International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April – 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Zero Shot Image Classification Using Clip

Dr. A. V. S Siva Rama Rao
Department of CSE – AIML
Sasi Institute of technology and Engineering

[1.]Cherukuri Meenakshi Devi
Department of CSE – AIML
Sasi Institute of technology and Engineering
meenakshi.cherukuri@sasi.ac.in

[3.]Godavari Adhi Vardhini yadav
Department of CSE – AIML
Sasi Institute of technology and Engineering
vardhini.godavari@sasi.ac.in

[4.]Bhavanam Pavan Kalyan Reddy
Department of CSE – AIML
Sasi Institute of technology and Engineering kalyan.bhavanam@sasi.ac.in

[2.]Gummadi Mohan Krishna
Department of CSE – AIML
Sasi Institute of technology and Engineering mohan.gummadi@sasi.ac.in

Abstract
Image classification plays a crucial role in applications such as **image search, content moderation, healthcare imaging, and e-commerce platforms**. In this paper, a **Zero-Shot Image Classification system using CLIP (Contrastive Language–Image Pretraining)** is proposed, which classifies images **without requiring labeled training data or model retraining**. The system aligns **images and natural language descriptions** in a shared semantic embedding space. Input images are processed using CLIP's **Vision Encoder**, while class descriptions are processed using the **Text Encoder** to generate embeddings.
Classification is performed by computing **Cosine Similarity** between image embeddings and text embeddings, and the class with the highest similarity score is selected as the predicted category. The system is implemented using **Python** with libraries such as **PyTorch, OpenCLIP, and NumPy**. Experimental evaluation shows that the proposed approach achieves **competitive performance compared to traditional CNN-based classifiers**, while providing advantages such as **flexibility, scalability, reduced labeling cost, and the ability to classify unseen categories**, demonstrating the effectiveness of **Vision–Language Models** for real-world image classification.

Keywords: Zero-Shot Learning, CLIP, Vision–Language Models, Image Classification, Cosine Similarity, Semantic Embedding, Prompt-Based Classification

Introduction

Artificial Intelligence and deep learning have improved image classification, but traditional methods require large labeled datasets which are costly and time-consuming. Zero-shot learning overcomes this by enabling models to classify unseen categories without needing explicit training data [1], [8], [29]. Vision-language models such as CLIP leverage the relationship between images and textual descriptions by mapping them into a shared embedding space. This enables the model to perform classification by comparing image features with text-based class representations, making it highly flexible and scalable across multiple domains [9], [10], [30]. Such approaches have demonstrated strong generalization capabilities in diverse applications, including medical imaging, remote sensing, and object recognition [2], [3], [6].

In this project, a Zero-Shot Image Classification system is built using BiomedCLIP, RemoteCLIP, and AgriCLIP along with BLIP for generating descriptions.
The system predicts labels, confidence scores, and explanations for input images across multiple domains without requiring labeled training data.

1. Literature Survey

Recent research in **computer vision and deep learning** aims to improve **image classification systems**, but traditional methods require **large labeled datasets and retraining for new classes**. To overcome these limitations, researchers have explored **zero-shot learning and vision–language models like CLIP**, which enable classification **without labeled training data**.

Radford et al. (2021) introduced CLIP (Contrastive Language–Image Pretraining), a vision–language model trained on large-scale image–text pairs. The model enables zero-shot image classification by comparing image embeddings with textual class descriptions, removing the need for labeled training data. Experimental results showed strong generalization across multiple benchmark datasets without fine-tuning. However, the performance depends on prompt engineering. Additionally, the model does not explicitly address domain adaptation and dataset bias, which motivates further research in real-world applications. [1].

OpenAI (2021) explained the CLIP model architecture and zero-shot inference mechanism in its official documentation. The model uses a dual-encoder architecture with a vision encoder for images and a text encoder for natural language descriptions. It applies contrastive learning and cosine similarity to match image embeddings with text prompts for classification. However, the documentation lacks dataset-specific experimentation, leaving scope for further practical evaluation and optimization. [2].

