Visually Grounded Reasoning across Languages and Cultures

Fangyu Liu*  Emanuele Bugliarello*  Edoardo M. Ponti  Siva Reddy  Nigel Collier  Desmond Elliott

University of Cambridge  University of Copenhagen  Mila/McGill University  Mila/McGill University  University of Cambridge  University of Copenhagen

fl399@cam.ac.uk  emanuele@di.ku.dk 
Evaluation data for cross-lingual V&L transfer

Task
Predict if a caption is True/False for 2 images

Label: True
V&L Data

Languages
- Mostly in English
- Or in a few Indo-European Languages

Image sources
- Mostly from ImageNet, MS COCO and Visual Genome
- Reflecting North American and European cultures

Implications for V&L models
- Narrow linguistic/cultural domain
- No way to assess their real-world comprehension
Biases in Image Collections: An ImageNet Study

Concept selection

Candidate image retrieval

Manual cleanup
Overview of MaRVL Collection

Native speaker-driven protocol

1. Concept Selection
2. Image Selection
3. Annotation
Universal Concepts

Concepts that are shared across cultures

From the *Intercontinental Dictionary Series* (Key & Comrie, 2015)

18 chapters with concrete objects & events

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Semantic Field</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animal</td>
<td>Bird, mammal</td>
</tr>
<tr>
<td>Food and Beverages</td>
<td>Food, Beverages</td>
</tr>
<tr>
<td>Clothing and grooming</td>
<td>Clothing</td>
</tr>
<tr>
<td>The house</td>
<td>Interior, exterior</td>
</tr>
<tr>
<td>Agriculture and vegetation</td>
<td>Flower, fruit, vegetable,</td>
</tr>
<tr>
<td>Basic actions and technology</td>
<td>Utensil/tool</td>
</tr>
<tr>
<td>Motion</td>
<td>Sport</td>
</tr>
<tr>
<td>Time</td>
<td>Celebrations</td>
</tr>
<tr>
<td>Cognition</td>
<td>Education</td>
</tr>
<tr>
<td>Speech and language</td>
<td>Music (instruments), visual arts</td>
</tr>
<tr>
<td>Religion and belief</td>
<td>Religion</td>
</tr>
</tbody>
</table>

OUR BIAS
Language-Specific Concepts

Defined by native speakers

• Commonly seen or representative in their culture
• Ideally, physical and concrete

SPORT

https://tr.wikipedia.org/wiki/Tenis
https://tr.wikipedia.org/wiki/Basketbol

5–10

Top-5
86-96 concepts per language
72.4% by at least 3 annotators
MaRVL concepts are in more languages

MaRVL concepts are in more families

MaRVL concepts are in more macroareas
Images

Collected by native speakers

- Representative of the language population
- ......
Captions

MATCH 4 PAIRS AT RANDOM

WRITE CAPTION TRUE ONLY FOR 2 PAIRS

VALIDATE ANNOTATIONS

FINAL VALIDATION

右图中的人在发球，左图中的人在接球。
(The man in the right image is serving a ball while the man in the left image is returning a ball.)
Examples

两图中至少有一张图里面是一口鸳鸯锅
At least one of the two pictures shows a mandarin duck pot
**Label:** True

兩張圖都有很多薩拉瓦達
Both images contain a lot of masala vadas
**Label:** False
Human Quality Assessment

Fleiss' Kappa

>0.81: almost perfect annotator agreement (Landis and Koch, 1977)
Key Statistics & Limitations

Key statistics
- 423 concepts
  - 96 not in WordNet
- 5464 images
- 1390 unique captions
- 5560 data points

Limitations
- Low-resource language annotators
- Wikipedia as a proxy for concepts
- ......
Experimental Setup

Models
• 5 V&L BERTs from VOLTA (Bugliarello+, 2021)
• 2 new multilingual UNITER models
  ‣ mUNITER: initialised from mBERT
  ‣ xUNITER: initialised from XLM-R

Multilingual Multimodal Pre-training
Following M³P (Ni+, 2021)
• 104 Wikipedia (mBERT): MLM
• Conceptual Captions: MLM + MRM + ITM

Fine-tuning
• Train on English NLVR2 (Suhr+, 2019)
• Test on MaRVL
  ‣ Multilingual models in a “zero-shot”, cross-lingual fashion
  ‣ English models in a “translate-test” approach
Main Results

mUNITER and xUNITER are on par in NLVR2

Zero-shot transfer: -10–20%

Translate-test: -10%
Turkish is surprisingly easy
Disentangling Distribution Shifts: A zh Study

2 distribution shifts in MaRVL
- Cross-lingual transfer (XLT)
- Out-of-distribution (OOD) concepts

Controlled study on **MaRVL-zh** with xUNITER

<table>
<thead>
<tr>
<th>Test Language</th>
<th>en</th>
<th>zh</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLVR2</td>
<td>72.8</td>
<td>57.1*</td>
</tr>
<tr>
<td>OOD</td>
<td>64.4</td>
<td>63.3</td>
</tr>
</tbody>
</table>

Translate MaRVL-zh into en -8%

Translate NLVR2 1K* into zh -15%

Translate NLVR2 Train into zh -9%
Conclusions & Next Steps

Show that concepts and images in existing V&L datasets are not cross-lingual/cultural

New protocol for concept and image selection entirely driven by native speakers

Introduce MaRVL: A V&L reasoning dataset in 5 typologically diverse languages

Develop and benchmark multilingual V&L BERTs
  • Performance can be at chance level

Adapting V&L models to culture-specific concepts

Multicultural object detectors

Code, data & annotation guidelines are online
  • marvl-challenge.github.io
  • Add your own language!