Multimodal Visual Question Answering with Amazon Berkeley Objects Dataset
AIM-825 Visual Recognition Project

Varsha Yamsani(IMT2022506)
R Harshavardhan(IMT2022515)
Keshav Goyal(IMT2022560)

1

1. Introduction and Dataset

This assignment involves creating a multiple-choice Visual Question Answering (VQA) dataset using the Amazon Berkeley Objects (ABO) dataset, evaluating baseline models, fine-tuning using Low-Rank Adaptation (LoRA), and assessing performance using standard metrics.

We utilized the **Amazon-Berkeley Objects (ABO)** dataset, which comprises approximately 147,702 product listings with multilingual metadata and a total of 398,212 unique catalog images.

- **Image Data:** We used the downscaled catalog images (maximum 256 pixels) provided in the archive `abo-images-small.tar`.
- **Metadata:** Product listings and associated metadata were obtained from the `abo-listings.tar` archive.

The dataset is publicly available at the following link: ABO Dataset.

2. Data Curation and Methodology

To build a high-quality multimodal Visual Question Answering (VQA) dataset, we performed systematic data curation using the Amazon Berkeley Objects (ABO) dataset. The procedure involved several steps: decompressing metadata, verifying structural integrity, filtering for quality and consistency, linking with images, and generating visual questions using a large language model. The following subsections detail each step.

- **Metadata Extraction:**
  - The ABO metadata was provided in compressed *.json.gz* format. We first decompressed all metadata files using the Python *gzip* module
  - All *.json.gz* files were extracted to plain *.json* format.
  - The files were stored in a designated directory structure for further processing.
- **Structural Validation of JSON:**
  - To ensure compatibility and reliability, each JSON file was examined to determine its structure
  - If the content was a valid JSON array, it was parsed directly using *json.loads()*.
  - Otherwise, line-delimited JSON objects were parsed individually, line-by-line.
  - Files that could not be parsed or raised decoding errors were logged and skipped.

- **Metadata Structure:**
  - Figure 2 shows a sample metadata record from the ABO dataset. Each record is represented as a structured JSON object containing product-level details such as *item_name*, *bullet_point*, *brand*, *color*, and various image identifiers. These fields are often nested and include additional tags like *language_tag*, enabling language-specific filtering (e.g., only retaining English-language entries marked as *en_IN*).
  - The *bullet_point* field provides rich natural language descriptions that are leveraged for generating visual questions. Each product can have multiple associated images, referenced by *main_image_id* and *other_image_id*, facilitating multimodal alignment. Standardized values in fields such as *color* (e.g., "multi-colored") help ensure consistency during dataset preprocessing.
  - Additional fields such as *item_id*, *model_number*, and *item_keywords* help uniquely identify and group product information. The *node* field provides hierarchical category information (e.g., "Cases & Covers / Back & Bumper Cases"), which is useful for contextual organization and filtering.

