Ideals and Compromises in Multilingual Multimodal Learning

Desmond Elliott
University of Copenhagen

de@di.ku.dk @delliott

UNIVERSITY OF COPENHAGEN

Advances in Vision-and-Language

IMAGE CAPTIONING

A sandwich on a cutting board with a knife.

Cornia+ (CVPR'20), Ramos+ (arXiv'22)

MULTIMODAL REASONING

age contains twice the number of dogs as the and at least two dogs in total are standing.

True

CL'19), Liu and Bugliarello+ (EMNLP21)

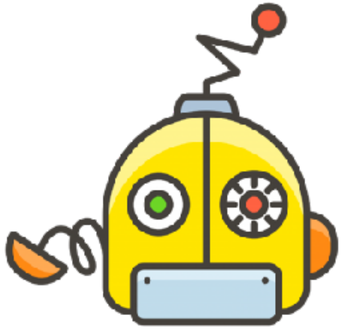
DATA

TEXT-TO-IMAGE SYNTHESIS

A baby daikon radish in a tutu walking a dog

Ramesh+ (ICML'20), Rombach+ (CVPR'22)

SELF-SUPERVISED LEARNING

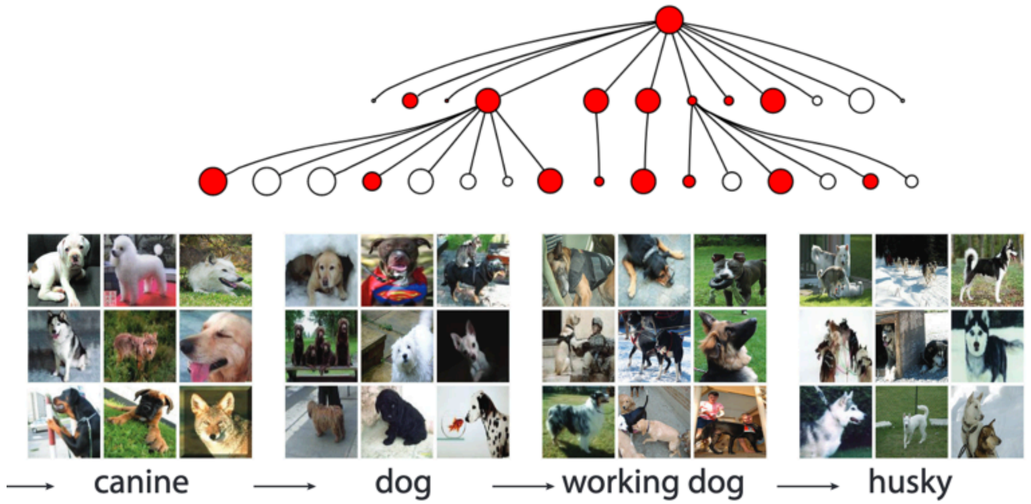


- CAPTIONING
- VQA
- REASONING
- GROUNDING

Typical Vision and Language

ImageNet (Deng+ CVPR'09)

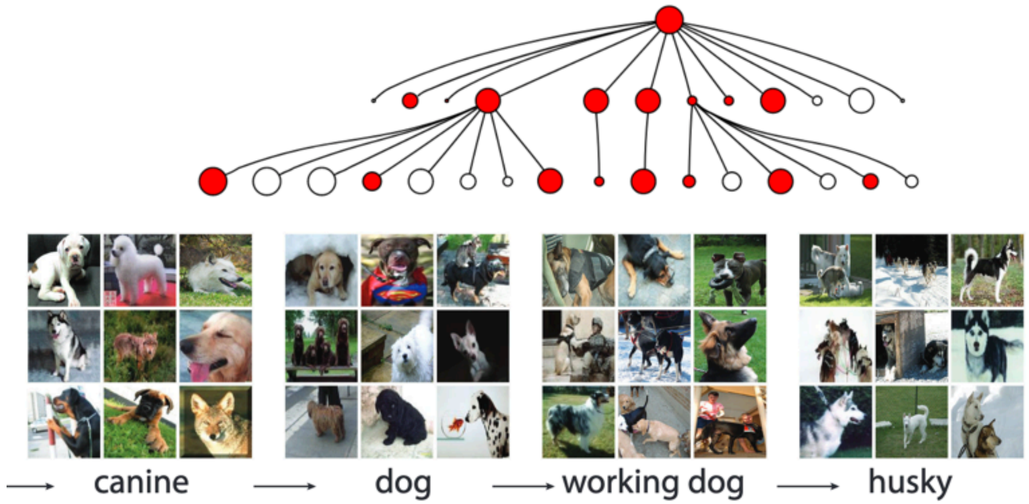
- Train visual encoders
- Millions of labelled images
- Derived from the WordNet hierarchy



Typical Vision and Language

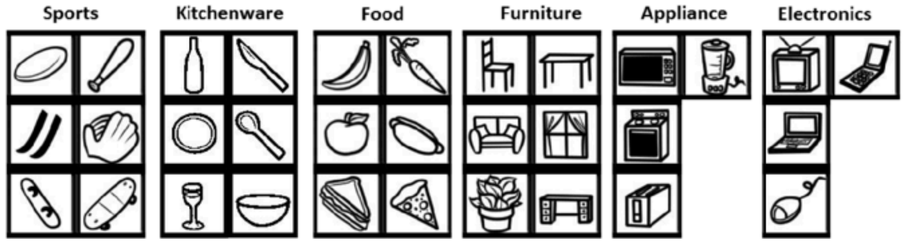
ImageNet (Deng+ CVPR'09)

- Train visual encoders
- Millions of labelled images
- Derived from the WordNet hierarchy



Common Objects in Context (Lin+ ECC V'14)

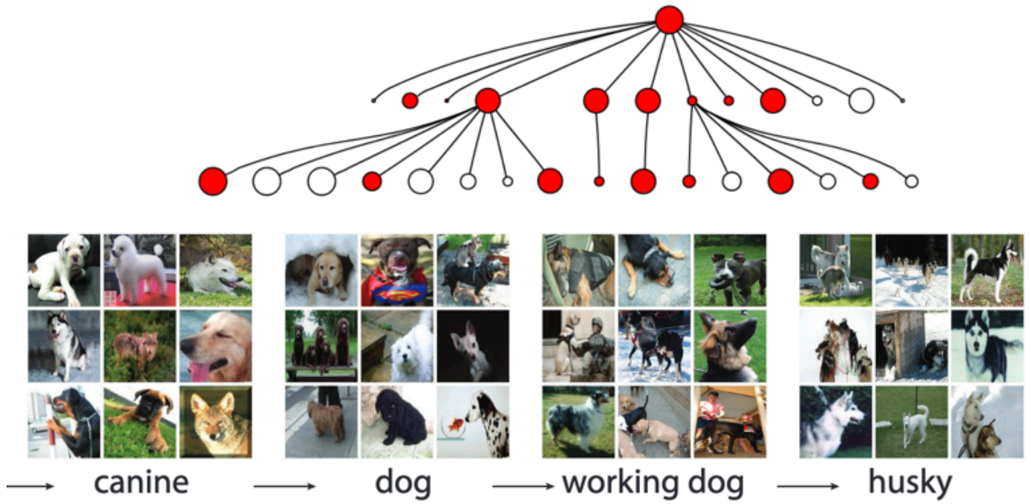
- Train and evaluate multimodal models
- 330K labelled images
  - 80 types of commonly occurring objects



Typical Vision and Language

ImageNet (Deng+ CVPR'09)

- Train visual encoders
- Millions of labelled images
- Derived from the WordNet hierarchy



Visual Genome (Krishna+ IJCV'17)

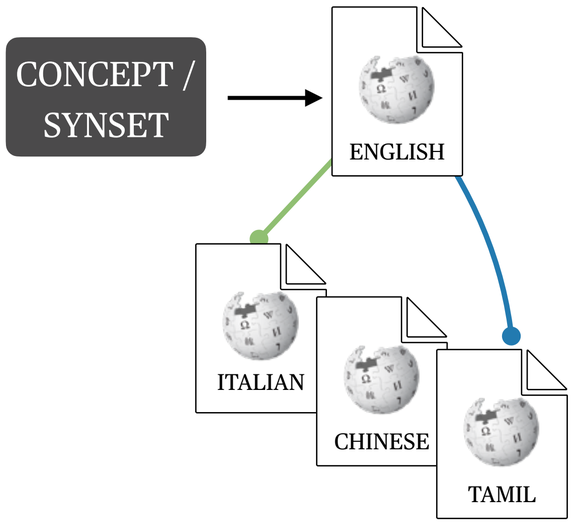
- Train and evaluate multimodal models
- 110K densely annotated images
- Derived from COCO and YFCC100M dataset

Which Languages Are Represented?

- ImageNet, COCO and Visual Genome are based on English WordNet hierarchy
- How cross-lingual are these concepts?
  - Idea: estimate cross-linguality using Wikipedia as a proxy

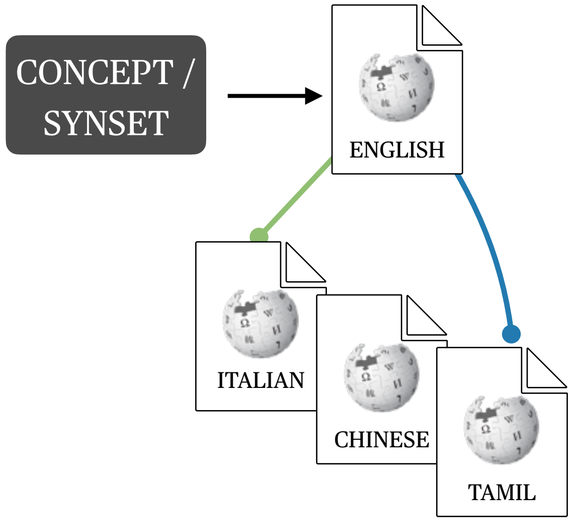
Which Languages Are Represented?

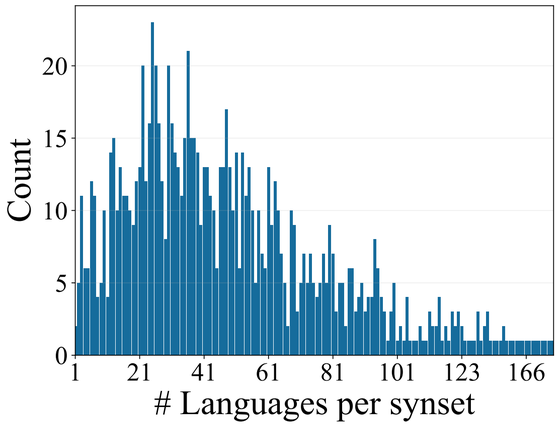
- ImageNet, COCO and Visual Genome are based on English WordNet hierarchy
- How cross-lingual are these concepts?
  - Idea: estimate cross-linguality using Wikipedia as a proxy



Which Languages Are Represented?

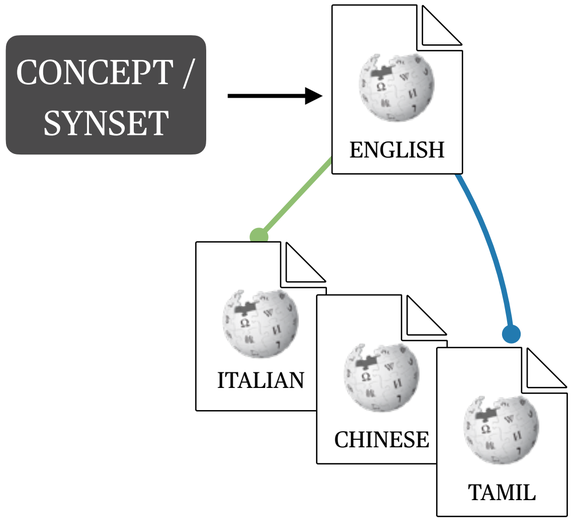
- ImageNet, COCO and Visual Genome are based on English WordNet hierarchy
- How cross-lingual are these concepts?
  - Idea: estimate cross-linguality using Wikipedia as a proxy

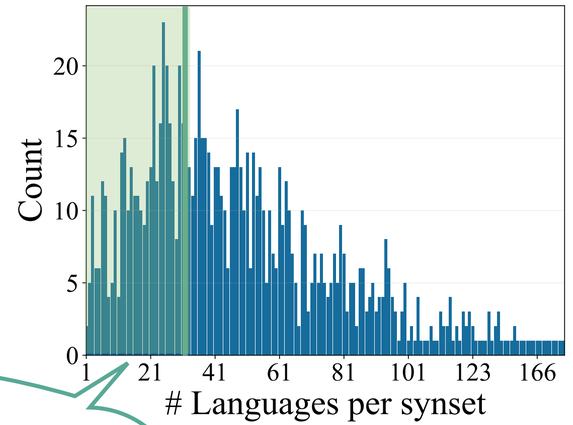




Which Languages Are Represented?

