International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 08 | Aug 2023 www.irjet.net p-ISSN: 2395-0072

Zero Shot Learning

Mudassir Ubaid[1], Mohd Tanveer Hasan[2]

[1]Student, Dept. of Computer Engineering, Zakir Hussain College of Engineering and Technology, Aligarh Muslim University, Aligarh, India
[2]Student, Dept. of Computer Engineering, Zakir Hussain College of Engineering and Technology, Aligarh Muslim University, Aligarh, India

---

Abstract - *This paper provides a comprehensive overview of zero-shot learning (ZSL), a subfield of machine learning that aims to recognize and classify new objects or concepts without prior exposure during training. ZSL utilizes semantic representations to enable models to generalize to new concepts, based on the knowledge acquired in related classes. The paper discusses the two primary types of semantic representations, attribute- based and semantic space-based methods, and examines re- cent developments in ZSL, such as generative models, novel semantic representation methods, and multi-modal ZSL. The potential applications of ZSL in a variety of fields, including natural language processing, computer vision, and robotics, are also highlighted. Finally, the paper discusses the future directions and challenges of ZSL research, including the need for large-scale datasets, improved evaluation metrics, and more robust semantic representation methods. Despite facing* ***ob****stacles, recent advances in ZSL have shown promising results in enabling models to recognize and classify new objects or concepts without prior exposure to them during training.*

Key Words: Zero-shot learning, Machine learning, Deep learning, Semantic embedding, Computer Vision, Unsupervised Learning, Image Classification, Few Shot Learning, Natural Language Processing, Data Augmentation, Embeddings, Generative Models, Transfer Learning, Convolutional Neural Networks (CNN).

1.Introduction

Zero-shot learning (ZSL) is a subfield of machine learning that aims to recognize and classify new objects or concepts without prior exposure during training. Traditional supervised learning methods require large amounts of labeled data to train models, which can be costly and time-consuming. ZSL, on the other hand, provides an alternative solution to this issue by enabling models to generalize to new concepts via semantic representations. ZSL is based on the concept of transfer learning, which generalizes the knowledge acquired in related classes to new ones. This is accomplished by employing semantic representations, which capture the relationships between classes and enable the model to reason about the properties of unseen classes. There are two primary types of semantic representations: attribute-based and semantic space-based methods. Each class is described by a collection of attributes, such as color, size, and shape, which are used to reason about the properties of unseen classes. Classes are mapped to a high-dimensional space based on their semantic relationships, such as co-occurrence and WordNet hierarchy, using semantic space-based methods. ZSL has been implemented successfully in a variety of fields, including natural language processing, computer vision, and robotics. ZSL has been utilized in natural language processing for tasks such as sentiment analysis and text classification, in which the model must recognize new concepts not present in the training data. ZSL has been utilized in computer vision for tasks such as image classification and object detection, in which the model must recognize new objects or attributes not present in the training data. ZSL has been utilized in robotics for tasks such as object manipulation and grasping, in which the robot must recognize new objects not present in the training data.

Despite its success, ZSL still faces a number of obstacles that must be resolved. Lack of large-scale datasets with sufficient labeled data for the seen classes and a di- verse set of unseen classes is one of the greatest obstacles. Another obstacle is the need for improved methods of semantic representation that can capture the complex relationships between classes. In addition, the evaluation of ZSL models remains an unresolved issue, as traditional classification metrics may not be appropriate for evaluating the performance of ZSL models.

Recent developments in ZSL have yielded promising out-comes in addressing a number of these obstacles. For instance, generative models such as generative adversarial networks (GANs) and variational autoencoders (VAEs) have been utilized to generate synthetic data for unseen classes, thereby enhancing the performance of ZSL models. In addition, recent studies have proposed novel se- mantic representation methods, such as knowledge graph- based and graph neural network-based methods, that can capture the complexity of the relationships between classes.

In this paper, we provide a comprehensive overview of the current techniques in ZSL, including the challenges and recent developments. First, we present the fundamental concepts of ZSL and the various types of semantic representations. Then, we examine the most recent developments in ZSL, such as generative models, novel semantic representation methods, and multi-modal ZSL. Further- more, we discuss the potential applications of ZSL in a variety of fields, including natural language processing, computer vision, and robotics.

Finally, we highlight the future directions and challenges of ZSL research, including the need for large-scale datasets, improved evaluation metrics, and more robust semantic representation methods.

ZSL is a promising area of research that has the potential to enable models to recognize and classify new objects or concepts without prior exposure to them during training. Recent advances in ZSL have shown promising results in addressing some of these challenges, despite the persistence of obstacles. This review paper provides a thorough understanding of ZSL and highlights its potential applications and future research directions.

2. Types of Zero Shot Learning

2.1 Attribute Based ZSL

Attribute-based zero-shot learning (ZSL) is a type of machine learning technique that enables a model to recognize previously untrained objects or categories. In attribute-based ZSL, the model is trained on a set of known categories, but it can classify new categories using the attributes associated with each category during inference.

The attributes are a collection of semantic descriptions that describe the visual qualities or characteristics of an object or category. For instance, the attribute "has wings" could be associated with the category "bird," while "four legs" could be associated with the category "dog." During inference, the model receives as input a new image and predicts its category based on the image's attributes. Even if the model has never seen the category before, it can still make a prediction based on the category's attributes.

Attribute-based ZSL is useful in situations where obtaining labeled data for all possible categories may be difficult or expensive. Utilizing semantic information in the form of attributes, it allows models to generalize to new categories.

2.2 Semantic embedding-based zero-shot learning

Based on Semantic Embedding Semantic embedding-based zero-shot learning (ZSL) is a machine learning technique that enables models to recognize and classify unseen classes using a semantic embedding space. Each class is represented as a vector of attributes or features within the semantic embedding space.

In ZSL based on semantic embeddings, the model is trained on a set of known classes and their respective semantic embeddings. The model predicts the class of an unseen example during inference by projecting it onto the semantic embedding space and locating the closest class vector.

Using a similarity measure, such as cosine similarity, the distance between the example and the class vector is computed. The predicted class is the one that is closest to the instance.

Utilizing techniques like principal component analysis (PCA) or metric learning, the semantic embedding space is typically learned. By acquiring knowledge of the embedding space, the model can effectively generalize to new classes that were possibly never encountered during training.

ZSL has been applied to a variety of tasks, including image classification, text classification, and natural language processing, on the basis of semantic embedding. It is useful when the number of seen classes is limited and new, unseen classes must be identified and classified.

2.3 Generalized ZSL

Generalized zero-shot learning (ZSL) is a type of machine learning approach that extends traditional ZSL methods to handle a scenario in which some of the test classes have been observed during training, in addition to new, unobserved classes.

In generalized ZSL, the model is trained using a set of seen classes and their corresponding semantic embeddings, as well as a set of unseen classes for which no visual examples are available during training. During inference, the model is evaluated against both observed and unobserved classes.

Generalized ZSL aims to teach a model that can effectively recognize and classify both seen and unseen classes. To accomplish this, the model must learn to transfer knowledge from seen classes to unseen classes without overfitting the seen classes.

