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

Oana Ignat, Longju Bai, Joan Nwatu, Rada Mihalcea
University of Michigan
{oignat,longju,jnwatu,mihalcea}@umich.edu

Abstract

Current foundation models have shown impressive performance across various tasks. However, several studies have revealed that these models are not effective for everyone due to the imbalanced geographical and economic representation of the data used in the training process. Most of this data comes from Western countries, leading to poor results for underrepresented countries. To address this issue, more data needs to be collected from these countries, but the cost of annotation can be a significant bottleneck. In this paper, we propose methods to identify the data to be annotated to balance model performance and annotation costs. Our approach first involves finding the countries with images of topics (objects and actions) most visually distinct from those already in the training datasets used by current large vision-language foundation models. Next, we identify countries with higher visual similarity for these topics and show that using data from these countries to supplement the training data improves model performance and reduces annotation costs. The resulting lists of countries and corresponding topics are made available at https://github.com/MichiganNLP/visual_diversity_budget.

Keywords: geo-diverse datasets, active learning, effective annotations, visual similarity, vision-language models

1. Introduction

Vision-language models have shown remarkable advances in recent years (Li et al., 2019; Zhang et al., 2021; Radford et al., 2021; Zellers et al., 2021; Li et al., 2022; Kirillov et al., 2023a; Huang et al., 2023b). These models have shown great performance on a variety of tasks, from lower-level tasks such as object detection, image segmentation (Kirillov et al., 2023a), and image and video classification to higher-level tasks such as image/video captioning (Li et al., 2022; Huang et al., 2023b), text-image/video retrieval (Radford et al., 2021), visual question answering and visual commonsense reasoning (Zellers et al., 2021, 2022).

At the same time, prior work has demonstrated that these models do not work well for everyone (De Vries et al., 2019). Specifically, models do not work well on out-of-domain data, and data from low-income and non-western countries (Nwatu et al., 2023). This is due to the imbalanced geographical and economic representation of the data used to train these models, as it comes mainly from North America and Western Europe (Shankar et al., 2017). One solution that Rojas et al. (2022) and Ramaswamy et al. (2023) propose is to collect more data from underrepresented countries. However, as Ramaswamy et al. (2023) highlights, annotation costs are a substantial bottleneck; when crowdsourcing the data, fair pay is about 1.08$ per image without including researcher time.

As a complementary solution, we investigate strategies to reduce the annotation budget while finding effective annotation data. Specifically, our paper aims to answer two main research questions.

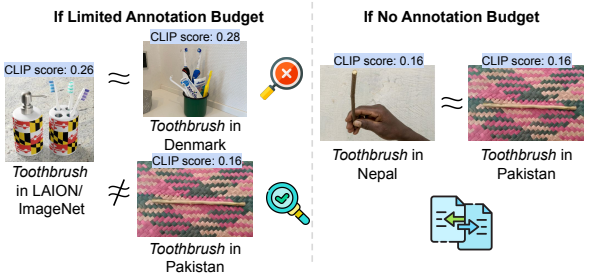

Figure 1: Vision-language models work poorly on data from underrepresented countries. This is primarily due to the diverse appearance of topics (objects and actions) across countries (e.g., "toothbrush"). However, collecting diverse global data is very expensive. As solutions to budget annotations, we propose to (1) annotate the images visually different from the ones in high-resource datasets such as LAION or ImageNet; (2) supplement data from low-resource countries with data from visually similar countries.

RQ1: Which countries are less represented in the training data of vision-language models?

We aim to find ways to effectively focus future annotation efforts on specific *countries* and their corresponding *topics* (objects and actions).[1] Our study highlights the visual diversity of common topics across countries and those that differ the most from the primarily Western data used to train most multimodal foundation models.

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

[1]Throughout the paper, for brevity, we use the term *country* to refer to a country or territory.

LREC-COLING 2024, pages 1239–1259

We obtain groups of countries that are visually similar in their representation of a given topic. This is particularly useful when there is not enough data for one of the countries in the group, and there is no annotation budget. We can supplement the data from this country by using data from the other countries in the group.

We summarize our contributions as follows. First, **we identify the data likely to most benefit from annotations** by finding which countries and corresponding topics are less represented in the training data of vision-language models. Second, across 52 countries and 94 topics, **we identify the groups of countries that are visually similar in their representation of a topic and show that they can be used to supplement training data effectively**. Third, **our main takeaways create opportunities for affordable and geo-diverse data collection**, encouraging contributions to creating datasets and models that work for everyone.

2. Related Work

Data Subset Selection/ Active Learning There have been numerous studies on the use of semi-supervised models to leverage a combination of limited labeled data and vast amounts of unlabeled data to improve model performance at lower costs (Hady and Schwenker, 2013; Oliver et al., 2018; Taha, 2023; Chen et al., 2022).

However, model-generated labels could be inconsistent and unrepresentative with semi-supervision, leading to reduced model performance (Ahfock and McLachlan, 2023; Elezi et al., 2022; Wang et al., 2021). While similar to semi-supervised learning in objective, active learning methods seek to capture the entire data distribution by focusing labeling efforts on the data points that provide the most information for training the best-performing models (Ren et al., 2021; Citovsky et al., 2021; Monarch, 2021; Yang et al., 2017) using approaches such as uncertainty-based sampling in Gal and Ghahramani (2016); Beluch et al. (2018) and geometric-based methods in Sener and Savarese (2018). Unsupervised subset selection methods like K-means and K-median core set in Har-Peled and Kushal (2005), which form the foundation for geometric-based active learning approaches are similar to our work which seeks to select a subset that is representative of the entire dataset using distance metrics. However, the objective of the selection is to include images from a low-resource dataset with the least similarity to data of the same class in a high-resource dataset.

Evaluating Disparities in Model Performance There exists a considerable body of literature evaluating the fairness and the unequal performance of vision and vision-language models on diverse groups categorized according to race (Gebru, 2020), gender (Buolamwini and Gebru, 2018), geolocation (Kim et al., 2021; Shankar et al., 2017; Goyal et al., 2022a) and income (De Vries et al., 2019; Nwatu et al., 2023).

Further analysis of these disparities reveals that factors such as ambiguous label definitions, domain shifts, annotator disagreement (Hall et al., 2023; Kalluri et al., 2023), as well as image properties relating to texture, lighting, and occlusion in vision and vision-language datasets (Gustafson et al., 2023) contribute to disparities in datasets which carry over to affect model performance.

Frameworks have been developed to facilitate the detection of bias through guided human-in-the-loop inspection, either in datasets Hu et al. (2020) or in models Goyal et al. (2022b). Our work focuses on exploring the presence of variations in image representations across demographic groups in existing datasets, to inform cost-effective methods for building balanced, diverse datasets.

Improving Representation in AI. Efforts toward improving equal representation in AI and equitable AI impact revolve around model adaptation, transfer learning, and dataset diversity. However, Salman et al. (2022); Kalluri et al. (2023); Dubey et al. (2021); Wang and Russakovsky (2023) suggest that transfer learning and model adaptation methods might not be enough to eradicate the issue of under-representation in AI models.

On the other hand, adding diverse data to training datasets tends to yield significant improvements in model performance across different groups (Ramaswamy et al., 2023; Rojas et al., 2022). The need for more diverse datasets has become apparent, leading to the development of datasets like GeoYFCC (Dubey et al., 2021), GeoDE (Ramaswamy et al., 2023), Dollar Street (Rojas et al., 2022), and Segment Anything (Kirillov et al., 2023b) that include data collected from diverse locations.

While advantageous, diverse datasets are expensive and resource-intensive to build. (Schumann et al., 2021; Garcia et al., 2023; Geigle et al., 2023) explored a less expensive alternative: revising or creating annotations for an existing dataset to improve inclusivity and reduce bias. Similarly, we seek to facilitate effective but less expensive annotations by leveraging the differences between high-resource and low-resource datasets to curate the best low-resource subset for annotation.

3. Methodology

We start by collecting two datasets that reflect the low-resource and high-resource settings. First, we compile a crowd-sourced geo-diverse dataset collected from a large number of countries, which we refer to as "low-resource data" due to the low number of images that could be collected for each country in the set and the difficulty of gathering more. Second, we also compile a web-scraped dataset used for training foundation models, which we refer to as "high-resource" due to its vast size consisting of billions of images (e.g., LAION-5B[2]) and the ease of gathering more data.

Next, we pre-process the data by mapping the topics between the two data sources, filtering out topics and countries with very few images. Finally, we utilize the collected data to generate visual representations through vision-language foundation models. These representations are then used to determine the visual similarity between images of topics in low-resource data and their corresponding topics in high-resource data.

3.1. Low-resource Multimodal Data

We combine two geographically diverse datasets: GeoDE (Ramaswamy et al., 2023) and Dollar Street (Rojas et al., 2022). For brevity, we call *topics* all the labels used for all the objects and actions in these two datasets.

GeoDE. The GeoDE dataset contains $61,940$ crowd-sourced images of 40 objects. The data is balanced across six regions (West Asia, Africa, East Asia, South East Asia, Americas, and Europe), each with 3-4 countries. These regions were chosen due to their scarcity in most public datasets. Using a combination of heuristics and manual validation, the authors selected the objects likely to be visually distinct across the six regions.