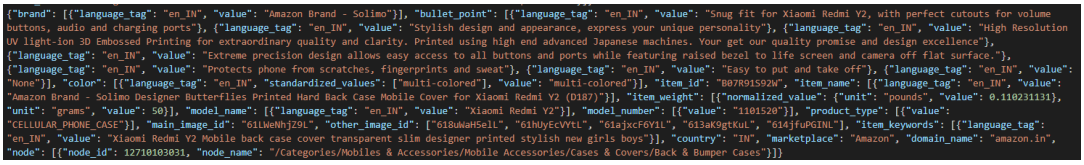
{"brand": [{"language_tag": "en_IN", "value": "Amazon Brand - Solimo"}], "bullet_point": [{"language_tag": "en_IN", "value": "Snug fit for Xiaomi Redmi Y2, with perfect cutouts for volume buttons, audio and charging ports"}, {"language_tag": "en_IN", "value": "Stylish design and appearance, express your unique personality"}, {"language_tag": "en_IN", "value": "High Resolution UV light-ion 3D Embossed Printing for extraordinary quality and clarity. Printed using high end advanced Japanese machines. Your get our quality promise and design excellence"}, {"language_tag": "en_IN", "value": "Extreme precision design allows easy access to all buttons and ports while featuring raised bezel to life screen and camera off flat surface."}, {"language_tag": "en_IN", "value": "Protects phone from scratches, fingerprints and sweat"}, {"language_tag": "en_IN", "value": "Easy to put and take off"}, {"language_tag": "en_IN", "value": "None"}], "color": [{"language_tag": "en_IN", "standardized_values": ["multi-colored"], "value": "multi-colored"}], "item_id": "B07R91S92W", "item_name": [{"language_tag": "en_IN", "value": "Amazon Brand - Solimo Designer Butterflies Printed Hard Back Case Mobile Cover for Xiaomi Redmi Y2 (D187)"}], "item_weight": [{"normalized_value": {"unit": "pounds", "value": 0.110231131}, "unit": "grams", "value": 50}], "model_name": [{"language_tag": "en_IN", "value": "Xiaomi Redmi Y2"}], "model_number": [{"value": "1101520"}], "product_type": [{"value": "CELLULAR_PHONE_CASE"}], "main_image_id": "61UweNhJ29L", "other_image_id": ["618uWaHSelL", "61hUyEcVYtL", "61ajxcF6Y1L", "613aK9gtKuL", "614jfuPGINL"], "item_keywords": [{"language_tag": "en_IN", "value": "Xiaomi Redmi Y2 Mobile back case cover transparent slim designer printed stylish new girls boys"}], "country": "IN", "marketplace": "Amazon", "domain_name": "amazon.in", "node": [{"node_id": 12710103031, "node_name": "/Categories/Mobiles & Accessories/Mobile Accessories/Cases & Covers/Back & Bumper Cases"}]}

Figure 1. Example metadata record from the ABO dataset.

- **Metadata Filtering:**
  - Next, we applied several filters to retain only relevant and linguistically valid product listings:
  - Only entries containing the keys *brand*, *bullet_point*, *color*, *model_name*, *item_name*, *product_type*, *main_image_id*, *item_keywords*, and *country* were retained.
  - Listings were restricted to those with *country* codes *IN* (India) and *US* (United States).
  - From nested metadata fields like *bullet_point*, *color*, and `item_keywords`, we filtered values based on their *language_tag*, retaining only entries with *en_US* or *en_IN*, or those without a language tag.
  - Standardized color values were also extracted for better textual description quality.

  The extracted and filtered information was saved into structured CSV files, where each row represented a single product listing with the following columns:
  - *main_image_id*
  - *overall_description* (from *bullet_point*)
  - *colour_description* (from both *color.standardized_values* and `color.value`)
  - *other_description* (from *product_type* and *item_keywords*)
  - *material_description* (from *material.value*, if available)
- **Image Linking:**
  To associate metadata with the actual product image:

