VATEX Video Captioning Challenge 2020: Multi-View Features and Hybrid Reward Strategies for Video Captioning

Xinxin Zhu, Longteng Guo, Peng Yao, Shichen Lu, Wei Liu, Jing Liu

NLPR, Institute of Automation, Chinese Academy of Sciences

University of Science and Technology Beijing Wuhan University

Challenges in VATEX dataset

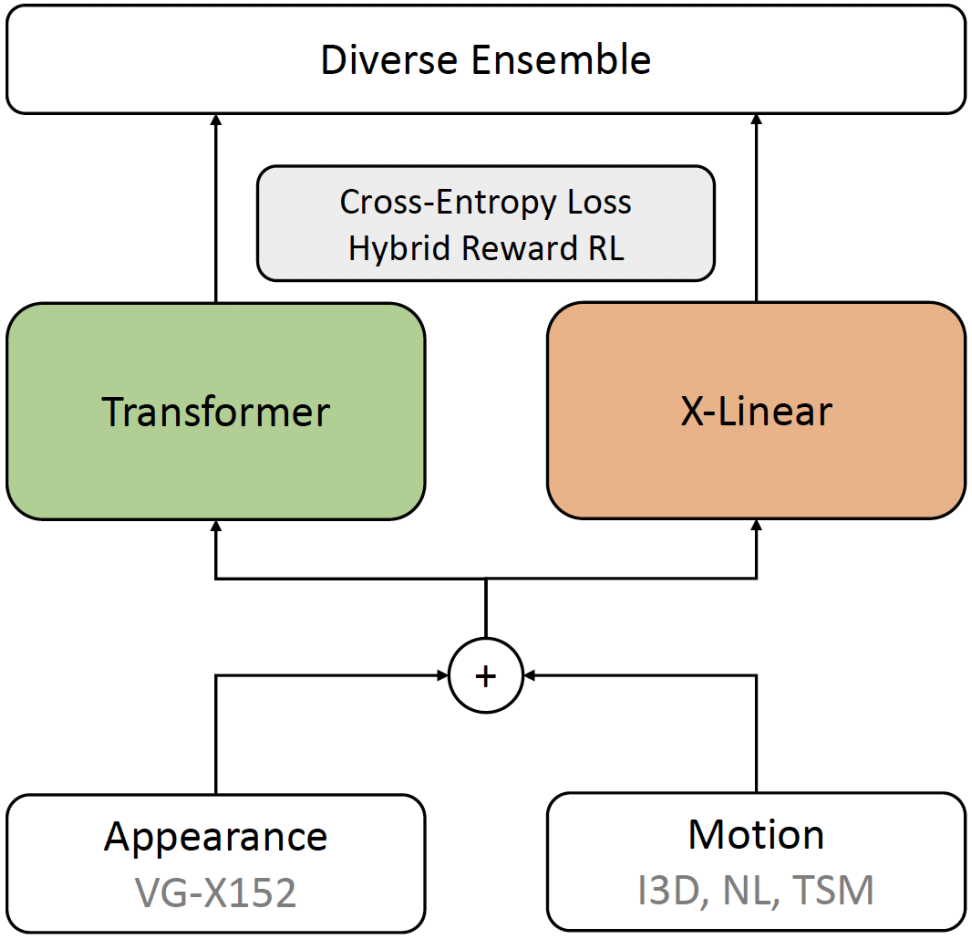
- Large variety of video -> difficulty in recognizing visual content
- Vast diversity of the captions -> difficulty in modeling language

Baseline: A man is surfing in the waves on a wave in the ocean
GT1: Man rides jet ski in wet-suit on rolling sea until he falls off as sun sets.
GT2: The watercraft are being used to quickly move through the water and over the waves.
GT3: A person jet skiing in wavy water and falling in the water after a little while.
GT4: A person on a jet ski going across the water and jumps off
GT5: In the ocean a man stands and rides a jet ski through the water and then falls off.

