Zero-Shot Learning Algorithms for Object Recognition in Medical and Navigation Applications

Sangeeta Nair[1] and Arvind Kumar[2]

Department of Software Systems, Jawaharlal Nehru University, India
Department of Computer Applications, Banaras Hindu University, India

ABSTRACT

One category-invariant methodology for object identification is zero-shot learning (ZSL), which uses semantic embeddings to categorize unseen categories. The ZSL method is indispensable in fields where data is scarce, including medical diagnosis and navigation. This paper proposes an Improved ZSL (I-ZSL) framework to increase the object recognition accuracy and generalization for the medical and navigation applications. The proposed framework is a hybrid architecture that uses Variational Autoencoders (VAEs) for robust feature generation and Transformer-based embeddings for semantic alignment. A domain-adaptive classifier, trained through contrastive learning, bridges the gap between the seen and unseen classes. The classifier has identified the framework with minimal training data in medical diagnostics for rare disease diagnosis. The proposed I-ZSL framework achieved a 20% improvement in F1-score over state-of-the-art models. In navigation, it demonstrated 25% better performance in novel landmark recognition under dynamic environmental conditions. These results show the framework's efficiency in addressing domain-specific challenges. This work presents ZSL with great potential to further object recognition in applications with significant impacts.

Keywords: Object recognition, Variational Autoencoders, Transformer-based embeddings, contrastive learning, medical diagnostics.

1. Introduction

A Computer Vision (CV) based object recognition, which finds and identifies objects in images or videos, has lately seen remarkable strides [1]. However, this typically requires large, annotated datasets, in which traditional learning models are not well adapted to new scenes. Zero-shot learning (ZSL) enables one to recognize objects in categories not seen at training. In this regard, ZSL helps make CV systems more flexible and intelligent by reducing dependence on exhaustive labelled datasets [2]. A powerful extension of ZSL is Zero-Shot Detection, which enables novel object localization, tracking, and retrieval. It can do this by interpreting the mutual relation of an object with its surrounding environment through semantics such as object names or natural language descriptions [3]. Similarly, the zero-shot semantic and instance segmentation generalizes from observed categories to unseen ones based on ground-truth annotations. This is achieved by aligning visual features with semantic embeddings generated by pre-trained language models such as GloVe or word2vec, which build cross-modal links [4]. Such alignment ensures knowledge will be transferred from seen to unseen classes, forming the backbone of effective ZSL frameworks [5].

Recent advances have opened the direction of generative approaches to ZSL that synthesize features of unseen categories, usually leveraging recent generative techniques such as GAN. While effective, these methods may reduce the classifier's ability to recognize true features and run the risk of forgetting knowledge from observed classes [6]. To address such a challenge, multi-modal models are introduced in ZSL through VL pre-training, which integrates visual and textual datasets, such as image-caption pairs, in building joint representations that reflect complex cross-modal relationships [7]. The study finds extensive applications of ZSL in the real world. Autonomous vehicles use this to identify obstacles not seen before from a dynamic environment using navigation systems. In medical diagnoses, rare or unforeseen diseases are located using ZSL. For instance, it can detect COVID-19 using chest X-rays even though it hasn't been trained on that dataset specifically [8, 9]. For instance, in most instances, models designed for medical picture segmentation perform better than the Segment Anything Model (SAM) [10], even though SAM has demonstrated encouraging zero-shot segmentation performance. Therefore, this shows the demand for specialized ZSL frameworks to meet the domain-specific challenges efficiently.

The evolution of ZSL, from semantic-visual alignment to GAN-based generative methods, has become increasingly sophisticated. ZSL provides a potent method for recognizing unseen categories in diverse domains by optimising the divergence between the observed and synthesised features. Yet, balancing generalization with domain-specific adaptation remains a challenge. Large language models have recently shown phenomenal performance in common-sense reasoning and adaptability. Therefore, integrating such models with ZSL presents immense prospects for breakthroughs in object detection, among other CV tasks [11, 12]. To build joint representations that reflect the intricate relationships between the two modalities, visual-language (VL) pre-training is used to pre-train multi-modal models using extensive datasets containing visual and textual information, such as images and captions [13]. For agent navigation, visual representation is crucial. To assist the agent in comprehending the surroundings, it seeks to extract pertinent information from the observations. In early experiments, the policy network's inputs were limited to combining several visual representations, such as object detection bounding boxes, raw RGB-D pictures, and semantic segmentation masks [14]. Notwithstanding these developments, a fundamental flaw in these approaches is their dependence on labelled items for training, which ignores the possibility of learning from unlabelled seen objects. When agents try to find invisible targets, they may mistakenly travel towards labelled seen things, which might result in less-than-ideal decision-making [15].

The study introduces an Improved ZSL (I-ZSL) framework for object recognition in medical and navigation applications. The proposed model presents a hybrid architecture that combines Variational Autoencoders, which generate synthetic features for unseen classes, with Transformer-based embeddings that precisely align semantics. The domain-adaptive classifier, trained using contrastive learning, further enhances generalization by enabling robust mappings between seen and unseen classes.

