Feature Fusion and Negation Understanding for Ad-hoc Video Search

Aozhu Chen, Ziyue Wang, Fan Hu, Kaibin Tian, Xirong Li

AIMC Lab, School of Information, Renmin University of China
(TRECVID team ID: RUCMM)
https://ruc-aimc-lab.github.io/

6 December 2022

Key question in Ad-hoc Video Search (AVS)

- **How to estimate the relevance of an unlabeled video w.r.t a specific text query ?**

Text query allows human to express

😊 what we do want：

A man is holding a knife in a kitchen location

😫 what we do NOT want：

A man is holding a knife in a non-kitchen location (730)

Our Solution

Based on two techniques: **Feature Fusion + Bidirectional Negation Learning**

Lightweight Attentional Feature Fusion: A New Baseline for Text-to-Video Retrieval

Fan Hu[1,2*], Aozhu Chen[1,2*], Ziyue Wang[1,2*], Fangming Zhou[1,2], Jianfeng Dong[3], and Xirong Li[1,2†]

[1] MoE Key Lab of DEKE, Renmin University of China
[2] AIMC Lab, School of Information, Renmin University of China
[3] College of Computer and Information Engineering, Zhejiang Gongshang University

Abstract. In this paper we revisit *feature fusion*, an old-fashioned topic, in the new context of text-to-video retrieval. Different from previous research that considers feature fusion only at one end, let it be video or text, we aim for feature fusion for both ends within a unified framework. We hypothesize that optimizing the convex combination of the features is preferred to modeling their correlations by computationally heavy multi-head self attention. We propose Lightweight Attentional Feature Fusion (LAFF). LAFF performs feature fusion at both early and late stages and at both video and text ends, making it a powerful method for exploiting diverse (off-the-shelf) features. The interpretability of LAFF can be used for feature selection. Extensive experiments on five public benchmark sets (MSR-VTT, MSVD, TGIF, VATEX and TRECVID AVS 2016-2020) justify LAFF as a new baseline for text-to-video retrieval.

Keywords: Text-to-video retrieval, video/text feature fusion

1 Introduction

Text-to-video retrieval is to retrieve videos *w.r.t.* to an ad-hoc textual query from many *unlabeled* videos. Both video and text have to be embedded into one or more cross-modal common spaces for text-to-video matching. The state-of-the-art tackles the task in different approaches, including novel networks for query representation learning [59, 65], multi-modal Transformers for video representation learning [3, 19], hybrid space learning for interpretable cross-modal matching [15, 60], and more recently CLIP2Video [17] that learns text and video representations in an end-to-end manner. Differently, we look into *feature fusion*, an important yet largely underexplored topic for text-to-video retrieval.

Given video/text samples represented by diverse features, feature fusion aims to answer a basic research question of *what is the optimal way to combine these features?* By optimal we mean the fusion shall maximize the retrieval performance. Meanwhile, the fusion process shall be explainable to interpret the importance of the individual features. As the use of each feature introduces extra

*Equal contribution.
†Corresponding author: Xirong Li (xirong@ruc.edu.cn)

LAFF [Hu et al., ECCV'22]

Focus on Feature Fusion

Learn to Understand Negation in Video Retrieval

Ziyue Wang*
AIMC Lab, School of Information, Renmin University of China
China

Aozhu Chen*
AIMC Lab, School of Information, Renmin University of China
China

Fan Hu
AIMC Lab, School of Information, Renmin University of China
China

Xirong Li†
MoE Key Lab of DEKE, Renmin University of China
China

Figure 1: Top-1 video retrieved by different models, *i.e.* W2VV++ [19], SEA [20], CLIP [28], CLIP* (fine-tuned by this work), CLIP4Clip [25] and our CLIP-*bnl*, which is CLIP re-trained with proposed negation learning. This paper presents the first study on a learning based method for handling negation in text-to-video retrieval (nT2VR). Data source: MSR-VTT [32].

