Exploring the Limits of Zero-Shot Learning - How Low Can You Go?

Hemanth Dandu* Karan Sharma‡ Suchendra M. Bhandarkar*‡
*Institute for Artificial Intelligence ‡Department of Computer Science
University of Georgia, Athens, Georgia 30602, USA
hemanthreg@gmail.com karan1234@gmail.com suchi@uga.edu

Abstract

Standard zero-shot learning (ZSL) methods use a large number of seen categories to predict very few unseen categories while maintaining unified data splits and evaluation metrics. This has enabled the research community to advance notably towards formulating a standard benchmark ZSL algorithm. However, the most substantial impact of ZSL lies in enabling the prediction of a large number of unseen categories from very few seen categories within a specific domain. This permits the collection and annotation of training data for only a few previously seen categories, thereby significantly mitigating the training data collection and annotation process. We address the difficult problem of predicting a large number of unseen object categories from very few previously seen categories and propose a framework that enables us to examine the limits of inferring several unseen object categories from very few previously seen object categories, i.e., the limits of ZSL. We examine the functional dependence of the classification accuracy of unseen object classes on the number and types of previously seen classes and determine the minimum number and types of previously seen classes required to achieve a prespecified classification accuracy for the unseen classes on three standard ZSL data sets. An experimental comparison of the proposed framework to a prominent ZSL technique on these data sets shows that the proposed framework achieves higher classification accuracy on average while providing valuable insights into the unseen class inference process.

Keywords: Zero-shot learning, unsupervised clustering, semantic embedding, deep learning

1. Introduction

Advances in *Deep Neural Network* (DNN) architectures have empowered computers to achieve human-level classification performance on object recognition tasks through the development of powerful and robust visual classifiers. Many problem domains are faced with a large and growing number of object categories resulting in a need to collect and annotate a large number of training images for each object category to enable the classifier to adapt to the naturally occurring variations in object appearances. Since classifiers trained on observed object instances lack the ability to deal with previously unseen classes, efficient collection and reliable annotation of training image data for a wide variety of object categories is of critical importance. To address the training data collection and annotation bottleneck, various *Zero-Shot Learning* (ZSL) methods have been proposed wherein training image data for every single object category is not strictly required.

Humans perform ZSL naturally, enabling recognition of at least 30,000 object classes [3]. The human ability to understand natural variations in object appearances stems from variety of factors. However, one of the prominent factors is an existing and ever evolving natural language (NL) knowledge base, which enables humans to connect unseen object categories to previously seen categories using high-level NL descriptions. To emulate the ZSL process on computers, previously unseen object categories are recognized by leveraging auxiliary information related to the unseen object categories derived typically from external data sources analogous to the human NL knowledge base such as *Wikipedia* and *WordNet* [15], or in some cases, manually engineered ontologies.

Several ZSL methods have been proposed in the literature; however, all of them use a proposed split of standard ZSL data sets [10, 29, 19] into *seen* and *unseen* classes to aid uniform research towards the formulation of a universal optimal ZSL framework [31]. The formal ZSL problem is typically posed as one of maximizing classification accuracy using specific categories within a standard data set as seen classes and the remaining as unseen classes where the number of seen classes is significantly higher than the number of unseen classes. While conventional ZSL methods have resulted in the formulation of several benchmark

approaches, the critical issue of mitigating the training data collection and annotation process has been largely skirted.

In this paper, we propose a new ZSL framework to infer the limits of inferring unseen object categories from very few seen object categories, i.e., test the limits of ZSL. By aiming to infer a *large* number of unseen object categories using *very few* seen object categories, the proposed ZSL framework aims to address the critical issue of optimizing the seen object categories and the number of annotated training images needed from each of the seen object categories to achieve a prespecified overall classification accuracy. We note the functional dependence of the classification accuracy on the number of previously seen classes drawn from the entire spectrum of classes in three widely used object classification data sets, i.e., *Animals-with-Attributes-2* (AWA2) [10], *CalTech-UCSD Birds-200-2011* (CUB) [29] and *Scene Understanding with Attributes* (SUN) [19]. We determine the optimal set of representative seen classes that allows one to infer a large number of previously unseen classes with a prespecified measure of accuracy.

The proposed framework significantly aids the training data collection and annotation process by identifying the key object categories from which this process can be initiated and determining the object categories at which it can be halted based on an expected or prespecified classification accuracy measure for a given problem. We evaluate the proposed framework in the more realistic generalized ZSL (GZSL) setting where the input images during prediction or inference can come from both, the seen and unseen classes [22]. We present valuable insights into the inference process for cases where the proposed framework performs exceptionally well, and for cases where it fails to infer the correct unseen category. We also compare the proposed framework with the well known *Attribute Label Embedding* (ALE) [1] scheme for ZSL, which has been shown in [31] to perform very well on the aforementioned AWA2, CUB and SUN data sets.

