Intelligence-Based Medicine 3-4 (2020) 100005

Contents lists available at ScienceDirect

Intelligence-Based Medicine

journal homepage: www.journals.elsevier.com/intelligence-based-medicine

Zero-shot learning and its applications from autonomous vehicles to COVID-19 diagnosis: A review

Mahdi Rezaei [a,*,1], Mahsa Shahidi [b,2]

[a] Institute for Transport Studies, The University of Leeds, Leeds LS2 9JT, United Kingdom
[b] Faculty of Computer and IT Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran

ARTICLE INFO

Keywords:
COVID-19 pandemic
SARS-CoV-2
Chest X-Ray (CXR)
Zero-shot learning
Deep learning
Semantic embedding
Machine learning
Autonomous vehicles
Supervised annotation

ABSTRACT

The challenge of learning a new concept, object, or a new medical disease recognition without receiving any examples beforehand is called Zero-Shot Learning (ZSL). One of the major issues in deep learning based methodologies such as in Medical Imaging and other real-world applications is the requirement of large annotated datasets prepared by clinicians or experts to train the model. ZSL is known for having minimal human intervention by relying only on previously known or trained concepts plus currently existing auxiliary information. This is ever-growing research for the cases where we have very limited or no annotated datasets available and the detection / recognition system has human-like characteristics in learning new concepts. This makes the ZSL applicable in many real-world scenarios, from unknown object detection in autonomous vehicles to medical imaging and unforeseen diseases such as COVID-19 Chest X-Ray (CXR) based diagnosis. In this review paper, we introduce a novel and broaden solution called Few / one-shot learning, and present the definition of the ZSL problem as an extreme case of the few-shot learning. We review over fundamentals and the challenging steps of Zero-Shot Learning, including state-of-the-art categories of solutions, as well as our recommended solution, motivations behind each approach, their advantages over each category to guide both clinicians and AI researchers to proceed with the best techniques and practices based on their applications. Inspired from different settings and extensions, we then review through different datasets inducing medical and non-medical images, the variety of splits, and the evaluation protocols proposed so far. Finally, we discuss the recent applications and future directions of ZSL. We aim to convey a useful intuition through this paper towards the goal of handling complex learning tasks more similar to the way humans learn. We mainly focus on two applications in the current modern yet challenging era: coping with an early and fast diagnosis of COVID-19 cases, and also encouraging the readers to develop other similar AI-based automated detection / recognition systems using ZSL.

1. Introduction

Object recognition is one of the highly researched areas of computer vision. Recent recognition models have led to great performance through established techniques and large annotated datasets. After several years of research, the attention over this topic has not only dimmed but it has been proven that there are still ways and rooms to refine models to eliminate existing issues in this area. The number of newly emerging unknown objects are growing. Some examples of these unseen or rarely-seen objects are futuristic object designs like the next generation of concept cars, other existing concepts but with restricted access to them (such as licensed or private medical imaging datasets), or rarely seen objects (such a traffic signs with graffiti on them), or fine-grained categories of objects (such as detection of COVID-19 in comparison with the easier task of detecting a common pneumonia). This brings the necessity of developing a fresh way of solving object recognition problems that concern lesser human supervision and lesser annotated datasets. Several approaches have tried to gather web images to train the developed deep learning models, but aside from the problem of the noisy images, the searched keywords are still a form of human supervision. One-Shot learning (OSL) and Few-shot learning (FSL) are two solutions that are able to learn new categories via one or a few images, respectively [70,76,104].

* Corresponding author. Institute for Transport Studies, The University of Leeds, 34-40 University Road, Woodhouse Lane, Leeds, LS2 9JT, UK.
E-mail address: m.rezaei@leeds.ac.uk (M. Rezaei).
[1] Assistant Professor in Computer Vision and Machine Learning; University Academic Fellow at the University of Leeds, United Kingdom; Honorary Academic Staff at Auckland University of Technology; and Senior Member of CeRV Research Centre, Auckland, New Zealand.
[2] Researcher at SYNTECH Technology & Innovation Centre, Qazvin, IR.

https://doi.org/10.1016/j.ibmed.2020.100005
Received 20 June 2020; Received in revised form 17 September 2020; Accepted 24 September 2020

(a) Positive COVID-19

(b) Axial chest CTs with visible GGO patches, consolidation, and parenchymal thickening

Fig. 1. Posterior-Anterior (PA)/Anterior-Posterior (AP) Chest X-Rays and the corresponding CT images of COVID-19 patients.

Natural language processing (NLP) is another major area of research in AI, and the application of Few-shot learning in the integration of NLP and object recognition has become a hot topic recently. [165] was the first FSL-based model to improve the performance of an NLP system. Zero-shot learning (ZSL) [7,38,80,178,188] is an emerging research which is completely free of any laborious task of data collection and annotation by experts. Zero-shot learning is a novel concept and learning technique without accessing any exemplars of the unseen categories during training, yet it is able to build recognition models with the help of transferring knowledge from previously seen categories and auxiliary information. The auxiliary information may include textual description, attributes, or vectors of word labels. This means the ZSL is interdisciplinary by nature with two inseparable components of visual and textual data.

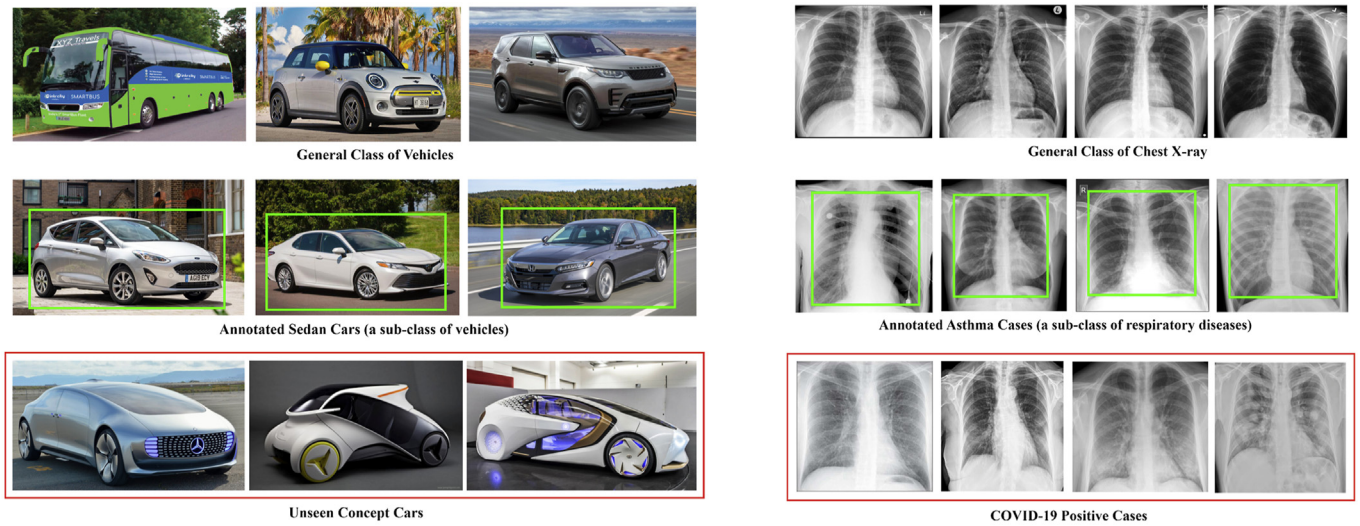
One of the interesting facts about ZSL is its similarity with the way human learns and recognises a new concept without seeing them beforehand. For example, a ZSL-based model would be able to automatically learn and diagnose COVID-19 patients, based on the existing chest X-ray images of patients with asthma and lungs inflammatory diseases which are already recognised and labelled by clinicians, plus some new auxiliary information about the COVID-19 attributes. Here, the auxiliary data can be the description of physicians and clinicians about the unique type of visual patterns, features, damages, or differences they have noticed on the Chest X-ray of patients with positive COVID-19 comparing to asthma X-ray images. A similar concept or approach is applicable in autonomous vehicles [132], where a self-driving car is responsible for automatic detection of surrounding cars including e.g. an unseen Tesla concept car based on the subgroup of labelled classic sedan cars plus auxiliary information about the common differences of concept cars than the classic cars; or recognising a Persian deer, based on the auxiliary information available for it and its appearance similarities or differences with other previously known deer. For instance, it belongs to a subgroup of the fallow deer, but with a larger body, bigger antlers, white spots around the neck, and also flat antlers for the male type.

Fig. 1(a) shows three examples of Posterior-Anterior (PA) and AP projection of chest X-rays of positive cases of COVID-19, and Fig. 1(b) represents their corresponding axial CT scans, taken from the COVID-ChestXRay dataset [27]. As it can be seen in the images, common evident anomalies may include unilateral or bilateral patchy ground-glass opacities (GGOs), patchy consolidations and parenchymal thickening. The goal of this research is to build an artificial intelligence based-model that can diagnoses COVID-19 without providing any visual exemplars in the training phase. In that case, the side (auxiliary) information should be provided to assist diagnosis in the test phase. In Fig. 2, the auxiliary information is provided in the form of textual descriptions for two examples of concept cars and COVID-19 X-rays. In Fig. 2(a) we aim at distinguishing new unseen concept cars (bottom row), using description on the exterior of the target and how it differs an already learned car from existing classic vehicle classification system such as in [133]. Similarly, visual differences and similarities between healthy Chest X-rays, Asthma cases, and COVID-19 positive cases are described in Fig. 2(b) as the auxiliary information.

Let's assume our pre-trained AI-based medical imaging system is capable of detecting Asthma cases, based on common deep learning techniques using a previously large dataset of labelled Asthma Chest X-ray images. However, these days we are facing an unknown COVID-19 pandemic with very limited annotated Chest X-rays. Obviously, we can not proceed on the same way of training traditional deep-learning methods, due to very sparse labelled images for COVID-19. The good point is that our medical experts and clinicians can provide some auxiliary information (textual descriptions) about common features and similarities among the COVID-19 positive chest X-rays to infer their findings. In Fig. 3, the side information is provided in form of what *"attributes"*: such as foggy effects, white spot features, blurred edges, and white/low-intensity pixel dominance in various areas of the chest X-ray images of COVID-19 patients.

Our idea behind the utilisation of ZSL models is to detect, understand, and recognise new concepts using an existing similar deep-learning based classifier, plus the integration of auxiliary information. This turns it to a completely new and efficient detector/recogniser or diagnosing system without the requirement of collecting a new dataset and a vast amount of costly and time-consuming labelling, especially when a speedy solution is crucial and life-saving; such as the recent global pandemic. In this research we will have four main contributions, as follows:

- We propose to categorise the reviewed approaches based on the embedding spaces that each model uses to learn/infer unseen objects/concepts as well as describing the variations to the data embedding inside those embedding spaces (Fig. 3 and Table 1).
- We evaluate the performance of the state-of-the-art models on famous benchmark datasets (Tables 3–5, Fig. 4). To the best of our knowledge, we are the first to include the evaluation of data-synthesising methods in the research field of applied Zero-shot learning.
- We study the motivation behind leveraging each space as a way to solve the ZSL challenge by reviewing current issues and solutions to them.



(a) Concept cars auxiliary information: *"The body of the car has a singular and unified shape with smoother curves. The wheels' colour, curves, and design match the body as a singular integrated piece. LED lights are omnipresent all around the car."*

(b) COVID-19 X-ray auxiliary information: *"Bilateral multifocal patchy GGOs and consolidation can be seen. Edges are blurred and the intensity sharpness of both lungs have decreased."*

Fig. 2. Similarities and differences between seen and unseen examples derived from textual descriptions and train and test images. The test images are concept cars (a) and COVID-19 symptoms (b).

- We provide sufficient technical justifications to support the ideas of using the proposed ZSL model as one of the best practices for COVID-19 diagnosis and other similar applications.

The rest of the materials in the article is organised as follows. In Section 2, we introduce the problem of Few-shot, One-shot and Zero-shot learning. In Section 3, we discuss about the test and train phases of the Zero-shot learning and generalised Zero-shot learning systems. Section 4 provides with embedding approaches followed by evaluation protocols in Section 5. In Section 6, we analyse the outcome of the experiments performed on different state-of-the-art methodologies. Further discussion about the applications of ZSL is investigated in Section 7. In Section 8, we discuss the outcome of this research, and finally, the concluding remarks in Section 9.

2. Few-shot/one-shot and zero-shot learning

Few-shot learning (FSL) is the challenge of learning novel classes with a tiny training dataset of one or a few images per category. FSL is closely related to knowledge transfer where a model, previously trained on large data, is used for a similar task with fewer training data. The more the transferred knowledge is accurate, the better FSL will generalise. Moreover, many approaches employ meta-learning to learn the challenge of few-shot or few-example learning [64,156]. The main challenge is to improve the generalisation ability as it often faces the overfitting problem.

In this type of learning, there is an auxiliary dataset that contains N classes each having K annotated samples of the new examples in the training phase. This makes the problem into a N-way-K-shot classification:

$$D_s = \{(x_i, y_i)\}_{i=1}^{N_t} \tag{1}$$

where x_i is the i^{th} training example and y_i is its corresponding label. $N_t = K \times N$ denotes the number of N categories and K defines the number of examples. Few-shot learning has $K > 1$ samples.

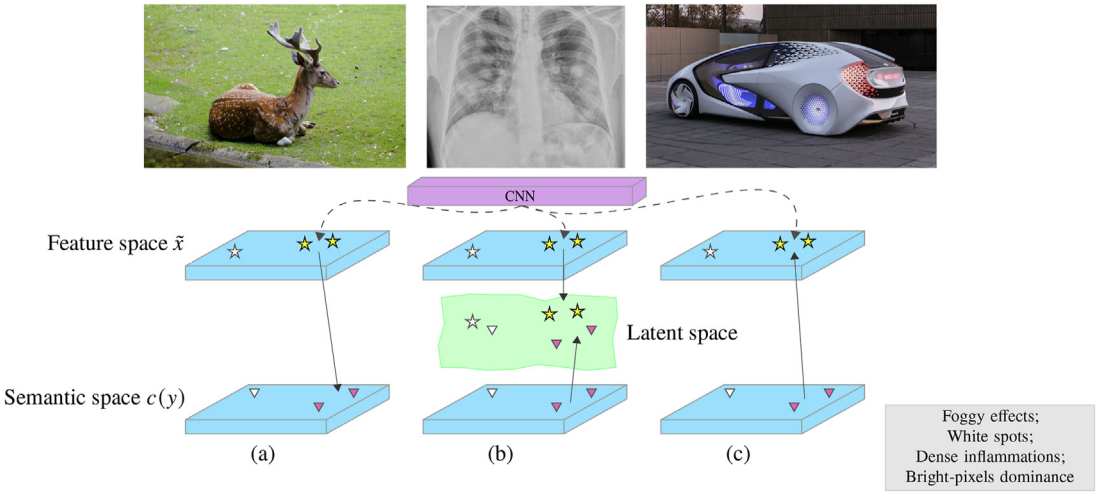


Fig. 3. Overview of ZSL models. Typical approaches use one of the three embedding types or a combination of them. (a) Semantic embedding models that map visual features to the semantic space. (b) Models that map visual and semantic features to an intermediate latent space. (c) Visual embedding models that map semantic features to the visual space.

Among the relevant research works, [163] uses the shared features among classes to compensate for the requirement for the large data, and follows a learning procedure based on boosted decision stumps. HDP-DBM [141] develops a compound of a deep Boltzmann machine and a hierarchical Dirichlet process to learn the abstract knowledge at different hierarchies of the concept categories. [156] Proposes prototypical networks that computes Euclidean distance between prototype representations of each class. It was not until recently that Few-shot learning was introduced in computer-aided diagnosis. For the first time, the idea of using additional information (attributes) in FSL, was introduced in [165]. [121] proposes a model to classify skin lesions. [68] uses FSL for Glaucoma Diagnosis from fundus images. [127] studies the problem of chest X-ray classification of five symptoms including Consolidation.

In the case of one-shot learning, there is only $K = 1$ example per class in the supporting set, thus it faces more challenge in comparison to the FSL. Bayesian Program Learning (BPL) framework [77] presents each concept of the handwritten characters as a simple probabilistic program. [14] proposes cross-generalisation algorithm. It replaces the features from the previously learned classes with similar features of the novel classes to adapt to the target task. In Bayesian learning, [41] depicts prior knowledge in the form of probability density function on the parameters of the model, and updates them to compute the posterior model. Matching Nets (MN) [172] uses non-parametric attentional memory mechanisms, and an *"episode"* during the training time. [25] captures salient features of general lung datasets using an encoder and augments multiple views for images, then uses the prototypical network for a 2-way, 1-shot classification.

Zero-shot learning is the extreme case of the FSL where $K = 0$. In other words, the difference between the two is the devoid of any visual examples of the target classes in the training phase of ZSL, while in few-shot learning, the support set contains few labelled samples of the novel categories. Also, auxiliary information in the form of class embeddings is one of the main components of Zero-shot learning. ZSL approaches might extend their solutions to one-shot or few-shot learning by either updating the training data with one or few generated samples from augmentation techniques, or by having access to a few of the unseen images during the training time [5], [17], [23], [59], [145], [147], [164], [171], [189], [199]. [145,189] both use auxiliary text-based information.

