Zero-Shot Image Classification Using CLIP For Interactive Kiosk Assistance

Samyukta K[1], Shruthiga K [2], Sowndarya S[3]

[1,2,3]*Sri Krishna College of Engineering and Technology, Affiliated to Anna University*

***Abstract*—Interactive kiosks have become essential in public spaces such as malls, museums, zoos, sanctuaries, and educational centers providing visitors with access to required information [1]. Most modern kiosks are equipped with touchscreen interfaces, QR code scanning [2], and voice recognition to improve user engagement. This paper presents the design of an image-classifying kiosk, where visitors can interact by means of their mobile phones to identify animals and converse with the kiosk to gain more knowledge. For image classification, traditional methods use Convolutional Neural Networks (CNN) models such as ResNet, EfficientNet and VGG uses hierarchical features from images to classify them into predefined categories [3] [4] [5]. But one major limitation with these models is that they require labeled datasets which are descriptive and require extensive pre-training, which are expensive and time-consuming [6] . In recent years, zero-shot learning has emerged as a groundbreaking approach in image classification, enabling models to recognize objects they have never encountered during training [7]. Our proposed system uses attribute-based zero-shot capabilities of a vision-language model called OpenAI's Contrastive Language-Image Pretraining (CLIP) on Animals with Attributes 2 (AWA2) dataset [8]. It effectively identifies both seen and unseen categories without explicit training on every possible class. It evaluates this by maximizing the cosine similarity between the correct pair of image and class-attribute embeddings [9]. The kiosk also utilizes a chat feature powered by LLaMA 3.3 70B Versatile model to dynamically generate and answer context- sensitive questions about the classified animal [10]. Therefore, our proposed application allows the user to experience zoos and sanctuaries not just for curiosity and recreation, but focuses on animal education, interpretation, and research [11].**

***Index Terms*—Zero-shot learning, Contrastive Language- Image Pretraining (CLIP) , ViT-B/32 image encoder, ViT-B/16 image encoder, LLaMA3.3 70B model, Animals with Attributes 2 (AWA2) Dataset, Kiosk Systems.**

I. INTRODUCTION

This project proposes the development of a Streamlit application that utilizes zero-shot image classification system with OpenAI's CLIP model on the AWA2 dataset. The system eliminates the need for manually labelled training data by mapping visual embeddings to semantic attributes using a binary predicate matrix. The CLIP image encoders which are ViT-B/32 and ViT-B/16, process an input image by dividing it into patches of dimension 32×32 or 16×16 and embeds them with positional encoders. It is passed through the transformer to extract a Classification Token (CLS) which generates the image embedding. The visual embedding is obtained by mapping the image features into the class-attributes embedding. This class-attributes embedding is a binary predicate matrix that is represented by 1 for presence of the feature and 0 for the absence. By employing cosine similarity, the model determines the best match between an image embedding and visual embedding, enabling it to classify animals with high accuracy [8]. This approach reduces data dependency while enhancing the adaptability of the classification model.

To further enrich the system's interpretability and usability, the project integrates the LLM, LLaMA3.3 70B Versatile model, which dynamically generates contextual descriptions and answers attribute-specific queries based on the classified images [10]. By combining CLIP's zero- shot image recognition with LLaMA's language understanding, the system provides a more interactive and informative classification experience. Users can obtain detailed explanations of an image's classification, including its physical traits, habitat, diet, and behavioural attributes [3].

This study also highlights the choice of a vision encoders of CLIP that significantly determines the outcome of the system. The vision transformer ViT

32/B under performs than the Vit 16/B due to increased number of patches. The accuracy between these two vision transformers details out that Vit16/B has a higher accuracy of 93.14 % while ViT 32/B has an accuracy of 90.62% [12]

A. Zero-Shot Learning

Zero-Shot Learning addresses the fundamental challenge in machine learning where models struggle to classify objects that lack labelled training [13]. Traditional classification methods require extensive datasets with predefined labels, making them impractical for recognizing new or rare categories. Zero-shot learning overcomes this limitation by utilizing semantic relationships, enabling models to classify from both seen classes to unseen ones. This is achieved by mapping image embeddings to a shared semantic space, where class attributes—such as textual descriptions and hierarchical taxonomies—help establish meaningful connections [7].