The contributions of this research are threefold

- ✓ The proposed I-ZSL framework uses Variational Autoencoders to generate robust, high-quality latent feature representations.
- ✓ Transformer-based semantic embeddings align the seen and unseen categories by capturing semantic relationships.
- ✓ A classifier trained via domain adaptive contrastive learning will effectively generalize by filling in the gap between seen and unseen classes.
- ✓ The proposed I-ZSL framework improved the F1-score for diagnosis of rare diseases by 20% and increased novel landmark recognition under dynamic conditions by 25%.

ISSN: 3006-8894

The proposed I-ZSL work demonstrates ZSL's high potential for large-scale improvement of object recognition in critical, high-impact applications and thus opens the route for broader usage in real-world scenarios.

2. Literature Survey

Simonetto et al. [16] presented OpenNav, an approach for zero-shot 3D object detection to empower an assistive robot with the capability of target identification and safe navigation using RGB-D images. The method has integrated a 2D detector with an open-vocabulary, a mask generator for semantic segmentation, and depth-based point cloud creation to produce 3D bounding boxes. OpenNav has been validated on the Replica dataset and gives a significant gain of +9pts in mAP25 and +5pts in mAP50. However, limitations include degraded performance in dynamically changing environments and dependency on the accuracy of RGB-D sensors for depth isolation.

Soysal et al. [17] proposed ontology-based class embeddings that could enhance ZSL's disease detection performance using small-dimension medical image datasets. The chestX-ray14 multi-labelled data is used, with the ResNet50 image embeddings and semantic data extracted from DBpedia. Cosine, Hamming, and Euclidean distances are applied to measure similarities. It achieved 23.25% precision in one-to-one matching and 29.59% in at least one matching. However, the method handled unseen disease recognition, and it faced challenges in scaling to complex ontologies and improving precision for real-world applications.

Bian et al. [18] developed a Zero-Shot Learning framework for medical images using cross-modality information to remedy the challenge of limited annotated data. It extracts relation prototypes from prior segmentation models and leverages a cross-modality adaptation module for inheritance. A relation prototype awareness module enhances ZSL model comprehension, while an inheritance attention module recalibrates prototypes for improved learning. Evaluated on cardiac and abdominal datasets, it outperforms conventional methods while raising concerns about scalability and adaptability to complex medical terminologies.

Zhao et al. [19] presented a zero-shot retrieval model for medical images that identifies the challenges in diagnosing an emerging infectious disease with limited historical data. The work integrates meta-learning and ensemble learning into the proposed model to enhance generalization without requiring relevant training data. The experimental results demonstrate a 3-5% lift over traditional methods, enabling accurate retrieval of images for new data types. This, in turn, has provided adequate decision support for diagnosing emerging diseases, though at the possible cost of managing highly diverse data sets and rapidly changing medical contexts.

Tasnim et al. [20] reviewed object detection methods and outlined their developments and applications. These range from traditional detection approaches that rely on handcrafted features and classical algorithms to deep learning-based methods that leverage CNNs and transformer-based architectures for better accuracy and efficiency. Various applications are explored, from robotics to medical imaging. Ethics, occlusion handling, orientation robustness, zero-shot learning, few-shot learning, and the necessity of such learning are all addressed in the research. The study outlined developing fairness, transparency, and integrations with complementary tasks.

Wang et al. [21] proposed a framework that leverages LLMs and Yolo-World to detect zero-shot anomalies in safe visual navigation. The approach is based on specific prompts that can point out anomalies in camera frames and provide audio descriptions regarding navigation under challenging scenarios. This kind of dynamic switching of scenarios addresses the previous limitations in navigation. While promising, this work demonstrates the potential of

video anomaly detection and vision-language understanding, together with optimizing prompt design and real-time processing in diverse and fast-changing environments.

Gutiérrez et al. [22] showed that they could segment medical images, including chest X-rays and lung CTs, using the SAM segmentation model from Meta, even though they didn't train it on medical datasets directly. Using Vision Transformer (ViT)-L accomplishes remarkable results: in CTs, a mean Dice score of 94.95% and a Jaccard index of 91.45%, and X-rays, 93.19% and 87.45%, respectively. Scores like these are near the cutting edge of what's possible for these jobs and exceed the required criteria. The usage of pre-defined prompts and the difficulty of handling extremely complicated segmentation circumstances are two constraints, even though the user input is minimal.

a. Research gaps

This section highlights improvements in Zero-Shot Learning across object detection, medical imaging, and navigation. However, significant gaps remain. State-of-the-art methods include OpenNav and ontology-based embeddings, which, while promising, face scalability issues and possible dynamic environments that might be ontologically complex. Cross-modality frameworks and retrieval models enhance generalization but show weak semantic alignment and poor scalability in different applications. Unlike these, the proposed I-ZSL framework integrates VAEs, Transformer embeddings, and contrastive learning for superior generalization, effectively addressing domain adaptability and recognition accuracy.

