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
Our Solutions to the Above Challenges...

• Encoder: Multi-View Video Features
  • To provide more comprehensive and discriminative 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

---

Decoder: SoTA captioning models

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

Learning: Hybrid Reward for RL

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

\[ \text{scores} = \alpha \times \text{CIDEr} + \beta \times \text{BLEU} + \gamma \times \text{METEOR} + \eta \times \text{ROUGE} \]

\[ \alpha + \beta + \gamma + \eta = 1 \]
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

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

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!