Dollar Street. The Dollar Street dataset contains $38,479$ images collected from 63 countries on four continents (Africa, America, Asia, and Europe). The images capture everyday household objects and actions (e.g., "toothbrush", "toilet paper", "cooking"). The data contains 291 unique topics, out of which we remove nineteen subjective topics following the work of De Vries et al. (2019) (e.g., "most loved item", "things I wish I had"). All the subjective topics are found in the Appendix. The number of images for a given country ranges from 45 in Canada to $4,704$ in India, with a median of 407 images per country.

[2] https://laion.ai/blog/laion-5b/

3.2. High-resource Multimodal Data

As high-resource datasets, we sample data from ImageNet (Deng et al., 2009) and LAION (Schuhmann et al., 2022). We chose these datasets due to their popularity in vision-language models.

ImageNet. ImageNet and ImageNet Large Scale Visual Recognition Challenge (ILSVRC) are pioneers in advancing object detection and classification progress. The imagenet21k dataset (Deng et al., 2009) contains around 21,000 WordNet (Fellbaum, 2000) synsets and more than 14 million annotated images. We use the processed version of ImageNet21k (Ridnik et al., 2021), with removed invalid classes and resized images. We also tried using ImageNet1k, but it did not have enough classes for our purpose, and we chose to use it to supplement the ImageNet21k data.

LAION. Large language-vision models such as CLIP or ALIGN have been trained on billions of image-text pairs unavailable to the public. LAION-5B (Schuhmann et al., 2022) was created to address this problem by open-sourcing a CLIP-filtered dataset[3] of 5,85 billion high-quality image-text pairs. We use LAION-400M (Schuhmann et al., 2021), a subset of LAION-5B that contains 400 million English image and text pairs.

3.3. Data Pre-processing

Combine GeoDE and Dollar Street. We pre-process and combine the low-resource datasets to increase the number of topics, images, and country diversity. First, we manually group and rename the topics from Dollar Street with the same meaning (e.g., "bathroom privacy", "bathroom/ toilet" are renamed "bathroom"). Next, we rename the topics from Dollar Street that match those in GeoDE (e.g., "bike" to "bicycle", "medication" to "medicine"). We remove three topics with less than 10 images per topic. Finally, we obtain a total of 99 unique topics, $93,060$ images, from 4 continents, 18 regions, and 83 countries.

Low-resource to High-resource Data Mapping. We map the 99 topics from the low-resource data to the high-resource data, ImageNet, and LAION by identifying the images with similar labels.

First, we map 51 topics from the low-resource data to an exact match to ImageNet21k or ImageNet1k. We could not find an exact match

[3] The data is filtered using OpenAI's CLIP ViT-L/14 by calculating the cosine similarity between the text and image embeddings and dropping those with a similarity below 0.3.

for 38 topics because these topics are too abstract (e.g., "jewelry", "source of cool", "religious building"). Instead, we find mappings for their hyponyms (e.g., for "jewelry", we map "bangle", "necklace", "bracelet" and "ring"). The remaining 10 topics for which we could not find any exact or hyponym mapping to ImageNet21k or ImageNet1k are mapped to LAION.

We map data in LAION by selecting the images with captions that contain the topic query. Because LAION data is web-crawled, we find that the images are lower quality than ImageNet and not always relevant to the topic query: e.g., the "TV" topic in LAION contains images of people on TV, not of the object TV. Therefore, to ensure the correctness of the mapping, we manually inspect the images and map a topic to LAION only when most images are relevant to the topic query. We map 64 topics to LAION. Note, however, that the number of hyponyms and the quality of LAION images limit how comprehensive the mapping process is. Two independent annotators check 20 random images from each topic and find that most noisy images come from LAION. Therefore, we decide to limit the amount of data from LAION and add more images from ImageNet. Specifically, we randomly sample around 200 images per topic from LAION and around 1,000 images per topic from ImageNet. Note that the high-resource data does not contain country information. We show the data before and after pre-processing and the topic mapping in our repository.[4]

Figure 2: Example images ("cooking pot") in low-resource data (left) vs. in high-resource data (right).

Data Filtering. The low-resource data is unbalanced, as the total number of images per country varies from 6,549 for Japan to 1 for Bulgaria and Venezuela, with a median of 345 images per country. The number of images per topic is also unbalanced, from 3,049 for "waste container" to 18 for "hanging clothes to dry". However, balancing the data by down-sampling significantly reduces the number of countries represented for each topic. Having numerous countries represented is essential for our setup. Therefore, we choose not to balance the data. Instead, we remove the (topic, country) pairs containing less than 10 images, considering this threshold a minimum for experiment significance. This also removes considerable data: 3,329/ 4,830 (topic, country) tuple pairs, 5/ 99 topics, and 31/ 83 countries. We show the removed topics and corresponding countries in our repository and highlight the need for more data for these pairs to obtain significant results.[4]

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

We show the statistics after the data collection and pre-processing in Table 1 and the image distribution of countries per topic in Appendix Figure 10.

3.4. Data Representation

We use an ensemble of three representations to compute the image similarity and to ensure the results generalize across representation types. We choose CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), and BLIP-2 (Li et al., 2023) due to their popularity as foundation models (Bommasani et al., 2022), i.e., their use in a multitude of models and their high zero-shot performance across various tasks and datasets, such as text-to-image retrieval, image question answering, human action segmentation, image-sentence alignment, image captioning (Cafagna et al., 2021; Saharia et al., 2022; Kirillov et al., 2023b; Huang et al., 2023a).

CLIP Representations. We use the pre-trained Vision Transformer ViT-B/32 (Dosovitskiy et al., 2021) from the CLIP model (Radford et al., 2021) to encode the visual representations of the images. The training dataset for CLIP was created from the results of numerous queries to various publicly available Internet sources. The dataset referred to as WebImageText WIT contains 400 million (image, text) pairs and is not available to the public.

ALIGN Representations. We also extract image features following the ALIGN (Jia et al., 2021) model setup, using a pre-trained EfficientNet (Tan and Le, 2019) as a vision encoder. Since the original code has not been released, our implementation is based on the Kakao Brain code that reproduced the original paper.[5] ALIGN was trained on

[4] https://github.com/MichiganNLP/visual_diversity_budget

[5] https://huggingface.co/docs/transformers/model_doc/align

1.8 billion image-text pairs collected following the methodology used for the Conceptual Captions dataset (Sharma et al., 2018). Since the emphasis was on scale instead of quality, the dataset underwent fewer post-processing steps, thus leading to a noisier dataset. This dataset is currently unavailable for public access.

BLIP-2 Representations. We also extract image features using BLIP-2 (Li et al., 2023), which uses ViT-g/14 from EVA-CLIP (Sun et al., 2023) as image encoder and removes the second last layer's output features to increase the performance. BLIP-2 was trained on a total of 129M images aggregated from the COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2017), CC3M (Sharma et al., 2018), CC12M (Changpinyo et al., 2021), SBU (Ordonez et al., 2011), and the LAION400M datasets (Schuhmann et al., 2021). Captions for the web images were generated using CapFilt (Li et al., 2022).

4. Mapping the Representation of Vision-Language Models

In this section, we address the first research question: **RQ1: Which countries are less represented in the training data of vision-language models?**

For each (topic, country) pair, we compute the cosine similarity between the average visual representations of all the corresponding images in the low-resource data and the average visual representations of all the corresponding images in the high-resource data. Note that the average is computed over all three visual representation types, i.e., CLIP, BLIP, and ALIGN. We select the (topic, country) pairs with a similarity score lower than a threshold computed as the average similarity score between all the image representations in the low-resource data and the corresponding representations in the high-resource data. This process is repeated for each visual representation type.[6] Finally, the (topic, country) pairs selected for all three visual representations are the ones we find to be consistently different from the high-resource data and, thus, the ones that benefit the most from annotations. We find 422 such (topic, country) pairs out of 1,501 unique (topic, country) pairs, potentially reducing the annotation budget to less than a third of the initial amount. We share the results in our repository.[4]

Visual similarity for each (topic, country) in low-resource data with corresponding topics in high-resource data. We compute a similarity heatmap where the rows are the topics, and the columns are countries. We sort the rows (countries) and columns (topics) from the least to the most similar based on the average similarity score per country and topic, leaving out the NaN values (the grey, empty cells). We show in Figure 3 the similarity heatmap for the CLIP representation and highlight the (topic, country) pairs we find to benefit the most from annotations based on consistently low similarity with the high-resource data across the three visual representations.

From Figure 3, we can also see that the countries with the fewest data are usually the ones with the most topics in need of annotations (e.g., from $Burundi$ to $Kenya$). Exceptions to this are countries such as $Nepal$, $Nigeria$, $Philippines$, and $Indonesia$, which have more data points (topics), but more than half of the topics require annotations, and countries such as $Czech\ Republic$, $France$ or $Austria$ which have very few topics and none require annotations. In Figure 3, we see a few topics in $United\ States$ that are marked to require annotations: "medicine", "spice", "ceiling", "clothes" and "makeup". We show in Appendix Figure 11 representative images from these topics from the two data sources, which explain the visual differences. For the rest of the topics, as expected, $United\ States$ data is similar to the high-resource data. We considered using the $United\ States$ as the high-resource data source. However, due to the lack of data on some topics and relatively few images per topic compared to other countries, it was not feasible.