3. ZSL test and training phases

ZSL models can be seen from two points of views in terms of training and test phase: Classic ZSL and Generalised ZSL (GZSL) settings. In the classic ZSL settings, the model only detects the presence of new classes at the test phase, while in GZSL settings, the model predicts both unseen and seen classes at the test time; hence, GZSL is more applicable for real-world scenarios [75,86,94,145,210]. The same idea can be applied to FSL to train in the generalised model, called generalised few-shot learning (GFSL) that detects both known and novel classes at the test time.

In the next paragraphs, we discuss two types of training approaches: Inductive vs. Transductive training.

Inductive Training: This training setting only uses the seen class of information to learn a new concept. The training data for the inductive setting is:

$$D_t = \{(x,\ y,\ c(y))\ |\ x \in X^S,\ y \in Y^S,\ c(y) \in C^S\} \tag{2}$$

where x represents image features, y is the class labels, and $c(y)$ denotes the class embeddings. Moreover, X^S and Y^S indicate seen class images and seen class labels, respectively. Inductive learning accounts for the majority of the settings used in ZSL and Generalised Zero-Shot Learning (GZSL). e.g. in [7,22,43,52,90,113,137,171,189,204,207].

Transductive Training: Although the original idea of zero-shot learning is more related to the inductive setting, in many scenarios, the transductive setting is used where either unlabelled visual or textual information, or both for unseen classes are used together with the seen class data e.g. in [6], [44], [52], [71], [90], and in [134,143,159,171,174,189,192,205,207]. The training data for transductive learning is:

$$D_t = \{(x,\ y,\ c(y))\ |\ x \in X^{S\cup U},\ y \in Y^{S\cup U},\ c(y) \in C^{S\cup U}\} \tag{3}$$

where $X^{S\cup U}$ denotes that images come from the union of seen and unseen classes. Similarly, $Y^{S\cup U}$ and $C^{S\cup U}$ indicate the train labels and class embeddings belong to both seen and novel categories.

According to [197], any approach that relies on label propagation will fall into the category of transductive learning. Feature generating network with labelled source data and unlabelled target data [189] are also considered as transductive methods. The transductive setting is seen as one of the solutions to the domain shift problem, since the provided unseen labelled information during training reduces the discrepancy between the two domains.

There is a slight nuance between the transductive learning and semi-supervised learning; in the transductive setting, the unlabelled data solely belong to the unseen test classes, while in semi-supervised setting, unseen test classes might not be present in the unlabelled data. Furthermore, the difference between FSL and the transductive ZSL learning is the existence of a few labelled examples of the unseen classes alongside annotated seen class examples in the few-shot learning. While in the transductive ZSL setting, the examples for the unseen classes are all unlabelled.

ZSL models are developed based on two high-level major strategies to be taken into account: a) defining the *"Embedding Space"* to combine visual and non-visual auxiliary data, and b) choosing an appropriate *"Auxiliary Data Collection"* technique.

a) *Embedding Spaces.* Fig. 3 demonstrates the overall structure of a ZSL system in terms of embedding spaces and auxiliary data types collection techniques. Such systems either map the visual data to the semantic space (Fig. 3a) or embed both visual and semantic data to a common latent space (Fig. 3b), or see the task as a missing data problem, and then map the semantic information to the visual space (Fig. 3c). Two or all of these approaches can also be combined and embedded together to boost up the benefits of each individual categories.

From a different point of view, semantic spaces can also be sub-categorised into euclidean and non-euclidean spaces. The intrinsic relationship between data points is better preserved when the geometrical relation between them is considered. These spaces are commonly based on clusters or graph networks. Some researchers may prefer manifold learning for the ZSL challenge. e.g. in [63], [83], [91], [134], [175], [181], [192], [193], [207], [210]. The Euclidean spaces are more conventional and simpler as the data has a flat representation in such spaces. However, the loss of information is a common issue of these spaces, as well. Examples of methods using Euclidean spaces are [43,80,106,137,187], and [145].

b) *Auxiliary Data Collection.* As mentioned before, Zero-shot learning is the challenge of learning novel classes without seeing their exemplars during the training. Instead, the freely available auxiliary information is used to compensate for the lack of visually labelled data. Such information can be categorised into two groups:

Human annotated attributes. The supervised way of annotating each image with its related attributes is an arduous process and requires time and expertise, but since they are manual, they yield noiseless and important attributes needed for learning and inference. There are several datasets in which side information in the form of attributes can be attained for each image. i.e. aPY [40], AWA1 [80], AWA2 [188], CUB [173], and SUN [118]. Several ZSL methods leverage the attributes as the side information [7,97,137], or visual attributes [40,79].

Unsupervised auxiliary information. There are several forms of auxiliary information that have minimum supervision and are widely used in the ZSL setting, such as human gazes [66], WordNet which is a large-scale lexical database of 117,000 English words [4,5], [7,83,100,102,123, 135,136,181,185], or Textual descriptions such as Web search [135], Wikipedia articles [4], [7], [37], [38], [43], [84], [112], [113], [123], [211], and sentence descriptions [129]. Textual side information needs to be transformed into class embeddings in order to be used at the training and testing stages. Word embedding and language embedding are the two representation techniques used for textual side information. As we gradually proceed, later we review on different embedding classes as well.

4. ZSL data embedding techniques

In this section, we first provide the task definition of ZSL and GZSL. Then we review the four recent approaches on the problem.

In the standard inductive setting as mentioned earlier in Section 3, the training set is

$$D_t = \{(x,\ y,\ c(y)) \mid x \in X^S,\ y \in Y^S,\ c(y) \in C^S\} \tag{4}$$

and the objective function to be minimised is as follows:

$$\mathbb{L} = \frac{1}{N}\sum_{n=1}^{N} L(y_n, f(x_n; W)) + \Omega(W) \tag{5}$$

where, $f(x,y;W) = \underset{y \in Y}{\operatorname{argmax}}\ F(x,y)$ is the mapping function. Through the training phase, the classifier $f : X \rightarrow Y^U$ is learned for ZSL to predict only the novel classes at the test time, or $f : X \rightarrow Y^U \cup Y^S$ for the GZSL challenge to estimate both novel classes and the previously learned seen classes. For instance, the classifier f can be a COVID-19 diagnoser.

We categorise the embedding methodologies into four categories based on the space they learn/infer target classes (like COVID-19 detection in Fig. 3):

1. Semantic Embedding: A semantic space with textual nature in which features are in the form of class embeddings.
2. Intermediate-Space Embedding: A space where both class embeddings and visual feature embedding present in conjunction.
3. Visual Embedding: A space where training and inferring is done with visual feature representations similar to the traditional recognition problems.
4. Hybrid Embedding Models: A combination of spaces are used in some models to bring together the advantages the different spaces have.

The majority of methods focus on the general tasks; however, they are scalable to disease classification.

4.1. Semantic embedding

Semantic embedding itself can be subcategorised into two tasks of *Attribute Classification* and *Label Embedding* which will be discussed here:

4.1.1. Attribute classifiers

Primitive approaches of Zero-Shot learning leverage manually annotated attributes in a two-stage learning schema. Attributes in an image are predicted in the first stage and labels of unseen classes would be chosen using similarity measures in the second stage. [79] uses a probabilistic classifier to learn the attributes and then estimates posteriors for test classes. [136] proposes a method to avoid manual supervision with mining the attributes in an unsupervised manner. [135] adopts DAP together with a hierarchy-based knowledge transfer for large-scale settings. [65]'s method is based on IAP, and uses Self-Organising and Incremental Neural Networks (SOINN) to learn and update attributes online. Later in IAP-SS by [65], an online incremental learning approach is used for faster learning of the new attributes. The Direct Attribute Prediction (DAP) [80] first learns the posteriors of the attributes, then estimates the posteriors of seen classes. On the other hand, Indirect Attribute Prediction (IAP) [80] first learns the posteriors for seen classes then uses them to compute the posteriors for the attributes. [179] uses a unified probabilistic model based on the Bayesian Network (BN) [110] that discovers and captures both object-dependent and object-independent relationships to overcome the problem of relating the attributes. ConSE [113] learns the probability of the training samples. It then predicts an unseen class by the convex combination of the class label embedding vectors. [59] uses a random forest approach for learning more discriminative attributes. Hierarchy and Exclusion (HEX) [31] considers relations between objects and attributes and maps the visual features [130,161] of the images to a set of scores to estimate labels for unseen categories. [8] takes on an unsupervised approach where they capture the relations between the classes and attributes with a three-dimensional tensor while using a DAP-based scoring function to infer the labels. LAGO by [12] also follows the DAP model. It learns soft and-or logical relations between attributes. Using soft-OR, the attributes are divided into groups, and the label class from unseen samples is predicted via a soft-AND within these groups. If each attribute comes from a singleton group, the all-AND will be used.

4.1.2. Label embedding

Instead of using an intermediate step, more recent approaches learn to map images to the structured euclidean semantic space automatically which would be the implicit way of representing knowledge. The compatibility function for linear mapping is:

$$F(x,y;w) = \theta(x)^T w c(y) \tag{6}$$

where $\theta(x)^T$ is the image embedding for training classes and w is the parameters in vector form to be learned. In the case of bilinear projection where it is more common, w takes the form of matrix:

$$F(x,y;W) = \theta(x)^T W c(y) \tag{7}$$

SOC [114] first maps the image features to the semantic embedding space, it then estimates the correct class using nearest neighbour. DeViSE by [43] uses a linear corresponding function with a combination of dot-product similarity and hinge rank loss used in [183]. ALE [6] optimises the ranking loss in [167] alongside the bi-linear mapping compatibility function. SJE [7] learns a bi-linear compatibility function using the structural SVM objective function [166]. ESZSL [137] introduces a better regulariser and optimises a close form solution objective function in a linear manner. ZSLNS [123] proposes a $l_{1,2}$ -norm based loss function. [17] takes on a metric learning approach and linearly embed the visual features to the attribute space. LAGO [12] is a probabilistic model that depicts soft and-or relations between groups of attributes. In a case where all attributes form all-OR group, It becomes similar to ESZSL [137] and learns a bilinear compatibility function. AREN [190] uses attentive region embedding while learning the bilinear mapping to the semantic space in order to enhance the semantic transfer. ZSLPP [38] combines two networks VPDE-net for detecting bird parts from images and PZSC-net that trains a part-based Zero-Shot classifier from the noisy text of the Wikipedia. DSRL [197] uses non-negative sparse matrix factorisation to align vector representations with the attribute-based label representation vectors so that more relevant visual features are passed to the semantic space.

Some approaches to ZSL use non-linear compatibility functions. CMT [157] uses a two-layer neural network, similar to common MLP networks by [131] alongside the compatibility function. In UDA [71] a non-linear projection from feature space to semantic space (word vector and attribute) is proposed in an unsupervised domain adaptation problem based on regularised sparse coding. [84] uses a deep neural network [161]

regression which generates pseudo attributes for each visual category via Wikipedia. LATEM [185] constructs a piece-wise non-linear compatibility function alongside a ranking loss. [23] regularises the model using structural relations of the cluster by which cluster centres characterise visual features. QFSL by [159] solves the problem in a transductive setting, and projects both sources and target images into several specified points to fight bias problem.

GFZSL [171] uses both linear and non-linear regression models and generates a probability distribution for each class. For transductive setting, it uses Expectation-Maximisation (EM) to estimate a Gaussian Mixture Model (GMM) of unlabelled data in an iterative manner.

Leveraging the non-euclidean spaces to capture the manifold structure of the data is another approach to the problem. Together with the knowledge graphs, the explicit relations between the labels will be demonstrated. In this setting, the side information mainly comes from a hierarchy ontology like WordNet. The mapping function will have the following form:

$$F(x, y; W) = \theta(X, A)^T W c(y) \tag{8}$$

where X is the $n \times k$ feature matrix and A is the adjacency matrix of the graph.

Propagated Semantic Transfer (PST) [134] first uses DAP model to transfer knowledge to novel categories, following the graph-based learning schema, it improves local neighbourhood in them. DMaP [91] jointly optimises the projecting of the visual features and the semantic space to improve the transferability of the visual features to the semantic space manifold. MFMR [193] decomposes the visual feature matrix into three matrices to further facilitate the mapping of visual features to the semantic spaces. To improve the representation of the geometrical manifold structure of the visual and semantic features, manifold regularisation is used. In [83] a Graph Search Neural Network (GSNN) [102] is used in the semantic space based on the WordNet knowledge graph to predict multiple labels per image using the relations between them. [181] distils both auxiliary information in forms of word embedding and knowledge graph to learn novel categories. DGP [63] proposes dense graph propagation to propagate knowledge directly through dense connections. In [210], a graphical model with a low dimensional visually semantic space is utilised which has a chain-like structure to close the gap between the high-dimensional features and the semantic domain.

4.2. Intermediate-Space Embedding

One of the methods of embedding is to measure the similarity between the visual and semantic features in a joint space.

4.2.1. Fusion-based models

Considering unseen classes as a fusion of previously learned seen concepts is called hybrid learning. Standard scoring function for hybrid models is defined as:

$$f(x, y; W) = \sum_{s \in S} \left(W, \theta^S(x)\right) c(y) \tag{9}$$

SSE [204] considers the histogram similarity between the seen class auxiliary information and seen visual data. SYNC [22] uses two spaces of semantic and model space, and the alignment is conducted with phantom classes. With the sparse linear combination of the classifiers for the phantom classes, the final classifier is learned. TVSE [192] learns a latent space using collective matrix factorisation with graph regularisation to incorporate the manifold structure between source and target instances, moreover, it represents each sample as a mixture of seen class scores. LDF [93] combines the prototypes of seen classes and jointly learns embeddings for both user-defined attributes and latent attributes.

4.2.2. Joint representation space models

Inferring unseen labels via measuring similarity between cross-modal data in a shared latent space is another workaround to the ZSL challenge. The first term in the objective function for standard cross-modal alignment approaches is:

$$\mathbb{L} = \min_{c(y)^S} \left\| x^S - c(y)^S y^S \right\|_F^2 \tag{10}$$

with y being a one-hot vector of corresponding class labels and $\| \ . \|_F^2$ is the Frobenius norm. Approaches to joint space learning are grouped into two categories, Parametric which follow a slow learning via optimising a problem and Non-parametric that leverage data points extracted from neural networks in a shared space. In parametric methods including [44] a multi-view alignment space is proposed for embedding low-level visual features. The learning procedure is based on the multi-view Canonical Correlation Analysis (CCA) [47]. [100] applies PCA and ICA embeddings to reveal the visual similarity across the classes and obtains the semantic similarity with the WordNet graph, followed by embedding the two outputs into a common space. MCZSL [4] uses visual part and multi-cue language embedding in a joint space. In [108] both images and words are represented by Gaussian distribution embedding. JLSE [205] decides on a dictionary learning approach to learn the parameters of source and target domains across two separate latent spaces where the similarity is computed by the likelihood of similarity independent to the class label. CDL [61] uses a coupled dictionary to align the structure of visual-semantic space using discriminative information of the visual space. In [73,138], a coupled sparse dictionary is leveraged to relate visual and attribute features together. It uses entropy regularisation to alleviate the domain shift problem.

There are several non-parametric methods. ReViSE [164] that combines auto-encoders with Maximum Mean Discrepancy (MMD) loss [49] in order to align the visual and textual features. DMAE [109] introduces a latent alignment matrix with representations from auto-encoders optimised by kernel target alignment (KTA) [29] and squared-loss mutual information (SMI) [195]. DCN [94] proposes a novel Deep Calibration Network in which an entropy minimisation principle is used to calibrate the uncertainty of unseen classes as well as seen classes.

To narrow the semantic gap, BiDiLEL [176] introduces a sequential bidirectional learning strategy and creates a latent space using the visual data, then the semantic representations of unseen classes are embedded in the previously created latent space. This method comprises both parametric and non-parametric models.

4.3. Visual embedding

Visual embedding is the other type of ZSL methods that performs classification in the original feature space and is orthogonal to semantic space projection. This is done by learning a linear or non-linear projection function. For linear corresponding functions, WAC-Linear [37] uses textual description for seen and unseen categories and projects them to the visual feature space with a linear classifier. [207] follows a transductive setting in which it refines unseen data distributions using unseen image data. To approximate the manifold structure of data, they use a global linear mapping for synthesising virtual cluster centres. [52] assigns pseudo labels to samples using reliability (with robust SVM) and diversity (via diversity regularisation). For learning a non-linear corresponding function in WAC-Kernel [36] and in order to leverage any kind of side information, a kernel method is proposed to predict a kernel-based on the representer theorem [144]. DEM [202] uses the least square embedding loss to minimise the discrepancy between the visual features and their class representation embedding vector in the visual feature space. OSVE [96] reversely maps from attribute space to visual space then trains the classifier using SVM [11]. In [60] the authors introduce a stacked attention network that corporates both global and local visual features weighted by relevance along with the semantic features. In [174] visual constraint is used in class centres in the visual space to avoid the domain shift problem.

4.3.1. Visual data augmentation

There are a variety of generative networks that augment unseen data, taking GAN [48] as an example, the first term in objective function would be:

$$\mathbb{L} = \max \ \mathbb{E}[\log D(x,\ c(y))] + \min \ \mathbb{E}[\log(1 - D(\tilde{x},\ c(y))] \tag{11}$$

$\tilde{x} = G(z, c(y))$ is the synthesised data of the generator and $z \in R^{d_z}$ is random Gaussian noise. The role of the discriminator D and generator G contradicts in loss function as the first one attempts to maximise the loss while the latter tries to minimise it. Another widely used generative neural network is the Variational AutoEncoder (VAE) [69]:

$$\mathbb{L} = \mathbb{E}_{q_\varphi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\varphi(z|x) \,||\, p_\theta(z)) \tag{12}$$

The first term is the reconstruction loss, and the latter is the Kullback-Leibler divergence that works as a regulariser.

