RUC_AIM3 at TRECVID 2020: Video to Text Description

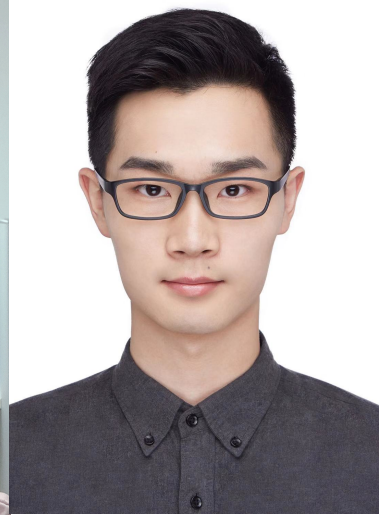
Yuqing Song, Yida Zhao, Shizhe Chen, Qin Jin

AI·M^3 , Renmin University of China

Team Members

RUC_AIM3 Team

Yuqing Song



Yida Zhao

Shizhe Chen

Qin Jin

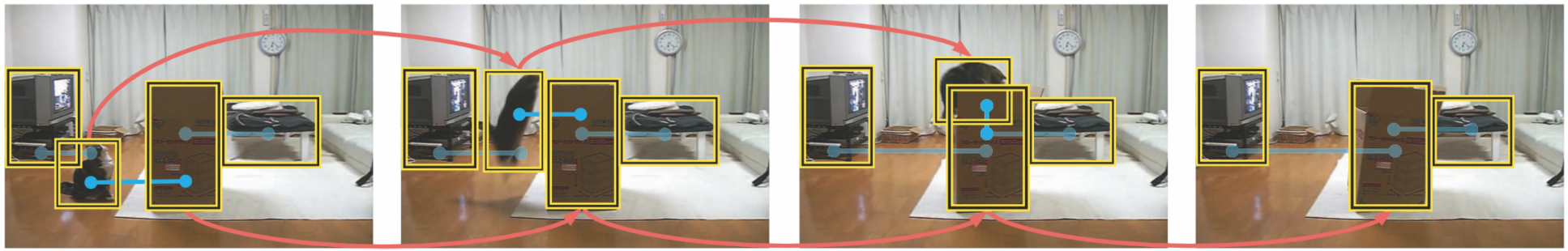
Outline

- Video Description Generation
- Video-Text Matching
- Conclusions

One

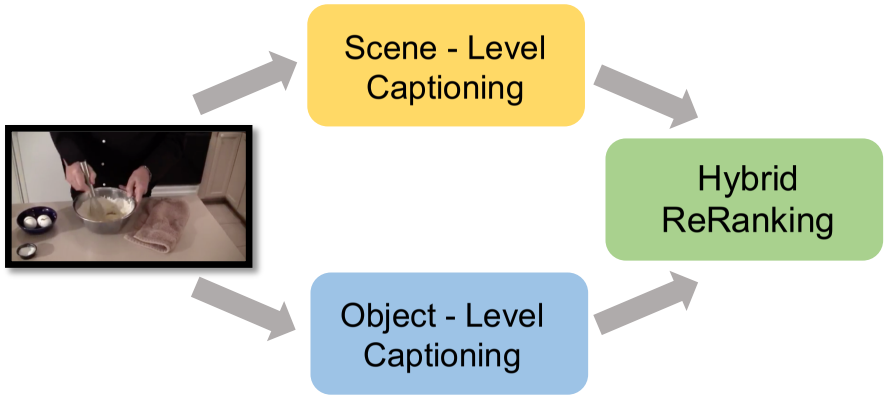
Video Description Generation

Video Description Generation



"A cat jumps into a box."

