RUC_AIM3 at TRECVID 2020: Ad-hoc Video Search

Yida Zhao, Yuqing Song, Shizhe Chen, Qin Jin*

AI·M^3 , Renmin University of China

RUC_AIM3 Team

MULTI MODAL MULTI LEVEL MULTI ASPECT

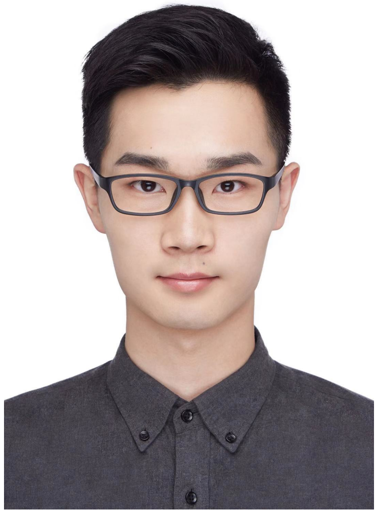
AI · M^3

中国人民大学多媒体计算实验室

- We learn, think, and express through multiple modalities
- AI system needs to have the ability to understand the multimodal world
- We focus on understanding from Multi-Level, Multi-Aspect, and Multi-Modal (M^3)

Team Members

RUC_AIM3 Team



Yida Zhao

Yuqing Song

Shizhe Chen

Qin Jin

Outline

● Task Introduction

● The Proposed System

● Conclusions

One

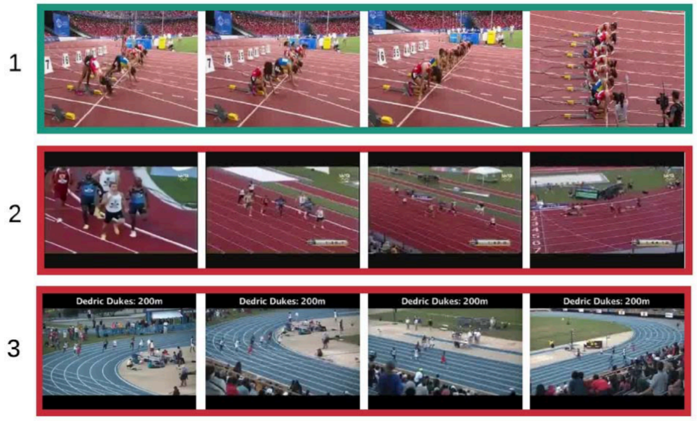
Task Introduction

Task Introduction

- Ad-hoc Video Search

  Given a text query, retrieve the most relevant top 1000 video clips from the V3C vimeo collection [1], which contains about one million video clips

Text: several woman are setting up in the blocks preparing to start a track race.



[1] V3c–a research video collection, ICMM, 2019

Task Introduction

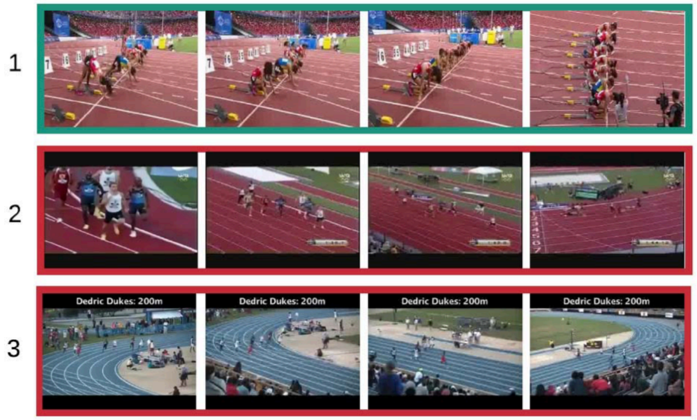
- Ad-hoc Video Search

  Given a text query, retrieve the most relevant top 1000 video clips from the V3C vimeo collection [1], which contains about one million video clips

