Visual Question Answering on the Indian Heritage in Digital Space Dataset Using the BLIP Model

Aryan S Phadnis[1], Vibhu V Revadi[1], Abhishek BR[1], Vinayak Neginhal[1], Dr Uday Kulkarni[1] and Shashank Hegde[1]

[1]*School of Computer Science and Engineering, KLE Technological University, Hubballi, India*
{*01fe22bcs126, 01fe22bcs010, 01fe22bcs282, 01fe22bcs199, uday_kulkarni, shashank.hegde*}*@kletech.ac.in*

Keywords: Visual Question Answering, Multimodal Framework, Bootstrapping Language-Image Pre-training model, Vision and Language Modalities, Fine-Tuning, Task-specific Adaptations, IHDS Dataset, Accuracy, F1 Score

Abstract: Visual Question Answering is a rapidly evolving domain in the field of artificial intelligence, which combines computer vision and natural language processing to understand and answer textual questions based on image content. Our approach involves the fine-tuning of the Bootstrapping Language-Image Pre-training model, a multimodal framework to address the problems between vision and language modalities. By using a pre-trained architecture, we can optimize the model for real-world applications through some task-specific adaptations. Our work highlights how such a model can address the practical challenges in Visual Question Answering tasks thus improving the alignment between the visual and textual modalities. Experimental results on the test dataset created using unseen images and questions from the IHDS dataset show an accuracy of 86.42% and a weighted F1 score of 0.89, showing the effectiveness of our approach in enhancing VQA systems for any diverse and complex dataset. The integration of domain-specific datasets highlights the versatility of using fine-tuned models for addressing distinct challenges while also maintaining robust performance. Our proposed methodology demonstrates adaptability to a domain and also establishes a foundation for applying multimodal frameworks to culturally rich datasets.

1 INTRODUCTION

Visual Question Answering (VQA)(Antol et al., 2015) is a relatively new field in artificial intelligence which combines the concepts of computer vision and natural language processing to develop systems which can answer questions based on an image. This task requires the integration of visual perception, contextual language modeling, and multimodal reasoning(Vaswani et al., 2023). Despite the advancements in the artificial intelligence field, achieving reliable and generalized performance in VQA tasks remains complicated due to diversity of images and the complexity of the human language(Li et al., 2024). Recent developments in multimodal architectures, such as Bootstrapping Language-Image Pre-training (BLIP) framework (Li et al., 2022), have shown promising progress in addressing these problems. BLIP leverages large-scale pre-training(Piergiovanni et al., 2022) on diverse vision-and-language datasets to learn strong representation that accurately display the relationships between visual and textual modalities(Jia et al., 2021). This foundation provides a strong and flexible backbone for a number of tasks, including VQA. The development of VQA systems tailored to datasets based on cultural heritage sites could potentially change the way people interact with and learn about history. Such a system can be used as an interactive tool for tourism by enabling tourists to ask questions about landmarks using the images of the landmark and receiving accurate responses. They can also be used for educational purposes by providing students and researchers with a deeper understanding of historical and cultural contexts. The integration of such cultural heritage datasets into research based on VQA preserves historical narratives while also fostering a deeper understanding and appreciation of cultural heritage.

The proposed methodology as discussed in this paper utilizes a fine-tuning approach(Lv et al., 2024) for the BLIP model. BLIP has shown remarkable performance in aligning the visual and textual modalities, leveraging its pre-trained to excel across a variety of applications. Our work as depicted in Figure 1, focuses on customizing BLIP to address a specific VQA dataset that is the Indian Heritage in Digital Space

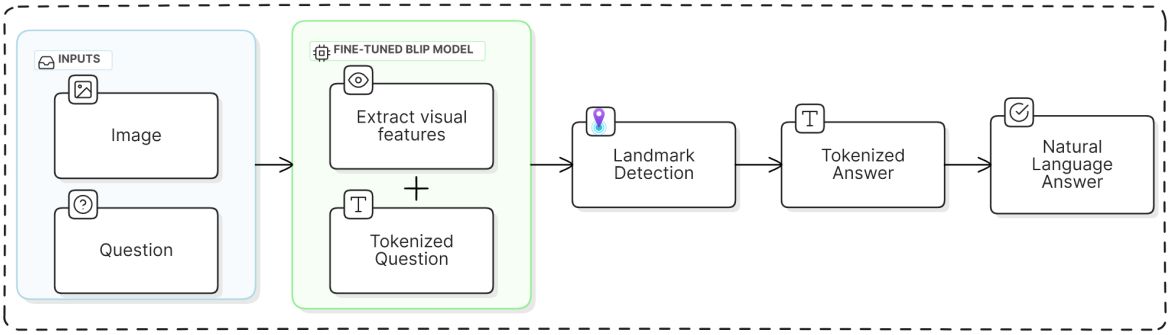

Figure 1: Workflow of proposed methodology.