- ImageNet, COCO and Visual Genome are based on English WordNet hierarchy
- How cross-lingual are these concepts?
  - Idea: estimate cross-linguality using Wikipedia as a proxy

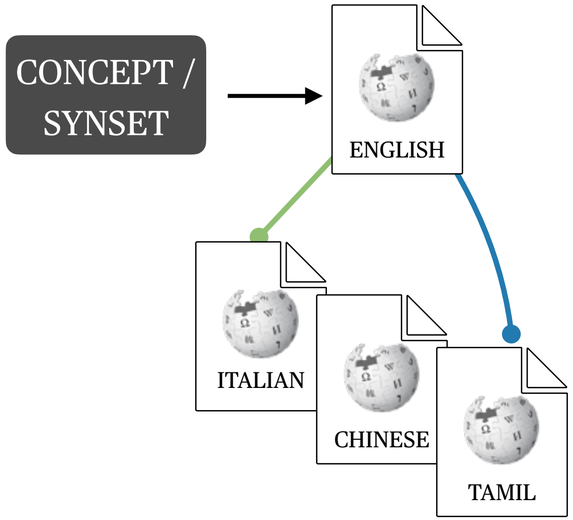


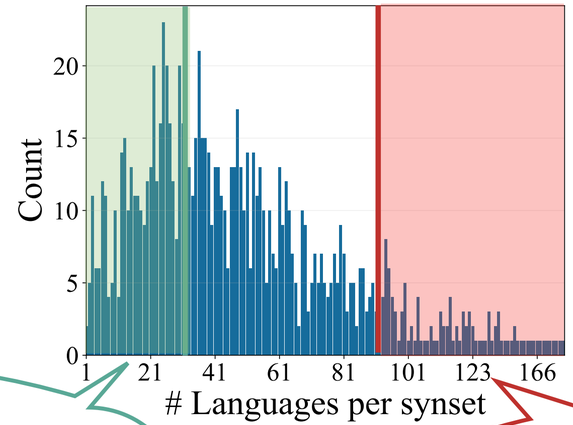


Most concepts in <30 languages

Which Languages Are Represented?

- ImageNet, COCO and Visual Genome are based on English WordNet hierarchy
- How cross-lingual are these concepts?
  - Idea: estimate cross-linguality using Wikipedia as a proxy



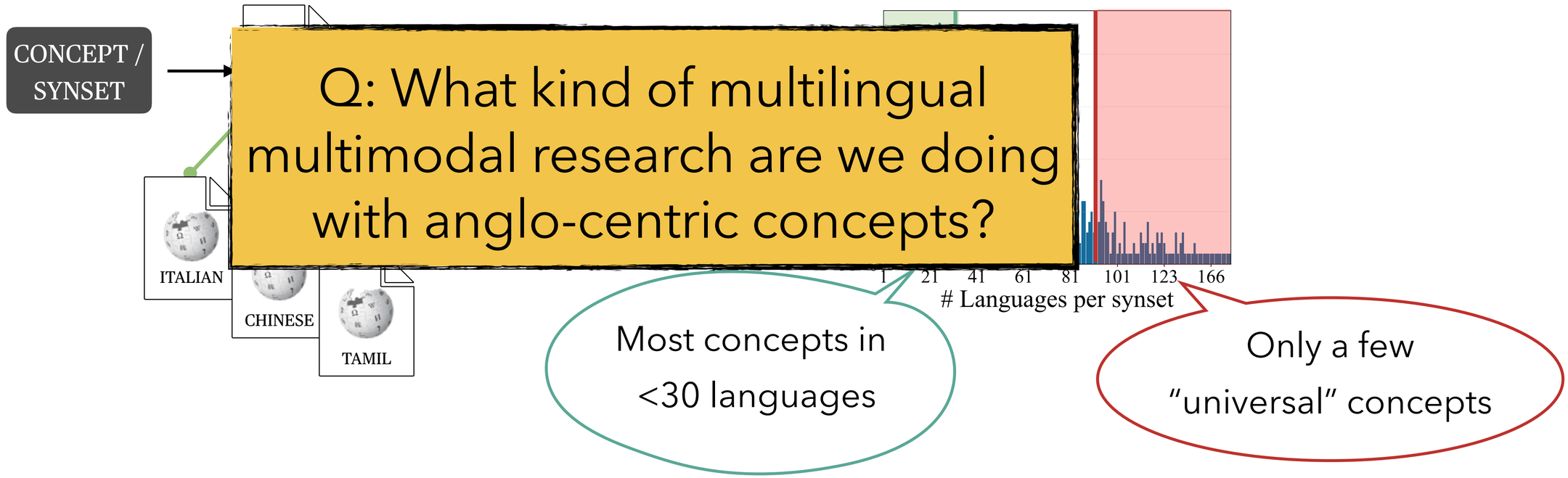


Most concepts in <30 languages

Only a few "universal" concepts

Which Languages Are Represented?

- ImageNet, COCO and Visual Genome are based on English WordNet hierarchy
- How cross-lingual are these concepts?
  - Idea: estimate cross-linguality using Wikipedia as a proxy



Concepts in Cultural Context

- Some concepts are most immediately understood within a cultural background

Culture: The way of life of a collective of people that distinguishes them from other people (Mora, 2013; Shweder et al. 2007).

Pilota / Jai-alai

Sanxian / Shamisen

Clavie

Today: Rethinking Vision and Language

Languages

- Mostly in English
- Or some Indo-European Languages

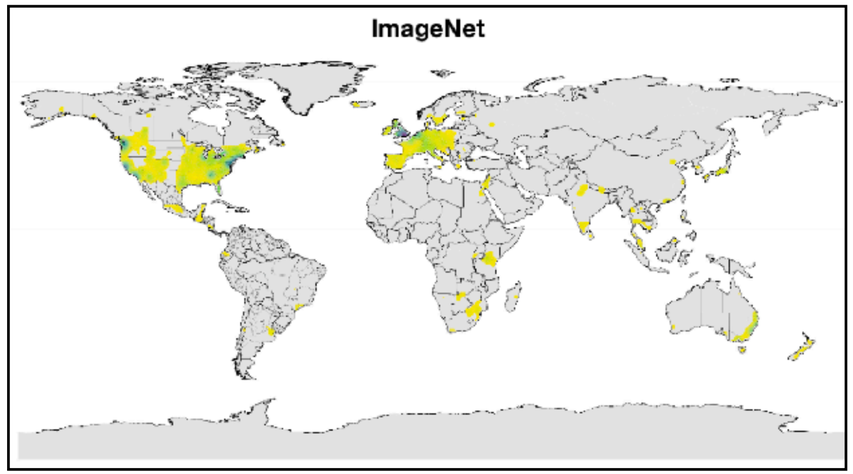
ENG: An unusual looking vehicle …

NLD: Een mobiel draaiorgel …

Example from van Miltenburg+ 2017

Image sources

- Mostly from ImageNet or COCO
- Reflecting North American and European cultures



Density map of geographical distribution of images in ImageNet (DeVries+, 2019)

Implications for V&L models

- Narrow linguistic/cultural domain
- No way to assess their real-world comprehension

Overview

1. Culture-specific Multilingual Multimodal Data
   Liu* and Bugliarello* et al. (EMNLP 2021)

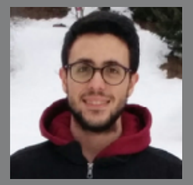
2. Measuring Progress with the IGLUE Benchmark
   Bugliarello et al. (ICML 2022)

3. Surprising Effectiveness of Machine Translated Data
   Liu et al. (Findings of EMNLP 2022)

Visually Grounded Reasoning across Languages and Cultures

Best Paper Award
EMNLP 2021


F. Liu*


E. Bugliarello*


E.M. Ponti


S. Reddy


N. Collier


D. Elliott

MaRVL Multicultural Reasoning over Vision and Language

Representative of annotators' cultures

5 typologically diverse languages

Independent, culture-specific annotations

MaRVL-id Bola basket

MaRVL-sw Mpira wa kikapu

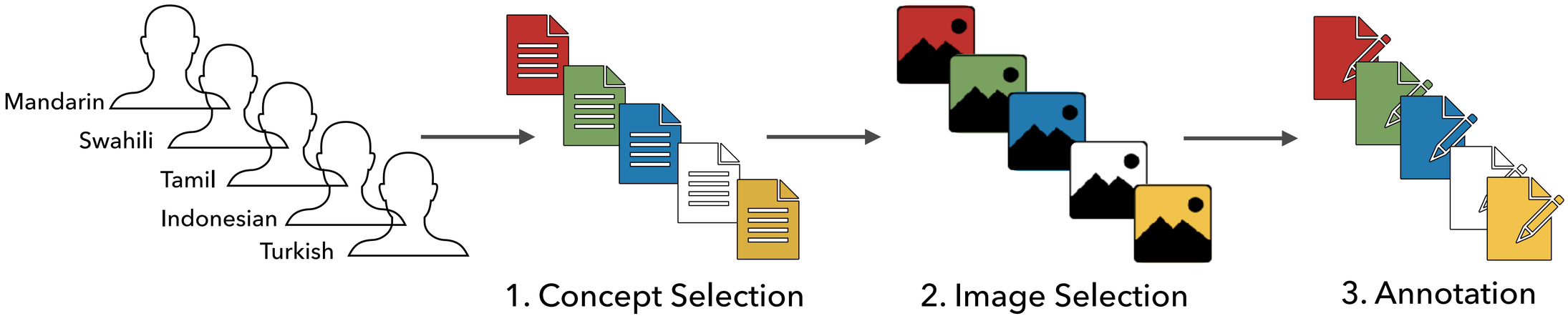
MaRVL-tr Basketbol

MaRVL-zh 篮球

MaRVL-ta கூடைப்பந்தாட்டம்

Collecting MaRVL data

Native speaker-driven protocol



What Type of Data?

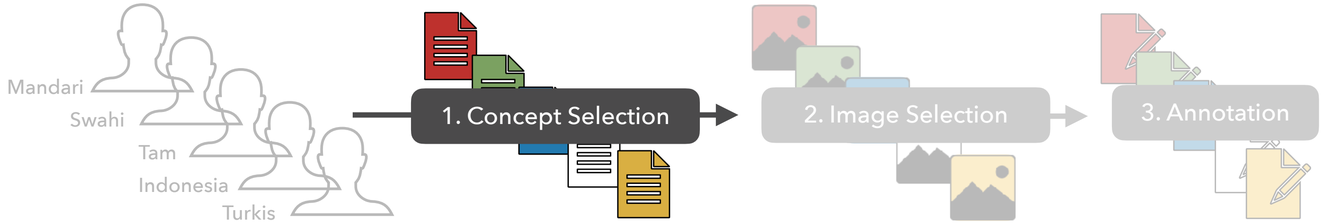
- **Visually-grounded reasoning** (Suhr+ ACL'19)
- **Task**: Predict whether the text x describes a pair of images v_1 v_2
- **Datapoint**: two images (v_1, v_2) paired with a sentence (x)

இரு படங்களில் ஒன்றில் இரண்டிற்கும் மேற்பட்ட மஞ்சள் சட்டை அணிந்த வீரர்கள் காளையை அடக்கும் பணியில் ஈடுப்பட்டிருப்பதை காணமுடி.

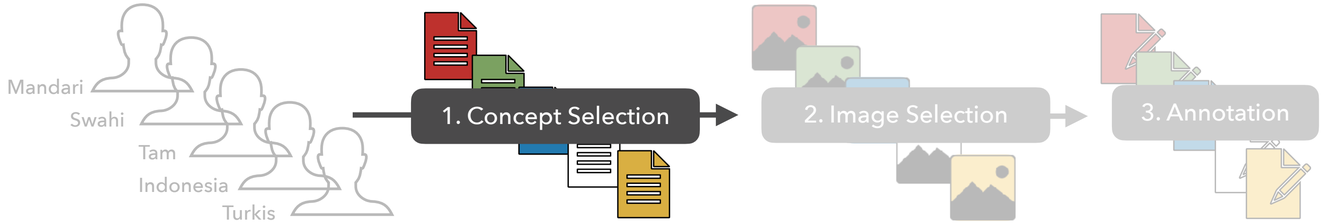
(In one of the two photos, more than two yellow-shirted players are seen engaged in bull taming.)

True

v_1 v_2 x Label



1. Concepts are commonly seen or representative in their culture
2. Concepts are ideally, physical and concrete



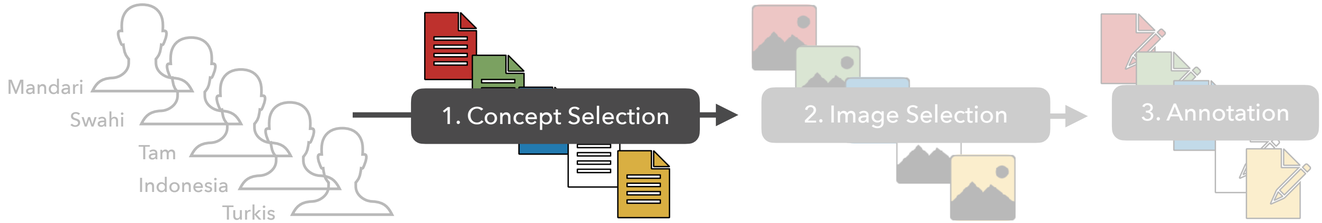
1. Concepts are commonly seen or representative in their culture
2. Concepts are ideally, physical and concrete

SPORT

Semantic Field

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

Intercontinental Dictionary Series
(Key & Comrie, 2015)



1. Concepts are commonly seen or representative in their culture
2. Concepts are ideally, physical and concrete

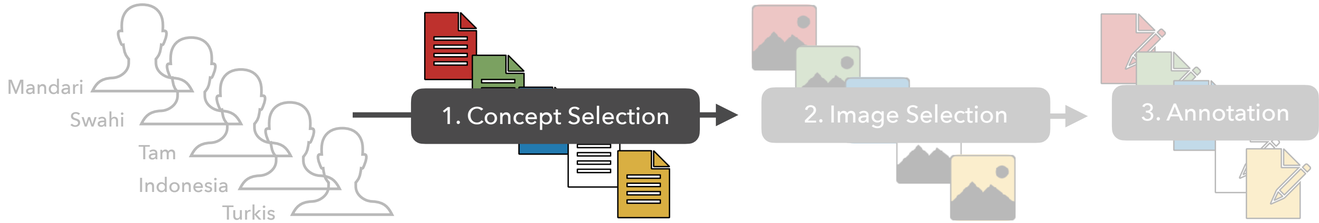
SPORT

Semantic Field

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

N native speakers

Intercontinental Dictionary Series
(Key & Comrie, 2015)



1. Concepts are commonly seen or representative in their culture
2. Concepts are ideally, physical and concrete

SPORT

Semantic Field

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

Intercontinental Dictionary Series
(Key & Comrie, 2015)

N native speakers

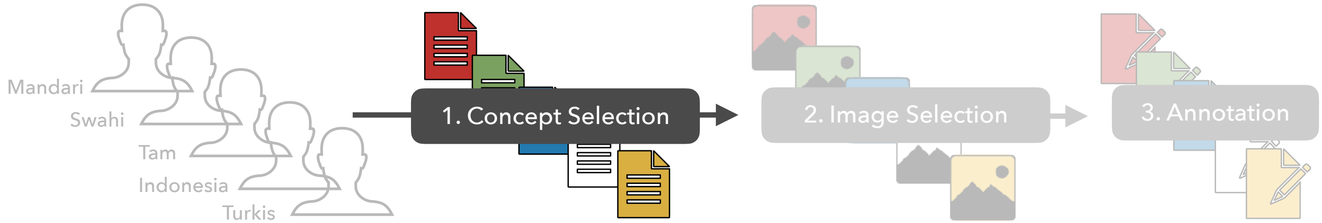
Find

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

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

....

