RUC_AIM3 at TRECVID 2023:

Video-to-Text Description

Kaiwen Wei, Zihao Yue, Liang Zhang, Qin Jin
Renmin University of China
Nov.14, 2023

Video-to-Text Description (VTT)

- automatically generate a single-sentence description in natural language for a given video.

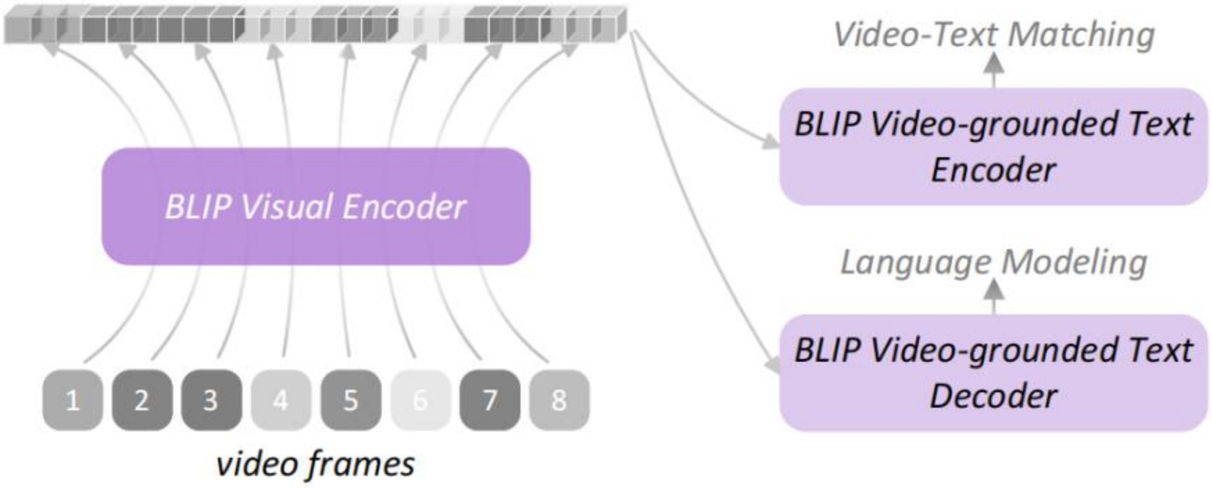
- *An Asian man playing an electronic guitar in an indoor setting.*

Effective Approach:

- Transferring image-text model pre-training model to VTT task.

Our last year's solution:

- BLIP4video: Transferring BLIP to VTT task
- Data Augmentation to obtain sufficient data
- Re-ranking for best candidate selection
- best CIDEr: 60.2, ranking 1st



Vision-Language Pre-training Model:

- Image-text model:

  lack of temporal modeling ability

- Video-text model:

  swinBERT(Lin, et.al.), mPLUG-2(Xu, et.al. 2023)

  obtain temporal modeling ability from large-scale video-text pre-training

We select mPLUG-2 (SOTA captioning model) this year.