(IHDS) dataset which includes various cultural heritage sites across India. The research work leveraging this dataset is supported under the Indian Heritage in Digital Space (DST/ICPS/IHDS/2018) initiative, a part of the Interdisciplinary Cyber Physical Systems Programme funded by the Department of Science and Technology, Government of India.The model is fine-tuned to answer domain-specific questions, such as identifying the landmarks, the locations or historical questions regarding the structure. Using this approach, we aim to connect the pre-training objectives to task-specific requirements, improving its ability to generate accurate responses based on the context(Coquenet et al., 2023). The proposed methodology focuses on adapting BLIP so that it optimizes performance on the IHDS Dataset, thus making sure that the relevance to the domain is maintained. The outputs obtained are evaluated across various question categories present in the dataset, showing that the model is capable of handling tasks such as object identification and reasoning related to Indian heritage. The proposed methodology is described in detail in Section 3 along with the challenges faced. Section 4 shows the results achieved by the model. The conclusion in Section 5 presents an analysis of the results and identifies future research directions for enhancing VQA systems in specialized domains.

2 BACKGROUND

Visual Question Answering has emerged as a pivotal area of research in artificial intelligence, integrating advancements in computer vision and natural language processing to enable systems to answer questions based on the content of an image. The early methods relied on human-made features and traditional machine learning models which struggled to capture the complex interactions between visual and language representations(Nguyen and Okatani, 2018). The advent of deep learning techniques, particularly Convolutional Neural Networks (CNNs)(O'Shea and Nash, 2015) led to significant progress by enabling automated feature extraction and improved visual representation. The introduction of transformers(Ondeng et al., 2023) has revolutionized the field by providing robust mechanisms for cross-modal understanding. Transformer-based models such as Generative Image-to-Text Transformer (GIT)(Wang et al., 2022) and Visual-Language Commonsense Bidirectional Encoder Representations from Transformers (VLC-BERT)(Ravi et al., 2022) have leveraged attention mechanisms and pre-training strategies to address challenges in tasks like image captioning(Salaberria et al., 2023) and VQA.

The overall workflow of our proposed methodology, as shown in Figure 1, illustrates how the input data is processed to obtain the resultant output. Among recent innovations, the BLIP model stands out for its use of the Multimodal Mixture of Encoder-Decoder (MED) framework (Li et al., 2022). BLIP effectively bridges the gap between vision and language by employing self-supervised pre-training strategies on large-scale datasets, making it a versatile foundation for various vision-language tasks.

The BLIP model is a state-of-the-art multimodal framework which integrates understanding of vision and language to perform tasks like Visual Question Answering (VQA). Figure 2 depicts the overall BLIP architecture which uses a Vision Transformer (ViT)(Ruan et al., 2022) as the image encoder, thus dividing the input images into different patches and then encoding these patches into a embedding sequences along with an additional [CLS] token which represents the global image feature. In comparison to other traditional object detector based methods, this approach is more computation-efficient making it suitable for large-scale tasks. This architecture integrates visual information in the form of text encoding

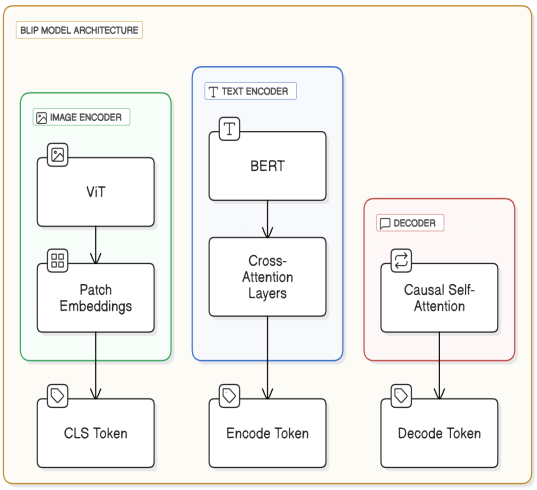

Figure 2: Bootstrapping Language-Image Pre-training Model Architecture.