Our Solutions to the Above Challenges…

- Encoder: Multi-View Video Features
  - To provide more comprehensive and dis-criminative video representation
- Decoder: more advanced captioning models
  - Better language generation ability
- Learning: Hybrid Reward For Reinforcement Learning
  - More balanced performance across metrics
- Ensemble: Diverse Ensemble

Method Overview



Encoder: Multi-View Video Features

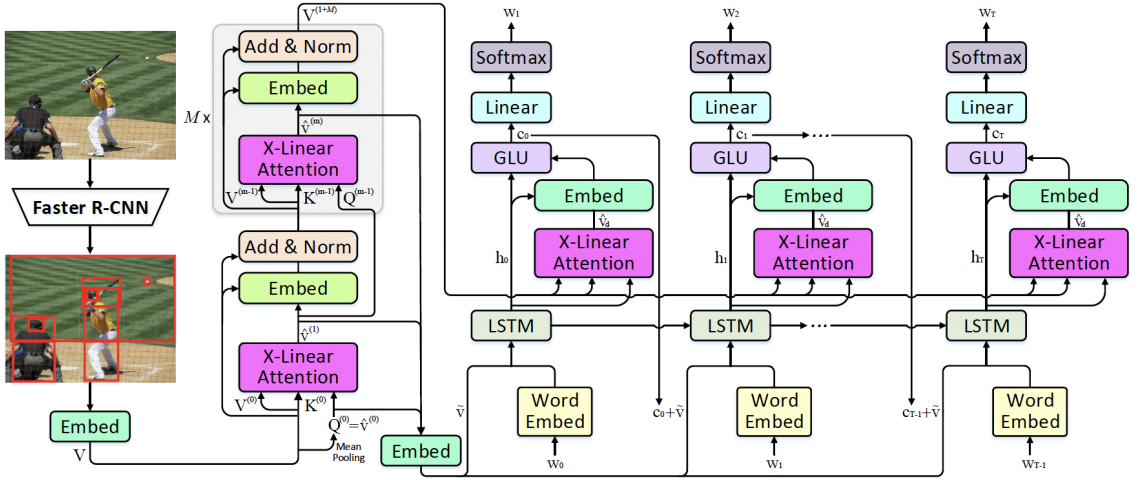
- Motion features
  - temporal dimension
  - I3D，Non-local models，TSM
  - Kinetics-600 pretrained
- Appearance features
  - spatial dimension
  - Faster R-CNN + ResNeXt-152
  - Visual Genome pretrained
- Better video features extraction
  - randomly cropping video frames
  - randomly selecting partial videos

Decoder: SoTA captioning models

- X-Linear
  - LSTM-based
- Transformer
  - Self-Attention-based

Decoder: SoTA captioning models

- X-Linear
  - LSTM-based
  - X-Linear Attention
  - Extend to video captioning

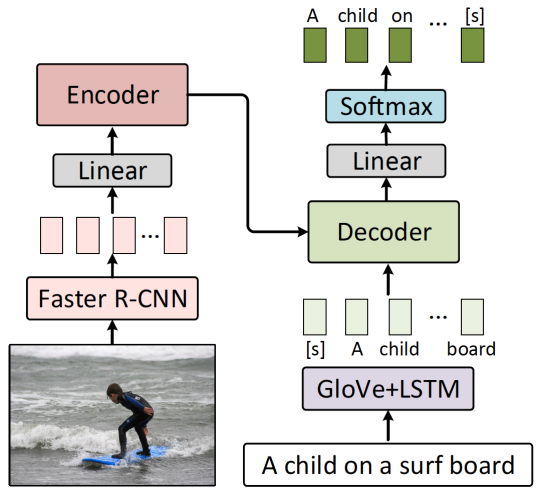


Model	B@1		B@2		B@3		B@4		M		R		C	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
LSTM-A (ResNet-152) [40]	78.7	93.7	62.7	86.7	47.6	76.5	35.6	65.2	27.0	35.4	56.4	70.5	116.0	118.0
Up-Down (ResNet-101) [2]	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
RFNet (ResNet+DenseNet+Inception) [13]	80.4	95.0	64.9	89.3	50.1	80.1	38.0	69.2	28.2	37.2	58.2	73.1	122.9	125.1
SGAE (ResNet-101) [36]	81.0	95.3	65.6	89.5	50.7	80.4	38.5	69.7	28.2	37.2	58.6	73.6	123.8	126.5
GCN-LSTM (ResNet-101) [38]	80.8	95.2	65.5	89.3	50.8	80.3	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
AoANet (ResNet-101) [12]	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
HIP (SENet-154) [39]	81.6	**95.9**	66.2	90.4	51.5	81.6	39.3	71.0	28.8	38.1	59.0	74.1	127.9	130.2
X-LAN (ResNet-101)	81.1	95.3	66.0	89.8	51.5	81.5	39.5	71.4	29.4	38.9	59.2	74.7	128.0	130.3
X-LAN (SENet-154)	81.4	95.7	66.5	**90.5**	52.0	82.4	40.0	**72.4**	**29.7**	**39.3**	**59.5**	**75.2**	130.2	132.8
X-Transformer (ResNet-101)	81.3	95.4	66.3	90.0	51.9	81.7	39.9	71.8	29.5	39.0	59.3	74.9	129.3	131.4
X-Transformer (SENet-154)	**81.9**	95.7	**66.9**	**90.5**	**52.4**	**82.5**	**40.3**	**72.4**	29.6	39.2	**59.5**	75.0	**131.1**	**133.5**

[1] Pan, Yingwei, et al. "X-Linear Attention Networks for Image Captioning." *CVPR* 2020.

