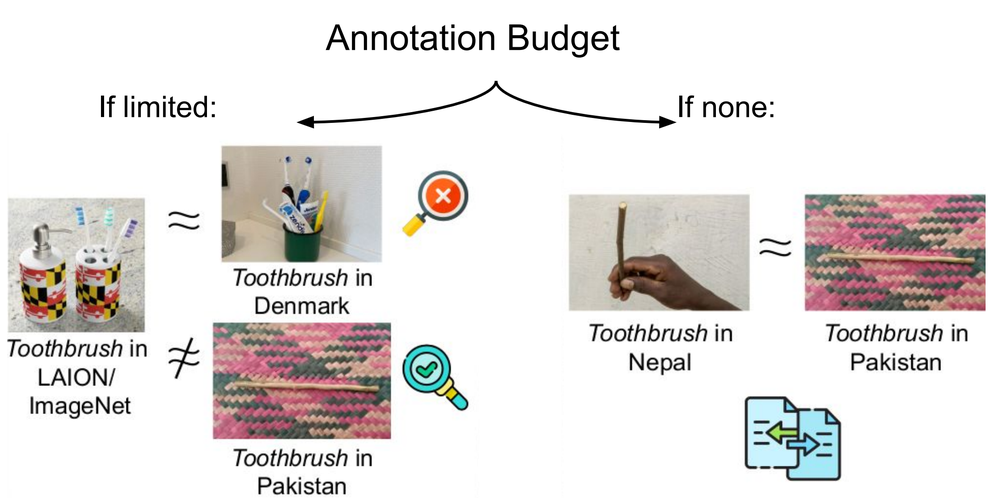
Annotations on a 💰 Leveraging Geo-Data Similarity to Balance Model Performance and Annotation Cost

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

Contact: oignat@umich.edu

Motivation

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



Contributions

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

Table 1. Data Statistics

github.com/MichiganNLP/visual_diversity_budget

Research Questions

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

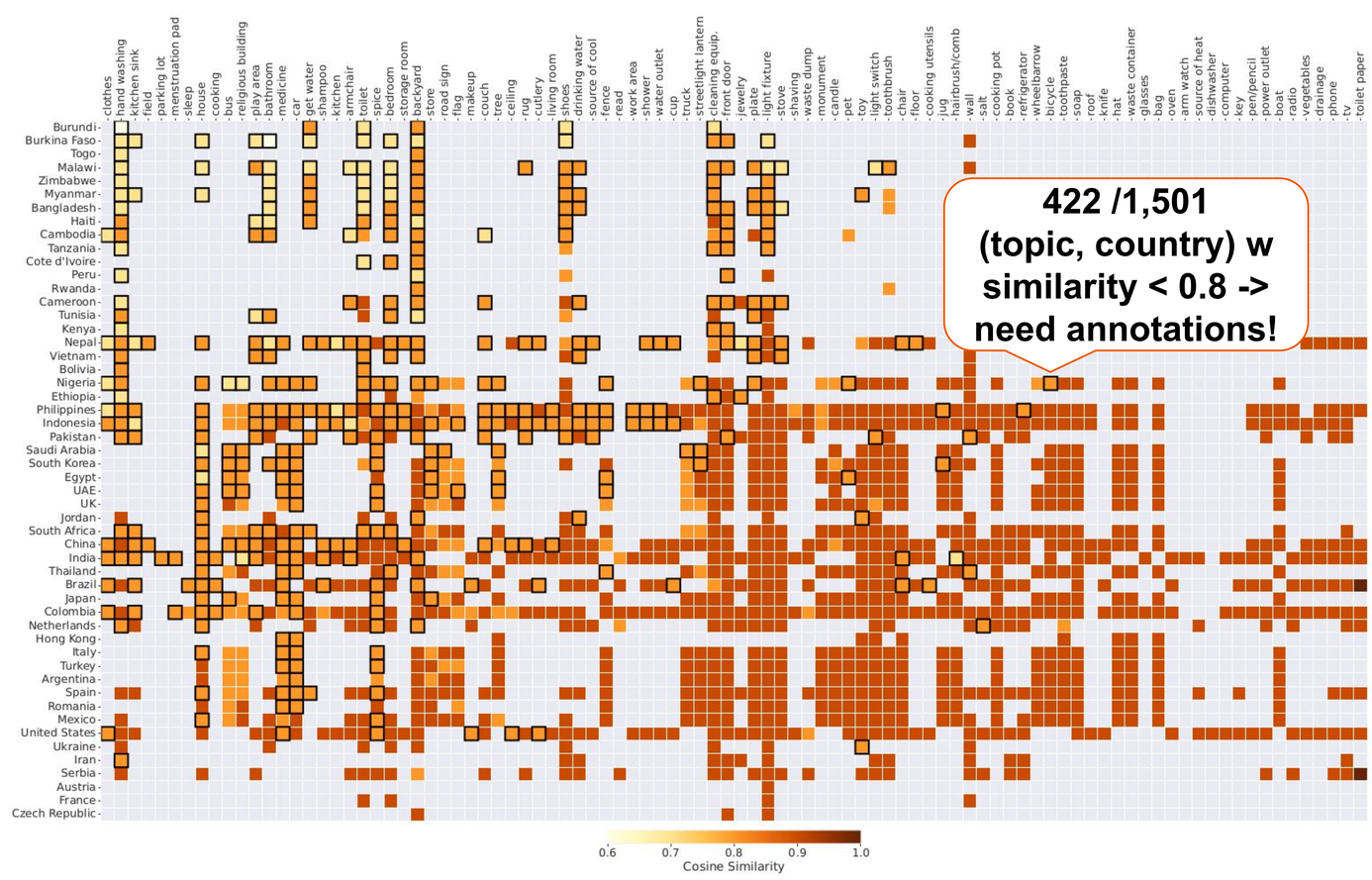
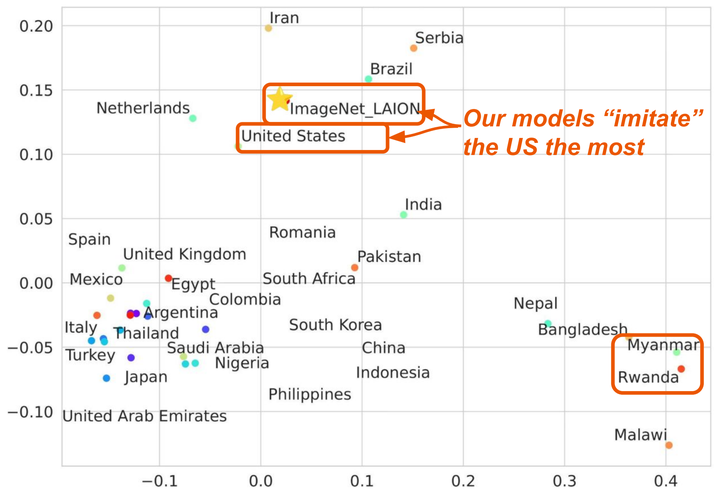