- Modeling semantic information from both temporal and spatial dimensions
- Combining the scene-level and object-level captions via hybrid reranking



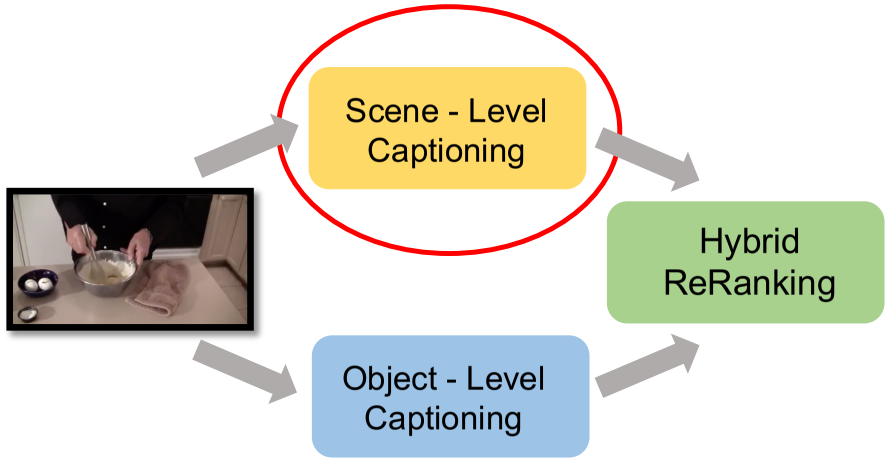
Video Description Generation

Scene-Level Captioning

- Scene-level video feature

$$V^S = \{v_1^S, v_2^S, \dots v_n^S\}$$

$$v_i^T = [\text{ResNeXt-101}(seg_i); \text{irCSN}(seg_i)]$$



Video Description Generation

Scene-Level Captioning

- Scene-level video feature

$$V^S = \{v_1^S, v_2^S, \dots v_n^S\}$$

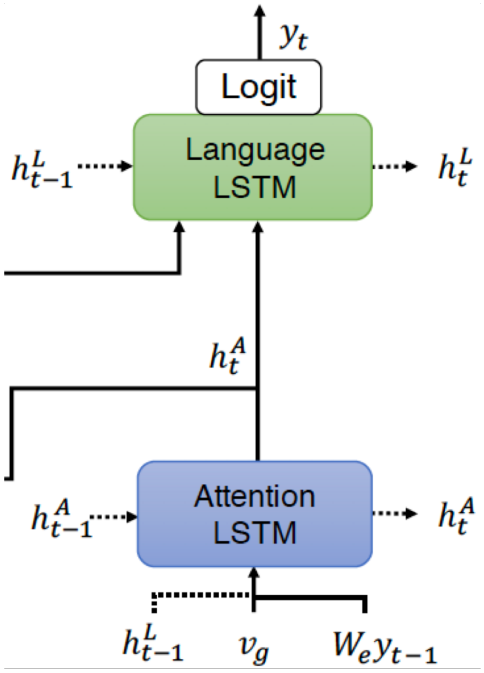
$$v_i^S = [\text{ResNeXt-101}(seg_i); \text{irCSN}(seg_i)]$$