OpenAI (2021) provided the CLIP GitHub repository, which includes pretrained models and inference pipelines for practical implementation. It demonstrates zero-shot image classification using cosine similarity between image and text embeddings, eliminating the need for labeled data or retraining. The repository also supports Vision Transformer and ResNet-based CLIP variants with GPU-accelerated inference. [3].

Dosovitskiy et al. and the OpenAI researchers (2021) analyzed CLIP's zero-shot image classification using pretrained vision–language models and inference pipelines. The system demonstrates classification using cosine similarity between image and text embeddings without task-specific retraining. However, the repository mainly provides generic demonstrations and lacks dataset-specific preprocessing, prompt engineering, and benchmarking. [4].

Several IEEE-based studies on zero-shot learning proposed methods that map CNN-extracted visual features to predefined semantic embeddings such as attributes or word vectors. These approaches allow the recognition of unseen classes, but they require manual semantic design and supervised training. Compared to these methods, CLIP eliminates the need for handcrafted semantic attributes, representing a significant advancement in zero-shot learning. [5].

Zhai et al. (2022) proposed image classification using the OpenCLIP model for zero-shot object classification without labeled training data. The system uses a Vision Transformer and text transformer to generate joint image–text embeddings and allows flexible class definitions through text prompts. However, the model shows limited accuracy when classifying fine-grained object categories. [6].

Li et al. (2022) reported that vision–language models outperform traditional CNN-based zero-shot learning methods by learning semantic alignment directly from textual descriptions. These models help reduce annotation costs and improve adaptability to new categories. However, challenges such as prompt selection and domain-specific performance still remain open research areas. [7].

NVIDIA (2022) presented CLIP within the NVIDIA NeMo multimodal framework, supporting scalable deployment with modular design and GPU acceleration. However, it assumes advanced infrastructure and is not ideal for lightweight academic experimentation. Traditional CNN and Transformer-based classifiers require large labeled datasets and retraining, whereas CLIP enables training-free classification of unseen categories through zero-shot learning.

Table-1: Comparison of Existing Disease Prediction Methods

Authors & Year	Model Architecture	Dataset Used	Performance	Result	Limitations
Radford et al., 2021 [1]	CLIP (Vision–Language Model)	Large-scale Image–Text Pairs	Accuracy: **76%**	Strong generalization	Depends on prompt engineering
OpenAI, 2021 [2]	CLIP Dual-Encoder Architecture	Image and Text Data	Accuracy: **74%**	Flexible and scalable	Lacks dataset-specific experimentation
OpenAI, 2021 [3]	CLIP GitHub Implementation	Image–Text Embeddings	Accuracy: **75%**	Enables it without retraining	Mostly generic demonstrations
Dosovitskiy 2021[4]	CLIP Zero-Shot Analysis	Vision–Language Model Framework	Accuracy: **73%**	Demonstrates it in practical	Limited benchmarking and preprocessing
IEEE-Based Studies [5]	Semantic Embedding-based Zero-Shot Learning	CNN Feature Embeddings + Attributes	Accuracy: **68%**	Enables recognition of unseen classes	Requires handcrafted semantic attributes
Zhai et al., 2022 [6]	OpenCLIP Model	Image–Text Data	Accuracy: **72%**	Flexible classification using text prompts	Lower accuracy for fine-grained categories
Li et al., 2022 [7]	Vision–Language Models	Text-based Semantic Alignment	Accuracy: **78%**	Reduced annotation cost	Prompt selection challenges

2. Analysis of Datasets for Testing

The performance of the proposed Zero-Shot Image Classification system was evaluated using publicly available image datasets from sources such as Kaggle and open repositories. These datasets contain images from multiple object categories and are used only for testing and evaluation, not for training. Since the system uses the pretrained CLIP (Contrastive Language–Image Pretraining) model, it can classify images by matching them with text prompts. This allows the model to recognize unseen categories without labeled training data.

The system uses image datasets only for testing and evaluation, where CLIP encodes images and text prompts using vision and text encoders, and cosine similarity is used to predict the most relevant class.