- Challenge

  The semantic matching between videos and texts

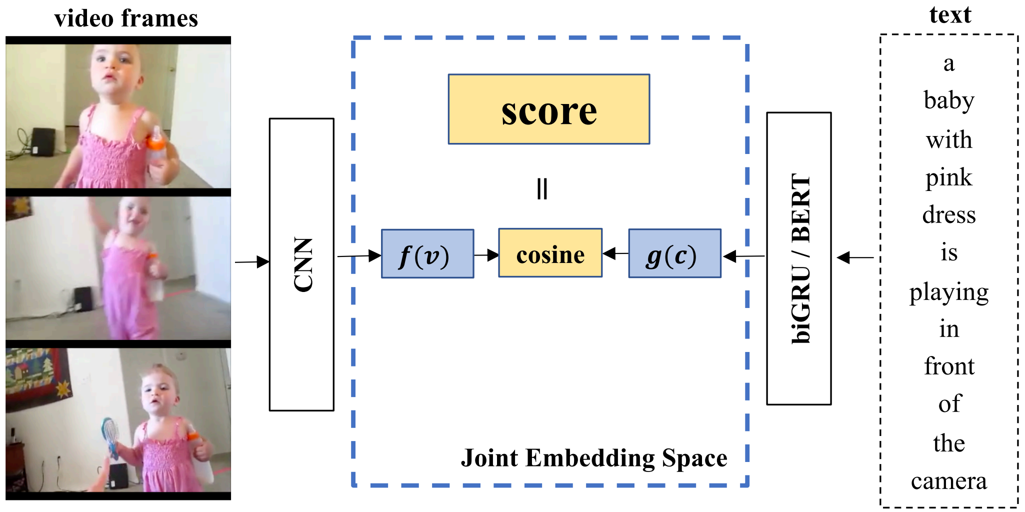
Text: several woman are setting up in the blocks preparing to start a track race.



[1] V3c–a research video collection, ICMM, 2019

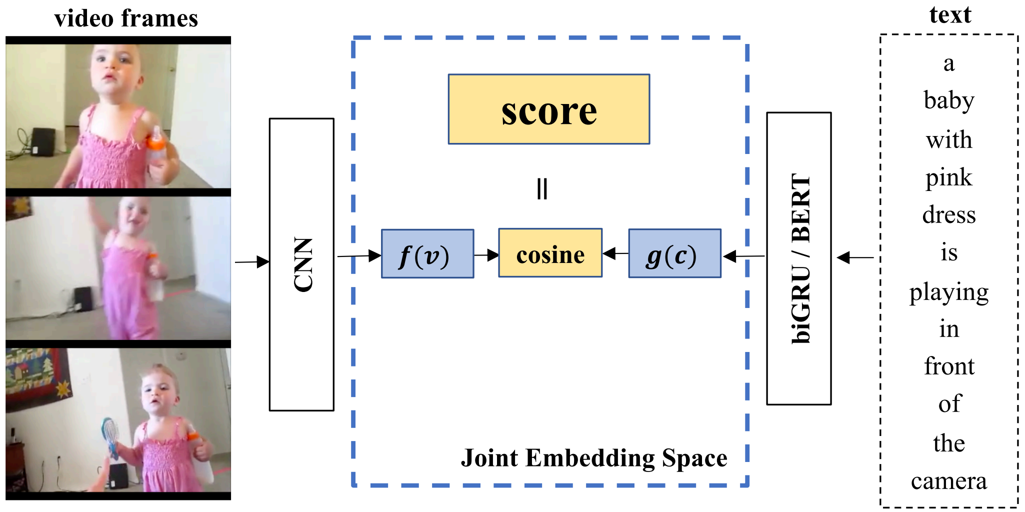
Task Introduction

- Video-Text Cross-modal Retrieval
  - Dominant approach: learning joint embedding space and global visual-semantic matching



Task Introduction

- Video-Text Cross-modal Retrieval
  - Dominant approach: learning joint embedding space and global visual-semantic matching