B. AWA2 Dataset

The AWA2 dataset is a benchmark for evaluating zero-shot learning algorithm by providing attributes. It serves a better substitute than Animals with Attributes (AwA) dataset, due to the better number of images than the old version and availability of the images publicly. It consists of 37322 images of 50 animals' classes with pre- extracted feature representations for each image. The classes share 85 different attributes (e.g., "has stripes," "lives in water") . Using the shared attributes, it is possible to transfer information between different classes. The dataset uses a binary predicate matrix that maps the visual features of the image to the semantic attributes of the dataset. The list of 85 attributes of AWA2 dataset are listed below [13]. "[Black, White, Blue, Brown, Gray, Orange, Red, Yellow, Patches, Spots, Stripes, Furry, Hairless, Toughskin, Big, Small, Bulbous, Lean, Flippers, Hands, Hooves, Pads, Paws, Longleg, Longneck, Tail, Chewteeth, Meatteeth, Buckteeth, Strainteeth, Horns, Claws, Tusks, Smelly, Flys, Hops, Swims, Tunnels, Walks, Fast, Slow, Strong, Weak, Muscle, Bipedal, Quadrapedal, Active, Inactive, Nocturnal, Hibernate, Agility, Fish, Meat, Plankton, Vegetation, Insects, Forager, Grazer, Hunter, Scavenger, Skimmer, Stalker, Newworld, Oldworld, Arctic, Coastal, Desert, Bush, Plains, Forest, Fields, Jungle, Mountains, Ocean, Ground, Water, Tree, Cave, Fierce, Timid, Smart, Group, Solitary, Nestspot, Domestic]"

C. Contrastive Language-Image Pretraining (CLIP)

CLIP is an advanced vision-language model developed by OpenAI that learns visual representations from natural language supervision [8]. CLIP is trained on 400 million image-text pairs collected from the internet, allowing it to generalize across diverse classification tasks without requiring task-specific fine-tuning. CLIP learns from text–image pairs that are already publicly available on the internet. Our proposed model uses vision transformers (ViT-B/32 and ViT-B/16) to allow CLIP to capture complex image relationships, where an image is embedded into a shared multimodal space, maximizing the similarity between matching pairs while minimizing the similarity between incorrect ones. This enables CLIP to perform zero- shot classification, predicting image labels by evaluating textual descriptions without having been explicitly trained on specific categories. CLIP has been benchmarked across over 30 computer vision datasets, demonstrating competitive performance in tasks such as object recognition and classification.

D. LLAMA3.3 70B Model

The LLaMA3.3 70B Versatile Model is a large-scale LLM designed for a wide range of natural language processing (NLP) tasks, including text generation, question answering, and contextual understanding [10]. Developed with 70 billion parameters, this model is built to provide highly accurate, context-aware, and coherent responses across various domains. With a context length of 128,000 tokens, Llama 3.3 70B can process extensive textual inputs, facilitating more coherent and contextually relevant outputs. We use LLaMA3.3 70B model to answer the user-prompted questions. The model analyses the question and answers it by using the attributes that are flagged positive for the identified image. The dynamically generated answer is provided as descriptive passage [14].

E. KIOSK Systems

Kiosk systems are interactive, self-service machines designed to provide users with information, services, or entertainment in public or commercial spaces. The kiosk developed here provides an intuitive user-interface for easy user interactions such as image uploading and dynamic question answering [11]

II. LITERATURE REVIEW

A. Zero-shot Learning

Zero-shot learning has gained significant attention due to its ability to classify unseen categories without requiring labelled training data. Unlike traditional models that depend on extensive annotated datasets, it leverages semantic attributes and textual descriptions. OpenAI's CLIP model [8] is a widely adopted framework for this task, aligning images and text in a shared embedding space to enable flexible and scalable classification. Research by [13] introduced benchmark datasets like AWA2, which provide attribute-based annotations for animal classification, further standardising zero-shot evaluation. Other contribution, introduced DeViSE

[15] an early deep learning approach to zero-shot learning that aligned visual and semantic embeddings.