RKT [175] leverages relational knowledge of the manifold structure in the semantic space, and generates virtually labelled data for unseen classes from Gaussian distributions generated by sparse coding. Then it projects them alongside the seen data to the semantic space via linear mapping. GLaP [90] generates virtual instances of an unseen class with the assumption that each representation obeys a prior distribution where one can draw samples from. To ease the embedding to the semantic space, GANZrl [162] proposes to increase the visual diversity by generating samples with specified semantics using GAN models. SE-GZSL [75] uses a feedback-driven mechanism for its discriminator that learns to map the produced images to the corresponding class attribute vectors. To enforce the similarity of the distribution of the sample and generated sample, a loss component was added to the VAE objective [69] function.

Synthesised images often suffer from looking unrealistic since they lack intricate details. A way around this issue is to generate features instead. [18] uses a GMMN model [89] to generate visual features for unseen classes. In [42] a multi-modal cycle consistency loss is used in training the generator for better reconstruction of the original semantic features. CVAE-ZSL [106] takes attributes and generates features for the unseen categories via a Conditional Variations AutoEncoder (CVAE) [158]. L_2 norm is used as the reconstruction loss. GAZSL [211] utilises noisy textual descriptions from Wikipedia to generate visual features. A visual pivot regulariser is introduced to help generate features with better qualities. f-CLSWGAN [187] combines three conditional GAN variants for a better data generation. f-VAEGAN-D2 [189] combines the architectures of conditional VAE [158], GAN [48] and a non-conditional discriminator for the transductive setting. LisGAN [87] generates unseen features from random noises using conditional Wasserstein GANs [9]. For regularisation, they introduce semantically meaningful soul samples for each class and force the generated features to be close to at least one of the soul samples. Gradient Matching Network (GMN) [143] trains an improved version of the conditional WGAN [51] to produce image features for the novel classes. It also introduces Gradient Matching (GM) loss to improve the quality of the synthesised features. In order to synthesise unseen features, SPF-GZSL [86] selects similar instances and combines them to form pseudo features using a centre loss function [182]. In Don't Even Look Once (DELO) by [209] a detection algorithm is conducted to synthesise unseen visual features to gain high confidence predictions for unseen concepts while maintaining low confidence for backgrounds with vanilla detectors.

Instead of augmenting data using synthesising methods, data can be acquired by gathering web images. [112] jointly uses web data which are considered weakly-supervised categories alongside the fully-supervised auxiliary labelled categories. It then learns a dictionary for the two categories.

4.4. Hybrid Embedding Models

Several works make use of both visual and semantic projections to reconstruct better semantics to confront domain shift issue by alleviating the contradiction between the two domains. Semantic AutoEncoder (SAE) [72] adds a visual feature reconstruction constraint. It combines linear visual-to-semantic (encoder) and linear semantic-to-visual (decoder). SP-AEN [24] is a supervised Adversarial AutoEncoder [101] which improves preserving the semantics by reconstructing the images from the raw 256 × 256 × 3 RGB colour space. BSR [153] uses two different semantic reconstructing regressors to reconstruct the generated samples into semantic descriptions. CANZSL [26] combines feature-synthesis with semantic embedding by using a GAN for generating visual features and an inverse GAN to project them into semantic space. In this way, the produced features are consistent with their corresponding semantics.

Some of the synthesising approaches utilise a common latent space to align the generated features space with the semantic space to facilitate capturing the relations between the two spaces. [97] introduces a latent-structure-preserving space where synthesised features from given attributes would suffer less from bias and variance decay with the help of Diffusion Regularisation. CADA-VAE [145] generates a visual feature latent space where both of visual and semantic features are embedded in this space by a VAE [69]. It uses Distribution Alignment (DA) loss and Cross-Alignment (CA) loss to align the cross-modal latent distributions.

GDAN [58] combines all three approaches and designs a dual adversarial loss. In this way, regressor and discriminator learn from each other.

A summary of the different approaches is reported in Table 1. The number of methods are growing with time and we can interpret that some areas like direct learning, common space learning and visual data synthesising are more popular in solving the task, while models combining different approach are fairly newer techniques thus have fewer works that are reported here.

5. Evaluation protocols

In this section, we review some of the standard evaluation techniques to analyse the performance of the ZSL techniques based on the common benchmark datasets in the field, also in terms of dataset splits, class embeddings, image embeddings, and various evaluation metrics. First, the benchmark datasets.

5.1. Benchmark datasets

There are several well-known benchmark datasets for Zero-shot learning that are frequently used.

North America Birds (NAB) [168] is a fine-grained dataset of birds consisting of 1011 classes and 48,562 images. Images are categorised based on their visual attributes. A new version of this dataset is proposed by [38] in which the identical leaf nodes are merged to their parent nodes where their only differences were genders and resulted in final 404 classes.

Attribute datasets. SUN Attribute [118] is a medium-scale and fine-grained attribute dataset consisting of 102 attributes, 717 categories and a total of 14,340 images of different scenes. CUB-200-2011 Birds (CUB) [173] is a 200 category fine-grained attribute dataset with 11,788 images of bird species that includes 312 attributes. Animals with Attributes (AWA1) [80] is another attribute dataset of 30,475 images with 50 categories and 85 attributes, the image features in this dataset are licensed and not available publicly. Later, Animals with Attributes 2 (AWA2) was presented by [188] which is a free version of AWA1 with more images than the previous one (37,322 images), with the same number of classes and attributes, but different images. aPascal and Yahoo (aPY) [40] is a dataset with a combination of 32 classes, including 20 pascal and 12 yahoo attribute classes with 15,339 images and 64 attributes in total.

A summary of the statics for the attribute datasets are gathered in Table 2.

ImageNet [32] is a large-scale dataset that contains 14 million

Table 1
Common ZSL and GZSL methods categorised based on their embedding space model, with further divisions in a top-down manner.

Models	Categories	Main Features	Description
Semantic Embedding	Two-Step Learning	Attributes classifiers	DAP-Based [8,12,79,80,135,136] IAP-Based [65,79,80,113] Bayesian network (BN) [179], Random Forest Model [59], HEX Graph [31]
	Direct Learning	Implicit knowledge representation	Linear [6,7,12,17,38,43,114,123,137,171,190,197] or Non-Linear [23, 71,84,159,171,185]Compatibility Functions
		Explicit knowledge representation	Graph Conv. Networks (GCN) [181], Knowledge Graphs [63,83,91,134], 3-Node Chains [210], Matrix Tri-Factorisation with Manifold Regularisation [193]
Cross-Modal Latent Embedding	Fusion-based Models	Fusion of seen class data	Combination of seen classes properties [22,93,204], Combination of seen class scores [192]
	Common Representation Space Models	Mapping of the visual and semantic spaces in a joint intermediate space	Parametric [4,44,61,73,100,108,138,205], Non-parametric [94,109, 164], or Both [176]
Visual Embedding	Visual Space Embedding	Learning of the semantic to visual projection	Linear [37,52,207] or Non-linear [36,60,96,174,202] Projection functions
	Data Augmentation	Image generation	Gaussian distribution [90,175], GAN [162], VAE [75]
		Visual feature generation	GAN [42,87,211], WGAN [143,187], CVAE [106,209], VAE + GAN [189], GMMN [18], Similar feature combination [86]
	Leveraging Web Data	Web images crawling	Dictionary learning [112]
Hybrid	Visual + Semantic Embedding	Reconstruction of the semantic features	AutoEncoder [72], Adversarial AutoEncoder [24], GAN with two reconstructing regressors [153], GAN an inverse GAN [26]
	Visual + Cross Modal Embedding	Feature generation with aligned semantic features	Semantic to visual mapping [97], VAE [145]
	All	Utilisation of generator and discriminator together with the regressor	GAN + Dual Learning [58]

Table 2
Statics of the attribute datasets accounting for the number of attributes, classes plus their splits and their total number of images.

Attribute Datasets	#attributes	y	y^U	y^S	#images
SUN [118]	102	717	580 + 65	72	14,340
CUB [173]	312	200	100 + 50	50	11,788
AWA1 [80]	85	50	27 + 13	10	30,475
AWA2 [188]	85	50	27 + 13	10	37,322
aPY [40]	64	32	15 + 5	12	15,339

images, shared between 21k categories with each image having one label that makes it a popular benchmark to evaluate models in real-world scenarios. Its organisation is based on WordNet hierarchy [105]. ImageNet is imbalanced between classes as the number of samples in each class vary greatly and is partially fine-grained. A more balanced version has 1k classes with 1000 images in each category.

There are several approaches in FSL setting for COVID-19 diagnosis, however ZSL is still new in the field of disease recognition, we introduce a dataset suited for the task of ZSL/GZSL that contains the required image and textual descriptions in one place.

COVID-ChestXRay [27] is a small and public dataset of CXR and CT scans suitable for ZSL and Few-shot learning experiences. At the time of this research, it had 444 unique clinical notes for a total of 16 categories, from no finding (normal cases) to other pneumonic cases like COVID-19, MERS, and SARS.

5.2. Dataset splits

Here we discuss the original splits of the datasets as well as the other splits proposed for the Zero-shot problem.

Standard Splits (SS). In ZSL problems, unseen classes should be disjoint to seen classes and test time samples are limited to unseen classes, thus the original splits aim to follow this setting. SUN [118] proposed to use 645 classes for training among which 580 of the classes are used for training, 65 classes are for validation and the remaining 72 classes will be used for testing. For CUB, [6] introduces the split of 150 training classes (including 50 validation classes) and 50 test classes. As for AWA1, [80] introduced the standard split of 40 classes for training (13 validation classes) and 10 classes for testing. The same splits are used for AWA2. In aPY, 20 classes of Pascal are used for training (15 classes for training and 5 for validation), while the 12 classes of Yahoo are used for testing.

Proposed Splits (PS). The standard split images from SUN, CUB, AWA1 and aPY overlap with some images of pre-trained ResNet-101 ImageNet model. To solve the problem, proposed splits (PS) is introduced by [186] where no test images are contained in the ImageNet 1K dataset. [186] proposes 9 ZSL splits for the ImageNet dataset; two of which evaluate the semantic hierarchy in distance-wise scales of 2-hops (1509 classes) and 3-hops (7678 classes) from the 1k training classes. The remaining six splits consider the imbalanced size of classes with increasing granularity splits starting from 500, 1K and 5K least-populated classes to 500, 1K and 5K most-populated classes, or All which denotes a subset of 20k other classes for testing.

Seen-Unseen relatedness. To measure the relatedness of seen samples to unseen classes, [38] introduces two splits Super-Category-Shared (SCS) and Super-Category-Exclusive (SCE). SCS is the easy split since it considers the relatedness to the parent category while SCE is harder and measures the closeness of an unseen sample to that particular child node.

5.3. Class embeddings

There exist several class embeddings, each suitable for a specific scenario. Class embeddings are in forms of vectors of real numbers which can further be used to make predictions based on the similarity between them and can be obtained through four categories: attributes, word embeddings, hierarchical ontology, and language modelling. The last three are done in an unsupervised manner thus do not require human labour.

5.3.1. Supervised attribute-embeddings

Human annotated attributes are done under the supervision of experts with a great amount of effort. Binary, relative and real-valued

attributes are three types of attributes embeddings. Binary attributes depict the presence of an attribute in an image thus value is either 0 or 1. They are the easiest type and are provided in benchmark attribute datasets AWA1, AWA2, CUB, SUN, aPY. Relative attributes [115] on the other hand, show the strength of an attribute in a given image comparing to the other images. The real-valued attributes are in continuous form thus they have the best quality [7]. In the SUN attribute dataset [118], they have achieved confidence through averaging the binary labels from multiple annotators.

5.3.2. Unsupervised word-embeddings

Also known as Textual corpora embedding. Bag of Words (BOW) [54] is a one-hot encoding approach. It simply shows the number of occurrences of the words in a representation called bag and is negligent of word orders and grammar. One-hot encoding approaches had a drawback of giving the stop words (like "a", "the" and "of") high relevancy counts. Later, Term Frequency-Inverse Document Frequency (TF-IDF) [142] used term weighting to alleviate this problem by filtering the stop words and to keep meaningful words. Word2Vec [103] is a widely used two-layered neural embedding model and has two variants, CBOW and skip-gram. CBOW predicts a target word in the centre of a context using its surroundings words while the skip-gram model predicts surrounding words using a target word. CBOW is faster in train and usually results in better accuracy for frequent words while Skip-gram is preferred for rare words and it works well with sparse training data. Global Vectors (GloVe) [119] is trained on Wikipedia. It combines local context window methods and global matrix factorisation. Glove learns to consider global word-word co-occurrence matrix statistics to build the word embeddings.

5.3.3. Hierarchy embedding

WordNet [105] is a large-scale public lexical database of 117,000 synsets. Synsents are a group of words that are semantically related to each other. i.e. synonyms, homonyms and meronymies of English words that are organised using the hierarchy distances with a graph structure, thus Approaches based on knowledge graphs often follow the WordNet to measure the similarity between the word meanings [4,5,7,83,100,102, 123,135,136,181,185].

5.3.4. Language modelling

In the general ZSL scenarios, word by word representations considered; however, with the advent of transfer learning in the natural language processing (NLP), and the introduction of contextual word embeddings, the boundaries of the capabilities of the embeddings has been pushed further. Unlike the traditional word embeddings, language models can capture the meaning of the words based on the context in which they appear. Several contextual representations have been introduced recently and showed great results. These existing pre-trained models can be fine-tuned on various ZSL tasks.

ELMo [120] is a contextual embedding model. Following morphological clues together with a deep bidirectional language model (biLM), ELMo learns the representations. Bidirectional Encoder Representations from Transformer (BERT) [33] is a multi-layer bidirectional Transformer encoder [170] trained upon BooksCorpus [212] dataset and English Wikipedia. It outperforms ELMo with having more parameters and layers. The pre-trained BERT model can be fine-tuned with just one additional output layer. However, BERT suffers from fine-tuning discrepancy due to ignoring the relation the masked positions have. XLNet [196] uses an autoregressive model to introduce a method that overcomes the shortcoming of BERT. In addition to the datasets used by BERT, XLNet pre-trains the model on Giga5 [116], ClueWeb 2012-B extended by [20] and Common Crawl[1]. ALBERT [81] increases the model size. It lowers the memory usage with two parameter reduction techniques. The first one is a factorized embedding parameterization. The second one is cross-layer parameter sharing. These two techniques result in lower memory usage and higher training speed than BERT. The data used for pre-training is the same as XLNeT.

In this article, we report the results of ZSL and GZSL using the same class embeddings as [186] that is Word2Vec trained on Wikipedia for ImageNet and per-class attributes for the attribute datasets, and for the seen-unseen relatedness task we follow [38] and consider TF-IDF for the CUB and NAB datasets.

5.4. Image embeddings

Existing models use either shallow or deep feature representation. Examples of shallow features are SIFT [99], PHOG [16], SURF [15] and local self-similarity histograms [148]. Among the mentioned features, SIFT is the commonly used features in ZSL models like [6,22,44].

Deep features are obtained from deep CNN architectures [161] and contain higher-level features. Extracted features are one of the followings:

4096-dim top-layer hidden unit activations (fc7) of the AlexNet [74], 1000-dim last fully connected layer (fc8) of VGG-16 [155], 4096-dim of the 6th layer (fc6) and 4096-dim of the last layer (fc7) features of the VGG-19 [155]. 1024-dim top-layer pooling units of the GoogleNet [160]. and 2048-dim last layer pooling units of the ResNet-101 [55].

In this paper, we consider the ResNet-101 network which is pre-trained on ImageNet-1K without any fine-tuning. That is the same image embedding used in [186]. Features are extracted from whole images of SUN, CUB, AWA1, AWA2, and ImageNet and the cropped bounding boxes of aPY. For the seen-unseen relatedness task, VGG-16 is used for CUB and NAB as proposed in [38].

5.5. Evaluation metrics

Common evaluation criteria used for ZSL challenge are:

Classification accuracy. One of the simplest metrics is classification accuracy in which the ratio of the number of the correct predictions to samples in class y is measured. However, it results in a bias towards the populated classes.

Average per-class accuracy. To reduce the bias problem for the populated classes, average per-class accuracies are computed by multiplying the division of the classification accuracy to division of their cumulative sum.

$$acc_y = \frac{1}{||y||}\sum_{y=1}^{||y||}\frac{\#\text{correct predictions in class } y}{\#\text{samples in class } y}\ [188] \tag{13}$$

Harmonic mean. For performance evaluation on both seen and unseen classes (i.e. the GZSL setting), the Top-1 accuracies for the seen and unseen classes are used to compute the harmonic mean:

$$H = \frac{2*acc_{y^S}*acc_{y^U}}{acc_{y^S} + acc_{y^U}}\ [188] \tag{14}$$

In this paper, we designate the Top-1 accuracies and the harmonic mean as the evaluation protocols [188].

6. Experimental results

As the main contributions of this research, and for the first time, we provide a comprehensive experiments of 21 state-of-the-art models in ZSL/GZSL domain that include the evaluations and comparisons of data-synthesising methods. In this section, first we provide the results for ZSL, GZSL and seen-unseen relatedness on attribute datasets, then we present the experimental results on the ImageNet dataset. A minor part of the results is reported from [188] for a more comprehensive comparison.