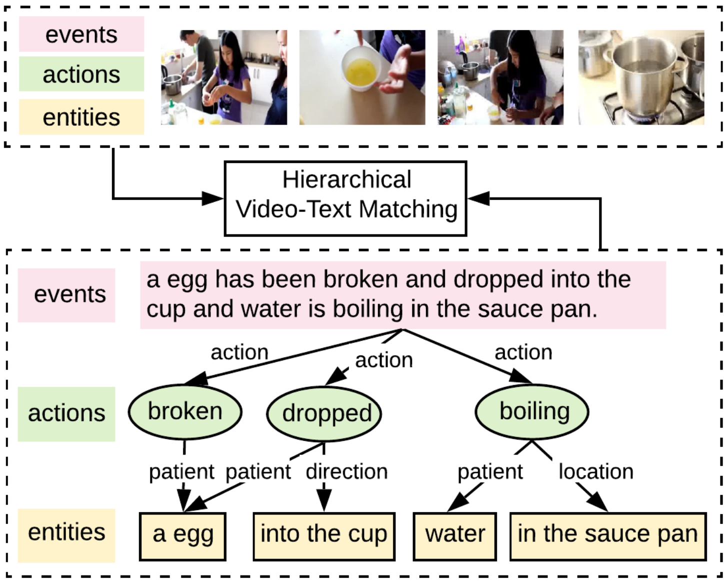
☹ One vector is hard to encode fine-grained details

Two

Our System
Two-branch Model

Two-branch Model

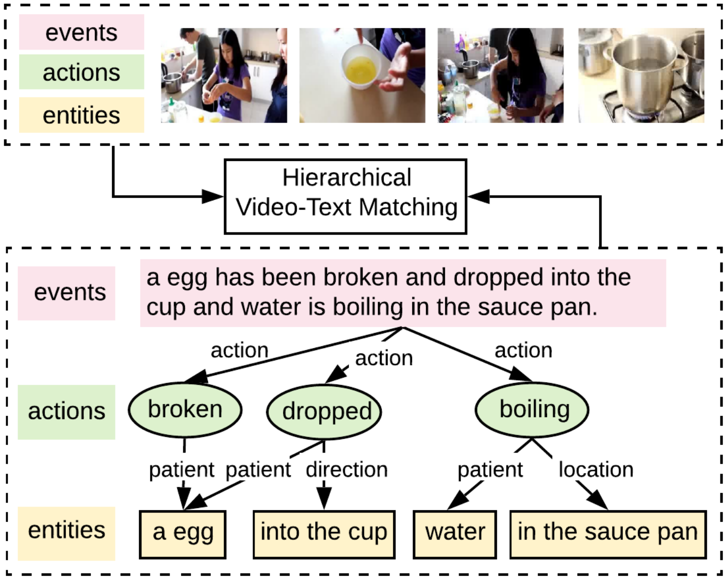
- Two-branch Matching Model
  - Global Matching
    - VSE++ [2]
    - Dual Encoding [3]
  - Fine-grained Matching
    - HGR [4]



[2] Vse++: Improving visual-semantic embeddings with hard negatives, BMVC, 2018
[3] Dual encoding for zero-example video retrieval, CVPR, 2019
[4] Fine-grained video-text retrieval with hierarchical graph reasoning, CVPR, 2020

Two-branch Model

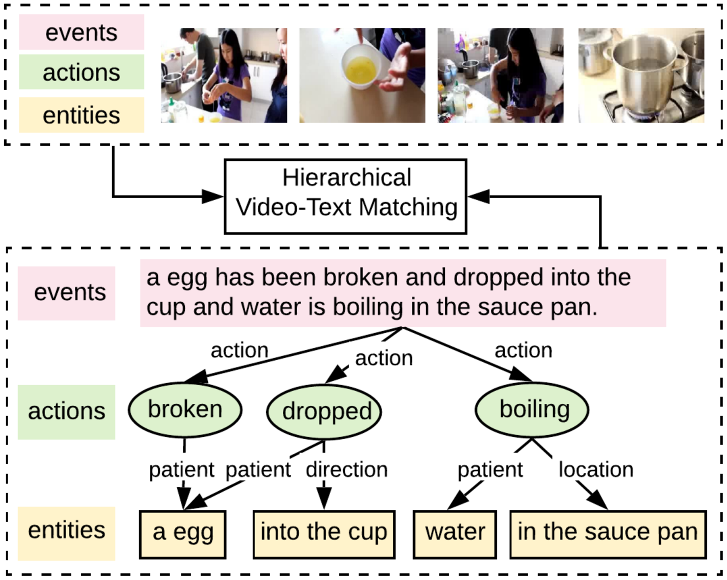
- Two-branch Matching Model
  - **Global Matching**
    - **VSE++ [2]**
    - **Dual Encoding [3]**
  - Fine-grained Matching
    - HGR [4]