ABSTRACT

Negation is a common linguistic skill that allows human to express what we do NOT want. Naturally, one might expect video retrieval to support natural-language queries with negation, e.g., finding shots of kids sitting on the floor and not playing with a dog. However, the state-of-the-art deep learning based video retrieval models lack such ability, as they are typically trained on video description datasets such as MSR-VTT and VATEX that lack negated descriptions. Their retrieved results basically ignore the negator in the sample query, incorrectly returning videos showing kids playing with dog. This paper presents the first study on learning to understand negation in video retrieval and make contributions as follows. By re-purposing two existing datasets (MSR-VTT and VATEX), we propose a new evaluation protocol for video retrieval with negation. We propose a learning based method for training a negation-aware video retrieval model. The key idea is to first construct a soft negative caption for a specific training video by partially negating its original caption, and then compute a bidirectionally constrained loss on the triplet. This auxiliary loss is weightedly added to a standard retrieval loss. Experiments on the re-purposed benchmarks

*Z. Wang and A. Chen contributed equally to this research.
†Corresponding author: Xirong Li (xirong@ruc.edu.cn).

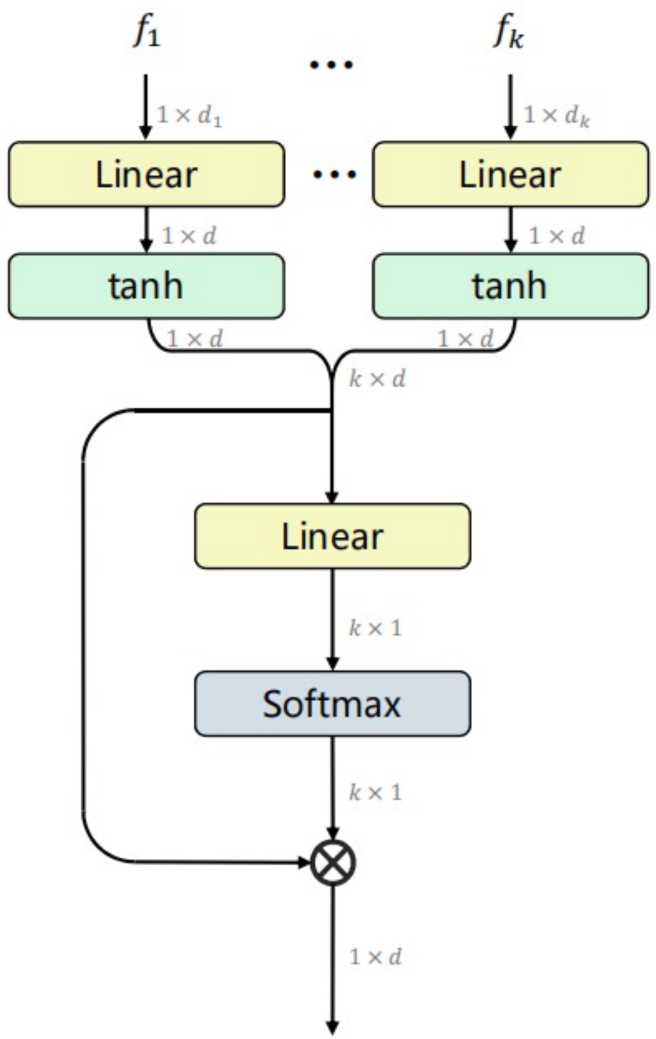
MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3547968

BNL [Wang et al., ACMMM'22]

Focus on Negation-Aware Video Retrieval

Technique 1 LAFF based Video Retrieval



Lightweight Attentional Feature Fusion (LAFF)

Transforming features into *d*-dimensional feature vector :