5-10



1. Concepts are commonly seen or representative in their culture
2. Concepts are ideally, physical and concrete

SPORT

Semantic Field

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

Intercontinental Dictionary Series
(Key & Comrie, 2015)

N native speakers

Find

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

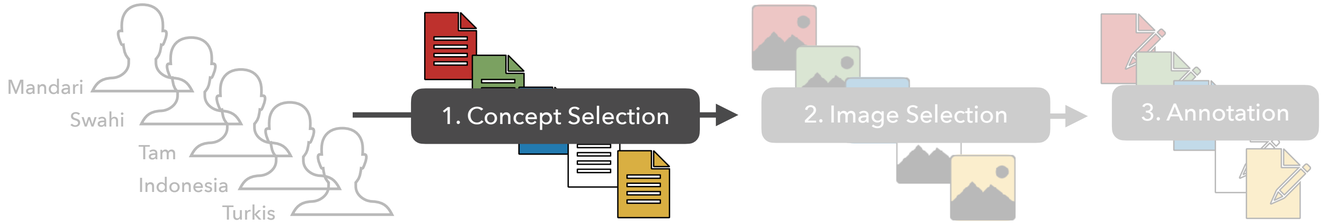
https://tr.wikipedia.org/wiki/Basketbol

....

5-10

Top-5 per semantic field

~90 concepts per language



1. Concepts are commonly seen or representative in their culture
2. Concepts are ideally, physical and concrete

SPORT

Semantic Field

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

Intercontinental Dictionary Series
(Key & Comrie, 2015)

N native speakers

Find

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

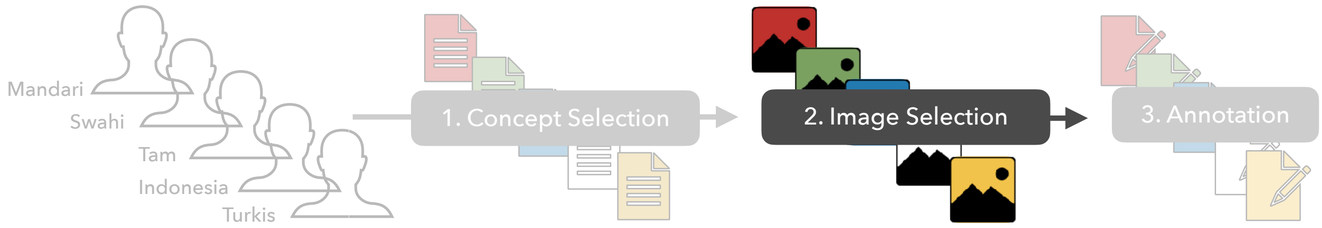
https://tr.wikipedia.org/wiki/Basketbol

....

5-10

Top-5 per semantic field

~90 concepts per language



- Representative of the language population
- NLVR2 (Suhr+ ACL'19) requirements

1. Contains more than one instance of the concept

2. Shows an instance of the concept interacting with other objects

3. Shows an instance of the concept performing an activity

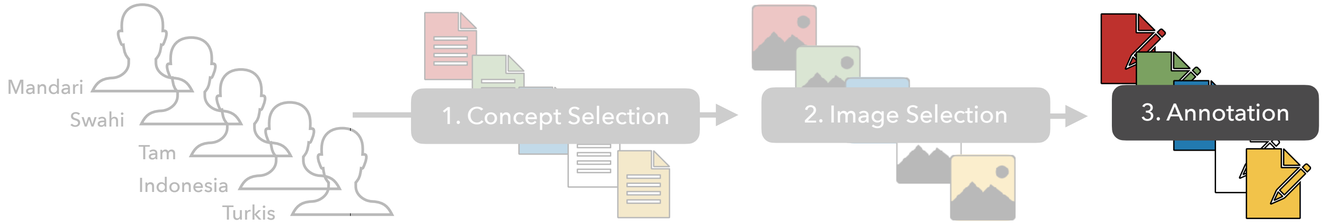
4. Displays a set of diverse objects or features

MaRVL-zh 花椰菜 (Cauliflower)

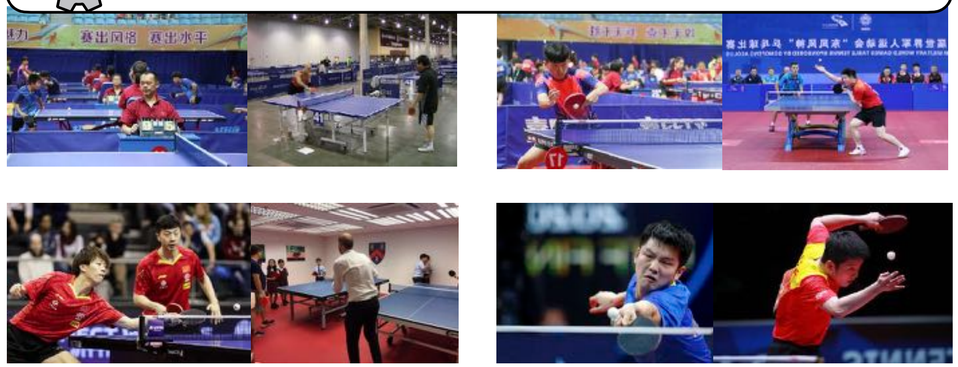
MaRVL-ta மோர் (Buttermilk)

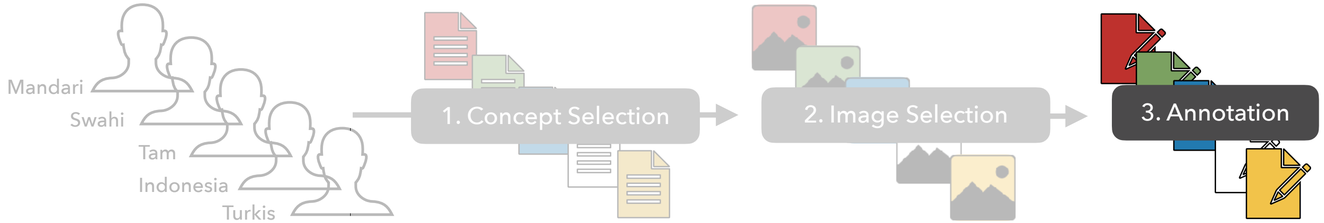
MaRVL-sw Jembe (Shovel)

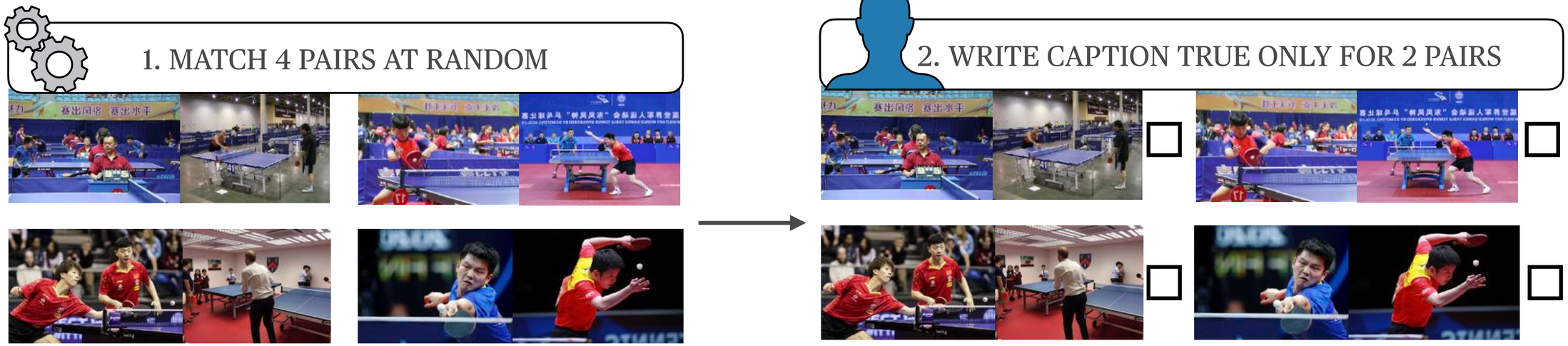
MaRVL-tr Rakı (Raki)

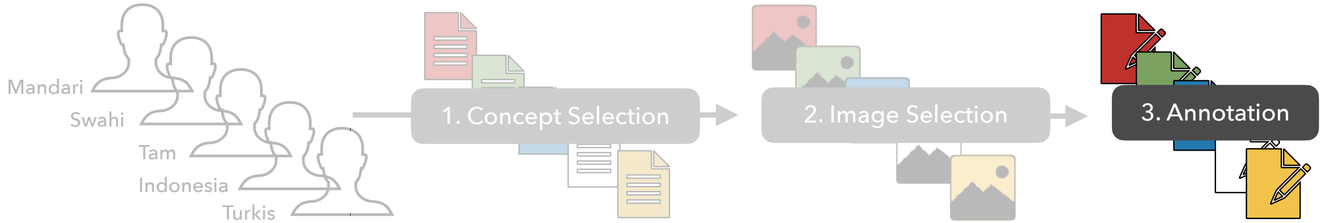


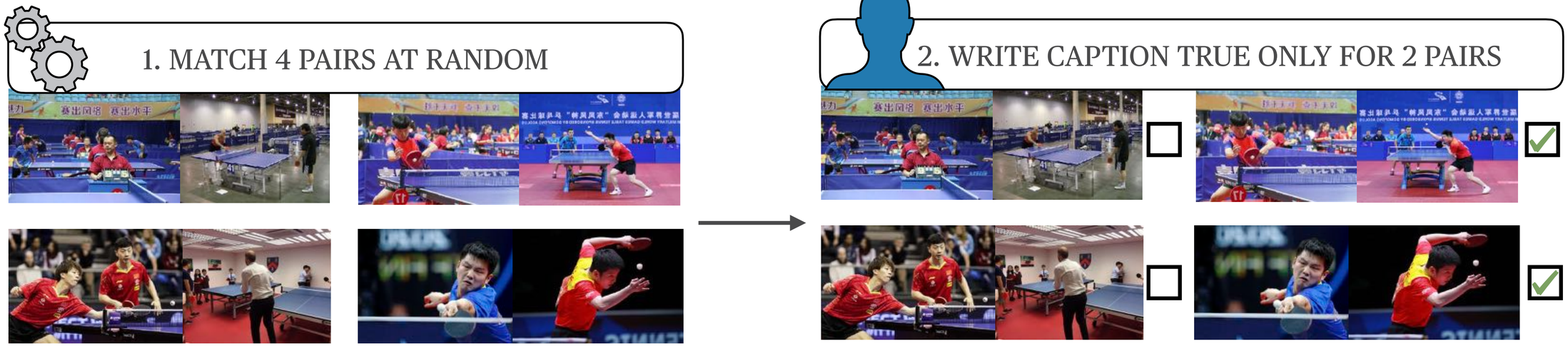
1. MATCH 4 PAIRS AT RANDOM



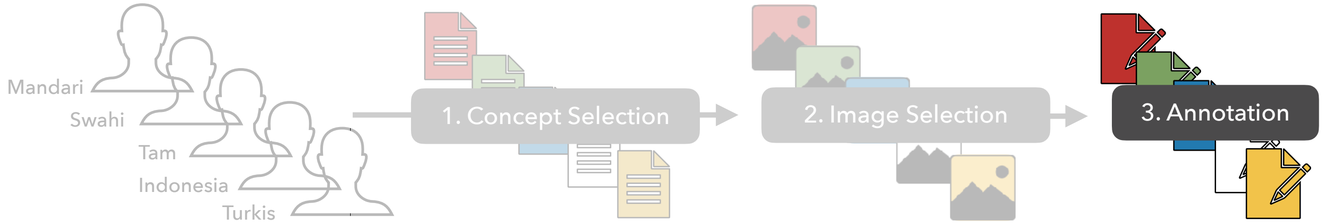

Mandari
Swahi
Tam
Indonesia
Turkis
1. Concept Selection
2. Image Selection
3. Annotation