There are various strategies for generalizing ZSL, such as domain adaptation and hybrid methods. The objective of domain adaptation methods is to adapt the model's representation to unseen classes by utilizing information available from seen classes. Hybrid methods combine various ZSL techniques, such as attribute-based and semantic embedding-based ZSL, to improve performance for both seen and unseen classes.

Generalized ZSL is useful in situations where new classes must be identified and classified, but some similar classes have already been encountered in training. It enables the model to utilize the available knowledge from the seen classes to improve its performance on the unseen classes.

2.4 Multi-modal zero-shot learning

Multi-modal zero-shot learning (ZSL) is a type of machine learning technique that combines information from multiple modalities, such as images, text, and audio, to recognize and classify unseen classes.

In multi-modal ZSL, the model is trained using a collection of known classes, their corresponding semantic embeddings, and multiple modalities of data. During inference, the model predicts the class of an unseen example by analyzing its semantic embeddings and the various data modalities associated with it.

In image classification, for instance, the model may be trained on a set of classes, their corresponding image features, and textual descriptions. During inference, the model may receive an image and textual description of an unseen class. Using the image characteristics and textual description, the model can then predict the semantic embedding of the unseen class and classify the image accordingly.

Multi-modal ZSL is useful when multiple data modalities are available and each provides complementary information about the objects or classes being classified. By integrating data from multiple modalities, the model's ac- curacy and generalization performance on unseen classes can be enhanced.

Multiple tasks, including image classification, speech recognition, and natural language processing, have utilized multimodal ZSL. It is a potent method that can be utilized to solve a variety of real-world issues where data from multiple sources is available.

3. Challenges in Zero Shot Learning

The semantic gap between the visual and semantic domains presents one of the main difficulties in zero-shot learning. The generalization problem, which requires the model to apply to classes that have not yet been seen and may lack training samples, presents another difficulty. While some methods use semantic embeddings to fill the gap between the visual and semantic spaces, others transfer knowledge from seen to unseen classes.

3.1 Evaluation Metrics

Measurement Metrics: Zero-shot learning (ZSL) employs evaluation metrics to measure the performance of ZSL models in recognizing and classifying unseen classes. Several evaluation metrics are commonly employed in ZSL:

1. **Top-1 Accuracy:** This metric measures the proportion of correctly predicted labels for a given set of un- seen classes. The predicted label of the model must match the ground-truth label for the classification of an example to be considered correct.

2. **Top-5 Accuracy:** This metric measures the percentage of instances for which the ground-truth label is among the top five labels predicted by the model. In other words, the model receives credit for identifying the correct class, even if its prediction is not the most confident.

3. **Harmonic Mean (H):** This metric takes both the Top-1 and Top-5 accuracies into account and provides a more balanced evaluation of the model's performance. It is the harmonic mean of the Top-1 and Top-5 precisions.

4. **Area Under the Curve (AUC):** This metric mea- sures the accuracy of the model's ranking of unseen classes. It is calculated as the area beneath the re- ceiver operating characteristic (ROC) curve where the true positive rate is plotted against the false pos- itive rate for various threshold values.

5. **Mean Average Precision (MAP):** This metric measures the model's precision across all unseen classes on average. It is calculated as the mean of each class's average precision scores.

In ZSL, evaluation metrics are essential because they quantify the model's performance in recognizing and classifying unseen classes. Using these metrics, researchers are able to compare the performance of various ZSL mod- els and identify improvement opportunities.

3.2 Data Bias

Data bias in zero-shot learning (ZSL) is when the training data used to construct the model is not representative of the actual distribution of the classes being recognized and classified. This can result in inaccurate predictions and poor generalization performance, especially for classes that have not been observed.

ZSL data can be biased in multiple ways. For instance, if the training data is skewed toward certain classes, such as animals or vehicles, the model may perform poorly on classes from other domains, such as food or sports. Similarly, if the training data is biased toward particular regions or cultures, the model may perform poorly when applied to classes from other regions or cultures.

When semantic embeddings used to represent classes are biased or insufficient, data bias can also occur. For instance, if the semantic embeddings are derived from biased sources, such as a specific language or cultural context, the model may struggle to generalize to unobserved classes in other contexts.

To reduce the impact of data bias in ZSL, researchers must design and curate their training and evaluation datasets with care. This requires ensuring that the training data is representative of the distribution of the classes being recognized and classified in the real world, and that the semantic embeddings are exhaustive and objective.

In addition, data augmentation techniques can be used to artificially increase the diversity of the training data, while domain adaptation techniques can be used to adapt the model's representation to the unseen classes. Researchers can then carefully monitor the performance of their ZSL models on various subgroups of classes in order to identify potential biases and enhance the model's generalization performance.

3.3 Hybrid Methodologies

Hybrid approaches in zero-shot learning (ZSL) combine multiple methods or modalities to improve the ZSL models' precision and generalization performance. These methods combine data-driven and knowledge-based approaches to overcome the limitations of each and achieve superior results.

Combining data-driven approaches, such as deep learning models, with knowledge-based approaches, such as semantic embeddings or attributes, is a common hybrid strategy in ZSL. For instance, a ZSL model may be trained using a deep neural network on a large amount of visual data, and then a knowledge-based approach, such as at- tributes or semantic embeddings, may be used to map the visual features to the corresponding class labels.

Combining multiple modalities of data, such as visual and textual data, to improve the model's precision is another hybrid approach in ZSL. For instance, a ZSL model may be trained on both visual and textual data using a multi-modal deep learning architecture, and then generalize to unseen classes using a knowledge-based approach.

Hybrid approaches may also employ transfer learning strategies to transfer knowledge from related tasks to the ZSL task. For instance, a ZSL model may be pre- trained on a related task, such as image classification or object detection, and then fine-tuned using a smaller set of labeled examples for the ZSL task.

Overall, hybrid approaches in ZSL can leverage the strengths of different methods and modalities to improve the accuracy and generalization performance of ZSL models, and are an active area of machine learning research.

4. Zero Shot Learning with semantic output codes

Zero Shot Learning through the use of semantic output codes: Zero-shot learning (ZSL) with semantic output codes is a method for classifying unseen classes using semantic representations. In this method, both the seen and unseen classes' semantic information is used to create a semantic codebook, which represents the semantic relationships between classes and attributes.

During training, the model is taught to predict the se- mantic code for each observed class example. This example's semantic code is a binary vector that encodes the presence or absence of each attribute. The semantic code is a condensed representation of the visual characteristics of an example that captures the most pertinent information for classification.

Once the model has been trained, it can be used to classify unobserved examples by predicting their semantic codes. The semantic code of an unseen example is then utilized to retrieve the semantic codebook class that is most similar. The retrieved class is then used as the class label prediction for the unseen example.

ZSL with semantic output codes offers numerous benefits. First, it can handle both visible and invisible classes within a single framework, making it more applicable to real-world applications. Second, it can utilize the semantic relationships between classes and attributes to enhance the accuracy of the model and reduce the impact of noisy data. Lastly, it can handle the open-set problem by assigning examples that do not belong to any of the seen or unseen classes to a separate "unknown" class.