B. Attribute Embedding for Classification

Attribute embeddings enable zero-shot learning by bridging the gap between seen and unseen classes. These embeddings represent semantic properties such as fur colour, diet, and habitat, which can describe multiple species and facilitate classification by capturing meaningful relationships between different categories. [16] introduced attribute-based recognition, where objects are identified based on semantic descriptors rather than explicit class labels, enabling more flexible and interpretable classification. [13] extended this approach by applying attribute-based classification to animal species, leading to the development of the AWA2 dataset, which is widely used for zero-shot learning research. . [17] explored attribute-guided generation models for improving zero-shot classification.

C. CLIP-based Zero-Shot Learning

CLIP-based zero-shot learning systems have improved by refining feature alignment and classification accuracy. [18] introduced CLIPPR (CLIP with Priors), a method that adapts CLIP for regression and classification tasks on unlabelled datasets. Their approach involves training an adapter network on top of CLIP, guided by a prior over the label distribution, achieving notable improvements in tasks like age regression and image classification without using labelled data. [10] developed RA- CLIP, an enhanced version of the original CLIP model that improves representation learning, leading to superior performance in various tasks. [19] explored extending CLIP to generate 3D shapes from text, demonstrating the model's adaptability to differentmodalities.

D. Cosine Similarity for Embedding Comparison

Cosine similarity is a metric for comparing high-dimensional embeddings. By measuring the angular distance between vectors, cosine similarity provides an effective way to assess the relationship between embeddings in a continuous space. [20] introduced cosine similarity in the context of word embeddings with Word2Vec, demonstrating its ability to capture meaningful relationships between words. [21] extended this concept to GloVe embeddings, further improving performance in natural language processing tasks by leveraging cosine similarity for vector comparisons. More recently, [22] explored the use of cosine similarity in biomedical and public health research, applying it to embedding-based classification tasks to improve automated categorization. This study adopts a similar methodology by computing cosine similarity between image embeddings and class embeddings, ensuring the selection of the most relevant class based on their alignment in the embedding space.

E. LLM

LLMs have expanded the scope of zero-shot classification by improving contextual understanding and question-answering capabilities.

[23] explored the use of LLMs in public health tasks, particularly in classifying and extracting key information from free-text data. This research underscores the potential of LLMs in improving decision-making and reducing manual effort in public health sectors. Other contributions to LLM research include [24] who introduced dynamic strategy planning for efficient LLM-based question answering.

III. PROPOSED SYSTEM

This project proposes a system with zero- shot image classification by integrating AI model to classify images without relying on extensive labelled datasets. The study utilizes OpenAI's CLIP model for vision-language alignment and the LLaMA3.3 70B model for contextual text generation. In our proposed system, we incorporate all 50 classes of the AWA2 dataset to enable a comprehensive zero-shot classification approach. The classification process involves generating visual embedding from classes and attributes, extracting image embeddings from uploaded image, computing cosine similarity, and mapping visual features. The model supports two vision backbones of CLIP, which are ViT-B/32 and ViT-B/16. The libraries used in this project include pandas for data handling, NumPy for numerical operations (cosine similarity), OpenAI's CLIP for image-text embeddings, transformers for model processing, and Streamlit for real-time deployment. The system flow is shown in the Fig. 1.

A. Image Input

The system is hosted on Streamlit, providing an intuitive and interactive interface . The user can upload an animal image in JPG format by scanning the QR code displayed in the user interface and clicking the Upload button. Meanwhile, in the KIOSK system, the Fetch button needs to be clicked to retrieve the uploaded image for processing. Unlike traditional image classification systems, no explicit image pre-processing (such as resizing, normalization, or augmentation) is performed reducing computational overhead [1]

