RUC_AIM3 at TRECVID 2021: Video to Text Description

Liang Zhang, Yuqing Song, Qin Jin

AI·M³ , Renmin University of China

Outline

- Video Description Generation
- Fill-in-the-Blanks
- Conclusions

One

Video Description Generation

Video Description Generation

- Automatically generate a sentence that best describes the video using natural language
- Learning video concepts from both temporal and spatial dimensions

"A man stands in a doorway using a pull up bar to do pull ups"

Video Description Generation

- Previous works show LSTM-based model perform better than vanilla Transformer. [1]
- We build our system with pretraining based transformer framework
- Modeling with video concepts

[1] Zhao Y, Song Y, Chen S, et al. RUC_AIM3 at TRECVID 2020: Ad-hoc Video Search & Video to Text Description[C]. TRECVID, 2020.

Video Description Generation

- Concept Enhanced Pretraining based Transformer Model (CE-PTM)

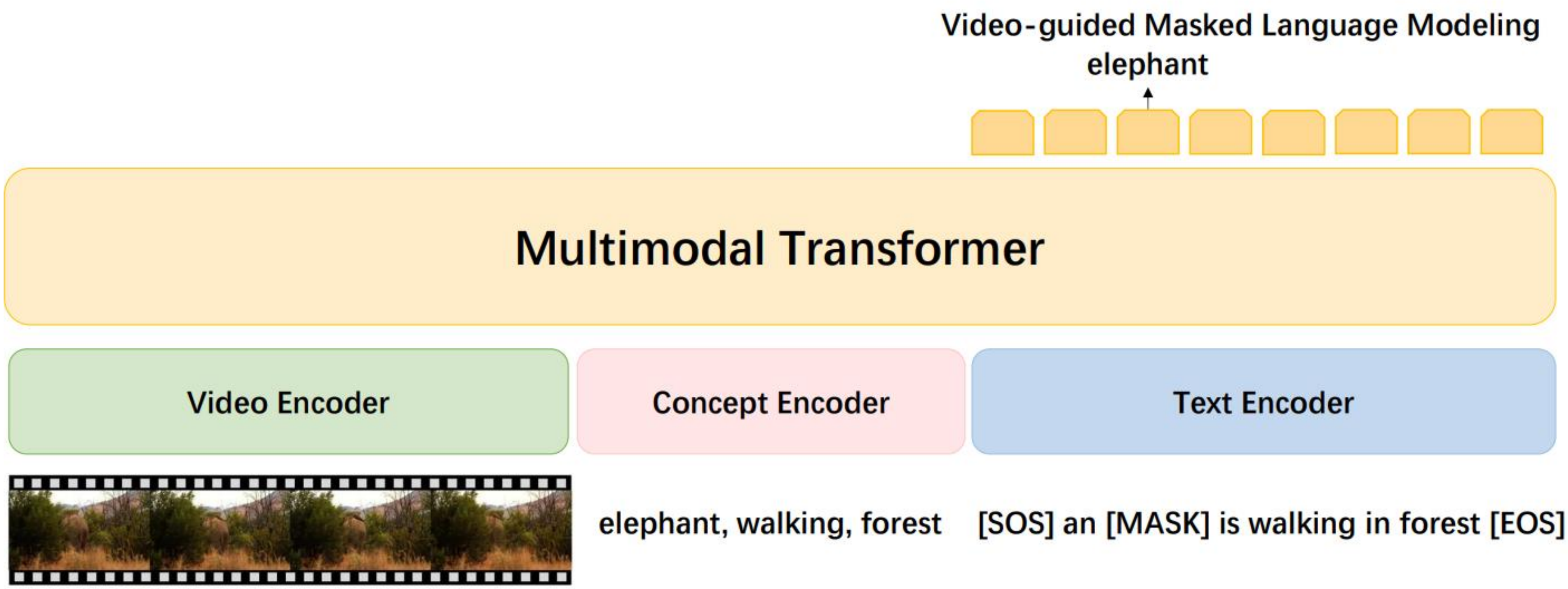


Figure 1: Architecture of CE-PTM.