- Language decoder with temporal attention

$$h_t^A = LSTM([v_g, W_e y_{t-1}, h_{t-1}^L], h_{t-1}^A; \theta^A)$$

$$z_t^S = softmax(h_t^A W V^S) V^S$$

$$h_t^L = LSTM([z_t^S, h_{t-1}^A], h_{t-1}^L; \theta^L)$$



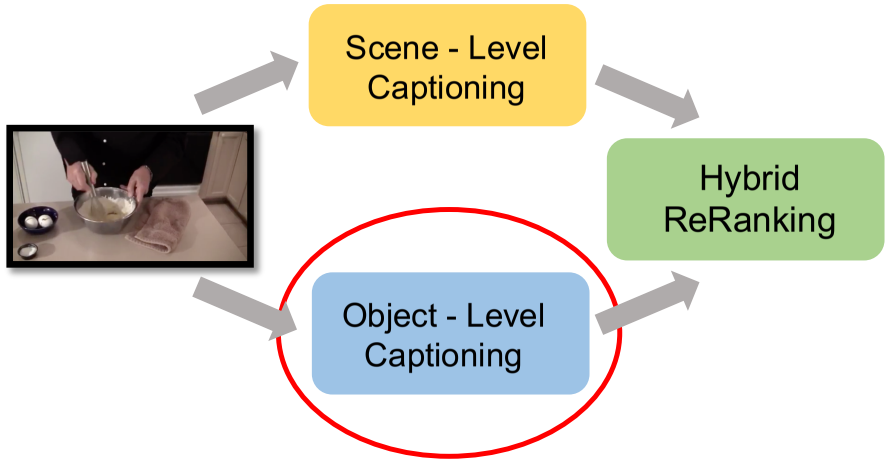
Video Description Generation

Object-Level Captioning

- Object-level video feature

$$V^O = \{v_1^O, v_2^O, \dots v_k^O\}$$

$$v_i^O = \text{Faster-RCNN(frame)}$$



Video Description Generation

Object-Level Captioning

- Object-level video feature

$$V^O = \{v_1^O, v_2^O, \dots v_k^O\}$$

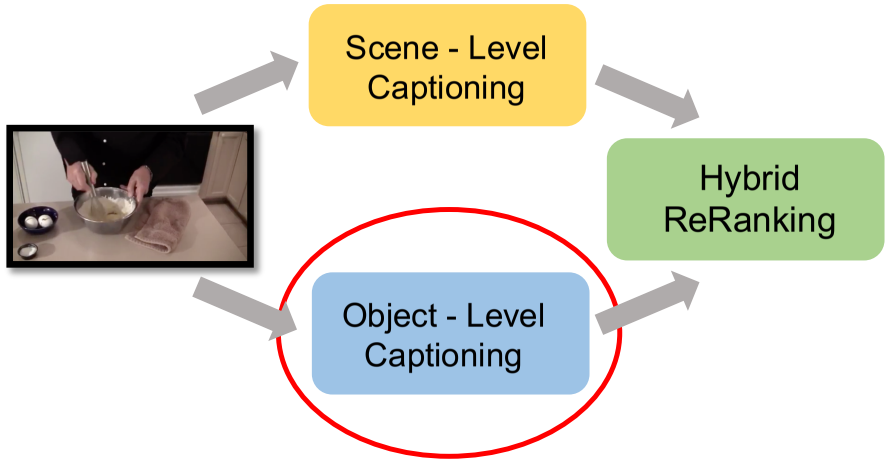
$$v_i^O = \text{Faster-RCNN(frame)}$$

- Language decoder with spatial attention

$$h_t^A = LSTM([v_g, W_e y_{t-1}, h_{t-1}^L], h_{t-1}^A; \theta^A)$$

$$z_t^O = softmax(h_t^A W V^O) V^O$$

$$h_t^L = LSTM([z_t^O, h_{t-1}^A], h_{t-1}^L; \theta^L)$$



Video Description Generation

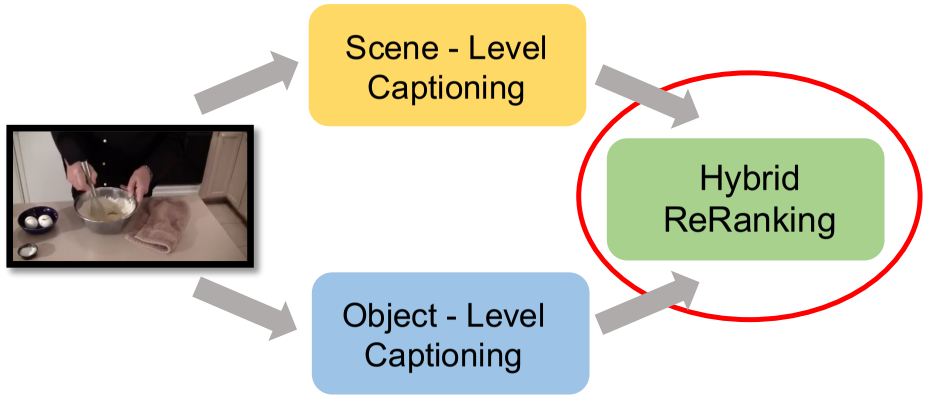
Hybrid ReRanking

- Scene-level video descriptions

  More suitable for videos with various action changes

- Object-level video descriptions

  More suitable for videos containing rich objects



Video Description Generation

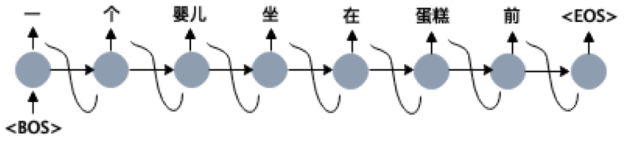
Hybrid ReRanking

- Language fluency score

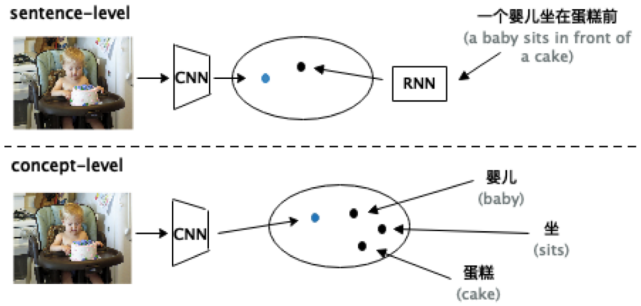
  The averaged log probabilities of each words generation by the pre-trained language model

- Content relevance score

  More suitable for videos containing rich objects



a) Language Model



b) Video-Text Matching Model

Video Description Generation Results

- **Training dataset** : TGIF, TRECVID16-18, MSRVTT, VATEX

  Validation dataset: TRECVID19

Video Description Generation Results

- **Training dataset** : TGIF, TRECVID16-18, MSRVTT, VATEX

  Validation dataset: TRECVID19

  Q1: What model architecture is suitable for short video captioning?

Video Description Generation Results

- **Training dataset** : TGIF, TRECVID16-18, MSRVTT, VATEX

  Validation dataset: TRECVID19

  Q1: What model architecture is suitable for short video captioning?