2. Related Work

Zero-shot learning (ZSL) approaches can be broadly classified into two categories, i.e., *inductive* ZSL (IZSL) and *transductive* ZSL (TZSL), based on the unseen class information available during the training process. In IZSL, one has access to labeled image data from the seen classes and auxiliary information (i.e., semantic attributes/descriptions) from both, seen and unseen classes during training. In the TZSL framework, one has access to auxiliary information from both, seen and unseen classes, labeled image data from the seen classes and unlabeled image data from the unseen classes during training. The IZSL and TZSL schemes can each be subclassified as conventional or generalized based on the model evaluation procedure used during testing. Conventional ZSL assumes that the input images during prediction or inference can only come from the unseen classes whereas generalized ZSL (GZSL) allows for the more practical real-word scenario where the input images during prediction or inference can come from both, the seen and unseen classes [22].

The *Directed Attribute Prediction* (DAP) and *Indirect Attribute Prediction* (IAP) models [11], use a two-stage IZSL approach where in the first stage, the attributes of an image are predicted, and in the second stage, the class label is inferred by searching for the class with the most similar set of attributes. Although widely cited, DAP and IAP suffer from domain shift [8] where the intermediate functions learned from the auxiliary information are observed to introduce an unknown bias in the absence of adaptation to the target domain. More recent ZSL schemes learn a compatibility function from the image feature space to the semantic or auxiliary space and are categorized based on learned compatibility function. *Attribute Label Embedding* (ALE) [1], *Deep Visual-Semantic Embedding* (DEVISE) [7], *Structured Joint Embedding* (SJE) [2] learn a linear compatibility function between the image space and semantic or auxiliary space coupled with an optimization algorithm. *Embarrassingly Simple Zero-Shot Learning* (ESZSL) [21] uses an additional regularization term to suppress noise in the auxiliary space whereas *Latent Embedding* (LATEM) [30] generalizes the notion of linear compatibility by learning a *piecewise* linear compatibility function resulting in significantly improved accuracy.

Cross-Modal Transfer (CMT) [24] approaches extend the linear compatibility-based approaches by learning *non-linear* projections from the image space to *Word2Vec* [14] space. Hybrid approaches [31] learn a joint embedding of both the image and semantic features in a combined intermediate space using a combination of linear and non-linear compatibility functions. *Semantic Similarity Embedding* (SSE) [32] uses a max-margin scheme to jointly optimize domain data and semantic data. The *Convex Combination of Semantic Embeddings* (CONSE) scheme [18] maps images into the semantic embedding space via a convex combination of the class label embedding vectors thereby obviating the need for additional training. Wang et al. [28] use a *Graph Convolutional Network* (GCN) and *GLoVe* text embedding [20] to generate a knowledge graph

embedding that exploits both, semantic embeddings and domain relationships to predict the object classifiers.

Recent ZSL approaches use the *Generative Adversarial Network* (GAN), where each class is represented by a probability distribution and the generator network synthesizes fake unseen features from noisy inputs and the discriminator network distinguishes the fake features from the real ones. The *Generative Framework for Zero-Shot Learning* (GFZSL) [26] models the class-conditional distributions of seen and unseen classes using a multivariate Gaussian distribution. The *Leveraging the Invariant Side GAN* (LisGAN) approach [13] uses a conditional Wasserstein GAN wherein fake unseen features are generated from random noise functions conditioned on the semantic descriptions and the discriminator learns to distinguish the fake features from the real ones via a minimax game. The *LsrGAN* approach [27] performs explicit knowledge transfer between the seen and unseen object categories using a novel *semantic regularized loss* (SR-Loss) function that exploits the semantic relationships between the two categories. The recent *TF-vaegan* [17] combines a *Variable Auto-Encoder* (VAE) and GAN and introduces a feedback loop from a semantic embedding decoder that iteratively refines the generated features during both the training and feature synthesis stages. The synthesized features and their corresponding latent embeddings from the decoder are then transformed into discriminative features and exploited during classification to reduce the ambiguities amongst the categories.

The proposed ZSL framework draws upon the two-stage DAP and IAP approaches [11] and hence falls within the IZSL category. However, the problem being addressed by the proposed ZSL framework is substantially different in that the proposed ZSL framework aims to predict a *large* number of unseen classes from *very few* seen classes and, consequently, has access to far less training data compared to conventional ZSL approaches. Hence, generative-adversarial-based and compatibility learning-based ZSL approaches which are training data-intensive would be expected to perform poorly in this scenario. The proposed ZSL framework permits one to examine and understand how classification accuracy measures change as a function of the number and types of the seen classes. The proposed framework also allows for selection of an optimal number and type of seen classes based on an expected overall classification accuracy measure which in turn allows one to optimize the training data collection and annotation process. The proposed framework is inspired by Sharma et al. [23] who address the scenario where the unseen classes significantly outnumber the seen classes; however, they use only embeddings from unstructured text corpora and k-means clustering which yields discrete categories. In contrast, the proposed framework incorporates embeddings from structured information and uses clustering based on partial membership and is experimentally compared to an extensively cited ZSL scheme on widely used ZSL evaluation data sets.