Video Description Generation

- Video Encoder: encode raw video frames into visual representations.
  - Extract video features $V^f = \{v_1^f, \ldots, v_{L_v}^f\} \in \mathbb{R}^{L_v \times 7168}$
    - Key frame sampling for every 8 frames
    - Off-the-shelf feature extractors

Table 1: Comparison between the video features.

Name	Type	Architecture	Pretrained Data	Dimension
I3D [8]	3D	CNN	Kinetics-400 [8]	1024
ResNeXt-101 [9]	2D	CNN	ImageNet [12]	2048
irCSN [10]	3D	CNN	IG-65M [13]	2048
Swin-Transformers [11]	2D	Transformer	ImageNet [12]	1536
CLIP ViT-B/32 [7]	2D	Transformer	WebImageText [7]	512

  - Encode into visual representations
    - One layer Transformer Encoder

$$V^r = \mathrm{TransformerEncoderLayer}(V^e + \mathrm{TE}(0), \theta_v)$$

Video Description Generation

- Text Encoder: encode sentences into textual representations
  - Learnable word embedding $T^e = \{t_1^e, ..., t_{L_t}^e\}, T^e \in \mathbb{R}^{L_t \times 512}$
    - Special token [SOS], [EOS] and [MASK]

Text Encoder

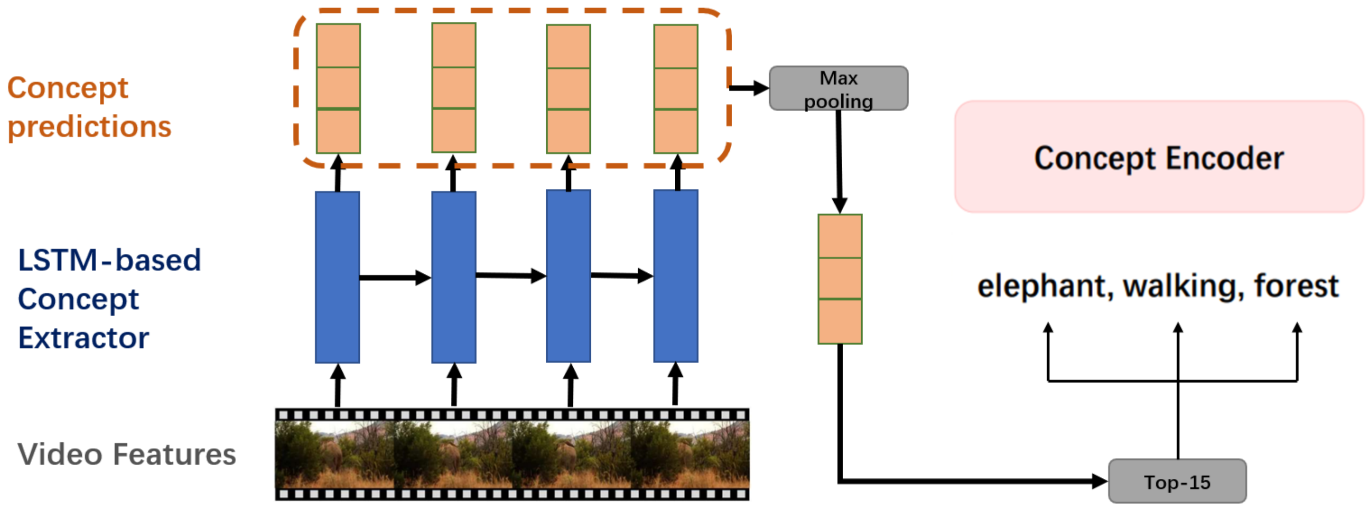
[SOS] an [MASK] is walking in forest [EOS]