[2] Vse++: Improving visual-semantic embeddings with hard negatives, BMVC, 2018
[3] Dual encoding for zero-example video retrieval, CVPR, 2019
[4] Fine-grained video-text retrieval with hierarchical graph reasoning, CVPR, 2020

Two-branch Model

- Two-branch Matching Model
  - Global Matching
    - VSE++ [2]
    - Dual Encoding [3]
  - **Fine-grained Matching**
    - **HGR [4]**



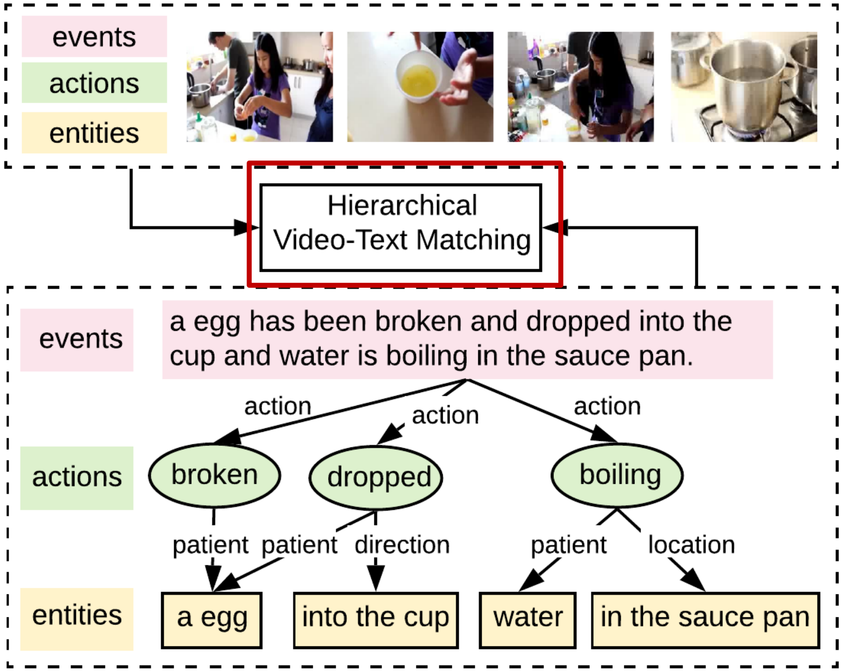
[2] Vse++: Improving visual-semantic embeddings with hard negatives, BMVC, 2018
[3] Dual encoding for zero-example video retrieval, CVPR, 2019
[4] Fine-grained video-text retrieval with hierarchical graph reasoning, CVPR, 2020

Hierarchical Graph Reasoning (HGR)