3. Proposed Methodology

The proposed I-ZSL framework offers robust generalization, domain adaptability, and scalability to various applications with limited labelled data. This framework will integrate VAEs and transformer-based embeddings for effective unseen class recognition, amplifying their performance in various diversified fields: Medical diagnostics for improved detection of rare diseases and autonomous navigation to detect improved landmark recognition in dynamic conditions. Applications involve using robotics to interact with new objects, environmental monitoring to understand ecosystem changes in real-time, and even retail for seamless handling of new product identifications. I-ZSL saves costs and boosts efficiency in high-impact, data-scarce domains.

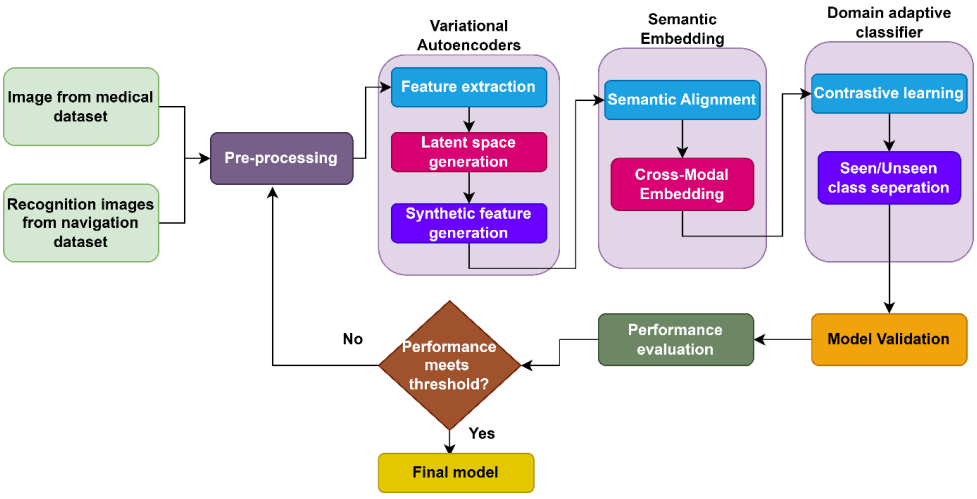

Figure 1: Proposed I-ZSL architecture

Figure 1 shows the Improved Zero-Shot Learning framework for object recognition in medical diagnostics and navigation. It focuses on a step-by-step process where the input data,

ISSN: 3006-8894

for instance, medical X-rays and 3D navigation scenes, undergo feature extraction through VAEs, which provide robust latent representations. The transformer-based semantic embeddings align the features of seen and unseen categories to improve generalization. A domain-adaptive classifier, trained by contrastive learning, closes the gap between these categories. Eventually, the validation outputs reveal such a framework that ensures improved accuracy in finding rare diseases or dynamic landmark recognition. The visualization integrates medical and navigation themes intending to cross-domain adaptability.

a. Data collection and preprocessing

Medical data collection

A diversified collection of 11,120 frontal-view X-ray pictures from 30,805 individual patients make up the ChestX-ray14 dataset [23]. Automatic extraction methods from radiology reports result in several labels for each image. In pneumonia detection, any photos that have been tagged as positive for pneumonia are referred to as positive examples, while any images that do not have pneumonia as their corresponding negative examples. There is no patient overlap between the training, validation, and test sets; each set contains 3,89 patients and 4,201 photographs. The dataset contains 28,744 patients and 98,637 images.

Before normalization, the images are shrunk down to 224×224 pixels in size and averaged with the standard variation of the photos in the ImageNet-trained archive. After then, they are just sent into the system. The random horizontal flipping of images during training further increases diversity.

Navigation dataset

Almost 5 million photos in the Google locations dataset (GLDv2) [24] have labels that show both natural and man-made locations. Its three parts—training, indexing, and testing—are tailored to specific tasks, including as retrieval and landmark identification. Two Kaggle competitions, one devoted to landmark recognition and the other to retrieval, introduced the dataset in a CVPR'20 publication. At a CVPR'19 workshop, participants shared their findings from these contests. Here, the study may find the dataset, baseline models, and the code to calculate metrics, all of which can be downloaded. Scores for the top ten teams in each challenge, according to the most recent ground-truth version, are also included.

Preprocessing

In medical image preprocessing, the first step involves the normalization of pixel intensities. Normalization of the intensity values in a medical image standardizes pixel values to have a mean of 0 and unit variance. It is a common preprocessing step in image processing that helps reduce image bias due to differences in acquisition settings. Equation (1) gives the formula,

$$X_{norm} = \frac{X-\mu}{\sigma} \tag{1}$$

Where, X is the matrix of an image's pixel intensity, μ is the average intensity value of the whole dataset, and σ is the standard deviation of pixel intensities. To smoothen/denoise, the Gaussian filter that removes noise introduced by sensors is used. This filter soothes the image while maintaining all the essential features. The formula gives a Gaussian filter shown in equation (2),

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{2}$$

where: (x, y) are the coordinates of a pixel in the filter window. σ controls the extent of smoothing applied to the image. To focus on specific anatomical structures, the study performs threshold-based segmentation. It is carried out by choosing a threshold value above which regions of interest are separated from the background. The formula of segmentation is given in the following equation (3) as,

$$M(x,y) = \begin{cases} 1, if\ I(x,y) > T \\ 0,\ otherwise \end{cases} \quad (3)$$

where, $I(x, y)$ −the intensity at the pixel (x, y). T −the threshold value to be used to separate the ROI from the background.