There are differences between the results obtained with each visual representation type regarding similarity score intervals and which (topic, country) pairs are similar to the high-resource data. However, the general similarity trend is consistent as most (topic, country) pairs have only low or high similarity scores across all three representations. This is also supported by the strong Pearson correlations between the scores obtained with the three representation types: CLIP and BLIP scores correlate 0.62, CLIP and ALIGN scores correlate 0.65, ALIGN and BLIP scores correlate 0.72. We show in the Appendix Figure 12, 13, and 14, the similarity heatmaps for each representation type: CLIP, ALIGN, and BLIP respectively.

Topic visual representation in high-resource and low-resource data. To show how the topic visual representations vary per low-resource and high-resource data, we perform a 2D transformation using Principal Component Analysis (PCA) (F.R.S., 1901). In Figure 4, we show the CLIP average representations per country in the low-resource and the corresponding high-

[6] Thresholds and data representations can be changed to fit the purpose of the analysis or application.

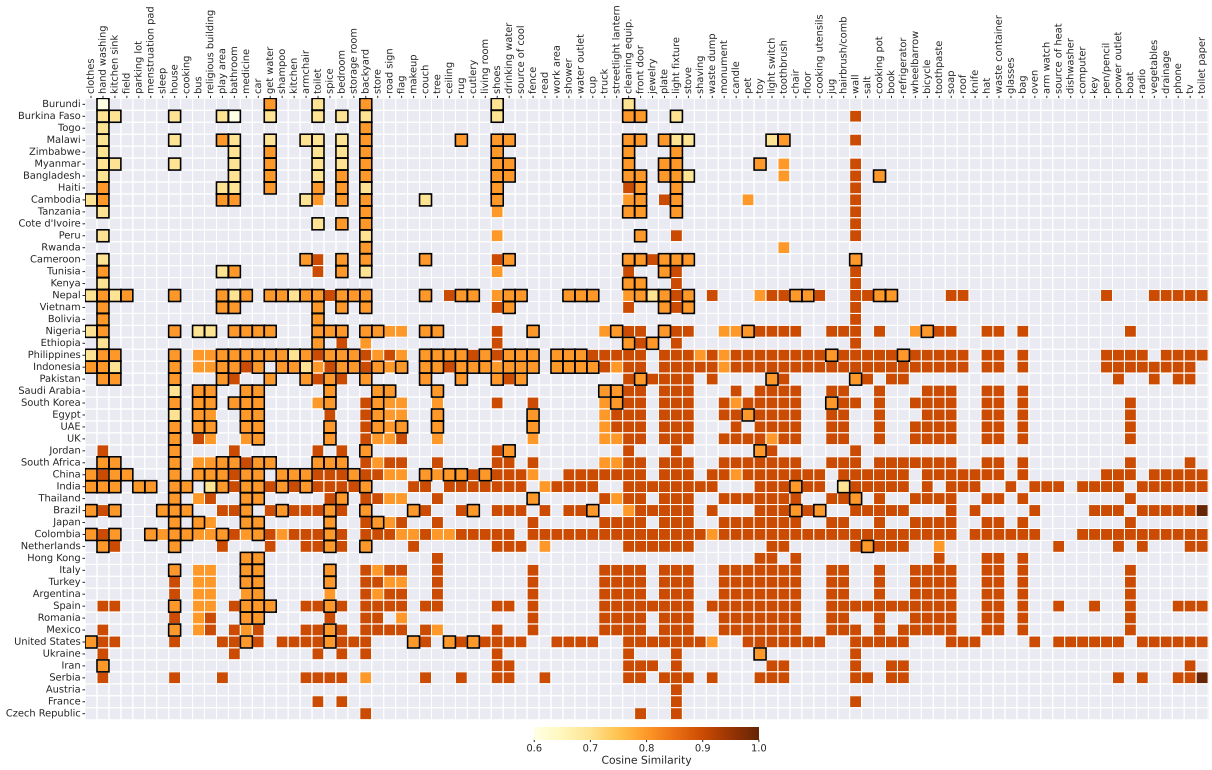

Figure 3: Similarity heatmap of (topic, country) pairs. Based on the average similarity score, rows and columns are sorted from the least to the most similar. The lighter the color, the lower the similarity between high-resource and low-resource data for that corresponding (topic, country) pair, the more beneficial it is to annotate. We highlight with *black* the pairs we determine to benefit the most from annotations. Grey cells have less than ten images and are therefore discarded. *Best viewed in color.*

resource data for the topic "toothbrush". We can observe that, for this topic, there is considerable visual diversity across countries. When comparing to the high-resource data, $ImageNet\_LAION$, we observe visually different countries, such as $Malawi$, $Rwanda$, and $Myanmar$, and countries very visually similar, such as $Netherlands$, $UnitedStates$, and $Brazil$. In addition, we observe many countries that tend to be clustered together, i.e., visually similar for this particular topic, such as $Mexico$, $Italy$, $Japan$, $South\ Korea$, and others. We examine more about the similarities between countries when answering **RQ2**, in the following section. In Appendix Figure 15, 16, 17 we show results for other topics ("hand washing", "toilet", "wall") in low-resource and high-resource data.

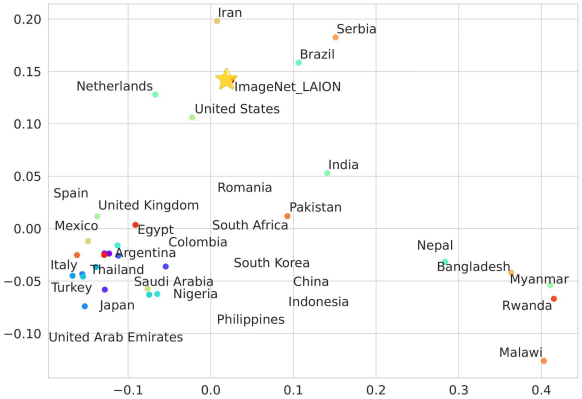

Figure 4: PCA for the topic "toothbrush" for all countries that contain this topic in the low-resource data and in the high-resource data. The high-resource data point is highlighted with *star* symbol. The data is represented as the average of the CLIP representations.

5. Cross-country Data Similarity for Improved Model Representation

We now turn to the second research question **RQ2: How can we leverage cross-country data similarity to improve the representation of vision-language models?**

We calculate the cosine similarity between the average visual representations of images for each topic across countries, and repeat this process for all three visual representations. Given a topic, the final visual similarity score between two countries is obtained by averaging the similarity values obtained for each visual representation type. For each (country, topic) pair, we obtain the visually similar countries, along with their similarity score, from the most to the least similar, and share them

in our repository.[4]

We calculate the average similarity score for each country across all corresponding topics and for each topic across all corresponding countries. We show the similarity score distribution for the top three and last three countries and topics in Figure 5, and for all countries and topics in the Appendix Figure 18 and 19.

As shown in Figure 5, $Burundi$ has the lowest similarity score of 0.775, indicating that it is the most different country compared to the others and needs its own annotations. On the other hand, $Argentina$ has the highest similarity scores of 0.907, indicating a high similarity to other countries. These results imply that annotating data from $Argentina$ would help other countries. The most visually different topic is "religious building" with a score of 0.76, and the most similar topic is "hat" with a score of 0.96. These results imply that "religious buildings" should be annotated more widely as their visual appearance varies across countries.

Finally, we investigate whether performance of similarity calculation depends on amount of annotated data. We find that at topic level the similarity scores are not correlated with the amount of annotated data (Pearson correlation coefficient is -0.02). We discuss more about the effect of data size on our analysis results in the Appendix.

Figure 5: Top three and last three countries (left) and topics (right) sorted by average similarity score.

Topic visual representation across countries in low-resource data. To show how the topic visual representations vary per country in the low-resource data, we perform a 2D transformation using Principal Component Analysis (PCA) (F.R.S., 1901). In Figure 6, we show the CLIP average representations per country for the topics with the most and least visual differences across countries: "religious building" and "hat", respectively. As expected, the representations for "religious building" are much more spread across countries than those for "hat", which tend to cluster together. In Appendix Figure 20, 21, and 22, we show representations for other topics visually different across countries: "get water", "house" and "backyard."