3. Methodology

Figure 1 gives a high-level depiction of the proposed framework. Broadly, the proposed framework comprises of the following components: deep feature extraction; incorporation of auxiliary information; clustering of auxiliary information; multi-label classification of deep features; and prediction of categories/classes. We present the details of each of the above components.

3.1. Deep Feature Extraction

A *ResNet-101* [9] DNN architecture pre-trained on the *ImageNet* data set [5] is used to extract deep-learned features from images, thus reducing the training time significantly while yielding useful features. Features are extracted from the last ResNet-101 layer, resulting in 2048 features for each image. The extracted visual features are split into training and testing sets using a stratified 80:20 split. Stratified sampling is used so that instances of seen and unseen classes are present in the test set, thus making it a GZSL setting.

3.2. Auxiliary Information

Auxiliary information is exploited to establish semantic relationships between seen and unseen classes, which is a critical aspect of ZSL. In the proposed framework, three sources of auxiliary information are used.

Attributes. All three data sets have labelled attribute information for each of their classes. The AWA2 data set [10] includes attributes such as *black*, *small*, *walks*, *smart* etc. with 85 such attributes for each class. The CUB data set [29] includes attributes such as *primary color*, *wing color*, *wing shape*, *size* etc. with 312 such attributes for each class. The SUN data set [19] includes attributes such as *man-made*, *natural light*, *medical activity*, *diving* etc. with 102 such attributes for each class.

Text Embeddings. The learned vector representations for words (i.e., word/text embeddings) derived from a large general text corpus can help to construct semantic relationships between the seen and unseen class labels. *FastText* [4] has been shown to perform better than other word embedding models such as *Word2Vec* [14] and *GloVe* [20] since it treats each word

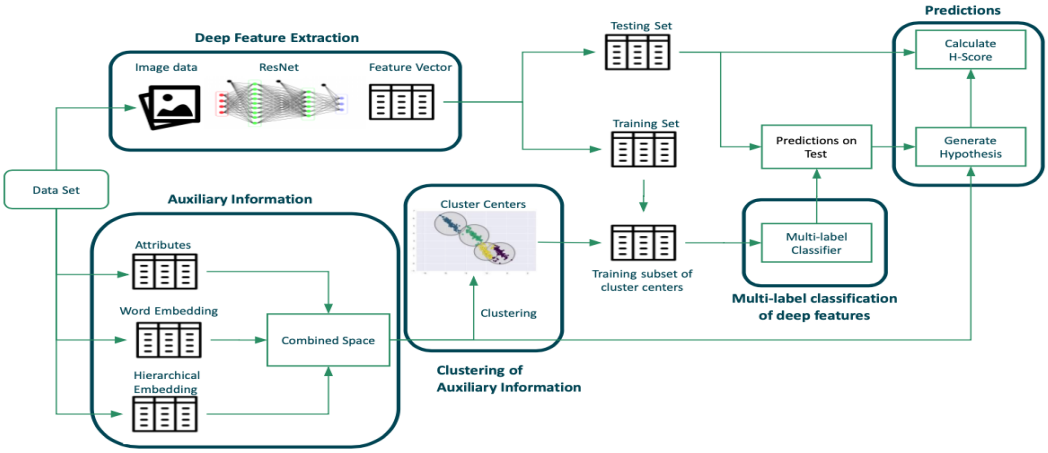

Figure 1. High-level schematic of the proposed framework.

as composed of character n-grams. Consequently, FastText can also generate vectors for a combination of words; for instance, *"polar bear"* has a unique FastText vector representation. In our framework, FastText embeddings pre-trained on a very large *Wikipedia* corpus are extracted for each class label, resulting in a 300-dimensional vector for each object category.

Hierarchy Embeddings. Creating a hierarchy of categories present in a data set allows us to derive taxonomy-based relationships between the classes and thereby improve ZSL performance. For the AWA2 and CUB data sets, Lee et al. [12] propose a two-stage approach for generating hierarchy embeddings where they first derive a top-down hierarchy using WordNet [15] and then create a flattened hierarchy by representing the probabilities of all the leaf nodes as a single probability vector. In the case of the AWA2 and CUB data sets, this results in a 61-dimensional vector and 193-dimensional vector respectively. The SUN data set, on the other hand, provides its own two-level hierarchy information for each of its 717 categories resulting in a 19-dimensional vector.

