclude that with 95% confidence, the probability p of the video frame features not following a Gaussian distribution is less than 3/300 = 0.01

Cross-modal Correspondence Modeling (CCM) Module

The CCM module is a crucial component of the CRET framework that aims to effectively align the local features of video and text modalities.

The key idea behind the CCM module is to use transformer decoders and shared center queries to bridge the semantic gap between the local video and text representations.

Cross-modal Correspondence Modeling (CCM) Module

Local Feature Extraction:

- The video encoder extracts a set of local patch-level features $V^l = \{V^l_{ij}\}$ from the video frames.

- The text encoder extracts a sequence of local token-level features $T^l = \{T^l_{ij}\}$ from the text captions.

Cross-modal Correspondence Modeling (CCM) Module

Local Feature Extraction:

Features $V^I = \{V^I_{ij}\}$

- The video encoder, which is a CaiT-S/24 transformer model, takes the video frames as input and divides each frame into p non-overlapping image patches.ù

- For each video frame, the CaiT model computes a set of patch embeddings, where each patch embedding represents the local visual information in that particular region of the frame.

Cross-modal Correspondence Modeling (CCM) Module

Local Feature Extraction:

Features $V^l = \{V^l_{ij}\}$

These patch-level features are collectively represented as $V^l = \{V^l_{ij}\}$, where:

- i indexes the video frame
- j indexes the patch within that frame
- V^l_{ij} is the embedding vector for the j-th patch in the i-th video frame

V^l is a 3D tensor, where the first dimension corresponds to the video frames, the second dimension corresponds to the patches within each frame, and the third dimension is the feature vector for each patch.

Cross-modal Correspondence Modeling (CCM) Module

Local Feature Extraction:

Features $T^l = \{T^l_{ij}\}$

The text encoder in the CRET framework, which uses a pre-trained BERT model, extracts two types of text features:

- Global text features (T^g): These are the features extracted from the special CLS token at the beginning of the text sequence. They capture the overall, high-level semantics of the text.

- Local text features (T^l): These are the token-level features extracted for each word in the text sequence

Cross-modal Correspondence Modeling (CCM) Module

Local Feature Extraction:

Features $T^l = \{T^l_{ij}\}$

This notation represents this set of local text features:

- T^l is the collection or set of all local text features extracted from the input text.
- Each individual element T^l_{ij} corresponds to the feature vector for the j-th token in the i-th text caption.
- The subscript '**i**' indexes the text caption, since there may be multiple captions or text inputs.
- The subscript '**j**' indexes the individual token or word within each text caption.

Cross-modal Correspondence Modeling (CCM) Module

Cross-modal Alignment:

- The CCM module uses C shared center queries **Q = {Q^c}** to compute the alignment between the local video and text features.

- For each center query **Q^c**, the module calculates the similarity scores between **Q^c** and the local features E **(either V^l or T^l)**.

- These similarity scores are then used as weights to aggregate the local features, producing the aligned video features **Z^v** and text features **Z^t**.

Cross-modal Correspondence Modeling (CCM) Module

Similarity Computation:

The final similarity score between a text-video pair is computed as the cosine distance between the aligned local features **Z^v** and **Z^t**.

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.

- Reported R@1, R@5, R@10

- **MdR (Median Rank):** Median rank of the correct item in the retrieved ranking

Experimental Evaluation

Implementation Details:

- **Video Encoder**: CaiT-S/24 transformer for spatial encoding, 3-layer transformer for temporal encoding.
- **Text Encoder**: BERT base-uncased model.
- **Optimizer**: Adam, Initial learning rate: 5e-5
- **Training**: 50 epochs, batch size of 20 per GPU, 8 NVIDIA V100 GPUs
- **Video Frame Sampling**: 4 frames per video, randomly sampled from 4 equal-length intervals

Conclusion

- The CRET framework successfully addresses the limitations of existing approaches in text-to-video retrieval by effectively modeling the local correspondences between video and text modalities.

- The key innovations of CRET include the Cross-modal Correspondence Modeling (CCM) module and the Gaussian Estimation of Embedding Space (GEES) loss.

Conclusion

- The CCM module uses transformer decoders and shared center queries to align the local visual and linguistic features, enabling the capture of fine-grained cross-modal relationships.

- The GEES loss leverages the assumption of Gaussian distributed video frame features to efficiently compute text-video similarities, mitigating the trade-off between information loss and computational cost in video frame sampling.

- Extensive experiments demonstrate that CRET outperforms state-of-the-art methods, including those pre-trained on additional datasets, in terms of both retrieval accuracy and efficiency.

Future Work

- Explore extending the CCM module and GEES loss to other model-based (MDB) methods, beyond the embedding-based (EDB) framework presented in this paper.

- Investigate the integration of multi-source expert features, such as motion, sound, and OCR, into the CRET framework to further boost retrieval performance.

Future Work

- Leverage large-scale video-text pretraining datasets, like HowTo100M, to potentially improve the CRET model's generalization capabilities.

- Apply the GEES loss to other video-related tasks, such as video captioning, to assess its broader applicability and generalizability.

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?