by inserting a Cross-Attention (CA) layer, whereas the decoder uses a causal attention mechanism thus enabling image-grounded text generation. General-purpose datasets like COCO lack the specificity that is required for domain-specific applications even though there have been several advancements. This study fine-tunes the BLIP model on the IHDS dataset to address limitations present in the existing VQA models for specific contexts like landmarks(Ramaswamy et al., 2023). The image-question pairs were tokenized by the `BlipProcessor` and the ground-truth answers were encoded for training. A dynamic beam search strategy (Huang et al., 2023) was implemented during the inference phase in order to improve accuracy of responses given by the model. BLIP efficiently captures both visual as well as textual modalities by leveraging its robust architecture and pre-training on large-scale datasets. The fine tuning process is used so that the model adapts to the IHDS dataset, thus providing accurate and context-aware answers for the questions related to the landmarks.

3 PROPOSED METHODOLOGY

We propose a fine-tuned BLIP model to answer questions based on the various Indian heritage sites that are part of the IHDS dataset. Our model is adjusted to fit the dataset's needs by utilizing BLIP's pre-trained capabilities. The steps involved in this methodology are creating a structured dataset, fine-tuning the pre-trained BLIP model, and incorporating a structured training and testing approach so that the model performs efficiently. The proposed method involves designing the model in such a way that it understands the differences associated with the cultural landmarks and their textual descriptions effectively. The process is divided into several stages, namely, dataset preparation, model fine-tuning, training and evaluation, and results and inference.

3.1 Introduction to Proposed Methodology

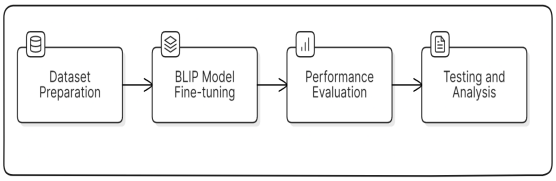

Figure 3: Steps involved in proposed methodology.

The overall flow of the steps involved is shown in Figure 3. A custom dataset of images and pairs of questions-answers was created. The dataset includes diverse landmarks of cultural heritage in India and is used for in the training and evaluation phases. The BLIP model, pre-trained on large-scale vision-and-language datasets, was fine-tuned to meet the requirements for the IHDS dataset. This includes optimizing the model to generate accurate answers for domain-specific questions. A well structured training configuration was used to fine-tune the BLIP model, using appropriate optimization techniques, dropout regularization(Salehin and Kang, 2023), and learning rate scheduling. To evaluate performance of the fine-tuned model, the custom dataset was used to get accuracy-based metrics, which included the F1 Score(Manning et al., 2008) and exact match accuracy(Risch et al., 2021). In the testing phase, unseen image-question pairs were used to generate predictions which were then compared with the actual answers. The results were saved in a CSV file for further analysis and interpretation. This methodology makes use of the BLIP model effectively for answering questions about the Indian heritage sites by leveraging its pre-trained capabilities while addressing the specific challenges faced in this domain.

3.2 Dataset Preparation

The IHDS dataset was modified according to the requirements of the challenge to evaluate the performance of the BLIP model's ability for visual question answering based on the Indian heritage sites. The

dataset contains multiple images along with natural language questions and answers. Every landmark has it's own directory containing the images and the metadata. The dataset is organized such that each directory corresponds to a unique landmark and each of the directories contain high-resolution images in standard formats and a `data.json` file, as depicted with an example in Figure 4.The `data.json` file consists of structured information which includes the questions associated with each image in the directory and the corresponding answers for each question.

```
{
    "question": "What is this structure?",
    "answer": "Stone Chariot at Hampi"
}
```

Figure 4: A sample from the dataset containing an image of the Hampi Chariot with the `json` file containing a question-answer pair for training.