Fig. Visual Similarity Heatmap between Low-Resource and High-Resource (topic, country) pairs.

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



Leverage cross-country similarity to supplement data

Supplementing data from underrep. countries with data from visually similar countries is effective!

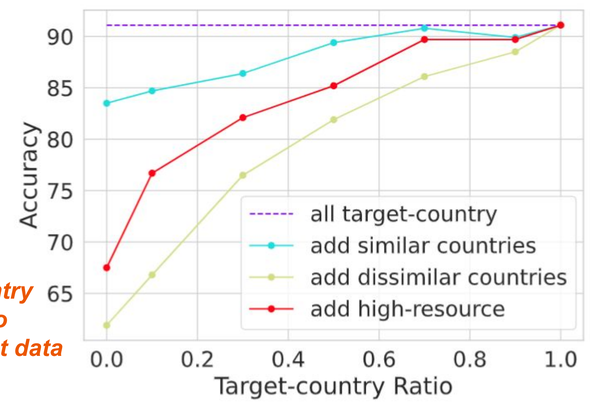


Fig. PCA for *toothbrush* for all countries that contain this topic in the low-resource data and in the high-resource data.

Burundi is the most different to other countries, therefore it needs more annotations

Argentina is the most similar to other countries, therefore it would help most other countries

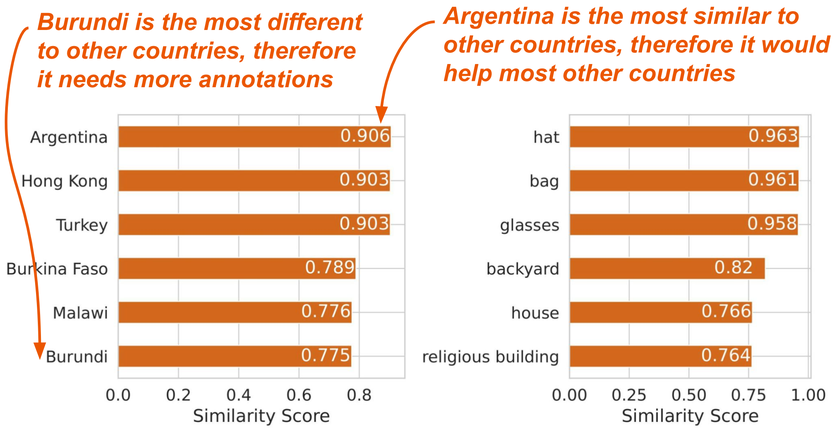
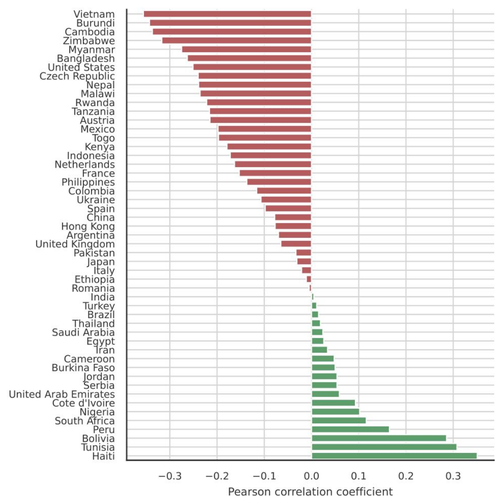


Fig. Top 3 and last 3 countries (left) and topics (right) sorted by average similarity score

Geographical distance does not usually correlate with visual similarity between countries



Main Takeaways

- **Focus annotations on unrepresented data** for a global performance.
- **Most countries < 10 img/ topic ->** annotate those first!
- **Leverage cross-country data similarity** to supplement unrepresented countries.
- **Geo distance not correlate w country sim. ->** need more info (history, income, …).