- Multi-level Video-Text Matching
  - Event
  - Actions
  - Entities

Global

Local

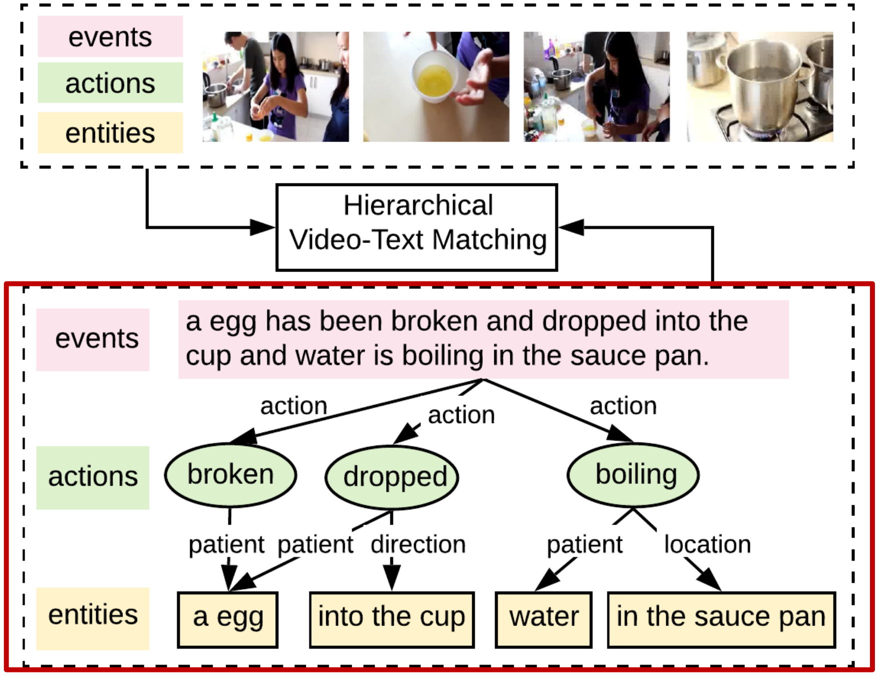


Hierarchical Graph Reasoning (HGR)

- Multi-level Video-Text Matching
  - Event
  - Actions
  - Entities

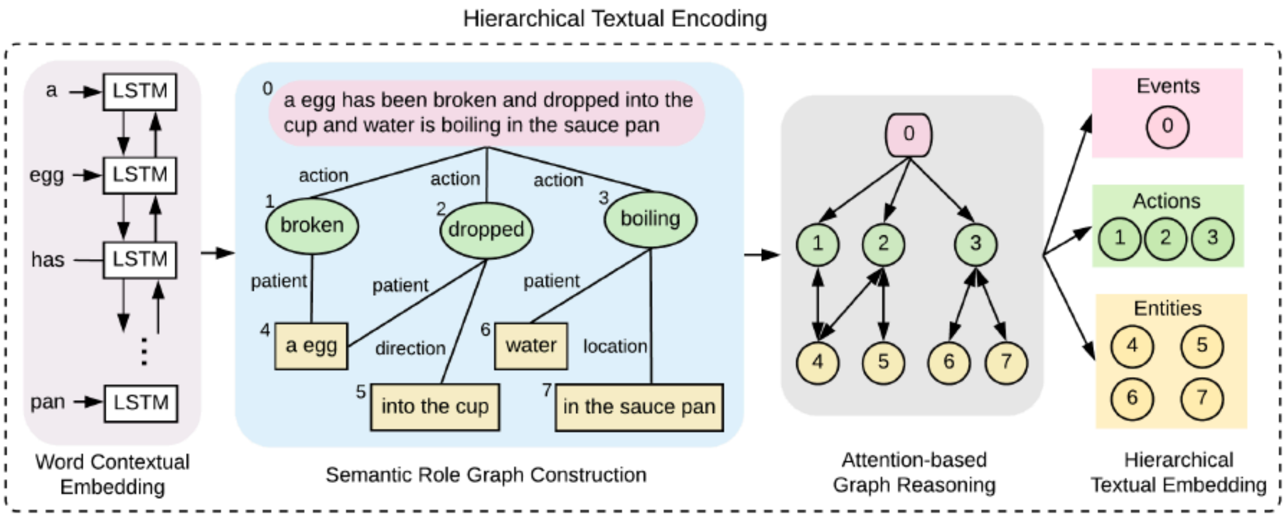
Global ↓ Local

- Hierarchical Textual Encoding
  - Decompose sentence into semantic role graph
  - Capture relationships via graph reasoning



Hierarchical Graph Reasoning (HGR)