1. MATCH 4 PAIRS AT RANDOM
2. WRITE CAPTION TRUE ONLY FOR 2 PAIRS

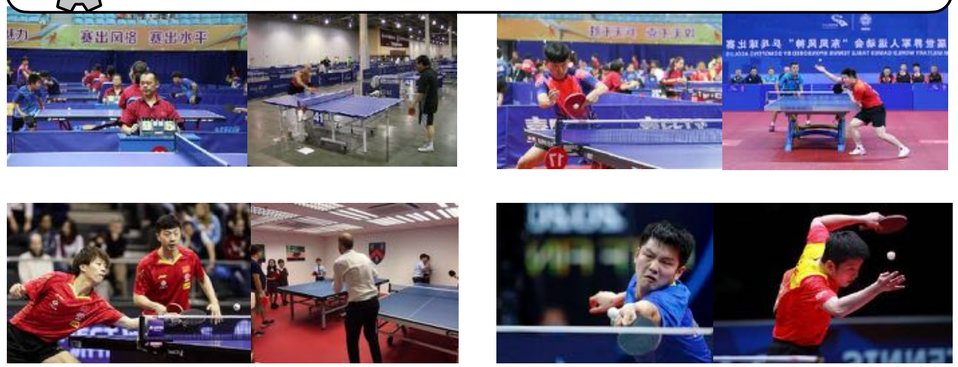




右图中的人在发球，左图中的人在接球。



1. MATCH 4 PAIRS AT RANDOM

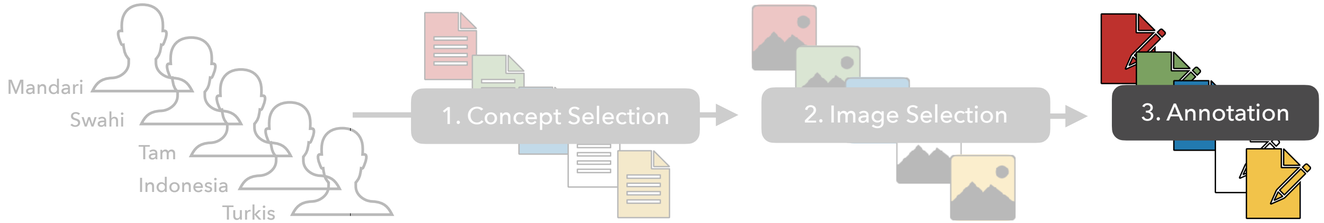


2. WRITE CAPTION TRUE ONLY FOR 2 PAIRS

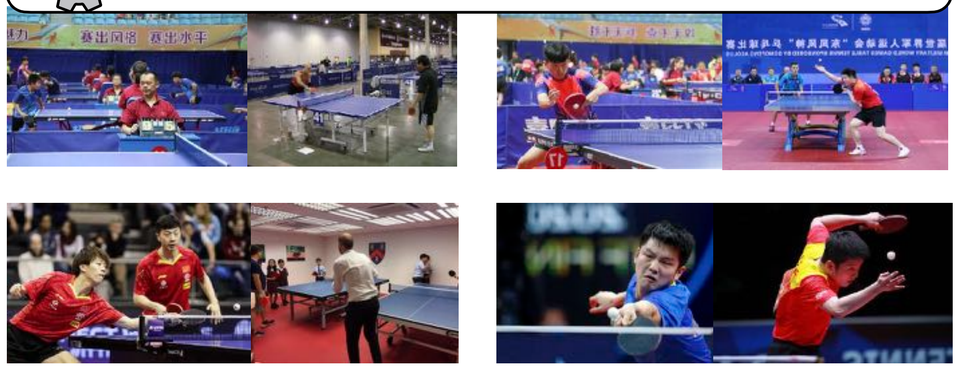
右图中的人在发球，左图中的人在接球。

3. VALIDATE ANNOTATIONS

右图中的人在发球，左图中的人在接球。



1. MATCH 4 PAIRS AT RANDOM

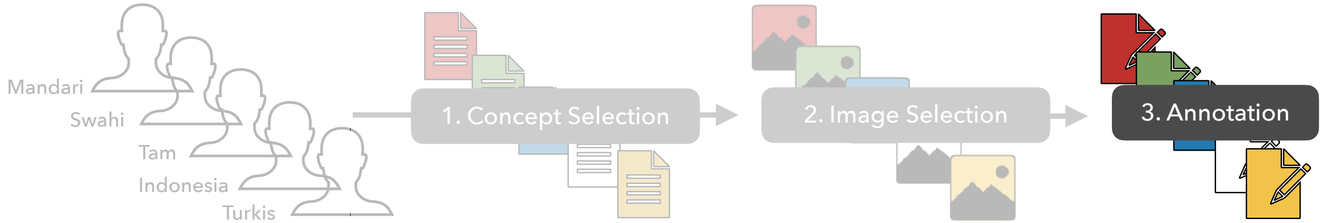


2. WRITE CAPTION TRUE ONLY FOR 2 PAIRS

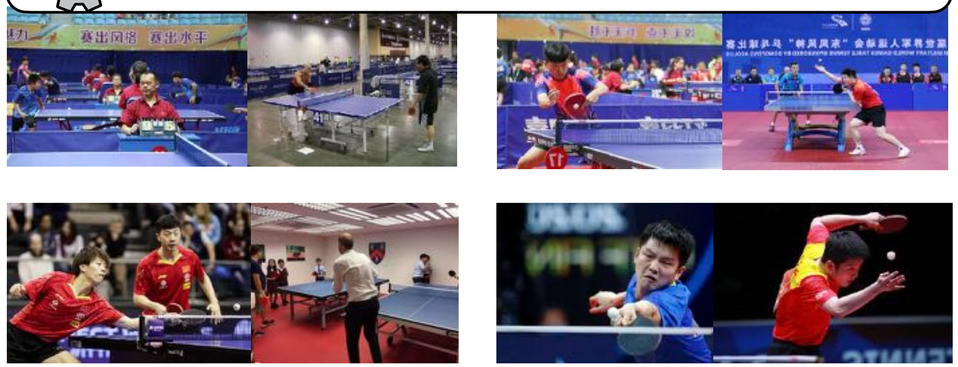
右图中的人在发球，左图中的人在接球。

3. VALIDATE ANNOTATIONS

右图中的人在发球，左图中的人在接球。



1. MATCH 4 PAIRS AT RANDOM

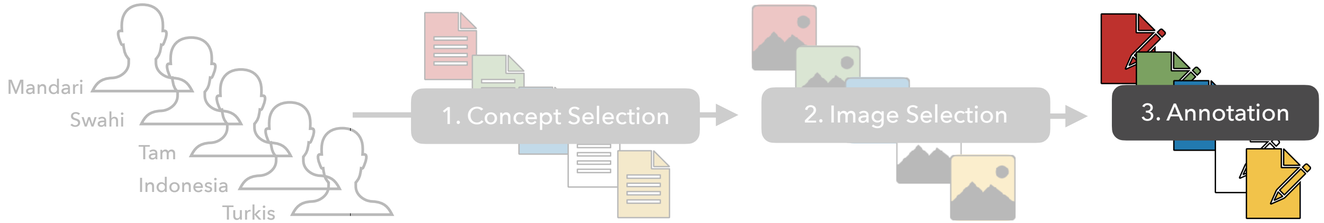


2. WRITE CAPTION TRUE ONLY FOR 2 PAIRS

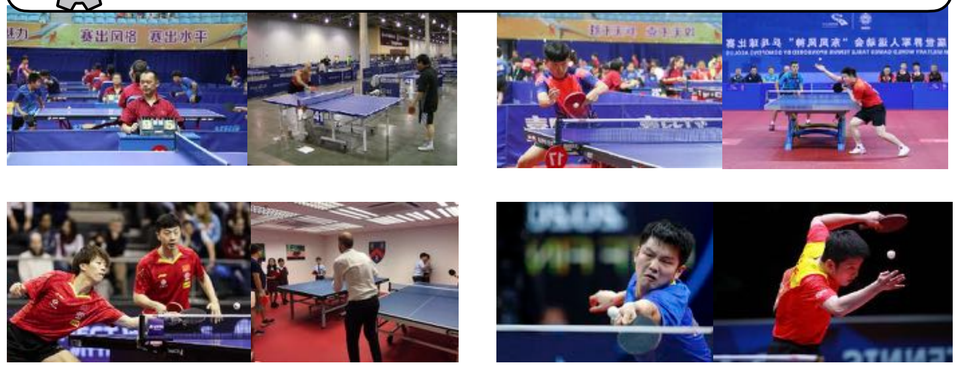
右图中的人在发球，左图中的人在接球。

3. VALIDATE ANNOTATIONS

右图中的人在发球，左图中的人在接球。



1. MATCH 4 PAIRS AT RANDOM



2. WRITE CAPTION TRUE ONLY FOR 2 PAIRS

右图中的人在发球，左图中的人在接球。

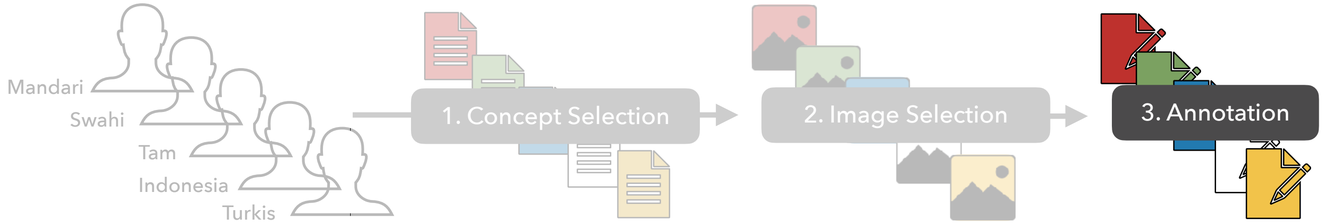
3. VALIDATE ANNOTATIONS

右图中的人在发球，左图中的人在接球。

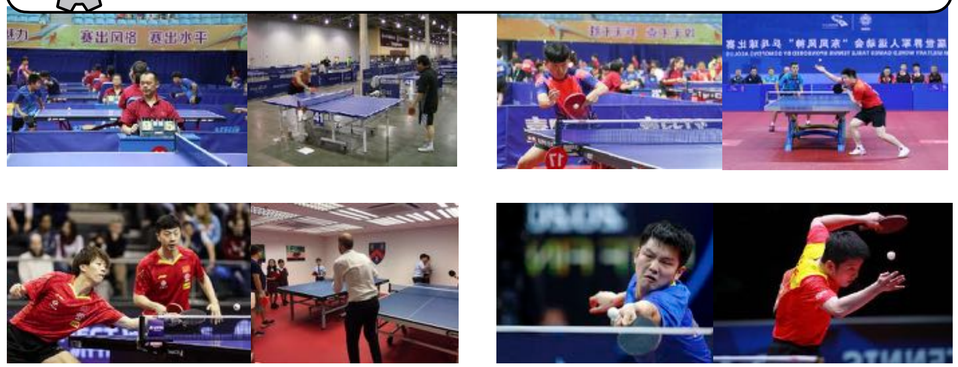
4. FINAL VALIDATION

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;
the man in the left image is returning a ball.)



1. MATCH 4 PAIRS AT RANDOM



2. WRITE CAPTION TRUE ONLY FOR 2 PAIRS

右图中的人在发球，左图中的人在接球。

3. VALIDATE ANNOTATIONS

右图中的人在发球，左图中的人在接球。

4. 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;
the man in the left image is returning a ball.)