  - ✓ AoANet [1]
  - ✓ X-LAN [2]
  - ✓ Transformer [3]
  - ✓ Our up-down model

[1] Attention on attention for image captioning, ICCV, 2019
[2] X-linear attention networks for image captioning, CVPR, 2020
[3] Attention is all you need, NIPS, 2017

Video Description Generation Results

- **Training dataset** : TGIF, TRECVID16-18, MSRVTT, VATEX

  Validation dataset: TRECVID19

  Q1: What model architecture is suitable for short video captioning?

Models	BLEU@1	BLEU@2	BLEU@3	BLEU@4	CIDEr	METEOR	SPICE
Trained with Cross-Entropy Loss							
AoANet	83.20	69.88	55.98	42.73	51.97	29.54	7.01
X-LAN	84.07	**71.95**	**58.94**	**46.65**	**58.63**	30.48	7.41
Transformer	**85.10**	70.98	56.78	43.63	51.61	**30.94**	7.60
Ours	82.14	67.58	53.08	40.68	53.32	30.50	**7.87**
Trained with Reinforcement Learning							
AoANet	85.83	71.89	57.15	43.62	60.39	30.41	7.71
X-LAN	**87.96**	**74.91**	**60.77**	**47.46**	59.69	**31.68**	7.92
Transformer	85.51	71.02	55.72	42.11	53.88	30.08	7.85
Ours	87.89	74.42	59.54	45.61	**62.06**	31.41	**8.00**

Video Description Generation Results

- **Training dataset** : TGIF, TRECVID16-18, MSRVTT, VATEX

  Validation dataset: TRECVID19

  Q1: What model architecture is suitable for short video captioning?

Models	BLEU@1	BLEU@2	BLEU@3	BLEU@4	CIDEr	METEOR	SPICE
Trained with Cross-Entropy Loss							
AoANet	83.20	69.88	55.98	42.73	51.97	29.54	7.01
X-LAN	84.07	**71.95**	**58.94**	**46.65**	**58.63**	30.48	7.41
Transformer	**85.10**	70.98	56.78	43.63	51.61	**30.94**	7.60
Ours	82.14	67.58	53.08	40.68	53.32	30.50	**7.87**
Trained with Reinforcement Learning							
AoANet	85.83	71.89	57.15	43.62	60.39	30.41	7.71
X-LAN	**87.96**	**74.91**	**60.77**	**47.46**	59.69	**31.68**	7.92
Transformer	85.51	71.02	55.72	42.11	53.88	30.08	7.85
Ours	87.89	74.42	59.54	45.61	**62.06**	31.41	**8.00**

Video Description Generation Results

- **Training dataset** : TGIF, TRECVID16-18, MSRVTT, VATEX

  Validation dataset: TRECVID19

  Q1: What model architecture is suitable for short video captioning?