Combined Semantic Space. The vector spaces of attributes, text embeddings, and hierarchy embeddings are combined into a unified space with reduced dimensionality, while retaining the most important information. Dimensionality reduction of the semantic space reduces the computational complexity of the clustering phase and creates robust clusters. *Principal Component Analysis* (PCA) is used for dimensionality reduction since it retains the variance in the input data while reducing the data dimensionality resulting in a compact combined semantic space. Figure 2 shows the t-distributed stochastic neighbour embedding (t-SNE) plots [25] of the combined semantic space for the AWA2 and CUB data sets. Since CUB is a fine-grained data set whereas AWA2 is a coarse-grained data set, we observe that the classes in the former are clustered closer to each other compared to those in the latter.

3.3. Clustering of Auxiliary Information

Clustering is used to identify object categories that are good representatives for a large number of similar object categories. The underlying hypothesis is that the resulting cluster centers would have a strong relationship with its cluster members, thus allowing us to infer cluster members using the cluster centers alone. We use two clustering algorithms, i.e., the *Gaussian Mixture Model* (GMM) [16] and *Affinity Propagation* (AP) [6], to identify the clusters and representative classes for the clusters where the clustering is performed in the combined reduced-dimensional semantic space.

Gaussian Mixture Model (GMM): The GMM can accommodate clusters that have different sizes and correlation structures within them, as opposed to k-means clustering where the number of clusters is denoted by the variable k. GMM-based clustering requires us to specify the number of clusters (i.e., GMM components) k before fitting the model. In our experiments, we start with $k = 5$ and end with a value of k that equals the total number of classes for a given data set in steps of 5.

Affinity Propagation (AP): AP is a clustering algorithm based on concept of *"message passing"* between data points. Unlike the GMM, AP does not require the number of clusters k in the final output to be specified in advance; the algorithm itself determines the optimal number of clusters k.

In this work, we focus primarily on the GMM-based clustering algorithm since we propose to study how changing number of seen classes, i.e., the value

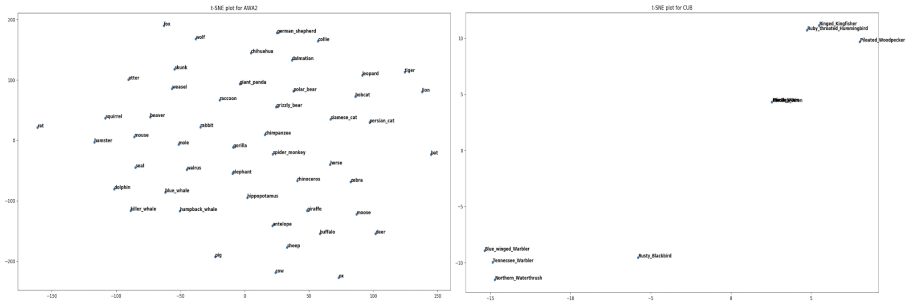

Figure 2. The *t*-SNE plots of the combined semantic space for the AWA2 (all classes) and CUB (10 random classes) data sets.

of k, impacts the classification accuracy.

3.4. Multi-label Classification of Deep Features

A trained visual classifier is used to label each new test image into one of the representative objects corresponding to the clusters. For each value of k, the training set is filtered for the class labels associated with the cluster centers. For example, in the AWA2 data set, for $k = 5$, the class labels *chimpanzee*, *hamster*, *humpback whale*, *bobcat*, and *ox* are the cluster centers. The image features associated with the class labels of the cluster centers alone are considered as the training set and a multi-class visual classifier is trained using this training set.

Visual Classifier. In our framework, we use a *Random Forest* (RF) classifier since it is highly scalable for a large number of classes and known to yield good results. The RF classifier builds an ensemble of decision trees using a bagging ensemble technique. Since we need to train multiple classifiers for varying values of k, when using GMM-based clustering, it is not feasible to train neural network- or SVM-based classifiers because of the extensive parameter tuning involved.

3.5. Generation of Predictions

Each test instance is classified into one of the representative clusters for a given value of k. Alternative hypotheses or predictions are then generated using a similarity measure in the combined semantic space.

Similarity measure. If the distance between two data points is small then there is a high degree of similarity between the classes and vice versa. We use the *cosine similarity measure* which computes the cosine of the angle between two vectors. The cosine similarity measure is advantageous because even if the two classes are far apart based on a standard distance metric, it is possible for their corresponding vectors to be closely aligned in terms of angular separation. The smaller the angular separation between the two vectors, the higher the cosine similarity measure.