  - One layer Transformer Encoder

$$T^r = \mathrm{TransformerEncoderLayer}(T^e + \mathrm{TE}(1) + \mathrm{PE}, \theta_t)$$

Video Description Generation

- Concept Encoder: getting Concept representations
  - Video concepts: Objects & Actions in the video
  - Automatically generate video-concepts pairs
    - Extract nouns and verbs from video captions
  - Train a LSTM-based Concept Extractor



Off-the-shelf Concept Extractor

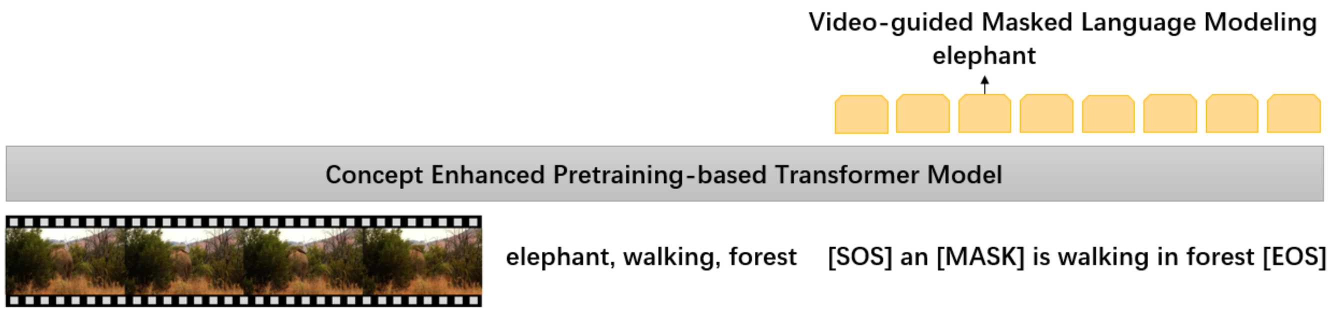
Video Description Generation

- Multimodal Transformer
  - Encode representations in different modalities
  - Four transformer encoder layers

$$H_0 = [V^r; C^r; T^r]$$
$$H_i = \text{TransformerEncoderLayer}(H_{i-1}, \theta_i), 0 < i <= 4$$

Video Description Generation

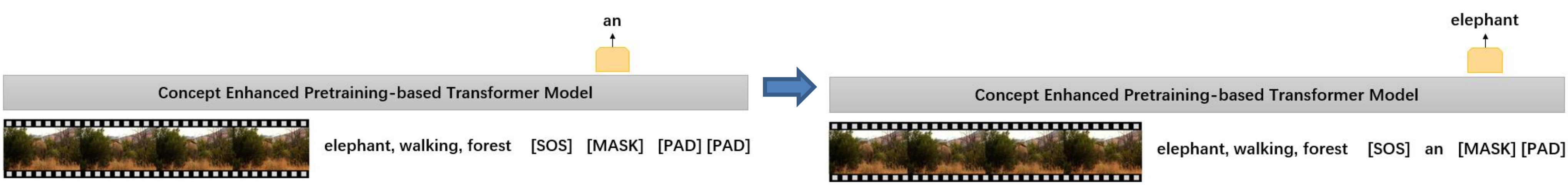
- Pretraining Task
  - Video-guided Masked Language Modeling (VMLM)
  - Predict masked words with video, concepts, and context
  - Select 15% masked words, and replace them with[3]:
    - [MASK] – 80% of the time
    - A random token 10% of the time
    - Keeping unchanged 10% of the time



[3] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

Video Description Generation

- Finetuning
  - Modify self-attention mask of VMLM
    - Avoid seeing future words
- Inference
  - Feed [MASK] and predict the current word autoregressively
  - Stop when [EOS] is predicted



Step 1

Step 2

Video Description Generation

- Reinforcement Learning (RL) [2]
  - CIDEr reward

$$L_{rl} = -\frac{1}{L_s} r(T^s) \sum_{i=1}^{L_s} \log p(T_i^s | V, C, T_{<i}^s)$$

- Hybrid Reranking
  - Visual relevance scoring by VSE++ model
  - Ensemble the captions generated from different models

[2] Rennie S J, Marcheret E, Mroueh Y, et al. Self-critical sequence training for image captioning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7008-7024.