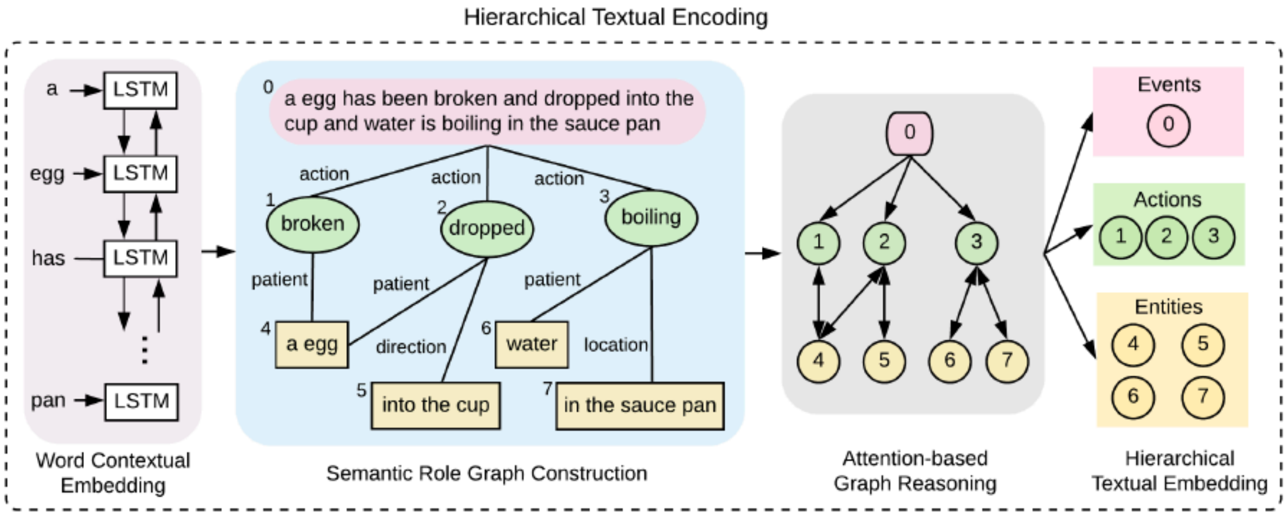
- Textual Graph Construction
  - Event node: the whole text query
  - Action node: verbs in the text
  - Entity node: noun phrases in the text



Hierarchical Textual Encoding
a
egg
has
pan
LSTM
LSTM
LSTM
LSTM
Word Contextual Embedding
0
a egg has been broken and dropped into the cup and water is boiling in the sauce pan
action
action
action
1
broken
2
dropped
3
boiling
patient
patient
patient
4
a egg
direction
6
water
location
5
into the cup
7
in the sauce pan
Semantic Role Graph Construction
0
1
2
3
4
5
6
7
Attention-based Graph Reasoning
Events
0
Actions
1
2
3
Entities
4
5
6
7
Hierarchical Textual Embedding

Hierarchical Graph Reasoning (HGR)

- Textual Graph Construction
  - Event node: the whole text query
  - Action node: verbs in the text
  - Entity node: noun phrases in the text
- Attentive Graph Reasoning —— Relational GCN


Hierarchical Textual Encoding
a
egg
has
pan
LSTM
LSTM
LSTM
LSTM
Word Contextual Embedding
0 a egg has been broken and dropped into the cup and water is boiling in the sauce pan
action
action
action
1 broken
2 dropped
3 boiling
patient
patient
patient
4 a egg
direction
6 water
location
5 into the cup
7 in the sauce pan
Semantic Role Graph Construction
0
1
2
3
4
5
6
7
Attention-based Graph Reasoning
Events
0
Actions
1 2 3
Entities
4 5
6 7
Hierarchical Textual Embedding

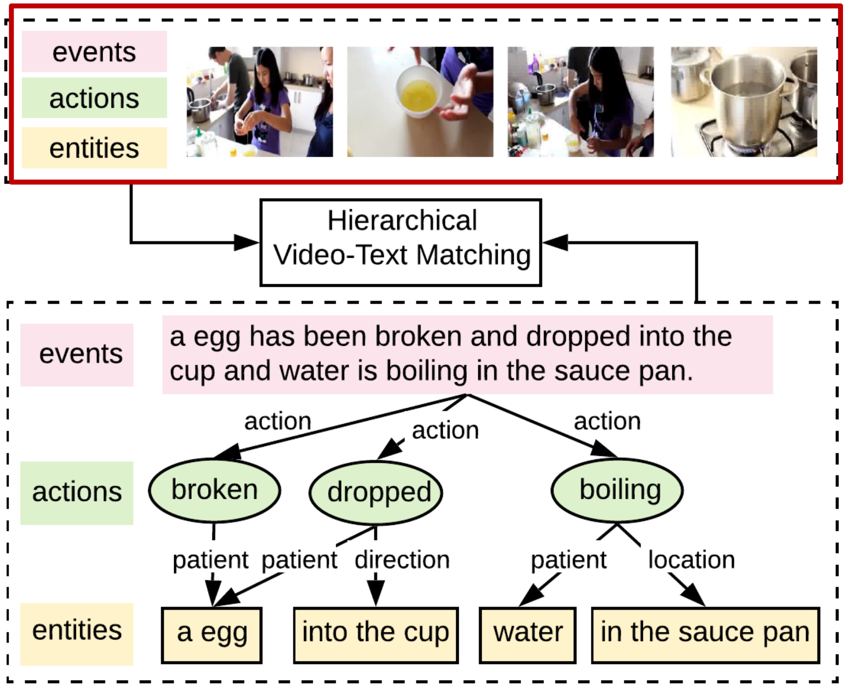
Hierarchical Graph Reasoning (HGR)