mPLUG-2

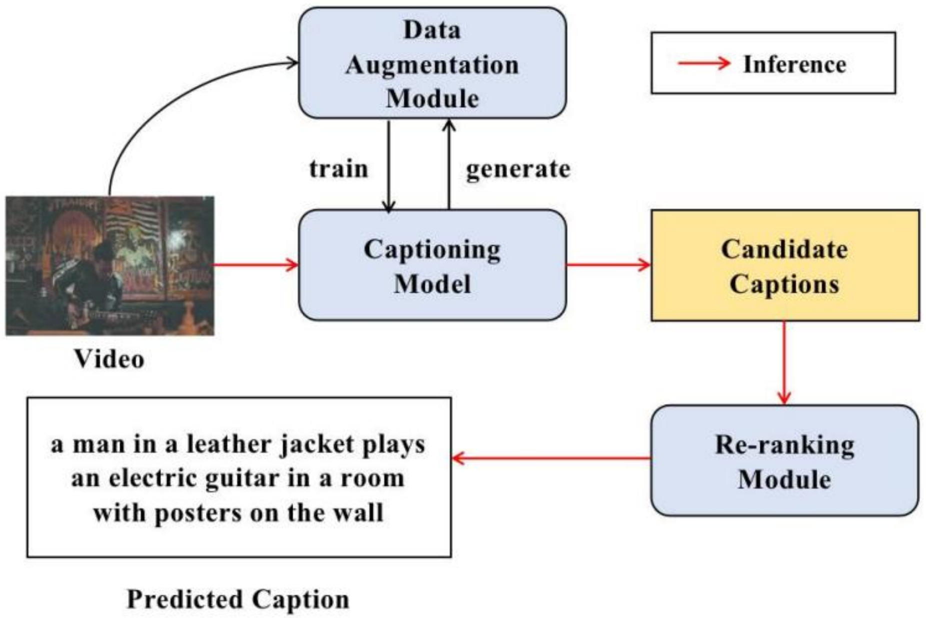
captioning model

Data Augmentation Module

improving both quality and quantity of data

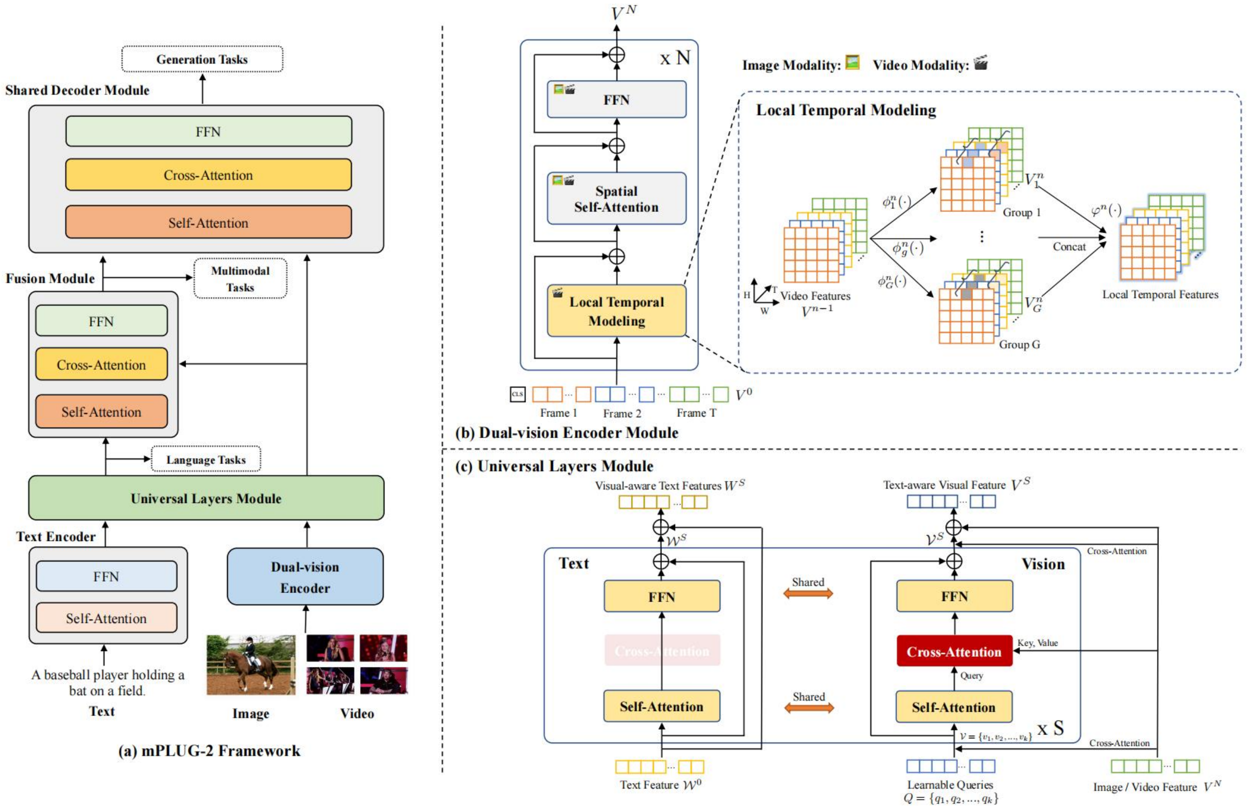
Re-ranking Module

best candidates selection



Methodology - mPLUG-2

mPLUG-2: Modularized Multi-modal model across video, image and text.



