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

Wang et.al.

Qi Huang
Outline

1. Motivation
2. VATEX Dataset Overview
3. Multilingual Video Captioning
4. Video-guided Machine Translation
5. Examples
6. Critique & Future Work
Motivation

- Previous video description datasets are **monolingual**, relatively **small**, with **restricted domains** and **linguistically simple**.
- They only enable video description tasks that are **single-modality on both input and output sides** (input: video frames; output: text).
- Can we have better video description datasets that are **multilingual, large, open domain** and linguistically complex?
- Can we design video description tasks that has **multi-modal input/output**?
VATEX achieves all of that

- 41,250 videos
- 825,000 captions
- **Parallel** description in English and Chinese
- Open domain, 600 classes
- Many more..
Comparison

Comparing to datasets used in seq2seq video2text:

- 10x increase in # sentences
- Open domains v.s. only movie clip

<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-MLevel[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>
Comparison

Comparing to MSR-VTT:

- **Unique sentence ensured with human effort**
- **Multilingual vs monolingual**
- **Linguistically more complicated (n-grams, POS tags..)**

<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>
Comparison

Comparing to MSR-VTT:

- Captions are **uniformly more complex** in caption length, # of unique token

![Graphs showing distribution of caption lengths, unique nouns, and unique verbs.](image)

(a) Distributions of caption lengths.  
(b) Distributions of unique nouns per caption.  
(c) Distributions of unique verbs per caption.

Figure 3: Statistical histogram distributions on MSR-VTT, VaTeX-en, and VaTeX-zh. Compared to MSR-VTT, the VaTeX dataset contains longer captions, each with more unique nouns and verbs.

![Graph showing type-caption curves.](image)

Figure 4: Type-caption curves. VaTeX has more lexical styles and caption diversity than MSR-VTT.
Data Collection

- Categorization and a large part of videos reused from Kinetics-600 dataset
- English caption collection:
  - Experienced, high approval rate AMT workers from English-speaking countries
  - Short, repeated, irrelevant and sensitive word captions are filtered out
  - 412,690 sentences with 2,159 workers
- Chinese caption collection:
  - Half of the captions are direct observation of videos (5/10)
  - Another half are Chinese translation of English captions, bootstrap by 3 commercial machine translation services, cross-approved by co-workers
Multilingual Video Captioning

Problem Setting: given sampled frames from video streams, output captions for each video stream sample

Baseline:

- Pretrained 3D CNN from I3D network to extract frame level features
- Bidirectional LSTM as Video Encoder
- LSTM with attention as caption decoder

\[ y_t, h_t = f_{dec}(y_{t-1}, c_t, h_{t-1}) \]
Multilingual Video Captioning

Multilingual Variants:

1. Shared Encoder
2. Shared Encoder-Decoder (word embedding are different for different languages)
Multilingual Video Captioning: Result

- Multilingual models consistently outperform baseline with reduced # parameters
Video-guided Machine Translation (VMT)

Problem Setting: given sampled frames from video streams and captions in a source language, output captions in the target language.

In following up experiments, some noun/verbs in source captions are randomly masked to test whether video information can help model disambiguate unknown tokens.
VMT: Model

Baseline: Encoder-decoder model without video information. Attend only to source caption features

Variant:
- Video information as a **average frame feature vector**
- Video information as **video encoder output**
- Video information as **attention over video encoder hidden states**

\[ y_t, h_t = f_{dec}^{tgt}([y_{t-1}, c_{t}^{src}, c_{t}^{vi}], h_{t-1}) \]
VMT: Result

- Actively attend to video information significantly boost MT performance over baseline – language dynamics are used as a query to retrieve related video features
- VMT is able to recover missing information with the help of video context
Multilingual Video Captioning: an example

Observation

- Base model and multilingual models all produce high-quality captions
- Information “women/girls” are preserved in base model for English, lost in shared enc-dec

Perhaps “一群女子” never appears in the training corpus for Chinese captions

Multilingual models encourage captions to converge, even at the cost of leaving out information.
VMT: example

Observation:

- Masked noun: in Chinese translation, “a man” is corrected into “a band”. Probably “a man” is much more common in training corpus.
- Disambiguate word: “cartwheel” is corrected from “making wheels” to “cartwheel”.

Video information can help reduce bias, disambiguate word meaning, and provide missing information.

**English:**
a [M] is putting on a [M] performance while using large string instruments.

[M]: band, rock.

**Ground Truth:**
一支乐队正在使用大型弦乐器表演摇滚乐。

**NMT:**
一个男人正在使用大型弦乐器表演。

**VMT:**
一支乐队正在使用大型弦乐器表演摇滚。
Critique & Future Work

Highlights:

- High-quality large scale multilingual video description dataset ready for use
- Data collection process is rigorous and can serve as a reference for future dataset creation
  - Data cross-validated by workers
  - Eliminate repeated data
  - Great visualization of linguistic properties of the dataset (histogram, type-caption curve, etc.)
- Empirical success:
  - Multilingual Video Caption: increase in performance and reduced parameters
  - Video-guided Machine Translation: video information help correct exposure bias, disambiguate rare words, and provide missing information
What's missing:

● Some questionable details:
  ○ Average VI averages frame feature vector directly, while attention is on encoder hidden states -- fair comparison?
  ○ Multilingual video captioning with shared weight encoder/decode: what’s the training scheme? Train English then Chinese? Iteratively? Will better training strategy benefit? How does swapping language embedding simple work?
  ○ Video-guided machine translation: visualize attention over video encoding? Vector encoding loss spatial information -- how does attention help if the key reference object appear in all frames?

● More experiments
  ○ Video-guided machine translation: English to Chinese?
  ○ Language model pretraining?
  ○ Video encoding that retain spatial information?
  ○ Since no metric is perfect -- test it with AREL learned reward?

● Future work
  ○ VMT looks like a really interesting task – improve machine translation quality on even harder dataset?
  ○ Single video + multilingual caption => single caption + multichannel video -- better video encoding?