Video Description Generation

- **Training dataset** : TGIF, TRECVID16-19, MSRVTT, VATEX

  Validation dataset: TRECVID20

Video Description Generation

- **Experiments**

PTM, CE-PTM: Proposed methods, 5 features

TOP1-in-2020: 2-layers LSTM Model, feature: irCSN + ResNext

BUTD: 2-layers LSTM Model, feature: 5 features[①]

① 5 features = irCSN+ResNext+CLIP+Swin+I3D

Video Description Generation

- **Experiments**

PTM, CE-PTM: Proposed methods, 5 features

TOP1-in-2020: 2-layers LSTM Model, feature: irCSN + ResNext

BUTD: 2-layers LSTM Model, feature: 5 features①

Models	BLEU@4	METEOR	CIDEr	SPICE
Trained with Cross-Entropy				
TOP1-in-2020 [3]	16.7	16.9	26.1	10.6
BUTD [24]	18.4	17.4	29.5	11.3
Ours PTM	**19.7**	18.4	33.8	12.3
Ours CE-PTM	19.6	**18.8**	**34.5**	**12.7**

① 5 features = irCSN+ResNext+CLIP+Swin+I3D

Video Description Generation

- **Experiments**

PTM, CE-PTM: Proposed methods, 5 features

TOP1-in-2020: 2-layers LSTM Model, feature: irCSN + ResNext

BUTD: 2-layers LSTM Model, feature: 5 features①

Conclusion 1: Encoding with different kinds of features make improvements.

Models	BLEU@4	METEOR	CIDEr	SPICE
Trained with Cross-Entropy				
TOP1-in-2020 [3]	16.7	16.9	26.1	10.6
BUTD [24]	18.4	17.4	29.5	11.3
Ours PTM	**19.7**	18.4	33.8	12.3
Ours CE-PTM	19.6	**18.8**	**34.5**	**12.7**

① 5 features = irCSN+ResNext+CLIP+Swin+I3D

Video Description Generation

- **Experiments**

PTM, CE-PTM: Proposed methods, 5 features

TOP1-in-2020: 2-layers LSTM Model, feature: irCSN + ResNext

BUTD: 2-layers LSTM Model, feature: 5 features①

Conclusion 2: Pretraining-based transformer model is more suitable for caption generation

Models	BLEU@4	METEOR	CIDEr	SPICE
Trained with Cross-Entropy				
TOP1-in-2020 [3]	16.7	16.9	26.1	10.6
BUTD [24]	18.4	17.4	29.5	11.3
Ours PTM	**19.7**	18.4	33.8	12.3
Ours CE-PTM	19.6	**18.8**	**34.5**	**12.7**

① 5 features = irCSN+ResNext+CLIP+Swin+I3D

Video Description Generation

- **Experiments**

Conclusion 3: Video concepts are helpful for generating more diverse captions that can be complementary to the classic models.

Models	BLEU@4	METEOR	CIDEr	SPICE
Trained with Cross-Entropy				
TOP1-in-2020 [3]	16.7	16.9	26.1	10.6
BUTD [24]	18.4	17.4	29.5	11.3
Ours PTM	**19.7**	18.4	33.8	12.3
Ours CE-PTM	19.6	**18.8**	**34.5**	**12.7**
Trained with Reinforcement Learning				
TOP1-in-2020 [3]	17.4	16.9	28.3	10.6
BUTD [24]	19.4	17.9	31.7	11.4
Ours PTM	21.3	18.8	35.4	**12.7**
Ours CE-PTM	**21.4**	**19**	**35.8**	**12.7**
Hybrid reranking				
BUTD [24]	20.2	18.5	34.7	12.2
PTM	21.4	19.0	37.1	13.0
PTM+CE-PTM	**21.6**	**19.3**	38.1	13.3
BUTD+PTM+CE-PTM	21.5	**19.3**	**38.5**	**13.4**

Video Description Generation

- **Submission**

  - Run 4: Our single best model.
  - Run 3: Ensemble of the BUTD models.
  - Run 2: Ensemble of the PTM and CE-PTM models.
  - Run 1: Ensemble of run2 and run3 by captions reranking.