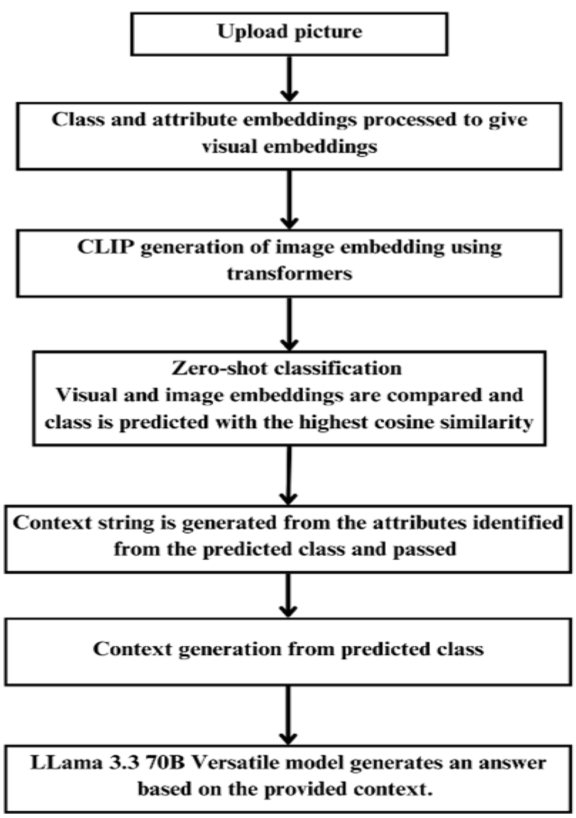


Fig. 1 Workflow of the proposed system

B. Feature Extraction using CLIP

Once the image is received, it is processed using OpenAI's CLIP model, which is designed to process both images, and attributes and classes [8]. This process can be divided into two steps.

1) Visual Embedding

In the AWA2 dataset, each animal class is described using an 85-dimensional attribute vector, where each attribute is represented as a binary value (1 for presence, 0 for absence) [25]. To embed these attributes into a continuous space, a binary predicate matrix (50 × 85) is constructed, where each row corresponds to a specific class and each column represents a distinct attribute. Attributes are mapped to a binary predicate matrix (e.g., "has fur" = 1, "can fly" = 0).These attributes are then converted into smaller feature vectors known as visual embeddings [7]. The Fig. 1 shows the binary predicate matrix for five classes which are antelope , grizzly bear , killer whale , beaver and dalmatian. comparing the image embedding with visual embeddings. This is achieved using cosine similarity, a mathematical measure that determines how closely the vectors are aligned [6]. The cosine similarity formula is stated in (1).

2) Image Embedding

$$cosine\ similarity = \frac{A.B}{||A||.||B||} \qquad .Eq\ 1$$

In the AWA2 dataset, an input image is first processed using the CLIP's Vision Transformers ViT-B/32 or ViT- B/16. The encoder transforms the image into a structured numerical representation by dividing it into patches, applying positional encodings, performing token extraction, and normalizing feature representations, known as image embeddings.

```
Loaded 50 classes and 85 attributes.
Sample Classes: ['antelope', 'grizzly+bear', 'killer+whale', 'beaver', 'dalmatian
Sample Attributes: ['black', 'white', 'blue', 'brown', 'gray']
Attribute Matrix Shape: (50, 85)
              black  white  blue  brown  gray  orange  red  yellow  patches  \
antelope          0      0     0      0     0       0    0       0        0
grizzly+bear      1      0     0      1     0       0    0       0        0
killer+whale      1      1     0      0     0       0    0       0        1
beaver            0      0     0      1     0       0    0       0        0
dalmatian         1      1     0      0     0       0    0       0        1

              spots  ...  water  tree  cave  fierce  timid  smart  group  \
antelope          0  ...      0     0     0       0      1      0      1
grizzly+bear      0  ...      0     0     1       1      0      1      0
killer+whale      1  ...      1     0     0       1      0      1      1
beaver            0  ...      1     0     0       0      1      1      1
dalmatian         1  ...      0     0     0       0      1      1      1

              solitary  nestspot  domestic
antelope             0         0         0
grizzly+bear         1         0         0
killer+whale         0         0         0
beaver               1         1         0
dalmatian            1         0         1

[5 rows x 85 columns]
Attribute Matrix Shape: (50, 85)
```

Fig. 2 Binary predicate matrix for five classes with their attributes.