However, ZSL with semantic output codes has limitations as well. For instance, it requires semantic annotations for both the visible and invisible classes, which is not always possible. Moreover, it may be sensitive to the quality of the semantic representations and the selection of the semantic codebook.

5. Zero Shot Learning with visual attributes

Zero-shot learning (ZSL) with visual attributes is a classification technique that employs a set of semantic attributes to represent the visual characteristics of objects and classify them into unseen classes. Suppose, for instance, that we wish to categorize images of animals according to their species. To represent the visual characteristics of various animals, we can use visual attributes like "has wings," "has fur," "has a beak," etc.

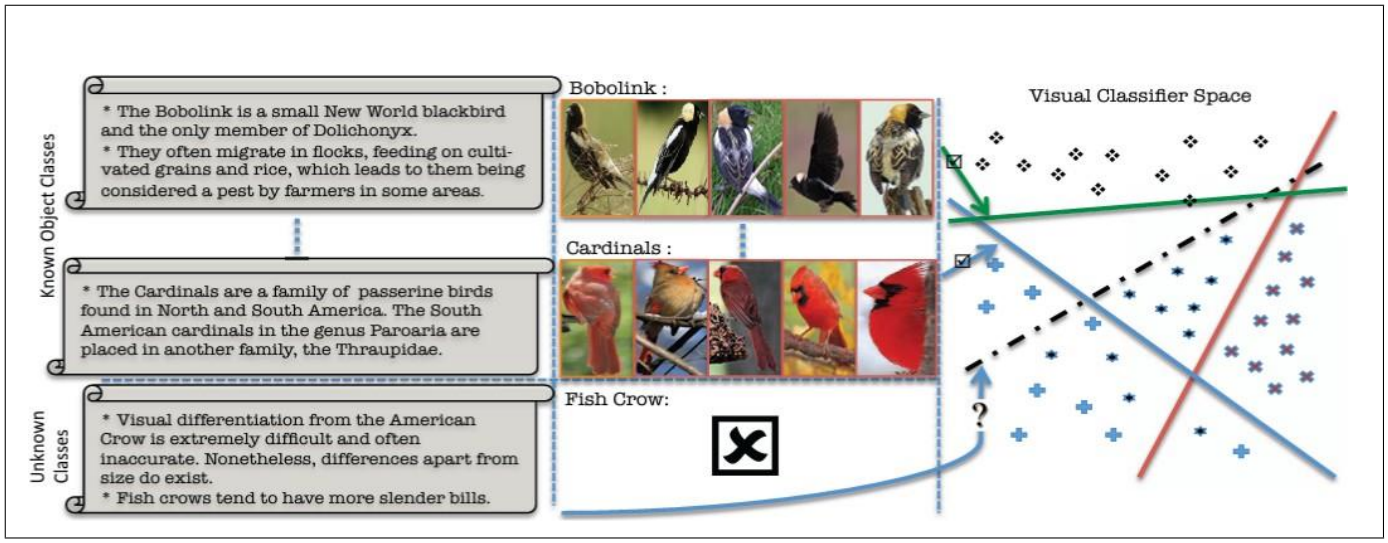

Figure 1: Zero-shot learning with textual description is a problem. Textual summaries for the various bird classes are shown on the left. Images for "seen classes" in the middle. Right: feature space classifier hyperplanes. To estimate a new classifier parameter using only a textual description is the objective.

During training, the model is taught to predict the attributes of each observed class example. The attributes are a binary vector encoding the presence or absence of each attribute for the provided example. The attribute vector is a condensed representation of the visual characteristics of the example that captures the most pertinent information for classification.

Once the model has been trained, it can be used to classify unobserved examples by predicting their attribute vector. The attribute vector of an unseen example is then utilized to retrieve the attribute space's most similar class. The retrieved class is then used as the class label prediction for the unseen example.

ZSL with visual attributes offers numerous benefits. First, it can handle both visible and invisible classes within a single framework, making it more applicable to real-world applications. Second, it can utilize the semantic relationships between attributes to enhance the accuracy of the model and reduce the impact of noisy data. Lastly, it can handle the open-set problem by assigning examples that do not belong to any of the seen or unseen classes to a separate "unknown" class.

However, ZSL with visual attributes has limitations as well. For instance, it requires accurate visual attribute annotations for both the visible and invisible classes, which is not always possible. In addition, it may be affected by the quality of the attribute annotations and the attribute space chosen.

6. Zero Shot Learning with Few Shot adaptation

Zero-shot learning (ZSL) with few-shot adaptation is a classification technique that combines the advantages of ZSL and few-shot learning to classify examples into unknown classes using limited labeled data. Suppose, for instance, that we wish to classify images of birds into various species but have only a few labeled examples for each species.

During training, both labeled examples from the seen classes and semantic representations of the unseen classes are used to train the model. Semantic representations may be attribute vectors or semantic embeddings. The model is then fine-tuned on a small number of labeled examples from the unseen classes, referred to as the support set, in order to accommodate the visual characteristics of the new classes.

After the model has been optimized, it can be used to classify examples from unseen classes by comparing them to the training set. In particular, for each unseen example, the model selects the semantically most similar examples from the training set. The class label for the unseen example is then predicted using the selected examples.

ZSL with few-shot adaptation offers numerous benefits. First, it can improve the model's accuracy by utilizing the semantic information of both the seen and unseen classes. Second, it can deal with the few-shot learning problem, in which there are few labeled examples for the unobserved classes, by adapting the model to the new classes with a small number of labeled examples. Lastly, it can handle the open-set problem by assigning examples that do not belong to any of the seen or unseen classes to a separate "unknown" class.

However, ZSL with few-shot adaptation has limitations as well. It requires accurate semantic representations for both the seen and unseen classes, which is not always possible. Additionally, it may be sensitive to the quality of the support set and the similarity metric chosen.

7. Zero Shot Learning for natural language processing

Zero-shot learning (ZSL) for natural language processing (NLP) is a technique that enables a machine learning model to predict unobserved tasks without the need for additional training data. Suppose, for instance, that we wish to train a model to classify movie reviews into different sentiment categories, such as positive or negative, without explicitly training on the specific task.

In ZSL for NLP, the model is trained on tasks that share semantic and linguistic characteristics with the unknown task. Frequently, these tasks are referred to as "auxiliary tasks." The model is trained to predict the labels of auxiliary tasks based on their inputs, such as text or audio data, during training.

Once the model has been trained, it can be utilized to make predictions on the unseen task by utilizing the semantic and linguistic features learned from the auxiliary tasks. If the unknown task is to classify movie reviews, for instance, we can encode the text inputs of the reviews into a semantic space that represents the text's meaning. The semantic space can be represented by a collection of semantic vectors, each of which corresponds to a particular sentiment category. The model can then predict the review's sentiment category by selecting the vector in the semantic space that is most similar.