	main_image_id	overall_description	colour_description	other_description	material_description
0	81+4dBN1jsL	Snug fit for Mi Redmi Go, with perfect cut-outs for volume buttons, audio and charging ports; Compatible with Mi Redmi Go; Easy to put & take off with perfect cutouts for volume buttons, audio & charging ports.; Stylish design and appearance, express your unique personality.; Extreme precision design allows easy access to all buttons and ports while featuring raised bezel to life screen and camera off flat surface.	multi-colored; Multicolor	CELLULAR_PHONE_CASE; cellphonecover; backcase; mobileguard; mobilecover; cellphonecase; phonecover; case; protectivecase; cover; backcover; phoneguard; mobilecase; protectivecover; phonecase; Back Cover; Designer Autumn Girl Mobile Cover; Designer Case; Hard Case; Mi Redmi Go Case; Printed Cover; cases and covers; fashion case; mobile cover; polycarbonate cover; Back Cover; Designer Autumn Girl Mobile Cover; Designer Case; Hard Case; Mi Redmi Go Case; Printed Cover; cases and covers; fashion case; mobile cover; polycarbonate cover; Back Cover; Designer Autumn Girl Mobile Cover; Designer Case; Hard Case; Mi Redmi Go Case; Printed Cover; cases and covers; fashion case; mobile cover; polycarbonate cover; Back Cover; Designer Autumn Girl Mobile Cover; Designer Case; Hard Case; Mi Redmi Go Case; Printed Cover; cases and covers; fashion case; mobile cover; polycarbonate cover; Back Cover; Designer Autumn Girl Mobile Cover; Designer Case; Hard Case; Mi Redmi Go Case; Printed Cover; cases and covers; fashion case; mobile cover; polycarbonate cover; Back Cover; Designer Autumn Girl Mobile Cover; Designer Case; Hard Case; Mi Redmi Go Case; Printed Cover; cases and covers; fashion case; mobile cover; polycarbonate cover; Back Cover; Designer Autumn Girl Mobile Cover; Designer Case; Hard Case; Mi Redmi Go Case; Printed Cover; cases and covers; fashion case; mobile cover; polycarbonate cover; Back Cover; Designer Autumn Girl Mobile Cover; Designer Case; Hard Case; Mi Redmi Go Case; Printed Cover; cases and covers; fashion case; mobile cover; polycarbonate cover; Back Cover; Designer Autumn Girl Mobile Cover; Designer Case; Hard Case; Mi Redmi Go Case; Printed Cover; cases and covers; fashion case; mobile cover; polycarbonate cover; Back Cover; Designer Autumn Girl Mobile Cover; Designer Case; Hard Case; Mi Redmi Go Case; Printed Cover; cases and covers; fashion case; mobile cover; polycarbonate cover; Back Cover; Designer Autumn Girl Mobile Cover; Designer Case; Hard Case; Mi Redmi Go Case; Printed Cover; cases and covers; fashion case; mobile cover; polycarbonate cover	Plastic

Figure 2. Example extracted metadata record from the ABO dataset.

  - A lookup map was created from the ABO image metadata (*images.csv*), mapping each *image_id* to its relative file path.
  - Listings without a matching image ID or missing image files were excluded from the next phase.
- **Visual Question Generation:**
  - We used the *Gemini 2.0 Flash* API by Google to generate visual questions based on product images and their associated metadata.
  - Each image was read and converted to byte format.
  - A prompt combining the image and a cleaned, structured textual description was sent to the LLM.
  - The prompt instructed the model to generate 5 diverse, image-dependent questions and answers, as per the following criteria:
    1. At least 2 simple questions (e.g., color, count).
    2. At least 2 moderately difficult questions (e.g., spatial relations, comparisons).
    3. 1 challenging question requiring subtle visual reasoning.
  - Answers were constrained to single-word responses to facilitate objective evaluation.
  - Only well-structured responses (exactly 5 question-answer pairs) were retained.

  To respect API limits:
  - A daily quota of 1500 requests was enforced.
  - A delay of 60 seconds was added between consecutive requests.
  - Progress was tracked using an index file to allow interruption and resumption without duplication.
- **Final Output:**

  The final curated dataset contains:
  - CSV files with image IDs, paths, and five generated question-answer pairs for each image.
  - Separate CSVs for filtered metadata and logs for images or entries that could not be processed.

  This process ensures a scalable, reproducible pipeline for curating large-scale multimodal VQA datasets from Amazon Berkley Dataset.