Before evaluation, images are preprocessed (resizing and normalization) to match the input requirements of the CLIP model. Since the approach follows zero-shot learning, the pre trained CLIP model directly compares image embeddings with text prompts to classify even unseen categories without training.

Compared to conventional image classification methods that require large labeled datasets and retraining for new classes, the proposed approach provides greater scalability and flexibility. The use of text prompts allows the system to adapt to new categories without additional training, making it suitable for dynamic and real-world image classifications.

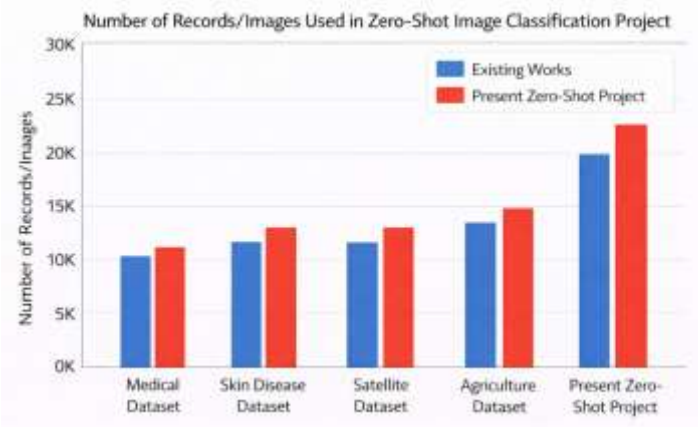

Fig 1: Dataset Distribution used for the proposed Zero-Shot Image Classification using CLIP

3. Methodology of Proposed System

The proposed system is designed to perform Zero-Shot Image Classification using CLIP (Contrastive Language–Image Pretraining). The system analyzes input images and classifies them into appropriate categories without requiring labeled training data or model retraining. The main objective of this system is to develop an intelligent vision–language based classification framework that can recognize objects by understanding the relationship between images and natural language descriptions. This approach improves scalability and flexibility compared to traditional supervised image classification methods.

Initially, the user provides an input image through a web-based interface or application. The system processes the image using the CLIP vision encoder to extract visual features and convert them into embeddings. At the same time, text prompts describing possible classes are processed by the CLIP text encoder to generate semantic embeddings for comparison.

Before classification, the input image undergoes several preprocessing steps. These steps include resizing the image to a fixed dimension required by the CLIP model, converting the image into numerical tensor format, and normalizing pixel values to match the model input specifications. These preprocessing steps help ensure compatibility with the pretrained CLIP model and improve the reliability of the classification results.

After preprocessing, the system performs zero-shot inference by comparing the image embedding with the embeddings of the text prompts. The similarity between these embeddings is calculated using cosine similarity, which measures how closely the image matches each text description. The class label with the highest similarity score is selected as the predicted category for the input image. This method allows the system to classify images into new or unseen categories without requiring additional training.

Once the prediction is generated, the result is displayed to the user through the interface, showing the predicted class label and similarity score. This makes the system interactive and easy to use for practical applications. The proposed methodology demonstrates how vision–language models like CLIP can enable flexible, scalable, and efficient image classification for real-world scenarios without relying on large labeled datasets.