Fleiss' kappa: 93%

Examples and Summary Statistics

MaRVL-tr **Kanun (çalgı)**

Görsellerden birinde dizlerinde kanun bulunan birden çok insan var

(In one of the images, there are multiple people with qanuns on their knees)

Label: True

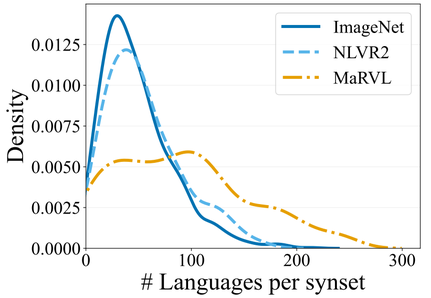
MaRVL-ta **மை (Vada)**

இரண்டு படங்களிலும் நிறைய மசால் வடைகள் உள்

(Both images contain a lot of masala vadas)

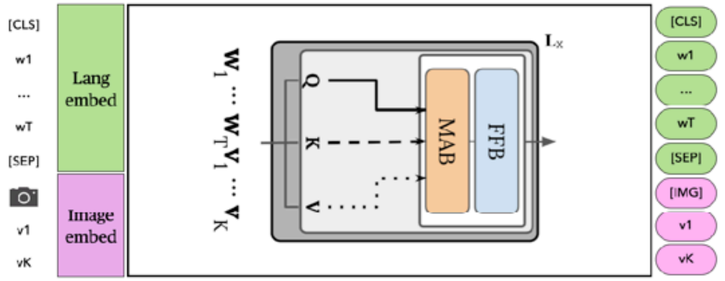
Label: False

- 5560 data points across 5 languages
- 423 concepts (96 not in WordNet)
- 1390 unique captions



Experimental Setup

- Multilingual multimodal models
    1. mUNITER: Initialised from mBERT
    2. xUNITER: Initialised from XLM-R



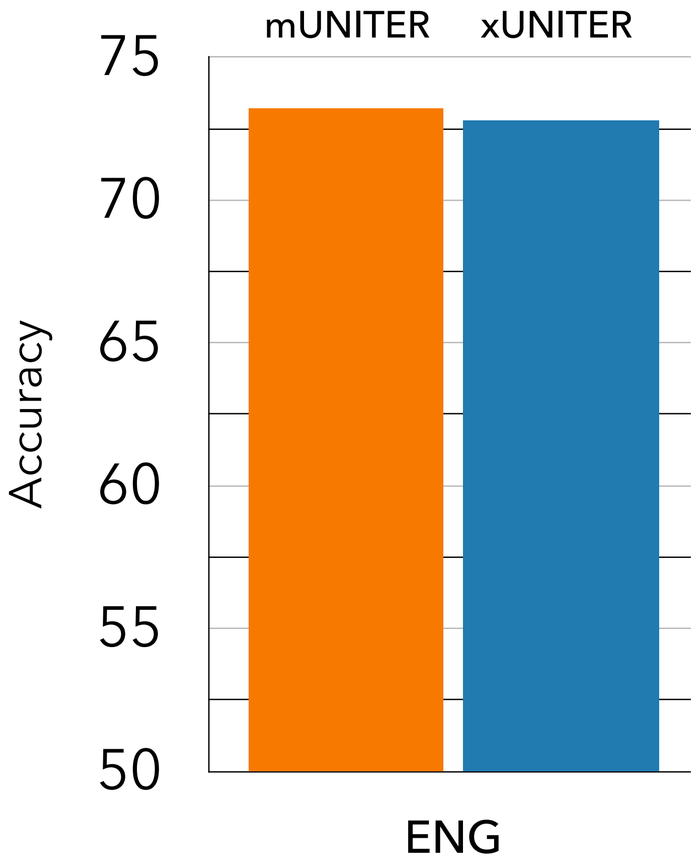
Pretraining Data

- Multimodal: Conceptual Captions in English (Sharma+, ACL'19)
- Language-only MLM: 104 Wikipedias

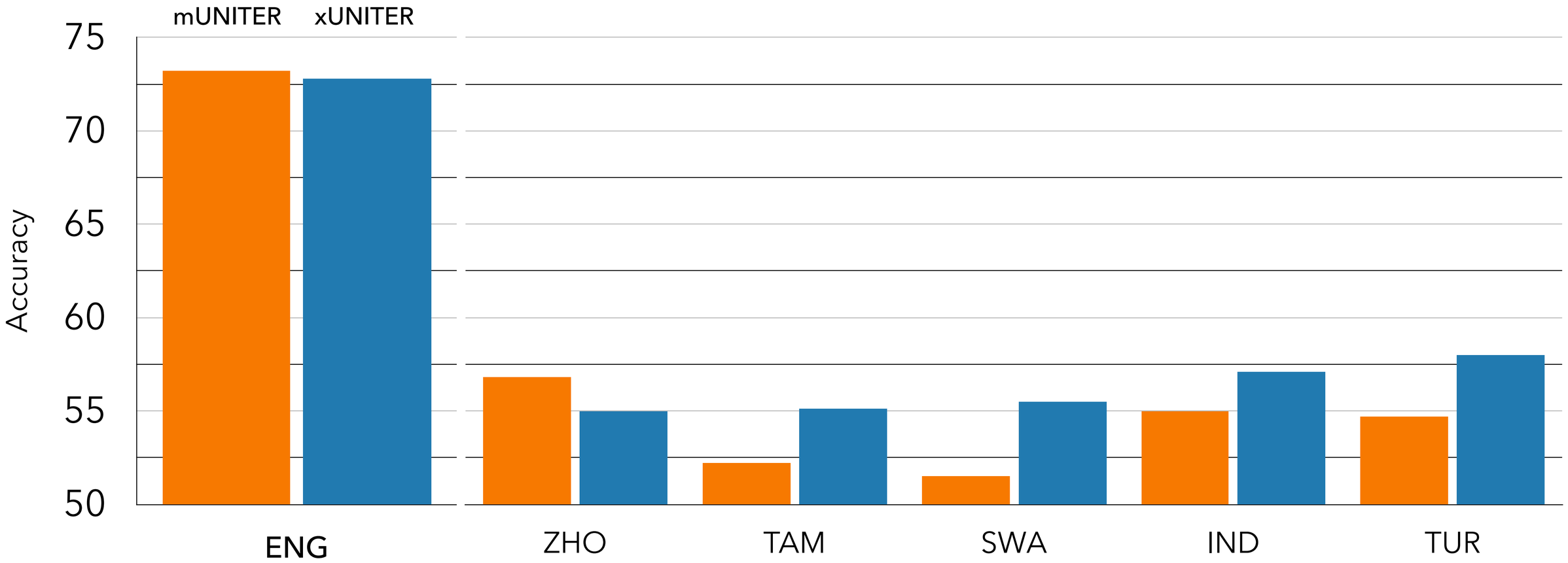
Fine-tuning

- Train on 86,373 examples in English NLVR2 (Suhr+, ACL'19)
- Zero shot cross-lingual evaluation on 5,560 datapoints in MaRVL

MaRVL Zero-shot Results



MaRVL Zero-shot Results



Zero-shot transfer: substantial drop in performance

Why is MaRVL so difficult?

There are two distribution shifts in MaRVL:

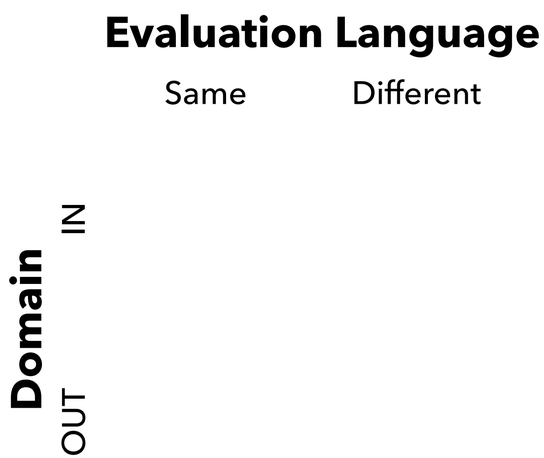
1. Cross-lingual transfer
2. Out-of-distribution concepts

Both analyses were completed by a native Mandarin Chinese speaker (Fangyu)

Why is MaRVL so difficult?

There are two distribution shifts in MaRVL:

1. Cross-lingual transfer
2. Out-of-distribution concepts

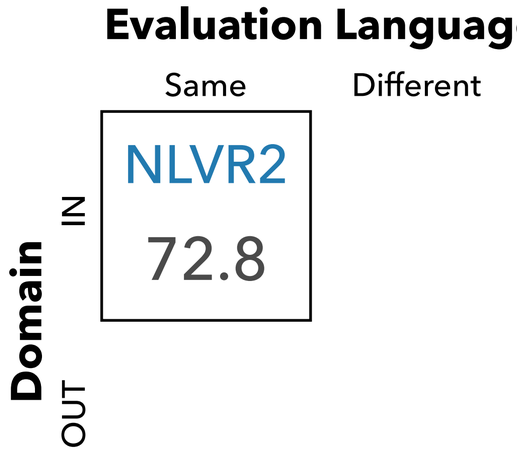


Both analyses were completed by a native Mandarin Chinese speaker (Fangyu)

Why is MaRVL so difficult?

There are two distribution shifts in MaRVL:

1. Cross-lingual transfer
2. Out-of-distribution concepts

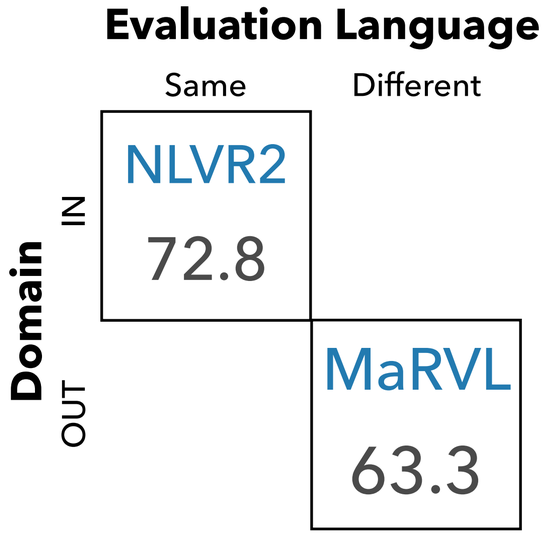


Both analyses were completed by a native Mandarin Chinese speaker (Fangyu)

Why is MaRVL so difficult?

There are two distribution shifts in MaRVL:

1. Cross-lingual transfer
2. Out-of-distribution concepts

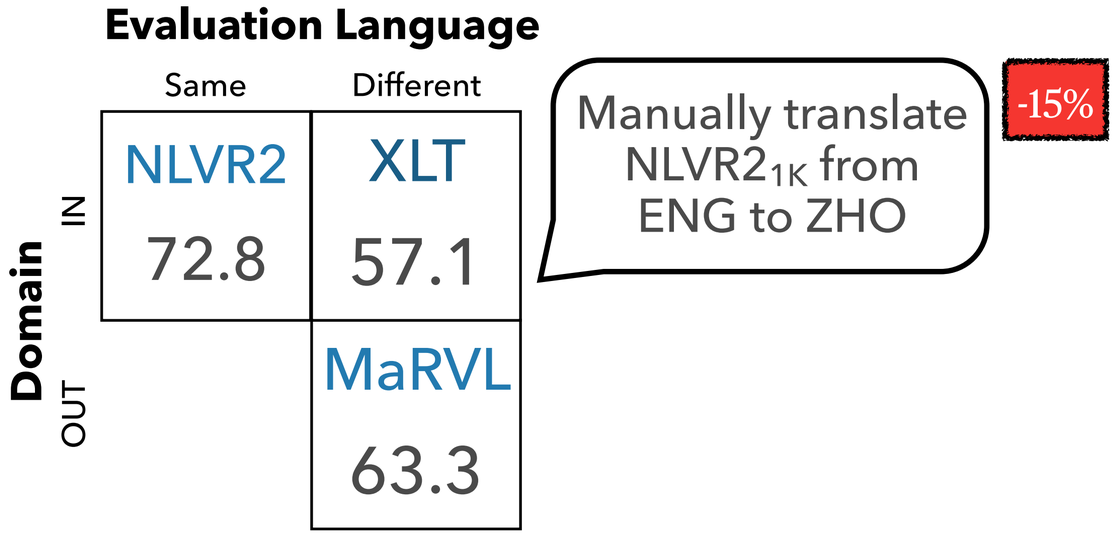


Both analyses were completed by a native Mandarin Chinese speaker (Fangyu)

Why is MaRVL so difficult?

There are two distribution shifts in MaRVL:

1. Cross-lingual transfer
2. Out-of-distribution concepts

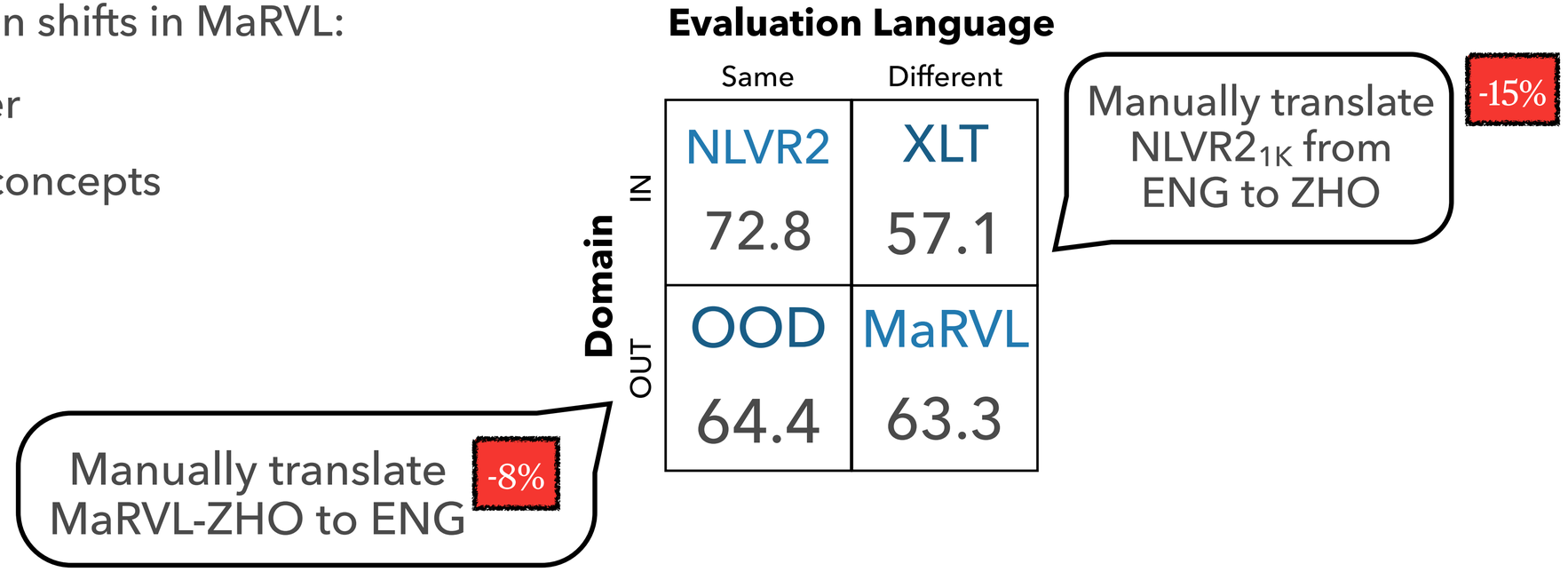


Both analyses were completed by a native Mandarin Chinese speaker (Fangyu)

Why is MaRVL so difficult?