$$f_i' = \sigma(Linear_{d_i \times d}(f_i))$$

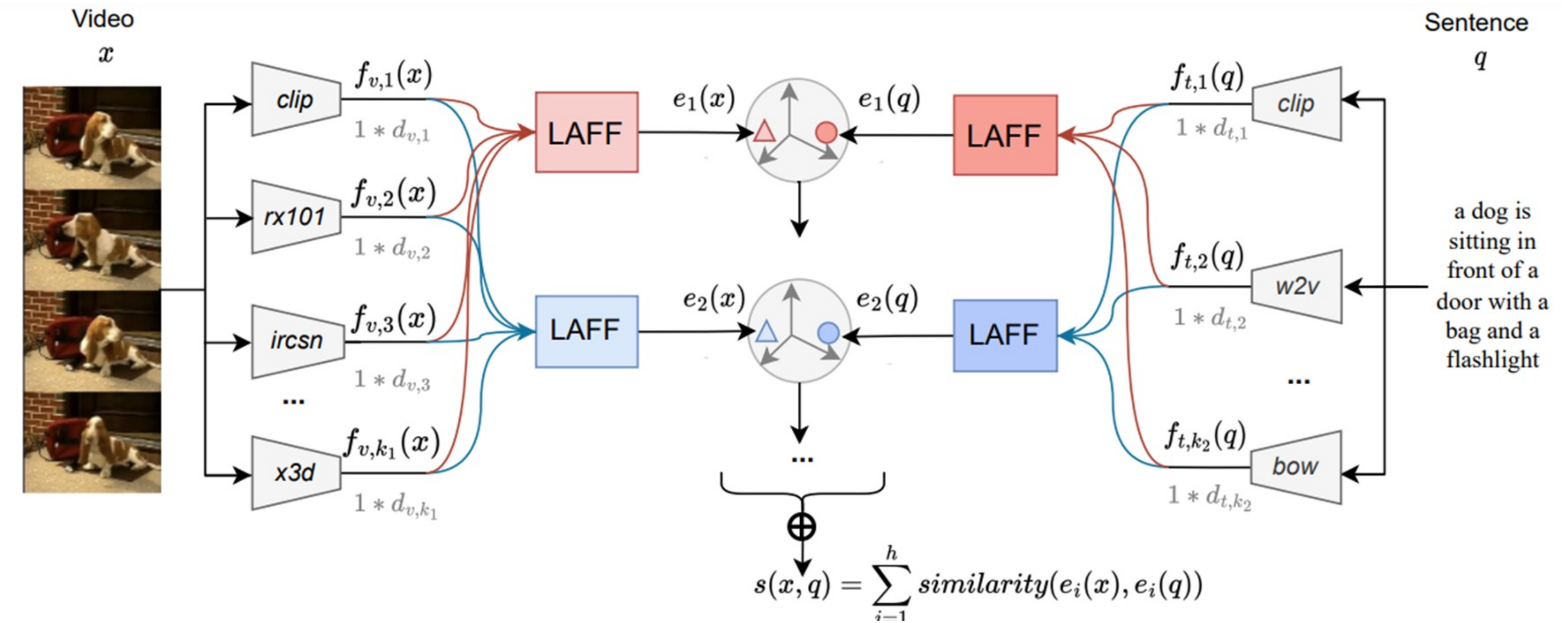
Aggregating the transformed features into a combined feature :

$$\bar{f}_v = LAFF(\{f_1, \ldots, f_k\}) = \sum_{i=1}^{k} a^i \times f_i'$$

$$\{a_1, \ldots, a_k\} = softmax(Linear_{d \times 1}(\{f_1', \ldots, f_k'\}))$$

Technique 1 LAFF based Video Retrieval

- ***How to use LAFF?***



It supports feature fusion at both text and video ends to exploit diverse (off-the-shelf) features.

Technique 2 BNL for Negation-Aware Video Retrieval

Re-purpose video-caption datasets

Video

Original caption *A man is playing the guitar while dancing with many other people*

Negated query *A man is playing the guitar while* ***not*** *dancing with many other people*

Video

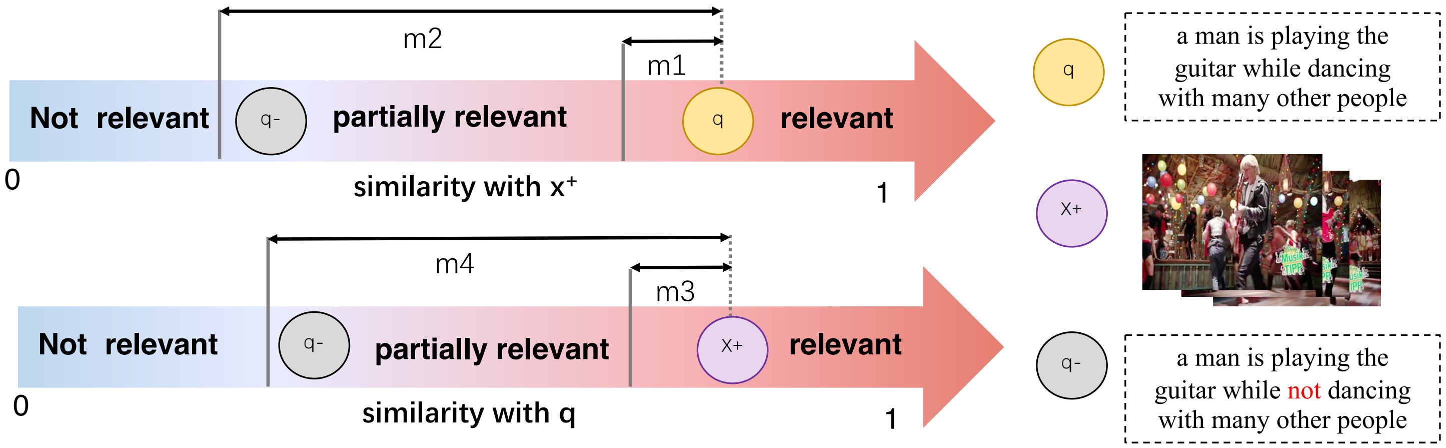
Original caption *A car is being flipped over*