ZSL for NLP has multiple benefits. First, it can reduce the quantity of labeled data required for training by utilizing the knowledge gained from related tasks. Second, it can deal with the zero-shot learning problem, which occurs when there are no labeled examples for the unseen task, by transferring the knowledge acquired from the auxiliary tasks. Lastly, it can manage the open-set problem, in which there may be examples that do not belong to any of the seen or unseen classes, by assigning them to a distinct "unknown" category.

8. Relevant Work

Our proposed work can be understood in terms of knowledge transfer and inductive learning. In general, the objective of knowledge transfer is to increase recognition by utilizing shared knowledge between classes. The majority of prior research focused on knowledge sharing exclusively within the visual domain, e.g. [12]; or on exporting semantic knowledge at the level of category similarities and hierarchies [10, 27]. Beyond the current state of the art, we investigate cross-domain knowledge sharing and transfer. We investigate how knowledge from the visual and textual domains can be utilized to discover cross-domain correlation, which facilitates the prediction of visual classifiers from textual description. Motivated by the practical need to learn visual classifiers for uncommon categories, researchers have investigated approaches for learning from a single image (one-shot learning [18, 9, 11, 2]) or even from no images (zero-shot learning). Recognizing object instances from previously unseen test categories (the zero-shot learning problem) can be accomplished, in part, by leveraging knowledge about shared attributes and components. Typically, an intermediate semantic layer is introduced to facilitate the sharing of knowledge between classes and the description of previously unseen classes, e.g. [22]. For example, given appropriately labeled training data, it is possible to learn classifiers for attributes occurring in training object categories. Then, these classifiers can be used to identify the same attributes in instances of objects from novel test categories. The recognition process can then proceed based on these acquired characteristics [17, 7]. These attribute-based "knowledge transfer" approaches utilize an intermediate visual attribute representation in order to describe categories of unseen objects. Typically, humans define attributes manually to describe shape, color, and surface material, e.g., furry, striped, etc. Therefore, an unseen category must be specified using the vocabulary of attributes currently in use. Rohrbach et al. [25] studied the extraction of valuable attributes from large text corpora. In [23], a method for interactively defining a vocabulary of attributes that are both human- readable and visually distinct was presented. Our work, in contrast, utilizes no explicit attributes. A new category's description is entirely textual. It has been investigated how linguistic semantic representations relate to visual recognition. In [4], it was demonstrated that there is a strong correlation between Word-Net-based semantic similarity between classes and confusion between classes. Large-scale image datasets, such as ImageNet[5] and Tiny Images [30], have utilized the linguistic semantics of WordNet nouns [19] to collect large-scale image datasets, such as ImageNet[5] and Tiny Images [30]. It has also been demonstrated that WordNet-based hierarchies are useful for learning visual classifiers, e.g. [27]. Barnard et al. [1] demonstrated that learning the joint distribution of words and visual elements facilitates semantic clustering of images, the generation of illustrative images from a caption, and the generation of annotations for new images. Recent research that focuses on generating textual descriptions of images and videos, such as [8, 16, 34, 14], has demonstrated a growing interest in the intersection between computer vision and natural language processing. This includes generating sentences about objects, actions, attributes, spatial relationships between objects, contextual information in the images, scene information, and so on. In contrast, our work differs fundamentally in two ways. We are not interested in generating textual descriptions from images; rather, we are interested in predicting classifiers from text in a zero- shot setting. In terms of the learning context, the textual descriptions we employ are at the level of the category and do not consist of image-caption pairs, as is common in datasets used for text generation from images, such as [21].

9. Problem Definition

Figure 1 depicts the learning environment. The information in our problem comes from two distinct domains, denoted by V and T, respectively: the visual domain and the textual domain. Similar to traditional visual learning problems, training data are given in the form V = (xi, li)N, where xi is an image and li 1 Nsc is its class label. The number of classes available at training is denoted by Nsc, where sc stands for "seen classes." As is customary in visual classification settings, we can learn Nsc binary one- versus-all classifiers for each of these classes. Consider a typical binary linear classifier in the feature space with the form fk(x) = cT k x, where x is the modified visual feature vector and ck Rdv is the linear classifier parameters for class k. Given a test image, l = arg maxk fk(x) deter- mines its class. Our objective is to be able to predict a classifier for a new category using only the learned classes and textual description(s) of that category. To accomplish this, the learning process must also incorporate a textual description of the observed classes (as depicted in Fig. 1). Depending on the domain, each class may have a few, a couple, or as few as one textual description. ti Tk denotes the textual training data for class k. In this paper, we assume we are dealing with the extreme case of having only one textual description per class, which ex- acerbates the difficulty of the problem. Nonetheless, the formulation we propose in this paper applies directly to the case of multiple textual descriptions per class. Similar to the visual domain, the unprocessed textual descriptions must undergo a process of feature extraction, which will be described in Section 5. Let's denote the extracted textual characteristic as T = tk Rdt k=1Nsc. Given a textual description t of a new unseen category C, the problem can be defined as predicting one-vs-all classifier parameters c(t) that can be used to directly classify any test image x as c(t). T x > 0 if x is a member of C c(t). T x 0 except if (1) Following the introduction of two potential frameworks for this problem and a discussion of their potential limitations, the proposed formulation is presented.

9.1. Regression Models

This problem can be formulated as a regression problem in which the objective is to use the textual data and the learned classifiers, (tk, ck)k=1Nsc, to learn a regression function from the textual feature domain to the visual classifier domain, i.e. a function c(): Rdt Rdv. Which regression model would be most appropriate for this problem? And would this approach to the problem yield reasonable results? A common regression model, such as ridge regression [13] or Gaussian Process (GP) Regression [24], learns the regressor to each dimension of the output domain (the parameters of a linear classifier) independently, i.e. a set of functions cj (): Rdt R. Obviously, this does not account for the relationship between the visual and textual domains. A structured prediction regressor, which would learn the correlation between the input and output domains, would be preferable. However, even a structure prediction model will only learn the correlation between the textual and visual domains from the input-output pairs (tk, ck). Here, the visual domain information is encapsulated in the pre- learned classifiers, and prediction lacks access to the original visual domain data. Instead, we must directly learn the correlation between the visual and textual domains and make predictions based on this information. The data points are the textual description-classifier pairs, and the number of classes is typically small compared to the dimension of the classifier space (i.e. Nsc dv). In such a context, it is inevitable that any regression model will suffer from an underfitting issue. This is best explained by GP regression, in which the predictive variance increases in regions of the input space devoid of data points. This will result in poor classification predictions in these regions.