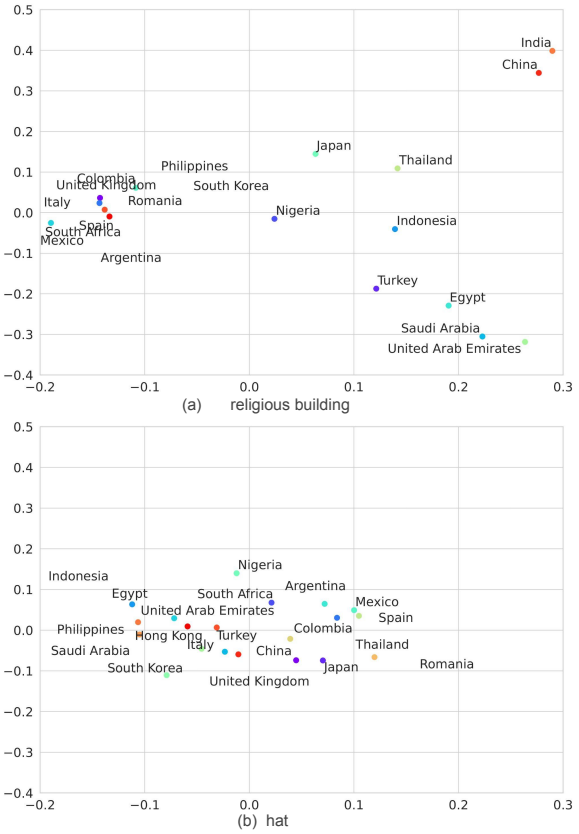

Figure 6: PCA for the topic "religious building" and "hat" for all countries in the low-resource data that contain this topic. The data is represented as the average of the CLIP representations.

Correlation between geographical distance and visual similarity across countries. We measure if the visual similarity between countries correlates with the geographical distance. The geographical distance between two countries is calculated using Vincenty's distance (Vincenty, 1975) between their capital cities.[7] The visual similarity between any two countries is calculated across all their shared topics. We compute the Pearson correlation coefficient (Freedman et al., 2007) over all countries and obtain a value of -0.01, indicating a weak negative correlation. A strong negative correlation is initially more expected as, intuitively, their visual similarity should increase as the distance between countries decreases. However, when we break down the correlation at the country level, the correlation coefficient varies significantly per country. In Figure 7, we show countries with weak to moderate positive correlations (e.g., $Haiti$ with 0.35, $Tunisia$ with 0.30), countries with weak to moderate negative correlations (e.g., $Vietnam$ with -0.35, $Burundi$ with -0.34), most countries have values close to 0, indicating no correlation between visual similarity and geographi-

[7] https://github.com/rahulbot/distances-between-countries

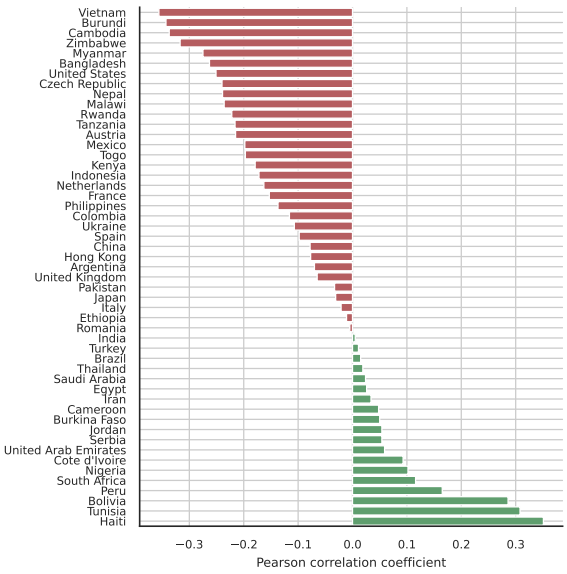

Figure 7: Pearson correlation coefficient between the visual similarity and the geographical distance, across countries. Most countries do not have a significant correlation between visual similarity and location.

cal distance. Upon close examination of the results, we determine the reasons behind this result: countries with positive correlation are often visually similar to countries from different continents (e.g., $Tunisia$ is more similar to $Bolivia$ with an average similarity 0.94 and distance $9,773$ than to $Austria$ with with an average similarity 0.80 and distance $1,362$). We hypothesize this might be due to history, climate, and/or income differences, which could contribute more to visual similarity than distance alone. Our analysis shows that geographical location does not generally correlate with visual similarity. Therefore, collecting globally diverse annotations on a budget requires considering other complementary information, such as the country's income, culture, history, and climate. Our results on which countries are similar to each other provide valuable insights into how to distribute the annotation budget effectively and can be used along with this complementary information.

Augmenting with data from visually similar countries significantly improves model performance. We train a classifier to predict the topic of the input images and measure the accuracy while controlling for the countries. Specifically, we input the CLIP visual representation in a linear layer, followed by a softmax to predict the topics of the input images.[8] We select one random country for each topic from the low-resource data, which we call target (topic, country) pairs. Next, we split the data into training and test sets in a 90-10% data split to include all the target (topic, country) pairs in both sets. Finally, we replace different ratios (100%, 90%, 70%, 50%, 30%, 10%, 0%) of the target-country data with images from: (1) the most *similar* countries to the target-country given the target-topic; (2) the most *dissimilar* countries to the target-country given the target-topic; (3) *high-resource* data corresponding to the target-topic.

The topic classification accuracy when using all the training target-country data is 91.1%, which is an upper bound. In Figure 8, we show the accuracy when adding data from (1), (2) and (3). The main takeaway is that **adding data from *similar* countries improves the performance more than adding data from *dissimilar* countries or *high-resource* data, and the gap in performance increases with the replacement ratio**. Additionally, supplementing with *high-resource* data is generally more beneficial than supplementing with data from *dissimilar* countries. We also compute the accuracy when no data is added, and find that adding data from *dissimilar* countries or from *high-resource* data can hurt the performance compared to not adding data, especially for high replacement ratios ($50\% - 90\%$). We show the results in the Appendix, in Figure 23.

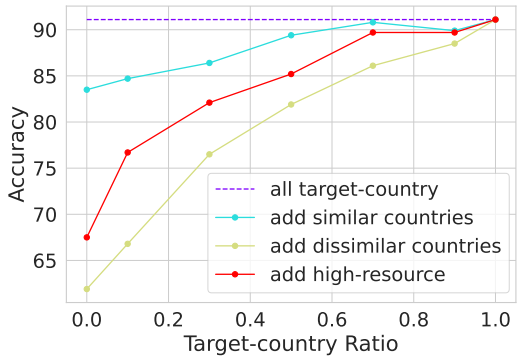

Figure 8: Topic classification accuracy (in %) for different target-country data ratios (e.g., target-country ratio 0.0% is equivalent to 100% replacement ratio). We replace different ratios of the target-country data with images from: (1) the most similar countries to the target-country given the target-topic; (2) the most dissimilar countries to the target-country given the target-topic; (3) high-resource data of the target-topic;

6. Main Takeaways

Our analyses provide multiple insights into the current state of vision-language annotations for various topics across different countries, and show the coverage limitations of existing large-scale datasets. We highlight the main takeaways and

[8] We set the learning rate as 5e-3, use AdamW as the optimizer, and conduct training over 250 epochs with a batch size of 512. Additionally, we use a cosine annealing schedule with 50 warm-up epochs.

propose actionable steps to help future work create more inclusive datasets and models.

We recommend focusing the annotation efforts on currently underrepresented data. To have more inclusive models and datasets, we need to collect more globally diverse annotations. Because annotations are expensive, we propose to focus future annotation efforts on specific countries and their topics. To assist with these efforts, we provide a list of countries and corresponding topics that are consistently unrepresented in the training data of vision-language models. Furthermore, most countries have less than ten images per topic. For most countries and corresponding topics – 3,329/ 4,830, we could not determine how similar they are to the high-resource data because of the lack of data. These countries have less than ten images per topic and, therefore, already need annotations. As an alternative solution, we recommend developing algorithms that can perform well with limited amount of data.

We can leverage cross-country data similarity to supplement data from unrepresented countries effectively. When we do not have a sufficient budget to annotate more data for a target country and topic, we propose using the available data from countries with similar visual representations of that given topic. We provide a list of similar countries for each target country and topic and show that using this data improves model performance more than using data from dissimilar countries or high-resource data.

Geographical distance does not correlate with visual similarity between countries. We compute the Pearson correlation coefficient between the visual similarity and the geographical distance between all countries and find a very weak negative correlation of -0.01. Therefore, collecting globally diverse annotations requires considering additional information. Multiple other factors, such as income, history, or cultural heritage, can contribute to the visual similarity between countries. We find this hypothesis worth investigating in depth in future work.

Visual similarity between countries and topics depends on the context. While examining images of topics across countries, we notice visually similar topics with very different backgrounds, which influence the visual similarity score. For example, in Figure 9, many countries have the same type of toothbrush, but because their storage place is different, their visual similarity score is low. In this paper, we measure similarity at the context level, considering both the topic and the context (e.g., background, storage space). However, as future work, we propose to investigate further which type of similarity to consider when we annotate diverse data: either at the topic level, by extracting the segmentation mask of the topic, or at the context level, by considering the entire image.

Figure 9: The context of the topic influences the visual similarity. For example, although the same type of toothbrush is depicted, their storage place differs, i.e., on a piece of wood, in a plastic container in the bathroom, in a plastic container tied to a tree, near a brick wall. Therefore, visual diversity is measured not only at the topic level but also at the context level.