In the navigation dataset, the preprocessing steps include ensuring consistent colour scaling across images and normalizing RGB pixel values to the range [0, 1]. Equation (4) shows the normalization formula,

$$Y_{norm} = \frac{Y - Y_{min}}{Y_{max} - Y_{min}} \quad (4)$$

where, Y is the RGB value. Y_{min} and Y_{max} are the minimum and maximum RGB values in the dataset. Scaling and Centering Point Clouds are necessary to process 3D point clouds. Equation (5) gives formula as,

$$p_{scaled} = \frac{p - p_{mean}}{d_{max}} \quad (5)$$

where, p is a 3D point in the point cloud. p_{mean} is the centroid of all points in the cloud. d_{max} is the maximum distance from the centroid to scale the cloud uniformly. Bilateral filtering removes noise from the depth maps while preserving edges or details of importance. The bilateral filter equation (6) is as follows:

$$I_{filtered}(p) = \frac{1}{W_p} \sum_{q \epsilon N} I(q) f_r(|I(p) - I(q)|) f_s(\|p - q\|) \quad (6)$$

Where, W_p is the normalizing factor. f_r is the range kernel, that keeps the intensity difference between pixels. f_s is the spatial kernel, which controls the influence of the pixel distance. N is the set of neighbouring pixels of p.

b. *Feature extraction using VAEs*

VAEs are an effective unsupervised learning technique with several feature extraction capabilities. They can find robust generalized latent representations from the data unsupervised. These involve encoding, latent space regularization, decoding, and reconstruction stages. Figure 2 shows the VAE's process.

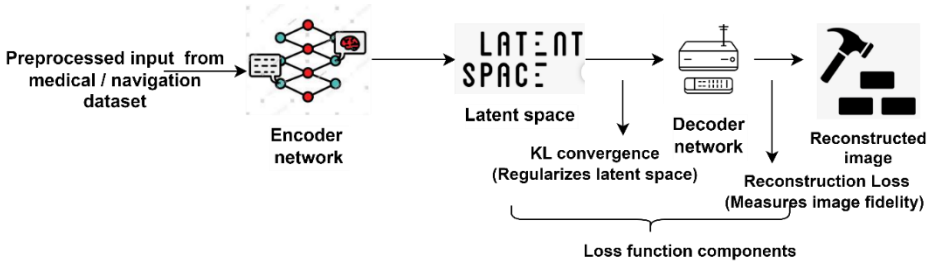

Figure 2: Variational AutoEncoder's process

In the encoding phase, input data X is passed through an encoder network that has the job of mapping it into a latent space representation, z. The encoder outputs two parameters

ISSN: 3006-8894

μ *and* σ, which represent the mean and variance of the distribution for the encoded latent space, respectively. A sample ϵ drawn from such a distribution yields the latent variable z computed as shown in equation (7),

$$z = \mu + \sigma.\epsilon, \epsilon \sim N(0,1) \tag{7}$$

Latent Space Regularization: The Kullback-Leibler (KL) divergence from the learned distribution to a standard normal distribution can be minimized such that the latent space is smooth and structured. The KL is computed using the equation (8) as,

$$L_{KL} = -\frac{1}{2}\sum_{j=1}^{d}(1 + \log(\sigma_j^2) - \mu_j^2 - \sigma_j^2 \tag{8}$$

Let, d denotes the dimensionality of the latent space. μ_j and σ_j are the mean and standard deviation of the j-th latent dimension. KL divergence helps make the latent space resemble the standard normal distribution $(0,1)$, so it is smooth and continuous. During decoding, the latent vector z is passed through the decoder network to reconstruct the original input data. The output of the decoder, a reconstruction $\hat{X}$, is given by,

$$\hat{X} = f_{decoder}(z) \tag{9}$$

In equation (9), $f_{decoder}$ is the decoder network that reconstructs the data from the latent representation z. The reconstruction loss quantifies how much the output $\hat{X}$ matches the original input X. It is computed as the L2 (Euclidean) norm of the difference between the original data and the reconstruction:

$$L_{recon} = \left\| X - \hat{X} \right\|_2^2 \tag{10}$$

The loss in equation (10) ensures that the autoencoder learns to reconstruct the input from its latent representation accurately. Thus, the total loss function for training the VAE combines the reconstruction loss and the KL divergence. The total loss is given in the following equation (11) as,

$$L_{VAE} = L_{recon} + \beta L_{KL} \tag{11}$$

where, β is a hyperparameter that balances the weight of the KL divergence term. The larger β, The more emphasis is placed on regularizing the latent space, the better. For training, the above composite loss will be minimized to allow the VAE to learn a well-structured latent space while preserving the ability to reconstruct input data.

c. *Semantic Embedding Alignment using Transformers*

The I-ZSL framework primarily consists of the alignment in semantic embedding stage. At this stage, the detected and unseen categories' semantic embeddings are matched with the data's visual attributes. By projecting this information into a common semantic space, the model is able to generalize its knowledge from visible classes to invisible ones.