Testing. The test set is split into two subsets. The first subset denotes the seen classes, comprising of data pertaining to class labels present in the training set. The second subset denotes the unseen classes, comprising of data pertaining to class labels absent from the training set. For each of these subsets, we determine top prediction using the trained visual classifier and then find the closest class label in the combined semantic space using the cosine similarity measure. Classification accuracy is computed for each subset separately followed by the computation of the *harmonic score* or *H-Score* (HS) which is the harmonic mean of the *seen class accuracy* (SCA) and *unseen class accuracy* (UCA) as shown in equation (1). Since our aim is to attain high classification accuracy on both the seen and unseen classes, the harmonic mean is a better quantifier of overall classification accuracy.

$$HS = \frac{2 \times (SCA \times UCA)}{(SCA + UCA)} \tag{1}$$

4. Results: Comparison and Discussion

Since the experimental setup for the proposed framework differs from that of standard ZSL, it is hard to compare our results with those of other ZSL frameworks described in the literature. To achieve a fair comparison of the proposed framework with other ZSL approaches, we adapt a well known ZSL framework, i.e., *Attribute Label Embedding* (ALE) [1] to our experimental setup.

ALE uses a bilinear compatibility function to associate visual and auxiliary information by embedding each class in the space of attribute vectors. A comparison study performed in [31] shows that ALE outperforms other ZSL frameworks in the GZSL setting. Recent generative methods described in the literature could potentially perform better than ALE but such a comparison study using same data sets, experimental conditions, and evaluation metrics has not been performed yet. Hence, we chose to compare the proposed approach with the ALE framework. Once the clusters centers are determined for each value of k, the ALE procedure is performed on the appropriate

training and testing sets. Table 1 and Table 2 show a summary of the comparison between the proposed approach and ALE using both, GMM-based clustering and AP-based clustering algorithms.

Results on the AWA2 data set. The AWA2 data set has 50 classes and the training set consists of 560 images per class on average. We study performance of the proposed approach for $k = 25$ which renders half the classes as seen and half the classes as unseen for the model. The seen classes exhibit an average classification accuracy of 85% whereas the unseen classes exhibit an average classification accuracy of 27% on the test set.

Among the input classes, we identify three cases of seen classes and three cases of unseen classes:

Seen Classes Case 1: Classes exhibiting $\geq$ 90% classification accuracy on the test set. In the AWA2 data set, 16 of 25 seen classes fall in this category and are expected to aid very well in the inference of unseen classes related to these seen classes. These seen classes have a good number of images to train on and the classifier is able to clearly identify distinguishing features for each class. For example, *humpback whale* is a seen class that exhibits 100% classification accuracy on the test set. Ideally, we would want all seen classes to fall into this category.

Seen Classes Case 2: Classes exhibiting classification accuracy $\geq$ 60% but $<$ 90% on the test set. Seven of the 25 seen classes fall in this category and are expected to aid reasonably in the inference of unseen classes. Although these classes have a reasonable number of images to train on, they are very close to each other but are still considered as seen classes. For example, *ox*, *moose*, and *cow* fall in this category. It is understandable why the visual classifier is unable to clearly distinguish between these categories, since they are quite similar compared to other categories in this data set.

Seen Classes Case 3: Classes exhibiting classification accuracy $<$ 60% on the test set. Two of the 25 seen classes fall in this category and are expected to negatively impact the unseen class inference process. Since the training sets for these 2 classes have only 160 images on average, the visual classifiers are insufficiently trained to learn the class discriminative features.

Unseen Classes Case 1: Classes exhibiting $>$ 60% classification accuracy on the test set. Six of the 25 unseen classes fall in this category comprising of cases when a particular unseen class is inferred from a seen class falling in Case 1 of seen classes. For example, *blue whale* is an unseen class that exhibits 100% accuracy on the test set, and it is inferred from *humpback whale* which falls under Case 1 of the seen classes.

Unseen Classes Case 2: Classes exhibiting classification accuracy $\geq$ 1% but $\leq$ 60% on the test set. Six of the 25 unseen classes fall in this category. These are cases where the unseen classes are inferred from the seen classes falling in Case 2 and Case 3 of seen classes. These unseen classes exhibit poor performance since the seen classes they are inferred from are not clearly distinguishable by the visual classifier. For example, the unseen class *deer* is inferred from the seen class *moose* which falls in Case 2 of the seen classes.

Unseen Classes Case 3: Classes that exhibit 0% accuracy on the test set. Twelve of the 25 unseen classes fall in this category. Our inference procedure only allows one unseen class inference per seen class; hence, unseen classes that are farther away from seen classes cannot be inferred and fall in this category. For example, *antelope* is an unseen class that falls in the *moose* cluster, but only *deer* can be inferred from *moose* since it is the closest neighbor to *moose* based on the cosine similarity measure.