7. Conclusion

In this paper, we addressed the need for balanced data representation used to train vision-language models. Because data annotations are expensive, we proposed to annotate primarily images from unrepresented countries. To find which countries are less represented in the training data of vision-language models, we compared the visual similarity of images across 94 topics and 52 countries found in crowd-sourced and web-scraped data. We used three visual representations, CLIP, BLIP-2, and ALIGN, to ensure the results generalize across representation types. Additionally, we proposed to leverage cross-country data similarity to improve model performance. We found visually similar countries for each country and corresponding topics and made them available in our repository: https://github.com/MichiganNLP/visual_diversity_budget. Finally, our analysis offers multiple takeaways for future work to make informed decisions on what global data to annotate and how to leverage cross-country data similarity to improve model representation. Through our work, we hope to contribute to building more inclusive and affordable vision-language models and datasets to help democratize AI globally.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback and are also grateful to the members of the Language and Information Technologies (LIT) lab at the University of Michigan for the insightful discussions during the project's early stages. This material is partly based on work supported by the Automotive Research Center ("ARC") at the University of Michigan. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of ARC or any other related entity.

8. Bibliographical References

Daniel Ahfock and Geoffrey J McLachlan. 2023. Semi-supervised learning of classifiers from a statistical perspective: A brief review. *Econometrics and Statistics*, 26:124–138.

William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. 2018. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9368–9377.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2022. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Laura Burdick, Rada Mihalcea, Ryan L. Boyd, and James W. Pennebaker. 2017. Multimodal analysis and prediction of latent user dimensions. In *Social Informatics*.

Michele Cafagna, Kees van Deemter, and Albert Gatt. 2021. What vision-language modelssee'when they see scenes. *arXiv preprint arXiv:2109.07301*.

Alan Kam Leung Chan, Chinasa T. Okolo, Zachary Terner, and Angelina Wang. 2021. The limits of global inclusion in ai development. *ArXiv*, abs/2102.01265.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568.

Yu-Wei Chao, Zhan Wang, Rada Mihalcea, and Jia Deng. 2015. Mining semantic affordances of visual object categories. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4259–4267.

Yanbei Chen, Massimiliano Mancini, Xiatian Zhu, and Zeynep Akata. 2022. Semi-supervised and unsupervised deep visual learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*.

Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopalan, Afshin Rostamizadeh, and Sanjiv Kumar. 2021. Batch active learning at scale. *Advances in Neural Information Processing Systems*, 34:11933–11944.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. 2019. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 52–59.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.

Abhimanyu Dubey, Vignesh Ramanathan, Alex 'Sandy' Pentland, and Dhruv Kumar Mahajan. 2021. Adaptive methods for real-world domain generalization. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14335–14344.

Ismail Elezi, Zhiding Yu, Anima Anandkumar, Laura Leal-Taixe, and Jose M Alvarez. 2022. Not all labels are equal: Rationalizing the labeling costs for training object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14492–14501.

Meng Fang and Trevor Cohn. 2017. Model transfer for tagging low-resource languages using a bilingual dictionary. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 587–593, Vancouver, Canada. Association for Computational Linguistics.

Christiane D. Fellbaum. 2000. Wordnet : an electronic lexical database. *Language*, 76:706.

David Freedman, Robert Pisani, and Roger Purves. 2007. Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*.

Karl Pearson F.R.S. 1901. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *International conference on machine learning*, pages 1183–1192. PMLR.

Noa Garcia, Yusuke Hirota, Yankun Wu, and Yuta Nakashima. 2023. Uncurated image-text datasets: Shedding light on demographic bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6957–6966.

Aparna Garimella, Rada Mihalcea, and James W. Pennebaker. 2016. Identifying cross-cultural differences in word usage. In *International Conference on Computational Linguistics*.

Timnit Gebru. 2020. Race and gender. *The Oxford handbook of ethics of al*, pages 251–269.

Gregor Geigle, Radu Timofte, and Goran Glavaš. 2023. Babel-imagenet: Massively multilingual evaluation of vision-and-language representations. *arXiv preprint arXiv:2306.08658*.

Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. 2022a. Vision models are more robust and fair when pretrained on uncurated images without supervision. *ArXiv*, abs/2202.08360.

Priya Goyal, Adriana Romero-Soriano, Caner Hazirbas, Levent Sagun, and Nicolas Usunier. 2022b. Fairness indicators for systematic assessments of visual feature extractors. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.

Laura Gustafson, Megan Richards, Melissa Hall, Caner Hazirbas, Diane Bouchacourt, and Mark Ibrahim. 2023. Pinpointing why object recognition performance degrades across income levels and geographies. *ArXiv*, abs/2304.05391.

Mohamed Farouk Abdel Hady and Friedhelm Schwenker. 2013. Semi-supervised learning. *Handbook on Neural Information Processing*, pages 215–239.

Melissa Hall, Bobbie Chern, Laura Gustafson, Denisse Ventura, Harshad Kulkarni, Candace Ross, and Nicolas Usunier. 2023. Towards reliable assessments of demographic disparities in multi-label image classifiers. *arXiv preprint arXiv:2302.08572*.

Sariel Har-Peled and Akash Kushal. 2005. Smaller coresets for k-median and k-means clustering. In *Proceedings of the twenty-first annual symposium on Computational geometry*, pages 126–134.

Xiao Hu, Haobo Wang, Anirudh Vegesana, Somesh Dube, Kaiwen Yu, Gore Kao, Shuo-Han Chen, Yung-Hsiang Lu, George K. Thiruvathukal, and Ming Yin. 2020. Crowdsourcing detection of sampling biases in image datasets. In *Proceedings of The Web Conference 2020*, WWW '20, page 2955–2961, New York, NY, USA. Association for Computing Machinery.

Hanyao Huang, Ou Zheng, Dongdong Wang, Jiayi Yin, Zijin Wang, Shengxuan Ding, Heng Yin, Chuan Xu, Renjie Yang, Qian Zheng, et al. 2023a. Chatgpt for shaping the future of dentistry: the potential of multi-modal large language model. *International Journal of Oral Science*, 15(1):29.

Xinyu Huang, Youcai Zhang, Jinyu Ma, Weiwei Tian, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Lei Zhang. 2023b. Tag2text: Guiding vision-language model via image tagging.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Tarun Kalluri, Wangdong Xu, and Manmohan Chandraker. 2023. Geonet: Benchmarking unsupervised adaptation across geographies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15368–15379.

Zu Whan Kim, Andre F. de Araújo, Bingyi Cao, Cameron S Askew, Jack Sim, Mike Green,

N'Mah Fodiatu Yilla, and Tobias Weyand. 2021. Towards a fairer landmark recognition dataset. *ArXiv*, abs/2108.08874.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. 2023a. Segment anything. *ArXiv*, abs/2304.02643.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023b. Segment anything. *arXiv preprint arXiv:2304.02643*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *ArXiv*, abs/1908.03557.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Robert Munro Monarch. 2021. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster.

Joan Nwatu, Oana Ignat, and Rada Mihalcea. 2023. Bridging the digital divide: Performance variation across socio-economic factors in vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. 2018. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in neural information processing systems*, 31.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing. *Computational Linguistics*, 45(3):559–601.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Vikram V. Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron B. Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. 2023. Beyond web-scraping: Crowd-sourcing a geographically diverse image dataset. *ArXiv*, abs/2301.02560.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. 2021. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40.

T. Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. 2021. Imagenet-21k pretraining for the masses. *ArXiv*, abs/2104.10972.

William Gaviria Rojas, Sudnya Diamos, Keertan Ranjan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. 2022. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. In *Neural Information Processing Systems*.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494.

Hadi Salman, Saachi Jain, Andrew Ilyas, Logan Engstrom, Eric Wong, and Aleksander Madry. 2022. When does bias transfer in transfer learning? *arXiv preprint arXiv:2207.02842*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *ArXiv*, abs/2210.08402.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs. In *Proceedings of the NeurIPS Data Centric AI Workshop*.

Candice Schumann, Susanna Ricco, Utsav Prabhu, Vittorio Ferrari, and Caroline Pantofaru. 2021. A step toward more inclusive people annotations for fairness. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 916–925.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv: Machine Learning*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale.

Kamal Taha. 2023. Semi-supervised and unsupervised clustering: A review and experimental evaluation. *Information Systems*, page 102178.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Thaddeus Vincenty. 1975. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Scientific Research and Essays*, 23:88–93.

Angelina Wang, Arvind Narayanan, and Olga Russakovsky. 2020. Revise: A tool for measuring and mitigating bias in visual datasets. *International Journal of Computer Vision*, 130:1790 – 1810.

Angelina Wang and Olga Russakovsky. 2023. Overcoming bias in pretrained models by manipulating the finetuning dataset. *arXiv preprint arXiv:2303.06167*.

Xudong Wang, Long Lian, and Stella X. Yu. 2021. Unsupervised selective labeling for more effective semi-supervised learning. In *European Conference on Computer Vision*.

Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z Chen. 2017. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*, pages 399–407. Springer.

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 2022. Merlot reserve: Neural script knowledge through vision and language and sound. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16354–16366.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. Merlot: Multimodal neural script knowledge models. In *Neural Information Processing Systems*.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition (CVPR), pages 5579–5588.

Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, Yandong Guo, and Lei Zhang. 2023. Recognize anything: A strong image tagging model.

Dora Zhao, Jerone Andrews, and Alice Xiang. 2023. Men also do laundry: Multi-attribute bias amplification. In *International Conference on Machine Learning*, pages 42000–42017. PMLR.

A. Subjective Topics

The 19 subjective topics that we remove: “favorite home decorations”, “favourite item in kitchen”, “favourite sports clubs”, “how the most loved item is used”, “icons”, “idols”, “latest furniture bought”, “looking over the shoulder”, “most loved item”, “most loved toy”, “most played songs on the radio”, “music idol”, “next big thing you are planning to buy”, “playing with most loved toy”, “thing I dream about having”, “things I wish I had”, “using most loved item”, “youth culture”, “what I wish I could buy”.

B. Data Stats

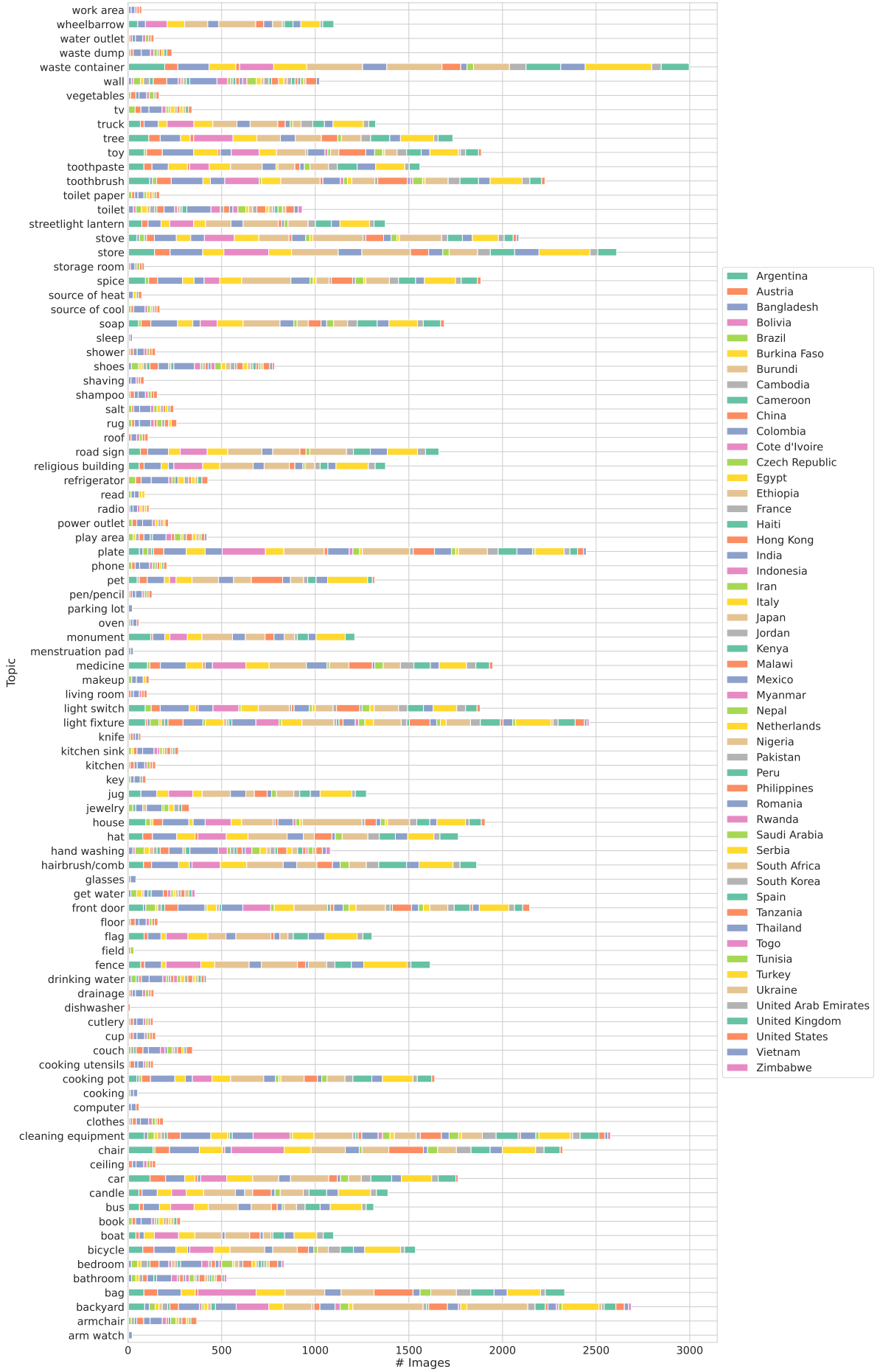

Figure 10: The distribution of countries per topic.

C. Research Question 1

C.1. USA respresentations.

Figure 11: Representative images from the visually different topics in low-resource USA data/ L, and high-resource data/ H. In H, "clothes" and "makeup" are shown on people, while in L they are separated in dressers and containers; in H, "spice" is in in large baskets in markets, while in L they are in small containers in people's houses; in H, "ceiling" is shown in public spaces, while in L is in private homes; in H "medicine" is usually in bottles, while in L can be in various forms.

C.2. The effect of data size on the data analysis results.

In Figure 5, $Burundi$ has the lowest similarity score of 0.775 and has very little data in the heatmap of Figure 3, only 97 images. Note however there are many counter-examples worth considering, such as countries with fewer images and high similarity scores (e.g., $Austria$ has eleven images and a similarity score of 0.862, $Bolivia$ has 37 images and a similarity score of 0.874), or on the opposing spectrum, countries with more images and low similarity scores (e.g., $Malawi$ has 390 images and a similarity of 0.776, $Burkina\ Faso$ has 253 images and a similarity score of 0.789).

Furthermore, at topic level, the similarity scores and data size are not correlated (Pearson correlation score is -0.02). Similar to the country level, there are topics with many images and low similarity scores (e.g., $religious\ building$ has 1,375 images and a similarity of 0.764) and topics with few images and high similarity scores (e.g., $glasses$ has 42 images and a similarity of 0.95).

In general, while data size can have an influence on our analysis results, we believe our work provides helpful strategies for annotation when data size is insufficient. Our paper is a call to action for future work to collect more globally diverse data to improve the robustness of the results.

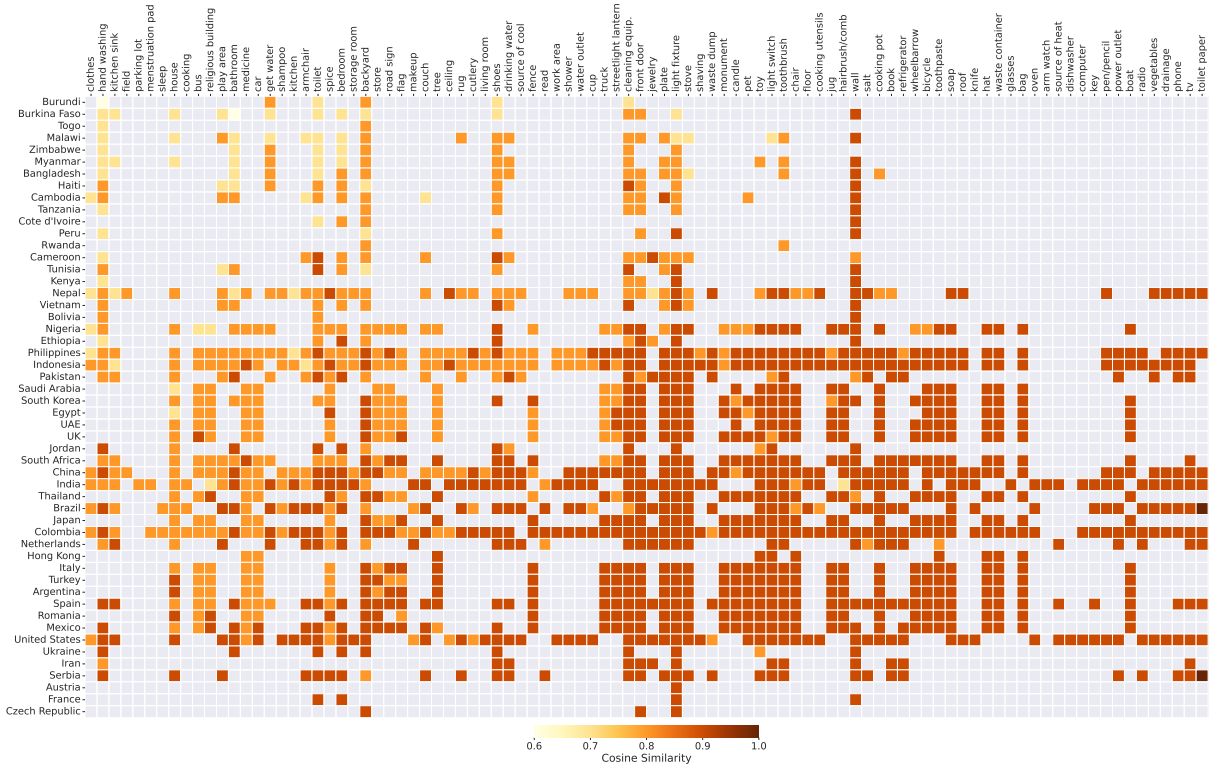

Figure 12: Similarity heatmap of (topic, country) pairs with CLIP visual representations. The darker, the less similarity between high-resource and low-resource data for that corresponding (topic, country), the more beneficial it is to annotate. Empty cells do not have any images for (topic, country). *Best viewed in color.*