Models	BLEU@1	BLEU@2	BLEU@3	BLEU@4	CIDEr	METEOR	SPICE
Trained with Cross-Entropy Loss							
AoANet	83.20	69.88	55.98	42.73	51.97	29.54	7.01
X-LAN	84.07	**71.95**	**58.94**	**46.65**	**58.63**	30.48	7.41
Transformer	**85.10**	70.98	56.78	43.63	51.61	**30.94**	7.60
Ours	82.14	67.58	53.08	40.68	53.32	30.50	**7.87**
Trained with Reinforcement Learning							
AoANet	85.83	71.89	57.15	43.62	60.39	30.41	7.71
X-LAN	**87.96**	**74.91**	**60.77**	**47.46**	59.69	**31.68**	7.92
Transformer	85.51	71.02	55.72	42.11	53.88	30.08	7.85
Ours	87.89	74.42	59.54	45.61	**62.06**	31.41	**8.00**

Video Description Generation Results

Q2: Which is more important for video description generation, temporal information or spatial information?

Video Description Generation Results

Q2: Which is more important for video description generation, temporal information or spatial information?

Table 5: Results with different visual features on TRECVID VTT 2019 dataset.

Models	Loss	BLEU@1	BLEU@2	BLEU@3	BLEU@4	CIDEr	METEOR	SPICE
Trained with Scene-level Video Features								
X-LAN	XE	57.67	39.87	26.35	16.78	30.23	16.16	10.94
Ours	XE	59.53	38.93	24.81	15.47	30.29	15.47	10.71
X-LAN	RL	**66.67**	**45.13**	29.21	18.13	36.01	17.30	11.60
Ours	**RL**	66.52	45.01	**29.24**	**18.19**	**36.30**	**17.37**	**11.63**
Trained with Object-level Video Features								
X-LAN	XE	57.81	39.55	26.06	16.62	28.85	15.95	10.64
Ours	XE	61.18	40.02	25.50	15.80	32.17	17.00	**11.64**
X-LAN	RL	65.85	44.40	28.64	17.78	32.96	17.02	11.18
Ours	**RL**	**65.87**	**44.84**	**29.13**	**18.04**	**35.15**	**17.29**	11.62
Hybrid Reranking								
Ours	**RL**	**67.75**	**46.48**	**30.30**	**18.80**	**38.45**	**17.96**	**12.32**

Video Description Generation Results

Q2: Which is more important for video description generation, temporal information or spatial information?

Table 5: Results with different visual features on TRECVID VTT 2019 dataset.

Models	Loss	BLEU@1	BLEU@2	BLEU@3	BLEU@4	CIDEr	METEOR	SPICE
Trained with Scene-level Video Features								
X-LAN	XE	57.67	39.87	26.35	16.78	30.23	16.16	10.94
Ours	XE	59.53	38.93	24.81	15.47	30.29	15.47	10.71
X-LAN	RL	**66.67**	**45.13**	29.21	18.13	36.01	17.30	11.60
Ours	**RL**	66.52	45.01	**29.24**	**18.19**	**36.30**	**17.37**	**11.63**
Trained with Object-level Video Features								
X-LAN	XE	57.81	39.55	26.06	16.62	28.85	15.95	10.64
Ours	XE	61.18	40.02	25.50	15.80	32.17	17.00	**11.64**
X-LAN	RL	65.85	44.40	28.64	17.78	32.96	17.02	11.18
Ours	**RL**	**65.87**	**44.84**	**29.13**	**18.04**	**35.15**	**17.29**	11.62
Hybrid Reranking								
Ours	**RL**	**67.75**	**46.48**	**30.30**	**18.80**	**38.45**	**17.96**	**12.32**

Video Description Generation Results

Q2: Which is more important for video description generation, temporal information or spatial information?