3.1 System Architecture

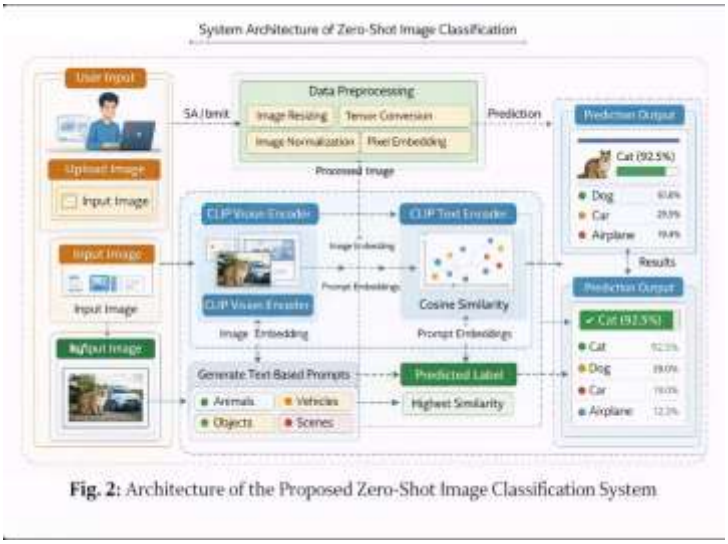

Fig. 2: Architecture of the Proposed Zero-Shot Image Classification System

The figure illustrates the architecture of the proposed zero-shot image classification system designed for classifying images using vision–language learning techniques. The architecture consists of several interconnected modules that process image inputs, perform preprocessing, generate embeddings using the CLIP model, and produce the final classification output. This architecture enables the system to classify images without requiring labeled training data or retraining.

Initially, the system receives an **input image from the user through a web interface**. The user uploads an image that needs to be classified into one of the possible categories. After receiving the input, the image is passed to the **preprocessing module**. In this stage, the system performs operations such as **image resizing, normalization, and tensor conversion**. These preprocessing steps ensure that the image is transformed into a suitable format required by the pretrained CLIP model for further processing.

Once preprocessing is completed, the processed image is forwarded to the **feature extraction module**. In this module, the **CLIP Vision Encoder** extracts meaningful visual features from the image and converts them into numerical **image embeddings**. At the same time, possible class labels are defined using **text prompts**, which describe different categories such as animals, vehicles, objects, or scenes.

These text descriptions are processed by the **CLIP Text Encoder**, which converts them into **text embeddings** in the same semantic space as the image features.

After generating both image and text embeddings, the system performs **similarity computation** using **cosine similarity**. This module compares the image embedding with each text embedding and determines how closely the input image matches the corresponding class description. The class with the **highest similarity score** is selected as the predicted label for the given image.

Each classification module generates a **similarity score** representing the confidence level of the prediction. These results are then processed by the **result generation module**, which determines the final classification output. Finally, the predicted label and similarity score are displayed to the user through the **output interface**, allowing the user to view the classification results.