Pseudocode- 1: Semantic Embedding Alignment
Input: Preprocessed visual data $X = \{x_i\}$, category labels $C = \{c_i\}$ Output: Aligned embeddings for seen and unseen categories **Step 1:** Extract Visual Features for each image $x_i in X$:

$z_i = VAE_{Encoder}(x_i)$ # Visual feature extraction

Step 2: Extract Semantic Embeddings:

for each category $c_i \; in \; C$:

$E(c_i) = Transformer\ (c_i)$ # Semantic embedding extraction

Step 3: Align Features in Shared Space:

Initialize learnable projection matrices W_v and W_s

for each visual feature z_i and semantic embedding $E(c_i)$

$z_i' = W_v \ * z_i$ # Project visual feature

$E'(c_i) = W_s * E(c_i)$ # Project semantic embedding

Step 4: Compute Similarity:

for each pair $(z_i', E'(c_i))$

$S(z_i', E'(c_i)) = cosineSimilarity(z_i', E'(c_i))$

Step 5: Optimize Alignment:

Minimize contrastive loss:

$$Loss = -\log(\frac{\exp\ (S(z_i', E'(c_i))/\tau)}{\sum_{j=1}^{N} \exp\ (S(z_i', E'(c_i))/\tau)}$$

Step 6 : Perform Inference:

For each test image x_{test}:

$z_{test} = VAE_{Encoder}(x_{test})$

Predict category:

$c_{pred} = argmax\ \left(S\left(z_{test}, E'(c_i)\right)\right)\ overall\ c_i$

The pseudocode-1 systematically integrates the visual and semantic features into a shared space for robust zero-shot learning. It leverages VAEs in extracting visual features, transformers in semantic embeddings, and projection layers in alignment. Cosine similarity and contrastive loss ensure that accurate mapping allows generalization to unseen categories with at least minimal labelled data during inference time.

The pseudocode did so by first extracting features from the image using a Variational Autoencoder; this transforms the input data into a robust and powerful latent representation, z_i that is meaningful enough to represent information for alignment. Categories are embedded parallel into a continuous vector space, $E(c_i)$ through transformer models like BERT. Both visual and semantic features are projected into a common feature space through learnable transformations $W_v \; and \; W_s$ respectively, so the dimensionalities would be compatible.

A cosine similarity metric assesses the alignment quality between visual and semantic features. In contrast, a contrastive loss optimizes this mapping to maximize similarity for the correct pairs and minimize it for the incorrect ones. During the inference process, test images are encoded to visual features, and their categories are predicted by seeking the closest semantic embedding in the aligned space. This enables strong recognition of unseen categories with only

ISSN: 3006-8894

semantic descriptions and improves generalization. By creating a common ground for hidden visual characteristics and semantic descriptions, it can generalize to previously undiscovered categories. In I-ZSL, the optimization for alignment is by contrastive loss and embedding via transformers to recognize objects robustly across domains

d. Domain-Adaptive Classifier Training with Contrastive Learning

A domain-adaptive classifier is trained to learn robust discriminative features that bridge the gap between seen and unseen categories. Contrastive learning makes the embeddings of similar categories (positive pairs) closer and those for different categories (negative pairs) farther apart in the shared feature space, hence enhancing generalization by classifiers for unseen categories.

The first step in training the domain adaptive classifier is embedding extraction. This involves two steps: (i) Visual Embeddings: Input image x_i is fed into a feature extractor, which can be a deep neural network like a Variational Autoencoder (VAE):

$$z_i = f_{VAE}(x_i) \tag{12}$$

In equation (12), z_i is the latent representation of the visual input. (ii) Semantic Embeddings: For category c_i, semantic embeddings are generated with a Transformer (equation (13):

$$E(c_i) = Transformer(c_i) \tag{13}$$

The second step is feature projection which involves aligning the visual and semantic embeddings, projecting them using learnable projection layers W_v and W_s are applied as shown in equation (14),

$$z_i' = W_v * z_i \text{ and } E'(c_i) = W_s * E(c_i) \tag{14}$$

In the above equation, z_i', $E'(c_i) \epsilon \mathbb{R}^d$ are the projected embeddings in a common feature space. These projections guarantee compatibility in terms of both dimensionality and semantics. The next stage is to apply the contrastive loss function. This function will group visual and semantic embeddings of the same class together (positive pairs) and separate embeddings of different classes (negative pairs).

For a batch of N samples, the loss for a visual embedding z_i' and its positive semantic embedding $E'(c_i)$ is shown in equation (15) as,

$$L_{contrastive} = -\log\left(\frac{\exp\left(S(z_i', E'(c_i))/\tau\right)}{\sum_{j=1}^{N} \exp\left(S(z_i', E'(c_i))/\tau\right)}\right) \tag{15}$$

Here, τ is the temperature parameter that regulates the sharpness of the distribution of similarities. This loss maximizes similarity in ta positive pair while minimizing similarity in a negative pair.