Table 5: Results with different visual features on TRECVID VTT 2019 dataset.

Models	Loss	BLEU@1	BLEU@2	BLEU@3	BLEU@4	CIDEr	METEOR	SPICE
Trained with Scene-level Video Features								
X-LAN	XE	57.67	39.87	26.35	16.78	30.23	16.16	10.94
Ours	XE	59.53	38.93	24.81	15.47	30.29	15.47	10.71
X-LAN	RL	**66.67**	**45.13**	29.21	18.13	36.01	17.30	11.60
Ours	**RL**	66.52	45.01	**29.24**	**18.19**	**36.30**	**17.37**	**11.63**
Trained with Object-level Video Features								
X-LAN	XE	57.81	39.55	26.06	16.62	28.85	15.95	10.64
Ours	XE	61.18	40.02	25.50	15.80	32.17	17.00	**11.64**
X-LAN	RL	65.85	44.40	28.64	17.78	32.96	17.02	11.18
Ours	**RL**	**65.87**	**44.84**	**29.13**	**18.04**	**35.15**	**17.29**	11.62
Hybrid Reranking								
Ours	**RL**	**67.75**	**46.48**	**30.30**	**18.80**	**38.45**	**17.96**	**12.32**

Video Description Generation Results

Q2: Which is more important for video description generation, temporal information or spatial information?

Table 5: Results with different visual features on TRECVID VTT 2019 dataset.

Models	Loss	BLEU@1	BLEU@2	BLEU@3	BLEU@4	CIDEr	METEOR	SPICE
Trained with Scene-level Video Features								
X-LAN	XE	57.67	39.87	26.35	16.78	30.23	16.16	10.94
Ours	XE	59.53	38.93	24.81	15.47	30.29	15.47	10.71
X-LAN	RL	**66.67**	**45.13**	29.21	18.13	36.01	17.30	11.60
Ours	**RL**	66.52	45.01	**29.24**	**18.19**	**36.30**	**17.37**	**11.63**
Trained with Object-level Video Features								
X-LAN	XE	57.81	39.55	26.06	16.62	28.85	15.95	10.64
Ours	XE	61.18	40.02	25.50	15.80	32.17	17.00	**11.64**
X-LAN	RL	65.85	44.40	28.64	17.78	32.96	17.02	11.18
Ours	**RL**	**65.87**	**44.84**	**29.13**	**18.04**	**35.15**	**17.29**	11.62
Hybrid Reranking								
Ours	**RL**	**67.75**	**46.48**	**30.30**	**18.80**	**38.45**	**17.96**	**12.32**

Video Description Generation Results

- Four runs for the final submission:
  - Run4: Our single best model

Table3. Results of TRECVID 2020 VTT description generation subtask.

Submissions	CIDEr
Run4	0.284

Video Description Generation Results

- Four runs for the final submission:
  - Run4: Our single best model
  - Run3: Ensemble of the captioning models trained on object-level visual features

Table3. Results of TRECVID 2020 VTT description generation subtask.

Submissions	CIDEr
Run4	0.284
Run3	0.277

Video Description Generation Results

- Four runs for the final submission:
  - Run4: Our single best model
  - Run3: Ensemble of the captioning models trained on object-level visual features
  - Run2: Ensemble of the captioning models trained on scene-level visual features

Table3. Results of TRECVID 2020 VTT description generation subtask.

Submissions	CIDEr
Run4	0.284
Run3	0.277
Run2	0.289

Video Description Generation Results

- Four runs for the final submission:
  - Run4: Our single best model
  - Run3: Ensemble of the captioning models trained on object-level visual features
  - Run2: Ensemble of the captioning models trained on scene-level visual features
  - Run1: Ensemble of run2 and run3 via hybrid reranking

Table3. Results of TRECVID 2020 VTT description generation subtask.

