VATEX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research

Xin (Eric) Wang
UC Santa Cruz
@ACL 2020 ALVR Workshop
Why \textsc{VATeX}?
There are thousands of languages on earth!
Why VATEX?

Unique and Fine-grained
Why VATeX?

257 Classes → 600 Classes
Why VATeX?

200K Captions → 826K Captions
41.3K Unique Video Clips
826K Unique Captions in English & Chinese
600 Human Activities
Comparison with other Video Description Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MLingual</th>
<th>Domain</th>
<th>#classes</th>
<th>videos:clips</th>
<th>sent</th>
<th>sent/clip</th>
</tr>
</thead>
<tbody>
<tr>
<td>TACoS[45]</td>
<td>-</td>
<td>cooking</td>
<td>26</td>
<td>127:3.5k</td>
<td>11.8k</td>
<td>-</td>
</tr>
<tr>
<td>TACoS-MLLevel[46]</td>
<td>-</td>
<td>cooking</td>
<td>67</td>
<td>185:25k</td>
<td>75k</td>
<td>3</td>
</tr>
<tr>
<td>Youcook[16]</td>
<td>-</td>
<td>cooking</td>
<td>6</td>
<td>88:-</td>
<td>2.7k</td>
<td>-</td>
</tr>
<tr>
<td>Youcook II[72]</td>
<td>-</td>
<td>cooking</td>
<td>89</td>
<td>2k:15.4k</td>
<td>15.4k</td>
<td>1</td>
</tr>
<tr>
<td>MPII MD[47]</td>
<td>-</td>
<td>movie</td>
<td>-</td>
<td>94:68k</td>
<td>68.3k</td>
<td>1</td>
</tr>
<tr>
<td>M-VAD[56]</td>
<td>-</td>
<td>movie</td>
<td>-</td>
<td>92:46k</td>
<td>55.9k</td>
<td>-</td>
</tr>
<tr>
<td>LSMDC[48]</td>
<td>-</td>
<td>movie</td>
<td>-</td>
<td>200:128k</td>
<td>128k</td>
<td>1</td>
</tr>
<tr>
<td>Charades[52]</td>
<td>-</td>
<td>indoor</td>
<td>157</td>
<td>10k:10k</td>
<td>27.8k</td>
<td>2-3</td>
</tr>
<tr>
<td>VideoStory[22]</td>
<td>-</td>
<td>social media</td>
<td>-</td>
<td>20k:123k</td>
<td>123k</td>
<td>1</td>
</tr>
<tr>
<td>ActyNet-Cap[30]</td>
<td>-</td>
<td>open</td>
<td>200</td>
<td>20k:100k</td>
<td>100k</td>
<td>1</td>
</tr>
<tr>
<td>MSVD[14]</td>
<td>✓</td>
<td>open</td>
<td>-</td>
<td>2k:2k</td>
<td>70k</td>
<td>35</td>
</tr>
<tr>
<td>TGIF[34]</td>
<td>-</td>
<td>open</td>
<td>-</td>
<td>-.100k</td>
<td>128k</td>
<td>1</td>
</tr>
<tr>
<td>VTW[69]</td>
<td>-</td>
<td>open</td>
<td>-</td>
<td>18k:18k</td>
<td>18k</td>
<td>1</td>
</tr>
<tr>
<td>MSR-VTT[66]</td>
<td>-</td>
<td>open</td>
<td>257</td>
<td>7k:10k</td>
<td>200k</td>
<td>20</td>
</tr>
<tr>
<td>VATEX (ours)</td>
<td>✓</td>
<td>open</td>
<td>600</td>
<td>41.3k:41.3k</td>
<td>826k</td>
<td>20</td>
</tr>
</tbody>
</table>
Meet VATEX!

• A person wearing a bear costume is inside an inflatable play area as they lose their balance and fall over.
• A person in a bear costume stands in a bounce house and falls down as people talk in the background.
• A person dressed in a cartoon bear costume attempts to walk in a bounce house.
• A person in a mascot uniform trying to maneuver a bouncy house.
• A person in a comic bear suit falls and rolls around in a moon bounce.
• A person dressed as a teddy bear stands in a bouncy house and then falls over.
• Someone dressed in a bear costume falling over in a bouncy castle.
• A person dressed up as a bear is standing in a bouncy castle and falls down.
• A man in a bear costume is balancing in a bouncy castle before they tumble to the floor.
• A man in costume was trying to stand straight on a bouncy castle but fell.

• 一个人穿着熊的布偶外套倒在了蹦床上。
• 一个人穿着一套小熊服装在充气蹦床上摔倒了。
• 一个穿着熊外衣的人在充气垫子上摔倒了。
• 一个穿着深色衣服的人正在蹦床上。
• 在一个充气大型玩具里,有一个人穿着熊的衣服站了一下之后就摔倒了。

• 一个打扮成泰迪熊的人站在充气房上,然后摔倒了。
• 有个穿着熊装的人在充气城堡摔倒了。
• 一个装扮成熊的人站在充气蹦床上,然后摔倒了。
• 一个穿着熊服装的人在一个有弹性的城堡里平衡,然后他们就倒在了地板上。
• 一个穿着布偶熊的人试图站在一个充气城堡上,但却摔倒了。
\textbf{VATEX vs. MSR-VTT}

- Distributions of Caption Lengths.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{chart.png}
\caption{Caption Length Distributions for MSR-VTT, VATEX-en, and VATEX-zh.}
\end{figure}
VATEX VS. MSR-VTT

• Distributions of Caption Lengths.
**VATEX vs. MSR-VTT**

- Distributions of Unique Nouns Per Caption.
\textbf{VATEX \textit{vs.} MSR-VTT}

- Distributions of Unique Nouns Per Caption.
**VATEX vs. MSR-VTT**

- Distributions of Unique Verbs Per Caption.