Figure 3 shows a comparison of the H-scores obtained on the all data sets by the proposed model and ALE for all values of k. Our model performs significantly better than ALE for all values of k on this data set. The H-score increases monotonically with increasing k which is expected since the increasing number of seen classes enables more unseen classes to be inferred with greater accuracy. Figures 4 and 5 show the qualitative results for the unseen and seen cases, respectively. We observe that the proposed model performs well when a seen class has a sufficient number of training images and the unseen class being inferred from it is proximal to the seen class and performs poorly when the unseen class being inferred is distant from any of the representative classes.

To determine the optimal number of classes required to achieve reasonable performance, we choose a k value that results in a greater than average H-score performance for the proposed model. For $k = 20$, we achieve a H-score of 46% with the proposed model whereas the average H-score on the AWA2 data set is 45% across all categories and k values. Thus, on the AWA2 data set, we need to have at least 20 seen classes to reasonably infer the unseen classes with greater than average accuracy with the proposed model.

Results on the CUB data set. The CUB data set has 200 classes with 47 images per class on average in the training set. This is a small number of images to train on for each seen class. The proposed model performs better than ALE for values of $k \leq 65$. For $65 < k \leq 115$, the proposed model and ALE exhibit comparable performance and for $k > 115$, the

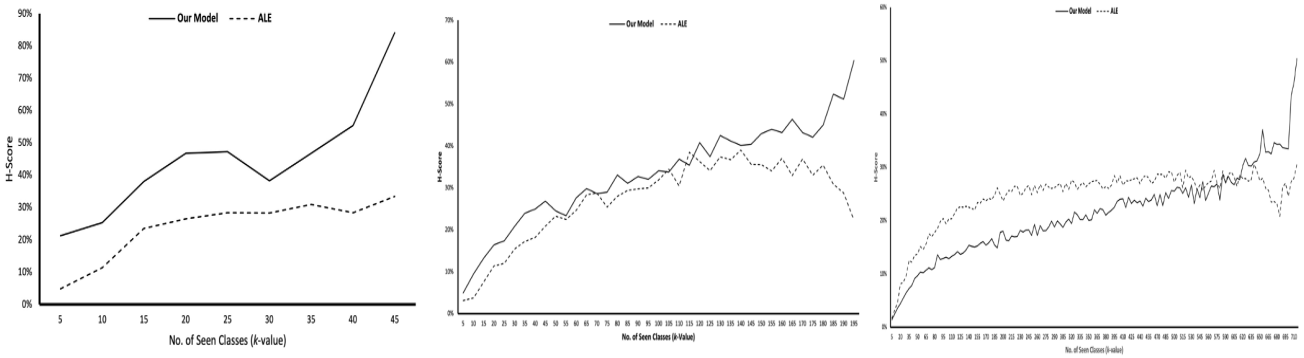

Figure 3. H-score comparison between the proposed model and ALE on the (from left to right) AWA2, CUB, and SUN data sets when using GMM-based clustering.

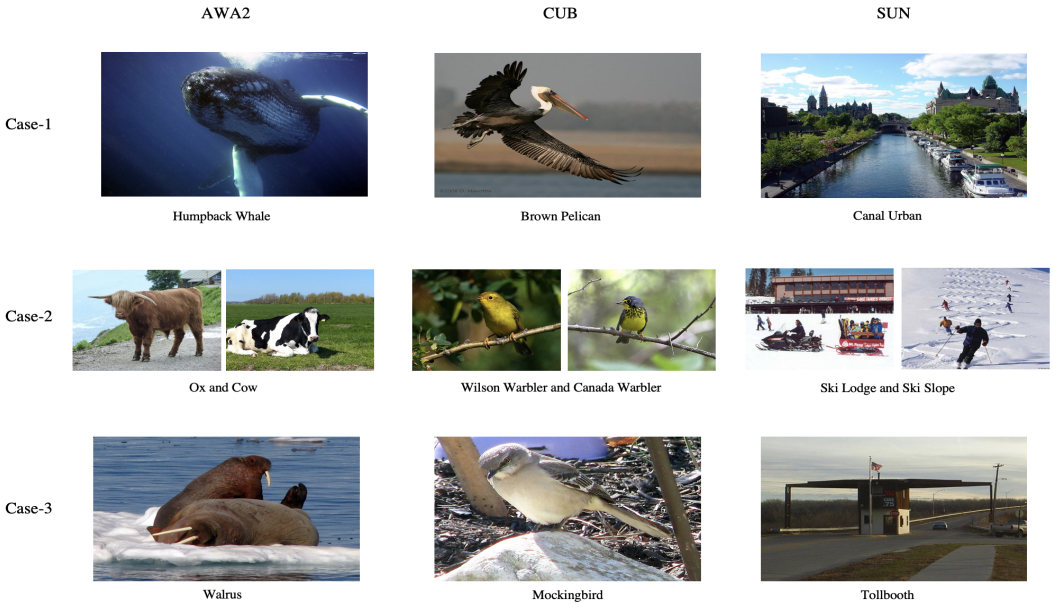

Figure 4. Qualitative results of seen cases from AWA2, CUB, and SUN data sets. Case-1 includes categories with clearly identifiable discriminative features. Case-2 include categories for which we are unable to clearly distinguish between similar categories. Case-3 include categories that have fewer images to train on and we are unable to identify distinguishing features.