9.2. Knowledge Transfer Models

Alternately, the problem could be stated as domain adaptation from the textual to the visual domain. In the context of computer vision, domain adaptation research has centered on transferring categories learned from a source domain with a given distribution of images to a target domain with a different distribution, such as images or videos from different sources [33, 26, 15, 6]. We require an approach that learns the correlation between the textual domain features and the visual domain features and then uses this correlation to predict a new visual classifier given the textual domain features. In particular, [15] introduced a method for learning cross- domain transformation. In this study, a regularized asymmetric transformation between two domains was discovered. The method was used to transfer learned categories between distinct visual data distributions. The source and target domains do not need to share the same feature spaces or dimensionality, which is an attractive feature of [15] over other domain adaptation models. We can formulate the zero-shot learning problem as a domain adaptation, inspired by [15]. This is possible through the acquisition of a linear (or nonlinear kernalized) transfer function W between T and V. By optimizing with a suitable regularizer over constraints of the form tTWx l if t T and x V belong to the same class, and tTWx u otherwise, the transformation matrix W can be learned. Here l and u are model parameters. This transfer function serves as a compatibility function between the textual and visual features, returning high values if they belong to the same class and low values if they belong to different classes. It is evident that this transfer function can serve as a classifier. Given a textual feature t and a test image represented by x, a classification decision can be obtained using the formula tT Wx b, where b is a decision boundary that can be set to (l + u)/2. Consequently, the desired predicted classifier in Equation 1 can be obtained as c(t) = tT W (note that the feature vectors have been replaced with ones). Due to the fact that W was learned using only seen classes, it is unclear how the predicted classifier c(t) will behave for unseen classes. There is no assurance that this

classifier will place all observed data on one side of the hyperplane and the new, unseen class on the opposite side.

10. Formulation of the Problem

10.1 Primary Objective

The proposed formulation aims to predict the hyperplane parameter c of a one-vs-all classifier for a new unseen class, given a textual description encoded by t and training-phase knowledge acquired from seen classes. Our solution architecture is depicted in Fig. Three elements are taught during the training phase: Classifiers: Individual one-vs-all classifiers ck are learned for each observed class. Probabilistic Regressor: Given (tk, ck), a regressor is discovered that can be used to provide a prior estimate for preg(c|t) (Details in Section 4.3). Transfer function for domains Given T and V, a domain transfer function, encoded in the matrix W, is learned to capture the relationship between the textual and visual domains (Details in Section 4.2). Each of these elements contains a subset of the problem's information. How can such knowledge be combined to predict a new classifier given a textual description? The newly developed classifier must be consistent with the observed classes. The new classifier must place all observed instances on one side of the hyperplane and adhere to the learned domain transfer function. The resulting constrained optimization problem is as follows: c^(t) =argmin c,i

cTc t TWc ln(preg(c|t)) + i s.t.: (cTxi) i, i 0, i = 1 ⋯ N t TWc l , , , l: hyperparameters (2)

The first term acts as a regularizer on top of the classifier c. The second term ensures that the predicted classifier is highly correlated with tT W. Given the prediction of the regressor, the third term favors a classifier with a high probability. The constraints cTxi i require all observed data instances to lie on the negative side of the predicted classifier hyperplane, allowing for some misclassification via the slack variables i. The constraint t TWc l ensures that the correlation between the predicted classifier and t TW is no less than l, thereby ensuring a minimum correlation between text and visual features.

10.2 Domain Transfer Functionality

To learn the domain transfer function W, we adapted the following strategy from [15]. Let T represent the textual feature data matrix, and let X represent the visual feature data matrix, with each feature vector replaced by a 1. Observe that adding a 1 to the feature vectors is necessary for our formulation, as we require tTW to function as a classifier. We must address the following optimization issue: min W r(W) + i ci(TWXT) (3) where ci's are loss functions over the constraints and r() is a regularizer matrix. The optimal W in Eq. 3 can be computed using inner products between data points in each domain separately, resulting in a kernalized non-linear transfer function; thus, its complexity is independent of the dimensions of either domain. The optimal answer to the equation 3 is W = TK 1 2 T LK 1 2 X XT, where KT = TTT and KX = XXT.

L is determined by minimizing the following minimization problem: min L [r(L) +].

p cp(K1 2 TLK1 2 X)], (4) where cp(K1 2 TLK1 2

X)=(max(0,(l eiK1 2 TLK1 2 Xej)))2 for same class pairs of index i,j, or = (max(0,(eiK1 2 TLK1 2 Xej u)))2 otherwise, where ek is a vector of zeros except a one at the kth element, and u>l (note any appropriate l, u could work. In this instance, we used l = 2 and u = 2). We utilized Frobenius norm regularization. A second-order BFGS quasi-Newton optimizer is used to minimize this energy. Using the aforementioned transformation, W is computed after L has been computed.

10.3. Probabilistic Regressor

There are numerous possible regressors, but we require one that provides a probabilistic estimate preg(c|(t)). For the reasons outlined in Section 3, we also require a structure prediction method that can predict all classifier dimensions simultaneously. We use the Twin Gaussian Process (TPG) [3] for these reasons. Utilizing Gaussian Process priors, TGP encodes the relationships between inputs and structured outputs. This is accomplished by minimizing the Kullback-Leibler divergence between the marginal GP of the outputs (in our case, classifiers) and observations (textual features). The solution to the following nonlinear optimization problem yields the estimated regressor output (c (t)) in TGP: [3] c (t) = argmin (t) [KC (c, c) 2kc(c) Tu log(KC (c, c) kc(c) T(KC + cI) 1kc(c))] (5) in which u = (KT + tI)1kt(t) and = KT (t, t) k(t)Gaussian kernels Tu, KT (tl, tm), and KC (cl, cm) for input feature t and output vector c. kc(c)=[KC (c, c1), ⋯ , KC (c, cNsc)]T. kt(t)=[KT (t, t1), ⋯ , KT (t, tNsc)] t and c serve as regularization parameters to prevent overfitting. This optimization problem is solvable using a second-order BFGS quasi- Newton optimizer with cubic polynomial line search for optimal step size selection [3]. In this case, the dimension of the classifier is predicted jointly. In this instance, preg(c|t) is a normal distribution. preg(c|t) = N (c = ~c(t), c = I) (6) TGP does not provide predictive variance, unlike Gaussian Process Regression, which explains why c = I. Nevertheless, it has the benefit of handling the dependence between the dimensions of the classifiers c and the textual features t. As a quadratic program, solving for c ln p(c|t) is a quadratic term in c that has the form ln p(c|t) (c c (t)) according to the TGP definition of preg(c|t).T(c c~(t)) = cTc 2cTc~(t)+~c(t) Tc~(t) (7) We reduce ln p(c|t) to 2cTc (t) because 1) c (t)Tc (t) is a constant (and thus has no effect on the optimization) and2) cTc is already included as a regularize in equation 2.