The dataset was preprocessed in various steps to prepare it for training and evaluation. First, in order to parse the data, the `LandmarkVQADataset` class was implemented to traverse the directory of each location in the dataset, parse the `data.json` files, and pair each image with the `data.json` file which contains the questions and answers for that landmark. Next, during the tokenization process, the `BlipProcessor`(Li et al., 2022) from Hugging Face Transformers was used to tokenize the natural language questions and convert them into input tensors which can be used by the BLIP model. By structuring the dataset in this way and following these steps for preprocessing the data, we can optimize the fine-tuning of the BLIP model. This preparation made sure that the model's requirements are compatible while it also facilitated efficient training and evaluation.

3.2.1 Optimization and Learning Rate Scheduling

The `AdamW` optimizer(Loshchilov and Hutter, 2019) was used for training the model and an exponential learning rate was used to improve convergence by gradually reducing the learning rate by using a weighted decay after each epoch. The learning stability and efficiency could be balanced by using this combination of optimizer and scheduler. The training process optimized the following cross-entropy loss function(Mao et al., 2023):

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log(p_i), \tag{1}$$

The optimization process follows the update rule:

$$\theta_{t+1} = \theta_t - \eta\nabla_\theta L, \tag{2}$$

where θ_t represents the model parameters at time t, η is the learning rate, and $\nabla_\theta L$ is the gradient of the loss function with respect to the parameters.

For learning rate scheduling, a cosine annealing strategy using the equation 3 was implemented to smooth the learning rate decay. The learning rate at time t is given by:

$$\eta_t = \eta_{\max} \cdot \frac{1}{2}\left(1 + \cos\left(\frac{t}{T}\pi\right)\right), \tag{3}$$

where $\eta_{\max}$ is the maximum learning rate, t is the current step, and T is the total number of steps. This ensures a gradual reduction in learning rate, promoting better convergence.

3.2.2 Batch Size and Gradient Accumulation

The batch size was set to 2 during the training and validation phase. To simulate the effects of a larger batch size and to stabilize the training without exceeding the GPU memory limitation, a gradient accumulation(Aburass and Dorgham, 2023) was employed of step size 4. A larger effective batch size per optimization step could be processed due to this approach, ensuring a balance between computational efficiency and model performance.

3.2.3 Regularization and Early Stopping

In order to prevent overfitting, a regularization method was needed, so a dropout regularization applied to the model's layers during training. A patience value was also used to implement early stopping, which monitored the validation loss so that training can be terminated if no significant improvement was observed. Thus the computation resources are used efficiently and overfitting is also avoided.

3.3 Inference Process

The inference process involves generating predictions for unseen image-question pairs. These predictions

are then compared with the ground truth answers for the purpose of evaluation. For each test sample, the image and its corresponding question are first preprocessed using the `BlipProcessor`. The natural language question is tokenized and the image is converted to a format that is compatible with the BLIP model. The predictions are generated by passing the preprocessed inputs through the fine-tuned BLIP model. The outputs provided by the model are decoded using the `BlipProcessor` which convert the tokenized predictions into natural language text. The generated predictions are saved to a `.csv` file along with the corresponding image IDs and questions for further evaluation.

To ensure accurate predictions for unseen questions, the beam search strategy is adapted, allowing the model to compare multiple possible answers and then choose the most likely one. The fine-tuning process was designed such that the model's outputs align with the domain-specific IHDS dataset, thus improving its ability to answer the questions related to the cultural landmarks. Rigorous preprocessing ensured that the input pair of questions and images were formatted consistently so that any errors during inference were reduced.This process allows the fine-tuned BLIP model to answer the questions accurately with relevant context even on a diverse range of unseen questions about Indian heritage sites.

3.4 Challenges and Limitations

The IHDS dataset consisted only of directories containing images for each heritage site without any questions or answers associated with the landmarks, which are necessary for VQA tasks. Manual annotation was required for creating the question-answer pairs for each landmark in the dataset. This process ensured thorough annotations which aligned with the domain-specific requirements of the dataset. The batch size could not be set higher than 4 due to GPU memory limitations. However, training with batch size set to 4 resulted in sharp drops in the F1 Score and the accuracy, thus it required further reduction of the batch size to 2.

The limited dataset diversity constrained the model's ability to generalize across different landmarks and question. Even thought the model performed well on distinct landmarks, it struggled with complex queries. In order to improve the performance augmentation techniques(Ma et al., 2024) which include random flips and color jittering were applied which improved the robustness of the model's responses. Future improvements could involve expanding the dataset to increase diversity and refining the hyperparameters in order to enhance the contextual understanding as well as the overall accuracy.