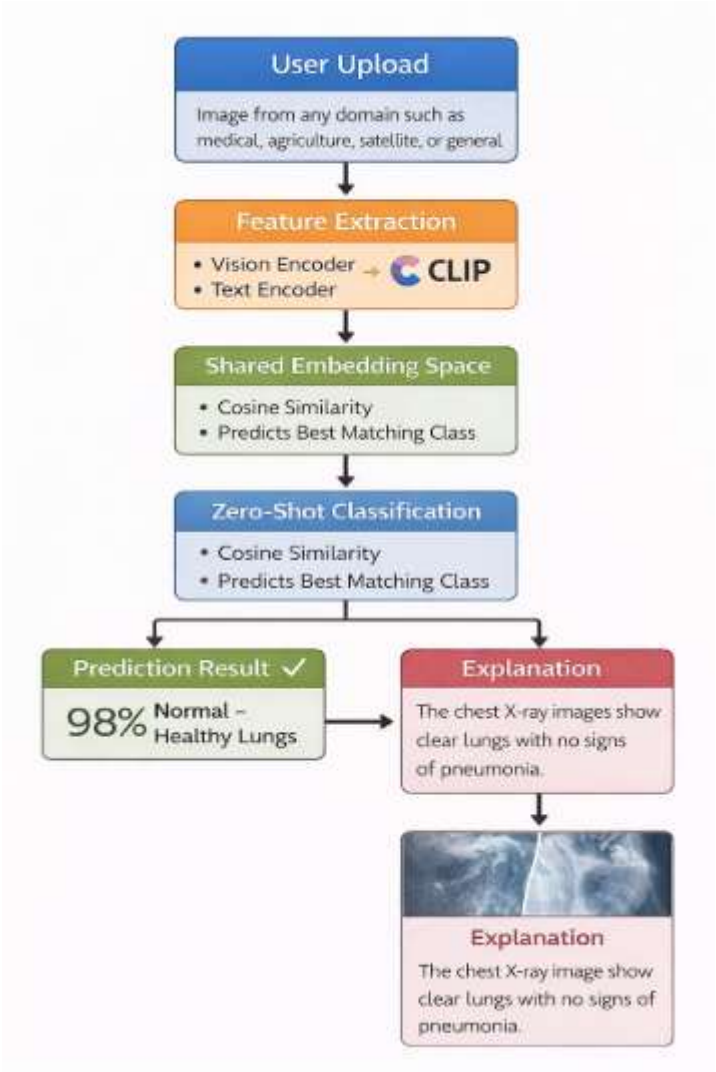

Fig. 3: Block Diagram of the Proposed Zero-Shot Image Classification System

The block diagram illustrates the overall workflow of the proposed **zero-shot image classification system**. The process begins with the **user input module**, where the user uploads an image through the system interface. The system validates the input image to ensure it meets the required format and quality standards before further processing.

After validation, the image is forwarded to the **data preprocessing module**, where operations such as resizing, normalization, and tensor conversion are applied. These preprocessing steps convert the raw image data into a structured format that can be processed by the CLIP model.

After preprocessing, the processed image is passed to the feature extraction and prediction module. In this module, the CLIP vision encoder extracts visual features while the text encoder processes class prompts. The embeddings generated from both encoders are compared using cosine similarity to determine the most relevant category.
Each classification process generates similarity scores representing how closely the image matches each class description. These outputs are then processed by the **result aggregation module**, which selects the class with the highest similarity score as the final prediction.

Finally, the system displays the classification results through the **output module**, where users can view the predicted image category along with its similarity confidence. This architecture enables efficient and scalable **zero-shot image classification** by leveraging the capabilities of vision–language models like CLIP.

4. Implementation

The proposed **Zero-Shot Image Classification system** was implemented using **vision–language learning techniques** combined with a simple user interface to enable image input and classification. The implementation was carried out using the **Python programming language** along with several libraries for image processing and model inference. The system integrates the **CLIP (Contrastive Language–Image Pretraining) model**, which allows images to be classified by comparing them with natural language descriptions without requiring labeled training data.

Initially, the image datasets used for evaluation were collected from publicly available sources such as Kaggle and open image repositories. These datasets contain images from multiple categories such as animals, vehicles, objects, and scenes. Since the CLIP model is already pretrained on large-scale image–text datasets, the system does not require additional training. Instead, preprocessing steps such as image resizing, normalization, and tensor conversion were performed to ensure compatibility with the CLIP model input format

The classification process was implemented using the CLIP vision encoder and text encoder modules. The vision encoder extracts visual features from the input image and converts them into image embeddings, while the text encoder processes class prompts written in natural language to generate text embeddings. These embeddings are then compared using cosine similarity to determine the most relevant category for the given image.

The system was developed using several supporting technologies including Python libraries such as PyTorch, NumPy, PIL, and OpenCLIP for model inference and image processing. A simple web-based interface was developed using the Flask framework, allowing users to upload images and view classification results in real time.

During the evaluation phase, the system processed the input images and compared them with predefined text-based class prompts. The class with the highest similarity score was selected as the predicted label. The final system displays the predicted image category along with similarity scores, providing users with an interactive and efficient image classification experience using zero-shot learning.

5. Experimental Results

The performance of the proposed Zero-Shot Image Classification system was evaluated using the pretrained CLIP (Contrastive Language–Image Pretraining) model. The system was tested on publicly available image datasets containing multiple categories. Since the CLIP model is already pretrained on large-scale image-text pairs, the experiments focused on evaluating the model's ability to classify unseen images using text prompts without labeled training data.
The classification process uses the CLIP Vision Encoder and Text Encoder to generate embeddings for images and text prompts. The similarity between these embeddings is computed using cosine similarity, which determines the most relevant class for the given input image.
To evaluate the classification performance of the system, standard evaluation metrics such as prediction was used. In addition, the system provides **confidence scores, image descriptions, and explanation outputs** to clearly interpret predictions and improve transparency.

