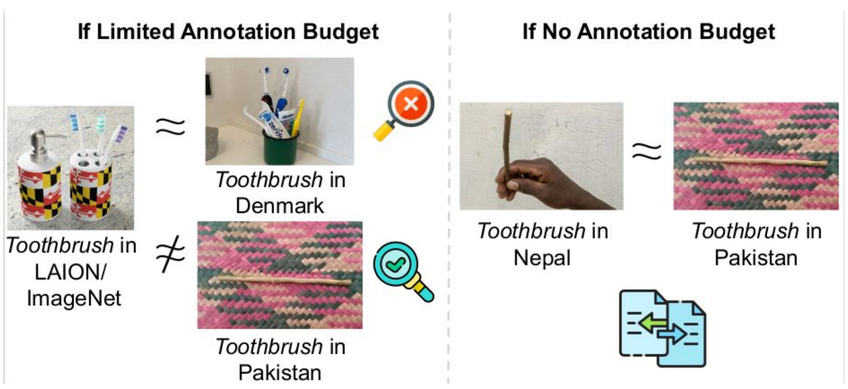
Annotations on a $Budget$: Leveraging Geo-Data Similarity to Balance Model Performance and Annotation Cost

Oana Ignat, Longju Bai, Joan Nwatu, Rada Mihalcea*

Contact: oignat@umich.edu

Motivation

- Vision-language models **work poorly on data from underrepresented countries.**
- This is primarily due to the diverse appearance of topics across countries (see *toothbrush* example).
- Collecting diverse global data is **expensive** (~1$/img).



Contributions

We investigate **how to reduce the annotation budget by finding the most effective data to annotate**.

1. We identify which **countries and topics are less represented in training data of vision-language models.**
2. We identify the groups of **countries that are visually similar and show they can be used to supplement training data effectively.**
3. Our work creates **opportunities for affordable and geo-diverse data collection**, encouraging contributions to **creating inclusive datasets and models.**

Dataset

Low-Resource:
- Diverse countries
- Crowd-sourced data
- **Dollar Street & GeoDE**

High-Resource:
- Mostly Western data
- Web-crawled
- **ImageNet & LAION**

Cooking pot in low-resource data (top) vs. in high-resource data (bottom)

# unique topics	94
# unique countries	52
# unique (topic, country) pairs	1,501
# images in low-resource data	80,801
# images in high-resource data	103,006
average # images per (topic, country)	53.8
median # images per (topic, country)	30

Tab 1. Statistics for the collected number of topics, countries, and images collected from low-resource and high-resource data after data pre-processing.

https://github.com/MichiganNLP/visual_diversity_budget

Research Questions

RQ1: Which countries are less represented in vision-language models?

Fig 1. Similarity heatmap of (topic, country) pairs, sorted from the least to the most similar. **The lighter the color, the lower the similarity between high-resource and low-resource data for that corresponding (topic, country), the more beneficial it is to annotate.** We highlight the (topic, country) pairs we determine to benefit the most from annotations, **based on consistently low similarity with the high-resource data when using CLIP, BLIP, ALIGN repres.** Grey cells do not have any images or have <10 images and are discarded.

RQ2: How can we leverage cross-country data similarity to improve the representation of vision-language models?

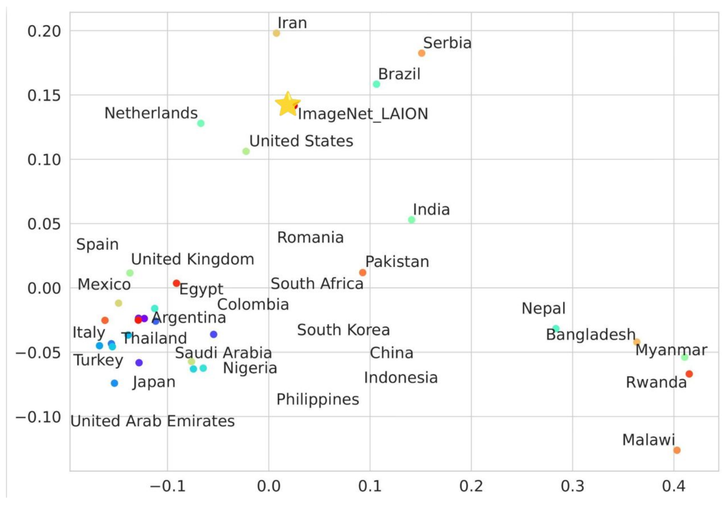


Fig 2. PCA for the topic *toothbrush* for all countries that contain this topic in the low-resource data and in the high-resource data. The data is repres. as the average of the CLIP representations.

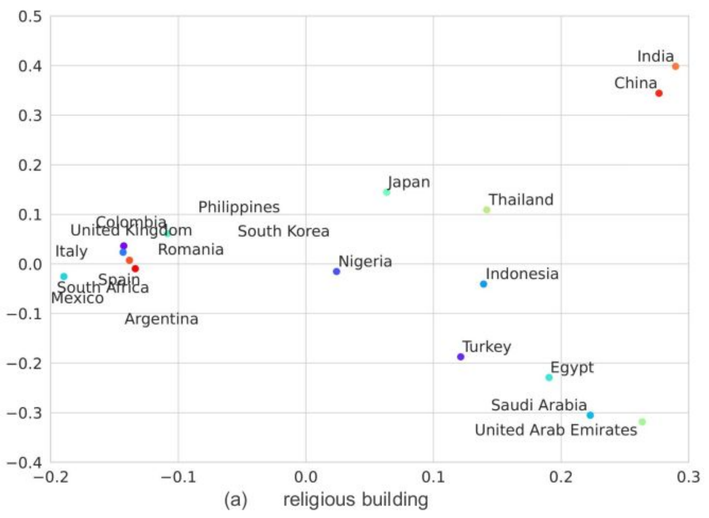


Fig 3. PCA for the topics *religious building* for all countries in the low-resource data that contain this topic.

Fig 4. Top 3 and last 3 countries (left) & topics (right) sorted by similarity score. **Implications**:
1. Annotating data from *Argentina* would help most countries, as it has highest similarity across countries
2. *religious buildings* should be annotated more widely as their visual appearance varies across countries.

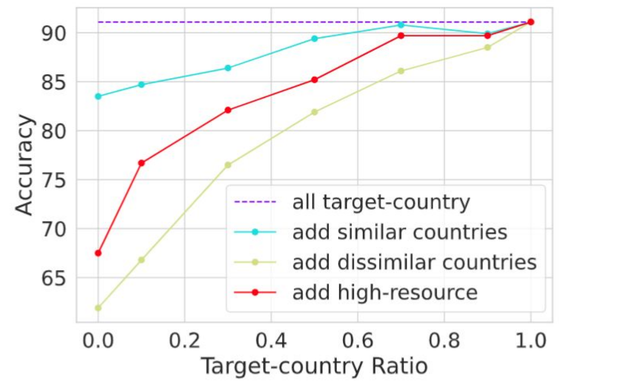


Fig 5. Topic classification accuracy after replacing different ratios of the target-country data with: (1) most similar countries to the target-country given the target-topic; (2) most dissimilar countries (3) high-resource data of the target-topic;

Main Takeaways

- **Focus annotation efforts on unrep. data for a global performance.**
- **Most countries < 10 img/topic -> annotate those 3,329/ 4,830 (country, topic) pairs**
- **Leverage cross-country data similarity to supplement unrep. countries.**
- **Geographical distance does not correlate w visual sim between countries.**