[1] https://commoncrawl.org/

Table 3
Zero-shot learning results for the Standard Split (SS) and Proposed Split (PS) on SUN, CUB, AWA1, AWA2, and aPY datasets. We measure Top-1 accuracy in % for the results. †and ‡denote inductive and transductive settings respectively.

	Methods	SUN		CUB		AWA1		AWA2		aPY	
		SS	PS	SS	PS	SS	PS	SS	PS	SS	PS
	DAP [80]	38.9	39.9	37.5	40.0	57.1	44.1	58.7	46.1	35.2	33.8
	IAP [80]	17.4	19.4	27.1	24.0	48.1	35.9	46.9	35.9	22.4	36.6
	ConSE [113]	44.2	38.8	36.7	34.3	63.6	45.6	67.9	44.5	25.9	26.9
	CMT [157]	41.9	39.9	37.3	34.6	58.9	39.5	66.3	37.9	26.9	28.0
	SSE [204]	54.5	51.5	43.7	43.9	68.8	60.1	67.5	61.0	31.1	34.0
	LATEM [185]	56.9	55.3	49.4	49.3	74.8	55.1	68.7	55.8	34.5	35.2
	ALE [6]	59.1	58.1	53.2	54.9	78.6	59.9	80.3	62.5	30.9	39.7
	DeViSE [43]	57.5	56.5	53.2	52.0	72.9	54.2	68.6	59.7	35.4	39.8
†	SJE [7]	57.1	53.7	55.3	53.9	76.7	65.6	69.5	61.9	32.0	32.9
	ESZSL [137]	57.3	54.5	55.1	53.9	74.7	58.2	75.6	58.6	34.4	38.3
	SYNC [22]	59.1	56.3	54.1	55.6	72.2	54.0	71.2	46.6	39.7	23.9
	SAE [72]	42.4	40.3	33.4	33.3	80.6	53.0	80.7	54.1	8.3	8.3
	GFZSL [171]	62.9	60.6	53.0	49.3	80.5	68.3	79.3	63.8	51.3	38.4
	DEM [202]	–	61.9	–	51.7	–	68.4	–	67.1	–	35.0
	GAZSL [211]	–	61.3	–	55.8	–	68.2	–	68.4	–	41.1
	f-CLSWGAN [187]	–	60.8	–	57.3	–	68.8	–	68.2	–	40.5
	CVAE-ZSL [106]	–	61.7	–	52.1	–	71.4	–	65.8	–	–
	SE-ZSL [75]	64.5	63.4	60.3	59.6	83.8	69.5	80.8	69.2	–	–
	ALE-tran [6]	–	55.7	–	54.5	–	65.6	–	70.7	–	46.7
‡	GFZSL-tran [171]	–	64.0	–	49.3	–	81.3	–	78.6	–	37.1
	DSRL [197]	–	56.8	–	48.7	–	74.7	–	72.8	–	45.5

6.1. Zero-shot learning results

For the original ZSL task where only unseen classes are being estimated during the test time, we compare 21 state-of-the-art models in Table 3, among which, DAP [80], IAP [80] and ConSE [113] belong to attribute classifiers. CMT [157], LATEM [185], ALE [6], DeViSE [43], SJE [7], ESZSL [137], GFZSL [171] and DSRL [197] are from compatibility learning approaches, SSE [204] and SYNC [22] are representative models of cross-modal embedding, DEM [202], GAZSL [211], f-CLSWGAN [187], CVAE-ZSL [106], SE-ZSL [75] are visual embedding models. From the hybrid or combination category, we compare the results of SAE [72]. Three transductive approaches ALE-tran [6], GFZSL-tran [171] and DSRL [197] are also presented among the selected models. Due to the intrinsic nature of the transductive setting, the results are competitive and in some cases better than the inductive methods, i.e. for GFZSL-tran [171] the accuracy is 9.9 % higher than CVAE-ZSL [106] for PS split of AWA1 dataset. However, in comparison with the inductive form of the same model, there are cases where the inductive model has better accuracies. i.e. in PS split of the aPY dataset, the performance is 38.4% vs 37.1% or for ALE-tran [6] model in PS split of SUN it's 58.1% vs 55.7%, also for PS split of CUB it is 54.9% vs 54.5% with its inductive type. GFZSL [171], a compatibility-based approach, has the best scores compared to other models of the same category in every dataset except for the CUB where SJE [7] tops the results in both splits. This superiority could be due to the generative nature of the model. GFZSL [171] performs the best on AWA1 both in inductive and transductive settings. Out

Table 4
Generalised Zero-Shot Learning results for the Proposed Split (PS) on SUN, CUB, AWA1, AWA2, and aPY datasets. We measure the Top-1 accuracy in % for seen (y^S), unseen (y^U) and their harmonic mean (H). †and ‡denote inductive and transductive settings, respectively.

	Methods	SUN			CUB			AWA1			AWA2			aPY		
		y^U	y^S	H	y^U	y^S	H	y^U	y^S	H	y^U	y^S	H	y^U	y^S	H
	DAP [80]	4.2	25.1	7.2	1.7	67.9	3.3	0.0	88.7	0.0	0.0	84.7	0.0	4.8	78.3	9.0
	IAP [80]	1.0	37.8	1.8	0.2	72.8	0.4	2.1	78.2	4.1	0.9	87.6	1.8	5.7	65.6	10.4
	ConSE [113]	6.8	39.9	11.6	1.6	72.2	3.1	0.4	88.6	0.8	0.5	90.6	1.0	0.0	91.2	0.0
	CMT [157]	8.1	21.8	11.8	7.2	49.8	12.6	0.9	87.6	1.8	0.5	90.0	1.0	1.4	85.2	2.8
	CMT* [157]	8.7	28.0	13.3	4.7	60.1	8.7	8.4	86.9	15.3	8.7	89.0	15.9	10.9	74.2	19.0
	SSE [204]	2.1	36.4	4.0	8.5	46.9	14.4	7.0	80.5	12.9	8.1	82.5	14.8	0.2	78.9	0.4
	LATEM [185]	14.7	28.8	19.5	15.2	57.3	24.0	7.3	71.7	13.3	11.5	77.3	20.0	0.1	73.0	0.2
	ALE [6]	21.8	33.1	26.3	23.7	62.8	34.4	16.8	76.1	27.5	14.0	81.8	23.9	4.6	73.7	8.7
	DeViSE [43]	16.9	27.4	20.9	23.8	53.0	32.8	13.4	68.7	22.4	17.1	74.7	27.8	4.9	76.9	9.2
†	SJE [7]	14.7	30.5	19.8	23.5	59.2	33.6	11.3	74.6	19.6	8.0	73.9	14.4	3.7	55.7	6.9
	ESZSL [137]	11.0	27.9	15.8	12.6	63.8	21.0	6.6	75.6	12.1	5.9	77.8	11.0	2.4	70.1	4.6
	SYNC [22]	7.9	43.3	13.4	11.5	70.9	19.8	8.9	87.3	16.2	10.0	90.5	18.0	7.4	66.3	13.3
	SAE [72]	8.8	18.0	11.8	7.8	54.0	13.6	1.8	77.1	3.5	1.1	82.2	2.2	0.4	80.9	0.9
	GFZSL [171]	0.0	39.6	0.0	0.0	45.7	0.0	1.8	80.3	3.5	2.5	80.1	4.8	0.0	83.3	0.0
	DEM [202]	20.5	34.3	25.6	19.6	57.9	29.2	32.8	84.7	47.3	30.5	86.4	45.1	11.1	75.1	19.4
	GAZSL [211]	21.7	34.5	26.7	23.9	60.6	34.3	25.7	82.0	39.2	19.2	86.5	31.4	14.2	78.6	24.1
	f-CLSWGAN [187]	42.6	36.6	39.4	43.7	57.7	49.7	57.9	61.4	59.6	52.1	68.9	59.4	32.9	61.7	42.9
	CVAE-ZSL [106]	–	–	26.7	–	–	34.5	–	–	47.2	–	–	51.2	–	–	–
	SE-GZSL [75]	40.9	30.5	34.9	41.5	53.3	46.7	56.3	67.8	61.5	58.3	68.1	62.8	–	–	–
	CADA-VAE [145]	47.2	35.7	40.6	51.6	53.5	52.4	57.3	72.8	64.1	55.8	75.0	63.9	–	–	–
	ALE-tran [6]	19.9	22.6	21.2	23.5	45.1	30.9	25.9	–	–	12.6	73.0	21.5	8.1	–	–
‡	GFZSL-tran [171]	0	41.6	0	24.9	45.8	32.2	48.1	–	–	31.7	67.2	43.1	0.0	–	–
	DSRL [197]	17.7	25.0	20.7	17.3	39.0	24.0	22.3	–	–	20.8	74.7	32.6	11.9	–	–

of cross-modal methods, SYNC [22] performs better than SSE [204] in SUN and CUB datasets, while for AWA1, AWA2 and aPY in SS split it has lower performance than SSE [204] in the proposed split. Visual generative methods have proved to perform better as they make the problem into the traditional supervised form, among which, SE-ZSL [75] has the most outstanding performance. For the proposed split in one case on CUB dataset, SE-ZSL [75] performs better than ALE-tran [6] which is its transductive counterpart where the accuracies are 59.6% vs 54.5%. In PS split of AWA1, CVAE-ZSL [106] stays at the top, with 1.9% higher accuracy than the second-best performing model. The accuracies for SS splits are higher than PS in most cases and the reason could be the test images included in training samples, especially for AWA1 and AWA2, as reported in [186].

6.2. Generalised Zero-Shot Learning results

A more real-world scenario where previously learned concepts are estimated alongside new ones is necessary to experiment. 21 state-of-the-art models, same as ZSL challenge, include: DAP [80], IAP [80], ConSE [113], CMT [157], SSE [204], LATEM [185], ALE [6], DeViSE [43], SJE [7], ESZSL [137], SYNC [22], SAE [72], GFZSL [171], DEM [202], GAZSL [211], f-CLSWGAN [187], CVAE-ZSL [106], SE-GZSL [75], ALE-train [6], GFZSL-tran [171], DSRL [197]. CADA-VAE [145] is added to the comparison as a model combining the visual feature augmentation approach with the cross-modal alignment. CMT* [157] has a novelty detection and is included in the report as an alternative version to CMT [157]. The reports in Table 4 are in PS splits. As shown in the table, the results on y^S are dramatically higher than y^U since in GZSL, the test search space includes seen classes as well as unseen classes, this gap is the most conspicuous in attribute classifiers like DAP [80] that performs poorly on AWA1 and AWA2, hybrid approaches and in GFZSL [171] where it results in 0% accuracy on SUN and CUB when training classes are estimated at test time. However for three models f-CLSWGAN [187], SE-GZSL [75] and CADA-VAE [145] in SUN dataset, the accuracy for y^U is higher than y^S, i.e. for SE-GZSL [75] it is 10.4% higher. For a fair comparison, the weighted average of training and test classes is also reported. According to harmonic means, the best model on all evaluated datasets is SE-ZSL [75], although the results haven't been reported for aPY. In some cases, the attribute classifier achieves the best results on y^S. Transductive models have fluctuating results in comparison with their inductive types. CADA-VAE [145] achieves the best performance in all of the harmonic means cases (results for aPY are not reported) and shows the best results, higher than all of the transductive methods.

6.3. Seen-unseen relatedness results

For fine-grained problems, sometimes it is important to measure the closeness of previously known concepts to novel unknown ones. For this purpose, a total of eleven models are compared in Table 5. MCZS [4], WAC-Linear [37], WAC-Kernel [36], ESZSL [137], SJE [7], ZSLNS [123], SynC$_{fast}$ [22], SynC$_{OVO}$ [22], ZSLPP [38], GAZSL [211] and CANZSL [26]. SCE is the hard split thus has lower accuracies compared to the SCS splits. The two variations reported for SYNC [22] model, SynC$_{fast}$ denotes the setting in which the standard Crammer-Singer loss is used, and SynC$_{fast}$ [22] depicts setting with one-versus-other classifiers. The first setting has better accuracies on CUB. CANZSL [26] outperforms all other models in both datasets and splits and improves the accuracy by 4% from 10.3% to 14.3% on SCE split of the CUB dataset and 35.6% vs 38.1% in SCS splits of NAB compared to the next best performing model is GAZSL [211]. Similar to previous experiments, in the seen-unseen relatedness challenge, models that contain feature generating steps have the highest results.

Table 5
Seen-Unseen relatedness results on CUB and NAB datasets with easy (SCS) and hard (SCE) splits. Top-1 accuracy is reported in.%

Methods	CUB		NAB	
	SCS	SCE	SCS	SCE
MCZSL [4]	34.7	–	–	–
WAC-Linear [37]	27.0	5.0	–	–
WAC-Kernel [36]	33.5	7.7	11.4	6.0
ESZSL [137]	28.5	7.4	24.3	6.3
SJE [7]	29.9	–	–	–
ZSLNS [123]	29.1	7.3	24.5	6.8
SynC$_{fast}$[22]	28.0	8.6	18.4	3.8
SynC$_{OVO}$[22]	12.5	5.9	–	–
ZSLPP [38]	37.2	9.7	30.3	8.1
GAZSL [211]	43.7	10.3	35.6	8.6
CANZSL [26]	45.8	14.3	38.1	8.9

6.4. Zero-shot learning results on ImageNet

ImageNet is a large-scale single-labelled dataset with an imbalanced number of data that possesses WordNet hierarchy instead of human-annotated attributes, thus is useful mean to measure the performance of various methods in recognition-in-the-wild scenarios. The performances of 12 state-of-the-art models are reported here. They are ConSE [113], CMT [157], LATEM [185], ALE [6], DeViSE [43], SJE [7], ESZSL [137], SYNC [22], SAE [72], f-CLSWGAN [187], CADA-VAE [145] and f-VAEGAN-D2 [189]. All of the Top-1 accuracies, except for the data generating models are reported from [186] experiments. As it can be understood from Fig. 4(a), feature generating methods have outstanding performance compared to other approaches. Although the results of f-VAEGAN-D2 [189] are available only for 2H, 3H and all splits, it still has the highest accuracies among other models. SYNC [22] and f-CLSWGAN [187] are the next best performing models with approximately the same accuracies. ConSE [113] is a representative model from attribute-classifier based models, as it is also superior to direct compatibility approaches. ESZSL [137], a model with linear compatibility function outperforms the other model within its category. However, in one case, SJE [7] has slightly better accuracy in L500 split setting. It can be interpreted from the figures that on coarse-grained classes, the results are conspicuously better, while fine-grained classes with few images per class have more challenges. However, if the test search space is too big then the accuracies decrease. i.e. M5K has lower accuracies compared to L500 splits, and on 20K split, it is the lowest.

The GZSL results are important in the way that they depict the models' ability to recognise both seen and unseen classes at the test time. The results for the SYNC [22] model is only reported in the L5K setting. As shown in Fig. 4(b), the trend is Similar to ZSL where populated classes have better results than the least populated classes, yet have poor results if the search spaces become too big like the decreasing trends in most and least populated classes. Moreover, data-generating approaches dominate other strategies. CADA-VAE [145] that has the advantages of both cross-modal alignment and data feature synthesising methods, evidently outperforms other models. In one case, i. e M500, it nearly has double the accuracy of f-CLSWGAN [187]. For the semantic embedding category, although ESZSL [137] had better results on ZSL, it falls behind approaches like ALE [6], DeViSE [43] and SJE [7].

7. Applications

During the very recent years, zero-shot learning has proved to be a necessary challenge to-be-solved for different scenarios and applications. The number of demands for learning without accessing to the unseen target concepts is also increasing each year.