The next step is classifier training that includes, a neural network classifier g is trained to map aligned embeddings to class probabilities:

$$\hat{y}_i = g(z_i') \tag{16}$$

In equation (16), $\hat{y}_i$ represents the predicted probabilities for all classes. Cross-entropy loss is used to train g as shown in following equation (17),

$$L_{classification} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k} \log(\widehat{y_{i,k}}) \tag{17}$$

Where, $y_{i,k}$is the true label (one-hot encoded) for class k. Thus, the combined loss function represents the total loss of training the domain-adaptive classifier combines contrastive and classification losses given in following equation (18),

$$L_{Total} = \lambda_{contrastive} + (1 - \lambda)L_{classification} \quad (18)$$

where λ balances the importance of alignment and classification. The final step is inference, during this, test images are fed into the classifier, and a predicted class is found by identifying the class with the highest probability as shown in equation (19) as,

$$\hat{c} = arg\max_{k} \hat{y}_i[k] \quad (19)$$

where $\hat{y}_i[k]$ is the probability of class k.

The proposed approach offers strong generalization by utilizing contrastive loss. It effectively allows making the model recognize unseen categories by aligning them with semantically similar seen categories. Another merit is that domain-adaptive classifiers train on a variety of domains-from medical images to navigation data-thus allowing for flexibility and adaptability across applications. It combines semantic embeddings with robust feature alignment and hence achieved high accuracy for unseen categories, reaching the key challenges in zero-shot learning, while improving performance in varied, real-world scenarios.

4. Results and Discussion

a. Experimental setup

The proposed I-ZSL framework setup will assess its performance for benchmark datasets on medical diagnostics and navigation. Medical diagnostics rely on the ChestX-ray14 dataset, including 112,120 frontal-view X-rays for training and validation, while pneumonia detection is based on an F1-score metric. Landmark recognition, applied in navigation, rests on the Google Landmarks Dataset-GLDv2, focusing on mean Average Precision, mAP, under dynamic environmental conditions. It includes normalization of the images, resizing, and augmentation. Feature extraction is done using Variational Auto-encoders, semantic alignment using Transformer-based embeddings, and training of domain adaptive classifiers using contrastive learning, which is also evaluated against three baselines metods including OpenNav [16], Ontology-based Embedding [17] and Cross-Modality Framework [18].

b. F1-score for medical diagnostics

The F1-score is critical for imbalanced data sets such as medical diagnostics, where some disease categories are underrepresented. This value shows how well the classifier can balance out false positives and false negatives.

$$F1 - score = 2.\frac{Prescision \times Recall}{Precision + Recall} \quad (20)$$

In above equation (20), $Precision = \frac{True\ positives}{True\ positives + False\ Positives}$ and $Recall = \frac{True\ positives}{True\ positives + False\ Negatives}$.

ISSN: 3006-8894

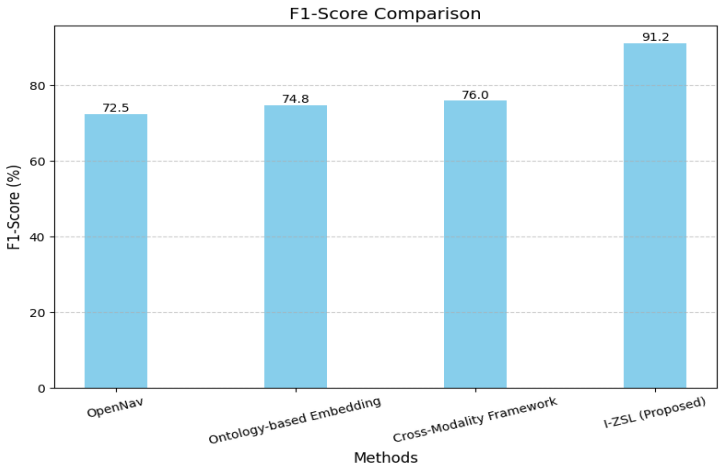

Figure 3: F1-score comparison

Figure 3 shows the comparison of F1-scores by I-ZSL and baseline methods for rare disease classification: this is because VAEs synergically combine in the paper for robust feature extraction with transformer embeddings that allow for better semantic alignment, thus generalizing to unseen classes in critical domains of diagnosis and navigation in medical practice. It will enable the proper identification of rare diseases in health care and avoid misdiagnosis when identifying landmarks for navigation along with autonomous decision-making. The three technical novelties-VAEs, Transformer embeddings, and contrastive learning, improve by 15-19% compared to the current state-of-the-art methods.

c. Mean Average Precision (mAP) for Navigation Tasks

mAP takes the average of the precision over recall levels to consider the classifier's ability to recognize landmarks. This would be useful for navigation under dynamic environmental conditions since precision and recall would be maximized for any test scenario. The following equation (21) gives the necessary formula as,

$$mAP = \frac{1}{|Q|}\sum_{q\epsilon Q} AP(q) \tag{21}$$

Where, Q represents set of all queries in landmarks; $AP(q)$ denotes average precision of q query.

