Visually Grounded Reasoning across Languages and Cultures

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

EMNLP 2021

MaRVL: Multicultural Reasoning over Vision and Language

marvl-challenge.github.io

Evaluation data for cross-lingual V&L transfer

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

Languages: Indonesian (id) Swahili (sw) Tamil (ta) Turkish (tr) Mandarin Chinese (zh)

• New protocol driven by native speakers
• Universal concepts from Intercontinental Dictionary Series

Our Approach

Existing V&L data

Languages
• Mostly in English or in another IE language

Limits of translation-based dataset creation

A street organ stands in front of a …

An unusual looking vehicle parked …

Example from van Miltenburg+ (INLG’17)

Image Sources
• Scraped from the web
• Reflect North American and European cultures

Density map of geographical distribution of images in ImageNet (DeVries+, CVPRW’19)

We should not simply translate existing V&L data!

86-96 concepts per language

MaRVL concepts are in more families

Both images contain a lot of masala vadas

Label: False

Challenges
• Cross-lingual transfer (XLT)
• Out-of-distribution (OOD) generalisation

XLT
72.8
57.1*

OOD
64.4
63.3

Translate MaRVL-zh into en
-8%

Translate NLVR2 en into zh
-15%

Results

Experiments

Models
• 5 English V&L BERTs from VOLTA (Bugliarello+, 2021)
• 2 new multilingual models: mUNITER & xUNITER

Fine-tuning
Train on English NLVR2 (Suhr+, 2019)
Test on MaRVL
• Multilingual models: zero-shot, cross-lingual
• English models: translate-test