Pre-trained with 2.5M video-text pairs and 14M image-text pairs.

A spatio-temporal modeling module to capture temporal information.

Comparison between captioning models:

- **Zero-shot:**

  mPLUG-2 significantly outperforms BLIP4video.

Approach	Model	CIDEr
Zero-shot	BLIP4video	28.9
	mPLUG-2	**44.8**
Fine-tuned	BLIP4video	50.5
	mPLUG-2	54.4
	mPLUG-2 + SCST	**57.1**

- **Supervised Fine-tuning:**

  Settings:

  - training set: VTT16-21
  - validation set: VTT22

  mPLUG-2 still outperforms BLIP4video.

  SCST further improve the performance.

Methodology - Data Augmentation

Back Translation:

English ➡ Chinese ➡ English

Pseudo Labeling:

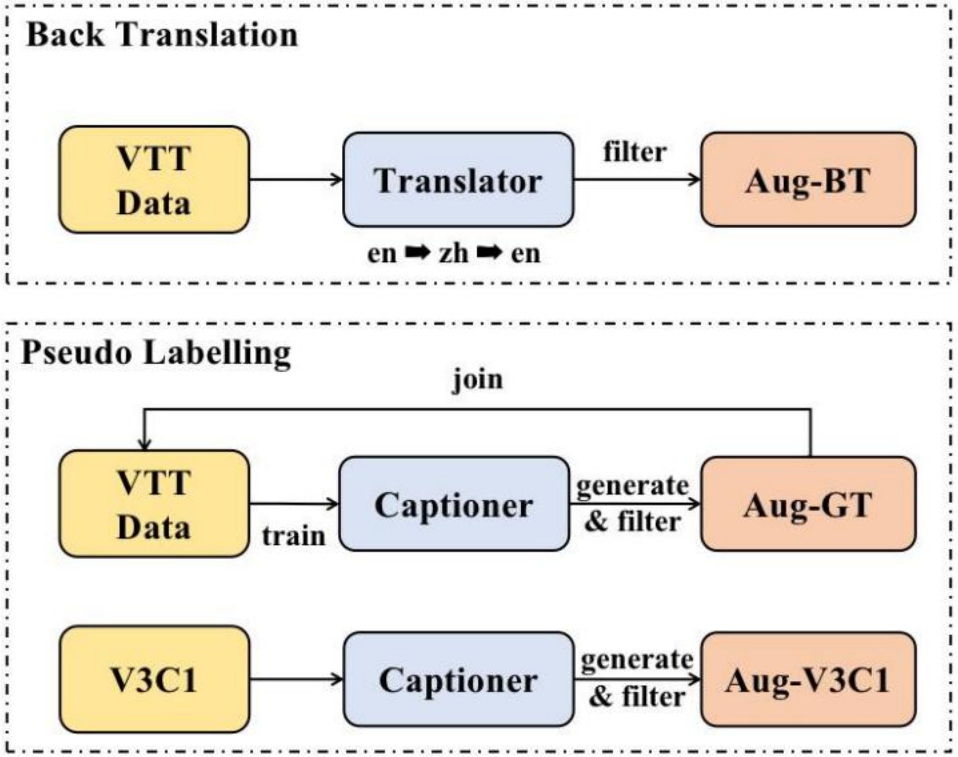
- For VTT videos

  Cycle the following procedures:

  - Generate pseudo descriptions by a captioner
  - Filter
  - Add to the training data
  - Re-train the captioner

- For V3C1 videos

  Generate descriptions for V3C1 by our finalized captioner



Pseudo labeling:

Round 1:

- Train Cap-0 by VTT16-21 using CE loss and SCST.
- Generate captions for VTT videos by Cap-0.

Round 2:

- Train Cap-1 by Aug-1 using CE loss.
- Generate captions for VTT videos by Cap-1.