Zero-shot learning is widely discussed in the computer vision field, such as object recognition in general, as in [133,140] where they aim to locate the objects beside recognising them. Several other variations of ZSL models are proposed for the same task purpose such as [13,30,126]. Zero-shot emotion recognition [200] has the task of recognising unseen emotions, while zero-shot semantic segmentation aims to segment the

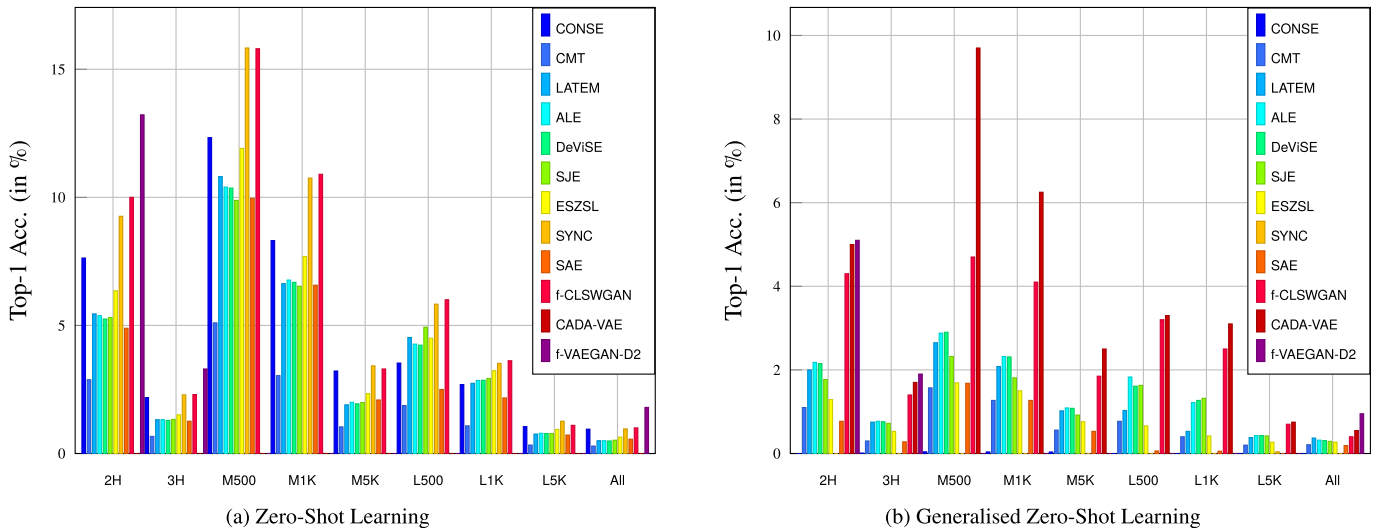


Fig. 4. ImageNet results measured with Top-1 accuracy in % for the 9 splits including 2 and 3 hops away from ImageNet-1K training classes (2H and 3H) and 500, 1K and 5K most (M) and least (L) populated classes, and All the remaining ImageNet-20K classes.

unseen object categories [19,177]. Moreover, on the task of retrieving images from a large scale set of data, Zero-shot has a growing number of research [98,194] along with sketch-based image retrieval systems [34, 35,150]. Zero-shot learning has an application on visual imitation learning to reduce human supervision by automating the exploration of the agent [82,117]. Action recognition is the task of recognising the sequence of actions from the frames of a video. However, if the new actions are not available when training, Zero-shot learning can be a solution, such as in [45,107,124,149]. Zero-shot Style Transfer in an image is the problem of transferring the texture of source image to target image while the style is not pre-determined and it is arbitrary [151]. Zero-shot resolution enhancement problem aims at enhancing the resolution of an image without pre-defined high-resolution images for training examples [154]. Zero-shot scene classification for HSR images [85] and scene-sketch classification has been studied in [191] as other applications of ZSL in computer vision. Zero-shot learning has also left its footprint in the area of NLP. Zero-Shot Entity Linking, links entity mentions in the text using a knowledge base [95]. Many research works focus on the task of translating languages to another without pre-determined translation between pairs of samples [50,53,62,78]. In sentence embedding [10] and in Style transfer of text, a common technique is to convert the source to another style via arbitrary styles like the artistic technique discussed in [21]. In the audio processing field, zero-shot based voice conversion to another speaker's voice [122] is an applicable scenario of ZSL.

In the era of the COVID-19 pandemic, many researchers have tried to work on Artificial Intelligence and Machine learning based methodologies to recognise the positive cases of the COVID-19 patients based on the CT scan images or Chest X-rays. Two prominent features in chest CT used for diagnosis are ground glass opacities (GGO) and consolidation which has been considered by some of the researchers such as [39,92,198], and [146]. [111] uses three CNN models to detect COVID-19, in which the ResNet50 shows a very high rate of classification performance. [146] introduces a deep-learning based system that segments the infected regions and the entire lung in an automatic manner. [184] shows that the increase in unilateral or bilateral Procalcitonin and consolidation with surrounding halo is prominent in chest CT of paediatric patients. [88] introduces the COVNet to extract the 2D local and 3D global features in 3D chest CT slices. The method claims the ability of classifying COVID-19 from community acquired pneumonia (CAP). [152] shows different imaging patterns of the COVID-19 cases depending on the time of infection. [208] classifies four stages to respiratory CT scan changes and shows the most dramatic changes to be in the first 10 days from the onset of initial symptoms. [201] introduces a deep learning based anomaly detection model which extracts the high-level features from the input chest X-ray image. [56] introduce COVIDX-Net to classify the positive cases for the COVID-19 in X-ray images. It uses 7 different architectures, which VGG19 outperforms the others. [3] proposes a COVID-CAPS that is based on the Capsule Networks [57] to avoid the drawbacks of CNN-based architectures as it captures better spatial information. It performs on a small dataset of X-ray images. [1] employs a class decomposition mechanism in DeTraC [2] which is a deep convolutional network that can handle image dataset irregularities of the X-ray images. [203] proposes a method for X-ray medical image segmentation using task driven generative adversarial networks. [128] proposes a 21-layer CNN called CheXNet, trained on the ChestX-ray14 dataset [180] to detect pneumonia with the localisation of the most infected areas from the X-ray images. [139] shows a possible diagnostic criteria could be the existence of bilateral pulmonary areas of consolidation found in the chest X-rays, and [169] uses DenseNet-169 for the purpose of feature extraction followed by an SVM classifier to detect Pneumonia from chest X-ray images.

A common weakness among the majority of the above-mentioned research works is that they either conduct their evaluations on a very limited number of cases due to the lack of comprehensive datasets (which puts the validity of the reported results under a question), or they suffer from underlying uncertainties due to unknown nature and characteristics of the novel COVID-19, not only for the medical community, but also for the machine learning and data analytic experts. In such an uncertain atmosphere with limited training dataset, we strongly recommend the adaptation of Zero-shot learning and its variances (as discussed in Fig. 4) as an efficient deep learning based solution towards COVID-19 diagnosis.

Diagnosis and recognition of the very recent and global challenge of COVID-19 disease caused by the severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2) is a perfect real-world application of Zero-shot learning, where we do not have millions of annotated datasets available; and the symptoms of the disease and the chest X-ray of infected people may significantly vary from person to person. Such a scenario can truly be considered as a novel unseen target or classification challenge. We only know some of the symptoms of the infected people with COVID-19 in forms of advices, text notes, chest X-ray interpenetration, all as the auxiliary data which have partial similarities with other lung inflammatory diseases, such as Asthma or SARS. So, we have to seek for a semantic relationship between training and the new unseen classes. Therefore, ZSL can help us significantly to cope with this new challenge like the induction of the SARS-CoV-2, from previously learned diagnosis of the Asthma, and the Pneumonia using written medical documents of

the respiratory tracts and chest X-ray images. In the case of the few-shot learning, a handful of the chest CT scans or X-ray of the positive cases of the COVID-19 can also be beneficial as further support-set alongside the chest X-ray images of SARS, Asthma and Pneumonia to infer the novel COVID-19 cases.

As a general rule and based on the recent successful applications, we can infer that in any scenarios that the goal is set to reduce supervision, and the target of the problem can be learned through side information and its relation to the seen data, the Zero-shot learning method can be conducted as one of the best learning techniques and practices.

8. Discussion

A typical zero-shot learning problem is usually faced with three popular issues that need to be solved in order to enhance the performance of the model. These issues are Bias, Hubness and domain shift; and every model revolves around solving one or more of the issues mentioned. In this section, we discuss efforts done by different approaches to alleviate bias, hubness and domain-shift and infer the logic each approach owns to learn its model.

Bias. The problem with ZSL and GZSL tasks is that the imbalanced data between training and test classes cause a bias towards seen classes at prediction time. Other reasons for bias could be high-dimensionality and the devoid of manifold structure of features. Several data generating approaches have worked on alleviating bias by synthesising visual data for unseen classes. [187] generates semantically rich CNN features of the unseen classes to make unseen embedding space more known. [106] generates pseudo seen and unseen class features, and then it trains an SVM classifier to mitigate bias. [143] improves the quality of the synthesised examples by using gradient matching loss. Models combining data generation or reconstruction along with other techniques have proved to be effective in alleviating bias. [97] uses an intermediate space to help discover the geometric structure of the features that previously didn't with the regression-based projections. [24] uses calibrated stacking rule. [145] generates latent feature sizes of 64 with the idea that low-dimensional representations tend to mitigate bias. [153] uses two regressors to calculate reconstruction to diminish the bias. Transductive-based approaches like [143] are also used to solve the bias issue. In [159], it forces the unseen classes to be projected into fixed pre-defined points to avoid results with bias.

Hubness [125]. In large-dimensional mapping spaces, samples (hubs) might end up falsely as the nearest neighbours of several other points in the semantic space and result in an incorrect prediction. To avoid the hubness, [176] proposes a stage-wise bidirectional latent embedding framework. When a mapping is done from high-dimensional feature space to a low-dimensional semantic space using regressors, the distinctive features will partially fade while in the visual feature space, the structures are better preserved. Hence, the visual embedding space is well-known for mitigating the hubness problem. [174,202] use the output of the visual space of the CNN as the embedding space.

Domain-shift. Zero-shot learning challenge can be considered as a domain adaptation problem. This is because the source labelled data is disjoint with the target unlabelled domain data. This is called project domain-shift. Domain adaptation techniques are used to learn the intrinsic relationships among these domains and transfer knowledge between the two. A considerable amount of works has been done through a transductive setting which has been successful to overcome the domain-shift issue. [44] a multi-view embedding framework, performs label propagation on graph a heuristic one-stage self-learning approach to assign points to their nearest data points. [71] introduces a regularised sparse coding based unsupervised domain adaptation framework that solves the domain shift problem. [206] uses a structured prediction method to solve the problem by visually clustering the unseen data. [174] uses a visual constraint on the centre of each class when the mapping is being learned. Since the pure definition of the ZSL challenge is the inaccessibility of unseen data during training, several inductive approaches tried to solve the problem as well. [72] proposes to reconstruct the visual features to alleviate this issue. [197] performs sparse non-negative matrix factorisation for both domains in a common semantic dictionary. MFMR [193] exploits the manifold structure of test data with a joint prediction scheme to avoid domain shift. [138] uses entropy minimisation in optimisation. [86] preserves the semantic similarity structure in seen and unseen classes to avoid the domain-shift occurrence. [87] mitigates projection domain-shift by generating soul samples that are related to the semantic descriptions.

These three common issues together with inferiorities of each methods will be a motivation to decide on a particular approach when solving the ZSL problem. Attribute classifiers are considered customised since human-annotations are used; however, this makes the problem a laborious task that has strong supervision. Compatibility learning approaches have the ability to learn directly by eliminating the intermediate step but often face with the bias and hubness problem. Manifold learning solves this weakness of the semantic learning approaches by preserving the geometrical structure of the features. Cross-modal latent embedding approaches take on a different point of view and leverage both visual and semantic features and the similarity and differences between them. They often propose methods for aligning the structures between the two modes of features. This category of methods also suffers from the hubness problem for the problems dealing with high-dimensional data. Visual space embedding approaches have the advantage of turning the problem into a supervised one by generating or aggregating visual instances for the unseen classes. Plus are a favourable approach for solving hubness problem due to the high-dimensionality of the visual space that can preserve information structure better and also bias problem by alleviating the imbalanced data by generating unseen class samples. Here a challenge would be generating more realistic looking data. Another different setting is transductive learning that presents solutions to bias problem, by creating balance in data by gathering unseen data, yet not applicable to many of the real-world problems since the original definition of ZSL limits the use of unseen data during the training phase.

Depending on the real-world scenarios, each way of solving the problem might be the most appropriate choice. Some approaches improve the solution by combining two or more methods to benefit from each one's strengths.

9. Conclusion

In this article, we performed a comprehensive and multi-faceted review on the Zero-Shot/Generalised Zero-shot Learning challenge, its fundamentals, and variants for different scenarios and applications such as COVID-19 diagnosis, Autonomous Vehicles, and similar complex real-world applications which involve fully/partially new concepts that have never/rarely seen before, besides the barrier of limited annotated dataset. We divided the recent state-of-the-art methods into four space-wise embedding categories. We also reviewed different types of side and auxiliary information. We went through the popular datasets and their corresponding splits for the problem of ZSL. The paper also contributed in performing the experiment results for some of the common baselines and elaborated on assessing the advantages and disadvantages of each group, as well as the ideas behind different areas of solutions to improve each group. Our evaluation reveals that data synthesis methods and combinational approaches yield the best performance, as by synthesising data, the problem shifts to the classic recognition/diagnosis problem, and by combining other methods, the model utilises the advantage of each embedding techniques. The models even outperform compatibility learning models in transductive setting. This means, the models consisting a visual data generation step, lead to better results than other approaches and settings. Furthermore, the accuracies improve when the unseen classes have closer semantic hierarchy and relatedness distance to the seen classes. Finally, we reviewed the current and potential real-world applications of ZSL and GZSL in the near future. To the best of

our knowledge, such a comprehensive and detailed technical review and categorisation of the ZSL methodologies, alongside with an efficient solution for the recent challenge of COVID-19 pandemic is not done before; hence, we expect it to be helpful in developing new research directions among AI and health-related research community.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Additional Notes and a Review on Mathematical Formulas

In this appendix, we provide a concise overview of the main specifications, mathematical formulas, and notations of the 26 state-of-the-art methods that we discussed and compared during this research, in a top-down matter.

DAP [80] acronym of Direct Attribute Prediction, first learns the posteriors of the attributes, then estimates

$$f(x) = \underset{U=1,\ldots,N_{y^U}}{\operatorname{argmax}} \prod_{m=1}^{M} \frac{p(a_m^{y^U} \mid x)}{p\left(a_m^{y^U}\right)} \tag{A.1}$$

where N_{y^U} and M are the number of the classes of y^U and the attributes of a, respectively. $a_m^{y^U}$ is the m^{th} attribute of the class y^U, $p\left(a_m^{y^U} \mid x\right)$ is the estimated attribute via attribute classifier for image x, and $p(a_m^{y^U})$ is the prior attributes computed for training classes with the MAP.

IAP [80] is an indirect approach as it first learns the posteriors for seen classes and then uses them to compute the posteriors for the attributes:

$$p(a_m \mid x) = \sum_{y^S=1}^{N_{y^S}} p(a_m \mid y^S)\, p(y^S \mid x) \tag{A.2}$$

where N_{y^S} is the number of training classes, $p(a_m \mid y^S)$ is the pre-trained attribute of the classes and $p(y^S \mid x)$ is probabilistic multi-class classifier to be learned.

ConSE [113] takes a probabilistic approach and predicts an unseen class by the convex combination of the class label embedding vectors. It first learns the probability of the training samples:

$$f(x,t) = \underset{y \in y^S}{\operatorname{argmax}}\ p^S(y \mid x) \tag{A.3}$$

in which y is the most probable label for the training sample. It then computes a weighted combination of the semantic embedding to its probability to find a label for a given unseen image.

$$\mathbb{L} = \frac{1}{Z} \sum_{i=1}^{N_T} p_s(f(x,\ t) \mid x)\ .\ s(f(x,\ t)) \tag{A.4}$$

In this function, Z is the normalisation factor and s combines N_T semantic vectors to infer unseen labels.

Linear corresponding functions are the simplest mapping functions that are typically used to map visual features to semantic spaces in vector form. If the mapping parameters are in the form of matrix, then it's called bi-linear corresponding (compatibility) function. These approaches often include other losses alongside the main mapping function.

ESZSL [137], introduces a better regulariser and optimises a close form solution objective function in a bi-linear manner.

$$\Omega = \alpha \|Wc(y)\|_{Fro}^2 + \beta \left|\left|\theta(x)^T W\right|\right|_{Fro}^2 + \gamma \|W\|_{Fro}^2 \tag{A.5}$$

α, β, and γ are the hyper-parameters. The first two terms are the Frobenius norm of the attribute features and visual features respectively, and the third term is the weight decay penalty of the matrix.

ZSLNS [123] proposes a $l_{1,2}$-norm based loss function and an optimiser based on [137] to help suppress the noise in textual data.

ALE [6] optimises the ranking loss in [167] with a bi-linear mapping compatibility function. The objective function used in ALE is similar to unregularised structured SVM (SSVM) [166].

$$\mathbb{L} = \frac{1}{N} \sum_{n=1}^{N} \max_{y \in y^S} \Delta(y_n, y) + F(x_n, y; W) - F(x_n, y_n; W) \tag{A.6}$$

were $F(.)$ is the compatibility function, W is the matrix with dimensions of image and label embeddings, and Δ is the loss of the mapping function. In spite of having different losses, the inspiration comes from WSABIE algorithm [183]. In ALE, rank 1 loss with a multi-class objective is used instead of all of the weighted ranks.

SJE [7] is similar to ALE. It learns a bi-linear compatibility function using the unregularised structural SVM objective function [166] and train their model on different supervised and unsupervised class embeddings.

DeViSE [43] uses the combination of dot-product similarity and hinge rank loss used in [183] as the objective function.

$$\mathbb{L} = \sum_{c(y)\neq c(j)} \max\left[0,\ \alpha - \theta(x)^T Wc(y) - \theta(x)^T Wc(j)\right] \tag{A.7}$$

Here, α is a hyperparmeter and $c(j)$ are randomly selected word embeddings.

ZSLPP [38] also uses bi-linear corresponding function for a part-based cross-modal framework. The visual part detectors detects bird parts from the images, while the zero-shot classifier performs prediction on the previously detected visual bird parts based on textual side information.

DSRL [197] uses a non-negative sparse matrix factorisation for better feature alignment while learning the compatibility function. And uses label propagation to predict unseen classes. The NMF is computed as follows:

$$\theta(x)^{S\cup U} = \min_{Z\geq 0,\ \varphi\geq 0} \frac{1}{2}\left\|x^{S\cup U} - Z\varphi\right\|_{Fro}^2 + \alpha||\varphi||_1\ \beta\|Z\|_1 \tag{A.8}$$

Here, φ and Z are dictionary and the latent representation of matrix respectively. α and β are the hyperparameters.