- Multi-level Video-Text Matching
  - Event
  - Actions
  - Entities

  Global ↓ **Local**

- Hierarchical Textual Encoding
  - Decompose sentence into semantic role graph
  - Capture relationships via graph reasoning

- Hierarchical Video Encoding
  - Guided by different levels of text to learn diverse video representations

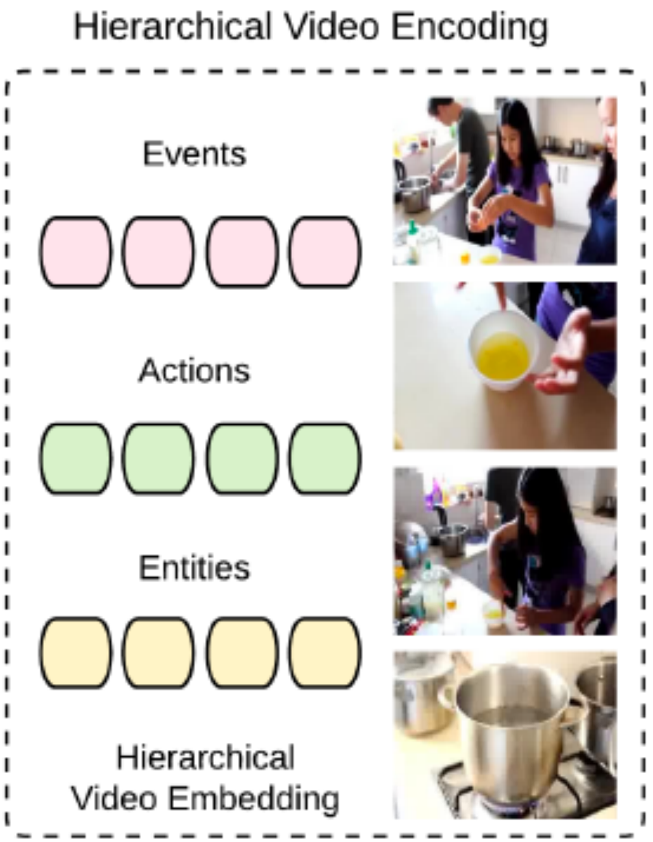


Hierarchical Graph Reasoning (HGR)

- Video Encoding

$$V = \{f_1, \dots, f_M\}$$

$$v_{x,i} = W_x^v f_i, \quad x \in \{e, a, o\}$$



Hierarchical Graph Reasoning (HGR)

- Global Matching at the event level

$$s_e = cos(v_e, c_e)$$

- Local Attentive Matching

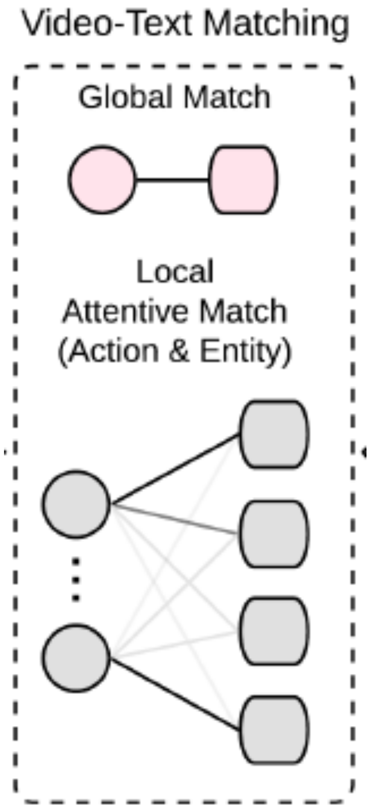
$$s_{ij}^x = cos(v_{x,j}, c_{x,i})$$

$$s_{x,i} = \sum_j \varphi_{ij}^x s_{ij}^x.$$

$$\varphi_{ij}^x = \mathrm{softmax}(\lambda([s_{ij}^x]_+ / \sqrt{\sum_j [s_{ij}^x]_+^2}))$$

$$s_x = \sum_i s_{x,i} \quad x \in \{a, o\}$$

- $s(v,c) = (s_e + s_a + s_o)/3$



Video-Text Matching Results

- **Video datasets** : TGIF, MSRVTT, VATEX

  Image dataset : MSCOCO (only for global matching models)