Table 3: Results of the submitted four runs on TRECVID VTT 2021 dataset.

Runs	BLEU@4	METEOR	CIDEr	SPICE	STS
4	3.9	31.6	33.6	11.9	44.1
3	3.7	31.1	32.4	11.6	44.2
2	**4.7**	**32.7**	35.9	**12.7**	**45.7**
1	4.6	32.5	**36.0**	12.6	45.6

Two

Fill in the Blanks

Fill-in-the-Blanks

- Complete video description with a blank based on the video content.

A man is jumping off a mountain using __**a parachute**__ in a sunny day.

Fill-in-the-Blanks

- Generate pseudo blanks using video captioning dataset
  - Extract verb and noun phrases
  - Randomly select one of the pseudo blanks to fill during training

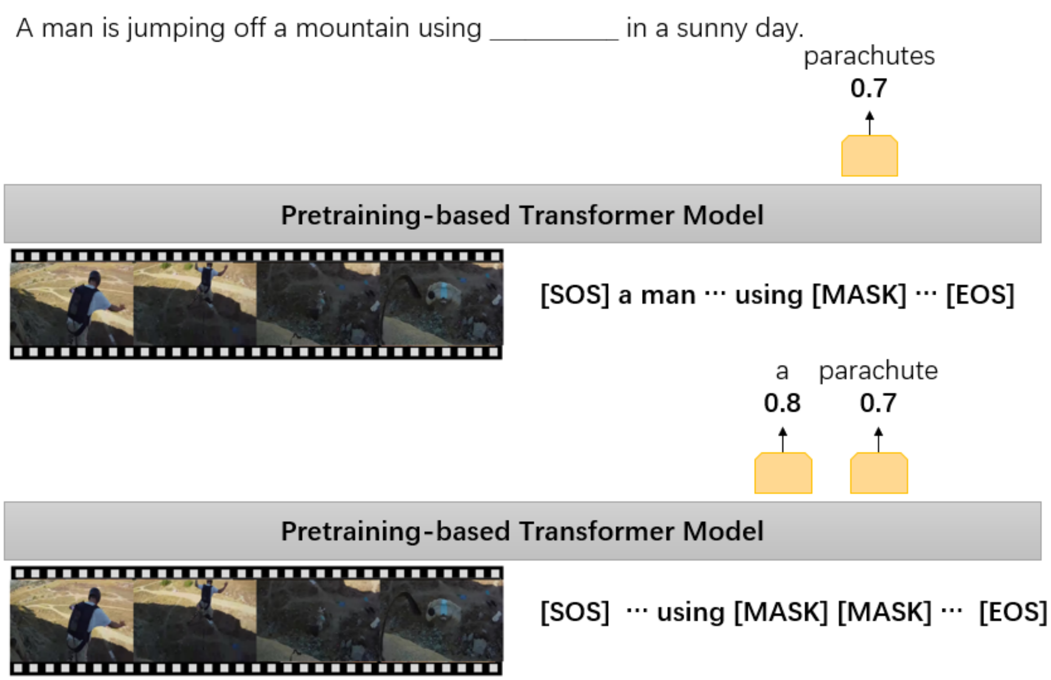
Fill-in-the-Blanks

- Use PTM to Fill-in-the-Blanks

- Four approaches to generate the blank phrases
  - Non-autoregressive Mask Generation (NMG)
  - Auto-regressive Mask Generation (AMG)
  - LSTM Decoder Generation (LDG)
  - Transformer Decoder Generation (TDG)