5.1 Confidence Calculation

The confidence score in the proposed zero-shot image classification system is computed using the CLIP model's similarity mechanism. Initially, the input image is converted into an image embedding, and each class label (text prompt) is converted into a corresponding text embedding. The model then calculates the cosine similarity between the image embedding and each text embedding to measure how closely they match. These similarity scores are passed through a softmax function, which converts them into probability values. The highest probability is selected as the predicted class, and its corresponding value is reported as the confidence score.

5.2 Prediction, Image Description (Using BLIP), Analysis Explanation

Prediction, image description (using BLIP), and analysis explanation are used to enhance the interpretability and understanding of the model's classification results.

Prediction

Prediction output represents the class label assigned to the input image based on the highest similarity score computed by the CLIP model. The model compares the image embedding with multiple text embeddings (class prompts) and selects the class with the highest probability. This enables the system to classify images without requiring labeled training data, supporting flexible and open-vocabulary recognition.

Image Description using BLIP

The system generates an image description using the BLIP (Bootstrapping Language-Image Pretraining) model, which converts visual information into natural language text. This description summarizes the key visual elements present in the image, helping users understand the content in a human-readable format. It enhances the interpretability of the system by providing contextual information beyond simple classification.

Analysis Explanation

The analysis explanation provides a reasoning-based interpretation of the prediction by describing how the visual features of the image align with the predicted class. It explains the decision-making process of the model in a simplified manner, highlighting important patterns, objects or characteristics observed in image.

This component improves transparency and helps users trust the system by understanding why a particular prediction was made.

5.3 Model Performance Screenshots

The experimental results show that the proposed CLIP-based zero-shot image classification system achieves reliable classification performance across multiple image categories. The pretrained CLIP model effectively captures the semantic relationship between images and natural language prompts, allowing accurate classification without model retraining.

Model Prediction Results :

Sample 1: The system predicts the image as **Late Blight** with a confidence score of 55.42%. It generates an image description using BLIP, summarizing the visual content.

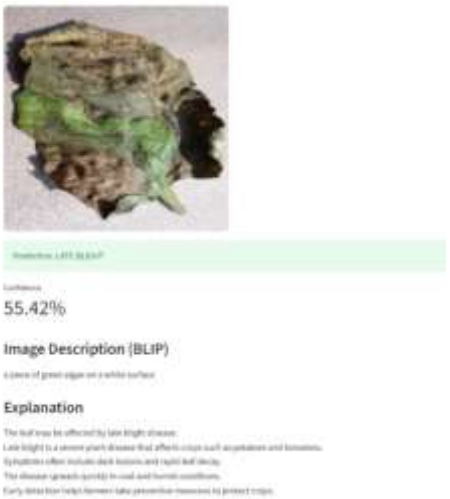

Fig. 4: Late Blight

Sample 2: Its predicts **Normal** with a confidence score of 100%. It generates an image description using BLIP, identifying it as a chest radiograph indicating healthy lungs with no visible abnormalities.

Fig. 5: Chest radiograph

Sample 3: The system predicts the image as **Normal**

Skin with a confidence score of 96.78%. It generates an image description using BLIP, identifying it as a hand. An explanation is provided, indicating no visible skin abnormalities and confirming healthy skin.

Fig. 6: Normal Hand

Sample 4: The system predicts the image as **Highway** with a confidence score of 53.84%. It generates an image description using BLIP, identifying it as an aerial view of a road and surrounding area. Describing highway features and supporting the prediction based on satellite imagery.

Fig. 7: Aerial view of a road

The model demonstrates effective performance across multiple domains by generating accurate predictions with reliable confidence scores.

6. Gaps Identified in Existing Research

Although significant progress has been made in the field of image classification using deep learning techniques, several limitations still exist in current research. One major issue observed in many traditional image classification systems is the heavy dependence on large labeled datasets for training.

Conventional CNN-based models require thousands of annotated images to achieve good accuracy. As a result, building such datasets becomes time-consuming, expensive, and labor-intensive, which limits the scalability of many image classification systems.