- **Video features** : ResNeXt-101, irCSN-152

  Image features : ResNeXt-101

Video-Text Matching Results

- Four runs for the final submission:
  - Run4: The global matching branch trained on video datasets

Table1. Results on TRECVID 2019 and 2020 AVS Main Task.

Submissions	2019	2020
Winner in 2019	0.163	-
Run4	0.177	0.354

Video-Text Matching Results

- Four runs for the final submission:
  - Run4: The global matching branch trained on video datasets
  - Run3: Run4 + global matching branch trained on image datasets

Table1. Results on TRECVID 2019 and 2020 AVS Main Task.

Submissions	2019	2020
Winner in 2019	0.163	-
Run4	0.177	0.354
Run3	0.193	0.350

Video-Text Matching Results

- Four runs for the final submission:
  - Run4: The global matching branch trained on video datasets
  - Run3: Run4 + global matching branch trained on image datasets
  - Run2: Run3 + fine-grained matching branch (HGR)

Table1. Results on TRECVID 2019 and 2020 AVS Main Task.

Submissions	2019	2020
Winner in 2019	0.163	-
Run4	0.177	0.354
Run3	0.193	0.350
Run2	0.195	0.357

Video-Text Matching Results

- Four runs for the final submission:
  - Run4: The global matching branch trained on video datasets
  - Run3: Run4 + global matching branch trained on image datasets
  - Run2: Run3 + fine-grained matching branch (HGR)
  - Run1: Run2 + BERT as text encoder

Table1. Results on TRECVID 2019 and 2020 AVS Main Task.

Submissions	2019	2020
Winner in 2019	0.163	-
Run4	0.177	0.354
Run3	0.193	0.350
Run2	0.195	0.357
Run1	**0.196**	**0.359**

Video-Text Matching Results

- Four runs for the final submission:
  - Run4: The global matching branch trained on video datasets
  - Run3: Run4 + global matching branch trained on image datasets
  - Run2: Run3 + fine-grained matching branch (HGR)
  - Run1: Run2 + BERT as text encoder

Table1. Results on TRECVID 2019 and 2020 AVS Main Task.

Submissions	2019	2020
Winner in 2019	0.163	-
Run4	0.177	0.354
Run3	0.193	0.350
Run2	0.195	0.357
Run1	**0.196**	0.359
Run5*	0.181	**0.361**

Video-Text Matching Results

- Four runs for the final submission:
  - Run4: The global matching branch trained on video datasets
  - Run3: Run4 + global matching branch trained on image datasets
  - Run2: Run3 + fine-grained matching branch (HGR)
  - Run1: Run2 + BERT as text encoder

Table2. Results on TRECVID AVS Progress Subtask.

Submissions	Results
Winner in 2019	0.177
Run4	**0.235**
Run3	0.208
Run2	0.220
Run1	0.223

Take Home Message

- We propose a two-branch model by combing the global matching and fine-grained matching for the AVS task

Take Home Message

- We propose a two-branch model by combing the global matching and fine-grained matching for the AVS task

- Training on additional image captioning dataset can improve the retrieval performance on 2019 AVS task, but not on 2020 AVS task

Take Home Message

- We propose a two-branch model by combing the global matching and fine-grained matching for the AVS task
- Training on additional image captioning dataset can improve the retrieval performance on 2019 AVS task, but not on 2020 AVS task
- Our models rank the 1st place on the TRECVID 2020 AVS Main Task.

THANKS ！

If you have any questions , please feel free to contact with us:
zyiday@ruc.edu.cn, syuqing@ruc.edu.cn,
cszhe1@ruc.edu.cn, qjin@ruc.edu.cn

http://jin-qin.com/AIM3-Lab.html