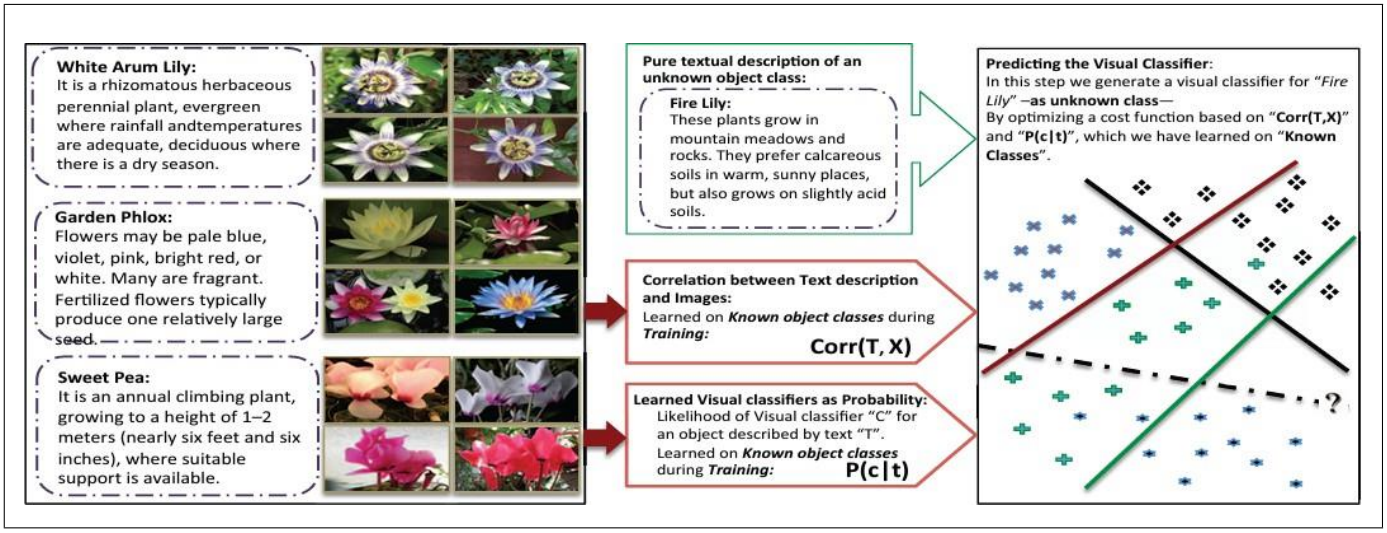

Figure 2: Illustration of the Proposed Solution Framework for the task Zero-shot learning from textual description.

In our context, the dot product is the superior measure of similarity between two hyperplanes. Consequently, 2cTc

(t) is minimized. Equation 2 reduces, given ln p(c|t) from the TGP and W, to a quadratic program on c with linear constraints. We tried various quadratic solvers, but the IBM CPLEX solver 2 provides the best speed and optimization for our problem.

11. Experiments

11.1. Datasets

This approach was tested using the CU200 Birds [32] (200 classes - 6033 images) and the Oxford Flower- 102 [20] (102 classes - 8189 images) image datasets, as they are among the largest and most widely used fine-grained datasets. Descriptive text were produced for each class in both datasets. The CU200 Birds image dataset was created based on birds with corresponding Wikipedia articles, so we created a tool to automatically extract Wikipedia articles given the class name. The tool successfully generated 178 articles automatically, while the remaining 22 were extracted manually from Wikipedia. Only when the article title is a different synonym for the same bird class do these mismatches occur. In contrast, the Flower image dataset was not created using the same criteria as the Bird image dataset; therefore, not all classes of the Flower dataset have corresponding Wikipedia articles. The tool was able to generate approximately 16 classes from Wikipedia out of 102, while the remaining 86 classes required manual generation of articles from Wikipedia, Plant Database 3, Plant Encyclopedia 4, and BBC articles 5. We intend to provide the extracted textual description as enhancements to these datasets. There is a sample textual description in the supplementary materials.

11.2. Extracting Textual Characteristics

Textual characteristics were extracted in two phases, as is typical in the literature on document retrieval. The initial phase is an indexing phase that produces textual features with tfidf (Term Frequency-Inverse Document Frequency) configuration (Term frequency as local weighting and inverse document frequency as global weighting). The tfidf is a measure of a word's significance within a corpus of text. The tfidf value increases proportionally with the number of times a word appears in the document, but is counterbalanced by the frequency of the word in the corpus, which helps to account for the fact that some words are more prevalent than others. Term's normalized frequency in the provided textual description was used [29]. The inverse document frequency is a measure of a term's frequency; in this study, the standard logarithmic idf [29] was utilized. In the second phase, the Clustered Latent Semantic Indexing (CLSI) algorithm [35] is used to reduce dimensionality. For document retrieval, CLSI is a low-rank approximation method for dimensionality reduction. In the Flower Dataset, tfidf features R8875, and final textual features R102 after CLSI. In the Birds Dataset, tfidf features are located in R7086, and after CLSI, textual features are located in R200.http://plants.usda.gov/java/
4http://www.theplantencyclopedia.org/wiki/Main Page
5http://www.bbc.co.uk/science/0/

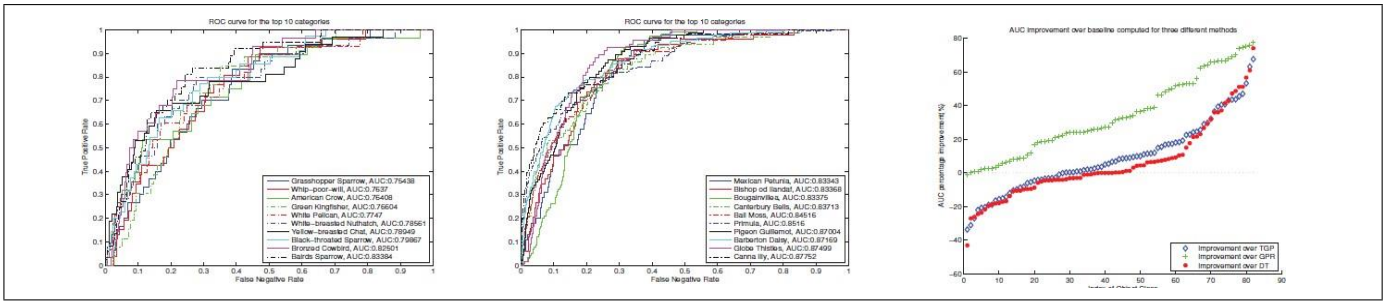

Figure 3: Zero-shot learning with textual description is a problem. Textual summaries for the various bird classes are shown on the left. Images for "seen classes" in the middle. Right: feature space classifier hyperplanes. To estimate a new classifier parameter using only a textual description is the objective.

11.3. Visual features

Classeme features [31] were utilized as the visual feature in our experiments as they provide an intermediate semantic representation of the input image. Classeme features are the output of a set of classifiers corresponding to a set of C category labels, which are drawn from a term list defined in [31] and are unrelated to our textual features. For each category c 1 C, a set of training images is collected by querying an image search engine with the category label. Following the extraction of a set of coarse feature descriptors (Pyramid HOG, GIST, etc.), a subset of feature dimensions was chosen [31] and a one-versus-all classifier c was trained for each category. The output of the classifier is real-valued and is such that if c(x) > c(y), then x is more similar to class c than y. The feature vector (descriptor) used to represent an image x is the classeme vector [1(x),), C (x)]. The Classeme characteristic has a dimension of 2569.