Negated query *A car* ***isn't*** *being flipped over*

Insert negation cue before verbs or after auxiliary verbs.

Technique 2 BNL for Negation-Aware Video Retrieval

Bidirectional Constraint Loss



- x^+ shall be closer q to than q $s(x^+,q) > s(x^+,q^-)$
- q^- shall not be pushed too far away from x^+

$$\ell_{bcl}(x^+,q,q^-) = \max(0, m_1 + s(x^+,q^-) - s(x^+,q)) + \max(0, -\boxed{m_2} - s(x^+,q^-) + s(x^+,q))$$

upper boundary

- q shall be closer to x^+ than q^-: $s(q,x^+) > s(q,q^-)$
- q^- shall not be pushed too far away from q

$$\ell_{bcl}(q,x^+,q^-) = \max(0, m_3 + s(q,q^-) - s(q,x^+)) + \max(0, -m_4 - s(q,q^-) + s(q,x^+)),$$

x^- : irrelevant video $x^\#$: hardest negative video x^+ : relevant video m_* : hyper parameter

q^- : soft negative cpation q : raw sentence $s(\cdot,\cdot)$: similarity

Technique 2 BNL for Negation-Aware Video Retrieval

- ***How to use BNL?***

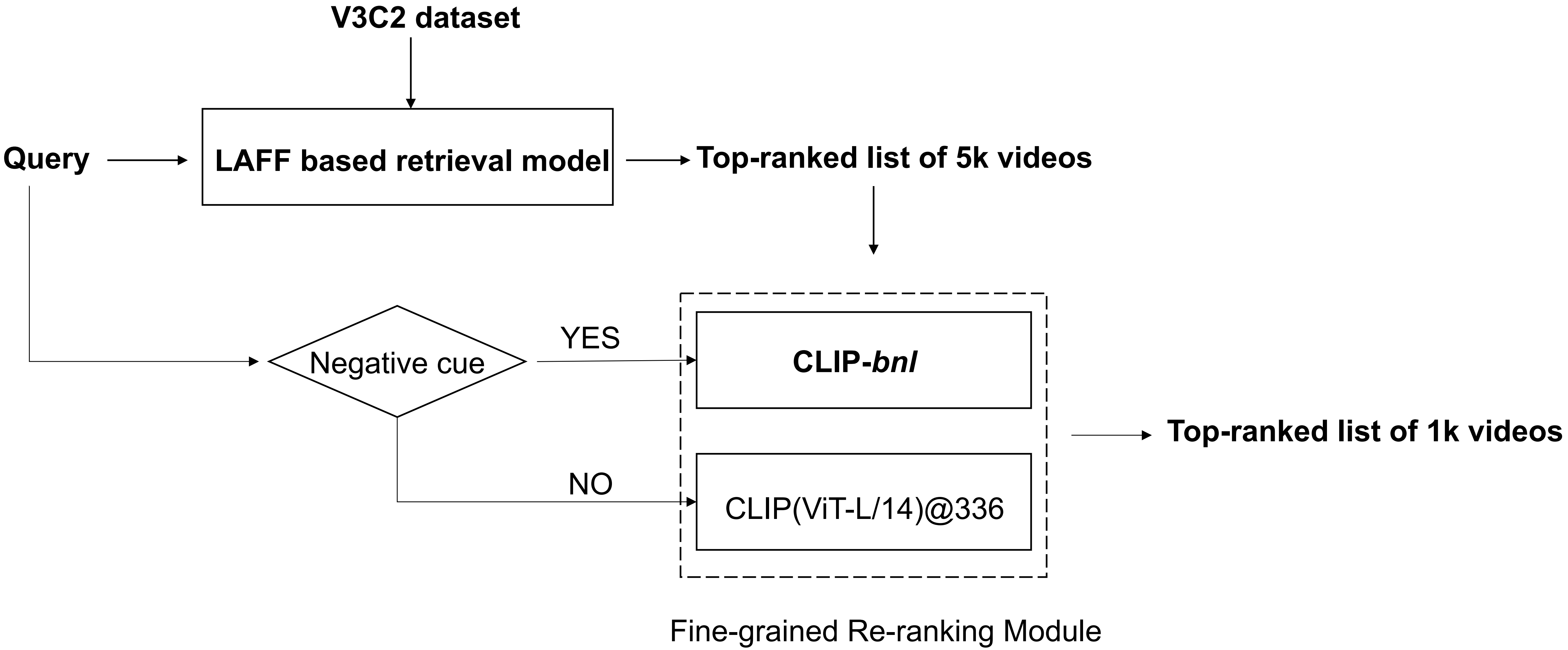
CLIP-*bnl*：

Using the BNL loss to retrain CLIP(ViT-B/32) by a negation-enriched version of MSR-VTT

CLIP-*bnl* is used in the following two manners:

- As a **cross-modal extractor** for both video and query representation.
- As a **re-ranking module** specifically used for queries that have negative cues automatically detected.