Fill-in-the-Blanks

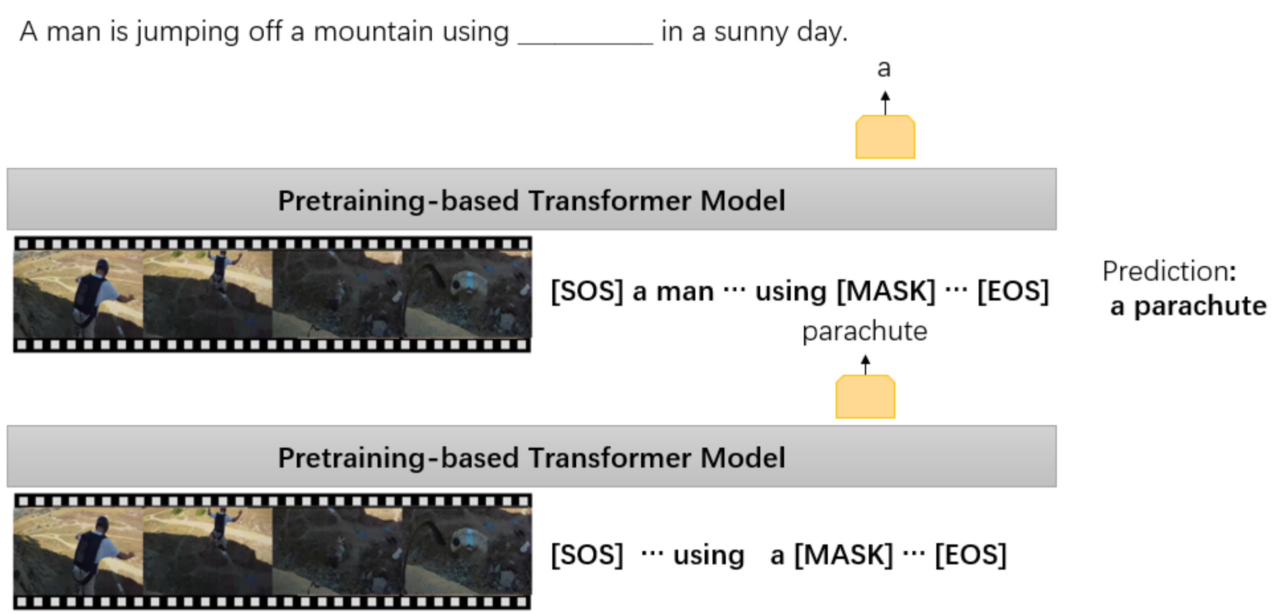
- Non-autoregressive Mask Generation (NMG)
  - Replace the blank phrase with 1-3 [MASK] tokens and make three predictions
  - For each prediction, we aggregate prediction scores of each token with mean pooling



Predictions	score
parachutes	0.7
a parachute	**0.75**

Fill-in-the-Blanks

- Autoregressive Mask Generation (AMG)
  - Feed [MASK] to generate the current blanked word
  - End until [BLANK_EOS] is generated