There are two distribution shifts in MaRVL:

1. Cross-lingual transfer
2. Out-of-distribution concepts



Both analyses were completed by a native Mandarin Chinese speaker (Fangyu)

Why is MaRVL so difficult?

There are two distribution shifts in MaRVL:

1. Cross-lingual transfer
2. Out-of-distribution concepts

Evaluation Language

Domain	Same	Different
IN	NLVR2 72.8	XLT 57.1
OUT	OOD 64.4	MaRVL 63.3

Manually translate $NLVR2_{1K}$ from ENG to ZHO -15%

Manually translate MaRVL-ZHO to ENG -8%

Both of these distributional shifts are challenging

Both analyses were completed by a native Mandarin Chinese speaker (Fangyu)

Take-away Messages

- Which concepts are important for multilingual multimodal learning?
- Devise a new protocol for data creation driven by native speakers
- MaRVL: V&L reasoning dataset in 5 typologically diverse languages
- Performance is just above chance on this challenging dataset

Data and Models: https://marvl-challenge.github.io/

IGLUE: A Benchmark for Transfer Learning across Modalities, Tasks, and Languages

ICML 2022


E. Bugliarello


F. Liu


J. Pfeiffer


S. Reddy


D. Elliott


EM. Ponti


I. Vulić

IGLUE: Scaling Up Multilingual Multimodal Data

- MaRVL represents an **idealised** dataset
  - Native-speaker driven with culturally-relevant concepts

- Challenging to scale this to many tasks and languages
  - *Compromise* on the data representing different cultures
    - Why? High-quality data collection costs money and time

The Four IGLUE Tasks & Datasets

NATURAL LANGUAGE INFERENCE

Given an *image*-premise, predict if a *text*-hypothesis `entails`, `contradicts`, or is `neutral` to it

XVNLI *

🌍 5 Languages: Arabic, French, Russian and Spanish

NLI

لاعب كرة السلة يرمي كرة بثلاث نقط.

↓

`contradiction`

ENG: The basketball player shoots a three pointer

QUESTION ANSWERING

Given an image and question about it, predict the answer

xGQA (Pfeiffer+, 2022)

🌍 8 Languages: Bengali, German, Indonesian, Korean, Mandarin, Portuguese, Russian

QA

갈색이 아닌 음식은 어떤 것입니까?

↓

`vegetables`

ENG: Which kind of food is not brown?

VISUAL REASONING

Given two images and a textual description, predict if the description applies to both images (`true`/`false`)

MaRVL (Liu&Bugliarello+, 2021)

🌍 6 Languages: Indonesian, Mandarin, Swahili, Tamil, Turkish

Reasoning

两张图加起来总共超过五个人在打鼓，并且两张图中 的人所打鼓的种类不同。

↓

`True`

ENG: In total, there are more than five people playing drums in the two images combined and people in the two images are playing different kinds of drums.

IMAGE–TEXT RETRIEVAL

IR: Given a caption, retrieve its image
TR: Given an image, retrieve its caption

xFlickr&CO *

🌍 8 high-resource languages

WIT (Srinivasan+, 2021)

🌍 11 diverse languages

Retrieval

Мужчины и женщины в черных платьях и костюмах держат ноты в руках и поют хором.

↓ ↑

ENG: A group of men and women dressed in formal black dresses and suits holding their music books and singing.

IGLUE Benchmark Overview

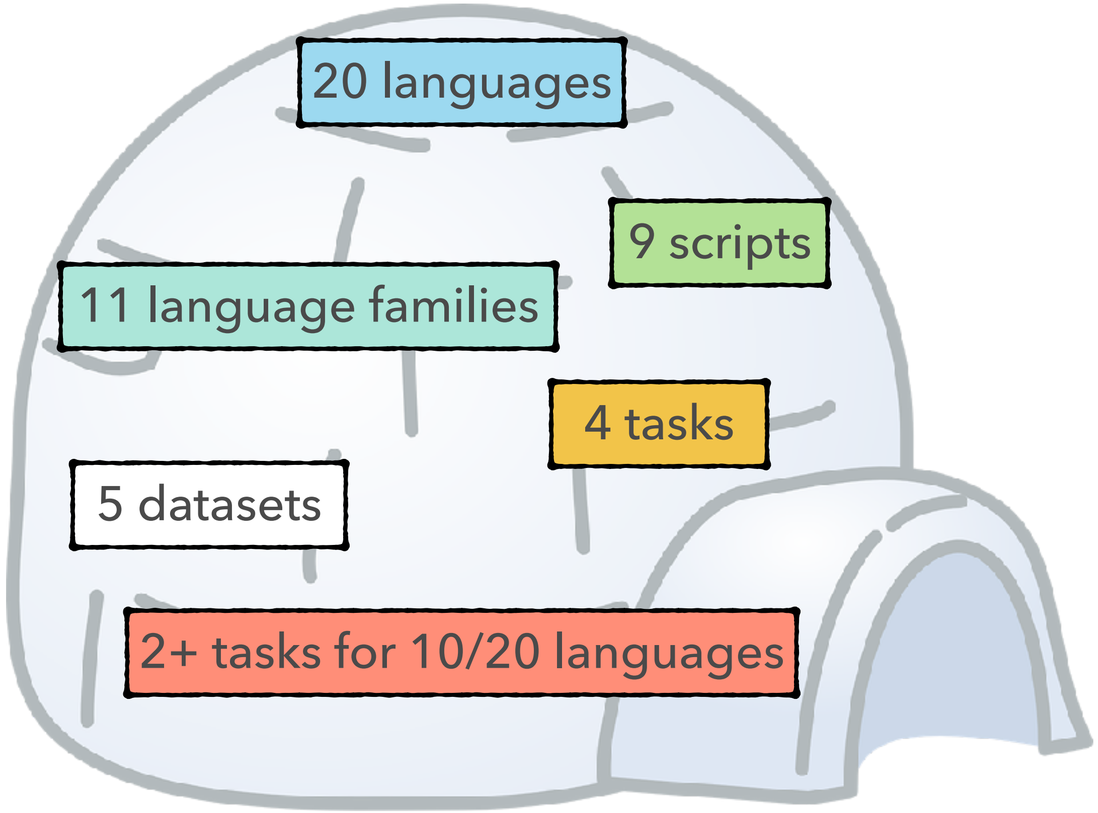


Table 2. Benchmark languages and tasks. English is only used used for training. Legend: ✔✔ train & test sets; ✔ test-only data.

	Language			NLI	QA	Reasoning	Retrieval	
Name	Code	Family	Script	XVNLI	xGQA	MaRVL	xFlickr&CO	WIT
English	ENG	Indo-E	Latin	✔✔	✔✔	✔✔	✔✔	✔✔
Arabic	ARB	Afro-A	Arabic	✔✔				✔
Bengali	BEN	Indo-E	Bengali		✔✔			
Bulgarian	BUL	Indo-E	Cyrillic					✔
Danish	DAN	Indo-E	Latin					✔
Estonian	EST	Uralic	Latin					✔
German	DEU	Indo-E	Latin		✔✔		✔✔	
Greek	ELL	Indo-E	Greek					✔
French	FRA	Indo-E	Latin	✔✔				
Indonesian	IND	Austron	Latin		✔✔	✔✔	✔✔	✔
Japanese	JPN	Japonic	Kanji				*✔✔	✔
Korean	KOR	Koreanic	Hangul		✔✔			✔
Mandarin	CMN	Sino-T	Hanzi		✔✔	✔✔	✔✔	
Portuguese	POR	Indo-E	Latin		✔✔			
Russian	RUS	Indo-E	Cyrillic	✔✔	✔✔		✔✔	
Spanish	SPA	Indo-E	Latin	✔✔			✔✔	
Swahili	SWA	Niger-C	Latin			✔		
Tamil	TAM	Dravidian	Tamil			✔		
Turkish	TUR	Turkic	Latin			✔✔	✔✔	✔
Vietnamese	VIE	Austro-A	Latin					✔

Experimental Setup

Models

mUNITER & xUNITER (Liu&Bugliarello+, 2021); M^{3}P (Ni+, 2021); UC^2 (Zhou+, 2021)

Experimental Setup

Models
mUNITER & xUNITER (Liu&Bugliarello+, 2021); M^{3}P (Ni+, 2021); UC^2 (Zhou+, 2021)

Fine-Tuning
Train on the English split
 On a V100 (16 GB) GPU for less than 12h

Experimental Setup

Models
mUNITER & xUNITER (Liu&Bugliarello+, 2021); M^3P (Ni+, 2021); UC^2 (Zhou+, 2021)

Fine-Tuning
Train on the English split
💻 On a V100 (16 GB) GPU for less than 12h

Zero-Shot Transfer
🌍 Evaluate on multilingual data

Translate-Test Transfer
🌍➡️ENG Evaluate on machine translated data

Experimental Setup

Models
mUNITER & xUNITER (Liu&Bugliarello+, 2021); M³P (Ni+, 2021); UC² (Zhou+, 2021)

Fine-Tuning
Train on the English split
💻 On a V100 (16 GB) GPU for less than 12h

Zero-Shot Transfer
🌍 Evaluate on multilingual data

Translate-Test Transfer
🌍➡️ENG Evaluate on machine translated data

Few-Shot Learning
🌍 After English fine-tuning, train on few samples in each target language
📈 Performance as a function of number of shots
💻 *Max-shot* setup: evaluate with all the few-shot samples (1 run per dataset–language pair)

Experimental Setup

Models
mUNITER & xUNITER (Liu&Bugliarello+, 2021); M³P (Ni+, 2021); UC² (Zhou+, 2021)

Fine-Tuning
Train on the English split
💻 On a V100 (16 GB) GPU for less than 12h

Zero-Shot Transfer
🌍 Evaluate on multilingual data

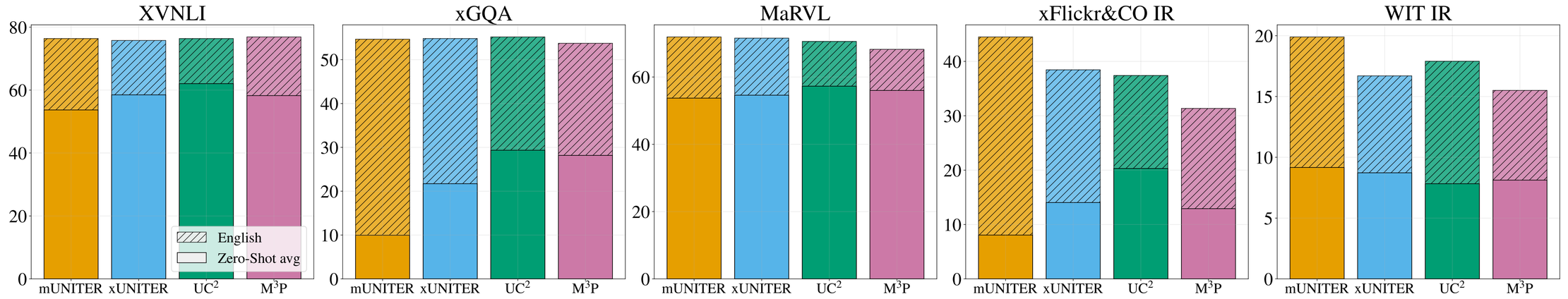
Translate-Test Transfer
🌍➡️ENG Evaluate on machine translated data

Few-Shot Learning
🌍 After English fine-tuning, train on few samples in each target language
📈 Performance as a function of number of shots
💻 *Max-shot* setup: evaluate with all the few-shot samples (1 run per dataset–language pair)

Evaluation Metrics
⚖️ Accuracy (XVNLI, xGQA, MaRVL) and Recall@1 (xFlickr&CO, WIT) – *equivalent* in our setup