Submissions	CIDEr
Run4	0.284
Run3	0.277
Run2	0.289
Run1	**0.303**

Two

Video-Text Matching

Video-Text Matching

Text pool

a man wearing goggles is swimming in a pool.

a red race car marked with the number seven is driven by a man in a blue and beige shirt.

a brown dog is bending down trying to drink from a jet of water.

a young child is throwing a big rock into a body of water.

Video pool

Video-Text Matching

- Global Matching Branch
  - VSE++ [4]
  - Dual-Encoding [5]
- Fine-grained Matching Branch
  - HGR [6]

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

Video-Text Matching

- Scaled matching score for “hub” texts [7]

[7] A strong and robust baseline for text-image matching, ACL, 2019

Video-Text Matching

- Scaled matching score for “hub” texts [7]

$$s'(v,t) = \frac{e^{\beta s(v,t)}}{\sum_{\overline{v} \in V \setminus \{v\}} e^{\beta s(\overline{v},t)}}$$

The text t is a “hub” text, which is close to multiple video queries

[7] A strong and robust baseline for text-image matching, ACL, 2019

Video-Text Matching Results

- Four runs for the final submission:
  - Run4: Dual-Encoding model with hubness optimization

Table4. Results of TRECVID 2020 VTT matching and ranking subtask.

Submissions	SetA	SetB	SetC	SetD	SetE
Run4	0.606	0.611	0.621	0.618	0.636

Video-Text Matching Results

- Four runs for the final submission:
  - Run4: Dual-Encoding model with hubness optimization
  - Run3: HGR model with hubness optimization

Table4. Results of TRECVID 2020 VTT matching and ranking subtask.

Submissions	SetA	SetB	SetC	SetD	SetE
Run4	0.606	0.611	0.621	0.618	0.636
Run3	0.627	0.621	0.620	0.620	0.641

Video-Text Matching Results

- Four runs for the final submission:
  - Run4: Dual-Encoding model with hubness optimization
  - Run3: HGR model with hubness optimization
  - Run2: Ensemble of global matching models with hubness optimization

Table4. Results of TRECVID 2020 VTT matching and ranking subtask.

Submissions	SetA	SetB	SetC	SetD	SetE
Run4	0.606	0.611	0.621	0.618	0.636
Run3	0.627	0.621	0.620	0.620	0.641
Run2	0.683	0.692	0.691	0.696	0.711

Video-Text Matching Results

- Four runs for the final submission:
  - Run4: Dual-Encoding model with hubness optimization
  - Run3: HGR model with hubness optimization
  - Run2: Ensemble of global matching models with hubness optimization
  - Run1: Ensemble of global and fine-grained matching models with hubness optimization

Table4. Results of TRECVID 2020 VTT matching and ranking subtask.

Submissions	SetA	SetB	SetC	SetD	SetE
Run4	0.606	0.611	0.621	0.618	0.636
Run3	0.627	0.621	0.620	0.620	0.641
Run2	0.683	0.692	0.691	0.696	0.711
Run1	**0.714**	**0.711**	**0.707**	**0.721**	**0.731**

Take Home Message

- The LSTM-based models are more suitable for short video captioning.
  Temporal information is more important than spatial information in video captioning.
  Combining temporal and spatial attention achieves the best results.

Take Home Message

- The LSTM-based models are more suitable for short video captioning.
  Temporal information is more important than spatial information in video captioning.
  Combining temporal and spatial attention achieves the best results.

- Fine-grained matching is better than global matching for video-text retrieval.
  Combining the global matching models and fine-grained matching models achieves the best results.

Take Home Message

- The LSTM-based models are more suitable for short video captioning.
  Temporal information is more important than spatial information in video captioning.
  Combining temporal and spatial attention achieves the best results.

- Fine-grained matching is better than global matching for video-text retrieval.
  Combining the global matching models and fine-grained matching models achieves the best results.

- Our models rank the 1st place on both video description generation and video-text matching subtasks in TRECVID 2020 VTT task.

AI · M³

中国人民大学多媒体计算实验室

THANKS !

If you have any questions , please feel free to contact with us:

syuqing@ruc.edu.cn, zyiday@ruc.edu.cn, cszhe1@ruc.edu.cn, qjin@ruc.edu.cn

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