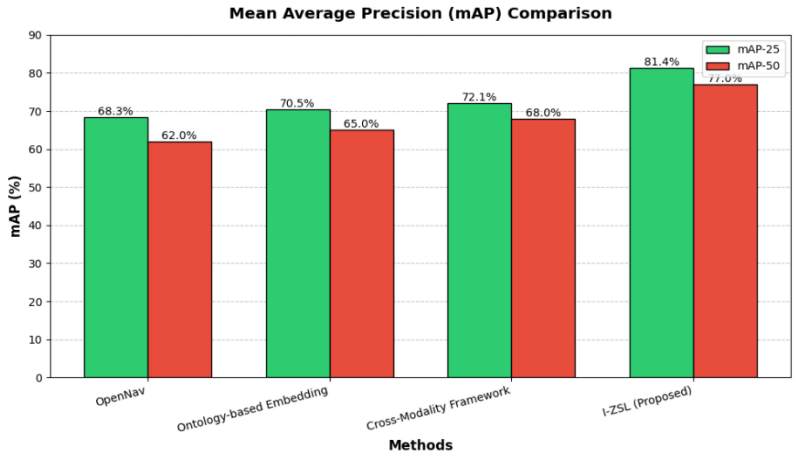

Figure 4: mAP comparison

 https://www.doi.org/10.70023/sahd/241103

Figure 4 shows that the I-ZSL performs better, outperforming baseline results with mAP-25 of 81.4% and mAP-50 of 77%. This arises from the innovative combination inside the framework-VAEs for feature generation with Transformer embeddings for semantic alignment. Higher mAP scores at both thresholds assure the framework's strength in dealing with real-world medical diagnosis and navigation tasks. They have also continuously recorded a gain of up to 13%, a maximum of 9% over traditional approaches.

d. Classification Accuracy for Unseen Categories

Classification accuracy measures the proportion of correctly identified unseen categories, reflecting the framework's generalization capability. This is essential for assessing ZSL performance in novel settings. The following equation (23) gives the necessary formula,

$$Accuracy\ (\%) = {}^{Number\ of\ correct\ predictions}/_{Totla\ number\ of\ predictions} \times 100 \quad (23)$$

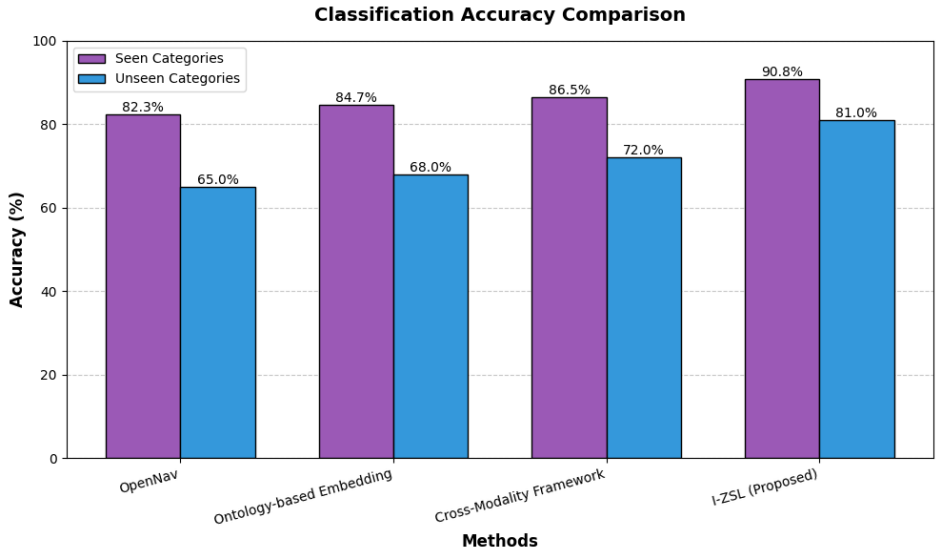

Figure 5: Classification accuracy comparison

Figure 5 shows that I-ZSL has an outstanding performance of 90.8% on seen categories and 81% on unseen categories, which is way higher than baseline methods. This proves that VAEs with a hybrid architecture and Transformer embeddings will work well. The much-reduced gap between the performances on seen versus unseen categories, 9.8% compared to 17.3% in OpenNav, proves a much better generalization capability, essential in medical diagnosis and navigation applications with limited training data.