Table 1. Comparison of average classification accuracy across k values between the proposed model and ALE when using GMM-based clustering.

	Proposed Model			ALE		
Data Set	**Avg. Seen**	**Avg. Unseen**	**Avg. H-Score**	**Avg. Seen**	**Avg. Unseen**	**Avg. H-Score**
AWA2	94%	32%	**45%**	90%	14%	24%
CUB	78%	22.86%	**33%**	70%	17.50%	27.19%
SUN	58.20%	15%	21.60%	41.50%	17.80%	**25%**

Table 2. Comparison of average classification accuracy between the proposed model and ALE when using AP-based clustering.

	Proposed Model			ALE		
Data Set	**Seen**	**Unseen**	**H-Score**	**Seen**	**Unseen**	**H-Score**
AWA2	96.44%	23.43%	**37.7%**	83.40%	10%	17.50%
CUB	91%	9.70%	**17.50%**	55%	8.33%	14.40%
SUN	83.10%	4.30%	8.20%	24.40%	8.20%	**12.30%**

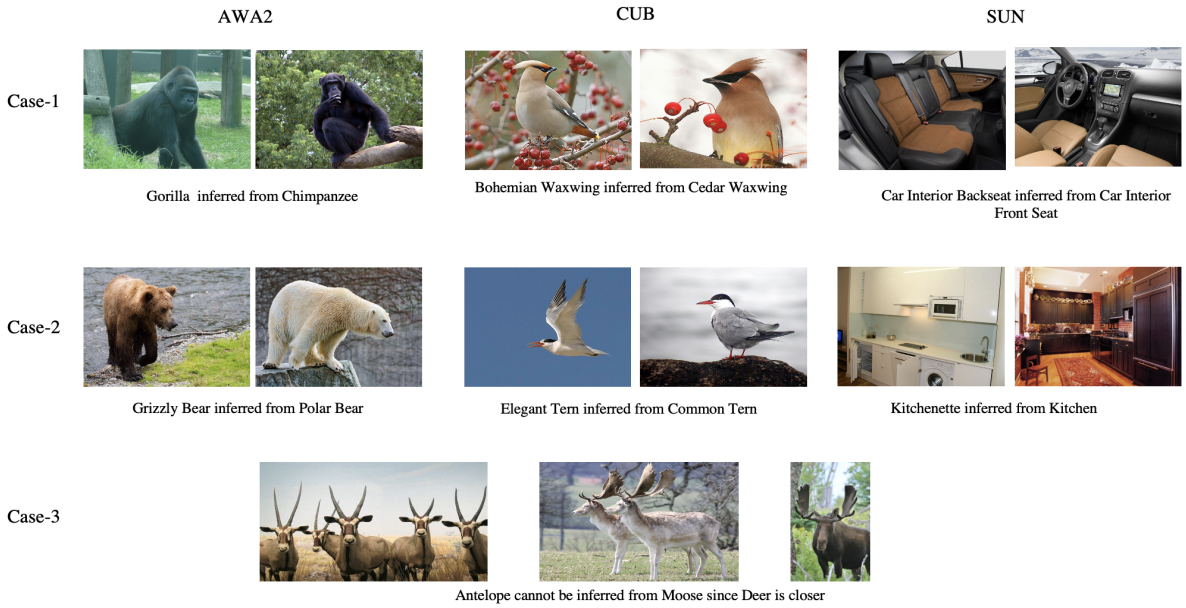

Figure 5. Qualitative Results of unseen cases from AWA2, CUB, and SUN data sets. Case-1 includes categories which are inferred from seen classes belonging to Case-1. Case-2 includes categories which are inferred from seen classes belonging to Case-2 and Case-3. Case-3 includes categories which are farther away from the seen classes and cannot be inferred.

proposed model significantly outperforms ALE on the CUB data set. Thus, across a large range of k values, the proposed model performs better than ALE on the CUB data set. For $k = 80$, we achieve a H-score of 33.5% with the proposed model whereas the average H-score on the CUB data set is 33% across all categories and k values. Thus, on the CUB data set, we need at least 80 seen classes to reasonably infer the unseen classes with higher than average accuracy with the proposed model.