Round 3:

- Train Cap-2 by Aug-1 using CE loss and SCST.
- Generate captions for V3C1 and VTT videos by Cap-2.

Table 2: Training data of our 4 captioners. CE refers to cross-entropy, and SCST refers to self-critical sequence training.

Model	Training data CE	Training data SCST
Cap-0	VTT16-21	VTT18-21
Cap-1	*Aug-1*	-
Cap-2	*Aug-1*	VTT18-21
Cap-3	*Aug-2*	VTT18-21

Augment	Data	Description
Aug-1	VTT16-21	VTT data from 2016 to 2021
	Aug-22	Augmentation data from Yue et al. (2022)
	Aug-GT-1	Pseudo labeling for VTT16-21 by *Cap-0*
Aug-2	VTT-22	VTT data 2022
	Aug-BT	Back translation for VTT16-21
	Aug-GT-2	Pseudo labeling for VTT16-21 by *Cap-1*
	Aug-GT-3	Pseudo labeling for VTT16-21 by *Cap-2*
	Aug-V3C1	Pseudo labeling for V3C1 by *Cap-2*

Our augmented data:

Filter:

- CIDEr score for VTT videos.
- VTM (Video-Text Matching) for V3C1 videos.

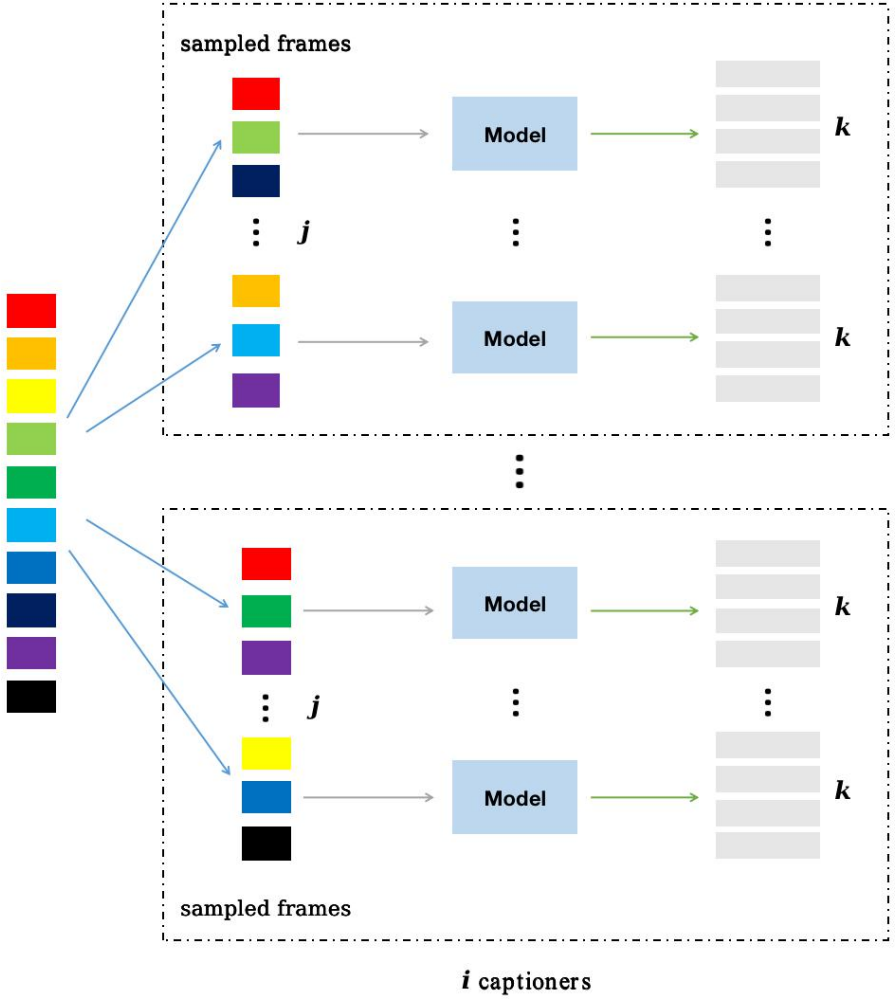
Aug-2: More training data filtered by higher threshold.