RUCMM Video Search Engine



Choice of (Pre-)Training Data

Three public datasets for training

Dataset	#Videos	#Sentences
MSR-VTT (CVPR2016)	10,000	200,000
TGIF (CVPR2016)	100,855	124,534
VATEX (ICCV2019)	32,239	259,909

#1 a crowd at a music festival
#2 a concert with people on the stage

One self-built video-text dataset for pre-training

Dataset	Frame/segment/Video Num	Sentence Num
V3C1-pseudo-caption	1,605,335/219,530/9,760	436,203

Choice of Video/ Text Feature

Seven video features & Six text features

Video Features	Dimensionality
irCSN	2048
ResNeXt101	2048
BEiT	2048
BLIP256	256
CLIP(B/32)	512
CLIP-*bnl*(B/32)	512
CLIP(L/14)@336	768

Text Features	Dimensionality
BoW	**10k+**
W2V	300
BLIP256	256
CLIP(B/32)	512
CLIP-bnl(B/32)	512
CLIP(L/14)@336	768

Heavy text encoders:

- BoW: High dimensions
- W2V: Big storage

Internal experiments

- ***Can we remove the bow and w2v when using LAFF？***

Heavy text encoders:

- BoW: High dimensions
- W2V: Big storage

Run id	TV16	TV17	TV18	TV19	TV20	TV21	MEAN
Run 4	0.282	0.368	0.197	0.255	0.361	0.365	0.305
Run 3	0.280	0.350	0.178	0.244	0.319	0.326	0.283

Run 4: LAFF
Run 3: LAFF (w/o BoW and W2V)

Submissions (fully automatic track)