- Finally, all non-yes/no questions were retained, resulting in a final dataset comprising 120,000 question-answer pairs aligned with product images. This dataset

	image_id	full_image_path	question	answer
0	71owAzvPFuL	abo-images-small/images/small\cd/cd58ca00.jpg	Is the cover patterned?	Yes
1	71owAzvPFuL	abo-images-small/images/small\cd/cd58ca00.jpg	What shape is the charm?	Ghost
2	71owAzvPFuL	abo-images-small/images/small\cd/cd58ca00.jpg	What color is the heart?	Red
3	71owAzvPFuL	abo-images-small/images/small\cd/cd58ca00.jpg	Is there a lens visible?	Yes
4	71owAzvPFuL	abo-images-small/images/small\cd/cd58ca00.jpg	What symbol is present?	Yin-yang

Figure 3. Example questions for one of the products.

was split into training, validation, and testing sets with 80,000, 20,000, and 20,000 samples respectively. The *image_id* column was removed during preprocessing to focus purely on the textual and visual modalities relevant to the model.

	full_image_path	question	answer
0	abo-images-small/images/small\00/0029c3b2.jpg	What shapes are printed on the case?	Stars
1	abo-images-small/images/small\6d/6d06d0cd.jpg	What is depicted on the case?	Man
2	abo-images-small/images/small/d7/d7763e4d.jpg	what shape is featured?	skull
3	abo-images-small/images/small\33/33360502.jpg	what shape above hearts?	oval
4	abo-images-small/images/small/cd/cd678bbf.jpg	where is the speaker grille located?	top

Figure 4. Sample entries from the final training dataset after preprocessing.

3. Baseline Evaluation

We evaluated several pretrained VQA models, including BLIP, BLIP-2, Granite Vision, ViLT, CLIP, and BakLlava, on our 20,000 test datapoints. Initially, accuracy was used as the primary evaluation metric; later, we expanded to additional metrics which will be discussed subsequently. Below are the observations for each model:

Granite Vision: Granite Vision achieved the highest baseline accuracy on our test dataset, reaching approximately 52.5%. However, a significant limitation was its long inference time—11 hours and 43 minutes for 14,177 images. Due to Kaggle's 12-hour session timeout, we were unable to process all 20,000 images in a single session. As a result, we had to split the inference into two separate runs, which collectively took approximately 17 hours to complete for the full dataset.

BakLlava: BakLlava obtained the second-best accuracy at 49.4%. It also demonstrated a reasonable inference time of just over 3.5 hours, making it a strong baseline candidate.

ViLT: The ViLT model yielded an accuracy of 26.11%. Although its inference time was relatively short (approximately 1 hour and 15 minutes), the low accuracy limited our motivation to extensively apply LoRA fine-tuning; only a few experiments were conducted.

BLIP-2: BLIP-2, the successor to BLIP, performed poorly in comparison, achieving an exact match accuracy of 24.9%. This is likely because BLIP-2 is primarily pre-

trained on general image-text pairs and captioning tasks, which do not fully prepare it for the complex reasoning and specific question types inherent to VQA.

BLIP: BLIP served as the main model for extensive LoRA fine-tuning. Its baseline accuracy was 46.6%, with an inference time under 3 hours, representing the best trade-off between accuracy and inference time among the models tested.

CLIP: CLIP performed the worst out of all the models tried, giving a accuracy of mere 2.6%, with an inference time of around 4.5 hours, and thus, this model wasn't considered for LoRA finetuning.

Baseline Results:

Model	Test Accuracy (%)	BERTScore(Precision)	ROUGE Score	BERT Cosine Similarity	Levenshtein Distance
CLIP	2.6	0.1364	0.0299	0.2732	6.16
BLIP	46.60	0.9143	0.4756	0.7437	2.73
BLIP-2	24.90	0.7608	0.2587	0.5522	4.14
ViLT	26.11	0.7812	0.2734	0.5721	3.98
Granite Vision	52.50	0.8559	0.5389	0.7734	3.06
BakLlava	49.40	0.9145	0.5038	0.7514	2.64

Table 1. Performance comparison of baseline VQA models on multiple evaluation metrics.

Other models such as VilBERT and VisualBERT, referenced in the question documentation, were also explored. However, these models require CNN-based feature extraction as a preprocessing step, which proved to be computationally expensive and time-consuming. Given the large size of our dataset, it was not feasible to include these models in our evaluation.