Results

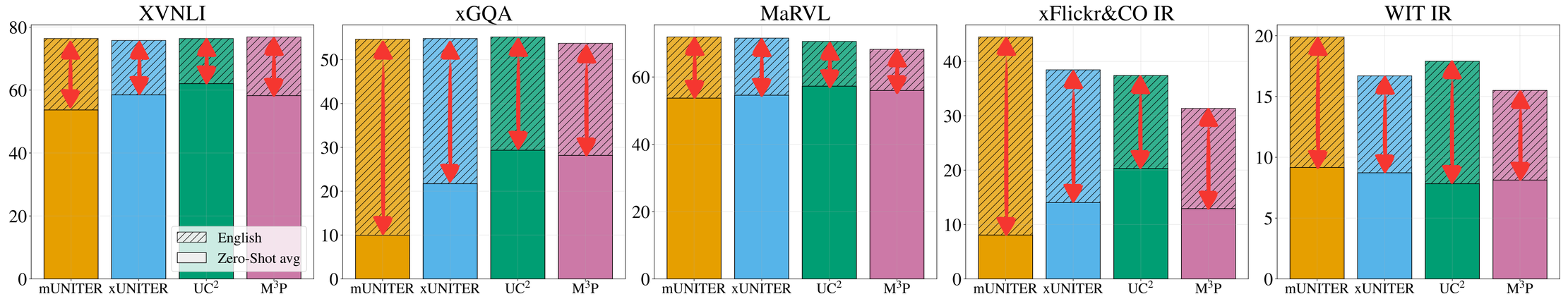
Zero-Shot Learning



Results

Zero-Shot Learning

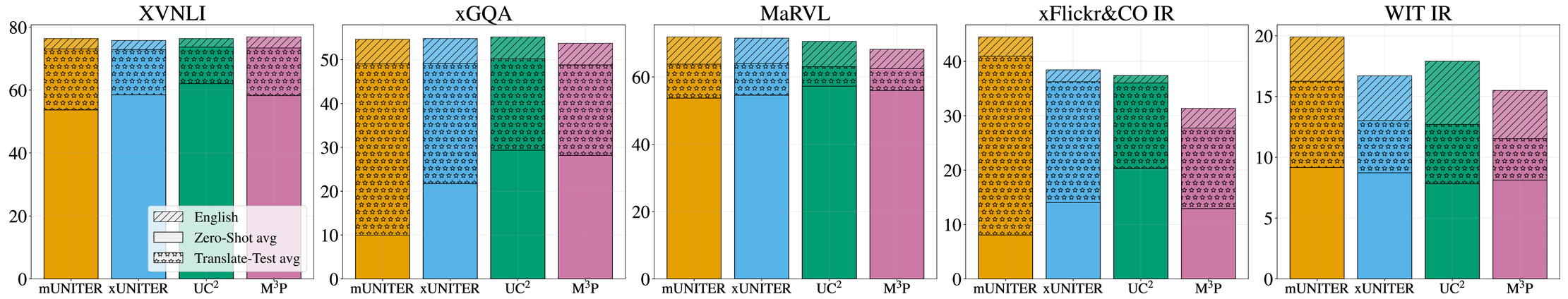
Large zero-shot transfer gap



Results

Zero-Shot Learning

Large zero-shot transfer gap

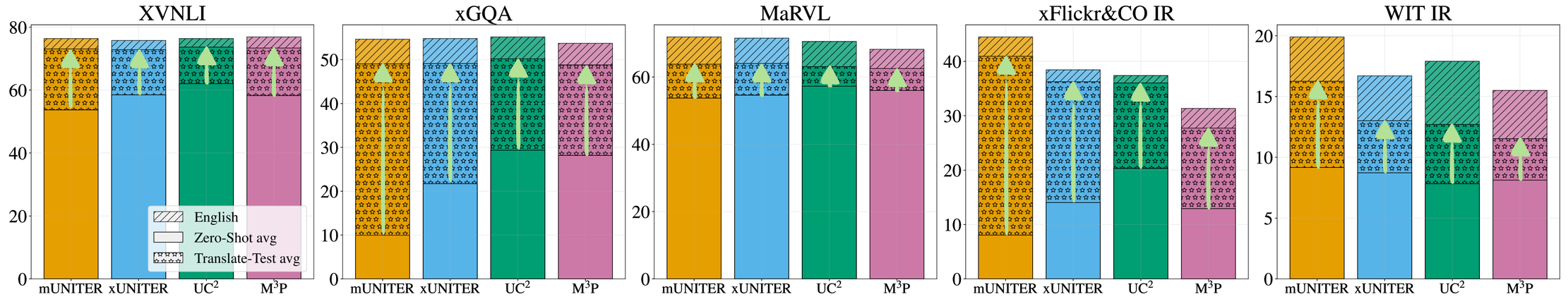


Results

Zero-Shot Learning

Large zero-shot transfer gap

Translate-test transfer ≫ zero-shot transfer

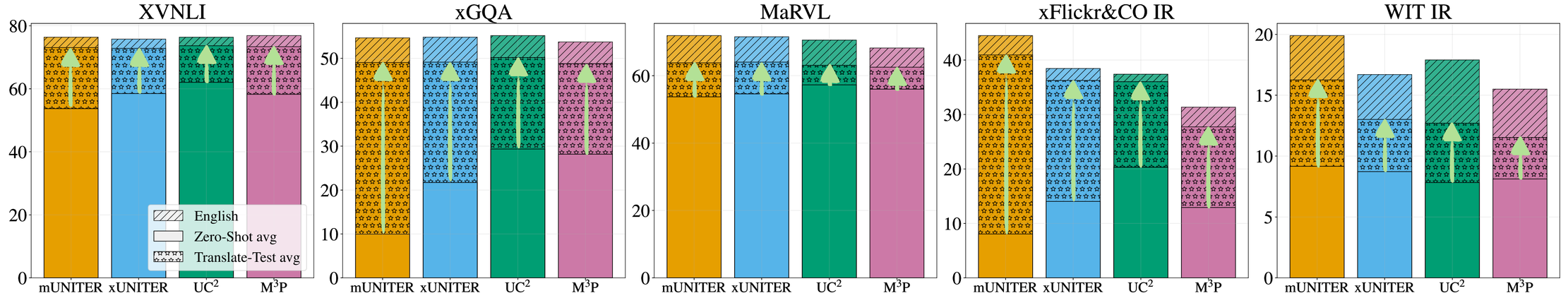


Results

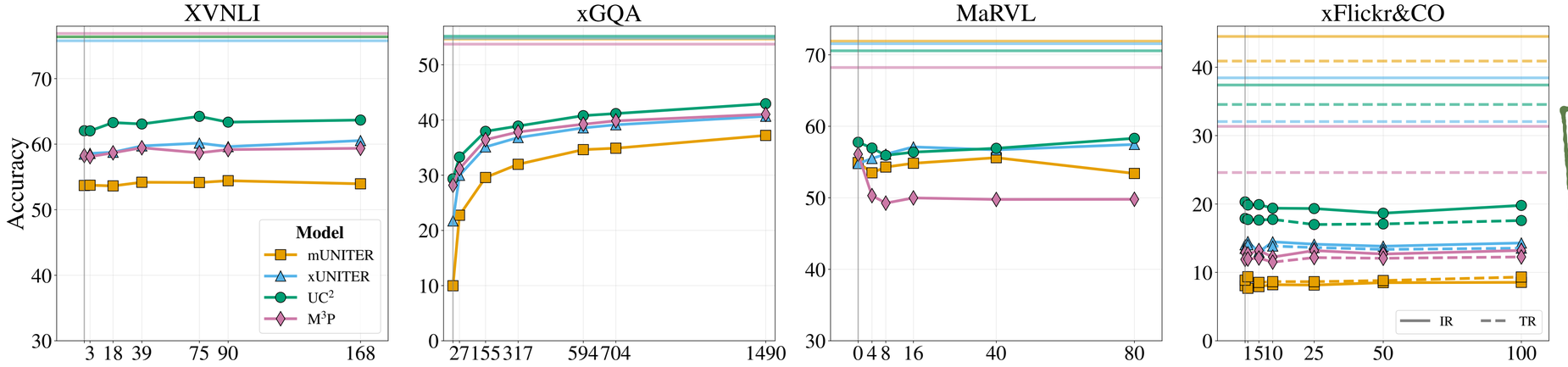
Zero-Shot Learning

Large zero-shot transfer gap

Translate-test transfer ≫ zero-shot transfer



Few-Shot Learning



Consistent but moderate gains

Take-away Messages

- Compromise (a little) on having culturally-specific grounded data

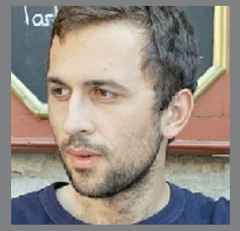
- But we end up with 5 datasets across 4 tasks in 20 languages
- Zero-shot & few-shot transfer setups
- Challenging benchmark with plenty of room for improvement!

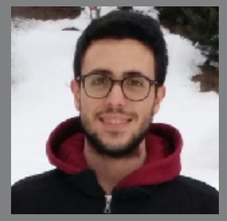
Data and Models: https://iglue-benchmark.github.io

Multilingual Multimodal Learning with Machine Translated Text

Findings of EMNLP 2022


C. Qiu


D. Oneață

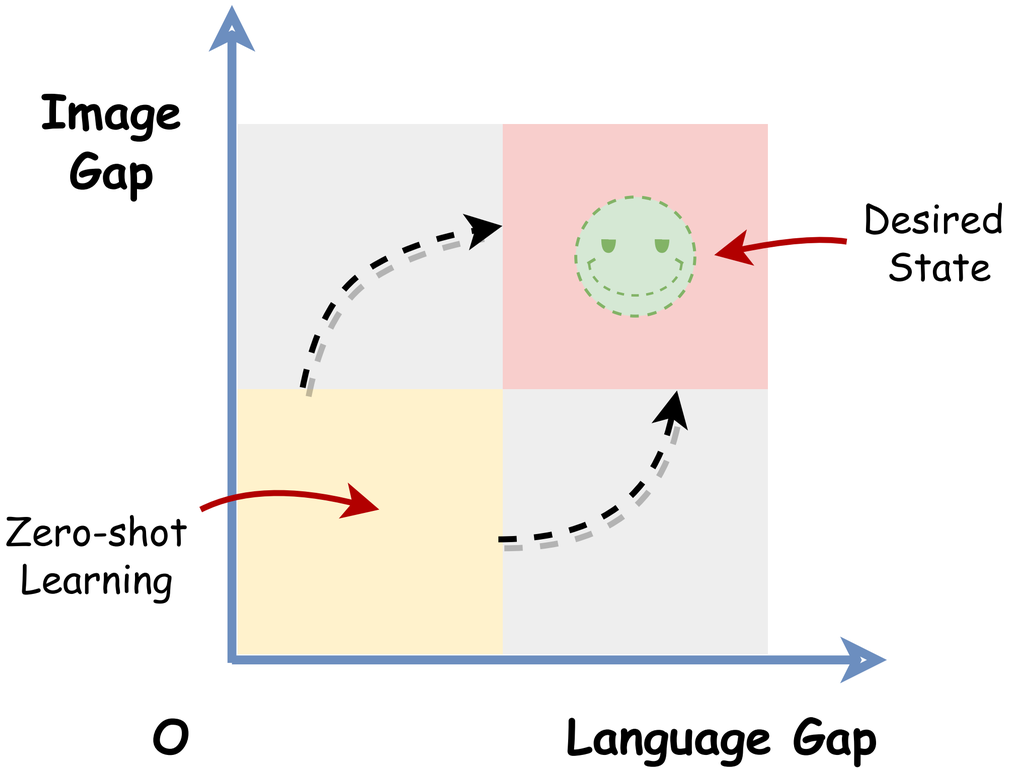

E. Bugliarello


S. Frank

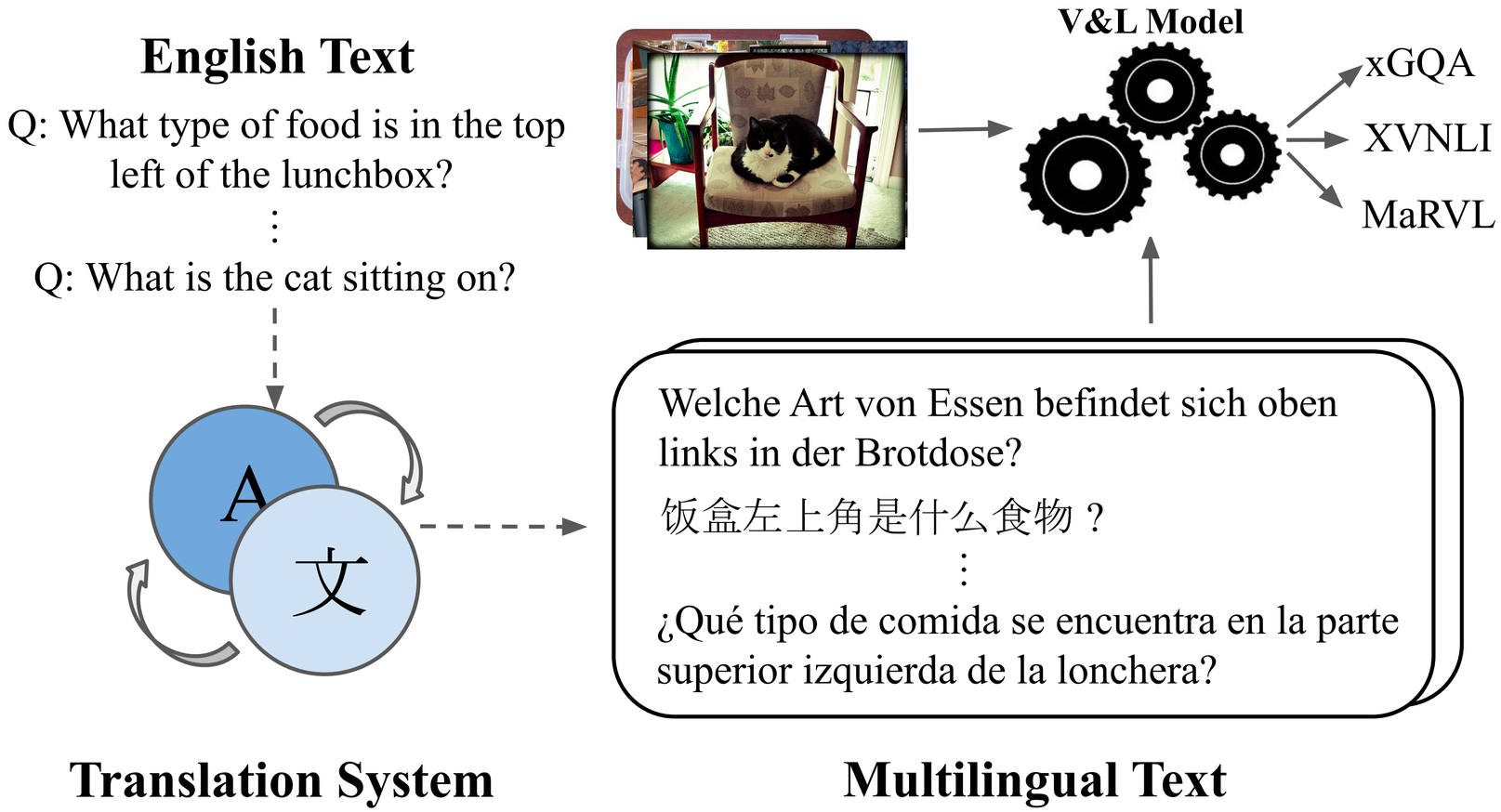

D. Elliott

Bridging the Multilingual Multimodal Gap

- Multilingual multimodal data is a sparse resource.
- Which pretraining strategies induce high-quality multilingual multimodal representations?
- Can we use machine translated data to bridge these gaps?



