The Data Part of the Thesis

Introduction

In previous chapters, we introduced the challenge of cultural bias and underrepresentation in Vision-Language Models (VLMs), emphasizing the limitations these models face when dealing with culturally diverse visual and textual data. To mitigate these limitations, numerous research initiatives have emerged, focusing on the creation of high-quality datasets that either benchmark or fine-tune models for better cultural comprehension.

This section delves into some of the most influential datasets developed to address cultural diversity in multimodal models. We also describe our own data-centric approaches, including the construction of a culturally-aware VQA dataset and the application of Retrieval-Augmented Generation (RAG) in our system.

1. CANDLE Dataset: Building Cultural Commonsense Knowledge

One of the foundational efforts in this domain is the **CANDLE dataset**, which was introduced to confront a specific gap in Language Models—namely, the lack of Cultural Commonsense Knowledge (CCSK). CCSK refers to the intuitive, context-sensitive understanding of cultural norms, practices, and behaviors that human beings acquire through social exposure but which language models often lack.

CANDLE introduces an **end-to-end pipeline** for extracting and organizing CCSK from web-scale corpora. The methodology leverages the **C4 dataset**—a massive web crawl corpus developed by Google—as its data source. Pre-trained LMs are used as sub-modules for specific subtasks, such as cultural facet classification, which helps categorize statements into culture-relevant domains like food, traditions, values, etc.

CANDLE serves primarily as a **textual knowledge base**, offering structured assertions like "In Japan, people remove their shoes before entering a house." However, it lacks visual elements, which limits its utility in training VLMs.

2. CultureVQA: Visual Extension of Cultural Knowledge

To bridge the gap between textual and visual cultural knowledge, researchers extended CANDLE to form **CultureVQA**, a benchmark designed specifically to evaluate the cultural understanding of Vision-Language Models in a **visual question answering** (VQA) setting.

CultureVQA enhances CANDLE by **automatically retrieving images** corresponding to the cultural assertions listed in the textual dataset. These images serve as the visual grounding for the creation of image-question pairs.

To ensure cultural authenticity and nuance, the team behind CultureVQA employed **human annotators from diverse cultural backgrounds** via Amazon Mechanical Turk (MTurk). These annotators, familiar with the respective cultural concepts depicted in the images, manually generated questions and answers. Initially, the authors explored using LLMs for question generation, but found that the results lacked subtlety and cultural specificity—prompting a shift to human-based annotation.

Limitations of CultureVQA include:

- **English-language bias**: Image and concept retrieval was limited to English-language websites.
- **Geo-regional simplification**: Cultural groups were approximated by region (e.g., Southeast Asia) due to scalability constraints.

Despite these challenges, CultureVQA presents a well-structured benchmark of **2,378 image-question pairs**, representing cultures from **11 countries across 5 continents**, covering topics such as **clothing, food, rituals, traditions, and beverages**. Each image is paired with 1–5 answers, making up a total of **7,206 answers**.

3. CVQA: A Multilingual and Multicultural Benchmark

While CultureVQA emphasizes cultural imagery, **CVQA** (Culturally-diverse Multilingual Visual Question Answering) takes cultural evaluation further by introducing **multilingualism** into the VQA domain. The dataset was carefully curated to challenge models not just on cultural imagery but also in processing **local languages and scripts**, a known weakness in most current MLLMs (Multimodal Large Language Models).

Key characteristics of CVQA:

- **9,000+ questions**
- Representing **28 countries**
- Written in **26 languages** and **11 different scripts**
- Organized by **country-language pairs**

Unlike previous datasets, CVQA uses **native speakers and cultural experts** during the annotation phase to ensure questions are both linguistically and culturally accurate. Questions span categories such as local history, geography, cuisine, and rituals. These questions are formatted in a **multiple-choice setting**, allowing objective benchmarking.

Some questions were derived from **UNESCO Cultural Heritage** materials—especially for Southeast Asia—while others were generated using **GPT-4**, followed by human review and refinement. Each question was paired with both a **local language version** and an **English translation**, making the dataset highly accessible and suitable for multilingual evaluation.

Benchmarking on CVQA has shown that current VLMs, including the most advanced models, often **fail to generalize well to non-English, culturally dense questions**, especially those embedded with regional or traditional nuance.

4. CultureBank: Grounded Cultural Knowledge from Social Media

Another innovative dataset is **CultureBank**, which diverges from traditional academic or curated sources by collecting **real-world cultural expressions from social media platforms**—primarily **TikTok and Reddit**. This approach provides a unique bottom-up perspective, capturing lived cultural experiences, humor, traditions, and behaviors that are often **invisible in formal datasets**.

CultureBank includes:

- **12,000 cultural descriptors** from TikTok
- **11,000 from Reddit**
- A structured **taxonomy** including:
    - Cultural group (e.g., “Californian” or “Egyptian Muslim”)
    - Context (e.g., wedding, festival)

    - Goal, Actor, Behavior, Topic

To curate the dataset, researchers developed a **generalizable pipeline** to process noisy, unstructured narratives from social media into structured cultural scenarios. These scenarios were then used in a **grounded evaluation dataset**, which includes:

- Simulated consulting sessions
- Client personas
- Associated questions

The team used a **Mixtral model** (a Mixture-of-Experts transformer) fine-tuned on GPT-4-generated data and refined it with **human oversight** to ensure consistency and cultural authenticity. This pipeline enabled scalable yet grounded dataset creation across a wide cultural spectrum, offering a unique benchmark for practical cultural applications like digital assistants or therapy bots.

Our Approaches

1. Vision Question Answering (VQA) Dataset Construction

In our own work, we began by designing a VQA dataset with the objective of fine-tuning a VLM to better interpret and extract cultural elements from images.

However, this approach prompted an important design question:
"Should we ask the same questions across all images, or tailor questions to each image based on its cultural context?"

To explore this, we investigated two main strategies:

a. Image-Specific Question Generation

We used a pretrained generative model (e.g., BLIP or MiniGPT) to generate questions for each image dynamically. These questions were conditioned on the visual content and aimed to probe potential cultural cues present in the image. However, the approach often produced generic or irrelevant questions, revealing the need for better cultural grounding.

b. Categorical Question Templates

In this method, we manually defined a taxonomy of **cultural categories**—such as clothing, food, festivals, and religious symbols. Images were first classified into one of these categories using a visual classifier. Then, a **predefined question set** relevant to that category was posed. This method offered better control and consistency, especially when aiming to benchmark performance across cultural dimensions.

This dual-experimentation helped us understand how question design directly impacts the model's ability to reason about cultural elements.

2. Retrieval-Augmented Generation (RAG) for Cultural Context