11.4 Methodology for Evaluating Experimental Results and Metrics:

Similar to zero shot learning literature, we evaluated the performance of an unseen classifier in a one-vs-all setting where the test images of unseen classes are regarded as the positives and the test images of seen classes as the negatives. We have computed the ROC curve and reported the area under the curve (AUC) as a comparative metric for various approaches. In a zero-shot learning environment, test data from the seen class are typically much larger than those from the unseen class. This renders other metrics, such as accuracy, useless, as high accuracy can be achieved even if all unseen class test data are incorrectly classified; therefore, we utilized ROC curves, which are not affected by this issue. In each fold, four-fifths of the classes were deemed "seen classes"

Table 1: Comparative Evaluation on the Flowers and Birds

baseline	Flowers (102) % improvement	Birds (200) % improvement
GPR	100%	98.31%
TGP	66%	51.81%
DA	54%	56.5%

Table 2: Percentage of classes that the proposed approach makes an improvement in predicting over the baselines (relative to the total number of classes in each dataset

Approach	Flowers Avg AUC (+/- std)	Birds Avg AUC (+/- std)
GPR	0.54 (+/- 0.02)	0.52 (+/- 0.001)
TGP	0.58 (+/- 0.02)	0.61 (+/- 0.02)
DA	0.62(+/- 0.03)	0.59 (+/- 0.01)
Approach	0.68 (+/- 0.01)	0.62 (+/- 0.02)

and used for training, while one-fifth of the classes were deemed "unseen classes" and their classifiers were predicted and evaluated. Within each of these class-folds, the observed classes' data are further divided into training and test sets. The approach's hyper-parameters were determined using an additional five-fold cross validation within the class-folds (i.e., the 80 Baselines: Since this work is the first to predict classifiers based solely on textual description, there are no previously published results with which to compare it. However, we designed three state-of-the-art benchmarks for comparison, which are intended to be consistent with our argument in Section

class	TGP (AUC)	DA (AUC)	Our (AUC)	%Improv.
2	0.51	0.55	0.83	57%
28	0.52	0.54	0.76	43.5%
26	0.54	0.53	0.76	41.7%
81	0.52	0.82	0.87	37%
37	0.72	0.53	0.83	35.7%

Table 3: Top-5 classes with highest combined improvement in Flower dataset

3. Specifically, we used: 1) A Gaussian Process Regressor (GPR) [24], 2) Twin Gaussian Process [3] as a structured regression method, and 3) Nonlinear Asymmetric Domain Adaptation [15]. Due to the fact that our formulation utilizes the TGP and DA baselines, it is essential to determine whether or not the formulation outperforms them. Notably, we also evaluate TGP and DA as alternative formulations for the problem, neither of which has been used in the same context previously. Results: Table 1 displays the average AUCs for the proposed method versus the three baselines for both datasets. As expected, GPR performed poorly in all classes of both data sets, as it is not a structure prediction method. The DA formulation performed slightly better than TGP on the Flower dataset, but slightly worse on the Bird dataset. On both datasets, the proposed method outperformed all baselines, with a significant difference on the flower dataset. It is also evident that the TGP performed better on the Bird dataset due to the larger number of classes (more points used for prediction). Fig. 3 displays the ROC curves for our method based on the best predicted unseen classes from the Birds dataset on the left and the Flower dataset in the center. Figure 4 depicts the AUC for each class in the Flower dataset. Additional findings are included in the supplementary materials. On the right side of Figure 3 is the improvement over the three baselines for each class, calculated as (our AUC- baseline AUC)/baseline AUC Table 2: Percentage of classes that the proposed method outperforms the baselines in predicting (relative to the total number of classes in each dataset). The dataset where our method produced the greatest average improvement. The purpose of this table is to demonstrate that both TGP and DA performed poorly in these instances, whereas our formulation based on both performed significantly better. This demonstrates that our formulation is not merely a combination of the two best approaches, but can significantly improve prediction performance. To evaluate the effect of the constraints in the objective function, we eliminated the constraints (cTxi) i, which attempts to ensure that all observed examples are on the negative side of the predicted classifier hyperplane, and evaluated the approach. With the constraints, the average AUC for the flower dataset (using a single fold) decreased to 0.59 from 0.61. Likewise, we assessed the impact of the constraint tT Wc l. The average AUC was reduced to 0.58 from 0.65 as a result of the constraint. This demonstrates the significance of this formulation constraint.

12. Conclusion and Future Activities

We investigated the problem of predicting visual classifiers from a textual description of classes in the absence of training images. We investigated and tested various formulations of the problem within the context of fine-grained categorization. We proposed a novel formulation that captures information between the visual and textual domains by transferring knowledge from textual features to visual features, which leads indirectly to the prediction of the visual classifier described in the text. Rather than using linear classifiers in the future, we intend to propose a kernel-based solution to the problem. In addition, we will investigate the prediction of classifiers based on complex- structured textual features.

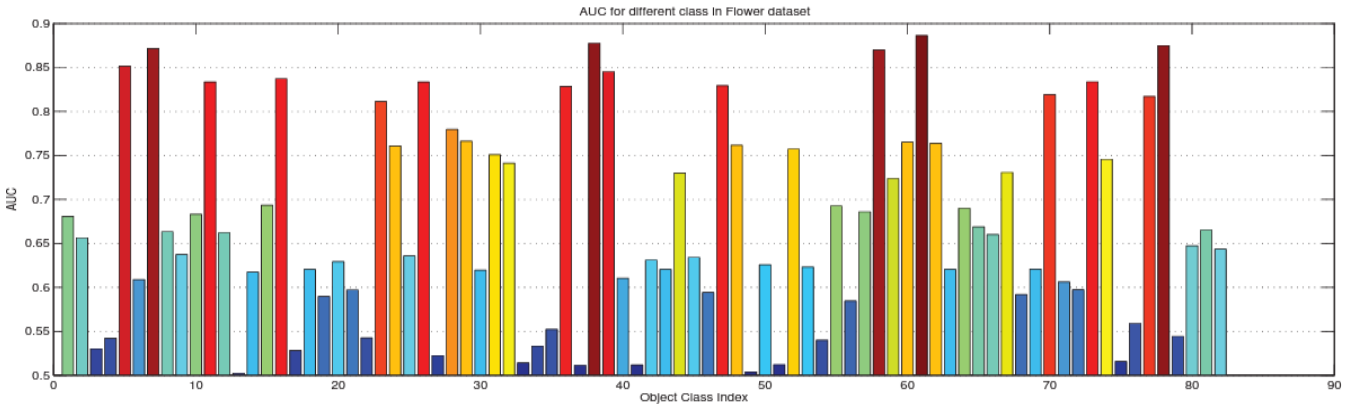

Figure 4: AUC of the predicated classifiers for all classes of the flower datasets.