4 RESULTS AND ANALAYSIS

This section presents the performance analysis of the VQA model fine-tuned on the IHDS dataset, which includes the dataset description, model training overview and the evaluation metrics.

4.1 Dataset Description

The IHDS dataset, contains images of notable Indian monuments such as the Agoreshwar Temple, Aihole Temple, Chandramouleshwar Temple, and the Hampi Chariot, which highlight India's rich architectural and cultural heritage. This dataset comprises 133,481 images, categorized into 50 distinct classes. However, due to computational constraints and the need for manual annotation of question-answer pairs for each class, we sampled 60 images per class for the training and evaluation processes, resulting in a curated subset of 3,000 images. This dataset was adapted and augmented in order to fit the requirements of the VQA task by adding image-question pairs specific to landmark identification and contextual queries. The questions span categories such as identification of the structure, historical context, and location, thus including diverse queries.

4.2 Model Training Overview

The BLIP model was fine-tuned using the IHDS dataset, where the answers were then tokenized with a padding of maximum length of 8 tokens which would be the labels used for supervised learning. The data set was then divided into training and validation sets with a 90:10 ratio, resulting in the training set containing 19,954 samples and the validation set containing 2,218 samples. A testing dataset with 250 testing samples was also created with images and questions that were not used during the training phase. For the training configuration, the batch size was set to 2, a learning rate of 2×10^{-5} with a weight decay of 0.01 and the AdamW optimizer was used. The training was carried out for a total of 20 epochs with gradient clipping by setting a maximum norm of 1.0 to prevent exploding gradients. If there was no improvement after 10 consecutive epochs then an early stopping was triggered. To improve the quality of the predicted answers, a beam search strategy is employed during this step. For questions that are shorter than 10 words, the beam size is set to 3. Otherwise, for longer questions

the beam size is increased to 5 for better contextual understanding.

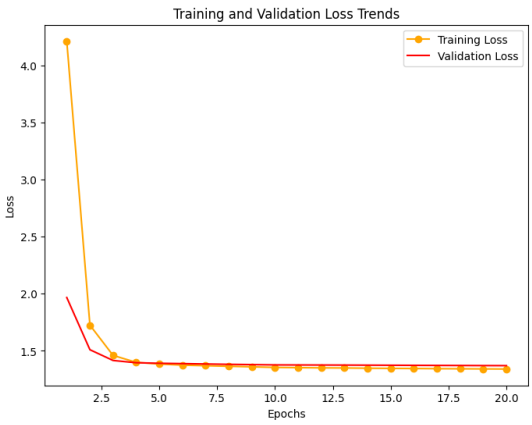

Figure 5: Training and Validation Loss Trends.

Figure 5 shows the trends of training and validation loss over 20 epochs. The consistent decline in both training and validation losses indicates effective learning by the model and convergence of the fine-tuning process. The two loss curves being close indicates minimal risk of overfitting, and shows that the model generalizes well to unseen data. Overall, the graph reflects stable convergence of the model.

4.3 Evaluation Metrics

The model's performance was assessed using two metrics namely, Exact Match Accuracy and F1 Score. Exact Match Accuracy measures the percentage of correctly predicted answers and the F1 score takes a weighted average that accounts for precision and recall. During the training phase, the model achieved an Exact Match Accuracy of 94.1% and an F1 Score of 0.939 on the validation set, indicating strong generalization during fine-tuning. When evaluated on the test dataset, which comprised of unseen locations and images, the model achieved an accuracy of 86.42% and a weighted F1 Score of 0.89. The performance drop on the test set highlights challenges in generalizing to unseen landmarks. Figure 6 provides example comparisons between the answers generated by the fine-tuned model and the correct answers from the test dataset. The examples illustrate the ability of the model to accurately predict answers for distinct landmarks and structured questions, thus validating the quantitative metrics.

Even though the model achieved strong overall performance, the analysis process revealed some frequent failure cases. Some questions that require distinct reasoning (e.g., 'What material was used for constructing this structure?') or contextual knowledge beyond visual information (e.g., 'When was this templ renovated?') often led to incorrect answers. Additionally, landmarks with similar architectural features caused misidentification of the landmarks. These errors highlight the need for training the model on a larger and more diverse dataset with temporal knowledge as well.