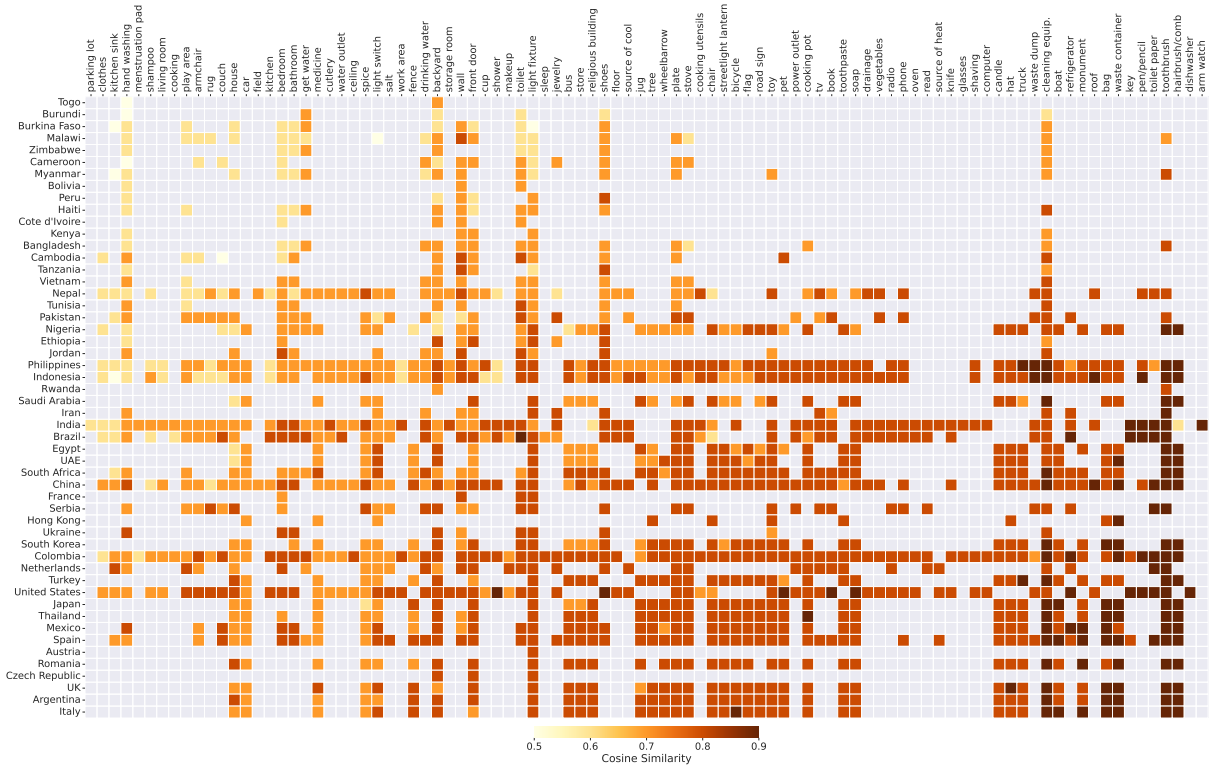

Figure 13: Similarity heatmap of (topic, country) pairs with ALIGN visual representations. The darker, the less similarity between high-resource and low-resource data for that corresponding (topic, country), the more beneficial it is to annotate. Empty cells do not have any images for (topic, country). *Best viewed in color.*

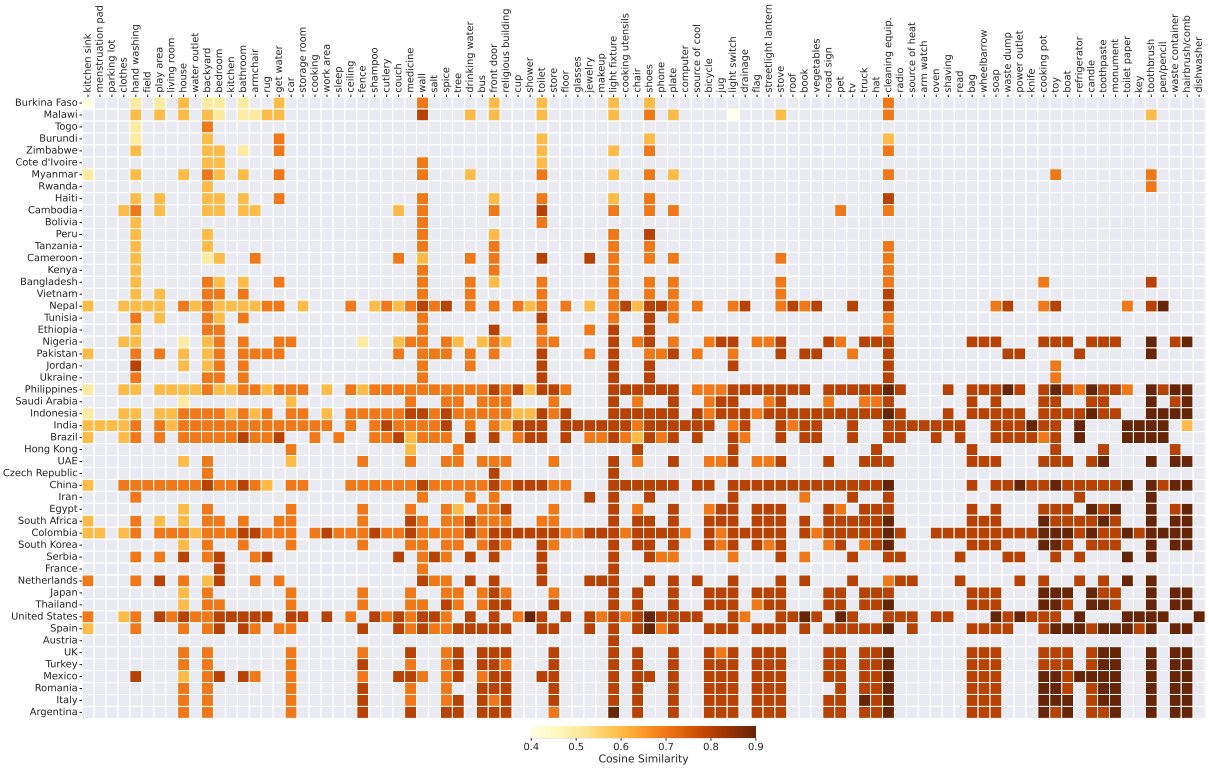

Figure 14: Similarity heatmap of (topic, country) pairs with BLIP visual representations. The darker, the less similarity between high-resource and low-resource data for that corresponding (topic, country), the more beneficial it is to annotate. Empty cells do not have any images for (topic, country). *Best viewed in color.*

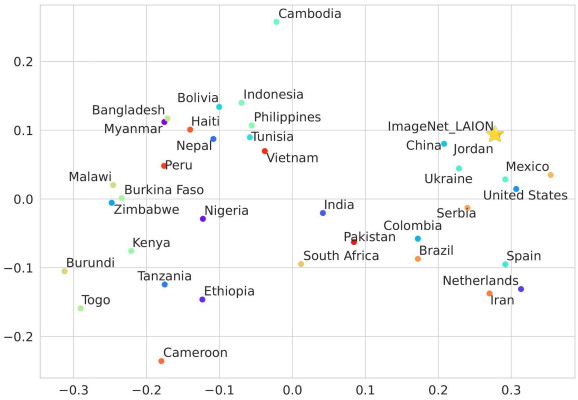

Figure 15: PCA for the topic “hand washing” for all countries that contain this topic in the low-resource data and in the high-resource data. The high-resource data point is highlighted. The data is represented as the average of the CLIP representations.

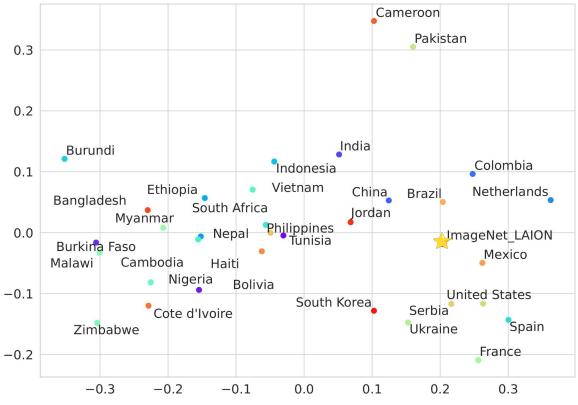

Figure 16: PCA for the topic “toilet” for all countries that contain this topic in the low-resource data and in the high-resource data. The high-resource data point is highlighted. The data is represented as the average of the CLIP representations.