Data	*Aug-1*		*Aug-2*	
	Count	Filter	Count	Filter
VTT16-21	45,820	-	51,820	-
Aug-22	34,660	CIDEr > 55	8,902	CIDEr > 80
Aug-GT	5,598	CIDEr > 55	13,220	CIDEr > 80
Aug-V3C1	-	-	4,392	VTM > 60
Aug-BT	-	-	12,860	CIDEr > 80
Total	86,078	-	91,194	-

Model	VTT Data	Aug-GT	Aug-BT	Aug-V3C1	CIDEr
Cap0	✓				57.1
Cap2	✓	✓			59.5
Cap2+	✓	✓		✓	59.6
Cap2+	✓	✓	✓		60.0
Cap3	✓	✓	✓	✓	**61.0**

Ablation study of data augmentation:

- Pseudo Labeling on VTT data helps a lot.
- Back Translation and Pseudo labeling on V3C1 also helps.
- Pooling all the augmentation data leads to the best performance.



$i \times j \times k$ candidates per video.

- i distinct captioning models
- j different random frame selection methods(TSN) for each video
- k sentences for each selection method

How to select the best caption from candidates?

Measure:

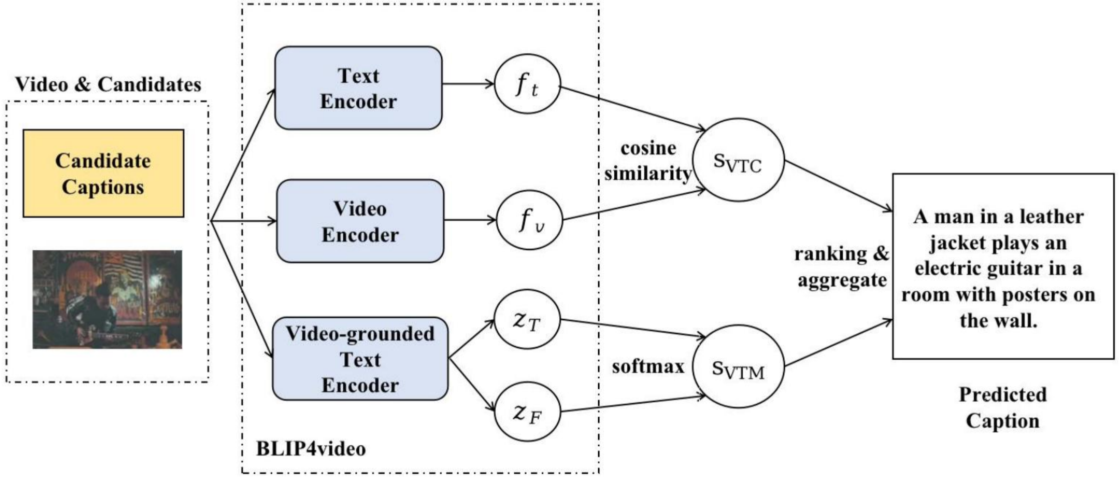
- VTM (Video-Text Matching)

  Fine-grained video-text alignment

- VTC (Video-Text Contrastive)

  Overall video-text alignment

Combine VTM and VTC for re-ranking. (Lowest ranking sum of VTM and VTC)



Our submissions:

4 runs: different captioners, different re-ranking strategies

C: CIDEr, B@4: BLEU@4, M:METEOR, SP:SPICE, ST:STS

Submission	Captioner		Re-ranking		Main Task					Robust Task				
	Cap-2	Cap-3	VTM	VTC	C	B@4	M	SP	ST	C	B@4	M	SP	ST
run1	✓		✓	✓	38.4	9.21	32.81	14.9	47.0	38.9	9.41	33.05	14.8	20.52
run2		✓	✓	✓	**39.4**	9.45	**33.25**	**15.2**	47.3	38.6	9.68	33.04	15.0	20.50
run3	✓	✓		✓	**39.4**	**9.48**	33.19	15.1	47.3	38.4	9.72	33.15	14.9	20.36
run4	✓	✓	✓	✓	**39.4**	**9.48**	33.16	**15.2**	**47.4**	**39.0**	**9.83**	**33.24**	**15.1**	**20.61**

- Run4 stands out marginally.
- Basically the same performance on the main task and robust task.
    - video input augmentation

Result

A view of a snow covered mountain from a paraglider on a sunny day in the mountains.