For better analysis in this paper, we have compared two different vision encoders of CLIP: ViT-B/32 and ViT-B/16 . For ViT-B/32 image encoder , the input image is divided into 32×32- pixel patches. For ViT-B/16 image encoder , the input image is divided into 16×16-pixel patches. Then both these patches are then passed through multiple self-attention layers, generating a high- dimensional image embedding that captures key visual features [12].

C. Zero-Shot Classification

To determine the most relevant class, the system performs zero-shot classification by where A represents the image embedding and B represents the visual embedding of a class. CLIP uses cosine similarity to measure how closely the vectors are related in the embedding space. Instead of comparing raw vector values (which can have different magnitudes), cosine similarity measures the angle between vectors. This ensures that embeddings with the same meaning but different magnitudes still have high similarity. The class with the highest similarity score is selected as the predicted label.

D. Context Generation

After identifying the predicted class, the system maps it to a binary predicate matrix, which assigns semantic attributes such as fur colour, habitat, and diet [25]. These attributes are then compiled into a context string, which serves as structured knowledge about the classified object. The context string is passed to the to LLaMA3.3 70B model, a state-of-the-art LLM, which provides a descriptive summary of the classified object [26].

E. Interactive Question-Answering using LLaMA3.3 70B

The LLaMA 3.3 70B system dynamically answers questions about the identified animal based on attributes. The generated content is structured in a JSON format for easier integration into applications, ensuring that attributes and class details are properly formatted [26]

F. Kiosk-Based Display System

The final content is displayed in the Streamlit UI, where users can read the generated context, ask questions, see the top -5 predictions by the model and interact with the system.

This feature makes the system applicable to education, retail, and automated information kiosks, where real-time AI responses can improve user interaction and accessibility. The kiosk displays the most potential use of zero-shot image classification with CLIP in real-world applications [1] [2].

IV. RESULT ANALYSIS

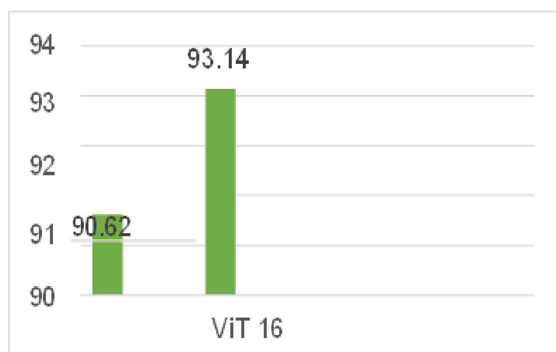
Top-1 accuracy is a key evaluation metric in this system as it measures the frequency with which the model's most confident prediction matches the correct class label. A high Top-1 accuracy indicates that the model can successfully differentiate between visual categories based on attribute embeddings. The top-1 accuracy formula used in calculating the accuracy is stated below

$$top - 1\ accuracy = \frac{Number\ of\ classified\ images}{Total\ number\ of\ images} \times 100 \quad \text{\_ Eq 2}$$

Table I. Comparison Table between ViT-B/32 and ViT- B/16

CLIP Image Transformers	*Correct Predictions*	*Top-1 Accuracy (in %)*
ViT-B/32	33820 / 37322	90.62
ViT-B/16	34760 / 37322	93.14

Fig. 3 Comparison graph between ViT-B/32 and ViT-B/16 image transformer accuracy



From table (I), the results indicate that the CLIP model's performance varies based on the selected Vision Transformer (ViT) backbone. The ViT-B/16 model achieves a higher top-1 accuracy of 93.14%, compared to ViT-B/32, which records 90.62% accuracy. This improvement is attributed to ViT-B/16's finer patch size (16×16-pixel patches), which allows for better feature extraction and more precise classification. The higher resolution embeddings in ViT-B/16 contribute to a more detailed representation of image features, confirming its superior feature extraction capability and higher precision in zero-shot classification tasks

These results emphasize the importance of choosing a more granular vision backbone, making ViT-B/16 the preferred choice for this project.