Fill-in-the-Blanks

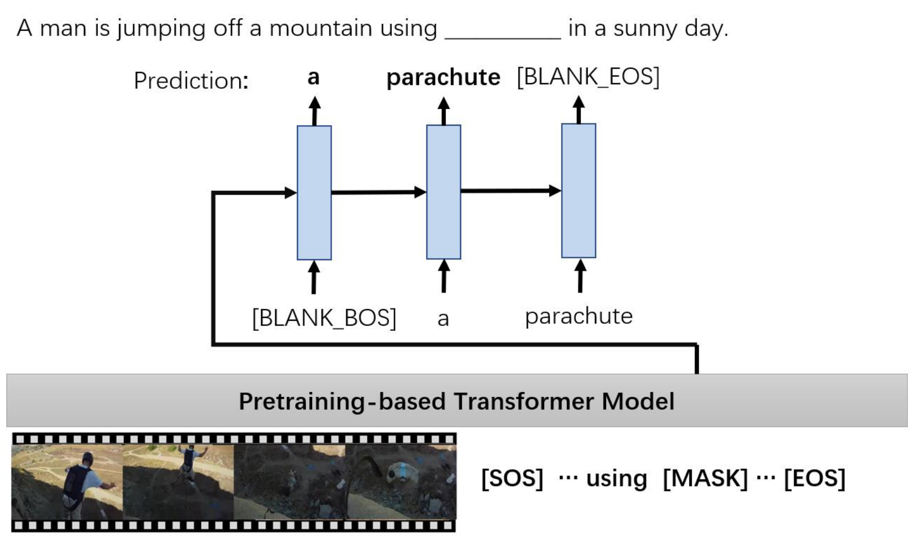
- LSTM Decoder Generation (LDG)
  - Replace blank phrase with one [MASK] token
  - Initialize the first hidden state with the output feature of the [MASK]
  - Generate the blank phrase in autoregressive manner

$$h_0 = h_{[MASK]}$$

$$h_i = \mathrm{LSTM}(y_{i-1}, h_{i-1}; \theta_l)$$

$$p(y_i|y_{<i}) = \mathrm{softmax}(h_i W_{ld}^T)$$

$$L_{xe} = -\frac{1}{N_b}\sum_{i=0}^{N_b} \log p(y_i|y_{<i})$$



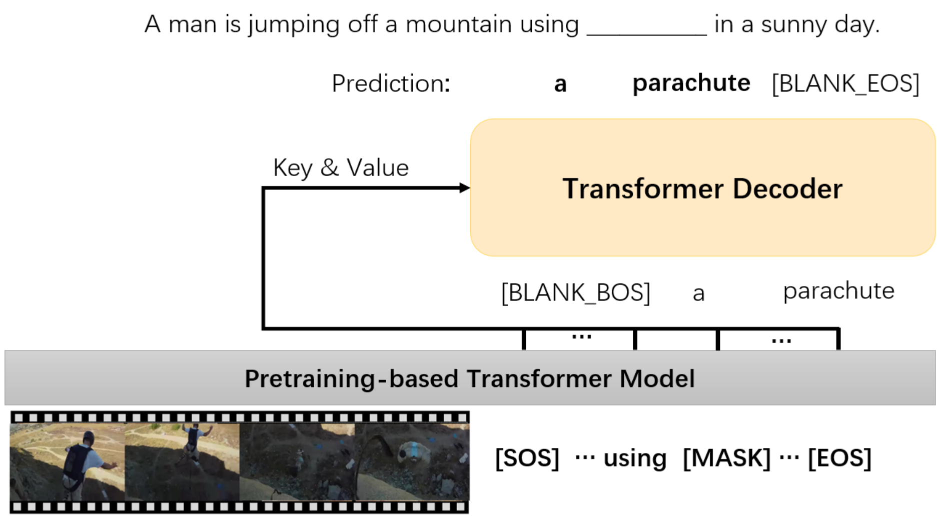
Fill-in-the-Blanks

- Transformer Decoder Generation (TDG)
  - Replace blank phrase with one [MASK] token
  - Features of the sentence from PTM are treat as the Key & Value
  - 4 transformer decoder layers

$$Y_i = \mathrm{TransformerDecoderLayer}(Y_{i-1}, H_t, H_t, \theta_i)$$

$$p(y_i|y_{<i}) = \mathrm{softmax}(Y_4 W_{tfd}^T)$$

$$L_{xe} = -\frac{1}{N_b}\sum_{i=0}^{N_b} \log p(y_i|y_{<i})$$



Fill-in-the-Blanks

- Transformer Decoder Generation (TDG)
  - Random Blank Filling (RBF)
    - Select 1~5 continuous words as a random blank
    - Not necessarily noun or verb phrases (different from pseudo blanks)
  - Finetuning on pseudo blanks