SAE [72] or Semantic Auto Encoder, uses an AutoEncoder to combine two linear mapping functions, one for the visual space and the other one for semantic space. In this way, the decoded visual feature, produces semantically meaningful features after the mapping to the semantic space. The objective to be minimised is as follows:

$$\mathbb{L} = \min_{W}\left|\left|\theta(x)^T - W^T c(y)\right|\right|^2 + \lambda\left|\left|W\theta(x)^T - c(y)\right|\right|^2 \tag{A.9}$$

where W^T and W are decoder and encoder projection matrices. And λ is a hyperparameter.

WAC-Linear [37] combines a regression function that solely maps semantic features to the visual space, and a knowledge transfer function, to map the textual descriptions to the visual space.

Some approaches use non-linear compatibility functions to solve the ZSL challenge.

CMT [157] uses a two-layer neural network, similar to common MLP networks by [131] that minimises the objective function

$$\mathbb{L} = \sum_{y\in y^S}\sum_{x\in X_y}\left\|c(y) - \theta^{(2)} tanh\left(\theta^{(1)}x^{(i)}\right)\right\|^2 \tag{A.10}$$

$\theta = (\theta^{(1)}, \theta^{(2)})$.

GFZSL [171] introduces both linear and non-linear regression models in a generative approach as it produces a probability distribution for each class. It then uses MLE for estimating seen class parameters and two regression functions for unseen categories.

$$\mu_y = f_\mu(c(y)) \tag{A.11}$$

$$\sigma_y^2 = f_\sigma^2(c(y)) \tag{A.12}$$

where μ is the Gaussian mean vector and σ is the diagonal Covariance matrix of the attribute vector. In its transductive setting, it uses Expectation-Maximisation (EM) that works like estimation of a Gaussian Mixture Model (GMM) of unlabelled data in an iterative manner. The inferred labels will be included in the next iterations.

LATEM [185] learns several mappings and selects one to be the latent variable for a pair of image and class. The selected latent embedding learns a piece-wise non-linear compatibility function alongside a ranking loss. Its compatibility function is

$$F(x, y; w) = \max_{1\leq i\leq K} w_i(\theta(x)^T \otimes c(y)) \tag{A.13}$$

$i = i, \ldots, K$ with $K \geq 2$ are the indexes over latent choices.

DEM [202] and WAC-Kernel [36] learn non-linear mapping in the inverse direction from different types of class embeddings. i.e. textual data. WAC-Kernel uses a kernel method for the integration of side information. The objective function for DEM is

$$\mathbb{L} = \left|\left|\theta(x)^T - Wc(y)\right|\right|_{Fro}^2 + \lambda||W||_{Fro}^2 \tag{A.14}$$

that looks like a ridge regression.

Some of the methods consider cross-modal feature similarity in a mutual space.

SSE [204] learns two embedding functions, one being ψ which is learned from seen class auxiliary information and the other one from seen data which is target class π embedding and predicts unseen labels via maximising the similarity between histograms:

$$\mathbb{L} = \underset{y\in y^u}{\text{argmax}}\ \pi(\theta(x))^T \psi(c(y)) \tag{A.15}$$

SYNC [22] considers the mapping between the semantic space of the external information and the model space. it introduces phantom classes to align the two spaces. The classifier is trained with the sparse linear combination of the classifiers for the phantom classes:

$$\mathbb{L} = \min_{w_c, v_r}\left\|w_c - \sum_{r=1}^{R} s_{cr} v_r\right\|_2^2 \tag{A.16}$$

where w_c and v_r are weighted graphs of the real and phantom classes respectively. While s_{cr} is the bipartite graph of those two previously graph

combinations.

MCZSL [4] combines compatibility learning with Deep Fragment embeddings [67] in a joint space. Their visual part and multi-cue language embedding are defined as follows, respectively:

$$\theta_i = E^{\text{visual}}\left[\text{CNN}_\theta(I_b) + b^{\text{visual}}\right] \tag{A.17}$$

$$c(y)_j = f\left(\sum_m E_m^{\text{language}} l_m + b^{\text{language}}\right) \tag{A.18}$$

In this equation, l_m and E_m^{language} are the language encoder for each modality. $f(.)$ is the language token from the m modality and ReLU, respectively. Also, E^{visual} is the visual encoder and $\text{CNN}_\theta(I_b)$ is the part descriptor extracted from bounding box I_b for the image part annotation b. Hence the complete objective function is as follows:

$$\mathbb{L} = \sum_i \sum_j \max\left(0, 1 - y_{ij}\theta_i^T c(y)_j\right) + \alpha\|w\|_2^2 \tag{A.19}$$

where w is the parameters of the two encoders and α is the hyperparameter.

Several methods decide to generate images or image features using different visual data synthesis techniques. some of them are VAE-based [69].

CADA-VAE [145] learns latent space features and class embedding by training VAE [69] for both visual and semantic modalities. It uses the Cross-Alignment (CA) Loss to align latent distributions in cross-modal reconstruction:

$$\mathbb{L}_{CA} = \sum_i^M \sum_{j\neq i}^M \left|x^{(j)} - D_j\left(E_i\left(x^{(i)}\right)\right)\right| \tag{A.20}$$

where, i and j are two different modalities. Wasserstein distance [46] is used between the latent distributions i and j to align the Latent distribution (LDA):

$$\mathbb{L}_{DA} = \sum_i^M \sum_{j\neq i}^M W_{ij} \tag{A.21}$$

where $W_{ij} = \left(\left\|\mu_i - \mu_j\right\|_2^2 + \left\|\eta_i^{\frac{1}{2}} - \eta_j^{\frac{1}{2}}\right\|_{\text{Fro}}^2\right)^{\frac{1}{2}}$.

μ and η are predictions of the encoder.

SE-GZSL [75] adds an extra loss term named "feedback loss" to the VAE objective [69] function that works as a discriminator to enforce the similarity of the generated sample to the original distribution. The regressor feedback term is as follows:

$$\mathbb{L}_{Reg} = -\mathbb{E}[\log \text{G}(\tilde{x}|z, c(y)] \tag{A.22}$$

where z is a random noise.

CVAE-ZSL [106] uses a Conditional variations AutoEncoder (CVAE) [158], conditioned on attributes, alongside the L_2 norm for reconstruction loss. The objective function of a CVAE is:

$$\mathbb{L} = \mathbb{E}_{q_\varphi(z|x,c)}\left[\log p_\theta(x \mid z, c)\right] - D_{KL}(q_\varphi(z \mid x, c)||\, p_\theta(z \mid c)) \tag{A.23}$$

where c is the condition. It then trains a SVM classifier [28] for unseen categories.

Other approaches introduce GAN-based [48] methods.

GAZSL [211] adds a visual pivot regulariser to GAN's objective function. This regulariser aims to reduce the noise of Wikipedia articles.

$$\Omega = \frac{1}{N}\sum_{n=1}^{n}\left\|\mathbb{E}_{\tilde{x}\sim p_g}[\tilde{x}] - \mathbb{E}_{x\sim p_{data}}[x]\right\|^2 \tag{A.24}$$

Due to the inaccessibility of data, empirical expectations

$$\mathbb{E}_{x\sim p_s}[\tilde{x}] = \frac{1}{N}\sum_{i=1}^{N_S} x^i \tag{A.25}$$

and

$$\mathbb{E}_{x\sim p}[\tilde{x}] = \frac{1}{N^U}\sum_{i=1}^{N^U} G_\theta(T_U, z_i) \tag{A.26}$$

are used instead; where N^S and N^U are the number of samples in class y^S and number of synthesised features in class y^U, respectively.

f-CLSWGAN [187] combines three conditional GAN [48] variants: GAN, conditional WGAN [51] and a classification loss, and name their method

f-CLSWGAN.

$$\mathbb{L} = \min_{G}\max_{D} \mathbb{L}_{WGAN} + \beta\, \mathbb{L}_{CLS} \tag{A.27}$$

The classification loss is like a regulariser for the enhancement of the generated features and, β is a hyperparameter.

CANZSL [26] uses GAN for generating visual features and an inverse GAN to project them back to the semantic space. In this way, the produced features are consistent with their corresponding semantic features.

f-VAEGAN-D2 [189] introduces a generative model that integrates VAE and WGAN. In this model, the decoder of VAE and the generator of the WGAN are the same component, and there are two discriminators (D_1, D_2) for this model. The full objective function to be optimised is as follows:

$$\mathbb{L} = \min_{G,E}\max_{D_1,D_2} \mathbb{L}_{VAE} + \mathbb{L}_{WGAN} \tag{A.28}$$

References

[1] Abbas A, Abdelsamea MM, Gaber MM. Classification of covid-19 in chest x-ray images using detrac deep convolutional neural network. 2020. https://doi.org/10.1101/2020.03.30.20047456. arXiv preprint arXiv:2003.13815.

[2] Abbas A, Abdelsamea MM, Gaber MM. Detrac: transfer learning of class decomposed medical images in convolutional neural networks. IEEE Access 2020b;8:74901–13. https://doi.org/10.1109/ACCESS.2020.2989273.

[3] Afshar P, Heidarian S, Naderkhani F, Oikonomou A, Plataniotis KN, Mohammadi A. Covid-caps: a capsule network-based framework for identification of covid-19 cases from x-ray images. 2020. arXiv preprint arXiv:2004.02696.

[4] Akata Z, Malinowski M, Fritz M, Schiele B. Multi-cue zero-shot learning with strong supervision. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 59–68. https://doi.org/10.1109/CVPR.2016.14.

[5] Akata Z, Perronnin F, Harchaoui Z, Schmid C. Label-embedding for attribute-based classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2013. p. 819–26. https://doi.org/10.1109/CVPR.2013.111.

[6] Akata Z, Perronnin F, Harchaoui Z, Schmid C. Label-embedding for image classification. IEEE Trans Pattern Anal Mach Intell 2015a;38:1425–38. https://doi.org/10.1109/TPAMI.2015.2487986.

[7] Akata Z, Reed S, Walter D, Lee H, Schiele B. Evaluation of output embeddings for fine-grained image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 2927–36. https://doi.org/10.1109/CVPR.2015.7298911.

[8] Al-Halah Z, Tapaswi M, Stiefelhagen R. Recovering the missing link: predicting class-attribute associations for unsupervised zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 5975–84. https://doi.org/10.1109/CVPR.2016.643.

[9] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: International conference on machine learning; 2017. p. 214–23. https://dl.acm.org/doi/10.5555/3305381.3305404.

[10] Artetxe M, Schwenk H. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. ACM Transactions on Intelligent Trans Assoc Comput Linguist 2019;7:597–610. https://doi.org/10.1162/tacl_a_00288.

[11] Arvanaghi M, Rezaei M. Facial age estimation using hybrid haar wavelet and color features with support vector regression. In: 2017 artificial intelligence and robotics (IRANOPEN); 2017. p. 6–12. https://doi.org/10.1109/RIOS.2017.7956436.

[12] Atzmon Y, Chechik G. Probabilistic and-or attribute grouping for zero-shot learning. 2018. arXiv preprint arXiv:1806.02664.

[13] Bansal A, Sikka K, Sharma G, Chellappa R, Divakaran A. Zero-shot object detection. In: Proceedings of the European conference on computer vision (ECCV); 2018. p. 384–400. https://doi.org/10.1007/978-3-030-01246-5_24.

[14] Bart E, Ullman S. Cross-generalization: learning novel classes from a single example by feature replacement. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), IEEE; 2005. p. 672–9. https://doi.org/10.1109/CVPR.2005.117.

[15] Bay H, Ess A, Tuytelaars T, Van Gool L. Speeded-up robust features (surf). Comput Vis Image Understand 2008;110:346–59. https://doi.org/10.1016/j.cviu.2007.09.014.

[16] Bosch A, Zisserman A, Munoz X. Representing shape with a spatial pyramid kernel. In: Proceedings of the 6th ACM international conference on Image and video retrieval; 2007. p. 401–8. https://doi.org/10.1145/1282280.1282340.

[17] Bucher M, Herbin S, Jurie F. Improving semantic embedding consistency by metric learning for zero-shot classiffication. In: European conference on computer vision, Springer; 2016. p. 730–46. https://doi.org/10.1007/978-3-319-46454-1_44.

[18] Bucher M, Herbin S, Jurie F. Generating visual representations for zero-shot classification. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 2666–73. https://doi.org/10.1109/ICCVW.2017.308.

[19] Bucher M, Vu TH, Cord M, Pérez P. Zero-shot semantic segmentation. 2019. arXiv preprint arXiv:1906.00817.

[20] Callan J, Hoy M, Yoo C, Zhao L. Clueweb09 data set. 2009.

[21] Carlson K, Riddell A, Rockmore D. Zero-shot style transfer in text using recurrent neural networks. 2017. ArXiv Machine Learning doi:abs/1711.04731.

[22] Changpinyo S, Chao WL, Gong B, Sha F. Synthesized classifiers for zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 5327–36. https://doi.org/10.1109/CVPR.2016.575.

[23] Changpinyo S, Chao WL, Sha F. Predicting visual exemplars of unseen classes for zero-shot learning. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 3476–85. https://doi.org/10.1109/ICCV.2017.376.

[24] Chen L, Zhang H, Xiao J, Liu W, Chang SF. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 1043–52. https://doi.org/10.1109/CVPR.2018.00115.

[25] Chen X, Yao L, Zhou T, Dong J, Zhang Y. Momentum contrastive learning for few-shot covid-19 diagnosis from chest ct images. 2020. arXiv preprint arXiv:2006.13276.

[26] Chen Z, Li J, Luo Y, Huang Z, Yang Y. Canzsl: cycle-consistent adversarial networks for zero-shot learning from natural language. 2019. https://doi.org/10.1109/WACV45572.2020.9093610. arXiv preprint arXiv:1909.09822.

[27] Cohen JP, Morrison P, Dao L. Covid-19 image data collection. arXiv 2003.11597 URL: https://github.com/ieee8023/covid-chestxray-dataset; 2020.

[28] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273–97.

[29] Cristianini N, Shawe-Taylor J, Elisseeff A, Kandola JS. On kernel-target alignment, in: advances in neural information processing systems. 2002. p. 367–73. https://doi.org/10.1007/3-540-33486-6_8.

[30] Demirel B, Cinbis RG, Ikizler-Cinbis N. Zero-shot object detection by hybrid region embedding. 2018. ArXiv Computer Vision and Pattern Recognition doi:arXiv:1805.06157.

[31] Deng J, Ding N, Jia Y, Frome A, Murphy K, Bengio S, Li Y, Neven H, Adam H. Large-scale object classification using label relation graphs. In: European conference on computer vision, Springer; 2014. p. 48–64. https://doi.org/10.1007/978-3-319-10590-1_4.

[32] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, Ieee; 2009. p. 248–55. https://doi.org/10.1109/CVPR.2009.5206848.

[33] Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. 2018. https://doi.org/10.18653/v1/N19-1423. arXiv preprint arXiv:1810.04805.

[34] Dey S, Riba P, Dutta A, Llados J, Song YZ. Doodle to search: practical zero-shot sketch-based image retrieval. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2019. p. 2179–88. https://doi.org/10.1109/CVPR.2019.00228.

[35] Dutta A, Akata Z. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2019. p. 5089–98. https://doi.org/10.1109/CVPR.2019.00523.

[36] Elhoseiny M, Elgammal A, Saleh B. Write a classifier: predicting visual classifiers from unstructured text. IEEE Trans Pattern Anal Mach Intell 2016;39:2539–53. https://doi.org/10.1109/TPAMI.2016.2643667.

[37] Elhoseiny M, Saleh B, Elgammal A. Write a classifier: zero-shot learning using purely textual descriptions. In: Proceedings of the IEEE international conference on computer vision; 2013. p. 2584–91. https://doi.org/10.1109/ICCV.2013.321.

[38] Elhoseiny M, Zhu Y, Zhang H, Elgammal A. Link the head to the" beak": zero shot learning from noisy text description at part precision. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), IEEE; 2017. p. 6288–97. https://doi.org/10.1109/CVPR.2017.666.

[39] Fang Y, Zhang H, Xie J, Lin M, Ying L, Pang P, Ji W. Sensitivity of chest ct for covid-19: comparison to rt-pcr. Radiology 2020:200432. https://doi.org/10.1148/radiol.2020200432.

[40] Farhadi A, Endres I, Hoiem D, Forsyth D. Describing objects by their attributes. In: 2009 IEEE conference on computer vision and pattern recognition, IEEE; 2009. p. 1778–85. https://doi.org/10.1109/CVPR.2009.5206772.

[41] Fei-Fei L, Fergus R, Perona P. One-shot learning of object categories. IEEE Trans Pattern Anal Mach Intell 2006;28:594–611. https://doi.org/10.1109/TPAMI.2006.79.

[42] Felix R, Kumar VBG, Reid I, Carneiro G. Multi-modal cycle-consistent generalized zero-shot learning. In: Proceedings of the European conference on computer vision (ECCV); 2018. p. 21–37. https://doi.org/10.1007/978-3-030-01231-1_2.