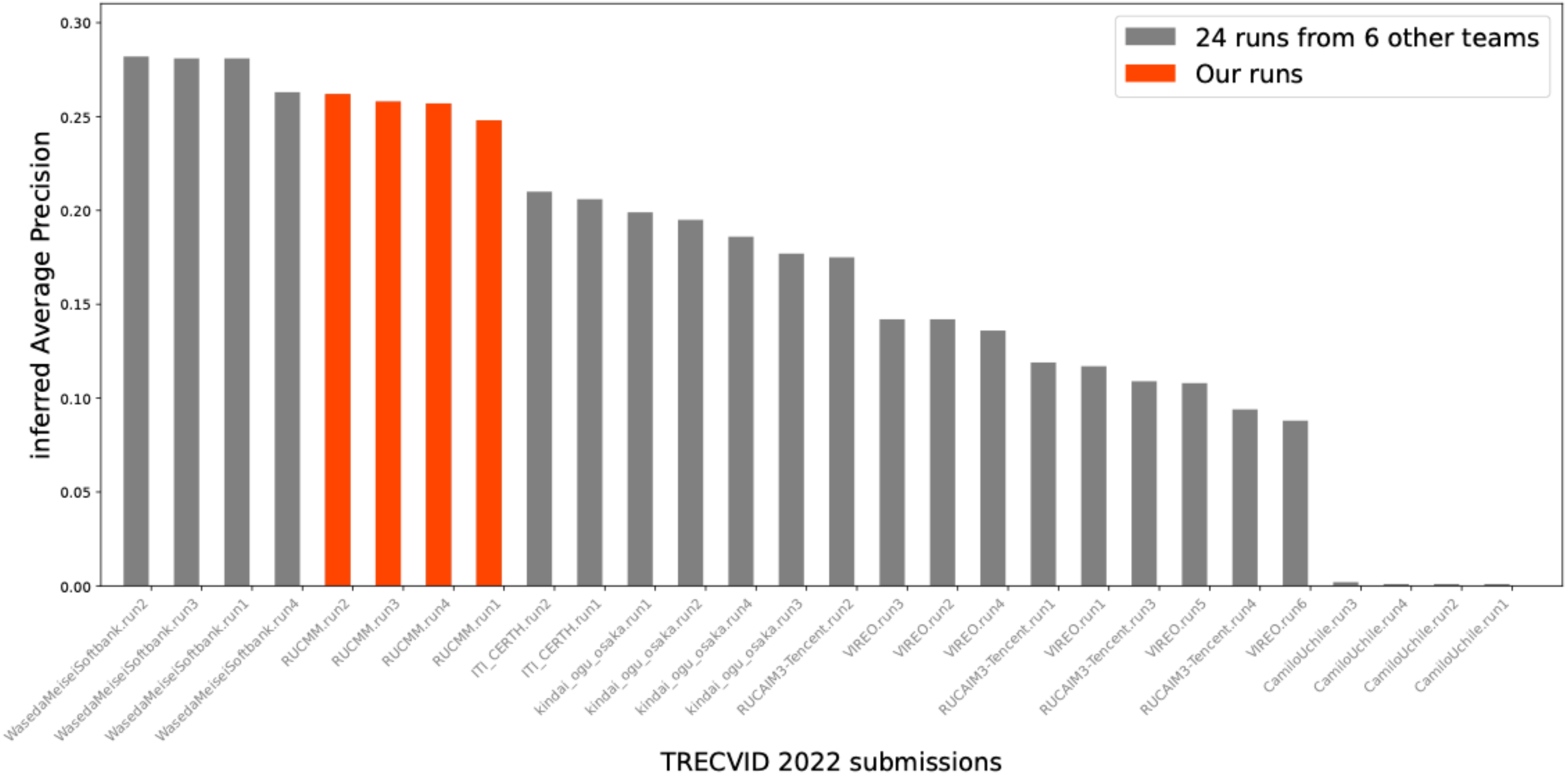
We submitted the following 4 runs:

- ***Run 4***: LAFF
- ***Run 3***: LAFF (w/o BoW and W2V)
- ***Run 2*** : Late average fusion of Run3 on test queries and narrative of queries.
- ***Run 1***: Late average fusion of multiple augmented query retrieval results.

NOTE: Search result reranking is applied on all Runs

Benchmark evaluation

Our submissions ranked the 2rd

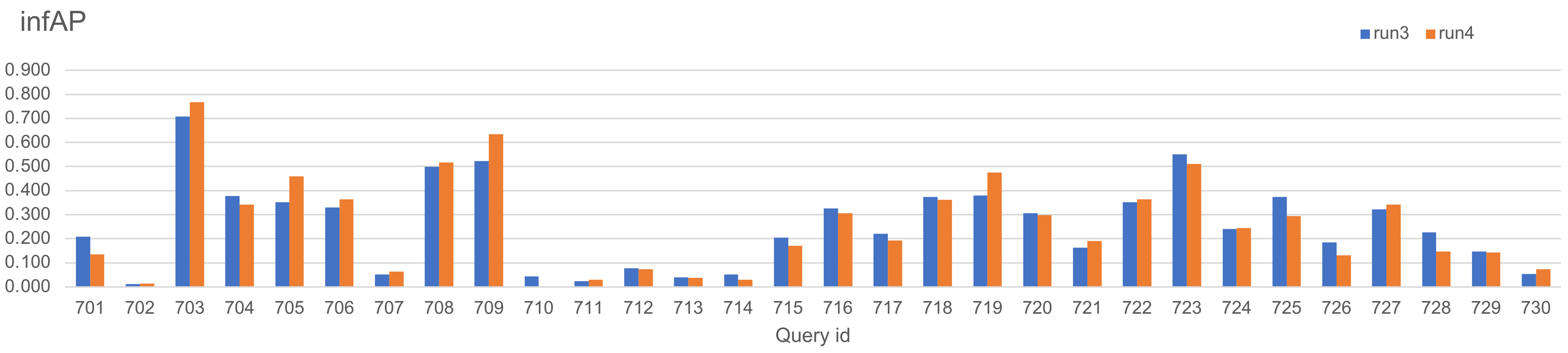


Retrospective experiments

- ***Can we remove the BoW and W2V when using LAFF ?***

Run id	TV22
Run 4	**0.257**
Run 3	**0.258**

- ***Run 4***: LAFF
- ***Run 3***: LAFF (w/o BoW and W2V)



Those heavy text encoders (BoW and W2V) can be removed.