V. CONCLUSION

This study demonstrates the effectiveness of zero-shot learning in image classification by integrating OpenAI's CLIP model and LLaMA3.3 70B for enhanced vision-language understanding. By leveraging cosine similarity and attribute-based classification, the system successfully classifies both seen and unseen objects without requiring extensive labelled datasets. The integration of a real-time interactive question-answering system further enhances user engagement, making AI-driven image classification more accessible and informative.
The results highlight the impact of Vision Transformer (ViT) architecture on classification accuracy, with ViT-B/16 outperforming ViT-B/32 by achieving a higher accuracy of 93.14. The comparison underscores the importance of transformer selection in optimizing zero-shot learning models.
Finally, the deployment of this system in an interactive kiosk format extends its usability to various real-world applications, such as education. The ability to dynamically classify images and generate context-aware responses makes the system a practical solution for AI-powered information retrieval, demonstrating the potential of zero-shot learning in real-world scenarios.

REFERENCES

[1] H. Alves, "Interactive kiosk systems: A modern approach to visitor engagement," Interactive Systems, pp. 156-168, 2022.
[2] "The Evolution of Kiosk Technology," Kiosk Marketplace Report, 2023.
[3] X. Z. S. R. a. J. S. K. He, "Deep Residual Learning for Image Recognition," pp. 770-778, 2016.
[4] M. T. a. Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Int. Conf. Machine Learning (ICML), 2019.
[5] K. S. a. A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," 2014.Neural Information Processing Systems (NIPS), 2013. I. E. D. H. a. D. F. A. Farhadi, "Describing objects by their attributes," in A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2009.
[6] J. B. T. T. L. G. T. Y. a. N. U. C. Kemp, "Learning systems of concepts with an infinite relational model," in AAAI Conf. Artificial Intelligence, 2006.
[7] H. N. a. S. H. C. H. Lampert, "Attribute-Based Classification for Zero-Shot Visual Object Categorization," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 3, p. 2014, 453–465. Z. Y. a. T. H. K. Zhang, "Diverse and Tailored Image Generation for Zero-shot Multi-label Classification," in Int. Conf. Computer Vision, 2024. E. S. a. Y. H. Y. Kahana, "CLIPPR: What did CLIP Learn to Reason About Images?" in Advances in Neural Information Processing Systems (NeurIPS), 2022.
[8] A. Radford, "[Learning Transferable Visual Models from Natural Language Supervision," in Machine Learning (ICML), 2021..A. Sanghi, "CLIP-Forge: Towards Zero-Shot Text-to-Shape Generation," in IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2022.
[9] C. E. a. N. K. H. Steck, "Is Cosine- Similarity of Embeddings Really About Similarity?" in ACM Web Conf. (WWW 2024 Companion), 2024.K. C. G. C. a. J. D. T. Mikolov, "Efficient Estimation of Word Representations in Vector Space," in Int. Conf. Learning Representations (ICLR), 2013.
[10] H. Xie, "RA-CLIP: Improving Representation Learning in CLIP- Based Zero-Shot Learning," in IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2023.R. S. a. C. D. M. J. Pennington, "GloVe: Global Vectors for Word Representation," in Empirical Methods in Natural Language Processing (EMNLP), 2014.
[11] W. C. Prakoso, C. BoonchaiThe gap in technology for visitor experience in a zoo: A case study in Indonesia," Zoo Biology, vol. 42, no. 1, p. 67– 74, 2023. W. C. a. R. Harris, "Cosine Similarity in Biomedical and Public Health Research: A Case Study," in Int. Conf. Biomedical Engineering, 2024.
[12] A. Dosovitskiy, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 2020.

[13] Y. Xian, "Zero-Shot Learning - A Comprehensive Evaluation of the Good, the Bad and the Ugly," IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI), vol. 40, no. 8, p. 2251–2265, 2018. H. S. a. S. Lam, "Large Language Models for Healthcare Text Classification: A Systematic Review," 2025. T. Parekh, "Dynamic Strategy Planning for Efficient Question Answering with Large Language Models," in Int. Conf. Natural Language Processing, 2024.
[14] T. Brown, "Language Models are Few-Shot Learners," 2020.
[15] A. Frome, "DeViSE: A deep visual- semantic embedding model," in H. N. a. S. H. C. Lampert, "Learning to Detect Unseen Object Classes by