[43] Frome A, Corrado GS, Shlens J, Bengio S, Dean J, Ranzato M, Mikolov T. Devise: a deep visual-semantic embedding model. In: Advances in neural information processing systems; 2013. p. 2121–9. doi:10.1.1.466.176.
[44] Fu Y, Hospedales TM, Xiang T, Gong S. Transductive multi-view zero-shot learning. IEEE Trans Pattern Anal Mach Intell 2015;37:2332–45. https://doi.org/10.1109/TPAMI.2015.2408354.
[45] Gao J, Zhang T, Xu C. I know the relationships: zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In: Proceedings of the AAAI conference on artificial intelligence; 2019. p. 8303–11. https://doi.org/10.1609/aaai.v33i01.33018303.
[46] Givens CR, Shortt RM, Others. A class of wasserstein metrics for probability distributions. Mich Math J 1984;31:231–40. https://doi.org/10.1307/mmj/1029003026.
[47] Gong Y, Ke Q, Isard M, Lazebnik S. A multi-view embedding space for modeling internet images, tags, and their semantics. Int J Comput Vis 2014;106:210–33. https://doi.org/10.1007/s11263-013-0658-4.
[48] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Advances in neural information processing systems; 2014. p. 2672–80. https://doi.org/10.5555/2969033.2969125.
[49] Gretton A, Borgwardt K, Rasch M, Schölkopf B, Smola AJ. A kernel method for the two-sample-problem. In: Advances in neural information processing systems; 2007. p. 513–20. https://doi.org/10.5555/2976456.2976521.
[50] Gu J, Wang Y, Cho K, Li VOK. Improved zero-shot neural machine translation via ignoring spurious correlations. 2019. https://doi.org/10.18653/v1/P19-1121. arXiv preprint arXiv:1906.01181.
[51] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. Improved training of wasserstein gans. In: Advances in neural information processing systems; 2017. p. 5767–77. https://doi.org/10.5555/3295222.3295327.
[52] Guo Y, Ding G, Han J, Gao Y. Zero-shot recognition via direct classifier learning with transferred samples and pseudo labels. In: Thirty-first AAAI conference on artificial intelligence; 2017. https://dl.acm.org/doi/abs/10.5555/3298023.3298158.
[53] Ha TL, Niehues J, Waibel A. Effective strategies in zero-shot neural machine translation. arXiv preprint arXiv:1711.07893, https://doi.org/10.1162/neco.1997.9.8.1735; 2017.
[54] Harris ZS. Distributional structure. 1954. https://doi.org/10.1080/00437956.1954.11659520.
[55] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–8. https://doi.org/10.1109/CVPR.2016.90.
[56] Hemdan EED, Shouman MA, Karar ME. Covidx-net: a framework of deep learning classifiers to diagnose covid-19 in x-ray images. 2020. https://doi.org/10.1016/j.cmpb.2020.105581. arXiv preprint arXiv:2003.11055 doi:.
[57] Hinton GE, Sabour S, Frosst N. Matrix capsules with em routing. OpenReview. 2018.
[58] Huang H, Wang C, Yu PS, Wang CD. Generative dual adversarial network for generalized zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2019. p. 801–10. https://doi.org/10.1109/CVPR.2019.00089.
[59] Jayaraman D, Grauman K. Zero-shot recognition with unreliable attributes. In: Advances in neural information processing systems; 2014. p. 3464–72. https://dl.acm.org/doi/10.5555/2969033.2969213.
[60] Ji Z, Fu Y, Guo J, Pang Y, Zhang ZM, Others. Stacked semantics-guided attention model for fine-grained zero-shot learning. In: Advances in neural information processing systems; 2018. p. 5995–6004. https://dl.acm.org/doi/abs/10.5555/3327345.3327499.
[61] Jiang H, Wang R, Shan S, Chen X. Learning class prototypes via structure alignment for zero-shot recognition. In: Proceedings of the European conference on computer vision. ECCV); 2018. p. 118–34. https://doi.org/10.1007/978-3-030-01249-6_8.
[62] Johnson M, Schuster M, Le QV, Krikun M, Wu Y, Chen Z, Thorat N, Viégas F, Wattenberg M, Corrado G, Others. Google's multilingual neural machine translation system: enabling zero-shot translation. ACM Transactions on Intelligent Trans Assoc Comput Linguist 2017;5:339–51. https://doi.org/10.1162/tacl_a_00065.
[63] Kampffmeyer M, Chen Y, Liang X, Wang H, Zhang Y, Xing EP. Rethinking knowledge graph propagation for zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2019. p. 11487–96. https://doi.org/10.1109/CVPR.2019.01175.
[64] Kang B, Liu Z, Wang X, Yu F, Feng J, Darrell T. Few-shot object detection via feature reweighting. In: Proceedings of the IEEE international conference on computer vision; 2019. p. 8420–9. https://doi.org/10.1109/ICCV.2019.00851.
[65] Kankuekul P, Kawewong A, Tangruamsub S, Hasegawa O. Online incremental attribute-based zero-shot learning. In: 2012 IEEE conference on computer vision and pattern recognition, IEEE; 2012. p. 3657–64. https://doi.org/10.1109/CVPR.2012.6248112.
[66] Karessli N, Akata Z, Schiele B, Bulling A. Gaze embeddings for zero-shot image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 4525–34. https://doi.org/10.1109/CVPR.2017.679.
[67] Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 3128–37. https://doi.org/10.1109/CVPR.2015.7298932.
[68] Kim M, Zuallaert J, De Neve W. Few-shot learning using a small-sized dataset of high-resolution fundus images for glaucoma diagnosis. In: Proceedings of the 2nd international workshop on multimedia for personal health and health care; 2017. p. 89–92. https://doi.org/10.1145/3132635.3132650.
[69] Kingma DP, Welling M. Auto-encoding variational bayes. 2013. https://arxiv.org/abs/1312.6114.
[70] Koch G, Zemel R, Salakhutdinov R. Siamese neural networks for one-shot image recognition. In: ICML deep learning workshop, Lille; 2015.
[71] Kodirov E, Xiang T, Fu Z, Gong S. Unsupervised domain adaptation for zero-shot learning. In: Proceedings of the IEEE international conference on computer vision; 2015. p. 2452–60. https://doi.org/10.1109/ICCV.2015.282.
[72] Kodirov E, Xiang T, Gong S. Semantic autoencoder for zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 3174–83. https://doi.org/10.1109/CVPR.2017.473.
[73] Kolouri S, Rostami M, Owechko Y, Kim K. Joint dictionaries for zero-shot learning. In: Thirty-second AAAI conference on artificial intelligence; 2018.
[74] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems; 2012. p. 1097–105. https://doi.org/10.1145/3065386.
[75] Kumar Verma V, Arora G, Mishra A, Rai P. Generalized zero-shot learning via synthesized examples. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 4281–9. https://doi.org/10.1109/CVPR.2018.00450.
[76] Lake B, Salakhutdinov R, Gross J, Tenenbaum J. One shot learning of simple visual concepts. Proceedings of the annual meeting of the cognitive science society. 2011. 10.1.1.207.8634.
[77] Lake BM, Salakhutdinov R, Tenenbaum JB. Human-level concept learning through probabilistic program induction. Science 2015;350:1332–8. https://doi.org/10.1126/science.aab3050.
[78] Lakew SM, Lotito QF, Negri M, Turchi M, Federico M. Improving zero-shot translation of low-resource languages. 2018. arXiv preprint arXiv:1811.01389.
[79] Lampert CH, Nickisch H, Harmeling S. Learning to detect unseen object classes by between-class attribute transfer. In: 2009 IEEE conference on computer vision and pattern recognition, IEEE; 2009. p. 951–8. https://doi.org/10.1109/CVPR.2009.5206594.
[80] Lampert CH, Nickisch H, Harmeling S. Attribute-based classification for zero-shot visual object categorization. IEEE Trans Pattern Anal Mach Intell 2013;36:453–65. https://doi.org/10.1109/TPAMI.2013.140.
[81] Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. Albert: a lite bert for self-supervised learning of language representations. 2019. arXiv preprint arXiv:1909.11942.
[82] Lázaro-Gredilla M, Lin D, Guntupalli JS, George D. Beyond imitation: zero-shot task transfer on robots by learning concepts as cognitive programs. Science Robotics 2019;4. https://doi.org/10.1126/scirobotics.aav3150. eaav3150.
[83] Lee CW, Fang W, Yeh CK, Frank Wang YC. Multi-label zero-shot learning with structured knowledge graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 1576–85. https://doi.org/10.1109/CVPR.2018.00170.
[84] Lei Ba J, Swersky K, Fidler S, Others. Predicting deep zero-shot convolutional neural networks using textual descriptions. In: Proceedings of the IEEE international conference on computer vision; 2015. p. 4247–55. https://doi.org/10.1109/ICCV.2015.483.
[85] Li A, Lu Z, Wang L, Xiang T, Wen JR. Zero-shot scene classification for high spatial resolution remote sensing images. IEEE Trans Geosci Rem Sens 2017a;55:4157–67. https://doi.org/10.1109/TGRS.2017.2689071.
[86] Li C, Ye X, Yang H, Han Y, Li X, Jia Y. Generalized zero shot learning via synthesis pseudo features. IEEE Access 2019a;7:87827–36. https://doi.org/10.1109/ACCESS.2019.2925093.
[87] Li J, Jin M, Lu K, Ding Z, Zhu L, Huang Z. Leveraging the invariant side of generative zero-shot learning. 2019. https://doi.org/10.1109/CVPR.2019.00758. arXiv preprint arXiv:1904.04092.
[88] Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, Bai J, Lu Y, Fang Z, Song Q, Others. Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct. Radiology 2020:200905. https://doi.org/10.1148/radiol.2020200905.
[89] Li Y, Swersky K, Zemel R. Generative moment matching networks. In: International conference on machine learning; 2015. p. 1718–27. https://dl.acm.org/doi/10.5555/3045118.3045301.
[90] Li Y, Wang D. Zero-shot learning with generative latent prototype model. 2017. arXiv preprint arXiv:1705.09474.
[91] Li Y, Wang D, Hu H, Lin Y, Zhuang Y. Zero-shot recognition using dual visual-semantic mapping paths. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 3279–87. https://doi.org/10.1109/CVPR.2017.553.
[92] Li Y, Xia L. Coronavirus disease 2019 (covid-19): role of chest ct in diagnosis and management. Am J Roentgenol 2020:1–7. https://doi.org/10.2214/AJR.20.22954.
[93] Li Y, Zhang J, Zhang J, Huang K. Discriminative learning of latent features for zero-shot recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 7463–71. https://doi.org/10.1109/CVPR.2018.00779.
[94] Liu S, Long M, Wang J, Jordan MI. Generalized zero-shot learning with deep calibration network. In: Advances in neural information processing systems; 2018. p. 2005–15. https://doi.org/10.1109/LSP.2020.2977498.
[95] Logeswaran L, Chang MW, Lee K, Toutanova K, Devlin J, Lee H. Zero-shot entity linking by reading entity descriptions. 2019. https://doi.org/10.18653/v1/P19-1335. arXiv preprint arXiv:1906.07348.

[96] Long Y, Liu L, Shao L. Towards fine-grained open zero-shot learning: inferring unseen visual features from attributes. In: 2017 IEEE winter conference on applications of computer vision (WACV), IEEE; 2017. p. 944–52. https://doi.org/10.1109/WACV.2017.110.
[97] Long Y, Liu L, Shao L, Shen F, Ding G, Han J. From zero-shot learning to conventional supervised classification: unseen visual data synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 1627–36. https://doi.org/10.1109/CVPR.2017.653.
[98] Long Y, Liu L, Shen Y, Shao L. Towards affordable semantic searching: zero-shot retrieval via dominant attributes. In: Thirty-second AAAI conference on artificial intelligence; 2018.
[99] Lowe DG. Distinctive image features from scale-invariant keypoints. Int J Comput Vis 2004;60:91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94.
[100] Lu Y. Unsupervised learning on neural network outputs: with application in zero-shot learning. 2015. arXiv preprint arXiv:1506.00990.
[101] Makhzani A, Shlens J, Jaitly N, Goodfellow I, Frey B. Adversarial autoencoders. 2015. arXiv preprint arXiv:1511.05644.
[102] Marino K, Salakhutdinov R, Gupta A. The more you know: using knowledge graphs for image classification. 2016. https://doi.org/10.1109/CVPR.2017.10. arXiv preprint arXiv:1612.04844.
[103] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems; 2013. p. 3111–9. https://dl.acm.org/doi/10.5555/2999792.2999959.
[104] Miller EG, Matsakis NE, Viola PA. Learning from one example through shared densities on transforms. In: Proceedings IEEE conference on computer vision and pattern recognition. IEEE; 2000. p. 464–71. https://doi.org/10.1109/CVPR.2000.855856. CVPR 2000 (Cat. No. PR00662).
[105] Miller GA. Wordnet: a lexical database for English. Commun ACM 1995;38:39–41. https://doi.org/10.1145/219717.219748.
[106] Mishra A, Krishna Reddy S, Mittal A, Murthy HA. A generative model for zero shot learning using conditional variational autoencoders. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops; 2018. p. 2188–96. https://doi.org/10.1109/CVPRW.2018.00294.
[107] Mishra A, Verma VK, Reddy MSK, Arulkumar S, Rai P, Mittal A. A generative approach to zero-shot and few-shot action recognition. In: 2018 IEEE winter conference on applications of computer vision (WACV), IEEE; 2018. p. 372–80. https://doi.org/10.1109/WACV.2018.00047.
[108] Mukherjee T, Hospedales T. Gaussian visual-linguistic embedding for zero-shot recognition. In: Proceedings of the 2016 conference on empirical methods in natural language processing; 2016. p. 912–8. https://doi.org/10.18653/v1/D16-1089.
[109] Mukherjee T, Yamada M, Hospedales TM. Deep matching autoencoders. 2017. arXiv preprint arXiv:1711.06047.
[110] Murphy KP. Machine learning: a probabilistic perspective. MIT press; 2012. https://doi.org/10.1080/09332480.2014.914768.
[111] Narin A, Kaya C, Pamuk Z. Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. 2020. arXiv preprint arXiv:2003.10849.
[112] Niu L, Veeraraghavan A, Sabharwal A. Webly supervised learning meets zero-shot learning: a hybrid approach for fine-grained classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 7171–80. https://doi.org/10.1109/CVPR.2018.00749.
[113] Norouzi M, Mikolov T, Bengio S, Singer Y, Shlens J, Frome A, Corrado GS, Dean J. Zero-shot learning by convex combination of semantic embeddings. 2013. ArXiv Machine Learning doi:arXiv:1312.5650.
[114] Palatucci M, Pomerleau D, Hinton GE, Mitchell TM. Zero-shot learning with semantic output codes. In: Advances in neural information processing systems; 2009. p. 1410–8. https://dl.acm.org/doi/10.5555/2984093.2984252.
[115] Parikh D, Grauman K. Relative attributes. In: 2011 international conference on computer vision, IEEE; 2011. p. 503–10. https://doi.org/10.1109/ICCV.2011.6126281.
[116] Parker R, Graff D, Kong J, Chen K, Maeda K. English gigaword fifth edition, linguistic data consortium. Google Scholar; 2011.
[117] Pathak D, Mahmoudieh P, Luo G, Agrawal P, Chen D, Shentu Y, Shelhamer E, Malik J, Efros AA, Darrell T. Zero-shot visual imitation. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops; 2018. p. 2050–3. https://doi.org/10.1109/CVPRW.2018.00278.
[118] Patterson G, Hays J. Sun attribute database: discovering, annotating, and recognizing scene attributes. In: 2012 IEEE conference on computer vision and pattern recognition, IEEE; 2012. p. 2751–8. https://doi.org/10.1109/CVPR.2012.6247998.
[119] Pennington J, Socher R, Manning C. Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP); 2014. p. 1532–43. https://doi.org/10.3115/v1/D14-1162.
[120] Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep contextualized word representations. 2018. https://doi.org/10.18653/v1/N18-1202. arXiv preprint arXiv:1802.05365.
[121] Prabhu VU. Few-shot learning for dermatological disease diagnosis. Ph.D. thesis. Georgia Institute of Technology; 2019.
[122] Qian K, Zhang Y, Chang S, Yang X, Hasegawa-Johnson M. Autovc: zero-shot voice style transfer with only autoencoder loss. In: International conference on machine learning; 2019. p. 5210–9.
[123] Qiao R, Liu L, Shen C, Van Den Hengel A. Less is more: zero-shot learning from online textual documents with noise suppression. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 2249–57. https://doi.org/10.1109/CVPR.2016.247.
[124] Qin J, Liu L, Shao L, Shen F, Ni B, Chen J, Wang Y. Zero-shot action recognition with error-correcting output codes. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 2833–42. https://doi.org/10.1109/CVPR.2017.117.
[125] Radovanović M, Nanopoulos A, Ivanović M. Hubs in space: popular nearest neighbors in high-dimensional data. J Mach Learn Res 2010;11:2487–531.
[126] Rahman S, Khan S, Porikli F. Zero-shot object detection: learning to simultaneously recognize and localize novel concepts. In: Asian conference on computer vision, Springer; 2018. p. 547–63. https://doi.org/10.1007/978-3-030-20887-5_34.
[127] Rajan D, Thiagarajan JJ, Karargyris A, Kashyap S. Self-training with improved regularization for few-shot chest x-ray classification. 2020. arXiv preprint arXiv:2005.02231.
[128] Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz C, Shpanskaya K, Others. Chexnet: radiologist-level pneumonia detection on chest x-rays with deep learning. 2017. arXiv preprint arXiv:1711.05225.
[129] Reed S, Akata Z, Lee H, Schiele B. Learning deep representations of fine-grained visual descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 49–58. https://doi.org/10.1109/CVPR.2016.13.
[130] Rezaei M. Creating a cascade of haar-like classifiers: step by step. 2014.
[131] Rezaei M, Isehaghi M. An efficient method for license plate localization using multiple statistical features in a multilayer perceptron neural network 2019; 7–13doi. https://doi.org/10.1109/aiar.2018.8769804.
[132] Rezaei M, Klette R. Look at the driver, look at the road: No distraction! no accident!. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2014. p. 129–36. https://doi.org/10.1109/CVPR.2014.24.
[133] Rezaei M, Terauchi M, Klette R. Robust vehicle detection and distance estimation under challenging lighting conditions. IEEE Trans Intell Transport Syst 2015: 2723–43. https://doi.org/10.1109/TITS.2015.2421482.
[134] Rohrbach M, Ebert S, Schiele B. Transfer learning in a transductive setting. In: Advances in neural information processing systems; 2013. p. 46–54. https://dl.acm.org/doi/10.5555/2999611.2999617.
[135] Rohrbach M, Stark M, Schiele B. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In: CVPR 2011, IEEE; 2011. p. 1641–8. https://doi.org/10.1109/CVPR.2011.5995627.
[136] Rohrbach M, Stark M, Szarvas G, Gurevych I, Schiele B. What helps where–and why? semantic relatedness for knowledge transfer. In: 2010 IEEE computer society conference on computer vision and pattern recognition, IEEE; 2010. p. 910–7. https://doi.org/10.1109/CVPR.2010.5540121.
[137] Romera-Paredes B, Torr P. An embarrassingly simple approach to zero-shot learning. In: International conference on machine learning; 2015. p. 2152–61. https://doi.org/10.1007/978-3-319-50077-5_2.
[138] Rostami M, Kolouri S, Murez Z, Owekcho Y, Eaton E, Kim K. Zero-shot image classification using coupled dictionary embedding. 2019. arXiv preprint arXiv:1906.10509.
[139] Rutigliano I, Gorgoglione S, Pacilio A, De Meco C, Sacco MC. Chronic eosinophilic pneumonia: a pediatric case. Clin Med Rev Case Rep 2019;6:264. https://doi.org/10.23937/2378-3656/1410264.
[140] Sabzevari R, Shahri A, Fasih AR, Masoumzadeh S, Rezaei Ghahroudi M. Object detection and localization system based on neural networks for robo-pong. In: Proceeding of the 5th international symposium on mechatronics and its applications, ISMA 2008; 2008. https://doi.org/10.1109/ISMA.2008.4648837.
[141] Salakhutdinov R, Tenenbaum JB, Torralba A. Learning with hierarchical-deep models. IEEE Trans Pattern Anal Mach Intell 2012;35:1958–71. https://doi.org/10.1109/TPAMI.2012.269.
[142] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Inf Process Manag 1988;24:513–23. https://doi.org/10.1016/0306-4573(88)90021-0.
[143] Sariyildiz MB, Cinbis RG. Gradient matching generative networks for zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2019. p. 2168–78. https://doi.org/10.1109/CVPR.2019.00227.
[144] Schölkopf B, Herbrich R, Smola AJ. A generalized representer theorem. In: International conference on computational learning theory, Springer; 2001. p. 416–26. https://doi.org/10.1007/3-540-44581-1_27.
[145] Schonfeld E, Ebrahimi S, Sinha S, Darrell T, Akata Z. Generalized zero-and few-shot learning via aligned variational autoencoders. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2019. p. 8247–55. https://doi.org/10.1109/CVPR.2019.00844.
[146] Shan F, Gao Y, Wang J, Shi W, Shi N, Han M, Xue Z, Shi Y. Lung infection quantification of covid-19 in ct images with deep learning. 2020. arXiv preprint arXiv:2003.04655.
[147] Sharmanska V, Quadrianto N, Lampert CH. Augmented attribute representations. In: European conference on computer vision, Springer; 2012. p. 242–55. https://doi.org/10.1007/978-3-642-33715-4_18.
[148] Shechtman E, Irani M. Matching local self-similarities across images and videos. In: 2007 IEEE conference on computer vision and pattern recognition, IEEE; 2007. p. 1–8. https://doi.org/10.1109/CVPR.2007.383198.
[149] Shen L, Yeung S, Hoffman J, Mori G, Fei-Fei L. Scaling human-object interaction recognition through zero-shot learning. In: 2018 IEEE winter conference on applications of computer vision (WACV), IEEE; 2018. p. 1568–76. https://doi.org/10.1109/WACV.2018.00181.
[150] Shen Y, Liu L, Shen F, Shao L. Zero-shot sketch-image hashing. In: Proceedings of the IEEE conference on computer vision and pattern recognition,; 2018. p. 3598–607. https://doi.org/10.1109/CVPR.2018.00379.