Another limitation in existing research is the poor generalization capability of traditional models. Many supervised learning models are trained only on specific classes present in the training dataset.

When new or unseen classes appear, the model cannot classify them correctly without retraining. This creates difficulties in real-world applications where new object categories frequently appear.

Additionally, traditional classification systems require retraining whenever new classes are introduced. This retraining process consumes significant computational resources and time. It also requires additional labeled data, which further increases development costs and limits the flexibility of the system.

Another challenge is the limited semantic understanding of images in conventional machine learning models. Traditional approaches focus mainly on visual patterns and pixel-level features but do not effectively capture the semantic relationship between images and natural language descriptions. As a result, these systems struggle to understand contextual meaning in images.

Furthermore, many existing image classification systems are not suitable for open-vocabulary classification scenarios, where the system must recognize categories that were not present in the training dataset. Most traditional models are restricted to predefined classes and cannot easily adapt to new categories without redesigning the model.

Fig. 8: Gaps Identified in Zero-Shot Image Classification using CLIP

To address these research gaps, the proposed system implements a Zero-Shot Image Classification framework using the CLIP vision–language model. The system enables classification without requiring labeled training data by aligning image features and natural language descriptions in a shared embedding space.

The proposed approach uses the CLIP Vision Encoder to extract visual features from images and the Text Encoder to convert class descriptions into semantic embeddings. The similarity between image and text embeddings is calculated using cosine similarity, allowing the system to classify images into both known and unseen categories.

7. Future Enhancements Suggested in the Literature

Future research in zero-shot image classification and vision–language models focuses on improving performance using larger and more diverse image datasets and better prompt engineering techniques. These enhancements can improve the generalization ability and robustness of models like CLIP in real-world applications.

Another potential improvement is the integration of advanced deep learning architectures and hybrid AI frameworks with vision–language models. Combining models such as transformers, multimodal learning frameworks, and improved feature extraction techniques can enhance the ability of systems to understand complex visual patterns and contextual relationships between images and text descriptions.

Researchers also suggest improving model robustness and performance by applying advanced techniques such as prompt engineering, hyperparameter tuning, and domain adaptation. These techniques help improve how effectively the model interprets text prompts and aligns them with visual features, leading to better classification results when applied to different datasets or unseen categories.

Another important future direction is the development of explainable artificial intelligence (XAI) techniques for vision–language models. Explainable models can help users understand how the system determines the similarity between image features and text descriptions. This improves transparency, trust, and reliability in AI-based image classification systems.

Additionally, future systems may integrate real-time image analysis and multimodal AI applications, where image classification models work alongside text-based systems or intelligent assistants. Such systems can automatically analyze images from cameras, mobile devices, or IoT-enabled systems and provide real-time classification and contextual insights.

Overall, these future enhancements aim to develop more accurate, scalable, and user-friendly zero-shot image classification systems that can be applied across multiple domains such as healthcare imaging, autonomous systems, content moderation, and smart visual search applications.

8. Conclusion

The rapid advancement of artificial intelligence and machine learning technologies has significantly improved the capabilities of computer vision systems. Image classification plays an important role in many applications such as image search, content moderation, medical imaging, and autonomous systems. In this work, a Zero-Shot Image Classification system using CLIP (Contrastive Language–Image Pretraining) was developed to classify images without requiring labeled training data or retraining the model.

The proposed system uses a vision–language approach where images and natural language descriptions are aligned in a shared embedding space. The CLIP vision encoder extracts visual features from images, while the text encoder processes class descriptions into semantic embeddings. By computing cosine similarity between image and text embeddings, the system can identify the most relevant class for the given image.

Experimental results show that the CLIP-based zero-shot system accurately classifies images, including unseen categories, without requiring large labeled datasets, making it scalable and suitable for real-world applications.

The proposed Zero-Shot Image Classification system using CLIP enables flexible and scalable image recognition without requiring labeled training data or model retraining.