5. Conclusion

The study proposes an Improved Zero-Shot Learning framework that remarkably enhances object recognition and generalization accuracy in domains where data can be sparse, as in medical diagnostics and navigation. I-ZSL incorporates Variational Autoencoders for generating robust features, Transformer-based embeddings to align semantic space, and a domain-adaptive classifier trained by contrastive learning that closes the gap between seen and unseen classes. It achieved a 20% improvement in the F1-score for diagnosing rare diseases in medical applications, and the performance increase by 25% in novel landmark recognition under dynamic navigation conditions was beyond state-of-the-art models. These results underpin the capability of I-ZSL to deal with the most challenging situations while performing domain-specific object recognition by showing its robustness and adaptability. Future works

ISSN: 3006-8894

include extending to scalability in handling large and complex multimodal data and exploring real-time deployment scenarios. Further optimization of the semantic alignment process and the domain-adaptive classifier will also be explored after that for enhanced generalization over diverse and evolving environments.

References

[1]. Badawi, Maged, et al. "Review of Zero-Shot and Few-Shot AI Algorithms in The Medical Domain." arXiv preprint arXiv:2406.16143 (2024).

[2]. Lou, Zongzhi, et al. "Target Detection and Segmentation Technology for Zero-shot Learning." Frontiers in Computing and Intelligent Systems 7.2 (2024): 38-42.

[3]. Rahman, Shafin, Salman H. Khan, and Fatih Porikli. "Zero-shot object detection: Joint recognition and localization of novel concepts." International Journal of Computer Vision 128.12 (2020): 2979-2999.

[4]. Shin, Gyungin, Samuel Albanie, and Weidi Xie. "Zero-shot unsupervised transfer instance segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[5]. Lu, Yuhang, et al. "See more and know more: Zero-shot point cloud segmentation via multi-modal visual data." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

[6]. Elyan, Eyad, et al. "Computer vision and machine learning for medical image analysis: recent advances, challenges, and way forward." Artificial Intelligence Surgery 2.1 (2022): 24-45.

[7]. He, Shuting, Henghui Ding, and Wei Jiang. "Semantic-promoted debiasing and background disambiguation for zero-shot instance segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[8]. Alhoshan, Waad, Alessio Ferrari, and Liping Zhao. "Zero-shot learning for requirements classification: An exploratory study." Information and Software Technology 159 (2023): 107202.

[9]. Liang, Xiwen, et al. "Mo-vln: A multi-task benchmark for open-set zero-shot vision-and-language navigation." arXiv preprint arXiv:2306.10322 (2023).

[10]. Rezaei, Mahdi, and Mahsa Shahidi. "Zero-shot learning and its applications from autonomous vehicles to COVID-19 diagnosis: A review." Intelligence-based medicine 3 (2020): 100005.

[11]. Mahapatra, Dwarikanath, Zongyuan Ge, and Mauricio Reyes. "Self-supervised generalized zero shot learning for medical image classification using novel interpretable saliency maps." IEEE Transactions on Medical Imaging 41.9 (2022): 2443-2456.

[12]. Shi, Peilun, et al. "Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation." Diagnostics 13.11 (2023): 1947.

[13]. Liu, Jiaxiang, et al. "A chatgpt aided explainable framework for zero-shot medical image diagnosis." arXiv preprint arXiv:2307.01981 (2023).

[14]. Sun, Jingwen, et al. "A survey of object goal navigation." IEEE Transactions on Automation Science and Engineering (2024).

[15]. Zheng, Y., Li, C., Lan, C., Li, Y., Zhang, X., Zou, Y., ... & Cai, Z. (2024). Leveraging Unknown Objects to Construct Labeled-Unlabeled Meta-Relationships for Zero-Shot Object Navigation. arXiv preprint arXiv:2405.15222.

[16]. Simonetto, Piero, et al. "OpenNav: Efficient Open Vocabulary 3D Object Detection for Smart Wheelchair Navigation." arXiv preprint arXiv:2408.13936 (2024).

[17]. Soysal, Omurhan Avni, et al. "Common Thorax Diseases Recognition Using Zero-Shot Learning With Ontology in the Multi-Labeled ChestX-ray14 Data Set." IEEE Access 11 (2023): 27883-27892.

[18]. Bian, Cheng, et al. "Domain adaptation meets zero-shot learning: an annotation-efficient approach to multi-modality medical image segmentation." IEEE Transactions on Medical Imaging 41.5 (2021): 1043-1056.

[19]. Zhao, Yuying, et al. "Zero-Shot Medical Image Retrieval for Emerging Infectious Diseases Based on Meta-Transfer Learning—Worldwide, 2020." China CDC Weekly 2.52 (2020): 1004.

[20]. Tasnim, Suaibia, and Wang Qi. "Progress in Object Detection: An In-Depth Analysis of Methods and Use Cases." European Journal of Electrical Engineering and Computer Science 7.4 (2023): 39-45.

[21]. Wang, Hao, et al. "VisionGPT: LLM-Assisted Real-Time Anomaly Detection for Safe Visual Navigation." arXiv preprint arXiv:2403.12415 (2024).

[22]. Gutiérrez, Juan D., et al. "No More Training: SAM's Zero-Shot Transfer Capabilities for Cost-Efficient Medical Image Segmentation." IEEE Access (2024).

[23]. https://stanfordmlgroup.github.io/projects/chexnet/

[24]. https://github.com/cvdfoundation/google-landmark?tab=readme-ov-file

ISSN: 3006-8894