[151] Sheng L, Lin Z, Shao J, Wang X. Avatar-net: multi-scale zero-shot style transfer by feature decoration. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 8242–50. https://doi.org/10.1109/CVPR.2018.00860.
[152] Shi H, Han X, Jiang N, Cao Y, Alwalid O, Gu J, Fan Y, Zheng C. Radiological findings from 81 patients with covid-19 pneumonia in wuhan, China: a descriptive study. The Lancet Infectious Diseases. 2020.
[153] Shibing X, Zishu G. Bi-semantic reconstructing generative network for zero-shot learning. 2019. arXiv preprint arXiv:1912.03877.
[154] Shocher A, Cohen N, Irani M. "zero-shot" super-resolution using deep internal learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 3118–26. https://doi.org/10.1109/CVPR.2018.00329.
[155] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. arXiv preprint arXiv:1409.1556.
[156] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. In: Advances in neural information processing systems; 2017. p. 4077–87. https://dl.acm.org/doi/10.5555/3294996.3295163.
[157] Socher R, Ganjoo M, Manning CD, Ng A. Zero-shot learning through cross-modal transfer. In: Advances in neural information processing systems; 2013. p. 935–43. https://dl.acm.org/doi/10.5555/2999611.2999716.
[158] Sohn K, Lee H, Yan X. Learning structured output representation using deep conditional generative models. In: Advances in neural information processing systems; 2015. p. 3483–91. https://dl.acm.org/doi/10.5555/2969442.2969628.
[159] Song J, Shen C, Yang Y, Liu Y, Song M. Transductive unbiased embedding for zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 1024–33. https://doi.org/10.1109/CVPR.2018.00113.
[160] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 1–9. https://doi.org/10.1109/CVPR.2015.7298594.
[161] Teimouri M, Delavaran MH, Rezaei M. A real-time ball detection approach using convolutional neural networks. In: The 23rd annual RoboCup international symposium; 2019. https://doi.org/10.1007/978-3-030-35699-6_25.
[162] Tong B, Klinkigt M, Chen J, Cui X, Kong Q, Murakami T, Kobayashi Y. Adversarial zero-shot learning with semantic augmentation. In: Thirty-second AAAI conference on artificial intelligence; 2018.
[163] Torralba A, Murphy KP, Freeman WT. Shared features for multiclass object detection. 2006. p. 345–61. https://doi.org/10.1007/11957959_18.
[164] Tsai YHH, Huang LK, Salakhutdinov R. Learning robust visual-semantic embeddings. In: 2017 IEEE international conference on computer vision (ICCV), IEEE; 2017. p. 3591–600. https://doi.org/10.1109/ICCV.2017.386.
[165] Tsai YHH, Salakhutdinov R. Improving one-shot learning through fusing side information. 2017. arXiv preprint arXiv:1710.08347.
[166] Tsochantaridis I, Joachims T, Hofmann T, Altun Y. Large margin methods for structured and interdependent output variables. J Mach Learn Res 2005;6:1453–84. =10.1.1.92.6373.
[167] Usunier N, Buffoni D, Gallinari P. Ranking with ordered weighted pairwise classification. In: Proceedings of the 26th annual international conference on machine learning, ACM; 2009. p. 1057–64. https://doi.org/10.1145/1553374.1553509.
[168] Van Horn G, Branson S, Farrell R, Haber S, Barry J, Ipeirotis P, Perona P, Belongie S. Building a bird recognition app and large scale dataset with citizen scientists: the fine print in fine-grained dataset collection. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 595–604. https://doi.org/10.1109/CVPR.2015.7298658.
[169] Varshni D, Thakral K, Agarwal L, Nijhawan R, Mittal A. Pneumonia detection using cnn based feature extraction. In: 2019 IEEE international conference on electrical, computer and communication technologies (ICECCT), IEEE; 2019. p. 1–7. https://doi.org/10.1155/2019/4180949.
[170] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Advances in neural information processing systems; 2017. p. 5998–6008. https://dl.acm.org/doi/10.5555/3295222.3295349.
[171] Verma VK, Rai P. A simple exponential family framework for zero-shot learning. In: Joint European conference on machine learning and knowledge discovery in databases. Springer; 2017. p. 792–808. https://doi.org/10.1007/978-3-319-71246-8_48.
[172] Vinyals O, Blundell C, Lillicrap T, Wierstra D, Others. Matching networks for one shot learning. In: Advances in neural information processing systems; 2016. p. 3630–8. https://doi.org/10.7551/mitpress/7503.001.0001.
[173] Wah C, Branson S, Welinder P, Perona P, Belongie S. The caltech-ucsd birds-200-2011 dataset. California Institute of Technology; 2011.
[174] Wan Z, Chen D, Li Y, Yan X, Zhang J, Yu Y, Liao J. Transductive zero-shot learning with visual structure constraint. 2019. arXiv preprint arXiv:1901.01570.
[175] Wang D, Li Y, Lin Y, Zhuang Y. Relational knowledge transfer for zero-shot learning. In: Thirtieth AAAI conference on artificial intelligence; 2016. https://dl.acm.org/doi/10.5555/3016100.3016198.
[176] Wang Q, Chen K. Zero-shot visual recognition via bidirectional latent embedding. Int J Comput Vis 2017;124:356–83. https://doi.org/10.1007/s11263-017-1027-5.
[177] Wang W, Lu X, Shen J, Crandall DJ, Shao L. Zero-shot video object segmentation via attentive graph neural networks. In: Proceedings of the IEEE international conference on computer vision; 2019. p. 9236–45. https://doi.org/10.1109/ICCV.2019.00933.
[178] Wang W, Zheng VW, Yu H, Miao C. A survey of zero-shot learning: settings, methods, and applications. ACM Trans Intell Syst Technol (TIST) 2019b;10:1–37. https://doi.org/10.1145/3293318.
[179] Wang X, Ji Q. A unified probabilistic approach modeling relationships between attributes and objects. In: Proceedings of the IEEE international conference on computer vision; 2013. p. 2120–7. https://doi.org/10.1109/ICCV.2013.264.
[180] Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 2097–106. https://doi.org/10.1109/CVPR.2017.369.
[181] Wang X, Ye Y, Gupta A. Zero-shot recognition via semantic embeddings and knowledge graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 6857–66. https://doi.org/10.1109/CVPR.2018.00717.
[182] Wen Y, Zhang K, Li Z, Qiao Y. A discriminative feature learning approach for deep face recognition. European conference on computer vision. Springer; 2016. p. 499–515. https://doi.org/10.1007/978-3-319-46478-7_31.
[183] Weston J, Bengio S, Usunier N. Large scale image annotation: learning to rank with joint word-image embeddings. Mach Learn 2010;81:21–35. https://doi.org/10.1007/s10994-010-5198-3.
[184] Xia W, Shao J, Guo Y, Peng X, Li Z, Hu D. Clinical and ct features in pediatric patients with covid-19 infection: different points from adults. Pediatr Pulmonol 2020;55:1169–74. https://doi.org/10.1002/ppul.24718.
[185] Xian Y, Akata Z, Sharma G, Nguyen Q, Hein M, Schiele B. Latent embeddings for zero-shot classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 69–77. https://doi.org/10.1109/CVPR.2016.15.
[186] Xian Y, Lampert CH, Schiele B, Akata Z. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. In: IEEE transactions on pattern analysis and machine intelligence; 2018. https://doi.org/10.1109/TPAMI.2018.2857768.
[187] Xian Y, Lorenz T, Schiele B, Akata Z. Feature generating networks for zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 5542–51. https://doi.org/10.1109/CVPR.2018.00581.
[188] Xian Y, Schiele B, Akata Z. Zero-shot learning-the good, the bad and the ugly. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 4582–91. https://doi.org/10.1109/CVPR.2017.328.
[189] Xian Y, Sharma S, Schiele B, Akata Z. f-vaegan-d2: a feature generating framework for any-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2019. p. 10275–84. https://doi.org/10.1109/CVPR.2019.01052.
[190] Xie GS, Liu L, Jin X, Zhu F, Zhang Z, Qin J, Yao Y, Shao L. Attentive region embedding network for zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2019. p. 9384–93. https://doi.org/10.1109/CVPR.2019.00961.
[191] Xie Y, Xu P, Ma Z. Deep zero-shot learning for scene sketch. In: 2019 IEEE international conference on image processing (ICIP), IEEE; 2019. p. 3661–5. https://doi.org/10.1109/ICIP.2019.8803426.
[192] Xu X, Shen F, Yang Y, Shao J, Huang Z. Transductive visual-semantic embedding for zero-shot learning. In: Proceedings of the 2017 ACM on international conference on multimedia retrieval. ACM; 2017a. p. 41–9. https://doi.org/10.1145/3078971.3078977.
[193] Xu X, Shen F, Yang Y, Zhang D, Tao Shen H, Song J. Matrix tri-factorization with manifold regularizations for zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 3798–807. https://doi.org/10.1109/CVPR.2017.217.
[194] Xu Y, Yang Y, Shen F, Xu X, Zhou Y, Shen HT. Attribute hashing for zero-shot image retrieval. In: 2017 IEEE international conference on multimedia and expo (ICME), IEEE; 2017. p. 133–8. https://doi.org/10.1109/ICME.2017.8019425.
[195] Yamada M, Sigal L, Raptis M, Toyoda M, Chang Y, Sugiyama M. Cross-domain matching with squared-loss mutual information. IEEE Trans Pattern Anal Mach Intell 2015;37:1764–76. https://doi.org/10.1109/TPAMI.2014.2388235.
[196] Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. Xlnet: generalized autoregressive pretraining for language understanding. In: Advances in neural information processing systems; 2019. p. 5753–63.
[197] Ye M, Guo Y. Zero-shot classification with discriminative semantic representation learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 7140–8. https://doi.org/10.1109/CVPR.2017.542.
[198] Ye Z, Zhang Y, Wang Y, Huang Z, Song B. Chest ct manifestations of new coronavirus disease 2019 (covid-19): a pictorial review. Eur Radiol 2020;1. https://doi.org/10.1148/radiol.2020200343.
[199] Yu X, Aloimonos Y. Attribute-based transfer learning for object categorization with zero/one training example. In: European conference on computer vision. Springer; 2010. p. 127–40. https://doi.org/10.1007/978-3-642-15555-0_10.
[200] Zhan C, She D, Zhao S, Cheng MM, Yang J. Zero-shot emotion recognition via affective structural embedding. In: Proceedings of the IEEE international conference on computer vision; 2019. p. 1151–60. https://doi.org/10.1109/ICCV.2019.00124.
[201] Zhang J, Xie Y, Li Y, Shen C, Xia Y. Covid-19 screening on chest x-ray images using deep learning based anomaly detection. 2020. arXiv preprint arXiv:2003.12338.
[202] Zhang L, Xiang T, Gong S. Learning a deep embedding model for zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 2021–30. https://doi.org/10.1109/CVPR.2017.321.
[203] Zhang Y, Miao S, Mansi T, Liao R. Unsupervised x-ray image segmentation with task driven generative adversarial networks. Med Image Anal 2020b;62:101664. https://doi.org/10.1016/j.media.2020.101664.

[204] Zhang Z, Saligrama V. Zero-shot learning via semantic similarity embedding. In: Proceedings of the IEEE international conference on computer vision; 2015. p. 4166–74. https://doi.org/10.1109/ICCV.2015.474.
[205] Zhang Z, Saligrama V. Zero-shot learning via joint latent similarity embedding. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 6034–42. https://doi.org/10.1109/CVPR.2016.649.
[206] Zhang Z, Saligrama V. Zero-shot recognition via structured prediction. In: European conference on computer vision. Springer; 2016b. p. 533–48. https://doi.org/10.1007/978-3-319-46478-7_33.
[207] Zhao B, Wu B, Wu T, Wang Y. Zero-shot learning posed as a missing data problem. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 2616–22. https://doi.org/10.1109/ICCVW.2017.310.
[208] Zheng C. Time course of lung changes at chest ct during recovery from coronavirus disease 2019 (covid-19). Radiology 2020;295:715–21. https://doi.org/10.1148/radiol.2020200370.
[209] Zhu P, Wang H, Saligrama V. Dont even look once: synthesizing features for zero-shot detection. 2019. arXiv preprint arXiv:1911.07933.
[210] Zhu P, Wang H, Saligrama V. Generalized zero-shot recognition based on visually semantic embedding. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2019. p. 2995–3003. https://doi.org/10.1109/CVPR.2019.00311.
[211] Zhu Y, Elhoseiny M, Liu B, Peng X, Elgammal A. A generative adversarial approach for zero-shot learning from noisy texts. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 1004–13. https://doi.org/10.1109/CVPR.2018.00111.
[212] Zhu Y, Kiros R, Zemel R, Salakhutdinov R, Urtasun R, Torralba A, Fidler S. Aligning books and movies: towards story-like visual explanations by watching movies and reading books. In: Proceedings of the IEEE international conference on computer vision; 2015. p. 19–27. https://doi.org/10.1109/ICCV.2015.11.