Refernces

[1] K. Barnard, P. Duygulu, and D. Forsyth. Clustering art. In CVPR, 2001. 2

[2] E. Bart and S. Ullman. Cross-generalization: Learning novel classes from a single example by feature replacement. In CVPR, 2005. 1, 2

[3] L. Bo and C. Sminchisescu. Twin gaussian processes for structured prediction. IJCV, 2010. 5, 6

[4] J. Deng, A. C. Berg, K. Li, and L. Fei-Fei. What does classifying more than 10,000 image categories tell us? In ECCV. 2010. 1, 2

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei- Fei. Imagenet: A large-scale hierarchical image database. In CVPR, 2009. 2

[6] L. Duan, D. Xu, I. W.-H. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. TPAMI, 2012. 3

[7] A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth. Describing objects by their attributes. In CVPR, 2009. 1, 2

[8] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young,C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images.In ECCV. 2010. 3

[9] L. Fe-Fei, R. Fergus, and P. Perona. A bayesian approach to unsupervised one-shot learning of object categories. In CVPR, 2003. 1, 2

[10] R. Fergus, H. Bernal, Y. Weiss, and A. Torralba. Semantic label sharing for learning with many categories.In ECCV. 2010. 2

[11] M. Fink. Object classification from a single example utilizing class relevance metrics. In NIPS, 2004. 2

[12] G. Griffin and P. Perona. Learning and using taxonomies for fast visual categorization. In CVPR, 2008. 2

[13] A. E. Hoerl and R. W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 1970. 3

[14] N. Krishnamoorthy, G. Malkarnenkar, R. Mooney, K. Saenko, U. Lowell, and S. Guadarrama. Generating natural-language video descriptions using text-mined knowledge. NAACL HLT, 2013. 3

[15] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In CVPR, 2011. 3, 4, 5, 6

[16] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In CVPR, 2011. 3

[17] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by betweenclass attribute transfer. In CVPR, 2009. 1, 2

[18] E. G. Miller, N. E. Matsakis, and P. A. Viola. Learning from one example through shared densities on transforms. In CVPR, 2000. 2

[19] G. A. Miller. Wordnet: A lexical database for english. COMMUNICATIONS OF THE ACM, 1995. 2

[20] M.-E. Nilsback and A. Zisserman. Automated flower classification over large number of classes. In ICVGIP, 2008. 2,6

[21] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs.In NIPS, 2011. 3

[22] M. Palatucci, D. Pomerleau, G. E. Hinton, andT. M. Mitchell. Zero-shot learning with semantic output codes. In NIPS, 2009. 2

[23] D. Parikh and K. Grauman. Interactively building a discriminative vocabulary of nameable attributes. In CVPR, 2011. 2

[24] C. E. Rasmussen and C. K. I. Williams. Gaussian Processes for Machine Learning. The MIT Press, 2005. 3, 6

[25] M. Rohrbach, M. Stark, G. Szarvas, and B. Schiele. Combining language sources and robust seman- tic relatedness for attribute-based knowledge transfer. In Parts and Attributes Workshop at ECCV, 2010. 2

[26] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In ECCV. 2010. 3

[27] R. Salakhutdinov, A. Torralba, and J. B. Tenen- baum. Learning to share visual appearance for multiclass object detection. In CVPR, 2011. 1, 2

[28] B. Saleh, A. Farhadi, and A. Elgammal. Object- centric anomaly detection by attribute-based reasoning. In CVPR, 2013. 1

[29] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. IPM, 1988. 6

[30] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. PAMI, 2008. 2

[31] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. In ECCV, 2010.

[32] [32] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical report, California Institute of Technology, 2010. 2, 6

[33] J. Yang, R. Yan, and A. G. Hauptmann. Cross-domain video concept detection using adaptive svms. In MULTIMEDIA, 2007. 3

[34] Y. Yang, C. L. Teo, H. Daum´e III, and Y. Aloimonos. Corpus-guided sentence generation of natural images. In EMNLP, 2011. 3

[35] D. Zeimpekis and E. Gallopoulos. Clsi: A flexible approximation scheme from clustered term-document matrices. In In SDM, 2005. 6

[36] Research progress of zero-shot learning: Xiaohong Sun, Jinan Gu, Hongying Sun - Applied Intelligence 51, 3600-3614, 2021.

[37] Chao, WL., Changpinyo, S., Gong, B., Sha, F. (2016). An Empirical Study and Analysis of Generalized Zero-Shot Learning for Object Recognition in the Wild. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science(), vol 9906. Springer, Cham. https://doi.org/10.1007/978-3-319-46475-6$_4$

[38] F. Pourpanah et al., "A Review of Generalized Zero-Shot Learning Methods," in IEEE Transac- tions on Pattern Analysis and Machine Intelligence, vol. 45, no. 4, pp. 4051-4070, 1 April 2023, doi: 10.1109/TPAMI.2022.3191696.

[39] A review of generalized zero-shot learning methods: Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi - Zhao Wang, QM Jonathan Wu - IEEE transactions on pattern analysis and machine intelligence, 2022.

[40] Distinguishing unseen from seen for generalized zero-shot learning: Hongzu Su, Jingjing Li, Zhi Chen, Lei Zhu, Ke Lu - Proceedings of the IEEE/CVF Con- ference on Computer Vision and Pattern Recognition, 7885-7894, 2022.

[41] A survey of zero-shot learning: Settings, methods, and applications: Wei Wang, Vincent W Zheng, Han Yu, Chunyan Miao - ACM Transactions on Intelligent Systems and Technology (TIST) 10 (2), 1- 37, 2019.

[42] Contrastive embedding for generalized zero-shot learning: Zongyan Han, Zhenyong Fu, Shuo Chen, Jian Yang - Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2371- 2381, 2021.

[43] Open world compositional zero-shot learning: Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, Zeynep Akata - Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 5222-5230, 2021.

[44] An embarrassingly simple approach to zero-shot learning: Bernardino Romera-Paredes, Philip Torr - International conference on machine learning, 2152-2161, 2015.

[45] Zero-shot learning-the good, the bad and the ugly: Yongqin Xian, Bernt Schiele, Zeynep Akata - Proceedings of the IEEE conference on computer vision and pattern recognition, 4582-4591, 2017.

[46] Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly: Yongqin Xian, Christoph H Lampert, Bernt Schiele, Zeynep Akata- IEEE transactions on pattern analysis and machine intelligence 41 (9), 2251-2265, 2018.

[47] Synthesized classifiers for zero-shot learning: Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, Fei Sha - Proceedings of the IEEE conference on computer vision and pattern recognition, 5327-5336, 2016.

[48] Unsupervised domain adaptation for zero-shot learning: Elyor Kodirov, Tao Xiang, Zhenyong Fu, Shaogang Gong - Proceedings of the IEEE international conference on computer vision, 2452- 2460, 2015.

[49] Multi-modal cycle-consistent generalized zero- shot learning: Rafael Felix, Ian Reid, Gustavo Carneiro - Proceedings of the European Conference on Computer Vision (ECCV), 21-37, 2018.

[50] Zero-shot learning via visual abstraction: Stanis- law Antol, C Lawrence Zitnick, Devi Parikh - Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13, 401-416, 2014.

[51] Knowledge-aware zero-shot learning: Survey and perspective: Jiaoyan Chen, Yuxia Geng, Zhuo Chen, Ian Horrocks, Jeff Z Pan, Huajun Chen - arXiv preprintarXiv:2103.00070, 2021.

[52] Semantic-guided multi-attention localization for zero-shot learning: Yizhe Zhu, Jianwen Xie, Zhiqiang Tang, Xi Peng, Ahmed Elgammal - Advances in Neural Information Processing Systems 32, 2019.