4. Fine Tuning with LoRA

Initially, we attempted to apply LoRA fine-tuning on BakLlava, the model that achieved the second-highest baseline accuracy. However, the integration process proved to be technically challenging and less straightforward. Consequently, we opted to use BLIP instead, as it offered the best trade-off between inference time and accuracy.
Due to the limited GPU hours available on Kaggle, we were able to experiment with only six LoRA configurations—comprising three different values of the rank parameter r across two different training epochs. We also conducted a limited LoRA experiment on ViLT; however, the improvement in accuracy was negligible as compared to BLIP, and therefore we did not pursue further fine-tuning with it.

Parameters	Test Accuracy (%)	ROUGE Score	BERT Cosine Similarity	Training Loss
r=8 epoch=3	62.3	0.6358	0.8163	8.202490155029297
r=8 epoch=4	62.3	0.6362	0.8161	8.202452362813648
r=16 epoch=3	62.4	0.6367	0.8165	8.150731213617842
r=16 epoch=4	62.3	0.6360	0.8162	8.150731196085612
r=32 epoch=3	62.5	0.6375	0.8163	8.150330848185222
r=32 epoch=4	62.5	0.6374	0.8162	8.150340938145123

Table 2. Performance comparison of LoRA based BLIP Model for different configurations.

When LoRA was applied to the ViLT model, we observed a significant improvement in accuracy—from 26.11% to 54.46%. The trainable parameters for ViLT are shown in the figure below. Similarly, for BLIP as well only a small chunk of these parameters were trainable but the accuracy increase was significantly visible. This represents a substantial enhancement over the baseline performance, demonstrating the effectiveness of LoRA in improving the capabilities of VQA models. However, despite this gain, the final accuracy still fell short of that achieved by BLIP with LoRA fine-tuning. This comparison highlights how parameter-efficient fine-tuning methods like LoRA can meaningfully boost model performance in Visual Question Answering tasks.

Figure 5. Figure showing trainable parameters of LoRA applied on ViLT.

5. Evaluation Metrics

To evaluate the performance of VQA-based models comprehensively, we used a combination of traditional and semantic metrics. These include Test Accuracy, BERTScore (Precision), ROUGE Score, BERT Cosine Similarity, Levenshtein Distance, and F1 Score. Each of these metrics captures different aspects of model performance:

- **Test Accuracy (%):** Measures the percentage of exact matches between the predicted and ground-truth answers. While it is a straightforward and interpretable metric, it is often too rigid for natural language tasks where semantically similar answers may not match exactly.
- **BERTScore (Precision):** Uses contextual embeddings from a pretrained BERT model to compute the similarity between predicted and reference answers. This metric captures semantic similarity and is useful for understanding how well the model preserves meaning, even if the wording differs.
- **ROUGE Score:** Primarily used in summarization tasks, ROUGE measures the overlap of n-grams between predicted and ground-truth text. In the context of VQA, it helps evaluate how much of the answer content is correctly retrieved by the model.
- **BERT Cosine Similarity:** Computes the cosine similarity between sentence embeddings of the predicted and ground-truth answers. It gives a continuous-valued measure of semantic closeness and is particularly helpful when evaluating free-form or open-ended responses.
- **Levenshtein Distance:** Also known as edit distance, it quantifies how many single-character edits (insertions, deletions, substitutions) are needed to change one string into another. A lower Levenshtein Distance indicates that the predicted answer is textually close to the ground truth.

- **F1 Score:** The harmonic mean of precision and recall, F1 Score is useful when evaluating models on partially correct answers. It is particularly relevant for multi-word answers, where partial matches still reflect some level of understanding.

By using this diverse set of metrics, we can capture both the exactness and semantic quality of model predictions. For instance, a model with lower accuracy but higher BERTScore or ROUGE may still be semantically correct, indicating its answers are contextually appropriate but not exact string matches. This is especially important in VQA, where multiple valid phrasings can exist for a single answer.

6. Important Links and Documents

- Dataset Source: https://amazon-berkeley-objects.s3.amazonaws.com/index.html
- Project Code Repository: Github Link