Retrospective experiments

- ***Is BNL Effective?***

  - As a cross-modal extractor for both video and query representation ✅

  - As a re-ranking module specifically used for queries that have negative cues automatically detected.

730 A man is holding a knife in a non-kitchen location

Model	CLIP-*bnl*	Rerank	Query 730
LAFF	✖	✖	0.070
	✖	✓	0.069
	✓	✖	0.094

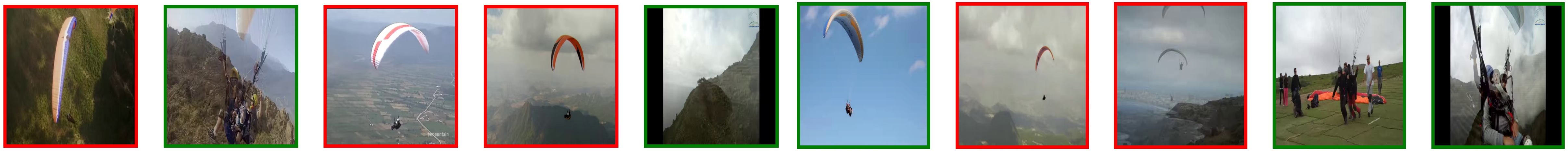
Using CLIP-*bnl* as a feature extractor can improve the performance of negation query.

Retrospective experiments

- ***Whether text augmentation is useful？***

Automatically appending noun / adjective based keywords at the end of each query

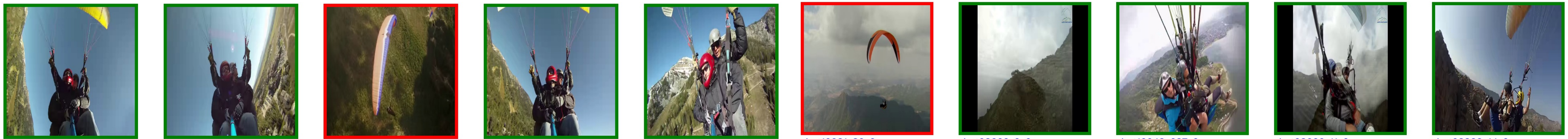
728 **Two adults** are seated in a flying paraglider in the air



shot10981_84_0 shot08833_12_0 shot10528_57_0 shot10981_80_0 shot08262_9_0 shot16662_69_0 shot10981_122_0 shot14909_17_0 shot14909_4_0 shot08262_11_0

Top-ranked list of 10 videos

Two adults are seated in a flying paraglider in the air **two adults**



shot17124_29_0 shot17124_33_0 shot10981_84_0 shot17124_34_0 shot17124_5_0 shot10981_80_0 shot08262_9_0 shot10048_687_0 shot08262_11_0 shot08833_14_0

Yes, but only work on the query with a simple sentence structure.

Retrospective experiments

Whether text augmentation is useful？

Automatically appending noun / adjective based keywords at the end of each query

726 Two teams playing a game where one team have their players wearing white t-shirts.

shot16629_175_0 shot16629_162_0 shot11492_10_0 shot11855_64_0 shot12992_134_0 shot11123_104_0 shot11123_107_0 shot11492_39_0 shot16629_163_0 shot07802_127_0

Top-ranked list of 10 videos

Two teams playing a game where one team have their players wearing white t-shirts. **white t-shirts.**

shot12992_134_0 shot11652_72_0 shot16869_15_0 shot11492_7_0 shot11492_10_0 shot11492_48_0 shot11492_5_0 shot11492_55_0 shot11492_39_0 shot11492_19_0

The context of query with a complex sentence structure is ignored.

Conclusions

- LAFF is an effective feature fusion block for video retrieval.
- BNL makes some favorable effects on training a negation-aware video retrieval model, but negation-aware is still hard.
- The query understanding is essential.

https://github.com/ruc-aimc-lab/laff

https://github.com/ruc-aimc-lab/nT2VR

caz@ruc.edu.cn