Fill-in-the-Blanks

- **Generate pseudo blanks on:**
  - Training dataset : TGIF, TRECVID16-19, MSRVTT, VATEX
  - Validation dataset: TRECVID20

- **Automatic Metrics**
  - Exactly Match (EM)
  - F1 Score over tokens (F1)

Fill-in-the-Blanks

- **Experiment:**
  - Non-autoregressive Mask Generation (NMG)
  - Auto-regressive Mask Generation (AMG)
  - LSTM Decoder Generation (LDG)
  - Transformer Decoder Generation (TDG)
    - Random Blank Filling (RBF)

Model	EM	F1
Single Model		
NMG	19.4	39.5
AMG	21.5	40.8
LDG	28.5	45.0
TDG	28.9	45.1
TDG+RBF	**29.1**	**45.8**
Hybrid reranking		
All models	**30.3**	**47.0**

Fill-in-the-Blanks

- **Experiment:**
  - Decoder-based method outperform those without additional decoder

Model	EM	F1
Single Model		
NMG	19.4	39.5
AMG	21.5	40.8
LDG	28.5	45.0
TDG	28.9	45.1
TDG+RBF	**29.1**	**45.8**
Hybrid reranking		
All models	**30.3**	**47.0**

Fill-in-the-Blanks

- **Experiment:**
  - Decoder-based method out perform those without additional decoder
  - Transformer Decoder performs slightly better than LSTM

Model	EM	F1
Single Model		
NMG	19.4	39.5
AMG	21.5	40.8
LDG	28.5	45.0
TDG	28.9	45.1
TDG+RBF	**29.1**	**45.8**
Hybrid reranking		
All models	**30.3**	**47.0**

Fill-in-the-Blanks

- **Experiment:**
  - Decoder-based method out perform those without additional decoder
  - Transformer Decoder perform slightly better
  - RBF pretraining brings improvements

Model	EM	F1
Single Model		
NMG	19.4	39.5
AMG	21.5	40.8
LDG	28.5	45.0
TDG	28.9	45.1
TDG+RBF	**29.1**	**45.8**
Hybrid reranking		
All models	**30.3**	**47.0**

Fill-in-the-Blanks

- **Submission:**
  - Run 2: The single best model with Transformer Decoder (TDG+RBF).
  - Run 1: Ensemble of all generation methods mentioned above via Hybrid Reranking.

Table 5: Evaluation results on TRECVID VTT 2021.

System	Automatic Metrics		Human Evaluation	
	EM	F1	Average	Average Z
human	-	-	**85.4**	**42.0**
Run2	14.1	38.7	**80.1**	**17.3**
Run1	**15.3**	**40.8**	79.5	13.0

Fill-in-the-Blanks

- **Submission:**
  - Run 2: The single best model with Transformer Decoder (TDG+RBF).
  - Run 1: Ensemble of all generation methods mentioned above via Hybrid Reranking.

- Gap between Automatic Metrics and Human Evaluation

Table 5: Evaluation results on TRECVID VTT 2021.

System	Automatic Metrics		Human Evaluation	
	EM	F1	Average	Average Z
human	-	-	**85.4**	**42.0**
Run2	14.1	38.7	**80.1**	**17.3**
Run1	**15.3**	**40.8**	79.5	13.0

Conclusion

- **Description Generation**
  - Pre-training based Transformer model can outperform the LSTM-based model
  - Video concepts can be helpful for generating more diverse captions that can be complementary to classic captioning models.
  - We ranked 1st in METEOR, CIDEr, SPICE, STS, and 2nd in BLEU4

- **Fill-in-the-Blanks**
  - Decoder-based methods perform better than models without decoder
  - Transformer decoder performs better than LSTM decoder
  - We ranked 1st in human evaluation

THANKS！

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

zhangliang00@ruc.edu.cn
syuqing@ruc.edu.cn
qjin@ruc.edu.cn

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