Image	Question	Generated Answer	Expected Answer
	What is this structure?	agra fort	agra fort
	What material was used to make this structure?	single block of granite	single block of granite
	Where is this located?	aihole, karnataka	aihole, karnataka

Figure 6: Result comparison.

5 CONCLUSION

In this research, we presented a VQA system tailored for Indian cultural heritage, leveraging the BLIP model to bridge image understanding and natural language processing. The application of different techniques like Exponential Moving Average (EMA), gradient clipping, and learning rate scheduling demonstrated significant improvements in the model's accuracy and stability. Experimental results validated the system's capability to understand and respond effectively to both visual and textual inputs, making it a promising tool for educational and cultural exploration. While the project achieved notable success, limitations such as dataset size, monolingual focus, and synonym handling were identified. Addressing these challenges in future work will involve expanding the dataset, incorporating multilingual capabilities, and integrating external knowledge sources to enhance comprehension and versatility. The proposed system opens avenues for real-world applications, especially on mobile devices and websites, enabling users to engage with cultural heritage interactively. With further optimization, the system could significantly contribute to the accessibility and appreciation of Indian landmarks and artifacts globally. Additionally, such a VQA model deployed as a public API would help interested developers to create immersive experiences, fostering global awareness of India's cultural heritage.

REFERENCES

Aburass, S. and Dorgham, O. (2023). Performance evaluation of swin vision transformer model using gradient accumulation optimization technique.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). Vqa: Visual question answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433.

Coquenet, D., Rambour, C., Dalsasso, E., and Thome, N. (2023). Leveraging vision-language foundation models for fine-grained downstream tasks.

Huang, T., Qasemi, E., Li, B., Wang, H., Brahman, F., Chen, M., and Chaturvedi, S. (2023). Affective and dynamic beam search for story generation.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., and Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision.

Li, J., Li, D., Xiong, C., and Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.

Li, P., Yang, Q., Geng, X., Zhou, W., Ding, Z., and Nian, Y. (2024). Exploring diverse methods in visual question answering. In *2024 5th International Conference on Electronic Communication and Artificial Intelligence (ICECAI)*, page 681–685. IEEE.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Lv, K., Yang, Y., Liu, T., Gao, Q., Guo, Q., and Qiu, X. (2024). Full parameter fine-tuning for large language models with limited resources.

Ma, J., Wang, P., Kong, D., Wang, Z., Liu, J., Pei, H., and Zhao, J. (2024). Robust visual question answering: Datasets, methods, and future challenges.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Mao, A., Mohri, M., and Zhong, Y. (2023). Cross-entropy loss functions: Theoretical analysis and applications.

Nguyen, D.-K. and Okatani, T. (2018). Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6087–6096.

Ondeng, O., Ouma, H., and Akuon, P. (2023). A review of transformer-based approaches for image captioning. *Applied Sciences*, 13(19).

O'Shea, K. and Nash, R. (2015). An introduction to convolutional neural networks. *CoRR*, abs/1511.08458.

Piergiovanni, A., Kuo, W., and Angelova, A. (2022). Pre-training image-language transformers for open-vocabulary tasks.

Ramaswamy, V. V., Lin, S. Y., Zhao, D., Adcock, A. B., van der Maaten, L., Ghadiyaram, D., and Russakovsky, O. (2023). Geode: a geographically diverse evaluation dataset for object recognition.

Ravi, S., Chinchure, A., Sigal, L., Liao, R., and Shwartz, V. (2022). Vlc-bert: Visual question answering with contextualized commonsense knowledge.

Risch, J., Möller, T., Gutsch, J., and Pietsch, M. (2021). Semantic answer similarity for evaluating question answering models.

Ruan, B.-K., Shuai, H.-H., and Cheng, W.-H. (2022). Vision transformers: State of the art and research challenges.

Salaberria, A., Azkune, G., Lopez de Lacalle, O., Soroa, A., and Agirre, E. (2023). Image captioning for effective use of language models in knowledge-based visual question answering. *Expert Systems with Applications*, 212:118669.

Salehin, I. and Kang, D.-K. (2023). A review on dropout regularization approaches for deep neural networks within the scholarly domain. *Electronics*, 12(14).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention is all you need.

Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., and Wang, L. (2022). Git: A generative image-to-text transformer for vision and language.