Translated Data Multilingual Multimodal Learning



Translation Methodology

- Translate 2.77M sentences from the English Conceptual Captions dataset into the twenty languages in the IGLUE Benchmark using M2M-100$_{\text{LARGE}}$.

Translation Methodology

- Translate 2.77M sentences from the English Conceptual Captions dataset into the twenty languages in the IGLUE Benchmark using M2M-100$_{\text{LARGE}}$.

- Filter out potentially bad data in the 55.4M translated sentences:
  - Complement of the token-to-type ratio (catch repeated tokens)
  - sBLEU against source sentence (avoid copied tokens)

Translation Methodology

- Translate 2.77M sentences from the English Conceptual Captions dataset into the twenty languages in the IGLUE Benchmark using M2M-100$_{\text{LARGE}}$.

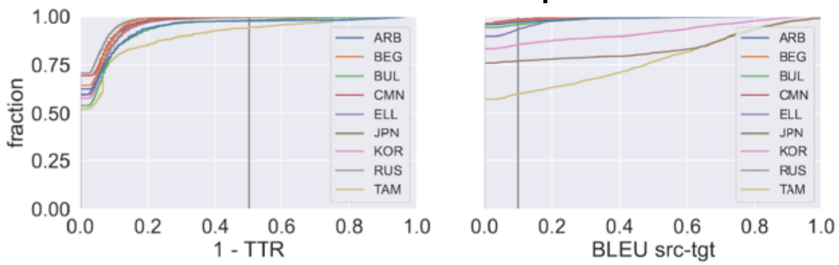
- Filter out potentially bad data in the 55.4M translated sentences:
  - Complement of the token-to-type ratio (catch repeated tokens)
  - sBLEU against source sentence (avoid copied tokens)

damask seamless floral pattern, ornament	→	*Mifano ya Mifano ya Mifano ya Mifano ya Mifano ya Mifano ya Mifano ya Mifano ya Mifano* (SWA)	BAD TRR
plaid, over garment , outfit idea cute fall outfit idea	→	*方格, over garment, cute fall* (CMN)	HIGH SBLEU

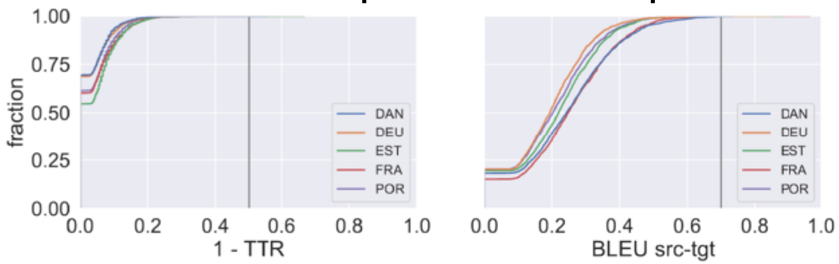
Filtering Thresholds and Resulting Corpus

- Filtering threshold depends on script and language family

Non-Latin script

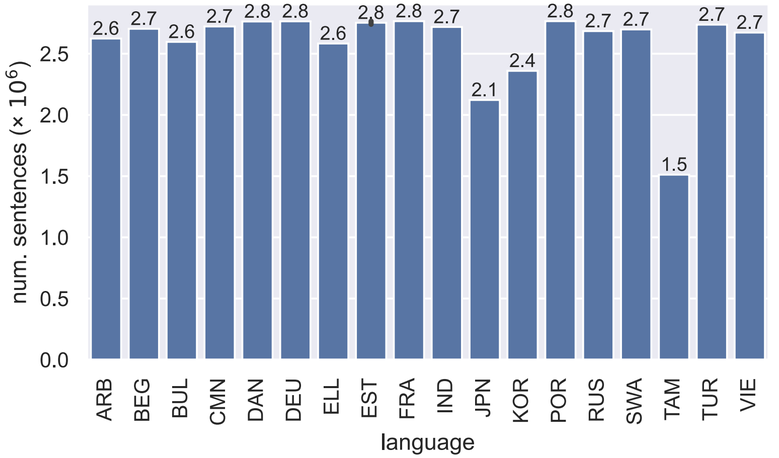


Indo-European Latin Script



- Lots of copied tokens in non-Latin script languages.

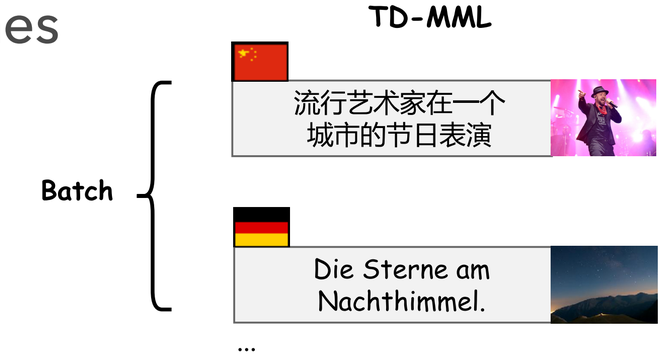
- Final pretraining corpus has 52M sentences
  - Japanese, Korean and Tamil are most affected by our filtering process.



Model

- TD-MML: XLM-R_{BASE} LM with 36 Faster R-CNN feature vectors
- Pretraining objectives with multilingual multimodal minibatches
  - MLM, MRC, ITM
  - UC^2 Visual Translation Masked Language Modelling

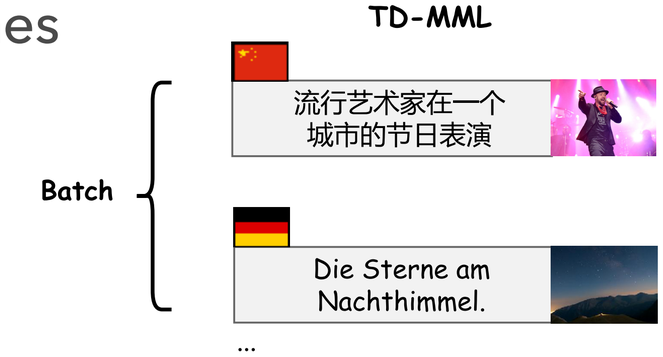
$$\mathcal{L}_{\mathrm{VTLM}}(\theta) = -\mathbb{E}_{(\mathbf{x}^{\mathrm{ENG}}, \mathbf{x}^{l}, \mathbf{v}) \sim D} \log P_\theta\left(\mathbf{x}_a^{\mathrm{ENG}}, \mathbf{x}_b^{l} \mid \mathbf{x}_{\setminus a}^{\mathrm{ENG}}, \mathbf{x}_{\setminus b}^{l}, \mathbf{v}\right)$$



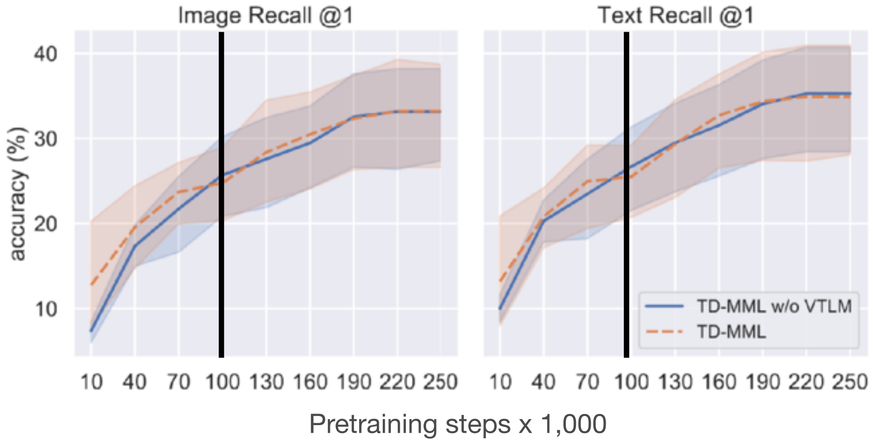
Model

- TD-MML: XLM-R_{BASE} LM with 36 Faster R-CNN feature vectors
- Pretraining objectives with multilingual multimodal minibatches
  - MLM, MRC, ITM
  - UC^2 Visual Translation Masked Language Modelling

$$\mathcal{L}_{\text{VTLM}}(\theta) = -\mathbb{E}_{(\mathbf{x}^{\text{ENG}}, \mathbf{x}^{l}, \mathbf{v}) \sim D} \log P_\theta\left(\mathbf{x}_a^{\text{ENG}}, \mathbf{x}_b^{l} \mid \mathbf{x}_{\setminus a}^{\text{ENG}}, \mathbf{x}_{\setminus b}^{l}, \mathbf{v}\right)$$



- **Crucial** to pretrain for much longer than the 100K parameter updates in xUNITER.
- Zero-shot ITM accuracy in Conceptual Captions development set on ENG and MT IND, TAM, CMN, JPN, SWA



Zero-shot IGLUE Results

- Clear improvement compared to our baselines.

Model	NLI	QA	Reasoning	Retrieval			
	XVNLI	xGQA	MaRVL	xFlickr&CO		WIT	
				IR	TR	IR	TR
mUNITER	53.69	9.97	53.72	8.06	8.86	9.16	10.48
xUNITER	58.48	21.72	54.59	14.04	13.51	8.72	9.81
UC^2	62.05	29.35	57.28	20.31	17.89	7.83	9.09
M^3P	58.25	28.17	56.00	12.91	11.90	8.12	9.98
TD-MML	64.84	**35.95**	**59.67**	**21.30**	**26.35**	**9.76**	10.35

Zero-shot IGLUE Results

- Clear improvement compared to our baselines.

Model	NLI	QA	Reasoning	Retrieval			
	XVNLI	xGQA	MaRVL	xFlickr&CO		WIT	
				IR	TR	IR	TR
mUNITER	53.69	9.97	53.72	8.06	8.86	9.16	10.48
xUNITER	58.48	21.72	54.59	14.04	13.51	8.72	9.81
UC^2	62.05	29.35	57.28	20.31	17.89	7.83	9.09
M^3P	58.25	28.17	56.00	12.91	11.90	8.12	9.98
TD-MML	64.84	**35.95**	**59.67**	**21.30**	**26.35**	**9.76**	10.35
CCLM3	74.64	42.36	65.91	67.35	65.37	27.46	28.66

- Far from concurrent CCLM3 (Zeng+arXiv22): Visual features from Swin-Transformer (Liu+ICCV21), ALBEF contrastive losses (Li+NeurIPS21), and 19M high-quality translated sentences from WikiMatrix (Schwenk+EACL21).

Take-away Messages

- Translations can **bridge** the gap between languages in MML
  - MaRVL data is high-quality: noteworthy improvement
- Important to be careful when using off-the-shelf translation systems
  - Lots of bad translation data in non-Latin script languages

Data and Models: https://github.com/danoneata/td-mml

Final Words

Connections to related work

- Machine Translation
  - Multimodal image text (Specia+ WMT16; Sulubacak+ MT20)
- Multimodal Learning
  - GD-VCR: Geographically-diverse common sense reasoning (Yin+ EMNLP21)
  - xGQA: Visually-grounded question answering (Pfeiffer+ fACL21)
  - MURAL: Multilingual image-text retrieval (Jain+ fEMNLP21)
  - CrossModal-3600: Multilingual culturally relevant image-text retrieval (Thapliya+EMNLP22)
- Computer Vision
  - Geographically diverse replacements for ImageNet (Asano+ 2021)

Some Open Questions for MML

- Pretraining data
  - Where is the naturally-occurring multilingual multimodal data?
    - Wikipedia (Srinivasan+2021), LAION-5B (Schuhmann+2021)
    - Will it be better than pretraining on translations?

- Cost-effective task data
  - How can we collect high-quality task data?
    - Multi30K cost €9,000 in 2016 for 155K image captions
    - MaRVL cost $5,000 in 2021 for 5K data points

Acknowledgements

- Emanuele Bugliarello
- Rita Ramos
- Wenyan Li
- Stella Frank
- Chen Qiu
- Dan Oneață
- Fangyu Liu
- Edoardo Ponti
- Siva Reddy
- Nigel Collier
- Jonas Pfeiffer
- Ivan Vulić

EuroHPC Joint Undertaking

Conclusions

Q: What kind of multilingual multimodal research are we doing with anglo-centric concepts?

- Multilingual multimodal research should use data that represents a broad spectrum of visual and linguistic contexts.
- Make informative progress when we compromise on some of our ideals.