Decoder: SoTA captioning models

- Transformer
  - Self-Attention-based
  - The SoTA on various NLP tasks
  - Multi-head attention
  - Extend to video captioning



Model	B@1		B@2		B@3		B@4		M		R		C	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
Google NIC [50]	71.3	89.5	54.2	80.2	40.7	69.4	30.9	58.7	25.4	34.6	53.0	68.2	94.3	94.6
M-RNN [51]	71.6	89.0	54.5	79.8	40.4	68.7	29.9	57.5	24.2	32.5	52.1	66.6	91.7	93.5
LRCN [25]	71.8	89.5	54.8	80.4	40.9	69.5	30.6	58.5	24.7	33.5	52.8	67.8	92.1	93.4
ADP-ATT [9]	74.8	92.0	58.4	84.5	44.4	74.4	33.6	63.7	26.4	35.9	55.0	70.5	104.2	105.9
LSTM-A [21]	78.7	93.7	62.7	86.7	47.6	76.5	35.6	65.2	27.0	35.4	56.4	70.5	116.0	118.0
SCST [10]	78.1	93.7	61.9	86.0	47.0	75.9	35.2	65.5	27.0	35.5	56.3	70.7	114.7	116.7
Up-Down [6]	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
RFNet [49]	80.4	95.0	64.9	89.3	50.1	80.1	38.0	69.2	28.2	37.2	58.2	73.1	122.9	125.1
GCN-LSTM [22]	-	-	65.5	89.3	50.8	80.3	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
SRCB-ML-Lab	81.1	95.4	66.0	89.8	51.5	81.3	39.7	71.3	28.4	37.3	58.5	73.1	125.3	126.7
h-p-hl	80.5	95.0	65.3	89.6	50.9	81.1	39.0	70.9	28.7	38.2	58.6	74.1	125.0	127.2
TecentAI.v2	81.1	95.5	65.7	90.0	50.8	80.9	38.6	70.1	28.6	37.7	58.7	73.7	125.4	127.8
lun	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
MT (ours)	**81.7**	**95.6**	**66.8**	**90.5**	**52.4**	**82.4**	**40.4**	**72.2**	**29.4**	**38.9**	**59.6**	**75.0**	**130.0**	**130.9**

[1] Zhu, Xinxin, et al. "Captioning transformer with stacked attention modules." Applied Sciences 2018.
[2] Yu, Jun, et al. "Multimodal transformer with multi-view visual representation for image captioning." TCSVT 2019.

Learning: Hybrid Reward for RL

- Hybrid reward, i.e. a linear combination of different metric scores, can result in a better overall result

$$scores = \alpha * CIDEr + \beta * BLEU + \gamma * METEOR + \eta * ROUGE$$
$$\alpha + \beta + \gamma + \eta = 1$$

Ensemble: Diverse Ensemble of Models

- Ensemble method
  - Average Ensemble
  - Weighted Ensemble
- Used models
  - Different architectures: X-Linear and Transformer
  - Initialization with different seeds
  - Different training settings
    - Learning rate
    - Scheduled sampling probability
    - Visual features
    - Hybrid reward

Results

Language	Method	CIDEr	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
Chinese	VATEX-team [15]	35.1	74.5	53.7	36.6	24.8	29.4	51.6
	X-Linear	56.8	81.6	63.5	45.9	32.2	31.9	56.1
	X-Linear+Transformer	**59.5**	**82.2**	**64.3**	**46.5**	**32.6**	**32.1**	**56.5**
English	VATEX-team [15]	45.1	71.3	53.3	39.6	28.5	21.6	47.0
	X-Linear	76.3	81.9	66.5	52.1	39.4	25.2	53.0
	X-Linear+Transformer	**81.4**	**83.1**	**68.0**	**53.6**	**40.7**	**25.8**	**53.7**

Table 1. The ensemble results of our ultimate models on Vatex test set and **X-Linear+Transformer** is our final submission on the leaderboard.

Thank you!