Results on the SUN data set. The SUN data set has 717 classes with 16 images per class on average in the training set. This makes it hard for the visual classifier to learn distinguishing features for each class because of the large number of classes and small number of images for each class. ALE performs better than the proposed model for values of $k \leq 560$ and the proposed model performs better than ALE for values of k beyond 560. For $k = 360$, we achieve a H-score of 22% with the proposed model and the average H-score on the CUB data set is 21.6% across all categories and k values. Thus, on the SUN data set, we need at least 360 seen classes to reasonably infer the unseen classes with greater than average accuracy with the proposed model.

5. Conclusions and Future Work

We have proposed a framework for generalized zero-shot learning (GZSL) that is simple yet very effective for scenarios where the number of unseen classes is significantly higher than that of seen classes. The proposed framework offers an intuitive approach to aid in the training data collection and annotation process for image recognition tasks by identifying representative classes using unsupervised clustering and a method to infer unseen classes using a simple cosine similarity measure. The proposed framework achieves accuracy figures that are 21% greater on the AWA2 data set and 6% greater on the CUB data set when compared to the well known Attribute Label Embedding (ALE) scheme for GZSL; whereas on the SUN data set, it exhibits performance that is comparable to that of ALE. The proposed framework allows us to determine the minimum number and types of categories needed to be considered as previously seen to achieve reasonable classification accuracy results on all the three data sets.

A current drawback of the proposed framework is its inability to infer unseen classes that are very distant from the representative classes in the semantic space which presents scope for future improvement. A potential solution could be a scheme to map the distance between each unseen class and representative class in a cluster to the classification probabilities obtained from the visual classifier allowing the framework to infer all unseen classes, regardless of the distance, with non-zero probability. We also intend to evaluate the scalability of the proposed framework on a very large data set such as ImageNet with ≥ 1000 classes and several hundred images per class.

References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 819–826, 2013.

[2] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 06 2015.

[3] I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94:115–147, 1987.

[4] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[5] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[6] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.

[7] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26, pages 2121–2129. Curran Associates, Inc., 2013.

[8] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Transductive multi-view zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(11):2332–2345, 11 2015.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[10] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36:453–465, 2014.

[11] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014.

[12] K. Lee, K. Lee, K. Min, Y. Zhang, J. Shin, and H. Lee. Hierarchical novelty detection for visual object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1034–1042, June 2018.

[13] J. Li, M. Jing, K. Lu, Z. Ding, L. Zhu, and Z. Huang. Leveraging the invariant side of generative zero-shot learning. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2019.

[14] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26, 10 2013.

[15] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, Nov. 1995.

[16] O. M. M. Mohamed and M. Jaïdane-Saïdane. Generalized gaussian mixture model. In *2009 17th European Signal Processing Conference*, pages 2273–2277, 2009.

[17] S. Narayan, A. Gupta, F. S. Khan, C. G. Snoek, and L. Shao. Latent embedding feedback and discriminative features for zero-shot classification. *arXiv preprint arXiv:2003.07833*, 2020.

[18] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. In *International Conference on Learning Representations*, 2014.

[19] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2751–2758, 2012.

[20] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[21] B. Romera-Paredes and P. H. S. Torr. An embarrassingly simple approach to zero-shot learning. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 2152–2161. JMLR.org, 2015.

[22] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2013.

[23] K. Sharma, H. Dandu, A. C. S. Kumar, V. Kumar, and S. M. Bhandarkar. Exploiting word embeddings for recognition of previously unseen objects. In *Pattern Recognition. ICPR International Workshops and Challenges*, pages 314–329, Cham, 2021. Springer International Publishing.

[24] R. Socher, M. Ganjoo, C. D. Manning, and A. Y. Ng. Zero-shot learning through cross-modal transfer. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'13, page 935–943, Red Hook, NY, USA, 2013. Curran Associates Inc.

[25] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[26] V. K. Verma and P. Rai. A simple exponential family framework for zero-shot learning. In *Machine Learning and Knowledge Discovery in Databases*, pages 792–808, Cham, 2017. Springer International Publishing.

[27] M. Vyas, H. Venkateswara, and S. Panchanathan. Leveraging seen and unseen semantic relationships for generative zero-shot learning. In *Computer Vision – ECCV 2020 - 16th European Conference, Proceedings*, pages 70–86, 2020.

[28] X. Wang, Y. Ye, and A. Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6857–6866, 2018.

[29] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[30] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent embeddings for zero-shot classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 69–77, 2016.

[31] Y. Xian, B. Schiele, and Z. Akata. Zero-shot learning — the good, the bad and the ugly. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 07 2017.

[32] Z. Zhang and V. Saligrama. Zero-shot learning via semantic similarity embedding. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4166–4174, 2015.