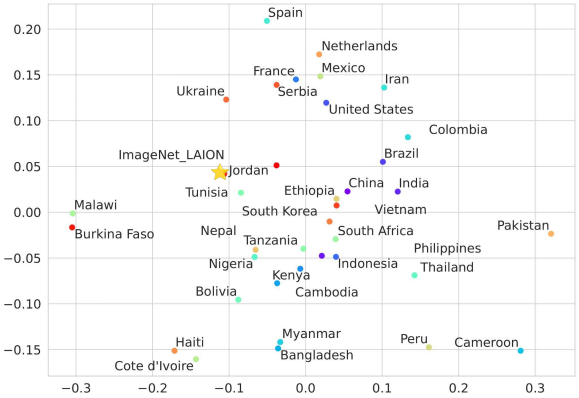

Figure 17: PCA for the topic “wall” for all countries that contain this topic in the low-resource data and in the high-resource data. The high-resource data point is highlighted. The data is represented as the average of the CLIP representations.

D. Research Question 2

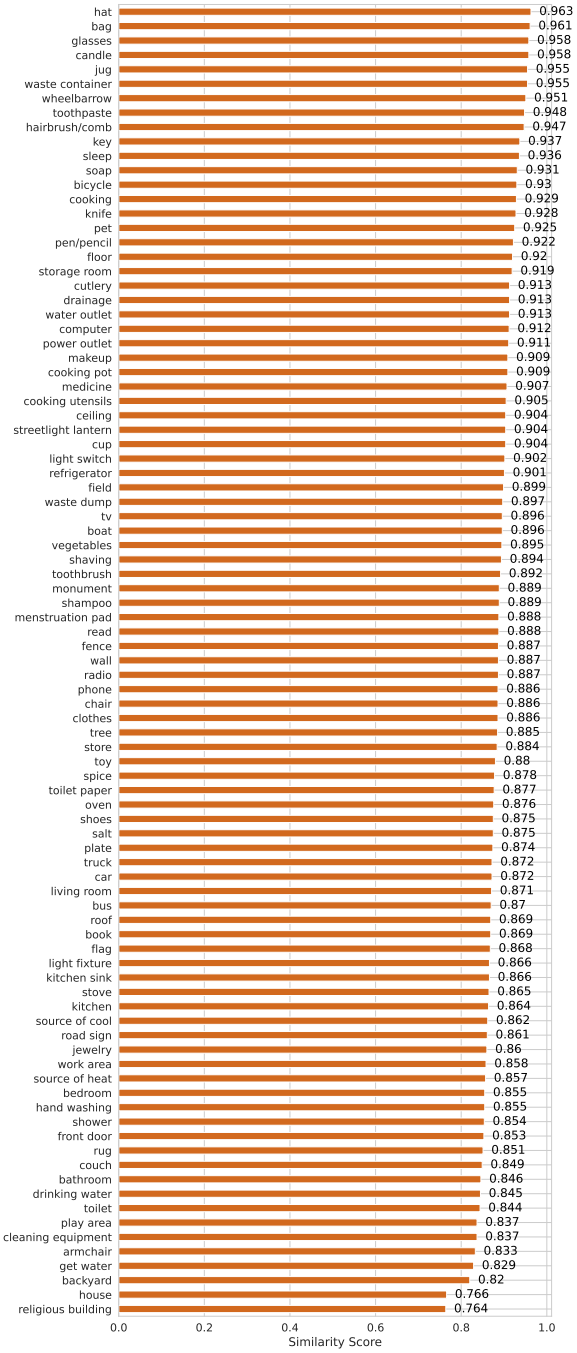

Figure 18: The distribution of average similarity scores per topic.

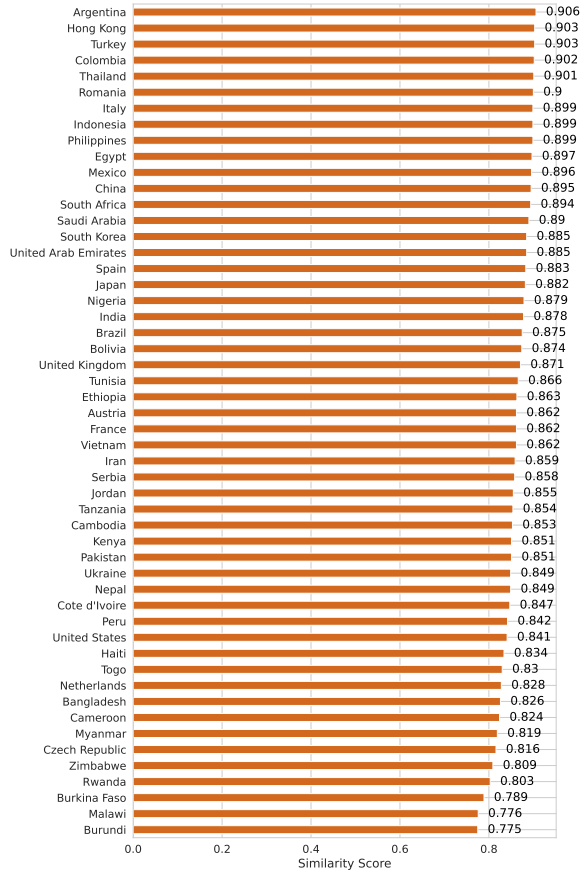

Figure 19: The distribution of average similarity scores per country.

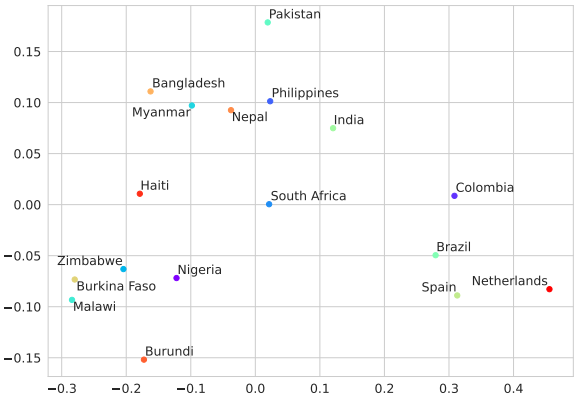

Figure 20: PCA for the topic “get water” for all countries that contain this topic in the low-resource data. The data is represented as the average of the CLIP representations.

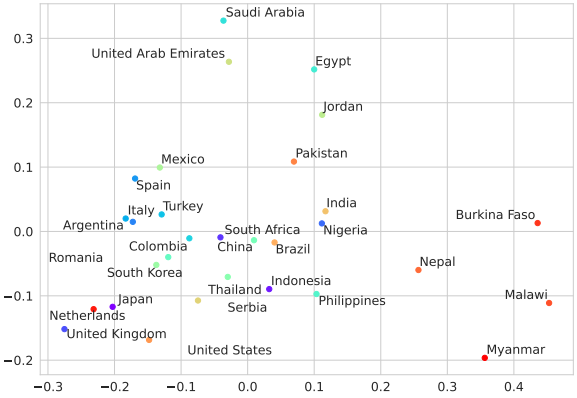

Figure 21: PCA for the topic “house” for all countries that contain this topic in the low-resource data. The data is represented as the average of the CLIP representations.

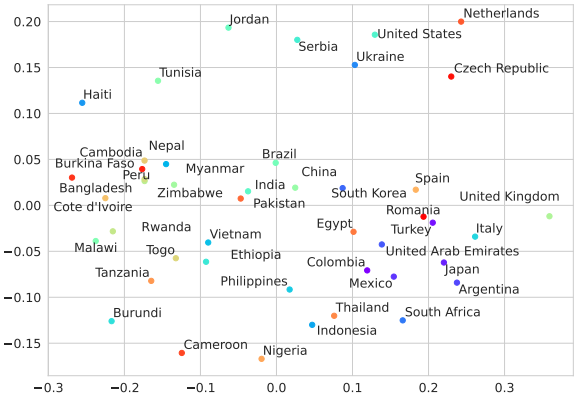

Figure 22: PCA for the topic “backyard” for all countries that contain this topic in the low-resource data. The data is represented as the average of the CLIP representations.

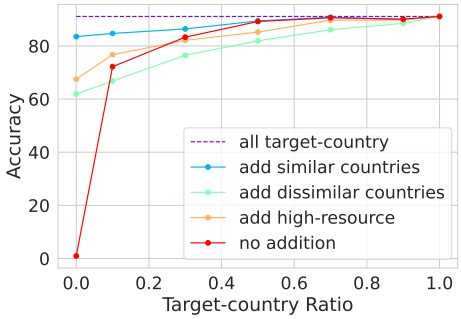

Figure 23: Topic classification accuracy (in %) for different target-country data ratios (e.g., target-country ratio 0.0% is equivalent to 100% replacement ratio). We replace different ratios of the target-country data with images from: (1) the most similar countries to the target-country given the target-topic; (2) the most dissimilar countries to the target-country given the target-topic; (3) high-resource data of the target-topic; (4) no replacement data/ addition.