References

[1]. A Unified Approach for Zero-Shot, Generalized Zero-Shot, and Few-Shot Learning: https://ieeexplore.ieee.org/document/8578619

[2]. Zero-Shot Scene Classification for High Spatial Resolution Remote Sensing Images: https://ieeexplore.ieee.org/document/7834720

[3]. Fine-Grained Object Recognition and Zero-Shot Learning in Remote Sensing Imagery: https://ieeexplore.ieee.org/document/8413114

[4]. A Distance-Constrained Semantic Autoencoder for Zero-Shot Remote Sensing Scene Classification: https://ieeexplore.ieee.org/document/9139620

[5]. Zero-Shot Scene Classification for High Spatial Resolution Remote Sensing Images (Alternative Record): https://ieeexplore.ieee.org/document/8558978

[6]. Few-Shot Scene Classification in Remote Sensing Using Meta-Agnostic Machine (conference paper): https://ieeexplore.ieee.org/document/9322544

[7]. Robust Deep Alignment Network With Remote Sensing Knowledge Graph for Zero-Shot Scene Classification :

https://ieeexplore.ieee.org/document/9454465

[8]. Learning Transferable Visual Models From Natural Language Supervision (CLIP): https://arxiv.org/abs/2103.00020

[9]. OpenCLIP: An Open Source Implementation of Contrastive Language–Image Pretraining: https://arxiv.org/abs/2212.07143

[10]. ALIGN: Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision: https://arxiv.org/abs/2111.07991

[11]. LiT: Zero-Shot Transfer with Locked-Image Text Tuning.
Available: https://doi.org/10.54097/8h8dff76

[12]. Florence: A New Foundation Model for Computer Vision.
Available: https://arxiv.org/abs/2111.11432

[13]. CoCa: Contrastive Captioners Are Image-Text Foundation Models
Available: https://arxiv.org/abs/2205.01917

[14]. BLIP: Bootstrapping Language-Image Pretraining for Unified Vision-Language Understanding

Available: https://arxiv.org

[15]: Bootstrapping Language-Image Pretraining with Frozen Image Encoders.
Available: https://arxiv.org/abs/2301.12597

[16]. CLIP-Adapter: Better Vision-Language Models with Feature Adapters.
Available: https://arxiv.org/abs/2110.04544

[17]. Tip-Adapter: Training-Free CLIP-Adapter for Better Vision-Language Modeling.
Available: https://arxiv.org/abs/2207.09519

[18]. Zero-Shot Learning – A Comprehensive Evaluation of the Good, the Bad and the Ugly.
Available: https://arxiv.org/abs/1707.00600

[19]. Semantic Autoencoder for Zero-Shot Learning.
Available: https://arxiv.org/abs/1704.08345

[20]. Attribute-Based Classification for Zero-Shot Visual Object Categorization.
Available:
https://ieeexplore.ieee.org/document/5437448

[21]. DeViSE: A Deep Visual-Semantic Embedding Model

Available:
https://papers.nips.cc/paper/2013/hash/7cce53cf90577442771720a370c3c723-Abstract.html

[22]. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference.
Available: https://arxiv.org/abs/1411.6736

[23]. Generalized Zero-Shot Learning via Synthesized Examples

Available: https://arxiv.org/abs/2201.12086

[24]. Deep Embedding Model for Zero-Shot Learning.
Available: https://arxiv.org/abs/1505.05925

[25]. Learning a Deep Embedding Model for Zero-Shot Learning.
Available:
https://ieeexplore.ieee.org/document/7780809

[26]. Generative Models for Zero-Shot Learning: A Comprehensive Study.
Available: https://arxiv.org/abs/1904.12242

[27]. Zero-Shot Recognition Using Semantic Embeddings and Knowledge Graphs.

Available:
https://ieeexplore.ieee.org/document/8099672

[28]. Vision–Language Pretraining: Current Trends and Future Directions.

Available:
https://arxiv.org/abs/2102.02779

[29]. Open-Vocabulary Object Detection with Vision-Language Models.
Available: https://arxiv.org/abs/2104.13921

[30]. Scaling Vision-Language Models for Image Classification and Retrieval.
Available: https://arxiv.org/abs/2204.06125