![Bar Chart](chart.png)

- MSR-VTT
- VATEX-en
- VATEX-zh
VATEX vs. MSR-VTT

- Distributions of Unique Verbs Per Caption.
### VATeX vs. MSR-VTT

<table>
<thead>
<tr>
<th>Dataset</th>
<th>sent length</th>
<th>duplicated sent rate</th>
<th>#unique $n$-grams</th>
<th>#unique POS tags</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>intra-video</td>
<td>inter-video</td>
<td>1-gram</td>
</tr>
<tr>
<td>MSR-VTT</td>
<td>9.28</td>
<td>66.0%</td>
<td>16.5%</td>
<td>29,004</td>
</tr>
<tr>
<td>VATeX-en</td>
<td>15.23</td>
<td>0</td>
<td>0</td>
<td>35,589</td>
</tr>
<tr>
<td>VATeX-zh</td>
<td>13.95</td>
<td>0</td>
<td>0</td>
<td>47,065</td>
</tr>
</tbody>
</table>
VATeX Tasks

- Multilingual Video Captioning

A little boy reaches into a basket and puts clothes into the washing machine.

一个小男孩站在洗衣机旁边，把篮子里的脏衣服扔进洗衣机里。
**VATEX Tasks**

- **Video-Guided Machine Translation**

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

  ![Diagram showing machine translation and video-guided machine translation](image)

  - Machine Translation
    - 一个 男人 站在 门口 使用 拉起酒吧 做 拉。
      - (pulling pub) (do pull)
    - 一个 男人 在 门口 使用 拉杆 做 向上运动。
      - (pull-up bar) (do upward movement)
VATeX Potentials

• Video-text retrieval

• Query-based moment retrieval in untrimmed videos

• Zero-/few-shot video captioning

• Cultural and linguistic differences

• ... ...
• Do NOT use any external corpora or pre-trained MT models. The participants may not build upon any existing pre-trained machine translation models for this challenge. The VMT model must be trained on our VATEX dataset from scratch.

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Videos</td>
<td>25,991</td>
<td>3,000</td>
<td>6,000</td>
</tr>
<tr>
<td>English Captions</td>
<td>259,910</td>
<td>30,000</td>
<td>60,000</td>
</tr>
<tr>
<td>Chinese Translations</td>
<td>259,910</td>
<td>30,000</td>
<td>60,000</td>
</tr>
<tr>
<td>Activity Label?</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Released?</td>
<td>yes</td>
<td>yes</td>
<td>English only</td>
</tr>
</tbody>
</table>
### VMT Leaderboard

<table>
<thead>
<tr>
<th>#</th>
<th>User</th>
<th>Entries</th>
<th>Date of Last Entry</th>
<th>Team Name</th>
<th>Corpus Bleu-4</th>
<th>Bleu-1</th>
<th>Bleu-2</th>
<th>Bleu-3</th>
<th>Bleu-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>tosho</td>
<td>3</td>
<td>06/15/20</td>
<td></td>
<td>0.366 (1)</td>
<td>0.631 (3)</td>
<td>0.419 (1)</td>
<td>0.302 (1)</td>
<td>0.225 (1)</td>
</tr>
<tr>
<td>2</td>
<td>zsyzsx1823</td>
<td>4</td>
<td>06/15/20</td>
<td></td>
<td>0.358 (2)</td>
<td>0.633 (1)</td>
<td>0.413 (2)</td>
<td>0.292 (2)</td>
<td>0.215 (2)</td>
</tr>
<tr>
<td>3</td>
<td>syuqing</td>
<td>6</td>
<td>06/15/20</td>
<td></td>
<td>0.353 (3)</td>
<td>0.632 (2)</td>
<td>0.409 (3)</td>
<td>0.287 (3)</td>
<td>0.209 (3)</td>
</tr>
<tr>
<td>4</td>
<td>acdart</td>
<td>2</td>
<td>05/11/20</td>
<td></td>
<td>0.314 (4)</td>
<td>0.598 (5)</td>
<td>0.371 (4)</td>
<td>0.250 (4)</td>
<td>0.175 (4)</td>
</tr>
<tr>
<td>5</td>
<td>zzxslp</td>
<td>1</td>
<td>04/15/20</td>
<td></td>
<td>0.311 (5)</td>
<td>0.599 (4)</td>
<td>0.368 (5)</td>
<td>0.247 (5)</td>
<td>0.172 (5)</td>
</tr>
<tr>
<td>6</td>
<td>Tcat</td>
<td>2</td>
<td>05/10/20</td>
<td></td>
<td>0.282 (6)</td>
<td>0.559 (6)</td>
<td>0.337 (6)</td>
<td>0.221 (6)</td>
<td>0.152 (6)</td>
</tr>
</tbody>
</table>
VMT Challenge 2020 Winner

VMT Challenge 2020

Winner

ACL AVLR Workshop, July 9, 2020

To Team T&T

Tosho Hirasawa, Zhishen Yang,
Mamoru Komachi, Naoaki Okazaki
VMT Challenge 2020 Second Place

VMT Challenge 2020
Second Place
ACL AVLR Workshop, July 9, 2020
To Team Overfit
Zhiyong Wu
VMT Challenge 2020 Third Place

VMT Challenge 2020

Third Place

ACL AVLR Workshop, July 9, 2020

To Team RUC_AIM3

Yuqing Song, Shizhe Chen, Qin Jin
Thanks!

vatex-challenge.org

Now VMT Challenge on CodaLab is open forever!

Organizers

Xin (Eric) Wang
UC Santa Cruz

An Yan
UC San Diego

Lei Li
ByteDance AI Lab

William Yang Wang
UC Santa Barbara