A white cat with blue eyes is sitting on a fence in a dark room with a green fence.

A person in a purple shirt is bouncing on the trampoline on a park on a sunny day.

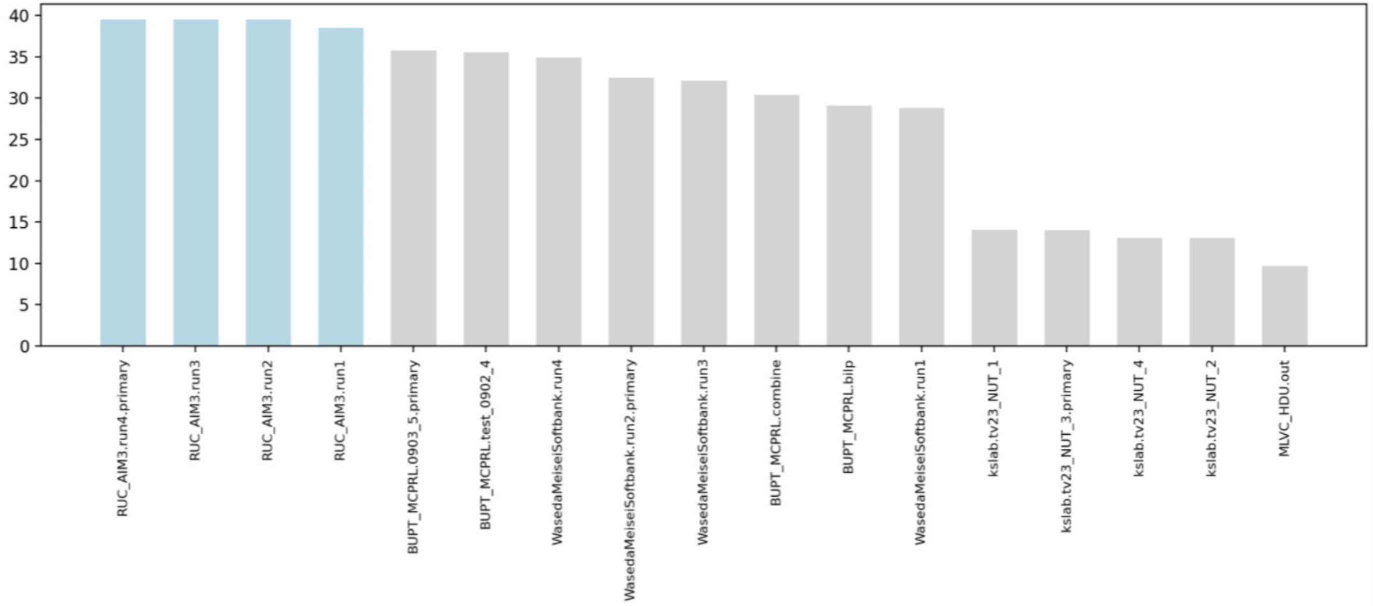
Conclusion

RUC_AIM3's solution for VTT

- Video-text model (mPLUG-2) achieves better performance on VTT task.
- We further improve our data augmentation and re-ranking strategies.

CIDEr rank 1st on the main task and robustness sub-task.

Best CIDEr score: 39.4



Limitations

Our model:

Lack detailed desription.

- More detailed test set, but less detailed training set

Increase minimum length:

- Repeat the same word many times

Data	Average Length
Aug-1	17.35
Aug-2	17.51
VTT23	24.68

Future work:

- **Generate more descriptive captions**

Limitations

Robust sub-task:

- introduce natural corruptions and perturbations
- robust enough to handle these perturbations

More challenging benchmark:

real-world conditions:

- inadequate lighting
- camera shake.

Thanks!

Feel free to contact us:

kaiwenwei@ruc.edu.cn
qjin@ruc.edu.cn